# Per-allele disease and complex trait effect sizes are predominantly African MAF-dependent in European populations

Jordan Rossen[1,2,&], Benjamin J. Strober[1,3], Kangcheng Hou[1,2,4], Gaspard Kerner[1,2], Alkes L. Price[1,2,5,&]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, USA

[3]Computational Health Informatics Program, Boston Children's Hospital, Boston, USA

[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[&]Corresponding authors: jordanerossen@gmail.com and aprice@hsph.harvard.edu

***Abstract***

Understanding genetic architectures of disease is fundamental to partitioning heritability, polygenic risk prediction, and statistical fine-mapping. Genetic architectures of disease in European populations have been shown to depend on European minor allele frequency (MAF): SNPs with lower MAF have larger per-allele effects, due to the action of negative selection. However, we hypothesized that (unadmixed) African MAF, which is not distorted by the out-of-Africa bottleneck, might better predict per-allele effect sizes of common genetic variation in European populations; we note that common variants explaining most disease heritability are typically much older than the split between African and non-African populations. To demonstrate this, we first analyze the proportion of non-synonymous SNPs, which are strongly impacted by negative selection. The proportion of non-synonymous SNPs is much better predicted by African MAF than European MAF; a mixture of African MAF with weight $w$=0.95 (95% CI: (0.93, 0.96)) and European MAF with

weight (1−*w*) is a more powerful predictor than either European MAF ($P<10^{-15}$, 3.65x greater increase in log-likelihood relative to a null model without MAF dependence) or African MAF ($P<10^{-15}$). Next, we consider the widely used $\alpha$ model, in which per-allele GWAS effect size variance is proportional to $[p_E(1-p_E)]^{\alpha}$, where $p_E$ is the European MAF. We propose a different model in which per-allele effect size variance is proportional to $[p_{mix}(1-p_{mix})]^{\alpha_{mix}}$, where $p_{mix}=w*p_A+(1-w)*p_E$, and $p_A$ is the African MAF. We fit the $\alpha_{mix}$ model by extending the baseline-LD model used in S-LDSC to include a grid of bivariate African and European MAF bins and identifying values of *w* and $\alpha_{mix}$ that fit mean effect size variance estimates from S-LDSC across bivariate MAF bins. We demonstrate that our approach provides conservative estimates of *w* in simulations. We applied this approach to summary statistics for 50 diseases/complex traits in European populations (average *N*=483K) and estimated best-fit parameters of *w*=0.96 (95% CI: (0.76, 1.16)) and $\alpha_{mix}$=−0.34 (95% CI: (-0.67, -0.02)), attaining a far better fit than the standard $\alpha$ model using $p_E$ only ($P<10^{-15}$, 4.53x greater decrease in mean-squared error relative to a null model without MAF dependence). We conclude that per-allele disease and complex trait effect sizes are predominantly African MAF-dependent in European populations.

41 ***Introduction***

42 Understanding genetic architectures is fundamental to partitioning heritability[1–5], polygenic
43 risk prediction[6–11], and statistical fine-mapping[12–16]. Recent studies have shown that genetic
44 variants with lower minor allele frequency (MAF) tend to have larger per-allele effects on diseases
45 and complex traits, due to negative selection against disease-associated variants that impact
46 fitness[17–20]. This relationship is frequently modeled using the $\alpha$ model[4,9,15,17–37], in which the per-
47 allele effect size variance of a SNP in a GWAS population (e.g. Europeans) is a function of its MAF
48 in the GWAS population. However, little attention has been paid to the possibility that the MAF in a
49 different population (e.g. Africans) might be informative for GWAS architectures (see ref.[28,34,38] and
50 Discussion).

51 We hypothesized that (unadmixed) African MAF might better predict per-allele effect sizes in
52 European populations, as it is not distorted by the out-of-Africa bottleneck and thus more
53 accurately reflects the consequences of negative selection; we note that common variants
54 explaining most disease heritability are typically older than the split between African and non-
55 African populations[39,40]. First, we demonstrate that the proportion of SNPs that are non-
56 synonymous is better predicted by African MAF than European MAF, relying on non-synonymous
57 variant type as a partial proxy for variant effect size. Second, we analyze European GWAS
58 summary statistics using a model that incorporates both African MAF and European MAF to
59 demonstrate that estimated per-allele effect variance is more heavily influenced by African MAF
60 than European MAF.

## ***Results***

*Overview of methods: defining the $\alpha_{mix}$ model*

We propose the $\alpha_{mix}$ model for quantifying the MAF-dependence of per-allele effects on diseases/traits. We model per-allele effect size variance as a function of a linear combination of (unadmixed) African and European MAF, instead of MAF in a single ancestry as in the widely used $\alpha$ model[4,9,17–37].

In detail, the $\alpha_{mix}$ model relates per-allele effect size variance to unadmixed African ($p_A$) and European ($p_E$) MAF:

$$\mathrm{E}[\beta^2|p_A, p_E] = \sigma_g^2[2p_{mix}(1-p_{mix})]^{\alpha_{mix}}, \quad (1)$$

where $\sigma_g^2$ is a variance parameter, $p_{mix}=w*p_A+(1-w)*p_E$, $w$ specifies the relative contributions of $p_A$ and $p_E$, and $\alpha_{mix}$ specifies the level of MAF dependence.

In the $\alpha_{mix}$ model, we define African MAF ($p_A$) using whole-genome sequencing data from All of Us[41] ($N$=46,672 African Americans, correcting for European admixture by extracting local genomic regions with two African haplotypes identified using RFMix2[42] (Methods)) and 1000 Genomes[43] ($N$=107 Yoruba), and define European MAF using whole-genome sequencing data from 1000 Genomes[43] ($N$=489 Europeans); we did not use All of Us to define $p_E$ to ensure consistency with the baseline-LD (v2.2) model[3,5]. In analyses of non-synonymous variation, we define African and European MAF using whole-genome sequencing data from All of Us ($N$=46,672 African Americans and $N$=70,346 European Americans). We note that Africa harbors a diverse set of populations[44–46], but we focused on the unadmixed African component of African Americans, due to (i) the large available sample size and (ii) its known genetic similarity ($F_{ST}$ < 0.001) to the Yoruba population[47], which is widely studied and has not been subject to recent extreme bottlenecks[43]. In this manuscript, our conclusions primarily apply to genetic variation with $p_E \geq 0.002$.

*Overview of methods: fitting the $\alpha_{mix}$ model*

We fit the $\alpha_{mix}$ model by stratifying SNPs into bins based on African and European MAF, estimating the mean per-allele effect size variance of each bin using S-LDSC[3,5] and estimating $\alpha_{mix}$ model parameters that best fit the S-LDSC estimates; this procedure is analogous to the approach used in ref.[18] to fit the $\alpha$ model.

In detail, we estimate $\alpha_{mix}$ via a generalized method of moments estimator. First, we define a grid of bivariate MAF bins, where each bivariate MAF bin is defined by an African MAF bin and a European MAF bin. Second, we estimate the mean effect size variance of each bivariate MAF bin using S-LDSC[3] with an extension of the baseline-LD (v2.2) model[5] designed to capture African MAF-dependence (Methods). Finally, we use a numerical optimization procedure to estimate values of $\alpha_{mix}$, $w$, and $\sigma_g^2$ that minimize the mean squared error (summed across bivariate bins and weighted by the number of SNPs in each bin) between the mean effect size variance estimated by S-LDSC and $\mathrm{E}[\beta^2|P_A, P_E]$ as computed using Equation 1 (averaging across values of $p_{mix}*(1-p_{mix})$ in each bivariate MAF bin). We fit Equation 1 using SNP bins (rather than individual SNPs) for two reasons. First, the expected effect size variance estimated by S-LDSC is often negative for individual SNPs, but is less frequently negative for bivariate MAF bins. Second, $p_{mix}*(1-p_{mix})$ for individual SNPs can approach 0 when $w$ is close to 1, but does not approach 0 when averaged across bivariate MAF bins. In analyses across diseases/traits, we estimate $w$ and $\alpha_{mix}$ jointly while estimating $\sigma_g^2$ separately for each trait; we also include analyses in which we estimate $\sigma_g^2$, $w$, and $\alpha_{mix}$ separately for each trait. We have publicly released custom code used to estimate $\alpha_{mix}$ and perform all analyses described in this manuscript (Code Availability).

We analyze summary association statistics for 50 approximately independent, heritable diseases/traits in European ancestry populations (29 UK Biobank and 21 other diseases/traits; average $N$=483K) (Methods and Supplementary Table 1). We have publicly released summary association statistics and S-LDSC output for each disease/trait (Data Availability).

We complement our analyses of disease/trait heritability by analyzing how the proportion of non-synonymous SNPs varies with African MAF and European MAF. We use the proportion of non-synonymous SNPs as a partial proxy for per-allele effect size variance, as non-synonymous variants are more strongly impacted by negative selection[18,48–52], and are highly enriched for heritability[11,18,52] and fine-mapped causal effects[14,53,54]. We model the probability that a SNP is non-synonymous as a function of $p_{mix}$ via the following logistic regression model:

$$\log\left(\frac{P(Y)}{1-P(Y)}\right) = \mu + \gamma \log(p_{mix}), \quad (2)$$

where $Y$=1 if the SNP is non-synonymous and 0 otherwise and $p_{mix}=w*p_A+(1-w)*p_E$ as above (Methods). To estimate $w$, we compute a profile likelihood, treating $\mu$ and $\gamma$ as nuisance parameters, and use a numerical optimization method to maximize the profile likelihood with respect to $w$ (Methods).

*The proportion of non-synonymous SNPs is predominantly African MAF-dependent in European and African ancestry populations*

We hypothesized that the probability that a SNP is non-synonymous might be better predicted by its African MAF than its European MAF, because African MAF is not distorted by the out-of-Africa bottleneck. To evaluate this hypothesis, we calculated the proportion of SNPs that are non-synonymous (PNS), stratified by bivariate African and European MAF bins (Figure 1a, Supplementary Figures 1-2 and Supplementary Table 2). In general, decreasing African MAF within a European MAF bin corresponds to a substantial decrease in PNS, whereas decreasing European MAF within an African MAF bin corresponds to a less substantial decrease in PNS. To quantify this finding, we fit logistic regression models for non-synonymous status as a function of $\log(p_A)$ stratified by $p_E$ (Figure 1b and Supplementary Table 3-4), and as a function of $\log(p_E)$ stratified by $p_A$ (Figure 1c, Supplementary Figure 3, and Supplementary Table 5-6). We observed consistently strong effects for $p_A$ within each $p_E$ bin ($\gamma$=−0.10, −0.13, and −0.13, for $p_E \geq 0.05$,

135 $0.005 \leq p_E < 0.05$, and $p_E < 0.005$, respectively; $P<10^{-15}$ for each bin), but generally weaker effects
136 for $p_E$ within each $p_A$ bin ($\gamma$=−0.028, −0.047, and −0.10, for $p_A \geq 0.05$, $0.005 \leq p_A < 0.05$, and $p_A <$
137 $0.005$, respectively; $P<10^{-15}$ for each bin); the relatively stronger effect for $p_A < 0.005$ may reflect
138 recent selection on variants that arose in Europeans after the out-of-Africa bottleneck[51], including
139 variants that arose via Neanderthal introgression[55–57]. This supports the hypothesis that African
140 MAF is highly predictive of PNS, even after accounting for European MAF.

141 To directly evaluate whether African MAF are more predictive of PNS than European MAF, we
142 fit a logistic regression model for non-synonymous status as a function of $\log(p_{mix})$ (where
143 $p_{mix}=w*p_A+(1-w)*p_E$) (see Equation 2) and compared different values of $w$ via a profile likelihood
144 approach, treating $\mu$ and $\gamma$ as nuisance parameters. We restricted to SNPs with $p_A \geq 0.002$ and/or
145 $p_E \geq 0.002$ (MAF thresholds are discussed in Methods). The maximum likelihood estimate (MLE) of
146 $w$ was 0.95 (95% CI: (0.93, 0.96)) (Figure 2a and Supplementary Table 7), indicating that PNS is
147 predominantly dependent on African MAF, but European MAF provides some additional
148 information ($P<10^{-15}$ for comparisons to European MAF-only or African MAF-only models).
149 Compared to a null model with no MAF dependence ($\gamma$=0), the $p_{mix}$ model attained a 3.65x larger
150 increase in log-likelihood than the European MAF-only model ($w$=0) and a 1.02x larger increase in
151 log-likelihood than the African MAF-only model ($w$=1) (Figure 2b and Supplementary Table 7).
152 Interestingly, the value of $w$ that produced the largest absolute $\gamma$ ($w$=0.84) (Figure 2c and
153 Supplementary Table 7) was smaller than the MLE value ($w$=0.95), because varying $w$ changes the
154 variance of $\log(p_{mix})$; this discrepancy became negligible when $\log(p_{mix})$ was standardized to have
155 unit variance (Supplementary Figure 4).

156 We performed 5 secondary analyses. First, we re-estimated the effects of conditional effects of
157 $p_A$ (stratifying by $p_E$) and $p_E$ (stratifying by $p_A$) on PNS while restricting to SNPs with both $p_A \geq 0.002$
158 and $p_E \geq 0.002$. Results were similar to Fig. 1b-c (Supplementary Figure 5). Second, we compared
159 the marginal effect of $p_A$ on PNS to the marginal effect of $p_E$ on PNS using logistic regression for

five sets of SNPs: SNPs with $p_A \geq 0.002$ and/or $p_E \geq 0.002$, SNPs with $p_A \geq 0.002$ and $p_E \geq 0.002$, SNPs with $p_A \geq 0.002$, SNPs with $p_E \geq 0.002$, and SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$. African MAF provided a better fit for all five SNP sets ($P < 10^{-15}$ for all tests, except $P = 0.01$ for SNPs with MAF $\geq 0.05$ in both ancestries) (Supplementary Figure 6). Third, we compared the marginal effect of $p_A$ on PNS for all SNPs with $p_A \geq 0.002$ to the marginal effect of $p_E$ on PNS for all SNPs with $p_E \geq 0.002$ (Supplementary Figure 7). Surprisingly, the $p_E$ model attained a higher McFaddens pseudo-$r^2$ and higher absolute slope than the $p_A$ model (McFadden's pseudo-$r^2$=0.007 for $p_E$ vs. 0.004 for $p_A$, $\gamma$=−0.17 for $p_E$ vs. −0.13 for $p_A$, $P < 10^{-15}$ for pseudo-$r^2$ and $\gamma$ differences); however, we note that this comparison varies both the set of SNPs evaluated and the explanatory variable, and thus does not directly indicate whether $p_A$ or $p_E$ better predicts PNS for any fixed set of SNPs. We hypothesize that the observed difference in $r^2$ is because PNS is harder to predict for African-specific SNPs than for European-specific SNPs. Fourth, we refit the logistic regression on $\log(p_{mix})$ using 4 alternative SNP sets: SNPs with $p_A \geq 0.002$ and $p_E \geq 0.002$, SNPs with $p_A \geq 0.002$, SNPs with $p_E \geq 0.002$, and SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$. In each case, the MLE estimate for $w$ was between 0.76-0.95, with tight confidence intervals (Supplementary Figure 8). Fifth, we refit the logistic regression on $\log(p_{mix})$ model while changing $p_A$ to max(0.05, $p_A$) and $p_E$ to max(0.05, $p_E$) (Supplementary Figure 9), rather than 0.002 as in the main analyses. The MLE for $w$ was 0.80, similar to the estimate obtained when restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$.

*Per-allele disease and complex trait effect sizes are predominantly African MAF-dependent in European populations*

We hypothesized that per-allele disease and complex trait effect sizes (e.g. in European populations, for which data is available in large sample size) might be better predicted by African MAF than by European MAF, because African MAF is not distorted by the out-of-Africa bottleneck. To evaluate this hypothesis, we used S-LDSC[3,5] to estimate mean per-allele effect size variances

185 across bivariate MAF bins (7 African MAF bins ranging from 0 to 0.50 x 5 European MAF bins
186 ranging from 0.05 to 0.50), scaled to a mean value of 1 across SNPs and meta-analyzed across 50
187 diseases/traits (average *N*=483K; Supplementary Table 1) (Methods); we excluded SNPs with
188 European MAF less than 0.05 because S-LDSC typically estimates SNP-heritability only for
189 common SNPs[3,5]. Results are reported in Figure 3 and Supplementary Tables 8-9. We determined
190 that per-allele effect size variance exhibited a much stronger dependence on African MAF
191 conditional on European MAF (1.1x-2.7x ratio within each of 5 common $p_E$ columns of Figure 3a,
192 restricted to common $p_A$ rows; solid red line in Figure 3b) than European MAF conditional on
193 African MAF (0.6x-1.4x ratio within each of 5 common $p_A$ rows of Figure 3a; dashed red line in
194 Figure 3b). The dependence on African MAF was much stronger without restricting to SNPs with $p_E$
195 > 0.05 (1.8x-6.7x ratio; 5 common $p_E$ columns of Figure 3a); the dependence on European MAF
196 was stronger for SNPs that are rare in Africans (1.2x ratio within the rare $p_A$ row of Figure 3a;
197 dashed blue line in Figure 3b), possibly reflecting recent selection on variants that arose in
198 Europeans after the out-of-Africa bottleneck[51], including variants that arose via Neanderthal
199 introgression[55–57]. These findings support the hypothesis that African MAF is highly predictive of
200 per-allele effect size variance, even after accounting for European MAF.

201 To directly evaluate whether African MAF are more predictive of per-allele effect size
202 variance than European MAF, we fit the $\alpha_{mix}$ model (Equation 1 and Methods) across the 50
203 diseases/traits, using a single *w* parameter and a single $\alpha_{mix}$ parameter. Results are reported in
204 Figure 4 and Supplementary Table 10. Best-fit estimates were *w*=0.96 (95% CI: (0.76, 1.16)) and
205 $\alpha_{mix}$=−0.34 (95% CI: (−0.67,−0.03)), indicating that per-allele effect size variance is best predicted
206 by placing much greater weight on African MAF than on European MAF ($P<10^{-15}$ vs. null
207 hypothesis of *w*=0 ($\alpha$ model); $P=9*10^{-6}$ vs. null hypothesis of *w*=0.5) (Figure 4a); we failed to reject
208 the null hypothesis that per-allele effect size variance depends on African MAF alone (*P*=0.71 vs.
209 null hypothesis of *w*=1). Relative to a null model with no MAF dependence ($\alpha_{mix} = 0$), $\alpha_{mix}$

attained a 4.53x greater decrease in weighted MSE than a European MAF-only model ($w$=0) and a 1.03x greater decrease in weighted MSE than an African MAF-only model ($w$=1) (Figure 4b). The relationship between $w$ and the MSE of the $\alpha_{mix}$ model was remarkably similar to the relationship between $w$ and the negative log-likelihood of our logistic regression of PNS (Figure 2b). We also estimated $\alpha_{mix}$ values for each value of $w$ (Figure 4c). Surprisingly, values of $w$ that were lower than the best-fit value of $w$=0.96 produced larger absolute $\alpha_{mix}$ estimates. This may be because at high values of $w$, $p_{mix}$ approaches 0 for some bins of SNPs, which results in overestimates of per-allele effect size variance under the $\alpha_{mix}$ model, which cause best-fit estimates of $\alpha_{mix}$ to be attenuated towards 0. This is analogous to the known phenomenon that the $\alpha$ model overestimates per-allele effect size variance at low values of $p$ according to evolutionary forward simulations, in which per-allele effect size variance plateaus at low values of $p$ (Figure 2 of ref.[19]). To investigate further, we fit the $\alpha_{mix}$ model while restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$. Accordingly, $\alpha_{mix}$ estimates were less dependent on $w$ when restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$ (best-fit $w$=0.94 and $\alpha_{mix}$=−0.22; Supplementπary Figure 10). Estimates of $\alpha_{mix}$ were closer to 0, providing further evidence that SNPs with $p_A < 0.05$ are the primary drivers of the European MAF-dependence of per-allele effect size variance.

We estimated $\alpha_{mix}$ and $w$ parameters separately for each disease/trait. Results are reported in Figure 5, Supplementary Figures 11-12, and Supplementary Table 11. The random effects meta-analysis estimate of $w$ was 1.00 (95% CI: (0.99, 1.00)) (similar to our joint estimate of 0.96; see above). Despite substantial noise in our estimates, $w$ estimates were significantly greater than 0.5 for 24/50 traits ($P$<0.05/50; greater dependence on African MAF vs. European MAF) and significantly greater than 0 for 32/50 traits ($P$<0.05/50; significant dependence on African MAF)—but significantly less than 1 for 0/50 traits ($P$<0.05/50; significant dependence on European MAF). The random effects meta-analysis $\alpha_{mix}$ estimate was =−0.31 (95% CI: (−0.26,−0.36)) (similar to our joint estimate of −0.34; see above). We determined that there was no statistically significant

heterogeneity in $w$ or $\alpha_{mix}$ in a random effects meta-analysis (between-trait variance estimate of 0) or a fixed-effect meta-analysis ($P$=0.97 and 0.98 for $w$ and $\alpha_{mix}$, respectively).

