Supplementary Information for:

Genomic signatures of adaptation in native lizards exposed to human-introduced fire ants

Braulio A. Assis*#[1,2], Alexis P. Sullivan*#[1,3], Stephanie Marciniak[2,4], Christina M. Bergey[1,2,5], Vanessa Garcia[1], Zachary A. Szpiech[1,6], Tracy Langkilde*[1,7], George H. Perry*[1,2,7]

[1] Department of Biology, Penn State University, University Park, PA, USA

[2] Department of Anthropology, Penn State University, University Park, PA, USA

[3] Institute for Systems Genetics, NYU Langone Health, New York City, NY, USA

[4] Department of Anthropology, McMaster University, Hamilton, Ontario, Canada, L8S 4L9

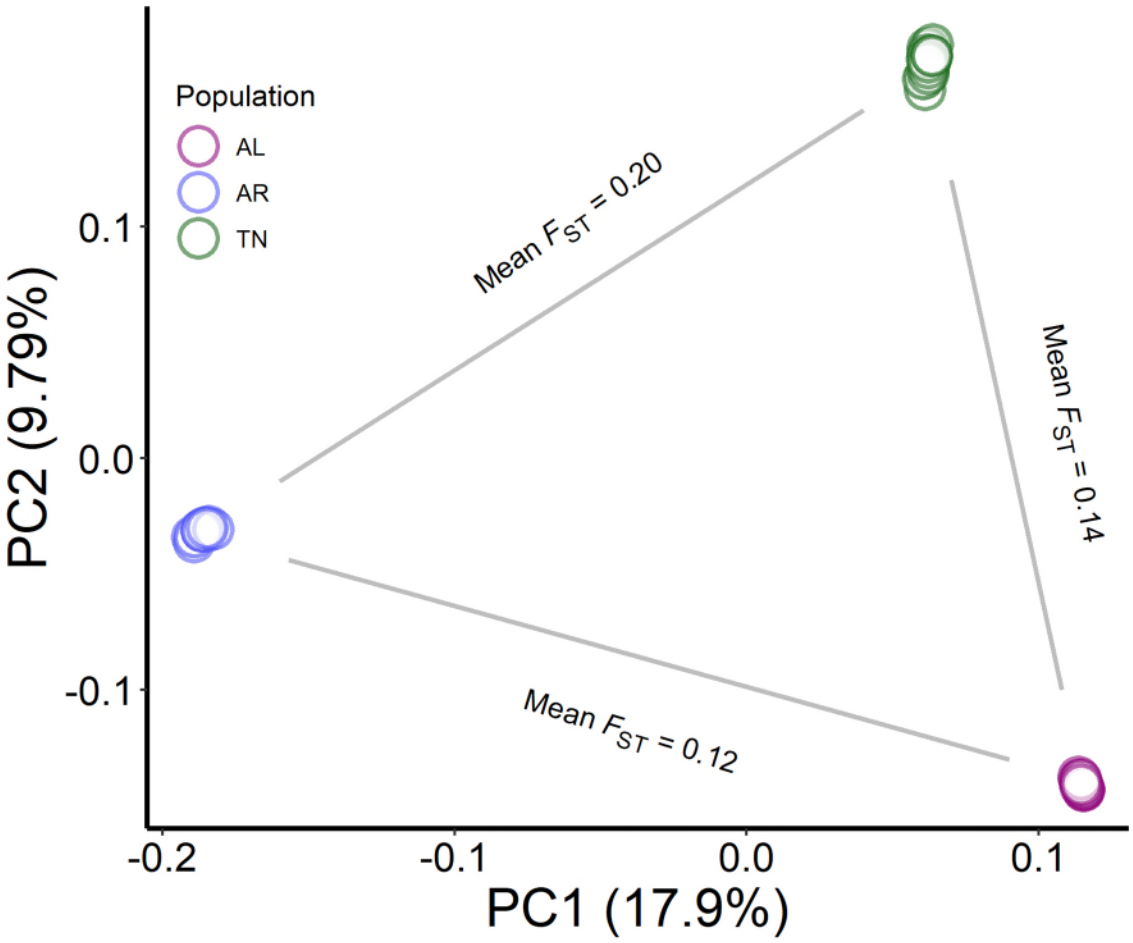
[5] Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[6] Institute for Computational and Data Sciences, Penn State University, University Park, PA, USA

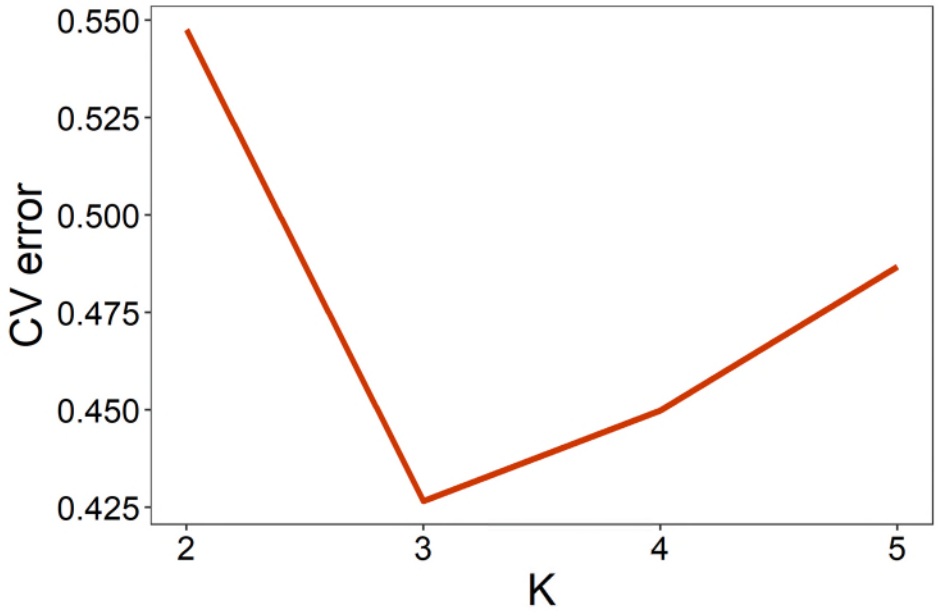
[7] Huck Institutes of the Life Sciences, Penn State University, University Park, PA USA

