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 African MAF (defined using African-ancestry segments in African Americans), 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 α 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 α_{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 α_{mix} that best 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 α_{mix}=−0.34 (95% CI: (-0.67, -0.02)), attaining a far better fit than the standard α 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.

42 ***Introduction***

43 Understanding genetic architectures is fundamental to partitioning heritability[1–5], polygenic
44 risk prediction[6–11], and statistical fine-mapping[12–16]. Recent studies have shown that genetic
45 variants with lower minor allele frequency (MAF) tend to have larger per-allele effects on diseases
46 and complex traits, due to negative selection against disease-associated variants that impact
47 fitness[17–20]. This relationship is frequently modeled using the α model[4,9,15,17–38], in which the per-
48 allele effect size variance of a SNP in a GWAS population (e.g. Europeans) is a function of its MAF
49 in the GWAS population. However, limited attention has been paid to the possibility that the MAF in
50 a different population (e.g. Africans) might be informative for GWAS architectures (see ref.[28,34,39]
51 and Discussion).

52 We hypothesized that (unadmixed) African MAF might better predict per-allele effect sizes in
53 European populations, as it is not distorted by the out-of-Africa bottleneck and thus more
54 accurately reflects the consequences of negative selection; we note that common variants
55 explaining most disease heritability are typically older than the split between African and non-
56 African populations[40,41]. First, we demonstrate that the proportion of SNPs that are non-
57 synonymous is better predicted by African MAF than European MAF, relying on non-synonymous
58 variant type as a partial proxy for variant effect size. Second, we analyze European GWAS
59 summary statistics using a model that incorporates both African MAF and European MAF to
60 demonstrate that estimated per-allele effect variance is more heavily influenced by African MAF
61 than European MAF.

Results

Overview of methods: defining the α_{mix} model

We propose the α_{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 α model[4,9,17–38].

In detail, the α_{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 σ_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 α_{mix} specifies the level of MAF dependence.

In the α_{mix} model, we define African MAF (p_A) using whole-genome sequencing data from All of Us[42] (N=46,672 African Americans, correcting for European admixture by extracting local genomic regions with two African haplotypes identified using RFMix2[43] (Methods)) and 1000 Genomes[44] (N=107 Yoruba), and define European MAF using whole-genome sequencing data from 1000 Genomes[44] (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[45–47], 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[48], which is widely studied and has not been subject to recent extreme bottlenecks[44].

Overview of methods: fitting the α_{mix} model

We fit the α_{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 α_{mix} model parameters that best fit the S-LDSC estimates; this procedure is similar to the approach used in ref.[18] to fit the α model.

In detail, we estimate α_{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 α_{mix}, w, and σ_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 α_{mix} jointly while estimating σ_g^2 separately for each trait; we also include analyses in which we estimate σ_g^2, w, and α_{mix} separately for each trait. We have publicly released custom code used to estimate α_{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[18,49–53], and are highly enriched for heritability[11,18,53] and fine-mapped causal effects[14,54,55]. 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 μ and γ 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 (γ=−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 (γ=−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[52], including
139 variants that arose via Neanderthal introgression[56–58]. 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 μ and γ 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 (γ=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 γ (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

160 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 $p_E \geq 0.002$,
161 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
162 provided a better fit for all five SNP sets ($P < 10^{-15}$ for all tests, except $P = 0.01$ for SNPs with MAF
163 ≥ 0.05 in both ancestries) (Supplementary Figure 6). Third, we compared the marginal effect of p_A
164 on PNS for all SNPs with $p_A \geq 0.002$ to the marginal effect of p_E on PNS for all SNPs with $p_E \geq$
165 0.002 (Supplementary Figure 7). Surprisingly, the p_E model attained a higher McFaddens pseudo-
166 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
167 p_A, γ=−0.17 for p_E vs. −0.13 for p_A, $P < 10^{-15}$ for pseudo-r^2 and γ differences); however, we note
168 that this comparison varies both the set of SNPs evaluated and the explanatory variable, and thus
169 does not directly indicate whether p_A or p_E better predicts PNS for any fixed set of SNPs. We
170 hypothesize that the observed difference in r^2 is because PNS is harder to predict for African-
171 specific SNPs than for European-specific SNPs. Fourth, we refit the logistic regression on log(p_{mix})
172 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
173 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
174 between 0.76-0.95, with tight confidence intervals (Supplementary Figure 8). Fifth, we refit the
175 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)
176 (Supplementary Figure 9), rather than 0.002 as in the main analyses. The MLE for w was 0.80,
177 similar to the estimate obtained when restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$.
178
179 *Per-allele disease and complex trait effect sizes are predominantly African MAF-dependent in*
180 *European populations*
181 We hypothesized that per-allele disease and complex trait effect sizes (e.g. in European
182 populations, for which data is available in large sample size) might be better predicted by African
183 MAF than by European MAF, because African MAF is not distorted by the out-of-Africa bottleneck.
184 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 (dashed blue line in Figure 3b), possibly reflecting
197 recent selection on variants that arose in Europeans after the out-of-Africa bottleneck[52], including
198 variants that arose via Neanderthal introgression[56–58]. These findings support the hypothesis that
199 African MAF is highly predictive of per-allele effect size variance, even after accounting for
200 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 α_{mix} model (Equation 1 and Methods) across the 50
203 diseases/traits, using a single *w* parameter and a single α_{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 α_{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 (α 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 ($\alpha_{mix} = 0$), α_{mix}