We performed 5 secondary analyses. First, we estimated $\alpha_{mix}$ and $w$ parameters separately for each disease/trait, while restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$ (Supplementary Figure 13). Our random effects meta-analysis estimates of $w$=0.53 (95% CI: (0.43, 0.63)) differed substantially from that of our main analysis, while our estimate of $\alpha_{mix}$=−0.33 (95% CI: (−0.19, −0.46)) was similar to the main analysis estimate. Identifying the reasons for this difference in $w$ estimates when estimating parameters separately for each disease/trait and restricting to common SNPs is a direction for future research. Second, we fit the $\alpha_{mix}$ model while estimating African MAF using a much smaller sample of Yoruba genomes ($N$=107 from 1000 Genomes YRI vs. $N$=46,779 African individuals and $N$=489 European individuals in the main analyses) (Supplementary Figure 14). Our estimates of $w$=0.96 (95% CI: (0.84, 1.07)) and $\alpha_{mix}$=−0.37 (95% CI: (−0.14, −0.59)) were very similar to those of the main analyses. Third, we fit the $\alpha_{mix}$ model while estimating per-allele effect size variances using the standard baseline-LD (v2.2) model (Supplementary Figure 15), rather than our extension of the baseline-LD (v2.2) model which includes annotations for 120 bivariate MAF bins. Our estimates of $\alpha_{mix}$=−0.34 (95% CI: (−0.25, −0.44)) and $w$=0.85 (95% CI: (0.76, 0.95)) were similar to those of the main analysis. Fourth, to test whether African LD-related annotations in the baseline-LD (v2.2) model were falsely inducing African MAF dependence, we fit the $\alpha_{mix}$ model while estimating per-allele effect size variances using 120 bivariate MAF bins rather than our extension of the baseline-LD (v2.2) model which includes all annotations from the standard baseline-LD (v2.2) model (Supplementary Figure 16). Our estimates of $w$=0.98 (95% CI: (0.86, 1.10)) and $\alpha_{mix}$=−0.38 (95% CI: (0, −0.76)) were very similar to those of the main analyses. Fifth, we fit the $\alpha_{mix}$ model using two disjoint subsets of the 50 diseases/traits from the main analysis: 29 UK Biobank diseases/traits and 21 diseases/traits from publicly available summary statistics (Supplementary Figures 17-18). Results from each subset of diseases/traits were very

similar to each other (e.g. $w$=0.97 (95% CI: (0.76, 1.17)) vs. $w$=0.96 (95% CI: (0.87, 1.05)), respectively).

*Simulations*

To test the validity of our $\alpha_{mix}$ estimation procedure, we simulated quantitative GWAS traits according to the $\alpha_{mix}$ model using real genotypes from 337,448 unrelated British individuals from UK Biobank and 5,907,305 SNPs with European MAF ≥ 0.05. We varied the true value of $w$ between 0 and 1, set $\alpha_{mix}$ to −0.38 (consistent with previous estimates in European populations[17–19]), set $\sigma_g^2$ such that $h^2$=0.5, set $p_{mix}$=max($w$*$p_A$ + (1-$w$)*$p_E$, $T$) with $T$=0.005 (to avoid unrealistically large values of $E[\beta^2|p_A, p_E]$ for SNPs with low values of $p_A$, consistent with ref.[19]; this implies that the generative model is different from the inference model, but our secondary analyses below show that this makes little difference), and set the number of causal SNPs to 10,000; other simulations settings were also evaluated. We fit the $\alpha_{mix}$ model (Equation 1 and Methods) using a single $w$ parameter and a single $\alpha_{mix}$ parameter across groups of 25 traits (which resulted in total $N*h^2$=4.2*$10^6$, vs. 4.1*$10^6$ in real trait analyses).

Results are reported in Figure 6 and Supplementary Table 12. We determined that our $w$ estimates were conservative in the sense that they were biased away from extreme values, e.g. average estimate ranging from 0.16 when true $w$=0.00 to 0.97 when true $w$=1.00 (Figure 6a); estimates were approximately unbiased (0.94) at true $w$=0.95, which was closest to our estimate for real traits. Estimates of $\alpha$ were biased downward (relative to true $\alpha$=−0.38) at most values of $w$, e.g. −0.48 when $w$=0.00 and −0.42 when $w$=0.95 (but 0.35 when $w$=1.00) (Figure 6b).

We performed 4 secondary analyses. First, we simulated effects based on MAF from $N$=337,448 unrelated British individuals, but performed inference using MAF from $N$=489 European individuals from 1000 Genomes to simulate the effect using noisy MAF estimates (Supplementary Figure 19). Results were very similar to those of the main analyses. Second, we varied $T$, the MAF

threshold used by the generative model (Supplementary Figures 20-21), which determines how high $E[\beta^2|p_A, p_E]$ is for SNPs with low values of $p_A$. Results with $T$=0.05 were similar to those of the main analyses, but with upward bias (rather than downward bias) in our $\alpha_{mix}$ estimates at $w$=0.75 and $w$=0.95 and slightly greater upward bias in $\alpha_{mix}$ at $w$=1.00. Results with $T$=1.5*10$^{-5}$ (equivalent to a minor allele count of 1) were similar to those of the main analyses, but with less bias in our estimates of $w$ at high values of $w$, and downward bias (rather than upward bias) in our estimates of $\alpha_{mix}$ at $w$=1.00. Third, we varied the number of causal SNPs per trait (Supplementary Figure 22-23). With 5,000 (instead of 10,000) causal variants per trait, results were similar to the main analyses, with slight downward bias at $w$=1. With 20,000 causal variants per trait, results were very similar to the main analyses. Finally, we estimated trait-specific values of $w$ and $\alpha_{mix}$, rather fitting a joint model across all 25 traits (Supplementary Figure 24). Results were very similar to the main analyses, but with greater uncertainty in mean parameter estimates.

297 ***Discussion***

298 We have shown that per-allele disease and complex trait effect sizes of SNPs that are
299 common in European populations are predominantly a function of African MAF, not European MAF.
300 This represents a significant departure from standard approaches for modeling MAF dependence,
301 where per-allele effect sizes estimated in a given population are related to MAF in the same
302 population[4,9,15–27,29–33,35–37]. We have provided two lines of evidence in favor of an African MAF-
303 dependence model. First, we showed that non-synonymous variant status is much better predicted
304 using African MAF than European MAF. This approach requires very few assumptions and is likely
305 robust to model misspecification, but is limited in that it only focuses a single, relatively narrow
306 subset of human genetic variation. Second, we used disease/trait GWAS summary statistic data
307 from European populations to show that per-allele effect size variance is also much better
308 predicted using African MAF than European MAF. This approach reached very similar conclusions
309 to our analysis of non-synonymous variation.

310 Our work is motivated in part by two recent population genetic studies exploring the
311 relationship between African MAF, European MAF, and GWAS effect sizes. Refs.[28,34] performed
312 simulations exploring the effect of bottlenecks on genetic architectures and MAF dependence,
313 using models in which per-allele effect sizes in one population are dependent on MAF in another
314 (Ref.[28] focuses specifically on the out-of-Africa bottleneck and Ref.[34] does not). However, neither
315 study analyzed empirical GWAS data or demonstrated that this model fits empirical data. Ref.[38]
316 used conditional allele frequency spectra to differentiate modes of selection on GWAS-associated
317 variants (emphasizing joint analyses of African MAF and European MAF, as we do in this study),
318 but did not quantify the dependence of per-allele effect sizes on African MAF or European MAF.

319 Our work has important implications both for analyses of European GWAS data and, more
320 generally, populations of all ancestries. Modeling MAF-dependent architectures has been shown to
321 impact studies of genetic architectures and negative selection[2,4,5,17–24,26,28,29,31,33,34,37,38,48–50,52,58,59],

polygenic risk prediction[9,11,60], and fine-mapping[16,60], and our results suggest that specifically modeling African MAF-dependent architectures will improve all of these endeavors, not only in European populations (the focus of this study) but in population of all ancestries as more GWAS data in diverse populations becomes available[61–66,41,67].

Our work has several limitations. First, our analyses focus exclusively on African and European MAF. We expect that drift from the out-of-Africa bottleneck causes MAF in all non-African ancestries to be less informative for per-allele effect size variance[43], although we have not explicitly tested this hypothesis. Second, our analyses of diseases/traits focus exclusively on European-ancestry GWAS data, which limits our disease/trait analyses to SNPs with $p_E > 0.05$; however, our analysis of non-synonymous variant status spans a larger set of SNPs ($p_A > 0.002$ and/or $p_E > 0.002$), providing some evidence that our findings are generalizable. Analyses of diseases/traits in non-European populations will be more informative as more GWAS data in diverse populations becomes available[61–66,41,67]. Third, our analyses focus on continental ancestries and do not consider subcontinental ancestries with distinct allele frequencies, demographic histories, and selective forces[44–46,68,69]; however, allele frequency differences between populations of the same continental ancestry are smaller than allele frequency differences between populations of different continental ancestry[70]. Fourth, our estimate of the degree of European MAF-dependence under the $\alpha$ model ($\alpha$=−0.50, 95% CI: (−0.61,−0.38), Figure 4c) is stronger than refs.[17–19] ($\alpha$ = −0.38 to −0.36 meta-analysis across traits) (though weaker than refs.[20,33]; $\alpha$ = −0.65 to −0.58). When estimating $\alpha$ using S-LDSC with the standard Baseline-LD (v2.2) model, which does not include African MAF annotations, (Supplementary Figure 15), we obtain an estimate that is consistent with refs.[17–19] ($\alpha$ = −0.41, 95% CI (−0.51, −0.32)); regardless of the S-LDSC model used, we find that per-allele effect size variances are predominantly African MAF-dependent, which is our main conclusion. Fifth, our approach to fitting the $\alpha_{mix}$ model suffers some biases in simulations (Figure 6); however, estimates of *w* were conservative (in the sense

347 that they were biased away from extreme values), and approximately unbiased for values of *w*
348 close to our estimate of w=0.96 for real traits, thus providing strong support for our conclusion that
349 per-allele effect sizes are predominantly African MAF-dependent. Sixth, we did not perform
350 simulations to assess whether the $\alpha_{mix}$ model is consistent with evolutionary models at appropriate
351 parameter settings; a previous study performed forward simulations[71] to show that the $\alpha$ model is
352 generally consistent with an evolutionary model involving a single population[19] (except that per-
353 allele effect size variance plateaus at low values of minor allele frequency; Figure 2 of ref.[19]), and
354 extending those forward simulations to incorporate African and non-African populations (impacted
355 by the out-of-Africa bottleneck) is an important future research direction. Seventh, analogous to
356 previous studies using the $\alpha$ model[4,18,19,21–26,29,31,32,34,35], our analyses using the $\alpha_{mix}$ model do not
357 account for non-infinitesimal disease/trait effects[72]; we note that failure to model non-infinitesimal
358 effects when estimating SNP-heritability leads to suboptimal precision, but does not induce bias[72].
359 Eighth, African MAF is unlikely to be informative of per-allele effects among SNPs which are
360 younger than the out-of-Africa bottleneck and which are common in European ancestry
361 populations, although these represent a small fraction of European common variation; modeling
362 MAF-dependence as a function of allele age is a potentially impactful future research direction.
363 Despite these limitations, we have shown that per-allele disease and complex trait effect size
364 variances are predominantly African MAF-dependent in European populations.

## Figures

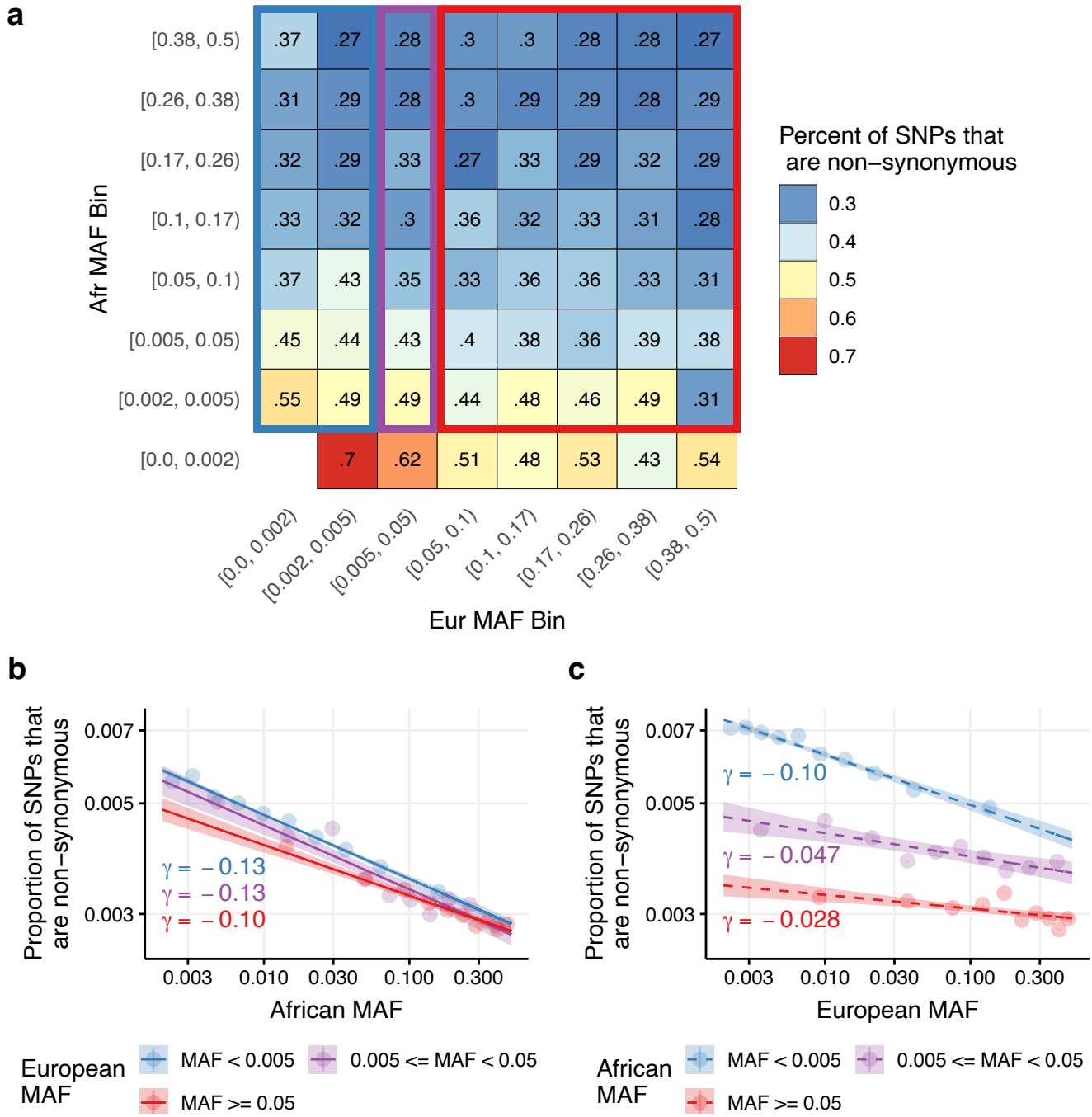



**Figure 1: The proportion of non-synonymous SNPs as function of unadmixed African and European minor allele frequencies:** We report **(a)** the percentage of SNPs that are non-synonymous in bivariate bins defined by unadmixed African MAF (rows) and European MAF (columns), **(b)** the probability that a SNP is non-synonymous as a function of African MAF after stratifying by European MAF categories (for SNPs with $p_A \geq 0.002$) as predicted by logistic regression, and **(c)** the probability that a SNP is non-synonymous as a function of European MAF after stratifying by African MAF categories (for SNPs with $p_E \geq 0.002$) as predicted by logistic regression. The blue, purple, and red rectangles in (a) correspond to the SNPs included in the blue, purple, and red lines in (b). The choice of MAF bin boundaries is discussed in Methods. Estimated regression coefficients for African MAF within European MAF categories are given in the bottom left of (b). Estimated regression coefficients for European MAF within African MAF categories are given below each line in (c). Shaded regions in (b) and (c) denote 95% confidence

intervals on the mean estimate. Logistic regression fits with one regression per plot are reported in Supplementary Figure 3. Numerical results, including standard errors for each bivariate bin in (a) and each bin in (b) and (c), are reported in Supplementary Tables 2-6.

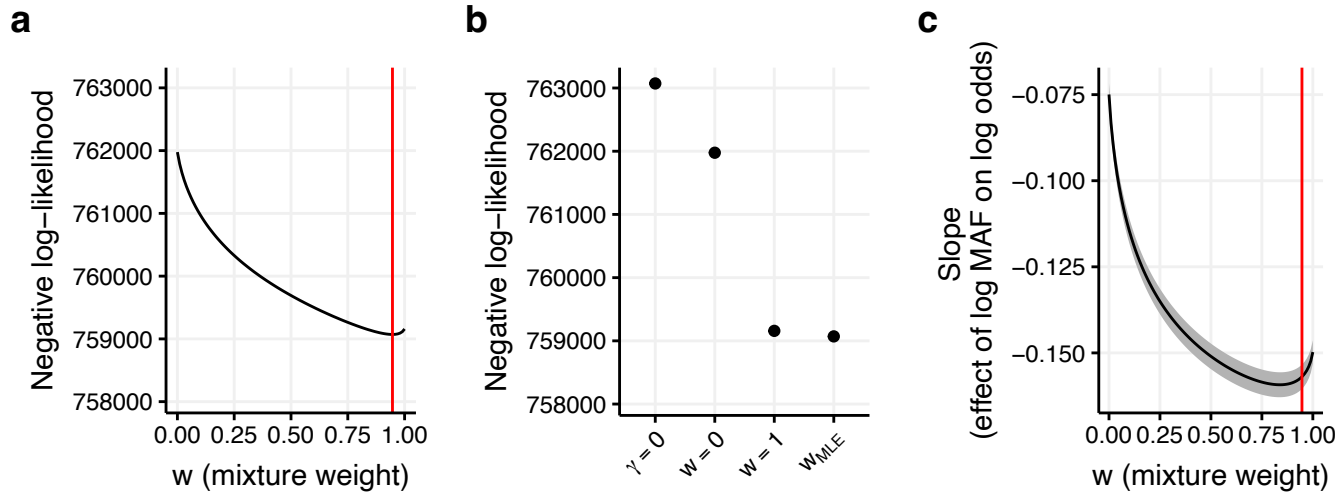



**Figure 2: Best-fit parameters for a logistic regression of non-synonymous status on log($p_{mix}$).** We report **(a)** the negative profile log-likelihood for various models, including a model without MAF dependence ($\gamma$=0), a model with European MAF dependence ($w$=0), a model with African MAF dependence ($w$=1) and the mixture model from equation 2 ($w_{MLE}$), **(b)** the negative profile log-likelihood with respect to $w$, treating the regression intercept and slope as nuisance parameters, and **(c)** the regression slope ($\gamma$) for values of $w$ between 0 and 1. Red lines in (b) and (c) denote the MLE estimate of $w$. The shaded region in (c) denotes 95% confidence intervals for $\gamma$. Numerical results are reported in Supplementary Table 7.

