Supplementary Material: Extending Tests of Hardy-Weinberg Equilibrium to Structured Populations

Wei Hao and John D. Storey*

Lewis-Sigler Institute for Integrative Genomics
Princeton University
Princeton, NJ 08544 USA

*Corresponding author: jstorey@princeton.edu

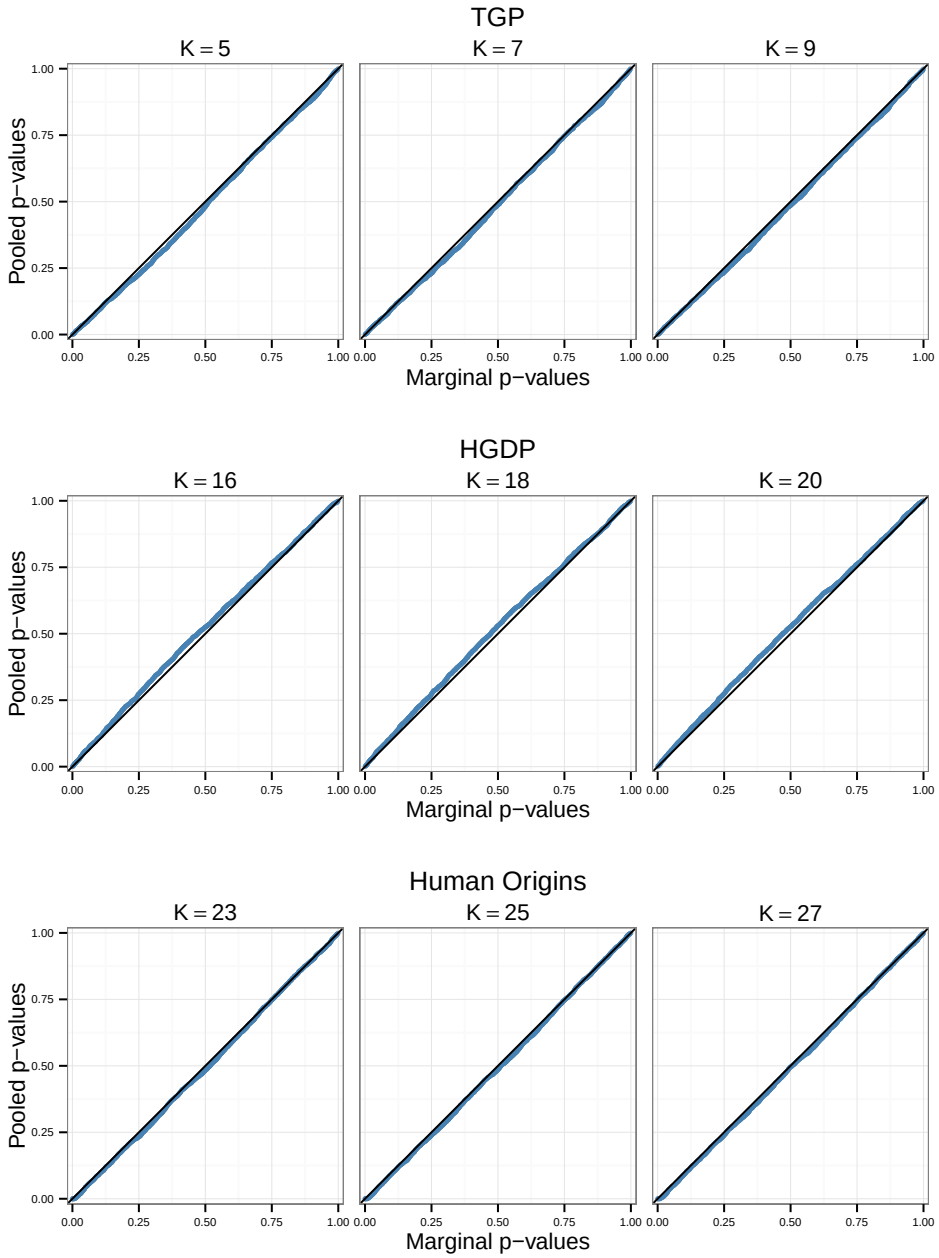

Figure S1: Quantile-quantile plots comparing p-values calculated using a SNP-specific marginal null distribution for each SNP versus the pooled null distribution for all three datasets. The quantile-quantile plots show a random sample of 5000 SNPs for each dataset to reduce clutter. The marginal null distribution was formed by simulating a new null dataset for each sample from the null distribution (10,000 simulated datasets total). The pooled null distribution was formed by pooling across all SNPs for 3 simulated null datasets.

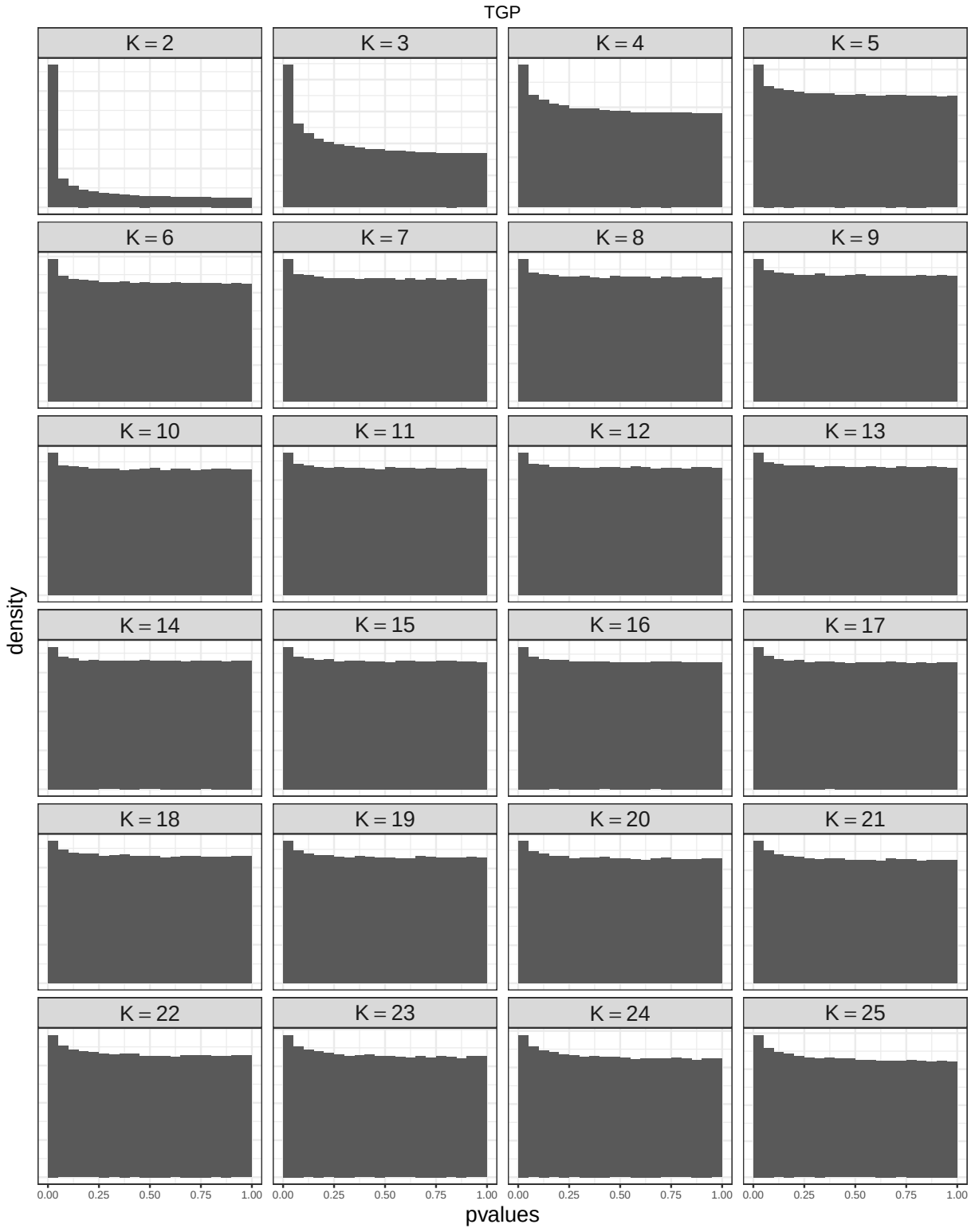

Figure S2: sHWE p-value histograms over a range of K for the TGP dataset. Note that for low K, population structure is only partially corrected for, resulting in a p-value distribution skewed towards zero. As K increases, skew towards zero reduces and the upper tail of the histogram becomes more Uniform(0,1), indicating that population structure is modeled more accurately resulting in minimal departures from sHWE.