210 attained a 4.53x greater decrease in weighted MSE than a European MAF-only model (*w*=0) and a
211 1.03x greater decrease in weighted MSE than an African MAF-only model (*w*=1) (Figure 4b). The
212 relationship between *w* and the MSE of the α_{mix} model was remarkably similar to the relationship
213 between *w* and the negative log-likelihood of our logistic regression of PNS (Figure 2b). We also
214 estimated α_{mix} values for each value of *w* (Figure 4c). Surprisingly, values of *w* that were lower
215 than the best-fit value of *w*=0.96 produced larger absolute α_{mix} estimates. This may be because at
216 high values of *w*, p_{mix} approaches 0 for some bins of SNPs, which results in overestimates of per-
217 allele effect size variance under the α_{mix} model, which cause best-fit estimates of α_{mix} to be
218 attenuated towards 0. This is analogous to the known phenomenon that the α model overestimates
219 per-allele effect size variance at low values of *p* according to evolutionary forward simulations, in
220 which per-allele effect size variance plateaus at low values of *p* (Figure 2 of ref.[19]). To investigate
221 further, we fit the α_{mix} model while restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$. Accordingly,
222 α_{mix} estimates were less dependent on *w* when restricting to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$
223 (best-fit *w*=0.94 and α_{mix}=−0.22; Supplementπary Figure 10). Estimates of α_{mix} were closer to 0,
224 providing further evidence that SNPs with $p_A < 0.05$ are the primary drivers of the European MAF-
225 dependence of per-allele effect size variance.

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

heterogeneity in w or α_{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 α_{mix}, respectively).

We performed 5 secondary analyses. First, we estimated α_{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 α_{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 α_{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 α_{mix}=−0.37 (95% CI: (−0.14, −0.59)) were very similar to those of the main analyses. Third, we fit the α_{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 α_{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 α_{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 α_{mix}=−0.38 (95% CI: (0, −0.76)) were very similar to those of the main analyses. Fifth, we fit the α_{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

260 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)),
261 respectively).
262
263 *Simulations of α_{mix} estimates*
264 To test the validity of our α_{mix} estimation procedure, we simulated quantitative GWAS traits
265 according to the α_{mix} model using real genotypes from 337,448 unrelated British individuals from
266 UK Biobank and 5,907,305 SNPs with European MAF ≥ 0.05. We varied the true value of w
267 between 0 and 1, set α_{mix} to −0.38 (consistent with previous estimates in European populations[17–
268 19]), set σ_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
269 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
270 the generative model is different from the inference model), and set the number of causal SNPs to
271 10,000; other simulations settings were also evaluated. We fit the α_{mix} model (Equation 1 and
272 Methods) using a single w parameter and a single α_{mix} parameter across groups of 25 traits
273 (which resulted in total N*h^2=4.2*10^6, vs. 4.1*10^6 in real trait analyses).
274 Results are reported in Figure 6 and Supplementary Table 12. We determined that our w
275 estimates were conservative in the sense that they were biased away from extreme values, e.g.
276 average estimate ranging from 0.16 when true w=0.00 to 0.97 when true w=1.00 (Figure 6a);
277 estimates were approximately unbiased (0.94) at true w=0.95, which was closest to our estimate
278 for real traits. Estimates of α were biased downward (relative to true α=−0.38) at most values of w,
279 e.g. −0.48 when w=0.00 and −0.42 when w=0.95 (but −0.35 when w=1.00) (Figure 6b). Secondary
280 analyses provided below indicate that these biases are due to imperfect estimation of mean effect
281 size variance via S-LDSC and misspecification of the inference model with respect to T.
282 We performed 6 secondary analyses. First, we simulated effects based on MAF from
283 N=337,448 unrelated British individuals, but performed inference using MAF from N=489 European
284 individuals from 1000 Genomes to simulate the effect using noisy MAF estimates (Supplementary