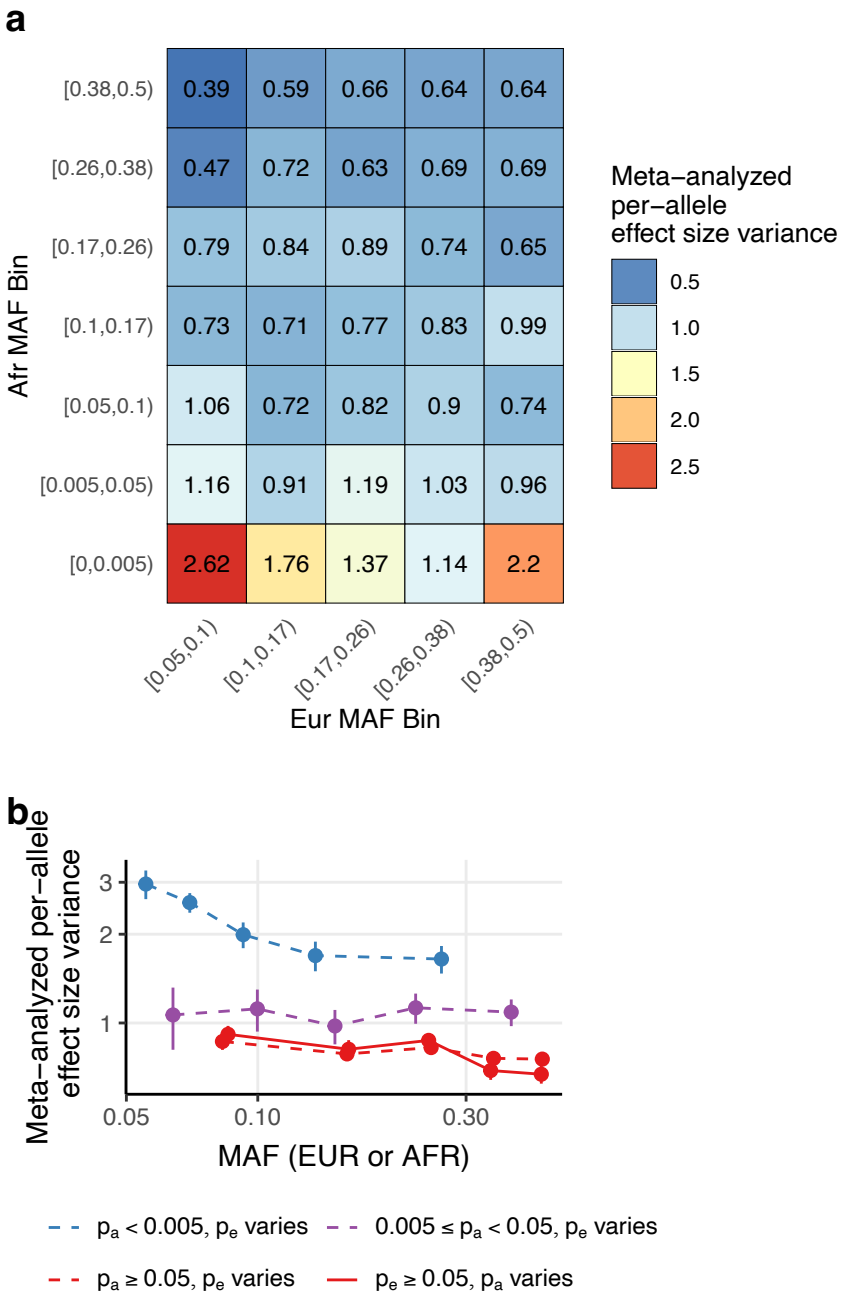



**Figure 3: Per-allele effect size variance across 50 diseases/traits as a function of African and European minor allele frequencies.** We report meta-analyzed scaled mean effect size variance as a function of **(a)** bivariate MAF bins defined by 7 African and 5 European MAF bins, and **(b)** African (solid line) and European (dashed lines) MAF quintiles, stratified by European and African MAF categories, respectively. The choice of MAF bin boundaries in (a) is discussed in Methods. Error bars in (b) denote 95% confidence intervals; error bars are smaller than dot size in some cases. Numerical results, including standard errors for each bivariate bin in (a) and each bin in (b), are reported in Supplementary Table 8-9.

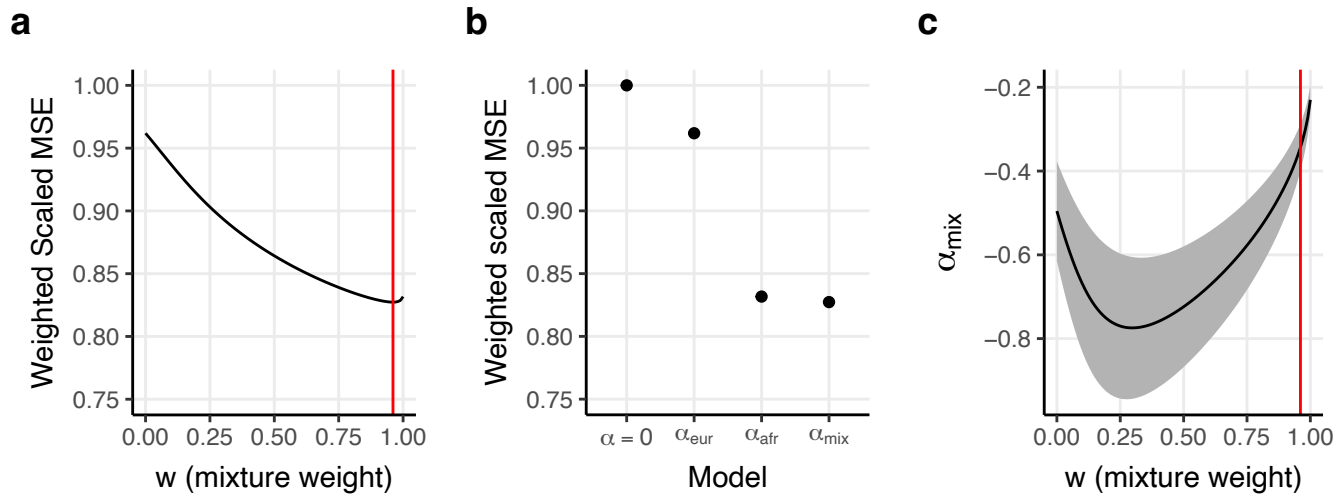



**Figure 4: Results of fitting the $\alpha_{mix}$ model across 50 diseases/traits.** We report **(a)** the weighted scaled mean squared error with respect to *w*, **(b)** the weighted scaled mean squared error for various models, including a model without MAF dependence ($\alpha$=0), a model with European MAF dependence ($\alpha_{eur}$), a model with African MAF dependence ($\alpha_{afr}$), and the $\alpha_{mix}$ model, and **(c)** estimates of $\alpha_{mix}$ as a function of *w*. Weighted scaled MSE weights the MSE for each bivariate MAF bin by the number of SNPs it contains and scales the sum over all traits such that the null model has a weighted scaled MSE of 1. Red lines in (a) and (c) denote our point estimate of *w*. The shaded region in (c) denotes 95% confidence intervals for $\alpha_{mix}$. Numerical results are reported in Supplementary Table 10.

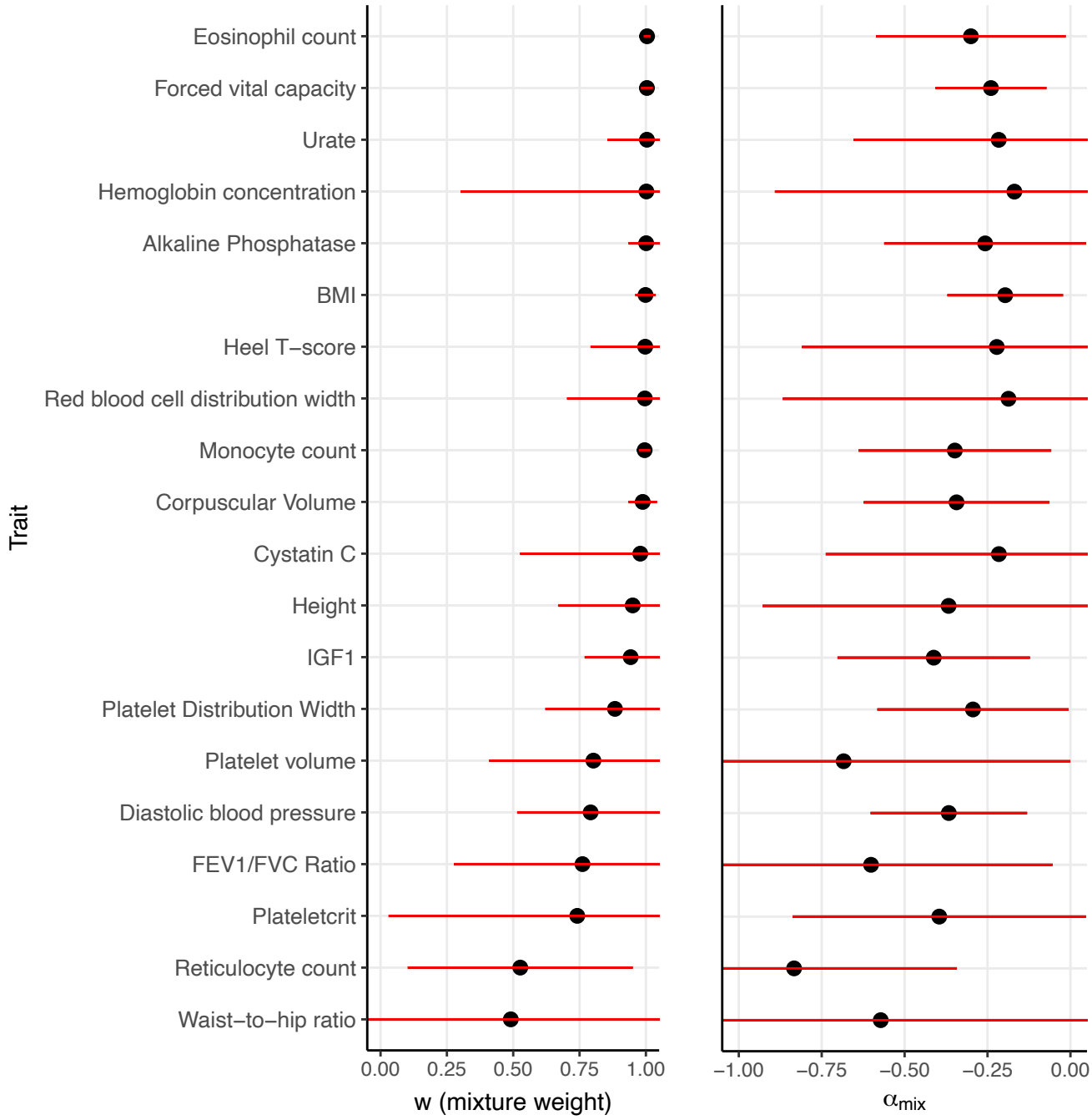



**Figure 5: Results of fitting the $\alpha_{mix}$ model separately for 20 out of 50 diseases/traits.** We report estimates of **(a)** *w* and **(b)** $\alpha_{mix}$. Diseases/traits are ordered by point estimate of *w*. The 20 diseases/traits with highest $N*h^2$ point estimates are displayed. Supplementary Figure 11 displays results for all 50. Error bars denote 95% confidence intervals. Numerical results for all 50 traits are reported in Supplementary Table 11.

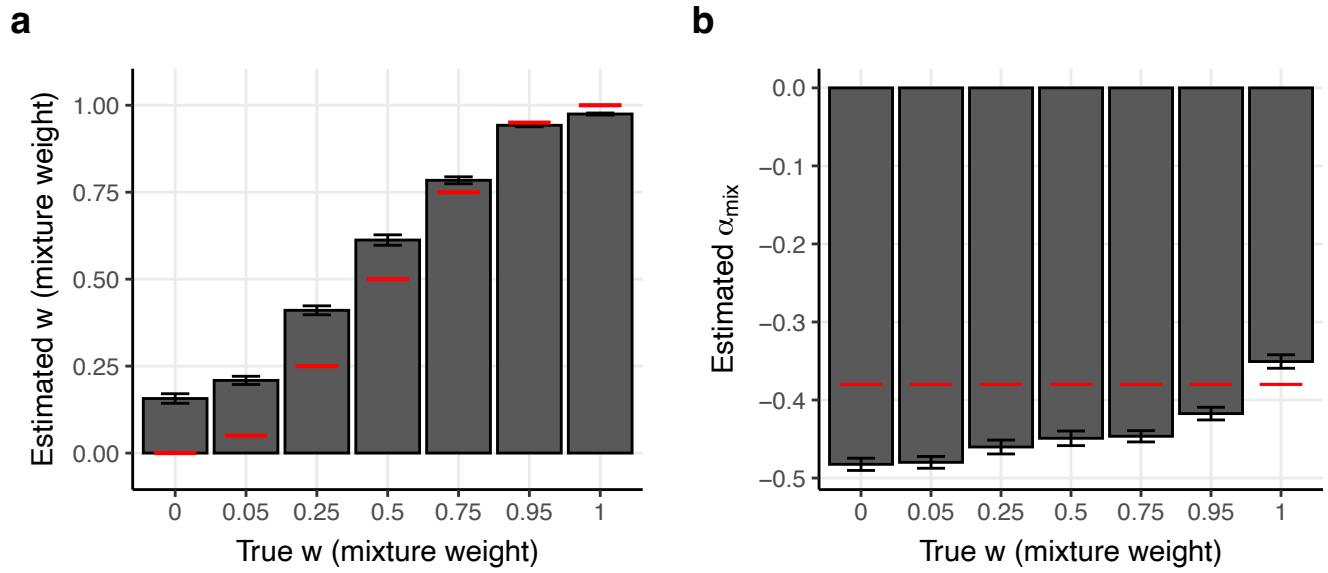



**Figure 6: Results of fitting the $\alpha_{mix}$ model in simulations.** We report **(a)** estimates of *w* and **(b)** $\alpha_{mix}$, at different values of true *w*. Bars denote mean estimates across 40 simulation replicates (each analyzing 25 traits). Red lines denote true parameter values. Error bars denote 95% confidence intervals. Numerical results are reported in Supplementary Table 12.

**Supplementary Tables:**

**Supplementary Table 1: Overview of 50 diseases/complex traits analyzed.** We report the SNP heritability ($h^2$) as estimated by LDSC and GWAS sample size.
(see .xlsx file)

**Supplementary Table 2: The proportion of non-synonymous SNPs in a grid of bivariate African and European MAF bins.** We report the proportion of non-synonymous SNPs, the proportion of synonymous SNPs, the number of SNPs, and the Wald standard error on the proportion of non-synonymous SNPs.
(see .xlsx file)

**Supplementary Table 3: The proportion of non-synonymous SNPs in African MAF deciles stratified by European MAF bins.** We report, for each decile, the African MAF, the proportion of non-synonymous variants, 95% confidence intervals on the proportion of non-synonymous variants, and the European MAF bin. We restricted to variants with $p_A \geq 0.002$
(see .xlsx file)

**Supplementary Table 4: Predictions from logistic regressions of SNP non-synonymous status on African MAF stratified by European MAF bins.** We report the predicted proportion of non-synonymous variants (and 95% confidence intervals) for a grid of 30,000 African MAF values from 0.002 - 0.5. In the regression African MAF was log transformed. We restricted to variants with $p_A \geq 0.002$.
(see .xlsx file)

**Supplementary Table 5: The proportion of non-synonymous SNPs in European MAF deciles stratified by African MAF bins.** We report, for each decile, the European MAF, the proportion of non-synonymous variants, 95% confidence intervals on the proportion of non-synonymous variants, and the African MAF bin. We restricted to variants with $p_E \geq 0.002$
(see .xlsx file)

**Supplementary Table 6: Predictions from logistic regressions of SNP non-synonymous status on European MAF stratified by African MAF bins.** We report the predicted proportion of non-synonymous variants and 95% confidence intervals for a grid of 30,000 European MAF values from 0.002 - 0.5. In the regression European MAF was log transformed. We restricted to variants with $p_E \geq 0.002$.
(see .xlsx file)

**Supplementary Table 7: Best-fit parameters for a logistic regression of non-synonymous status on log($p_{mix}$).** We report the log-likelihood, $\gamma$ estimate, $\gamma$ standard error, and 95% confidence interval on $\gamma$ across 101 values of $w$ from 0 to 1.
(see .xlsx file)

**Supplementary Table 8: Per-allele effect size variance across 50 diseases/traits for a grid of bivariate African and European MAF bins.** We report point estimates and standard errors of the meta-analyzed scaled mean effect size variance for each bin.
(see .xlsx file)

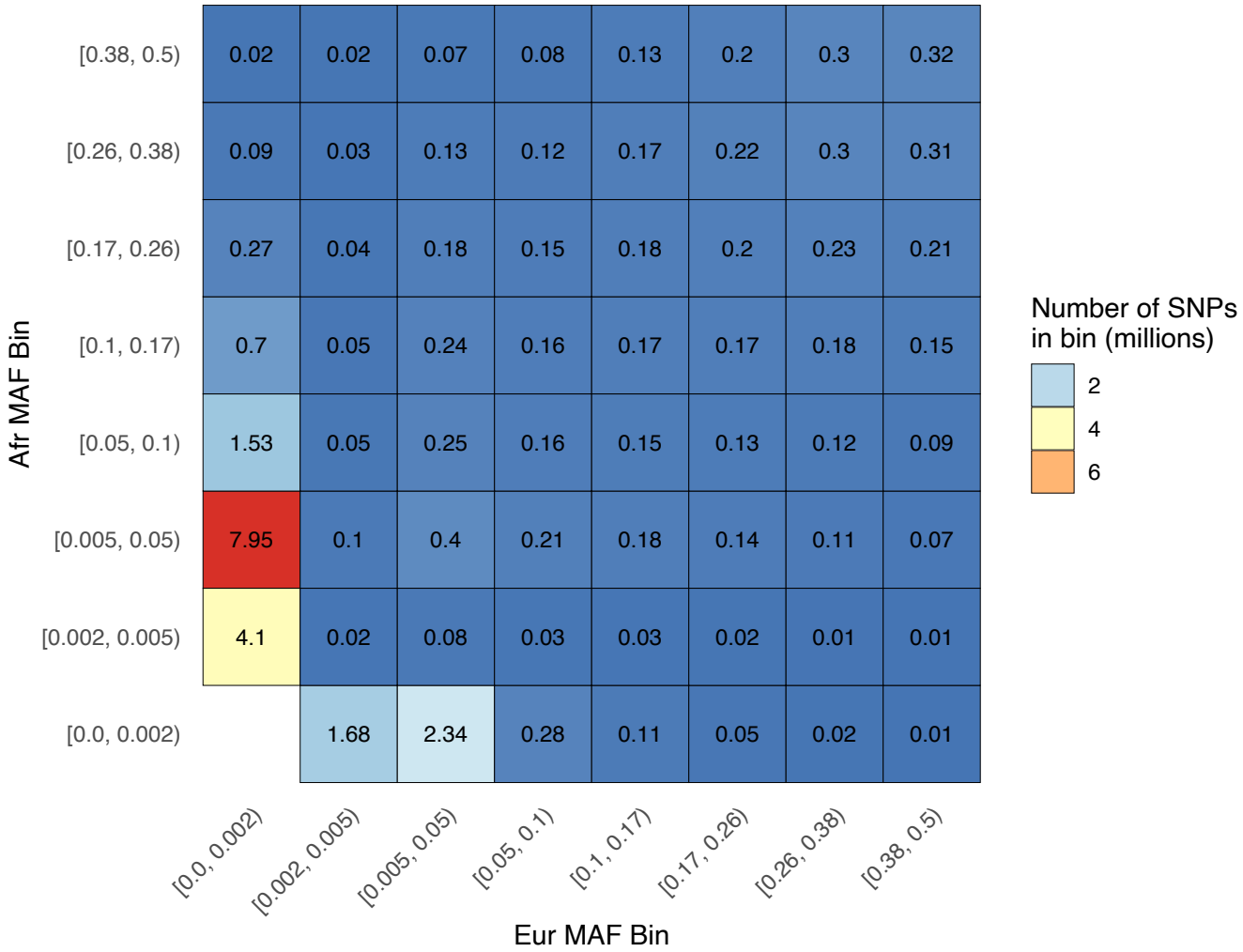
**Supplementary Table 9: Per-allele effect size variance across 50 diseases/traits as a function of MAF quintiles, stratified by African MAF.** We report, for each MAF quintile, the mean MAF, point estimate of the meta-analyzed scaled mean effect size variance, standard error of the meta-analyzed scaled mean effect size variance, and 95% confidence interval on the meta-analyzed scaled mean effect size variance.
(see .xlsx file)

**Supplementary Table 10: Results of fitting the $\alpha_{mix}$ model across 50 diseases/traits.** We report point estimates and standard errors on $\alpha_{mix}$ and point estimates and standard errors of the loss function across 101 values of *w* from 0 to 1.
(see .xlsx file)