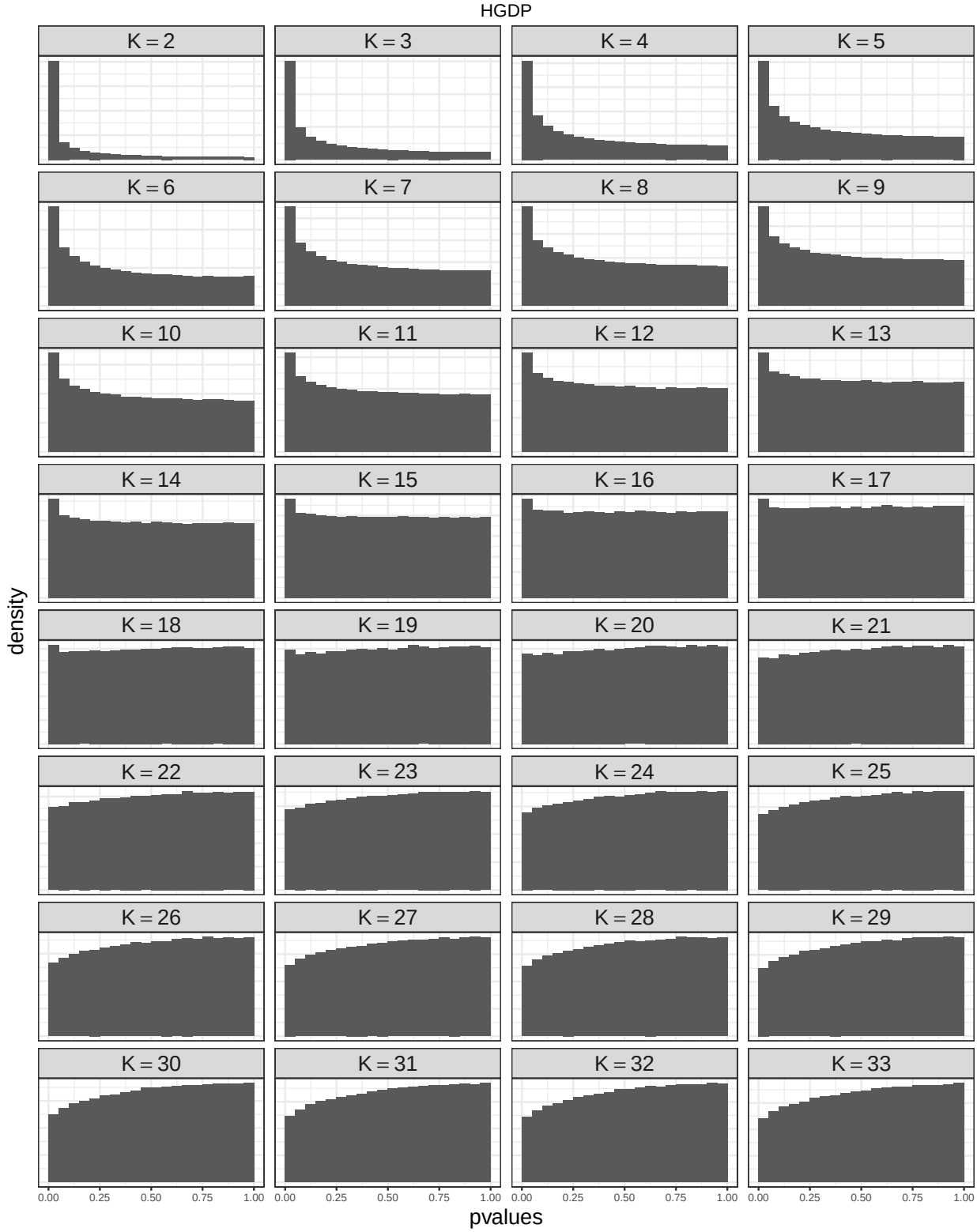

Figure S3: sHWE p-value histograms over a range of K for the HGDP dataset. Note that for low K, population structure is only partially corrected for, resulting in a p-value distribution skewed towards zero. As K increases, skew towards zero reduces and the upper tail of the histogram becomes more uniform, indicating that population structure is being better modeled.