*bmd5458@psu.edu; alexispsullivan@gmail.com; tll30@psu.edu; ghp3@psu.edu

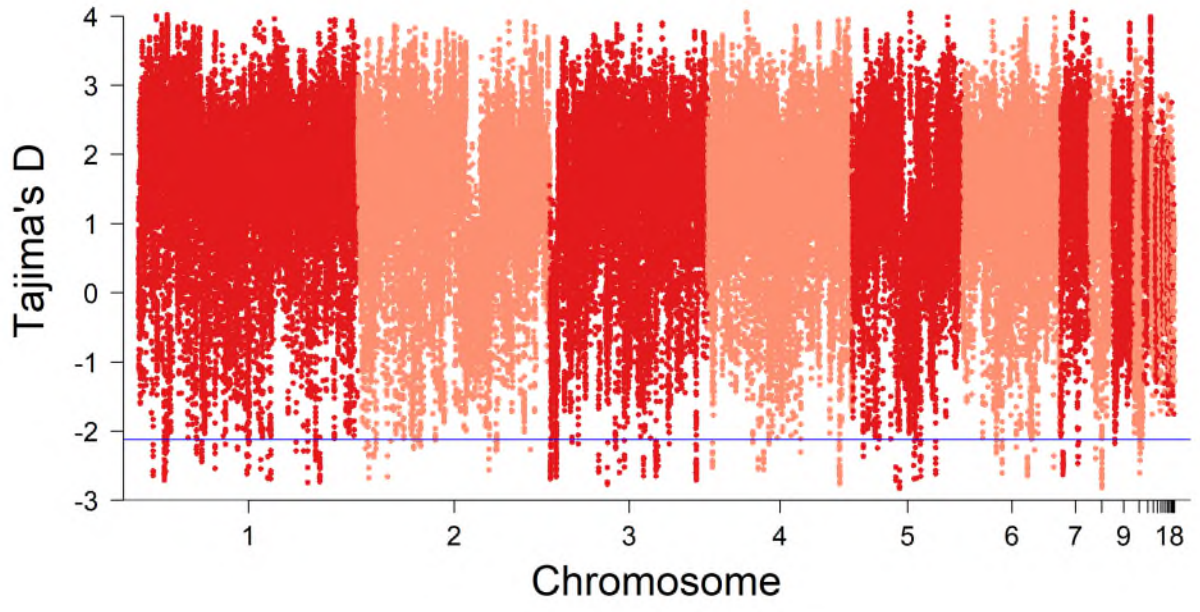
# Authors contributed equally



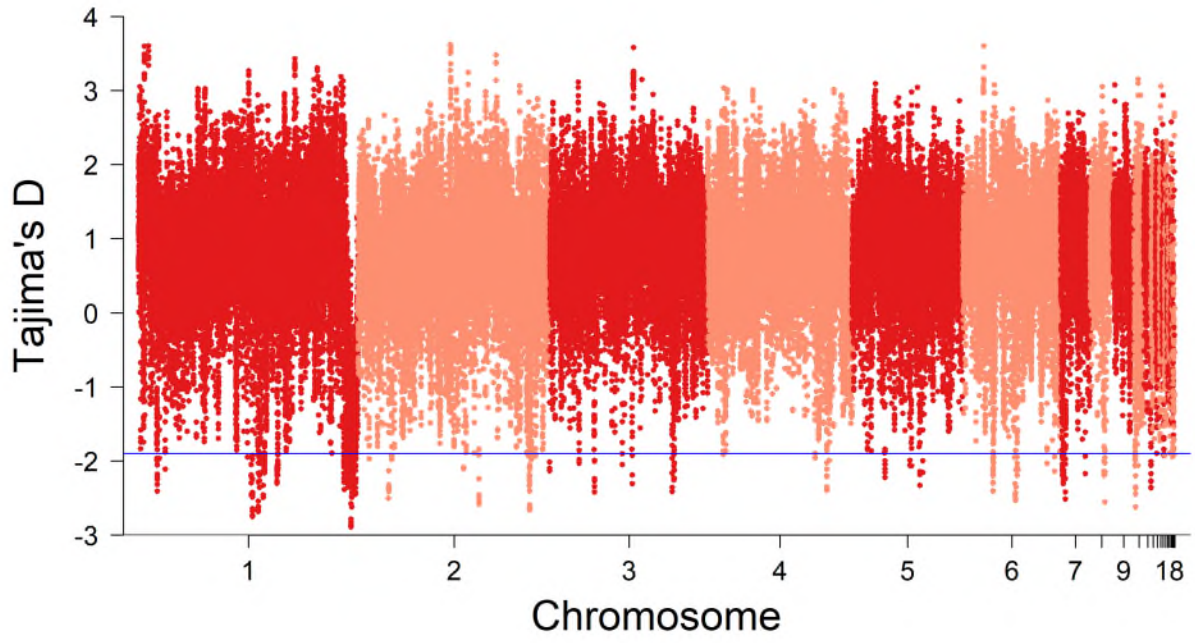
Supplementary figure 1: Principal Components Analysis and mean Weir & Cockerham F_{ST} for the three pairwise population comparisons based on the full set of 46,934,027 filtered SNPs (see Methods). Source data are provided as a Source Data file.



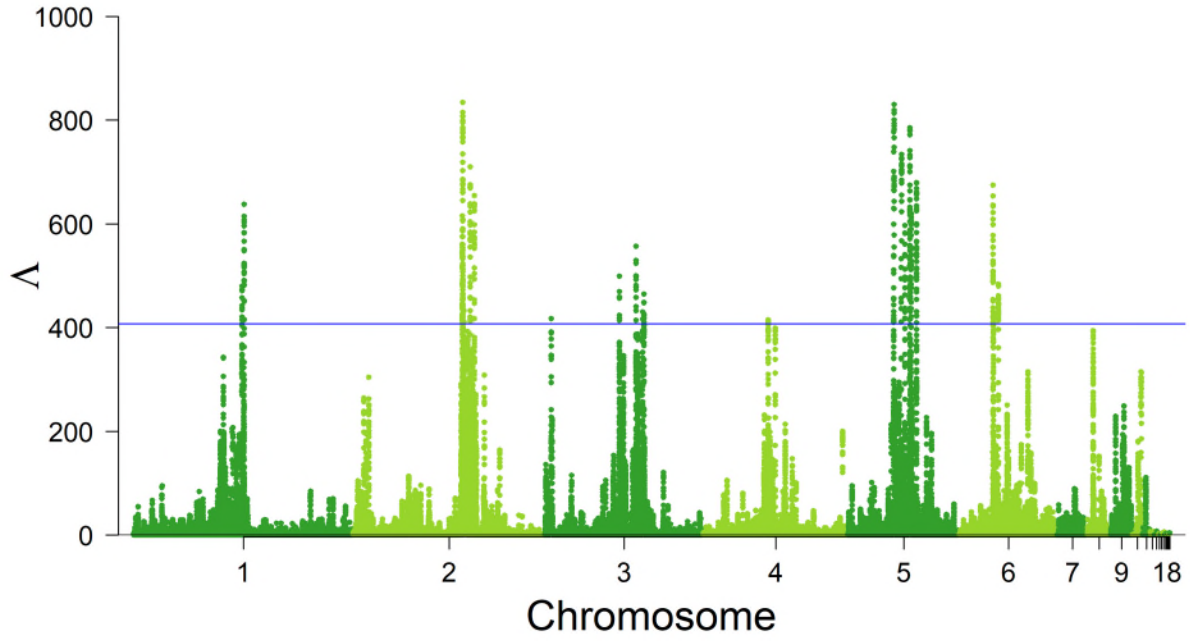
Supplementary figure 2: Cross-validation errors for admixture analysis with for K=1 through K=5 specified ancestral populations. Source data are provided as a Source Data file.