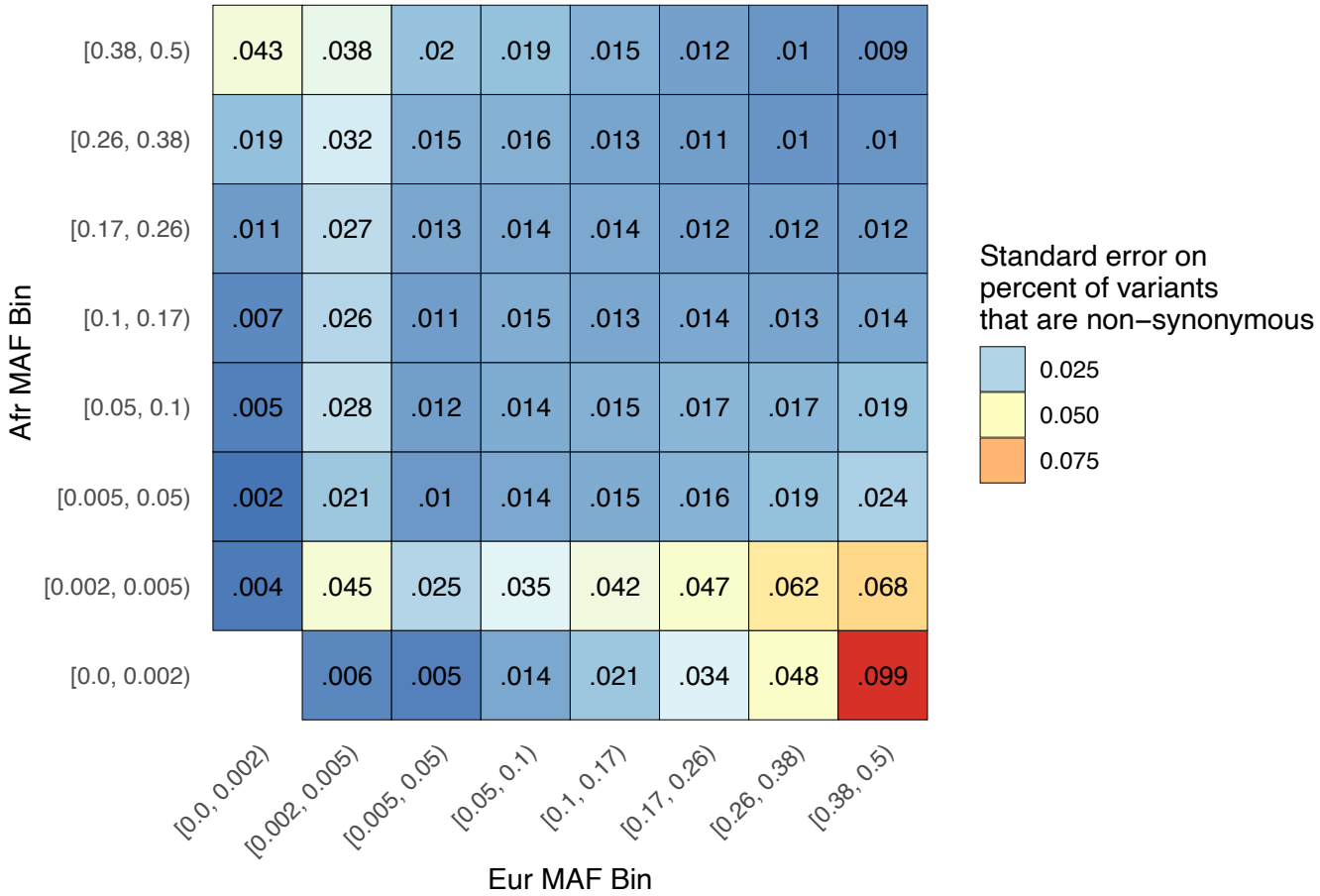
**Supplementary Table 11: Results of fitting the $\alpha_{mix}$ model separately for 50 diseases/traits.** We report point estimates and standard errors on *w*, $\alpha_{mix}$, and the loss function for each of 50 diseases/traits.
(see .xlsx file)

**Supplementary Table 12: Results of fitting the $\alpha_{mix}$ model in simulations.** We report point estimates and standard errors on the mean estimate of *w* and $\alpha_{mix}$.
(see .xlsx file)

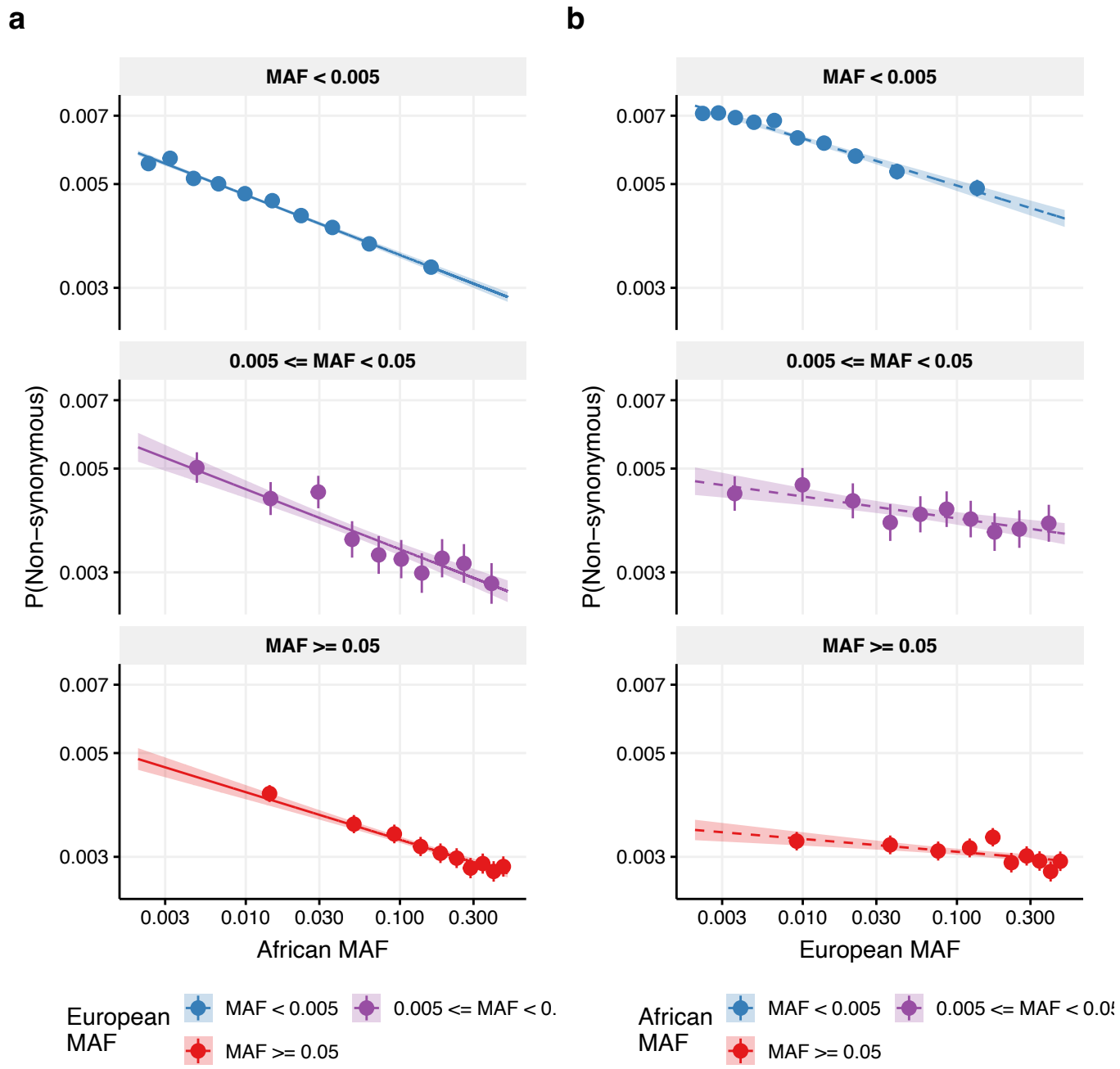
500 **Supplementary Figures:**



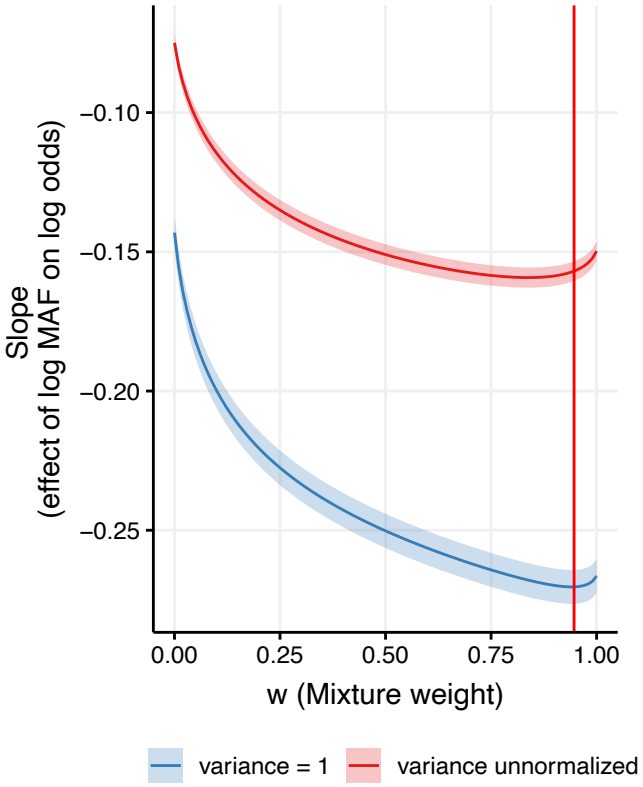

501
502 **Supplementary Figure 1: Number of SNPs in each AoU MAF stratum.** We report the number of
503 SNPs contained in each grid element given in Figure 1a. Each grid element contains SNPs with
504 unadmixed African and European MAF that fall in a pair of MAF intervals. A justification for MAF
505 interval boundaries is given in Methods.



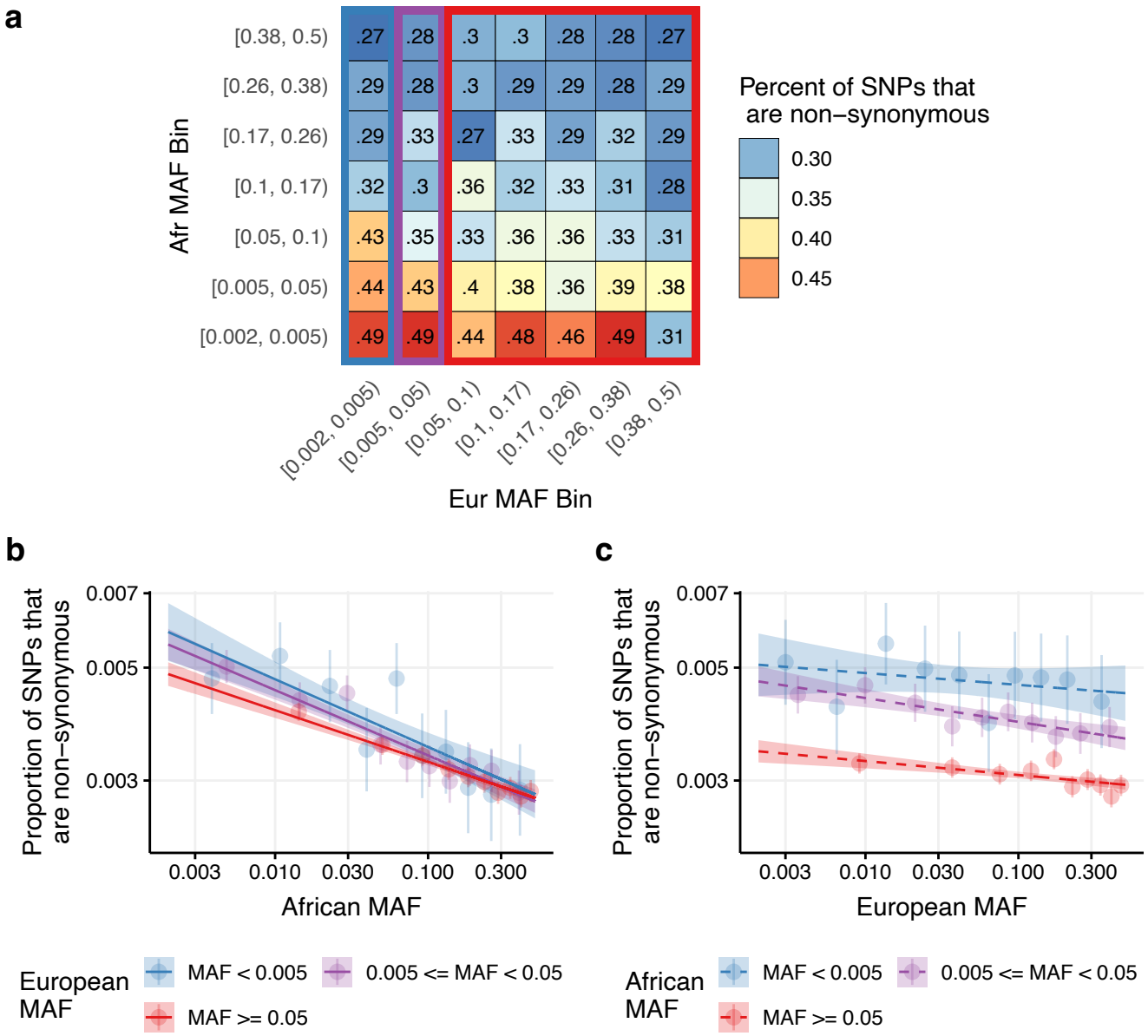

**Supplementary Figure 2: Standard error on the proportion of non-synonymous SNPs in each MAF stratum.** We report the binomial proportion standard error for each grid element given in Figure 1a. Each grid element contains SNPs with unadmixed African and European MAF that fall in a pair of MAF intervals. A justification for MAF interval boundaries is given in Methods.



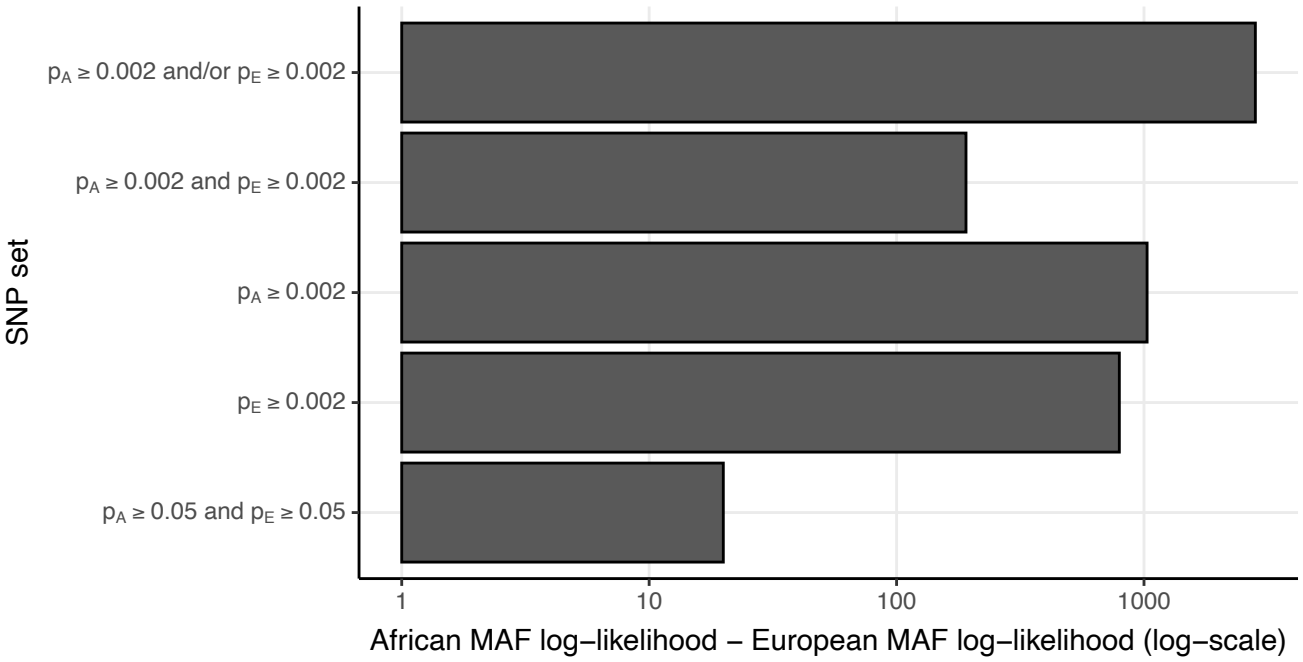

**Supplementary Figure 3: Logistic regression for non-synonymous status as a function of unadmixed African and European minor allele frequencies separated by MAF category.** We report **(a)** the probability that a SNP is non-synonymous as a function of African MAF after stratifying by European MAF categories (for SNPs with $p_A \geq 0.002$) as predicted by logistic regression, and **(b)** the probability that a SNP is non-synonymous as a function of European MAF after stratifying by African MAF categories (for SNPs with $p_E \geq 0.002$) as predicted by logistic regression. Error bars denote 95% Agresti-Coull confidence intervals. Shaded regions in (b) and (c) denote 95% confidence intervals on the mean estimate. Logistic regression fits match non-parametric MAF deciles remarkably well, indicating that our conclusions are unlikely to be impacted by model misspecification.



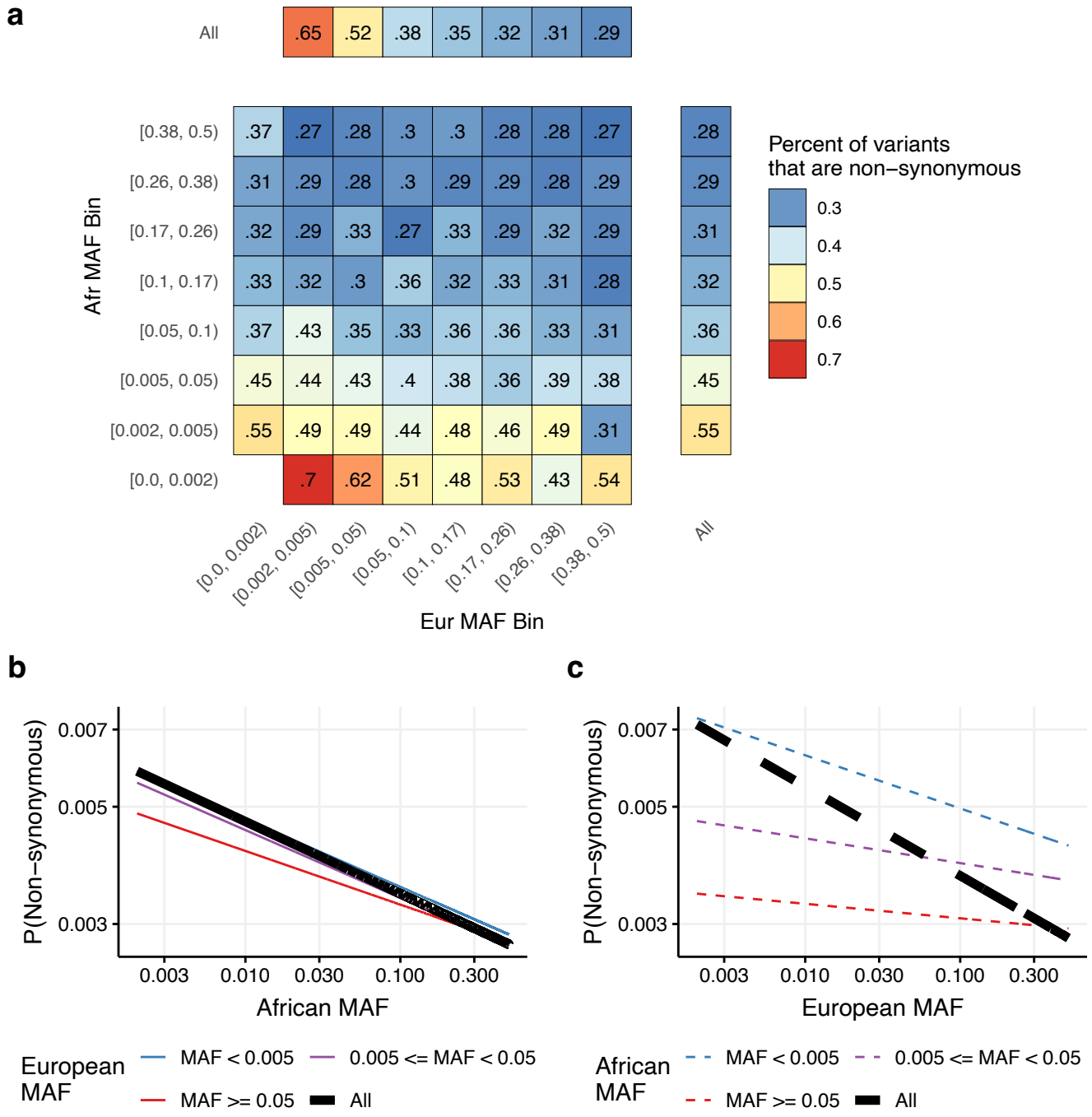

**Supplementary Figure 4: Best fit parameter for a logistic regression of non-synonymous status on log($p_{mix}$) while scaling log($p_{mix}$) to have unit variance.** We report the regression slope ($\gamma$) for values of w between 0 and 1. The blue curve denote results for a logistic regression on log($p_{mix}$) while scaling log($p_{mix}$). The red curve denotes results for a logistic regression on log($p_{mix}$) without scaling log($p_{mix}$). The vertical red line denotes the MLE of $w$=0.95 (the MLE of $w$ is not affected by scaling log($p_{mix}$)). With scaling, $w$=0.94 maximizes the absolute slope. Without scaling, $w$=0.84 maximizes the absolute slope.