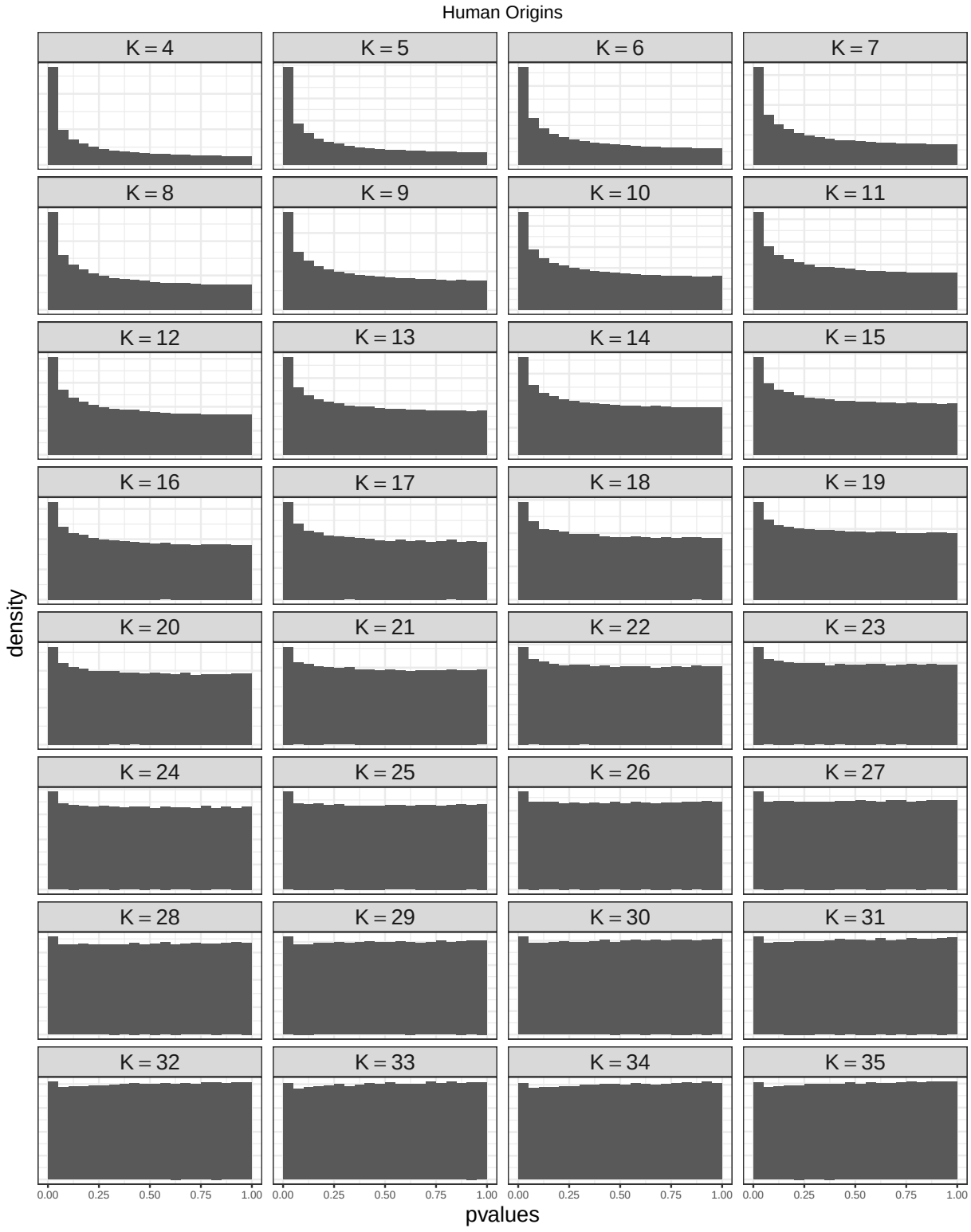

Figure S4: sHWE p-value histograms over a range of K for the Human Origins dataset. Note that for low K, population structure is only partially corrected for, resulting in a p-value distribution skewed towards zero. As K increases, skew towards zero reduces and the upper tail of the histogram becomes more uniform, indicating that population structure is being better modeled.

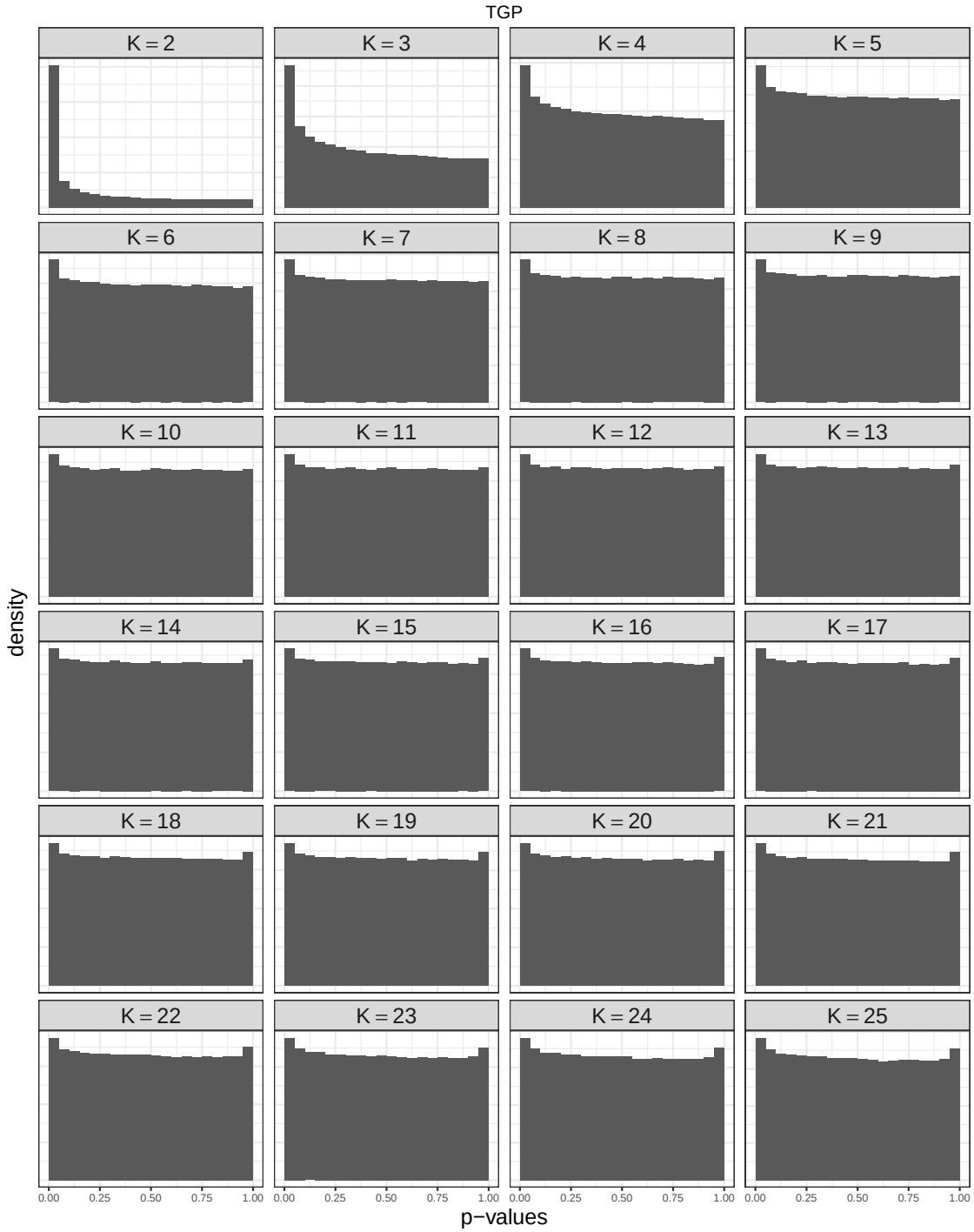

Figure S5: Structural HWE p-values for the TGP dataset using the truncated PCA approach to estimate allele frequencies. This is analogous to Figure S2.