285 Figure 19). Results were very similar to those of the main analyses. Second, we varied *T*, the MAF
286 threshold used by the generative model (Supplementary Figures 20-21), which determines how
287 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
288 main analyses, but with upward bias (rather than downward bias) in our α_{mix} estimates at *w*=0.75
289 and *w*=0.95 and slightly greater upward bias in α_{mix} at *w*=1.00. Results with $T=1.5*10^{-5}$
290 (equivalent to a minor allele count of 1) were similar to those of the main analyses, but with less
291 bias in our estimates of *w* at high values of *w*, and downward bias (rather than upward bias) in our
292 estimates of α_{mix} at *w*=1.00. We conclude that our estimates of *w* are insensitive to plateauing of
293 $\mathrm{E}[\beta^2|p_A, p_E]$ at low values of p_A. Third, we varied the number of causal SNPs per trait
294 (Supplementary Figure 22-23). With 5,000 (instead of 10,000) causal variants per trait, results
295 were similar to the main analyses, with slight downward bias at *w*=1. With 20,000 causal variants
296 per trait, results were very similar to the main analyses. Fourth, we applied our inference procedure
297 while using true values of β^2 (instead of estimates of mean effect size variance from S-LDSC;
298 Supplementary Figure 24). Estimates of *w* and α_{mix} were approximately unbiased, except for at *w*
299 ≥ 0.95, where estimates of α_{mix} were biased upwards. We conclude that in our main simulations,
300 bias in estimates of *w* and bias in estimates of α_{mix} at *w* ≤ 0.75 were caused by our approximation
301 of mean effect size variance. Fifth, we applied our inference procedure while using the true values
302 of β^2 (instead of estimates of mean effect size variance from S-LDSC), restricting the inference
303 procedure to causal variants, and thresholding p_{mix} for individual SNPs at the true value of *T*
304 (Supplementary Figure 25). We also performed inference using the same approach, but without
305 thresholding p_{mix} (Supplementary Figure 26). Estimates of α_{mix} were biased upwards at *w* ≥ 0.95
306 when p_{mix} was not thresholded, and were unbiased when p_{mix} was thresholded at *T*. We conclude
307 that the bias in our estimates of α_{mix} at *w* ≥ 0.95 in the main simulations was due to model
308 misspecification with respect to *T*. Finally, we estimated trait-specific values of *w* and α_{mix}, rather

309 fitting a joint model across all 25 traits (Supplementary Figure 27). Results were very similar to the

310 main analyses, but with greater uncertainty in mean parameter estimates.

311 ***Discussion***

312 We have shown that per-allele disease and complex trait effect sizes of SNPs that are
313 common in European populations are predominantly a function of African MAF, not European MAF.
314 This represents a significant departure from standard approaches for modeling MAF dependence,
315 where per-allele effect sizes estimated in a given population are related to MAF in the same
316 population[4,9,15–27,29–33,35–38]. We have provided two lines of evidence in favor of an African MAF-
317 dependence model. First, we showed that non-synonymous variant status is much better predicted
318 using African MAF than European MAF. This approach requires very few assumptions and is likely
319 robust to model misspecification, but is limited in that it only focuses a single, relatively narrow
320 subset of human genetic variation. Second, we used disease/trait GWAS summary statistic data
321 from European populations to show that per-allele effect size variance is also much better
322 predicted using African MAF than European MAF. This approach reached very similar conclusions
323 to our analysis of non-synonymous variation.