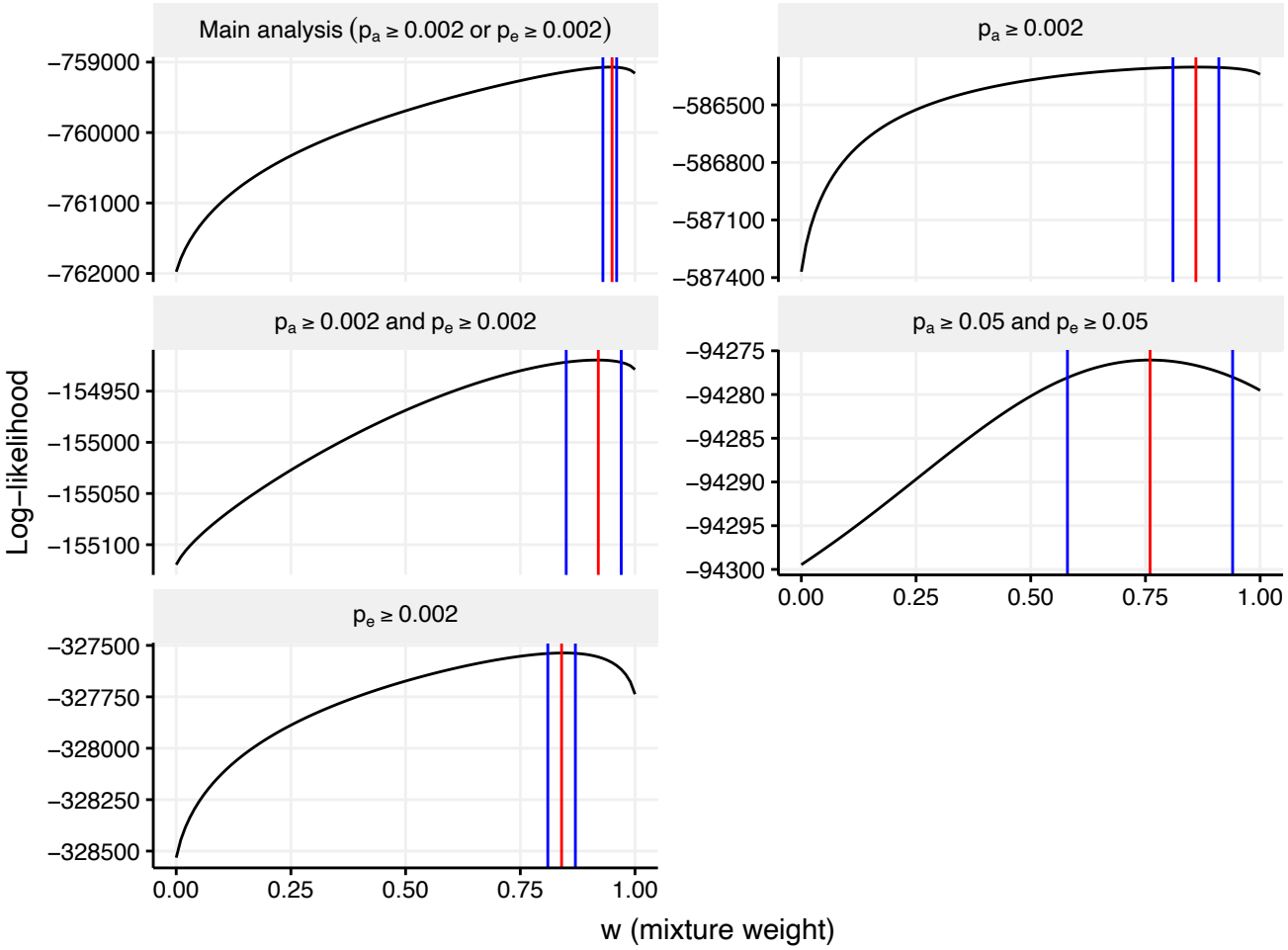

**Supplementary Figure 5: The proportion of non-synonymous SNPs as function of unadmixed African and European minor allele frequencies, restricted to variants with $p_A \geq 0.002$ and $p_E \geq 0.002$:** We report **(a)** the percentage of SNPs that are non-synonymous in bivariate bins defined by unadmixed African MAF (rows) and European MAF (columns), **(b)** the probability that a SNP is non-synonymous as a function of African MAF after stratifying by European MAF categories (for SNPs with $p_A \geq 0.002$ and $p_E \geq 0.002$) as predicted by logistic regression, and **(c)** the probability that a SNP is non-synonymous as a function of European MAF after stratifying by African MAF categories (for SNPs with $p_E \geq 0.002$ and $p_A \geq 0.002$) as predicted by logistic regression. The blue, purple, and red rectangles in (a) correspond to the SNPs included in the blue, purple, and red lines in (b). The choice of MAF bin boundaries is discussed in Methods. Shaded regions in (b) and (c) denote 95% confidence intervals on the mean estimate. Error bars denote 95% Agresti-Coull confidence intervals.



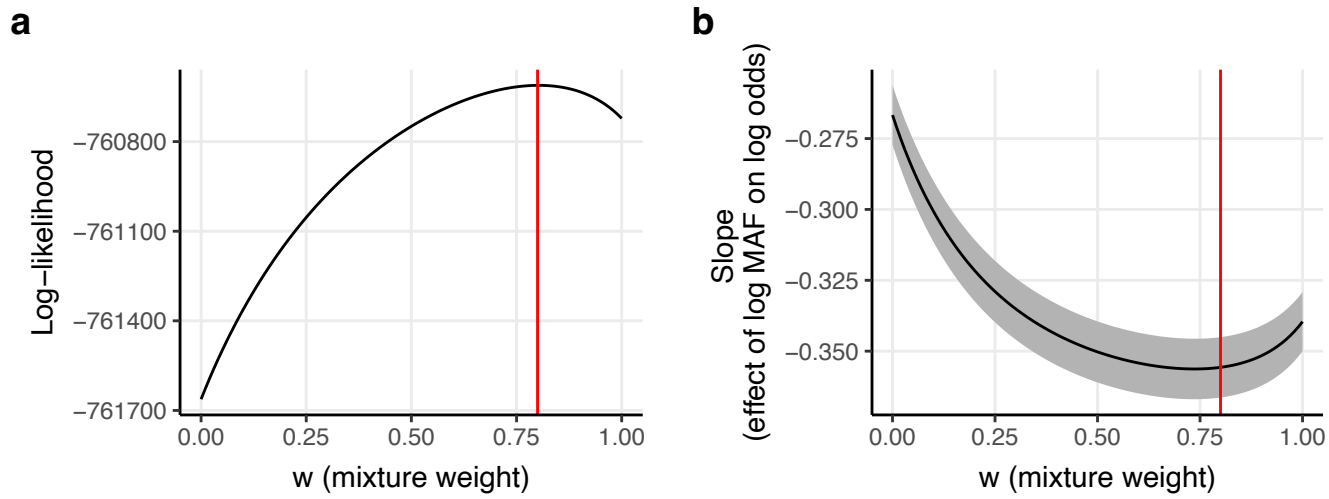

**Supplementary Figure 6: Difference in log-likelihood for a logistic regression of non-synonymous status on African MAF vs. a logistic regression of non-synonymous status on European MAF**. We report the log-likelihood for a logistic regression of PNS on log($p_A$) minus the log-likelihood for a logistic regression of PNS on log($p_E$), while varying the SNP set. All log-likelihood differences were significant via Vuong's test. For each SNP set, the exact same set of SNPs were included in the African MAF-only regression and the European MAF-only regression. For all SNPs, we thresholded MAF values at 0.002 after filtering for SNP set. That is, if $p_A < 0.002$, $p_A$ was set to 0.002, and if $p_E < 0.002$, $p_E$ was set to 0.002 (after restricting to the set of SNPs to be included in each SNP set).



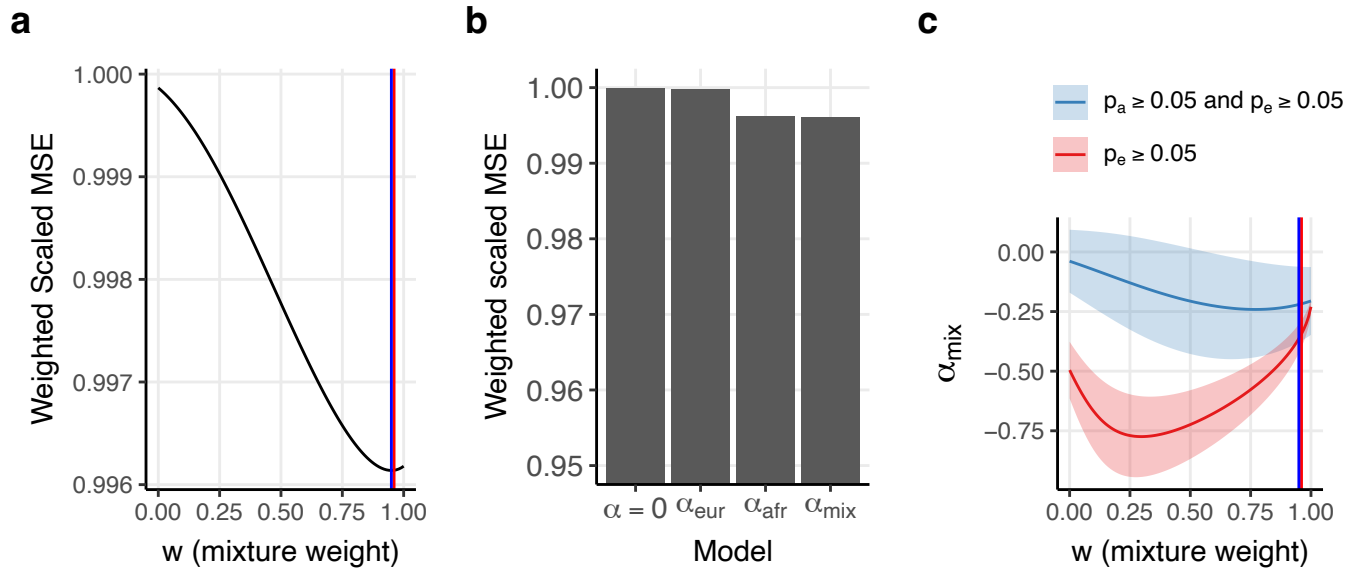

**Supplementary Figure 7: The proportion of non-synonymous SNPs as function of unadmixed African and European minor allele frequencies, highlighting the marginal effects of African MAF for SNPs with $p_A \geq 0.002$ and European SNPs with $p_E \geq 0.002$:** We report **(a)** the percentage of SNPs that are non-synonymous in univariate bins defined by unadmixed African MAF (rightmost column) and European MAF (top row) and bivariate bins defined by unadmixed African MAF (other rows) and European MAF (other columns), **(b)** the probability that a SNP is non-synonymous as a function of African MAF marginally (black) and after stratifying by European MAF categories (colored lines) (for SNPs with $p_A \geq 0.002$) as predicted by logistic regression, and **(c)** the probability that a SNP is non-synonymous as a function of European MAF marginally (black) and after stratifying by African MAF categories (colored lines) (for SNPs with $p_E \geq 0.002$) as predicted by logistic regression.



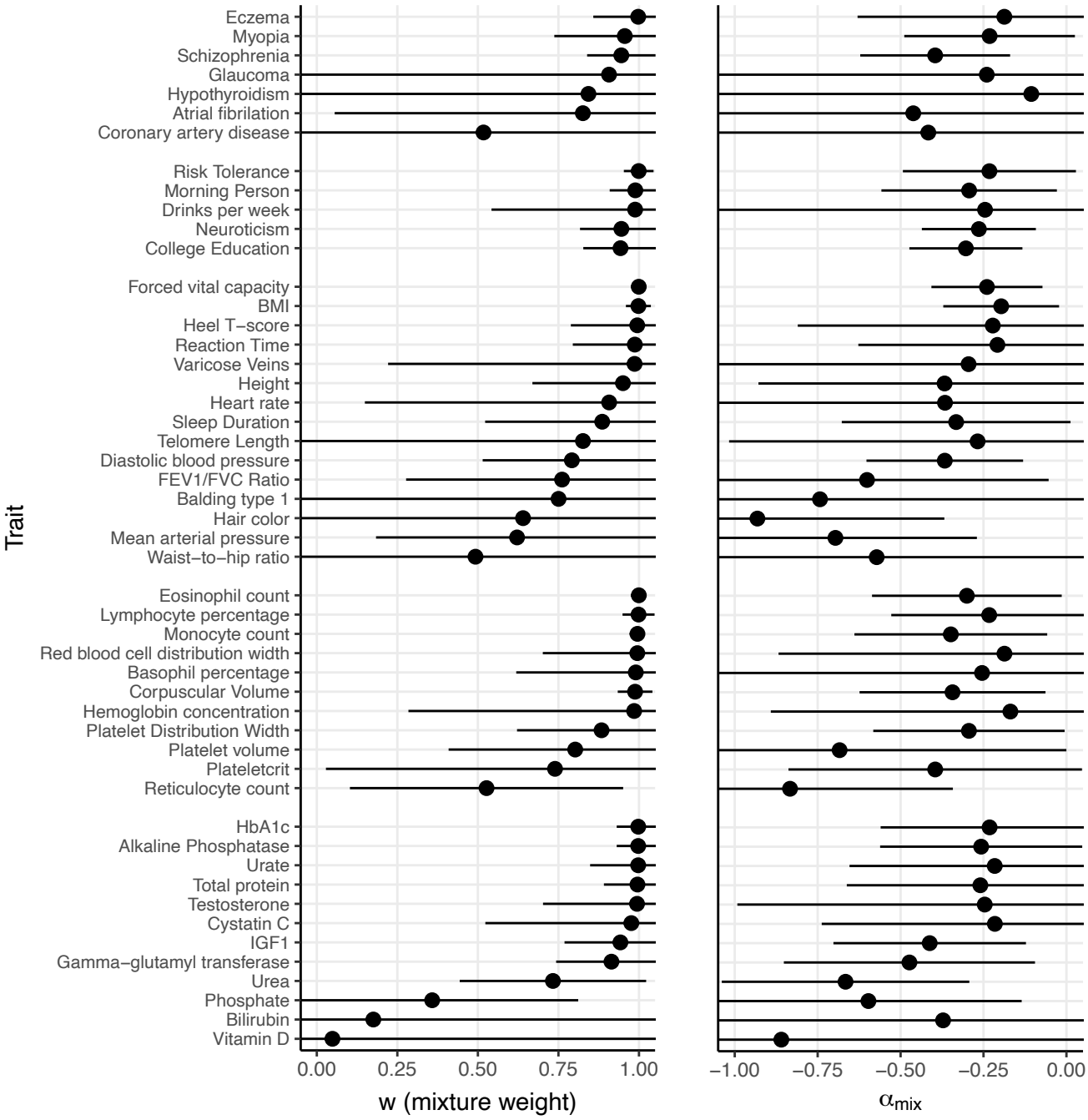

**Supplementary Figure 8: Log-likelihood for a logistic regression of non-synonymous status on log($p_{mix}$), varying the set of SNPs analyzed.** We report the profile log-likelihood with respect to *w*, treating the regression intercept and slope as nuisance parameters. Each curve corresponds to a distinct set of SNPs included while fitting the regression. Red lines denotes the MLE for *w*. Blue lines denote 95% confidence interval bounds for *w*.



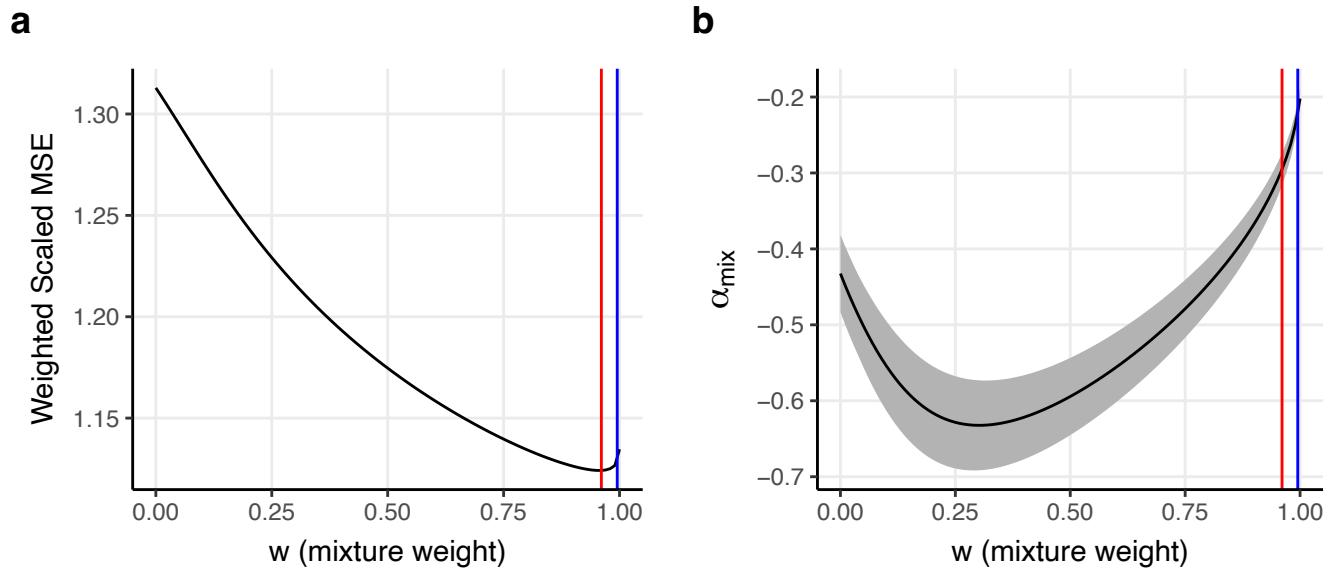

**Supplementary Figure 9: Best-fit parameters for a logistic regression of non-synonymous status on log($p_{mix}$), thresholding MAF at an alternate value of 0.05.** We report **(a)** the profile log-likelihood with respect to *w*, treating the regression intercept and slope as nuisance parameters, thresholding $p_A$ and $p_E$ at 0.05, and **(b)** the regression slope ($\gamma$) for values of *w* between 0 and 1, thresholding $p_A$ and $p_E$ at 0.05. If we "threshold" a SNP at a MAF of *c*, it means that if a SNP has $p_A$ < c, we set $p_A$ to c, and if a SNP has $p_E$ < c, we set $p_E$ to c. In the main analyses, we threshold all MAF values at 0.002 (after filtering). Red lines in (a) and (b) denote the MLE estimate of *w*. The shaded region in (b) denotes 95% confidence intervals for $\gamma$.



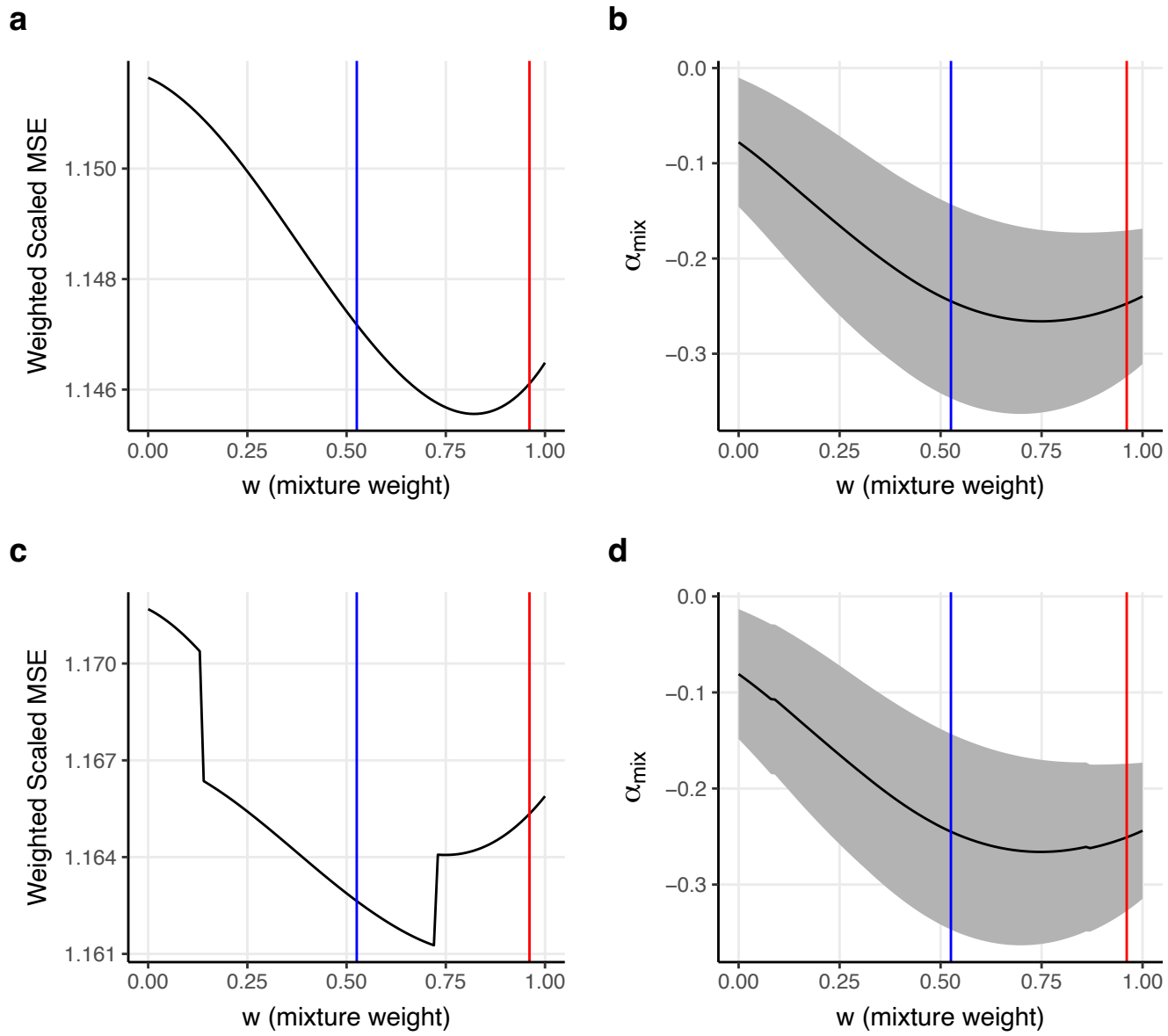

**Supplementary Figure 10: Results of fitting the $\alpha_{mix}$ model across 50 diseases/traits restricted to variants with $p_A \geq 0.05$ and $p_E \geq 0.05$.** We report **(a)** the weighted scaled mean squared error with respect to *w*, **(b)** the weighted scaled mean squared error for various models, including a model without MAF dependence ($\alpha$=0), a model with European MAF dependence ($\alpha_{eur}$), a model with African MAF dependence ($\alpha_{afr}$), and the $\alpha_{mix}$ model, and **(c)** estimates of $\alpha_{mix}$ as a function of *w*. Weighted scaled MSE weights the MSE for each bivariate MAF bin by the number of SNPs it contains and scales the sum over all traits such that the null model has a weighted scaled MSE of 1. Vertical lines in (a) and (c) denote point estimates of *w*. Blue lines correspond to estimates restricted to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$ and red lines correspond to estimates from the main analyses (restricted to SNPs with $p_A \geq 0.05$). Shaded regions in (c) denote 95% confidence intervals.