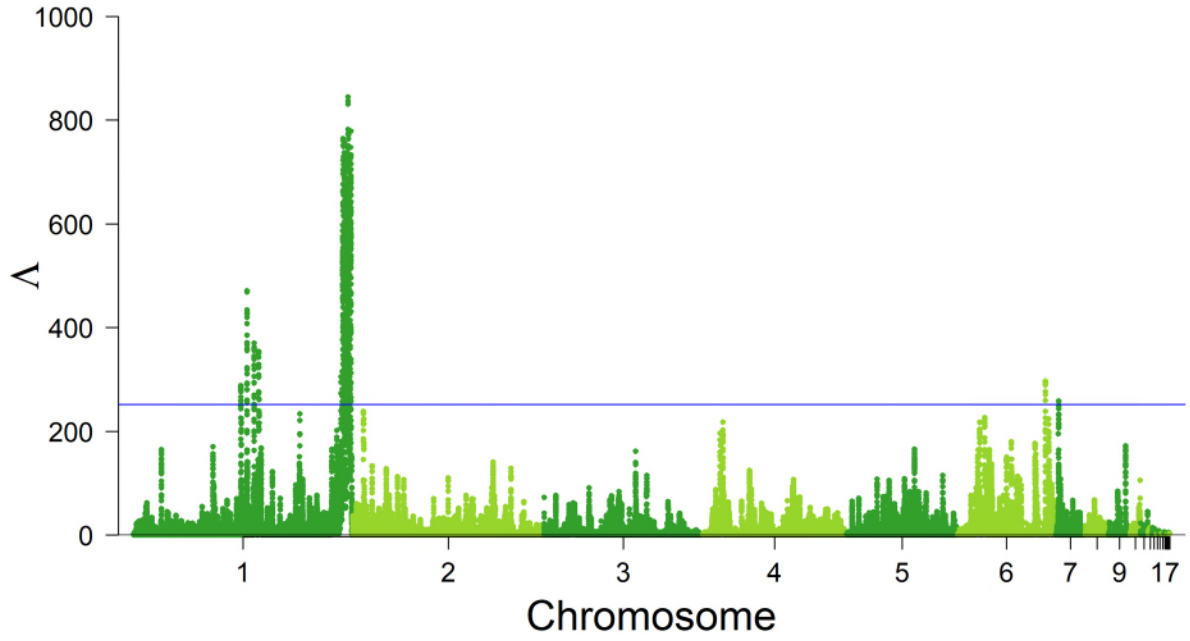
Supplementary figure 3: Genome-wide Tajima's D for the Arkansas population in 100 Kb windows and a 20 Kb step. Windows significantly under positive selection fall below the 0.5th percentile of the genome-wide distribution ($D < -2.11$, blue line). Source data are provided as a Source Data file.



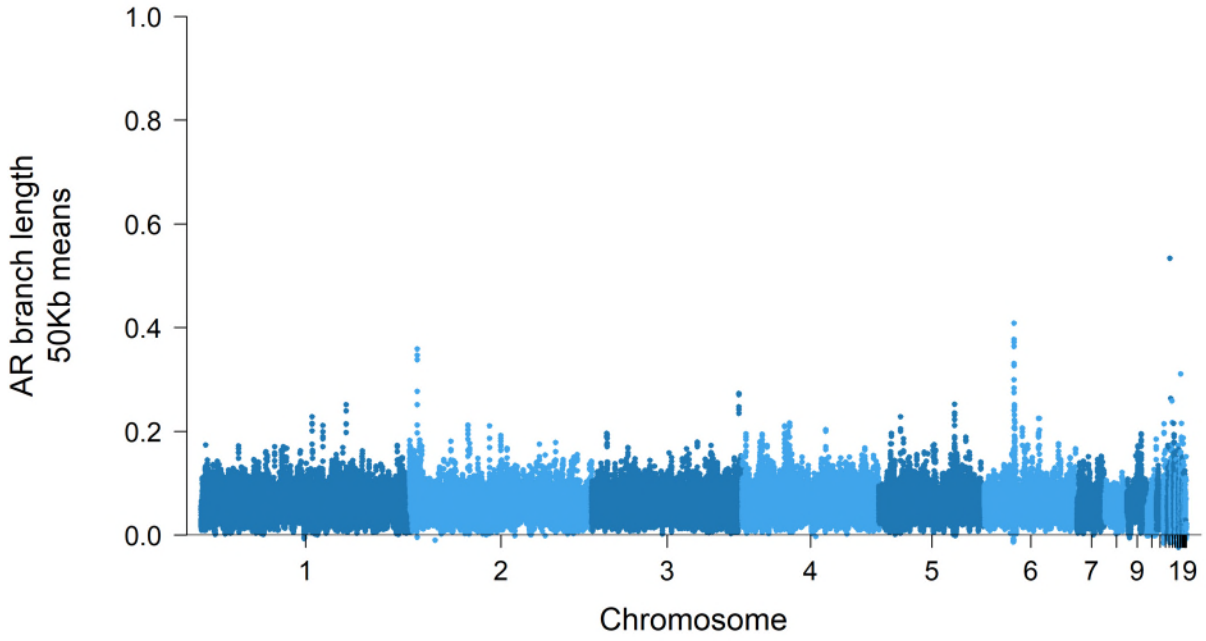
Supplementary figure 4: Genome-wide Tajima's D for the Tennessee population in 100 Kb windows and a 20 Kb step. Windows significantly under positive selection fall below the 0.5th percentile of the genome-wide distribution (D < -1.9, blue line). Source data are provided as a Source Data file.



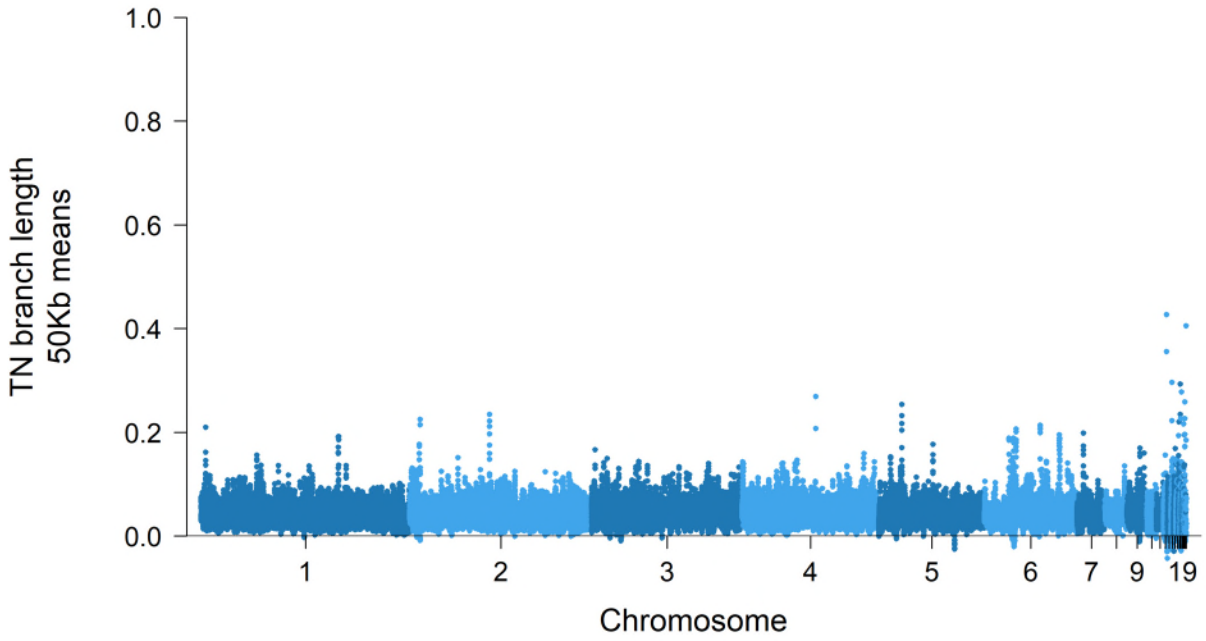
Supplementary figure 5: Genome-wide saltilassi's Λ statistic for the Arkansas population. Haplotypes significantly under positive selection fall above the 99.5th percentile of the genome-wide distribution (Λ > 407.56, blue line). Source data are provided as a Source Data file.