324 Our work is motivated in part by three recent population genetic studies exploring the
325 relationship between African MAF, European MAF, and GWAS effect sizes. Refs. [28,34] performed
326 simulations exploring the effect of bottlenecks on genetic architectures and MAF dependence,
327 using models in which per-allele effect sizes in one population are dependent on MAF in another
328 (ref. [28] focuses specifically on the out-of-Africa bottleneck, while ref. [34] does not). However, neither
329 study analyzed empirical GWAS data or demonstrated that this model fits empirical data. Ref. [39]
330 used conditional allele frequency spectra to differentiate modes of selection on GWAS-associated
331 variants (emphasizing joint analyses of African MAF and European MAF, as we do in this study),
332 but did not quantify the dependence of per-allele effect sizes on African MAF or European MAF.

333 Our work has important implications both for analyses of European GWAS data and, more
334 generally, populations of all ancestries. Modeling MAF-dependent architectures has been shown to
335 impact studies of genetic architectures and negative selection[2,4,5,17–24,26,28,29,31,33,34,37–39,49–51,53,59,60],

336 polygenic risk prediction[9,11,61], and fine-mapping[16,61], and our results suggest that specifically
337 modeling African MAF-dependent architectures will improve all of these endeavors, not only in
338 European populations (the focus of this study) but in population of all ancestries as more GWAS
339 data in diverse populations becomes available[62–67,42,68].

340 Our work has several limitations. First, our analyses focus exclusively on African and
341 European MAF. We expect that drift from the out-of-Africa bottleneck causes MAF in all non-
342 African ancestries to be less informative for per-allele effect size variance[44], although we have not
343 explicitly tested this hypothesis. Second, our analyses of diseases/traits focus exclusively on
344 European-ancestry GWAS data, which limits our disease/trait analyses to SNPs with $p_E > 0.05$;
345 however, our analysis of non-synonymous variant status spans a larger set of SNPs ($p_A > 0.002$
346 and/or $p_E > 0.002$), providing some evidence that our findings are generalizable. Analyses of
347 diseases/traits in non-European populations will be more informative as more GWAS data in
348 diverse populations becomes available[62–67,42,68]. Third, our analyses focus on continental
349 ancestries and do not consider subcontinental ancestries with distinct allele frequencies,
350 demographic histories, and selective forces[45–47,69,70]; however, allele frequency differences
351 between populations of the same continental ancestry are generally smaller than allele frequency
352 differences between populations of different continental ancestry[71]. Fourth, our estimate of the
353 degree of European MAF-dependence under the α model instead of the α_{mix} model (α=−0.50,
354 95% CI: (−0.61,−0.38), Figure 4c) was stronger than refs.[17–19] (α = −0.38 to −0.36 meta-analysis
355 across traits) (though weaker than refs.[20,33]; α = −0.65 to −0.58). When estimating α using S-LDSC
356 with the standard Baseline-LD (v2.2) model, which does not include African MAF annotations,
357 (Supplementary Figure 15), we obtain an estimate that is consistent with refs.[17–19] (α = −0.41, 95%
358 CI (−0.51, −0.32)); regardless of the S-LDSC model used, we find that per-allele effect size
359 variances are predominantly African MAF-dependent, which is our main conclusion. Fifth, our
360 approach to fitting the α_{mix} model suffers some biases in simulations (Figure 6); however,

361 estimates of *w* were conservative (in the sense that they were biased away from extreme values),
362 and approximately unbiased for values of *w* close to our estimate of *w*=0.96 for real traits, thus
363 providing strong support for our conclusion that per-allele effect sizes are predominantly African
364 MAF-dependent. Sixth, we did not perform simulations to assess whether the α_{mix} model is
365 consistent with evolutionary models at appropriate parameter settings; a previous study performed
366 forward simulations[72] to show that the α model is generally consistent with an evolutionary model
367 involving a single population[19] (except that per-allele effect size variance plateaus at low values of
368 minor allele frequency; Figure 2 of ref.[19]), and extending those forward simulations to incorporate
369 African and non-African populations (impacted by the out-of-Africa bottleneck) is an important
370 future research direction. Seventh, analogous to previous studies using the α model[4,18,19,21–
371 26,29,31,32,34,35,38], our analyses using the α_{mix} model do not account for non-infinitesimal disease/trait
372 effects[73]; we note that failure to model non-infinitesimal effects when estimating SNP-heritability
373 leads to suboptimal precision, but does not induce bias[73]. Eighth, African MAF is unlikely to be
374 informative of per-allele effects among SNPs which are younger than the out-of-Africa bottleneck
375 and which are common in European ancestry populations, although these represent a small
376 fraction of European common variation; modeling MAF-dependence as a function of allele age is a
377 potentially impactful future research direction. Despite these limitations, we have shown that per-
378 allele disease and complex trait effect size variances are predominantly African MAF-dependent in
379 European populations.