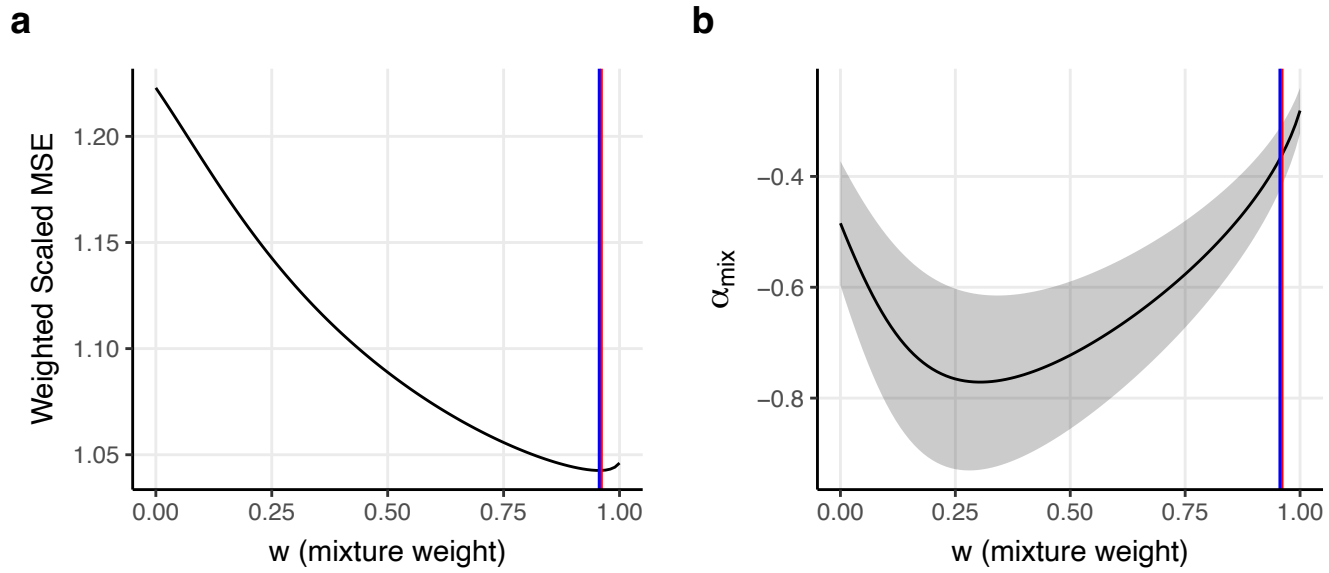

**Supplementary Figure 11: Results of fitting the $\alpha_{mix}$ model separately for 50 diseases/traits.** We report estimates of **(a)** *w* and **(b)** $\alpha_{mix}$. Diseases/traits are ordered by point estimate of *w* within disease/trait categories. Error bars denote 95% confidence intervals. Numerical results for all 50 traits are reported in Supplementary Table 11.



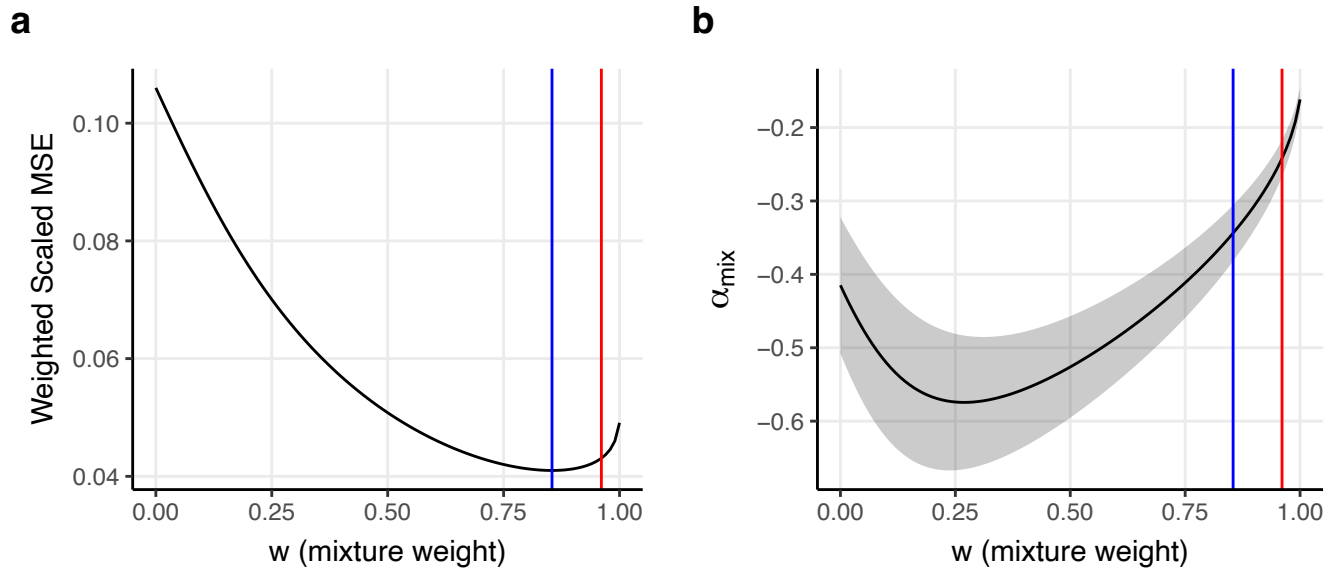

**Supplementary Figure 12**: **Results of fitting the $\alpha_{mix}$ model across 50 diseases/traits via a random effects meta-analysis of disease/trait-specific alpha-mix parameters.** We report **(a)** the weighted scaled mean squared error, summed across traits, with respect to *w*, and **(b)** estimates of $\alpha_{mix}$ as a function of *w*. Blue lines denote our meta-analyzed point estimate of *w*. Red lines denote the point estimate of *w* from the main analysis. The shaded region in (b) denotes 95% confidence intervals for $\alpha_{mix}$. Weighted scaled MSE was scaled by an arbitrary constant. The random effects meta-analysis estimate of *w* is not guaranteed to minimize the MSE because *w* is estimated separately for each trait, and some traits have MSE functions which are minimized at values other than the minimum of the shared MSE.



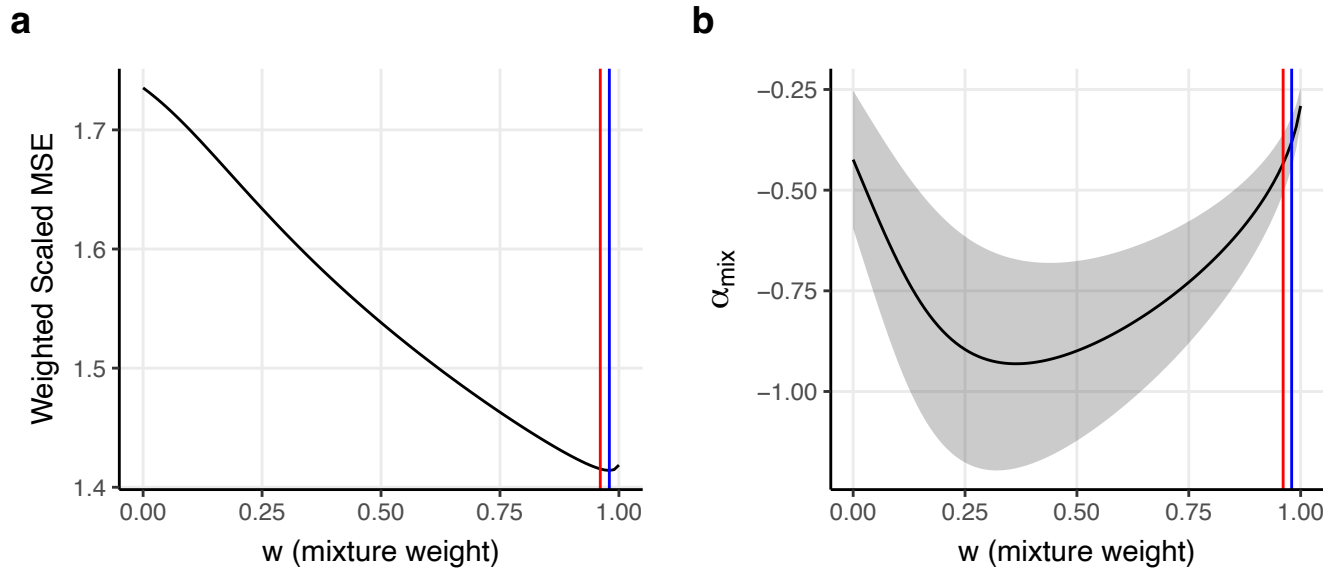

**Supplementary Figure 13**: **Results of fitting the $\alpha_{mix}$ model across 50 diseases/traits via a random effects meta-analysis of disease/trait-specific alpha-mix parameters restricted to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$.** We report **(a)** the weighted scaled mean squared error with respect to *w*, summed across diseases/traits while restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$, excluding UKB_460K.biochemistry_Phosphate, **(b)** $\alpha_{mix}$ as a function of *w*, estimated using a random effects meta-analysis across traits while restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$ excluding UKB_460K.biochemistry_Phosphate, **(c)** same as (a), including UKB_460K.biochemistry_Phosphate, and **(d)** same as (b), including UKB_460K.biochemistry_Phosphate. Blue lines denote our meta-analyzed point estimate of *w*. Red lines denote the point estimate of *w* from the main analysis. The shaded regions in (b,d) denote 95% confidence intervals for $\alpha_{mix}$.
The random effects meta-analysis estimate of *w* is not guaranteed to minimize the MSE because *w* is estimated separately for each trait, and some traits have MSE functions which are minimized at values other than the minimum of the shared MSE function.
Inclusion of UKB_460K.biochemistry_Phosphate had very little effect on our meta-analyzed estimates of *w* (difference of .0052) and $\alpha_{mix}$ (difference of .0040).
Weighted scaled MSE was scaled by an arbitrary constant.



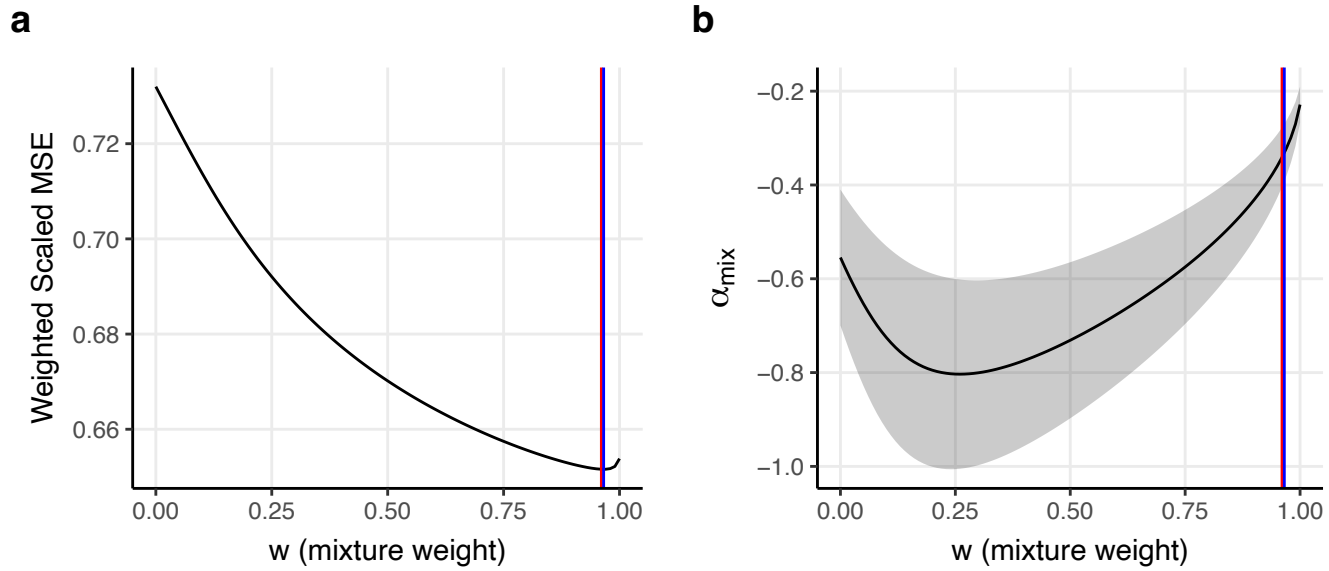

**Supplementary Figure 14: Results of fitting the $\alpha_{mix}$ model across 50 diseases/traits while estimating African MAF using *N*=107 Yoruban genomes.** We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of $\alpha_{mix}$ as a function of *w*. Blue lines denote our point estimate of *w* while estimating African MAF using *N*=107 Yoruban genomes. Red lines denote the point estimate of *w* from the main analysis. The shaded region in (b) denotes 95% confidence intervals for $\alpha_{mix}$. Yoruban genomes were used to define the 2-dimensional MAF bins which are used both while estimating per-allele effect size variance using S-LDSC and while estimating the parameters of the $\alpha_{mix}$ model.
Weighted scaled MSE was scaled by an arbitrary constant.



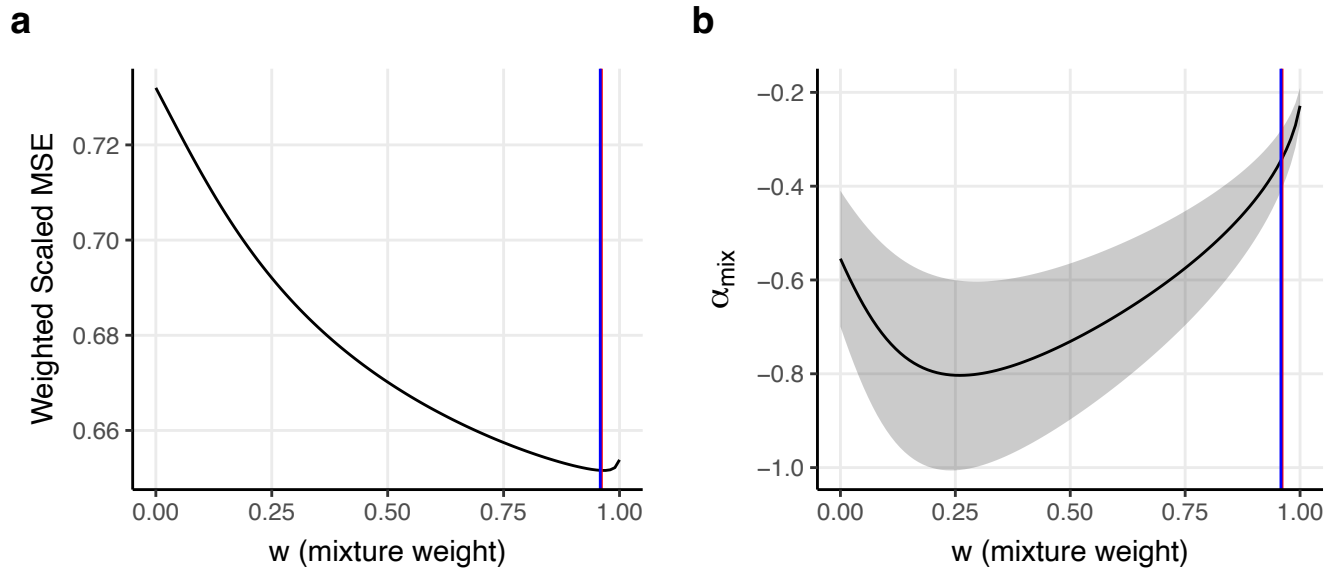

**Supplementary Figure 15: Results of fitting the $\alpha_{mix}$ model across 50 diseases/traits using the standard Baseline-LD (v2.2) model without annotations for African MAF.** We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of $\alpha_{mix}$ as a function of *w*. Blue lines denote our point estimate of *w* using the standard Baseline-LD (v2.2) model without annotations for African MAF. Red lines denote the point estimate of *w* from the main analysis. The shaded region in (b) denotes 95% confidence intervals for $\alpha_{mix}$. The European MAF bins included in the standard Baseline-LD (v2.2) model were retained.
Weighted scaled MSE was scaled by an arbitrary constant.



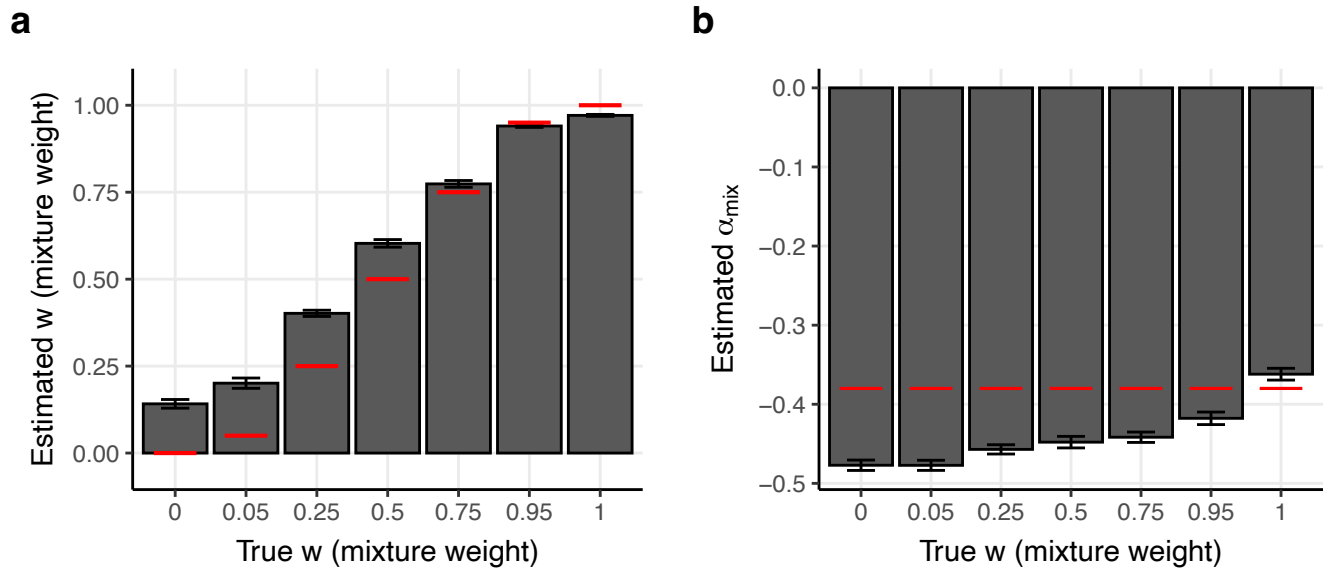

**Supplementary Figure 16: Results of fitting the $\alpha_{mix}$ model across 50 diseases/traits using 2-dimensional MAF bin annotations without including the BaselineLD-v2.2 model.** We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of $\alpha_{mix}$ as a function of *w*. Blue lines denote our point estimate of *w* using 2-dimensional MAF bin annotations without including the BaselineLD-v2.2 model. Red lines denote the point estimate of *w* from the main analysis. The shaded region in (b) denotes 95% confidence intervals for $\alpha_{mix}$. Weighted scaled MSE was scaled by an arbitrary constant.