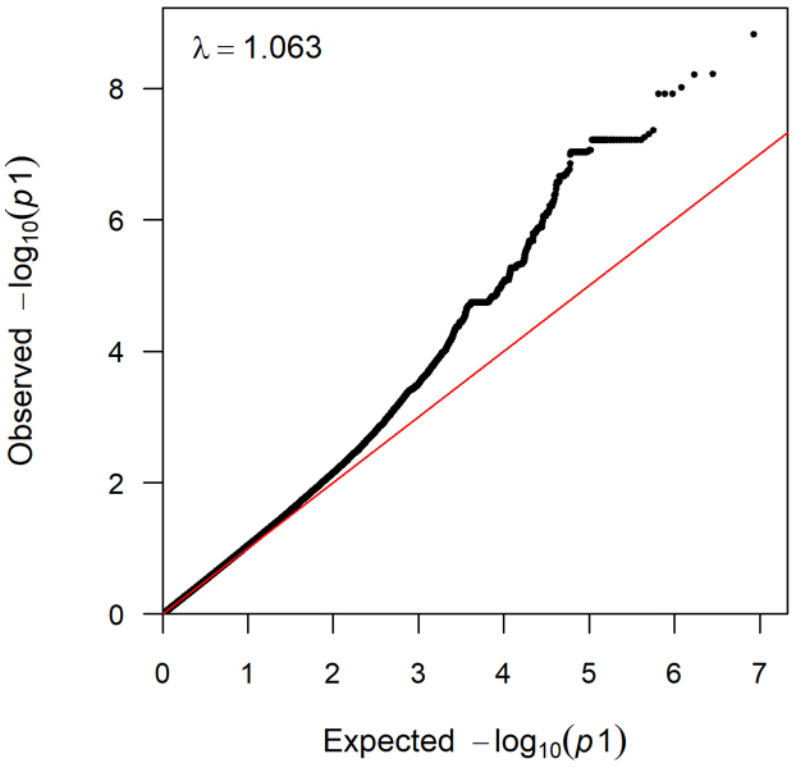
Supplementary figure 6: Genome-wide saltilassi's Λ statistic for the Tennessee population. Haplotypes significantly under positive selection fall above the 99.5th percentile of the genome-wide distribution (Λ > 252.27, blue line). Source data are provided as a Source Data file.



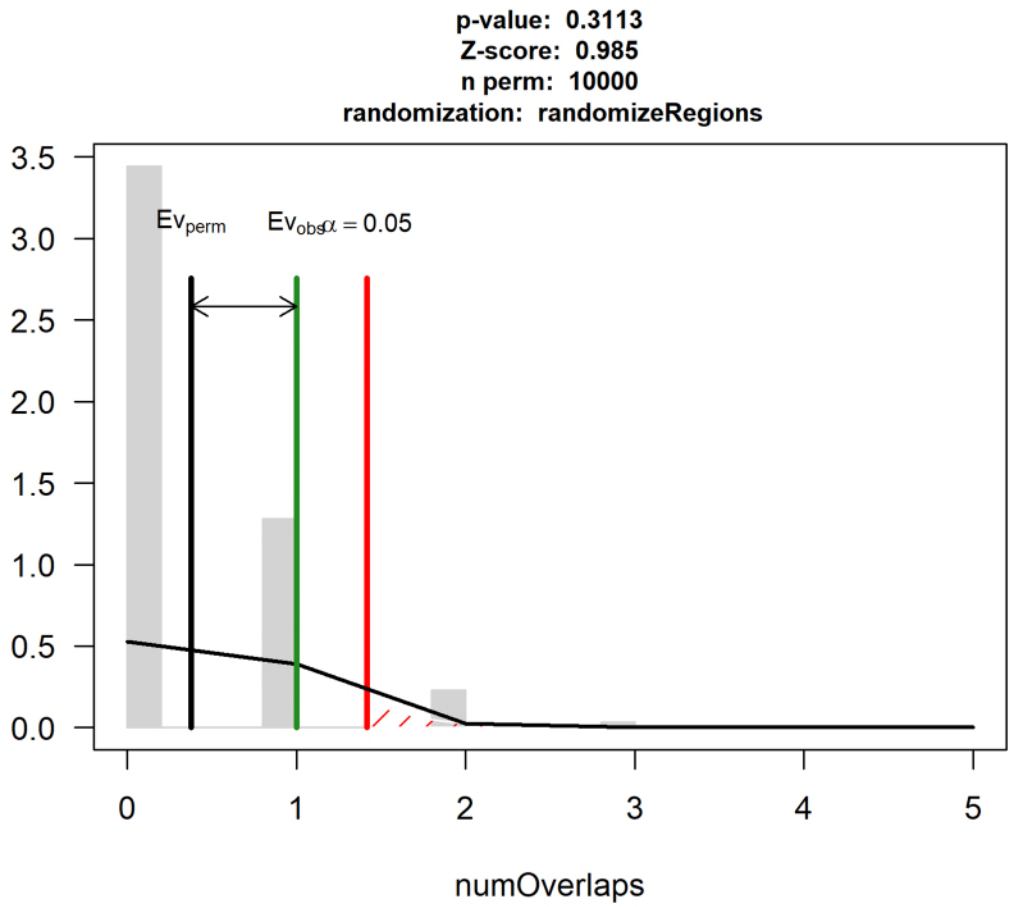
Supplementary figure 7: Locus-specific branch length for the Arkansas population (AR-AL F_{ST} + AR-TN F_{ST} – TN-AL F_{ST}) in 50Kb means and a 10Kb step. Source data are provided as a Source Data file.