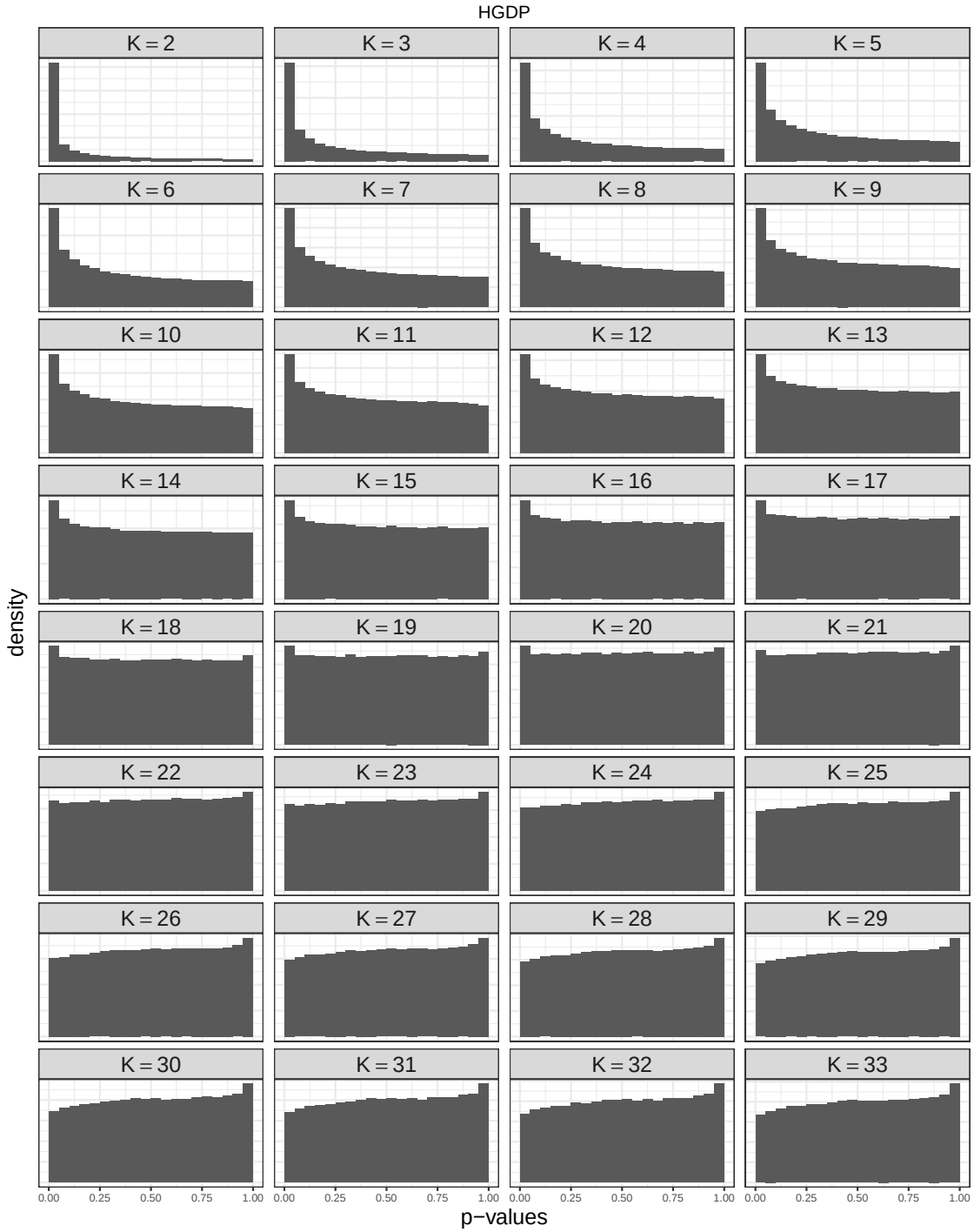

Figure S6: Structural HWE p-values for the HGDP dataset using the truncated PCA approach to estimate allele frequencies. This is analogous to Figure S3.

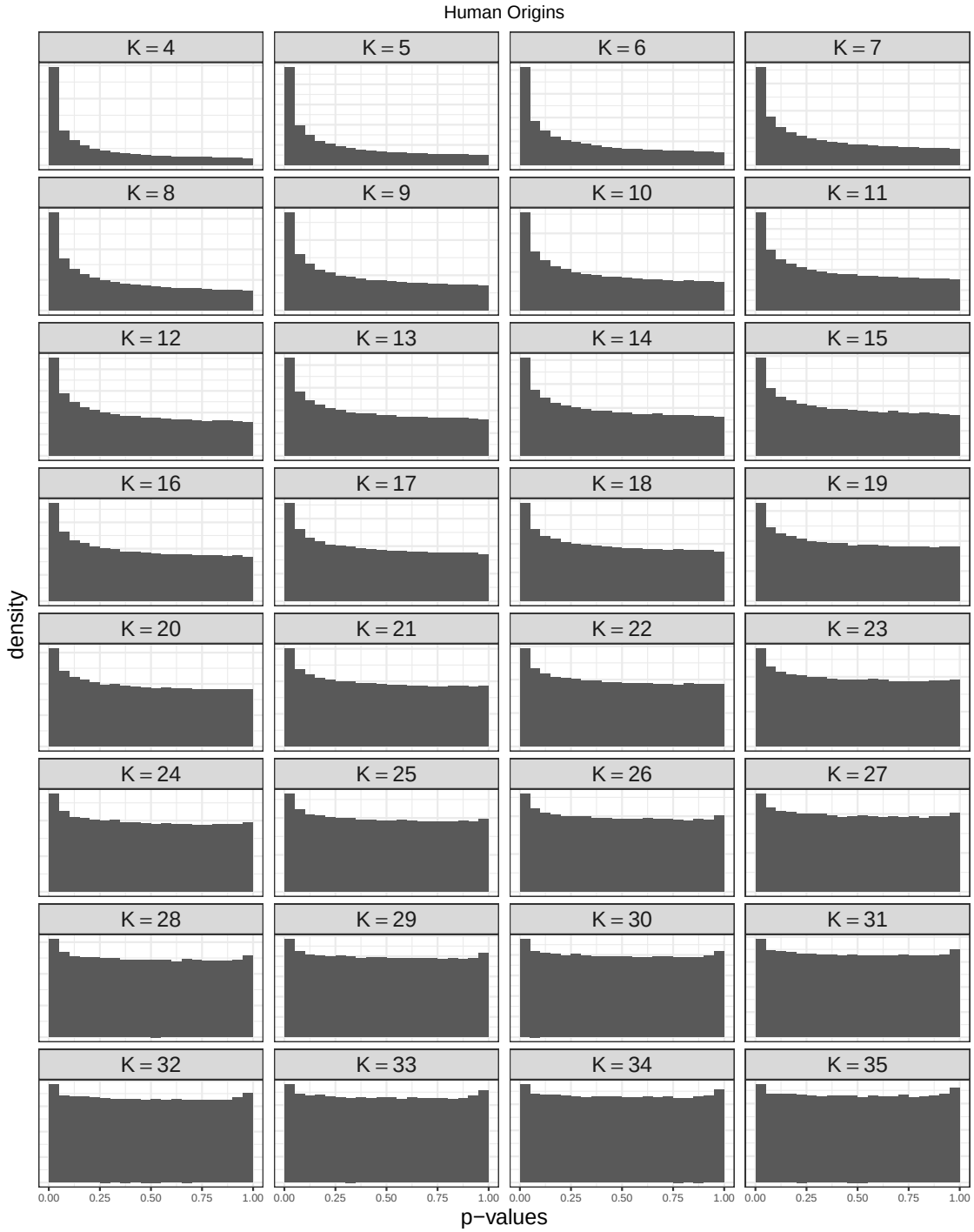

Figure S7: Structural HWE p-values for the HO dataset using the truncated PCA approach to estimate allele frequencies. This is analogous to Figure S4.