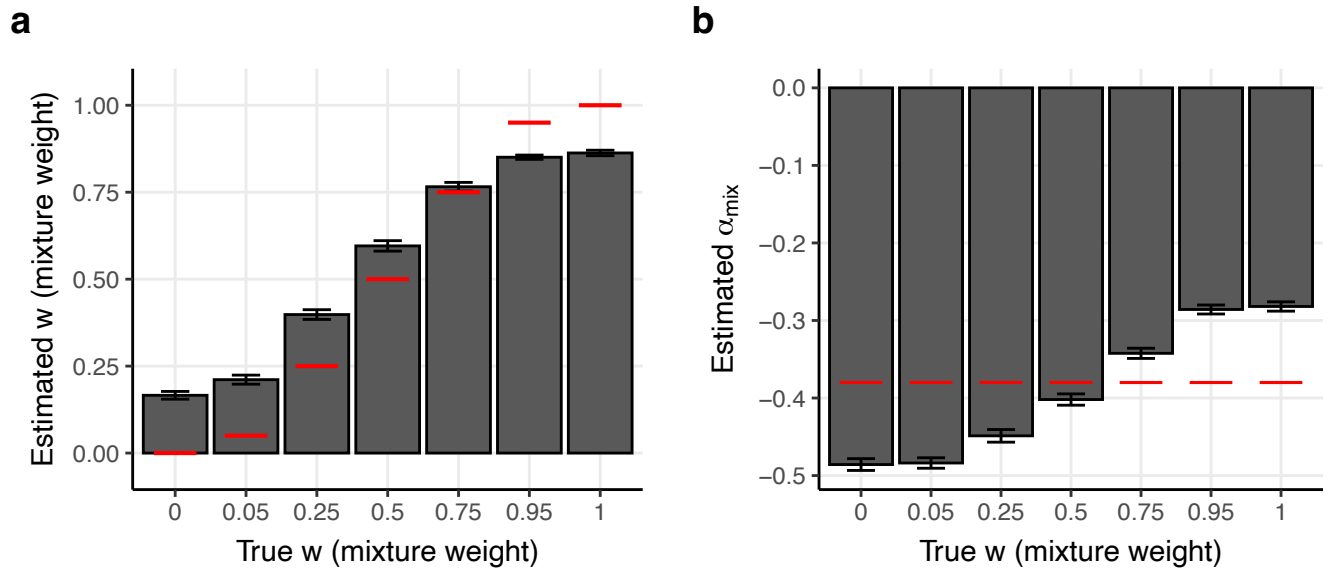

**Supplementary Figure 17: Results of fitting the $\alpha_{mix}$ model across 29 diseases/traits from UK Biobank.** We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of $\alpha_{mix}$ as a function of *w*. Blue lines denote our point estimate of *w* using 29 UK Biobank diseases/traits. Red lines denote the point estimate of *w* from the main analysis. The shaded region in (b) denotes 95% confidence intervals for $\alpha_{mix}$.
Weighted scaled MSE was scaled by an arbitrary constant.



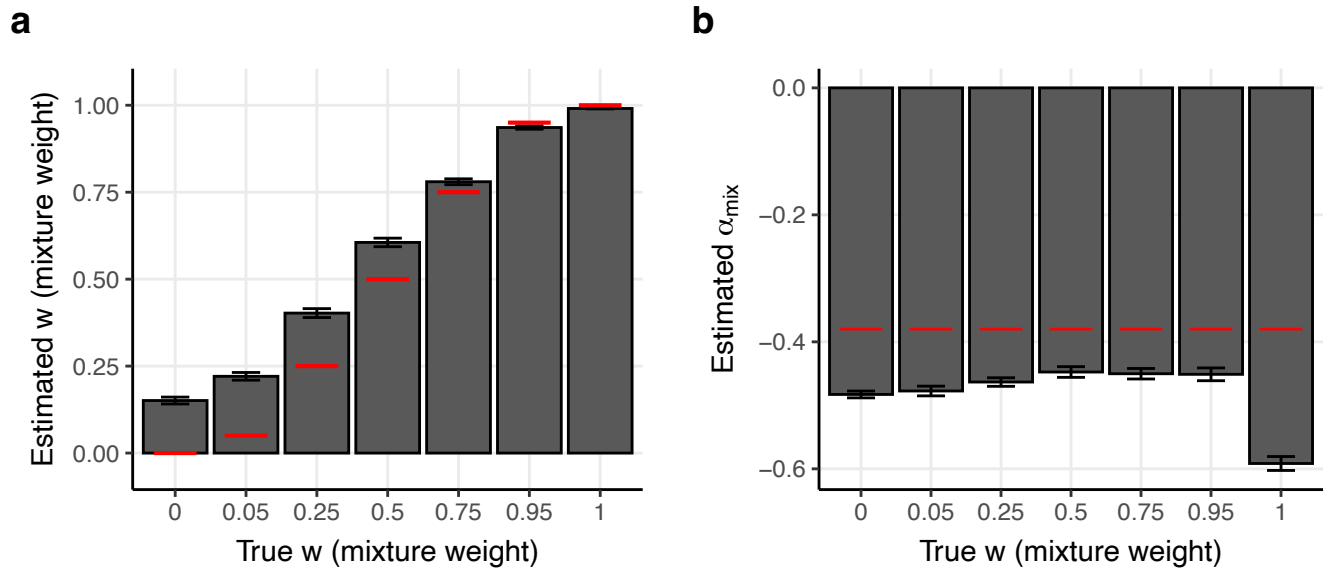

**Supplementary Figure 18: Results of fitting the $\alpha_{mix}$ model across 21 diseases/traits from publicly available summary statistics.** We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of $\alpha_{mix}$ as a function of *w*. Blue lines denote our point estimate of *w* using 21 diseases/traits from publicly available summary statistics (not including any diseases/traits from UK Biobank). Red lines denote the point estimate of *w* from the main analysis. The shaded region in (b) denotes 95% confidence intervals for $\alpha_{mix}$.
Weighted scaled MSE was scaled by an arbitrary constant.



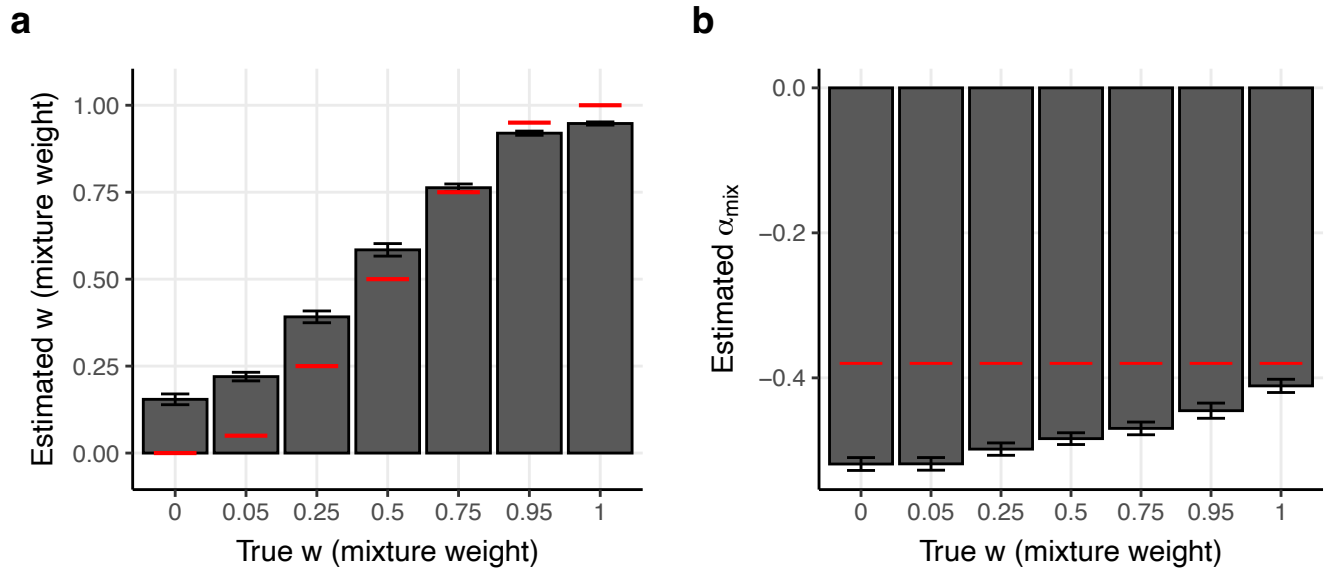

**Supplementary Figure 19: Results of fitting the $\alpha_{mix}$ model across groups of 25 traits in simulations, while using European MAF from UK Biobank for the generative model, and European MAF from 1000 Genomes for inference.** We report estimates of *w* **(a)** and $\alpha_{mix}$ **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value of each parameter. Error bars denote 95% confidence intervals. 1000 Genomes MAF was used both to define MAF bin annotations given to S-LDSC and during $\alpha_{mix}$ inference.



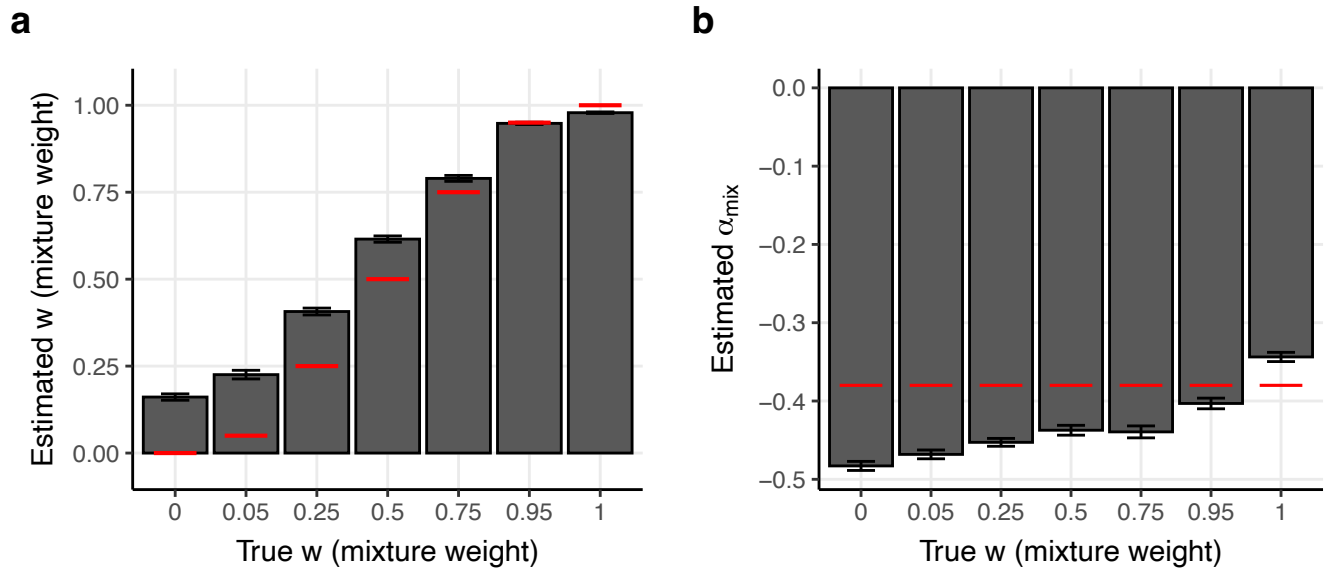

**Supplementary Figure 20: Results of fitting the $\alpha_{mix}$ model across groups of 25 traits in simulations, while thresholding MAF at *T*=0.05.** We report estimates of *w* **(a)** and $\alpha_{mix}$ **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value of each parameter. Error bars denote 95% confidence intervals. *T* is only used by the generative model, not by S-LDSC or in $\alpha_{mix}$ inference.



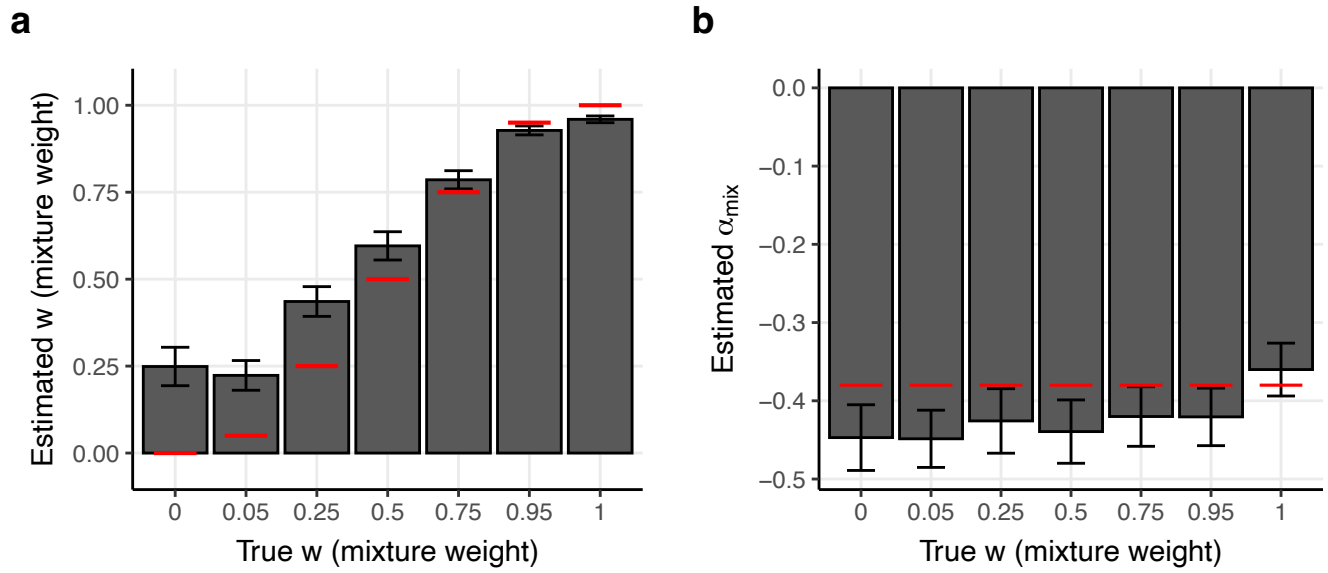

**Supplementary Figure 21: Results of fitting the $\alpha_{mix}$ model across groups of 25 traits in simulations, while thresholding MAF at *T*=1.5*10$^{-5}$ (equivalent to a minor allele count of 1 in our sample of genomes with local African ancestry from All of Us).** We report estimates of *w* **(a)** and $\alpha_{mix}$ **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value of each parameter. Error bars denote 95% confidence intervals. *T* is only used by the generative model, not by S-LDSC or in $\alpha_{mix}$ inference.




**Supplementary Figure 22: Results of fitting the $\alpha_{mix}$ model across groups of 25 traits in simulations for traits with 5,000 causal variants.** We report estimates of *w* **(a)** and $\alpha_{mix}$ **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value of each parameter. Error bars denote 95% confidence intervals.




**Supplementary Figure 23: Results of fitting the $\alpha_{mix}$ model across groups of 25 traits in simulations for traits with 20,000 causal variants.** We report estimates of *w* **(a)** and $\alpha_{mix}$ **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value of each parameter. Error bars denote 95% confidence intervals.




**Supplementary Figure 24: Results of fitting the $\alpha_{mix}$ model separately for each trait in simulations.** We report estimates of *w* **(a)** and $\alpha_{mix}$ **(b)**. Bars denote the mean estimate of each parameter across 100 replicates (each with a single trait). Red lines denote the true value of each parameter. Error bars denote 95% confidence intervals.

705 **Methods**

706 *Minor allele frequency estimation*

707 Throughout this manuscript we refer to unadmixed African MAF as African MAF. We used
708 continental ancestry group assignments provided by All of Us. Note that the African genetic
709 component of African ancestry individuals living in the United States is primarily of West African
710 origin[47] and, as a result, our African MAF estimates primarily represent West African ancestry. We
711 defined the MAF of a SNP in a specific ancestry to be the frequency of the less common (minor)
712 allele *in that ancestry*. As a result, for a single SNP, it is possible for the minor allele in Africans to
713 be different from the minor allele in Europeans.

714 We estimated unadmixed African MAF using All of Us[41] whole genome-sequencing data for
715 *N*=46,672 unrelated individuals labeled as African ancestry by All of Us (AoU). This cohort has
716 moderate levels of European admixture on average, so estimating unadmixed allele frequencies
717 (AF) is more complex than in an unadmixed cohort[73]. We restricted to 30,906,411 single-nucleotide
718 polymorphisms with either European MAF ≥ 0.001 or admixed African MAF ≥ 0.001 (as reported by
719 the Variant Annotation Table) in the All of Us v8 short read whole-genome sequencing Allele
720 Count/Allele Frequency threshold Plink dataset. We first estimated the local ancestry of each
721 individual using RFMix2 [49] with 1000 Genomes[43] as the reference dataset, restricting to HapMap3
722 variants. The local ancestry of SNPs that were not included in HapMap3 was interpolated between
723 that of the closest HapMap3 SNPs. Next, we computed the unadmixed African allele frequency of
724 each SNP using Plinkv1.9[74] while restricting to individuals with two African haplotypes.

725 We estimated unadmixed European MAF in 70,346 unrelated European ancestry individuals
726 from the v8 short read whole-genome sequencing Allele Count/Allele Frequency threshold Plink
727 dataset, restricting to the same set of variants used for African MAF estimation. These European
728 MAF estimates were used for analyses of non-synonymous variation (Figures 1-2), but not
729 analyses for analyses of diseases/traits (Figures 3-6).

730 For S-LDSC analyses of diseases/traits, unadmixed African AF was estimated as described
731 above, but using the All of Us v7.1 short read whole-genome sequencing Allele Count/Allele
732 Frequency threshold Plink dataset. AF were converted from GRCh38 to GRCh37 using
733 BCFtools/liftover[75]. The African MAF of SNPs in 1000 Genomes that could not be matched to any
734 of the SNPs from AoU, was calculated using *N*=107 individuals from Yoruba in 1000 Genomes
735 restricting to unrelated, QC'ed individuals from TGP2261 from ref.[76]. For S-LDSC analyses of
736 diseases/traits, European AF was calculated using *N*=489 European ancestry individuals from
737 1000 Genomes restricting to unrelated, QC'ed individuals from TGP2261 from ref.[76]. We used
738 1000 Genomes rather than the All of Us to ensure consistency between our MAF estimates and
739 those used in the Baseline-LD (v2.2) model, where MAF and LD scores are calculated using the
740 same 489 individuals from 1000 Genomes.
741 Throughout this manuscript, we stratified variants by African and European allele frequency
742 using a common set of bin boundaries, which are justified as follows. SNPs were first stratified by
743 common (MAF ≥ 0.05), low-frequency (0.005 ≤ MAF < 0.05), and rare (MAF < 0.005) categories. In
744 some analyses, we further stratified common variants into 5 or 10 bins with boundaries
745 corresponding to 1000 Genomes European ancestry common MAF quintiles or deciles as defined
746 by the Baseline-LD (v2.2) model. We used 5 common variant bins for visualization (e.g. Figures
747 1a, 3a) and in-text reporting of non-parametric MAF-dependence statistics (paragraph 1 of the *Per-*
748 *allele disease and complex trait effect sizes are predominantly African MAF-dependent in*
749 *European populations* subsection of Results) to improve interpretability. We used 10 common
750 variant bins when fitting the $\alpha_{mix}$ model (Figures 4-5) because using more bins decreases the
751 absolute difference between the MAF of each SNP in a bin and the mean MAF of the bin. In All of
752 Us analyses, rare SNPs were sometimes further stratified using a boundary of 0.002 to create a
753 rare SNP bin that avoids ascertainment bias (e.g. a SNP with unadmixed African MAF=0.0011 has
754 African American MAF around 0.0008 < 0.001 and thus would typically be excluded by our initial

MAF filter; however, a SNP with unadmixed African MAF=0.002 has African American MAF around 0.0016 > 0.001 and thus would typically be included in our dataset).

*Modeling non-synonymous status as a function of MAF*

We defined a set of non-synonymous SNPs in AoU by grouping the following values for the "consequence" field in the All of Us v8 Variant Annotation Table: frameshift_variant, inframe_deletion, inframe_insertion, missense_variant, start_lost, stop_gained, stop_lost. We computed the proportion of SNPs that are non-synonymous (PNS) in a stratum with respect to all SNPs, including SNPs in non-coding regions.