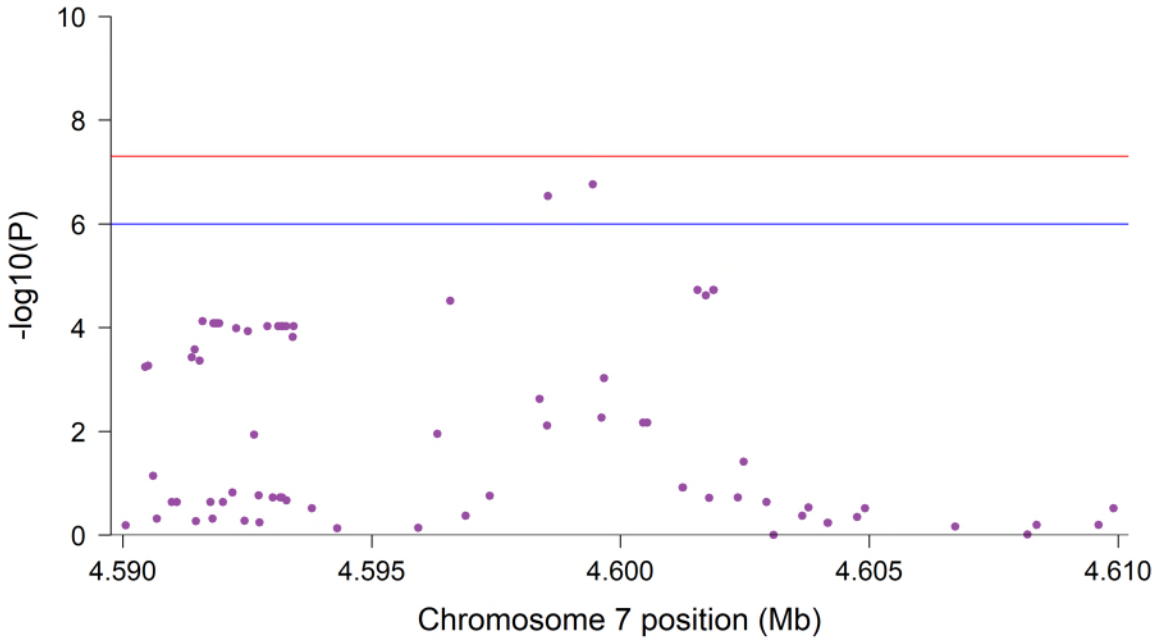
Supplementary figure 8: Locus-specific branch length for the Tennessee population (TN-AL F_{ST} + TN-AR F_{ST} – AR-AL F_{ST}) in 50Kb means and a 10Kb step. Source data are provided as a Source Data file.



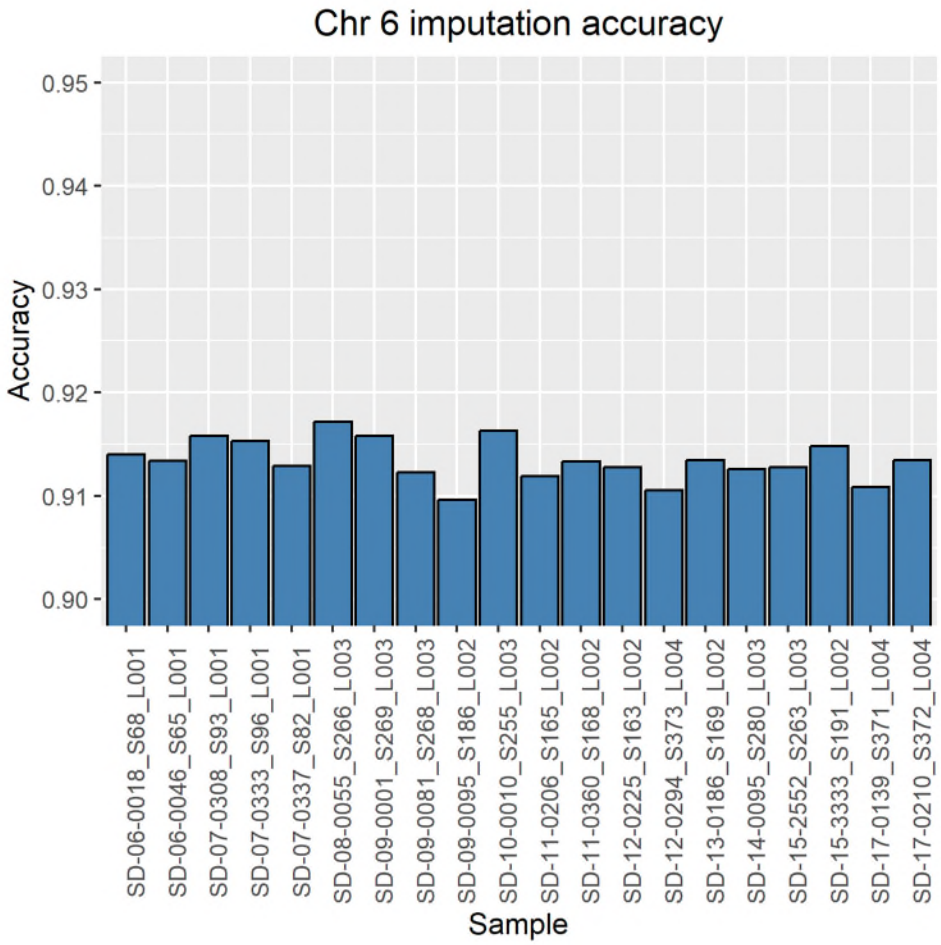
Supplementary figure 9: QQplot for genome-wide association for relative hind limb length. Covariates for each model were snout-to-vent length, sex, and Principal Components 1 through 4 for a population genetic relatedness Principal Components Analysis. Source data are provided as a Source Data file.



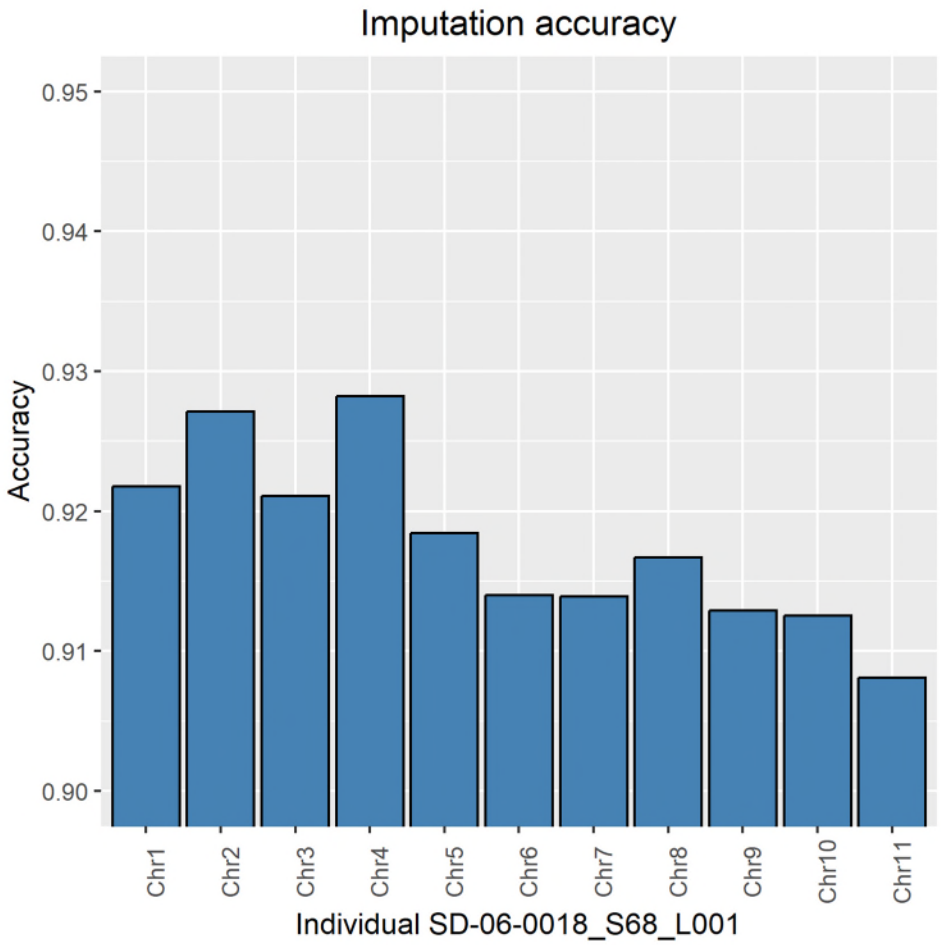
Supplementary figure 10: Permutation test (n=10,000) for number of overlaps between genomic regions significant for the Tajima's D statistic and genomic regions significantly associated with hind limb length versus the number of overlaps expected by chance. The number of overlaps was not significantly greater than the null distribution (p = 0.31). Source data are provided as a Source Data file.