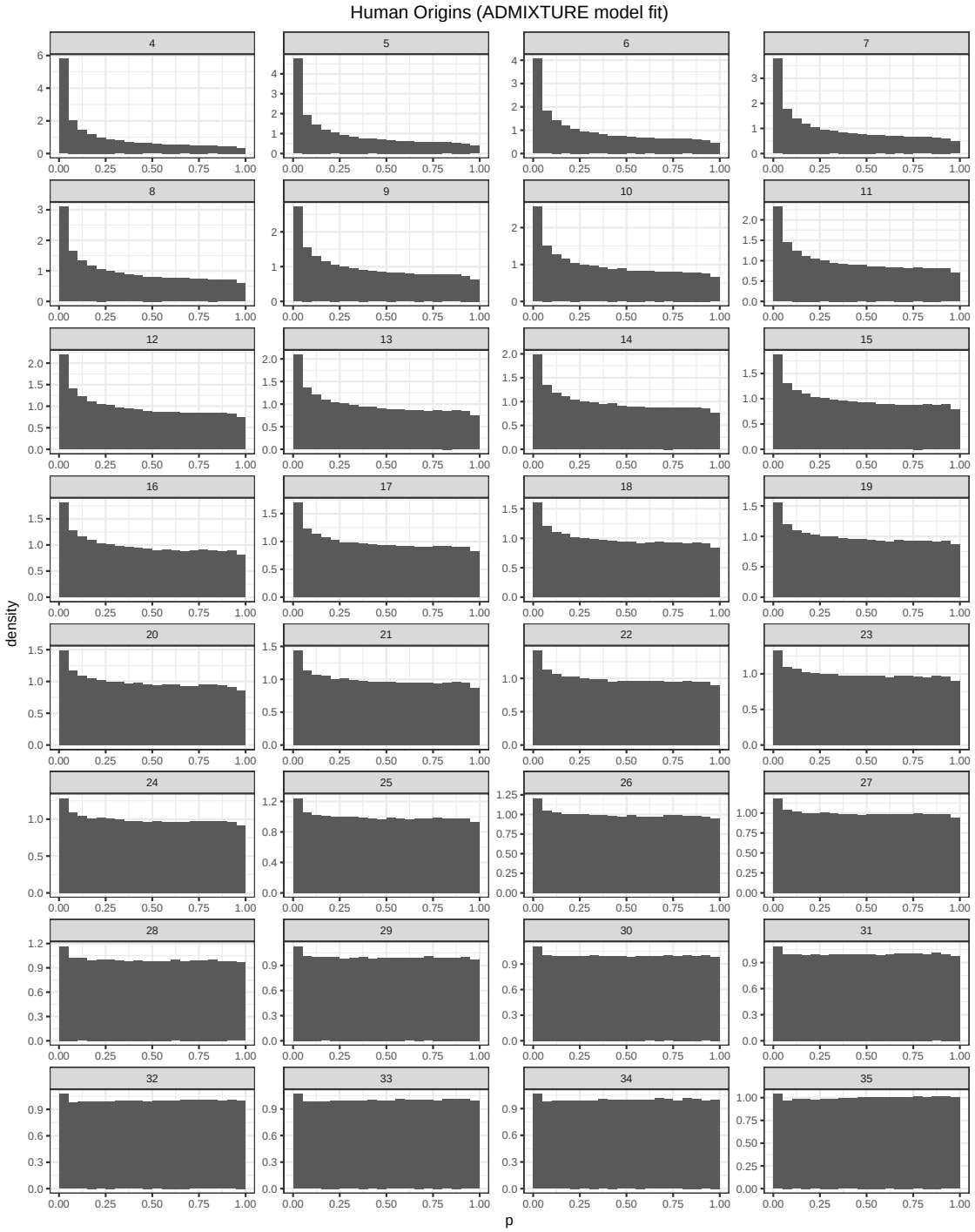

Figure S8: Structural HWE p-values for the HO dataset using the ADMIXTURE method to estimate allele frequencies. This is analogous to Figure S4.

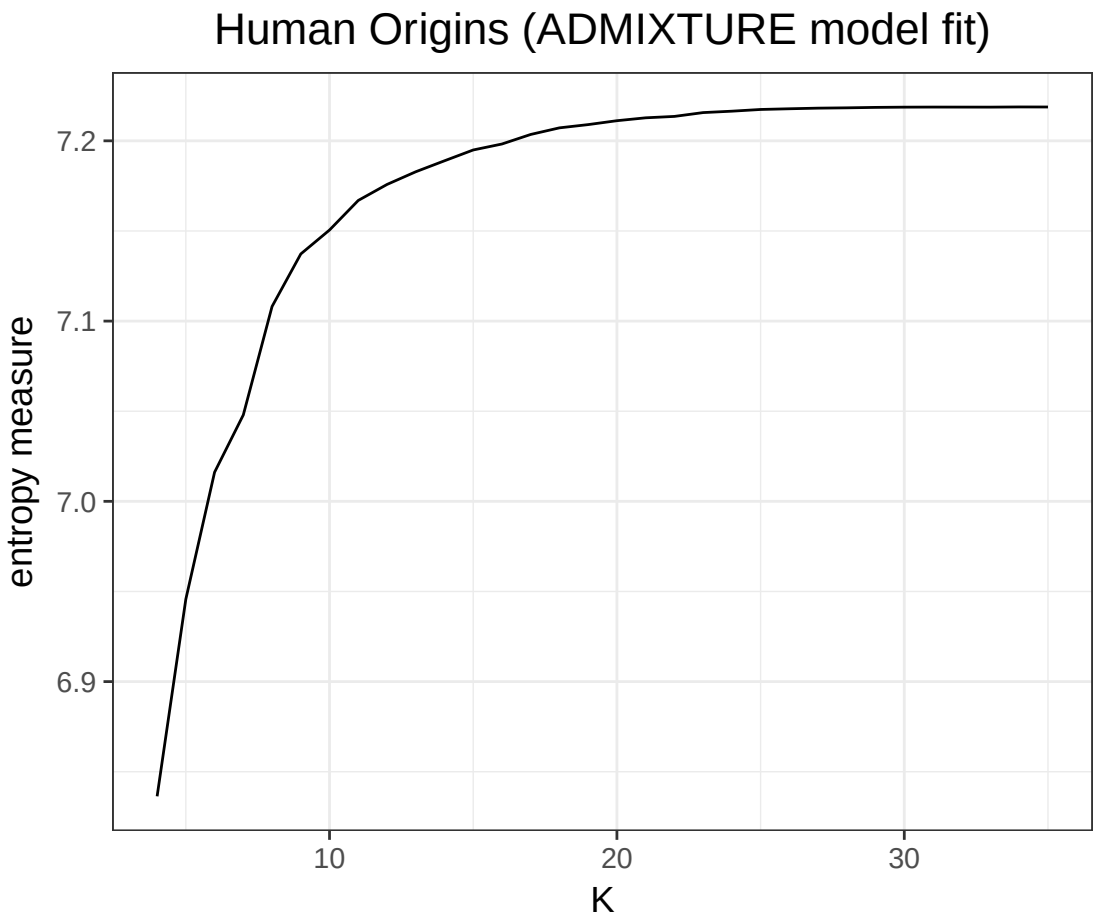

Figure S9: Entropy measure over a range of K for the HO dataset using the ADMIXTURE method to estimate allele frequencies.

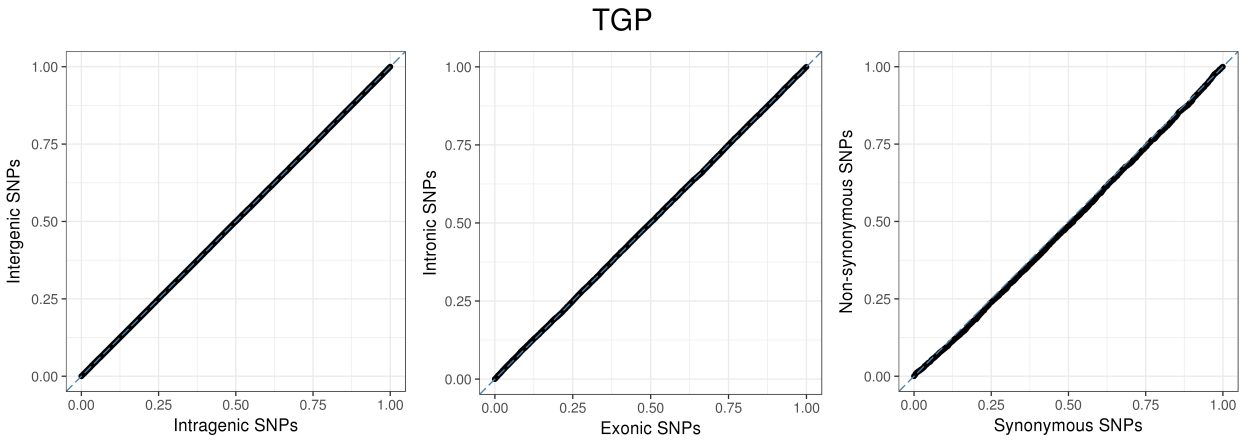

Figure S10: Quantile-quantile plots of structural HWE p-values for the TGP dataset at $K = 12$ comparing successively more fine-scale functional dichotomies for SNPs. The coarsest categories are intergenic SNPs versus intragenic SNPs. Then, among the intragenic SNPs, the comparison is between intronic SNPs versus exonic SNPs. Lastly, among the exonic SNPs, the comparison is between non-synonymous SNPs versus synonymous SNPs. Functional annotations are based on dbSNP build 147.