We modeled the probability that a SNP is non-synonymous as a function of $p_{mix}$ via the following logistic regression model:

$$\log\left(\frac{P(Y)}{1-P(Y)}\right) = \mu + \gamma \log(p_{mix}), \tag{2}$$

where $Y$=1 if the SNP is non-synonymous and 0 otherwise and $p_{mix}=w*p_A+(1-w)*p_E$. If $w$=0 or 1, this is equivalent to fitting a logistic regression using European or African, respectively. To estimate $w$, we computed a profile likelihood, treating $\mu$ and $\gamma$ as nuisance parameters, and used a numerical optimization method to maximize the profile likelihood with respect to $w$. The profile likelihood is given by:

$$L(w) = \max_{\gamma,\mu} L(w,\gamma,\mu), \tag{3}$$

where $L(w,\gamma,\mu)$ is the likelihood from equation 2. We calculated $\max_{\gamma,\mu} L(w,\gamma,\mu)$ for a specific choice of $w$ using discrete.discrete_model.Logit from the statsmodels[77] Python package and numerically identified the maximum of $L(w)$ with respect to $w$ using optimize.minimize_scalar from the SciPy Python package[78]. While fitting logistic regression models, values of $p_A$ and $p_E$ < 0.002 were set to 0.002, except for in Supplementary Figure 8, where a threshold of 0.05 is used.

Significance testing related to logistic regressions was performed in four ways depending on the specific analysis. We estimated P-values for $\gamma$ in $p_A$-only and $p_E$-only univariate logistic regressions using a two-tailed t-test. When comparing model fit for a $p_A$-only model to a $p_E$-only model on a shared set of SNPs, we computed p-values using Vuong's test because both models had the same number of degrees of freedom (precluding the use of a standard likelihood ratio test). When comparing $r^2$ and regression coefficients for a $p_A$-only model to a $p_E$-only model which were fit on different sets of SNPs, we computed p-values using a genomic block jackknife with 200 approximately equally sized blocks. All $p_{mix}$ logistic regression *P*-values and confidence intervals were computed by using a likelihood ratio test statistic with 1 degree of freedom. Throughout the manuscript, *P*-values less than $10^{-15}$ were reported as $P<10^{-15}$.

*Disease/Trait selection*

We first collected a set of 364 GWAS summary statistics in European ancestry cohorts across a wide range of complex traits and diseases. For each complex trait or disease, we estimated heritability using S-LDSC with baseline-LD (v2.2) LD scores. For each pair of complex traits and diseases, we estimated genetic correlation using S-LDSC[79] without partitioning based on functional annotations. Next, we filtered to traits with heritability z-score greater than 6 and $N*h^2$>20,000. Finally, we greedily selected traits based on $N*h^2$, subject to the constraint that for each pair of selected traits, the squared genetic correlation < 0.1.

*Estimation of per-allele effect size variance via stratified linkage disequilibrium score regression*

Stratified linkage disequilibrium score regression[3] (S-LDSC) estimates the common SNP heritability of a GWAS complex trait or disease partitioned across predefined functional annotations. In S-LDSC, the per-SNP heritability of SNP *i* is modeled as $h_i^2 = \sum_{d=1}^{D} a_d(i)\tau_d$, where *d* indexes a set of *D* functional annotations, $a_d(i)$ is the value of annotation *d* for SNP *i*, and $\tau_d$

803 represents the increase in $h_i^2$ associated with a unit increase in $a_d(i)$. S-LDSC requires the
804 specification of a set of *D* functional annotations that explain how heritability varies across the
805 genome.

806 We applied S-LDSC using annotations from a new extension of the Baseline-LD (v2.2)
807 model[5]. The Baseline-LD (v2.2) model includes standard functional annotations in addition to MAF
808 and linkage disequilibrium (LD) annotations that reduce MAF and LD-related bias. Our extension of
809 the Baseline-LD (v2.2) model incorporated new annotations based on bivariate MAF bins, where
810 each bivariate MAF bin is defined by an African MAF bin and a European MAF bin. Specifically, we
811 defined 12 univariate unadmixed African MAF bins: 10 for common African variants ($p_A \geq 0.05$), 1
812 for low-frequency African variants ($0.005 \leq p_A < 0.05$), and 1 for rare African variants ($p_A < 0.005$).
813 The boundaries of the 10 common variant bins were the same as those used for univariate
814 European common MAF in the Baseline-LD (v2.2) model. We then created 120 bivariate MAF bins
815 by stratifying SNPs based on the interaction between the 12 univariate African MAF bins and 10
816 univariate European MAF bins. We removed the 10 univariate European MAF bins from the
817 combined model to avoid collinearity of functional annotations. We calculated partitioned LD scores
818 using 489 European ancestry individuals from 1000 Genomes.

819 To estimate the *per-allele* effect size variance of variant *i*, we divided $\widehat{h^2}$ (as estimated by S-
820 LDSC) by the sample variance under Hardy-Weinberg equilibrium, $\widehat{\beta_i^2} = \frac{\widehat{h_i^2}}{2p_{E,i}(1-p_{E,i})}$, where $p_{E,i}$ is
821 the European MAF for SNP *i*. This procedure has been previously applied in ref.[80]. We calculated
822 the per-allele effect size variance for a specific trait and set of SNPs, *C*, as $\sum_{i \in C} \widehat{\beta_i^2}/n_C$, where $n_C$ is
823 the number of SNPs contained in *C*. When meta-analyzing across traits, we normalized $\widehat{\beta_i^2}$ within
824 each trait before taking the mean ($\frac{\sum_{i \in C} \widehat{\beta_i^2}/n_C}{\sum_i \widehat{\beta_i^2}/n}$, where *n* is the number of SNPs with $p_E > 0.05$), and
825 meta-analyzed these normalized quantities using random-effect meta-analysis as implemented in
826 stats.meta_analysis.combine_effects from the statsmodels Python package.

827

828 *Estimation of alpha-mix parameters*

829 We propose the $\alpha_{mix}$ model which relates per-allele effect size variance to unadmixed
830 African and European MAF:

831 $$\mathrm{E}[\beta_t^2 | p_A, p_E] = \sigma_{g,t}^2 [2p_{mix}(1 - p_{mix})]^{\alpha_{mix}} \quad (4)$$

832 where $\beta_t^2$ is the squared *per-allele* effect size of a single-nucleotide polymorphism (SNP) in trait *t*,
833 $\sigma_{g,t}^2$ is a variance parameter, $p_{mix}$=*w**$p_A$ + *(1-w)**$p_E$, *w* specifies the relative contributions of $p_A$ and
834 $p_E$, and $\alpha_{mix}$ specifies the level of MAF dependence. Equation 4 is identical to equation 1 with a
835 trait-specific subscript.

836 We estimated the parameters of the $\alpha_{mix}$ model via a generalized method of moments
837 estimator, with population moments from equation 4 and conditional sample moments estimated
838 via S-LDSC. We first partitioned variants with European MAF ≥ 0.05 into 120 two-dimensional MAF
839 bins based on the interaction between 12 African MAF bins and 10 European MAF bins. We then
840 defined one moment condition for each trait-bin pair: $E\left[\sigma_{g,t}^2\left[2\overline{p_{mix,k}}\left(1 - \overline{p_{mix,k}}\right)\right]^{\alpha_{mix}} - \widehat{\beta_{t,k}^2}\right] = 0$,
841 where $\widehat{\beta_{t,k}^2}$, is the S-LDSC estimate of the per-allele effect size variance (described in *Estimation of*
842 *per-allele effect size variance via stratified linkage disequilibrium score regression*) and $\overline{p_{mix,k}}$ is the
843 mean of $p_{mix}$ across all SNPs in bin *k*.

844 We next applied a numerical optimization procedure to minimize the sum of squared
845 moment conditions, weighted by the number of SNPs in each bin. When estimating alpha-mix
846 parameters for a single trait, we optimized the following loss function: $L_t\left(w, \alpha_{mix}, \sigma_{g,t}^2\right) =$
847 $\sum_k^{120} n_k\left[\sigma_{g,t}^2\left[2\overline{p_{mix,k}}\left(1 - \overline{p_{mix,k}}\right)\right]^{\alpha_{mix}} - \widehat{\beta_{t,k}^2}\right]^2$, using the differential evolution optimization method
848 with default parameters as implemented by the SciPy package to search over $\alpha_{mix}$ ∈ [-1.5, 1.5]
849 and $w$ ∈ [-0.5,1.5], while solving for $\sigma_{g,t}^2$ analytically (see below).

When estimating alpha-mix parameters jointly across *T* traits, we optimized the following loss function:

$$L\left(w, \alpha_{mix}, \sigma_{g,1}^2, \dots, \sigma_{g,T}^2\right) = \sum_t^T L_t\left(w, \alpha_{mix}, \sigma_{g,t}^2\right) = \sum_t^T \sum_k^{120} n_k \left[\sigma_{g,t}^2 \left[2\overline{p_{mix,k}}\left(1 - \overline{p_{mix,k}}\right)\right]^{\alpha_{mix}} - \beta_{t,k}^2\right]^2 \quad (5)$$

To avoid optimizing this function numerically with respect to *T* + 2 unknown parameters, note that for a fixed choice of *w* and $\alpha_{mix}$,

$$\mathrm{argmin}_{\sigma_{g,t}^2} L\left(w, \alpha_{mix}, \sigma_{g,1}^2, \dots, \sigma_{g,T}^2\right) = \frac{\sum_{k=1}^{120} n_k \widehat{\beta_{t,k}^2} \left[\overline{p_{mix,k}}\left(1 - \overline{p_{mix,k}}\right)\right]^{\alpha_{mix}}}{2\sum_{k=1}^{120} n_k \left[\overline{p_{mix,k}}\left(1 - \overline{p_{mix,k}}\right)\right]^{2\alpha_{mix}}} \quad (6)$$

so we can treat this as an optimization problem with two unknown parameters, *w* and $\alpha_{mix}$, and use equation 6 to identify $\sigma_{g,1}^2, \dots, \sigma_{g,T}^2$ analytically, in turn, for each trait. We minimized *w* and $\alpha_{mix}$ using the SciPy package implementation of the differential evolution method with default parameters, restricting $\alpha_{mix}$ to [-1.5, 1.5], and *w* to [-0.5, 1.5].

To estimate uncertainty of parameter estimates and perform significance testing, we wrapped the entire procedure in a genomic block-jackknife with 200 blocks (including both S-LDSC parameter estimation and GMM estimation). When estimating $\alpha_{mix}$ and *w* jointly across traits, we used a modified version of S-LDSC which forces jackknife block boundaries to be identical across summary statistics that include distinct sets of SNPs.

*Simulations*

We simulated quantitative traits using imputed genotypes for 337,448 unrelated British individuals from UK Biobank (version 3 of the Haplotype Reference Consortium imputed data) and 5,907,305 autosomal SNPs with $p_E \geq 0.05$ in 1000 Genomes and INFO score ≥ 0.60. Quantitative traits were simulated under a standard additive model: $Y = X\beta + \epsilon$, where $Y$ is a phenotype vector, $X$ is a genotype matrix, $\beta$ is a sparse vector with $n_{causal}$ non-zero entries representing per-allele causal effects, and $\epsilon$ is normally distributed residual variance. The non-zero elements of $\beta$ were sampled separately for each trait. The effects of non-zero elements of $\beta$ were sampled from a

normal distribution with mean 0 and variance $\sigma_g^2 * [2 * \max(T, p_{mix}) * (1 - \max(T, p_{mix}))]^{\alpha_{mix}}$, where $T$ represents a threshold below which varying MAF does not change effect size (see left side of Figure 2 from ref.[19]). When simulating effects, we used 1000 Genomes for European MAF, and All of Us and 1000 Genomes allele frequencies for African MAF (see the *Minor allele frequency estimation* subsection of Methods), except when stated otherwise. $\beta$ was scaled such that $var(X\beta) = var(Y)/2$ to set $h^2$=0.5. GWAS and computation of $X\beta$ were performed using Plink v2.0[81].

## Code Availability

We have publicly released custom code used to estimate $\alpha_{mix}$ and perform all analyses described in this manuscript at https://github.com/jordanero/alphamix.


## Acknowledgements

We are grateful to Xihong Lin, Po-Ru Loh, Xilin Jiang, Mashaal Sohail, and Bogdan Pasaniuc for helpful discussions. We thank All of Us participants for making this research possible. We also thank All of Us for providing access to the data used in this research. This research was conducted using the UK Biobank resource under application no. 16549. This research was funded by National Institutes of Health (NIH) grants R01 MH101244, R37 MH107649, R01 HG006399, U01 HG012009, F31 HG013040, and T32 GM135117. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

894 1. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat.*
895 *Genet.* **43**, 519–525 (2011).
896 2. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for
897 human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
898 3. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association
899 summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
900 4. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast
901 variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
902 5. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of
903 negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
904 6. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.
905 *Am. J. Hum. Genet.* **97**, 576–592 (2015).
906 7. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for
907 stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
908 8. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and
909 continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
910 9. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from
911 individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).
912 10. Kachuri, L. *et al.* Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals
913 ancestry-specific patterns of genetic architecture. *Nat. Genet.* **55**, 952–963 (2023).
914 11. Zheng, Z. *et al.* Leveraging functional genomic annotations and genome coverage to improve polygenic
915 prediction of complex traits within and between ancestries. *Nat. Genet.* **56**, 767–777 (2024).
916 12. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by
917 statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

13. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
14. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
15. Cui, R. *et al.* Improving fine-mapping by modeling infinitesimal effects. *Nat. Genet.* https://doi.org/10.1038/s41588-023-01597-3 (2023) doi:10.1038/s41588-023-01597-3.
16. Wu, Y. *et al.* Genome-wide fine-mapping improves identification of causal variants. *medRxiv* https://doi.org/10.1101/2024.07.18.24310667 (2025) doi:https://doi.org/10.1101/2024.07.18.24310667.
17. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
18. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
19. Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).
20. Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat. Commun.* **12**, 1164 (2021).
21. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved Heritability Estimation from Genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
22. Lee, S. H. *et al.* Estimation of SNP Heritability from Dense Genotype Data. *Am. J. Hum. Genet.* **93**, 1151–1155 (2013).
23. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).

24. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).
25. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
26. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).
27. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
28. Ashraf, B. & Lawson, D. J. Genetic drift from the out-of-Africa bottleneck leads to biased estimation of genetic architecture and selection. *Eur. J. Hum. Genet.* **29**, 1549–1556 (2021).
29. Zabad, S., Ragsdale, A. P., Sun, R., Li, Y. & Gravel, S. Assumptions about frequency-dependent architectures of complex traits bias measures of functional enrichment. *Genet. Epidemiol.* **45**, 621–632 (2021).
30. Zhou, D., Zhou, Y., Xu, Y., Meng, R. & Gamazon, E. R. A phenome-wide scan reveals convergence of common and rare variant associations. *Genome Med.* **15**, (2023).
31. Momin, Md. M. *et al.* A method for an unbiased estimate of cross-ancestry genetic correlation using individual-level data. *Nat. Commun.* **14**, (2023).
32. Hou, K. *et al.* Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).
33. Privé, F., Albiñana, C., Arbel, J., Pasaniuc, B. & Vilhjálmsson, B. J. Inferring disease architecture and predictive ability with LDpred2-auto. *Am. J. Hum. Genet.* **110**, 2042–2055 (2023).
34. Taylor, C. S. & Lawson, D. J. Heritability of complex traits in sub-populations experiencing bottlenecks and growth. *J. Hum. Genet.* **69**, 329–335 (2024).

965 35. Speed, D. & Evans, D. M. Estimating disease heritability from complex pedigrees allowing for
966 ascertainment and covariates. *Am. J. Hum. Genet.* **111**, 680–690 (2024).
967 36. Henches, L. *et al.* Polygenic risk score prediction accuracy convergence. *Hum. Genet. Genomics Adv.* **6**,
968 100457 (2025).
969 37. Hof, J. P. & Speed, D. LDAK-KVIK performs fast and powerful mixed-model association analysis of
970 quantitative and binary phenotypes. *Nat. Genet.* https://doi.org/10.1101/2024.07.25.24311005 (2025)
971 doi:10.1101/2024.07.25.24311005.
972 38. Patel, R. A. *et al.* Characterizing selection on complex traits through conditional frequency spectra.
973 *GENETICS* **229**, (2025).
974 39. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-Wide Inference of Ancestral
975 Recombination Graphs. *PLoS Genet.* **10**, e1004342 (2014).
976 40. Nait Saada, J., Tsangalidou, Z., Stricker, M. & Palamara, P. F. Inference of Coalescence Times and Variant
977 Ages Using Convolutional Neural Networks. *Mol. Biol. Evol.* **40**, msad211 (2023).
978 41. The All of Us Research Program Genomics Investigators *et al.* Genomic data in the All of Us Research
979 Program. *Nature* **627**, 340–346 (2024).
980 42. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach
981 for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
982 43. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature*
983 **526**, 68–74 (2015).
984 44. Tishkoff, S. A. *et al.* The Genetic Structure and History of Africans and African Americans. *Science* **324**,
985 1035–1044 (2009).
986 45. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**,
987 327–332 (2015).

46. Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
47. Price, A. L. *et al.* Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genet.* **5**, e1000519 (2009).
48. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
49. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
50. Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64–69 (2012).
51. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
52. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
53. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
54. Kanai, M. *et al. Insights from Complex Trait Fine-Mapping across Diverse Populations*. http://medrxiv.org/lookup/doi/10.1101/2021.09.03.21262975 (2021) doi:10.1101/2021.09.03.21262975.
55. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
56. Harris, K. & Nielsen, R. The Genetic Cost of Neanderthal Introgression. *Genetics* **203**, 881–891 (2016).
57. Juric, I., Aeschbacher, S. & Coop, G. The Strength of Selection against Neanderthal Introgression. *PLOS Genet.* **12**, e1006340 (2016).

1012 58. Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-
1013 genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
1014 59. Koch, E. *et al.* Genetic association data are broadly consistent with stabilizing selection shaping human
1015 common diseases and traits. Preprint at https://doi.org/10.1101/2024.06.19.599789 (2024).
1016 60. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-
1017 population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
1018 61. Fatumo, S. *et al.* Promoting the genomic revolution in Africa through the Nigerian 100K Genome
1019 Project. *Nat. Genet.* **54**, 531–536 (2022).
1020 62. Feng, Y.-C. A. *et al.* Taiwan Biobank: A rich biomedical research database of the Taiwanese population.
1021 *Cell Genomics* **2**, 100197 (2022).
1022 63. Ziyatdinov, A. *et al.* Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature*
1023 **622**, 784–793 (2023).
1024 64. Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. *Cell*
1025 *Genomics* **3**, 100361 (2023).
1026 65. Fatumo, S. *et al.* KidneyGenAfrica: A multi-cohort Genome-wide association study and polygenic
1027 prediction of kidney function in 110,000 continental and diasporan Africans. Preprint at
1028 https://doi.org/10.21203/rs.3.rs-5555692/v1 (2024).
1029 66. Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran
1030 Program. *Science* **385**, eadj1182 (2024).
1031 67. Lee, H. A. *et al.* Data Resource Profile: The Statistics of the Korea National Health and Nutrition
1032 Examination Survey (KNHANES) Biobank Project. *J. Korean Med. Sci.* **40**, e189 (2025).
1033 68. Fan, S. *et al.* Whole-genome sequencing reveals a complex African population demographic history and
1034 signatures of local adaptation. *Cell* **186**, 923-939.e14 (2023).

1035 69. Gouveia, M. H. *et al.* Unappreciated subcontinental admixture in Europeans and European Americans
1036 and implications for genetic epidemiology studies. *Nat. Commun.* **14**, 6802 (2023).
1037 70. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse
1038 human populations. *Nature* **467**, 52–58 (2010).
1039 71. Haller, B. C. & Messer, P. W. SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Mol. Biol. Evol.* **34**,
1040 230–240 (2017).
1041 72. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models.
1042 *PLoS Genet.* **9**, e1003264 (2013).
1043 73. Sharma, S. *et al.* Genetic ancestry and population structure in the All of Us Research Program cohort.
1044 Preprint at https://doi.org/10.1101/2024.12.21.629909 (2024).
1045 74. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage
1046 Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
1047 75. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants
1048 across genome assemblies. *Bioinformatics* **40**, (2024).
1049 76. Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E. & Leutenegger, A.-L. High level of inbreeding in final
1050 phase of 1000 Genomes Project. *Sci. Rep.* **5**, 17453 (2015).
1051 77. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in *Proceedings
1052 of the Python in Science Conference* 92–96 (SciPy, Austin, Texas, 2010). doi:10.25080/majora-92bf1922-
1053 011.
1054 78. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*
1055 **17**, 261–272 (2020).
1056 79. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**,
1057 1236–1241 (2015).

1058 80. Sullivan, P. F. *et al.* Leveraging base-pair mammalian constraint to understand genetic variation and
1059 human disease. *Science* **380**, eabn2937 (2023).
1060 81. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets.
1061 *GigaScience* **4**, s13742-015-0047–8 (2015).
1062