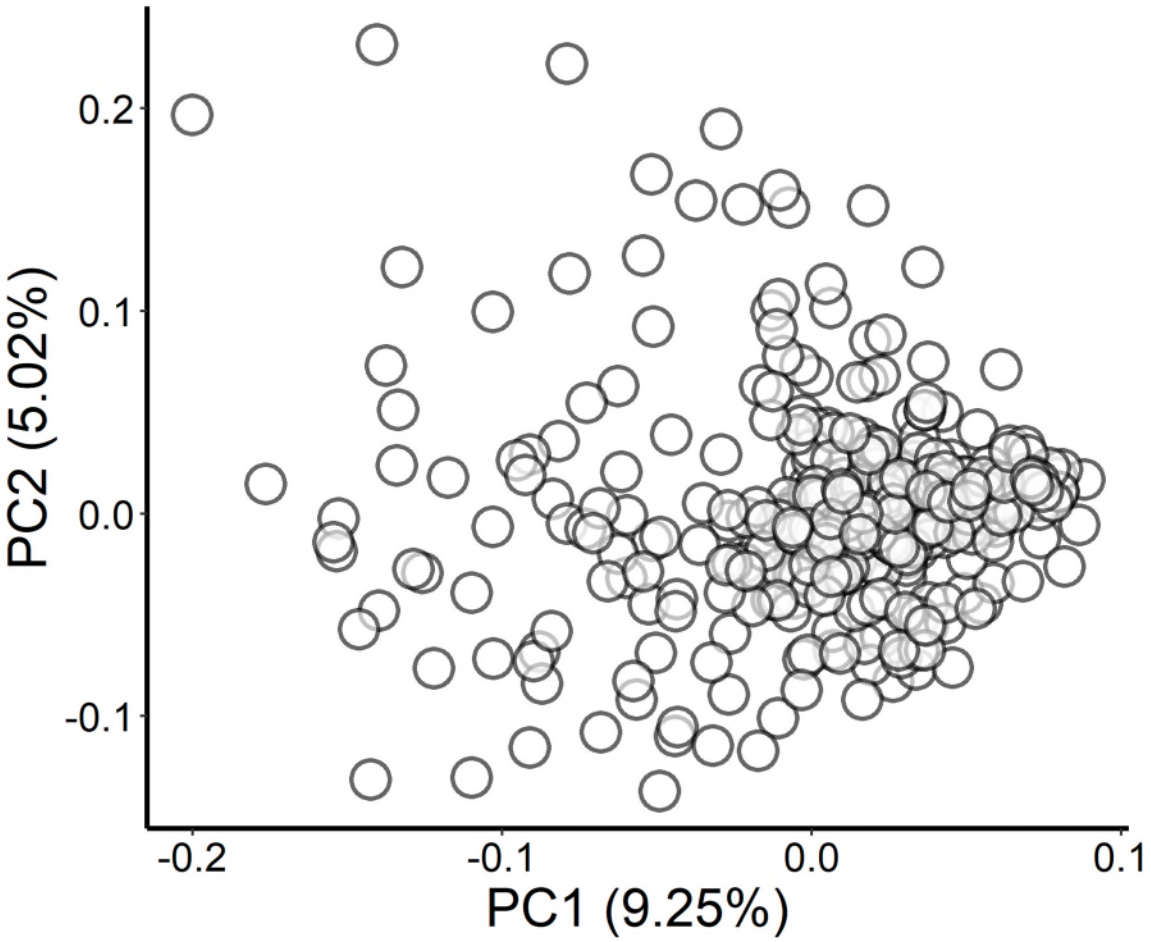
Supplementary figure 11: SNPs above the significance threshold of 1e-6 for the GWAS with hind limb length for n=381 individuals and overlapping the gene *PTPN11*. For the most significant SNP, $\beta \pm SE = -0.13 \pm 0.02$, $t = -5.22$, $p = 2.95e\text{-}07$. Source data are provided as a Source Data file.



Supplementary figure 12: For chromosome 6, and for each of 20 individuals, concordance rates for alleles sequenced at 13X, and imputed alleles based on a reference panel that did not contain that individual. Source data are provided as a Source Data file.



Supplementary figure 13: For chromosomes 1 through 11, concordance rates for alleles of an individual sequenced at 13X and imputed alleles based on a reference panel that did not contain that individual. Source data are provided as a Source Data file.