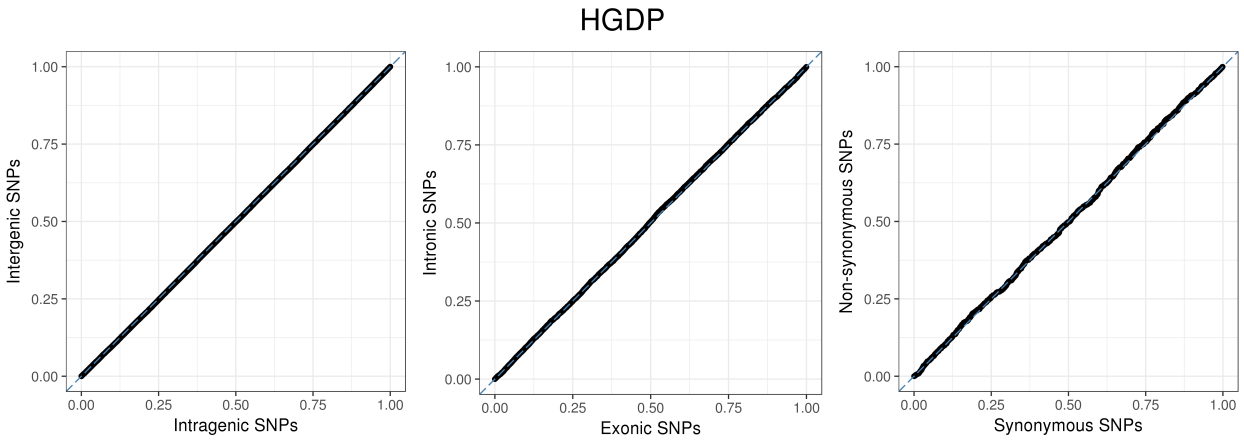

Figure S11: Quantile-quantile plots of structural HWE p-values for the HGDP dataset at $K = 16$ comparing successively more fine-scale functional dichotomies for SNPs. The coarsest categories are intergenic SNPs versus intragenic SNPs. Then, among the intragenic SNPs, the comparison is between intronic SNPs versus exonic SNPs. Lastly, among the exonic SNPs, the comparison is between non-synonymous SNPs versus synonymous SNPs. Functional annotations are based on dbSNP build 147.

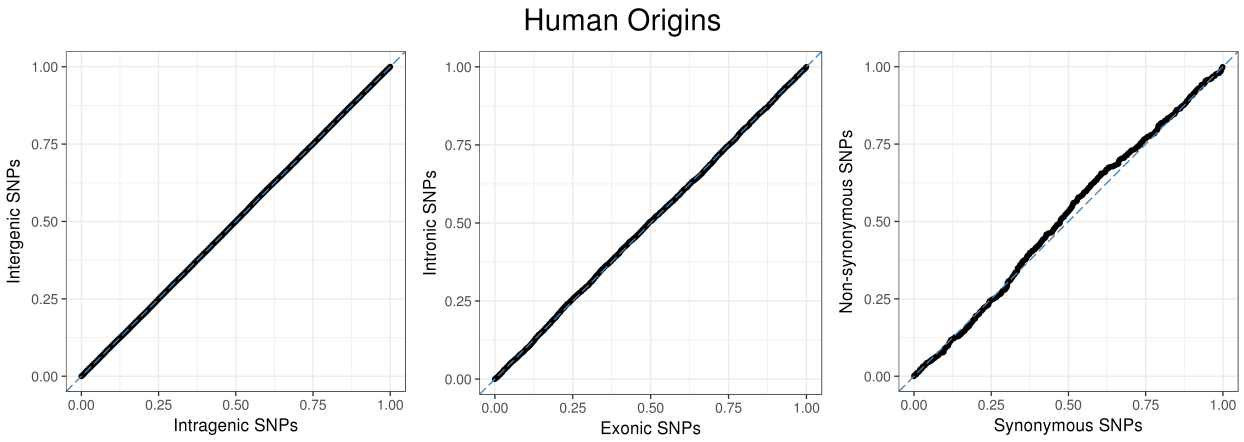

Figure S12: Quantile-quantile plots of structural HWE p-values for the Human Origins dataset at $K = 25$ comparing successively more fine-scale functional dichotomies for SNPs. The coarsest categories are intergenic SNPs versus intragenic SNPs. Then, among the intragenic SNPs, the comparison is between intronic SNPs versus exonic SNPs. Lastly, among the exonic SNPs, the comparison is between non-synonymous SNPs versus synonymous SNPs. Functional annotations are based on dbSNP build 147.

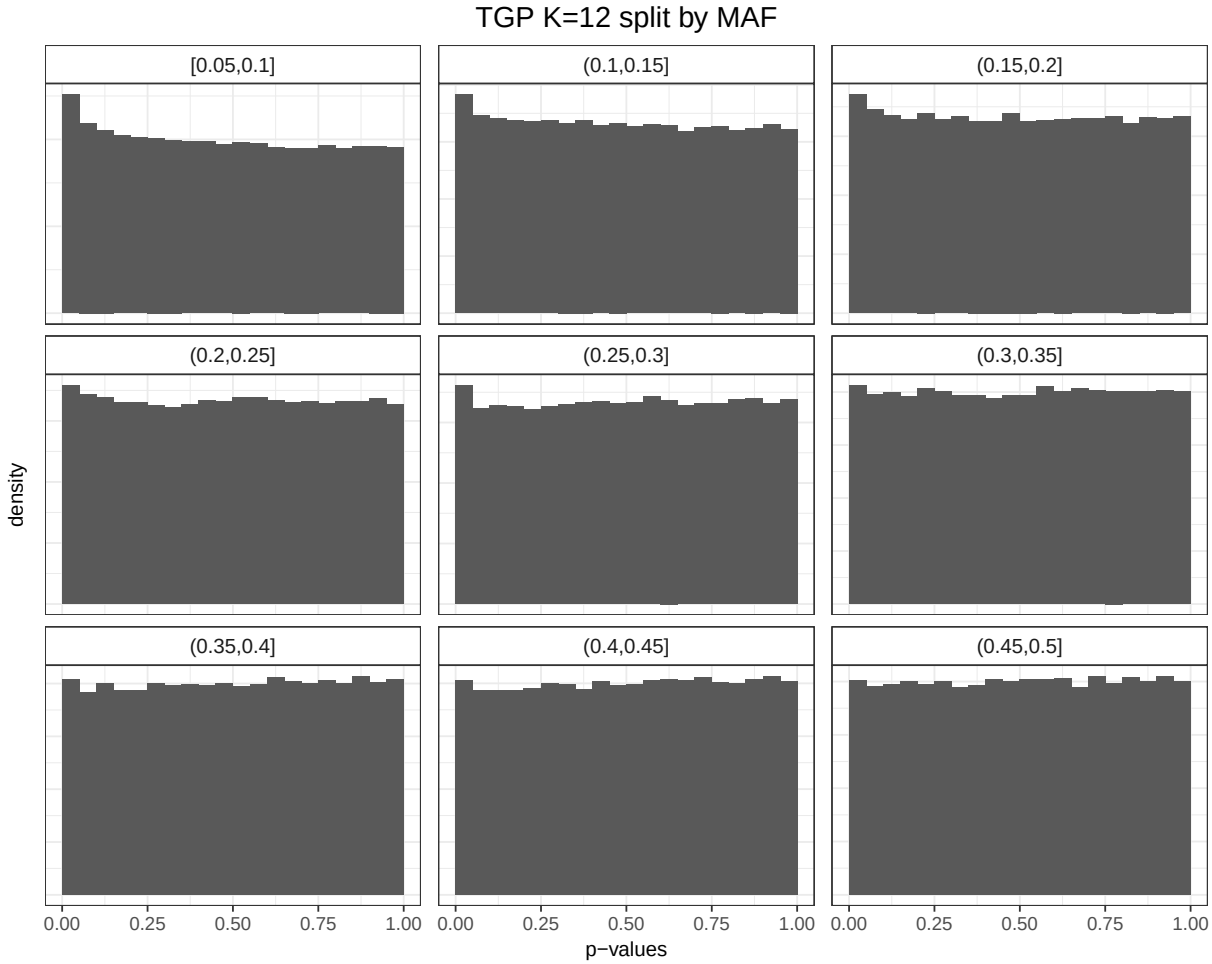

Figure S13: Structural HWE p-value histograms split by minor allele frequency for the TGP dataset ($K = 12$). There is minimal difference between bins.