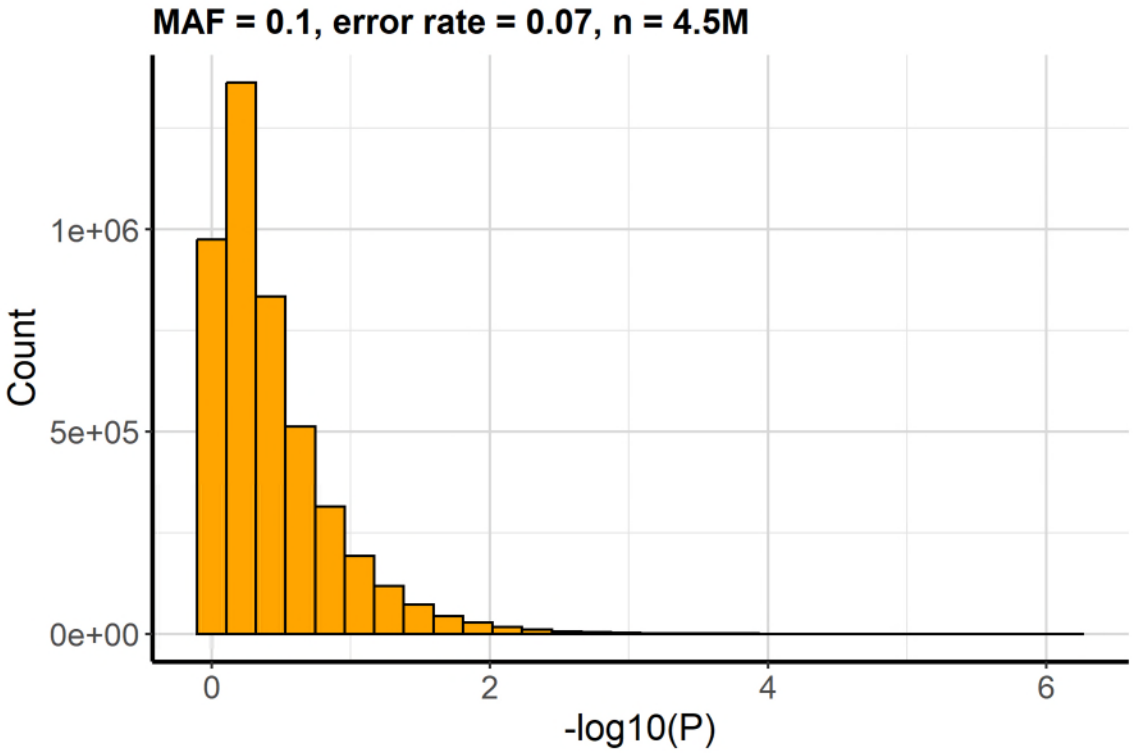
Supplementary figure 14: Principal Component Analysis for genetic similarity of 381 fence lizards from the AL population based on 458,533 SNPs genotyped at 4.39x. Source data are provided as a Source Data file.

Imputation error simulations

All simulations involved a randomly generated phenotype: a normally distributed variable with the same n, mean, and variance as relative limb length in our study. Additionally, for all simulations, we applied a genotyping error rate similar to that observed in the 100Kb window surrounding the Chr2: 129,424,718 locus, 7.29% (27 genotypes out of 381 in our sample).

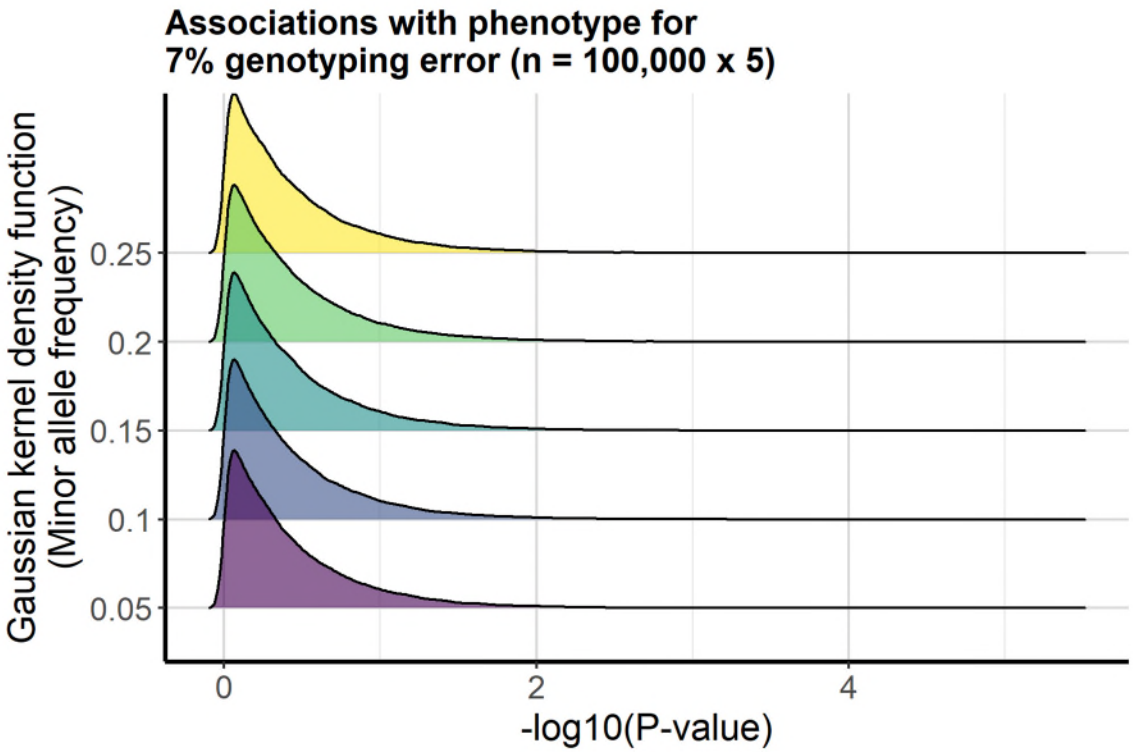
In our first set of simulations, we quantify the probability that this error rate would produce a false positive association with the phenotype in our study. We generated a population genotype distribution equivalent to the locus in question (minor allele frequency, MAF = 0.1). For each iteration, a new genotype distribution with MAF = 0.1 was generated, and 7% of the genotypes were randomly drawn and switched to another genotype. Finally, we tested the association between this new genotype distribution (with 7% error introduced) on our simulated phenotype. This process was repeated 4,500,000 times (compared to 4,245,544 SNPs tested in our study).

There was no occurrence of an association near our significance threshold ($p < 5e-8$). In 4,500,000 simulations, $p < 1e-5$ occurred 40 times, and $p < 1e-6$ occurred only once (6.8e-7).



Supplementary figure 15: For each of 4,500,000 simulations, a population with n=381 and randomly generated genotypes (minor allele frequency = 0.1) had 7% of genotypes randomly sampled and changed to another genotype. The association between each genotype distribution and a randomly generated phenotype with the same mean and variance as relative limb length in this study did not produce a significance level below the threshold for this study (< 5e-8). Source data are provided as a Source Data file.