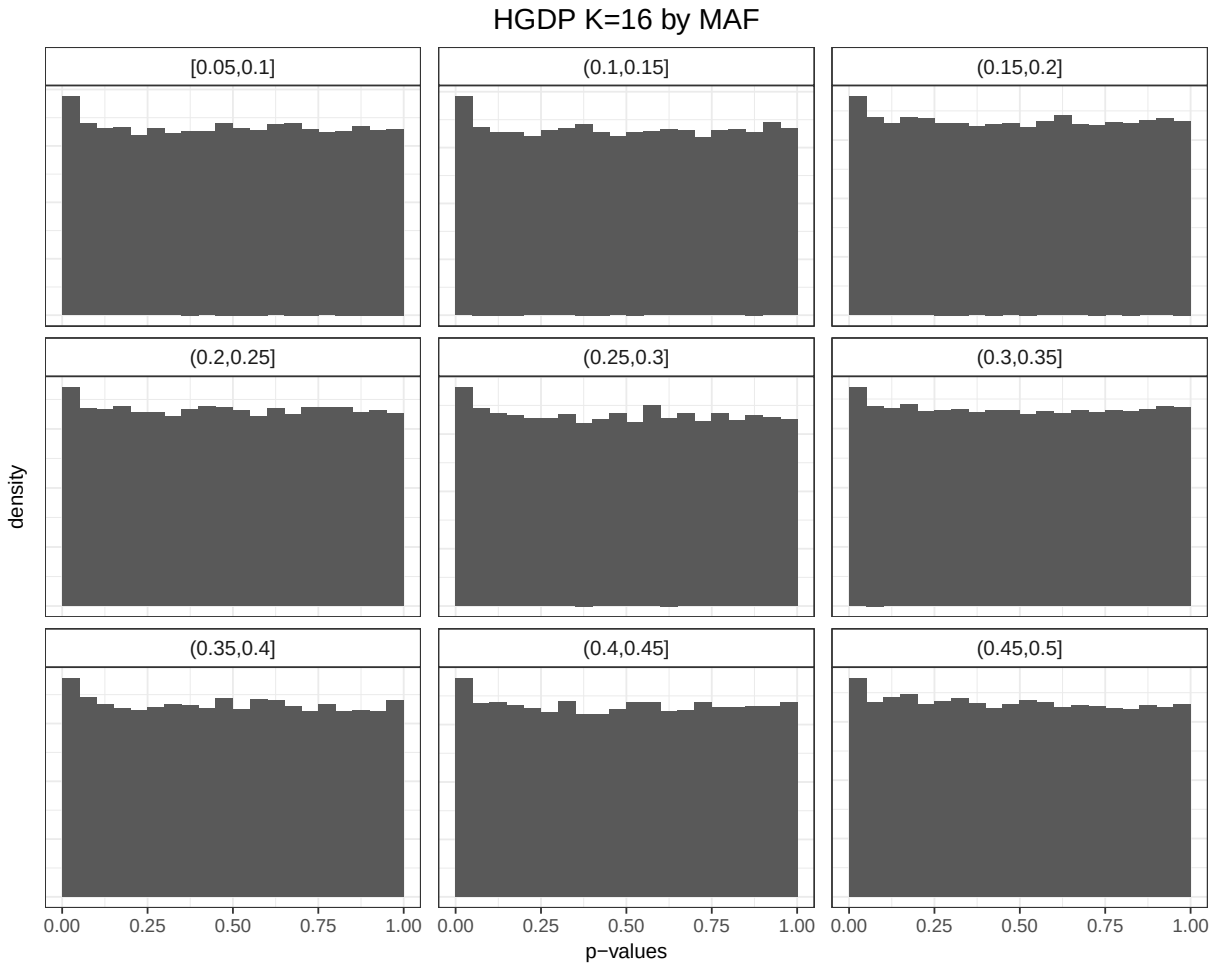

Figure S14: Structural HWE p-value histograms split by minor allele frequency for the HGDP dataset ($K = 16$). There is minimal difference between bins.

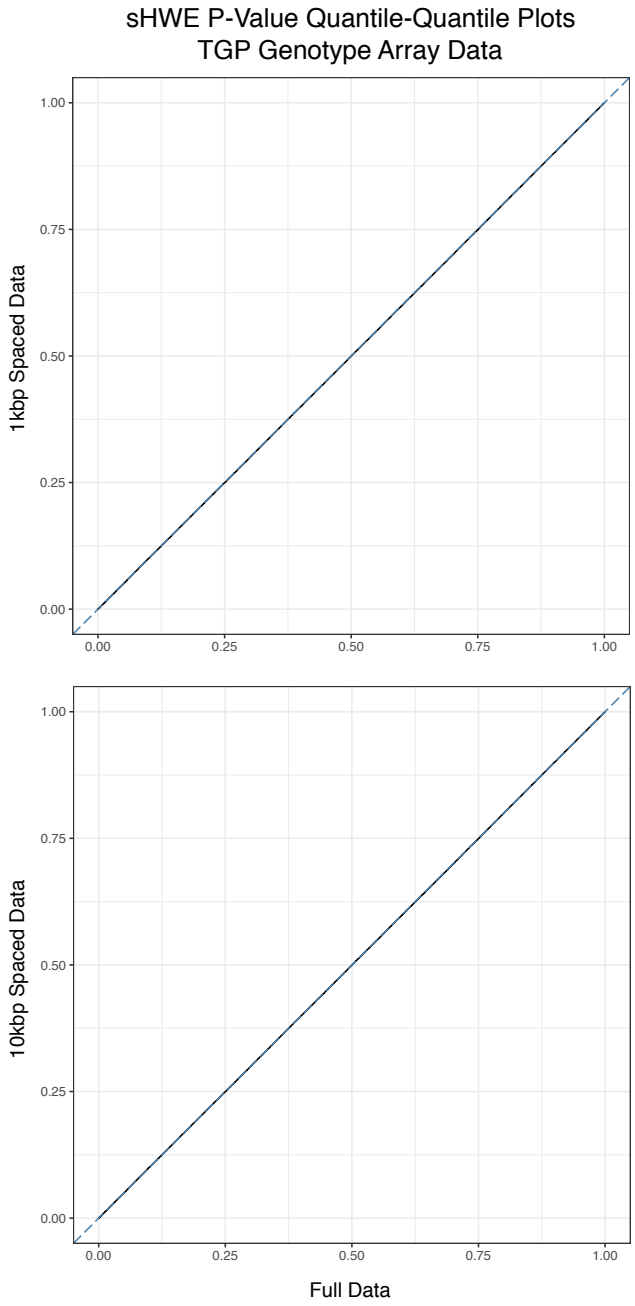

Figure S15: A quantile-quantile plot comparison of sHWE p-values from the full TGP genotyping array data set versus that where SNPs are at least 1kbp apart (top panel) and 10kbp apart (bottom panel).

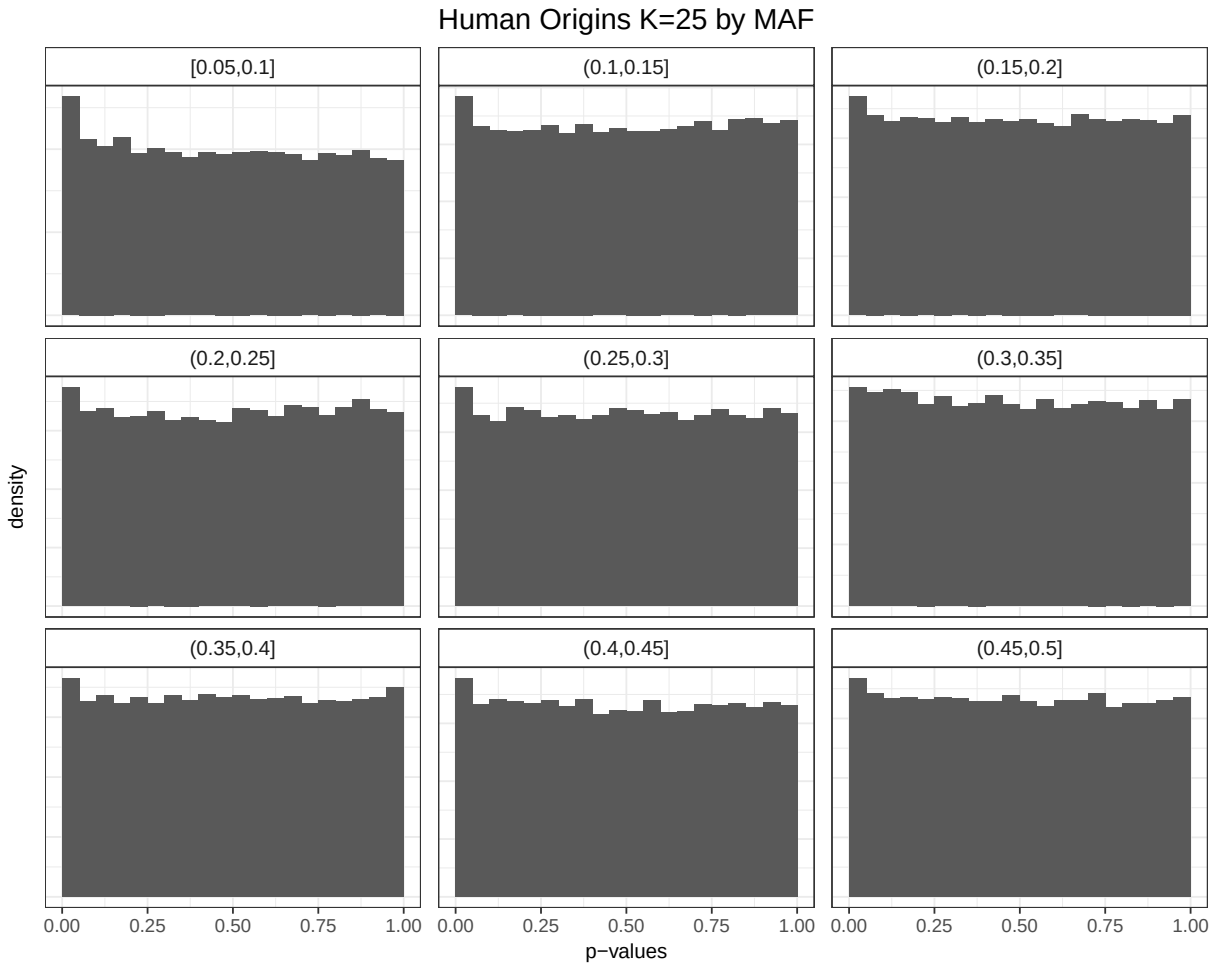

Figure S16: Structural HWE p-value histograms split by minor allele frequency for the Human Origins dataset ($K = 25$). There is minimal difference between bins.

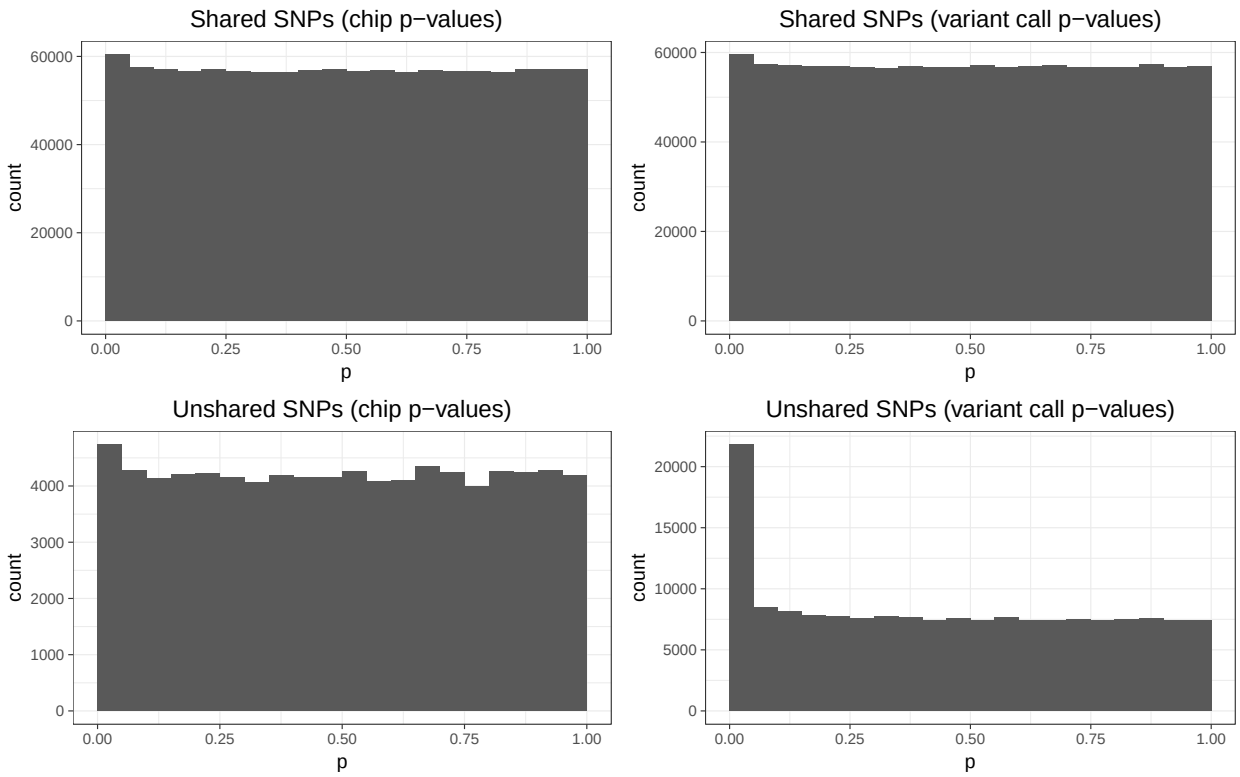

Figure S17: Histogram of sHWE p-values for the TGP chip and variant call datasets, using all SNPs and split between shared and unshared SNPs. The shared SNPs have similar distributions which shows that their results are conserved, while the unshared SNPs show that there are many more SNPs in the integrated callset which are significant.