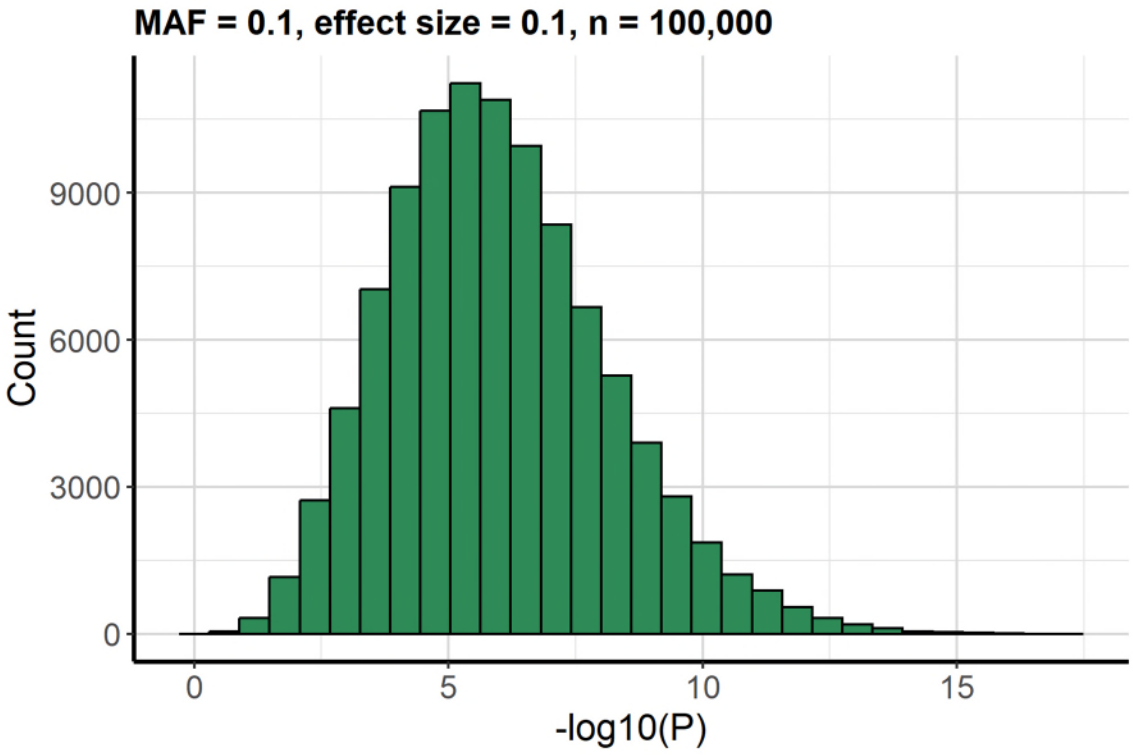
Next, we repeated this simulation but this time starting with different genotype distributions: MAF = 0.05, 0.1, 0.15, 0.2, and 0.25 (for each, n = 100,000). The smallest p-value observed across any of these simulations was p = 3.8e-6.



Supplementary figure 16: Density plots with distribution of p-values for the association between randomly generated genotype distributions of varying minor allele frequencies and the randomly generated phenotype (population size of n=381). For each of 100,000 iterations per MAF, 7% of genotypes were randomly sampled and changed to another genotype. Source data are provided as a Source Data file.

In summary, these simulations did not produce a false-positive result at the same significance as the genome-wide association study results reported in our study.

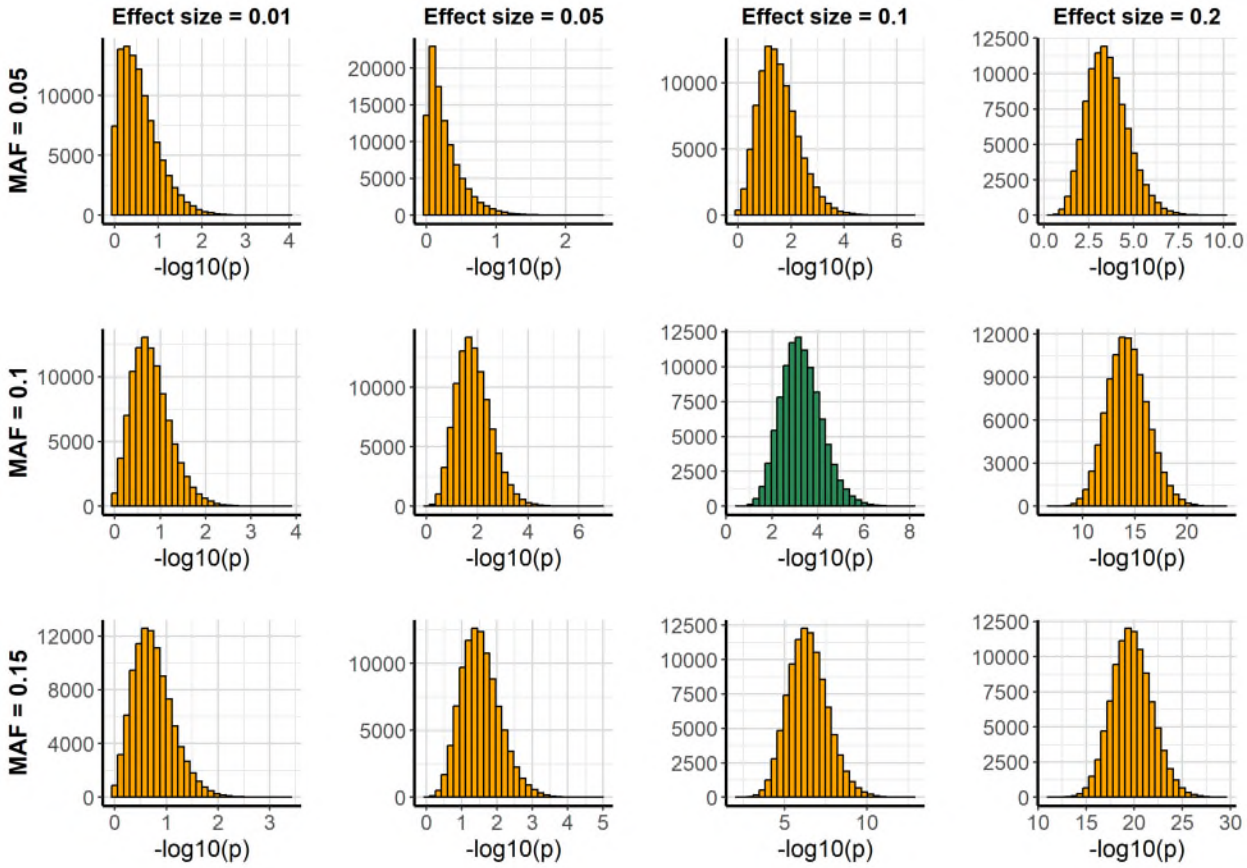
We then assessed the extent to which this imputation error rate may have produced false negative results. Here, we randomly generated a genotype distribution with the same allele frequency as the locus in question (MAF = 0.1), assigned to the alternate allele an effect size on the randomly generated phenotype equivalent to the locus in question (ß = 0.1), and then repeated this simulation 100,000 times. As expected, this allele distribution and effect size combination produced numerous significant association results among the permutations.



x

Supplementary figure 17: For each of 100,000 iterations, a population of size 381 had a randomly generated genotype distribution with a minor allele frequency of 0.1 and for which one allele had an assigned effect size of 0.1. With the assigned effect size, numerous significant associations ($P < 5e\text{-}8$) with the randomly generated phenotype occurred. Source data are provided as a Source Data file.

Then, we repeated this process with multiple genotype distributions and effect sizes, but now introducing 7% random error. While many significant associations do still occur among these permutations, there is an understandable shift in the p-value distribution, representing an increased proportion of false-negative associations.



Supplementary figure 18: With a population of size n=381, randomly generated genotype distributions with varying allele frequencies (MAF of 0.1 to 0.15) and allele effect sizes (0.01 to 0.2) had 7% of genotypes randomly sampled and assigned to a different genotype for 100,000 iterations. With the assigned effect sizes, significant associations with he randomly generated phenotype occurred despite the introduction of simulated errors. Source data are provided as a Source Data file.