Figures

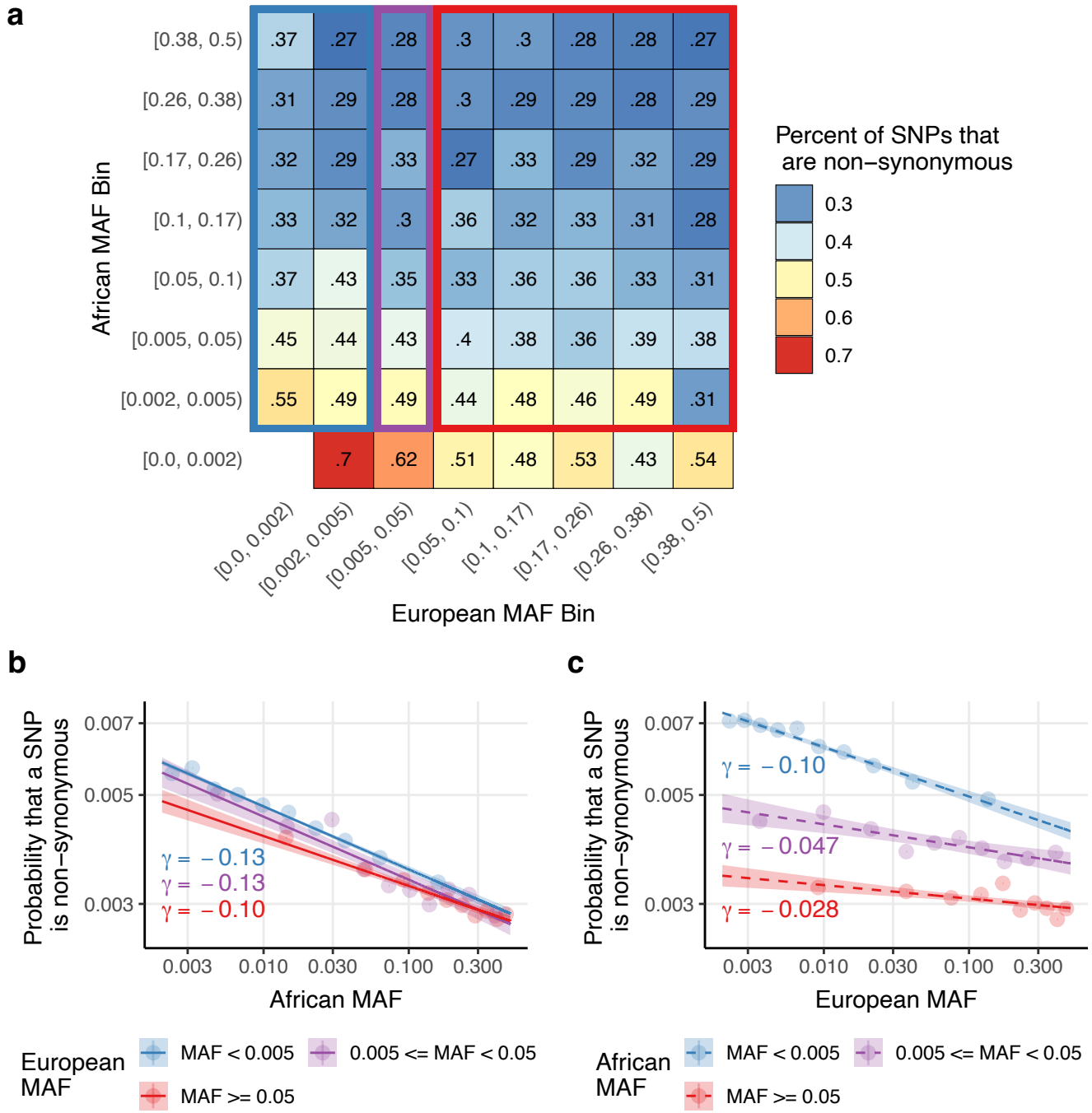


Figure 1: The proportion of non-synonymous SNPs as function of unadmixed African and European minor allele frequencies: We report **(a)** the percentage of SNPs that are non-synonymous in bivariate bins defined by unadmixed African MAF (rows) and European MAF (columns), **(b)** the probability that a 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.

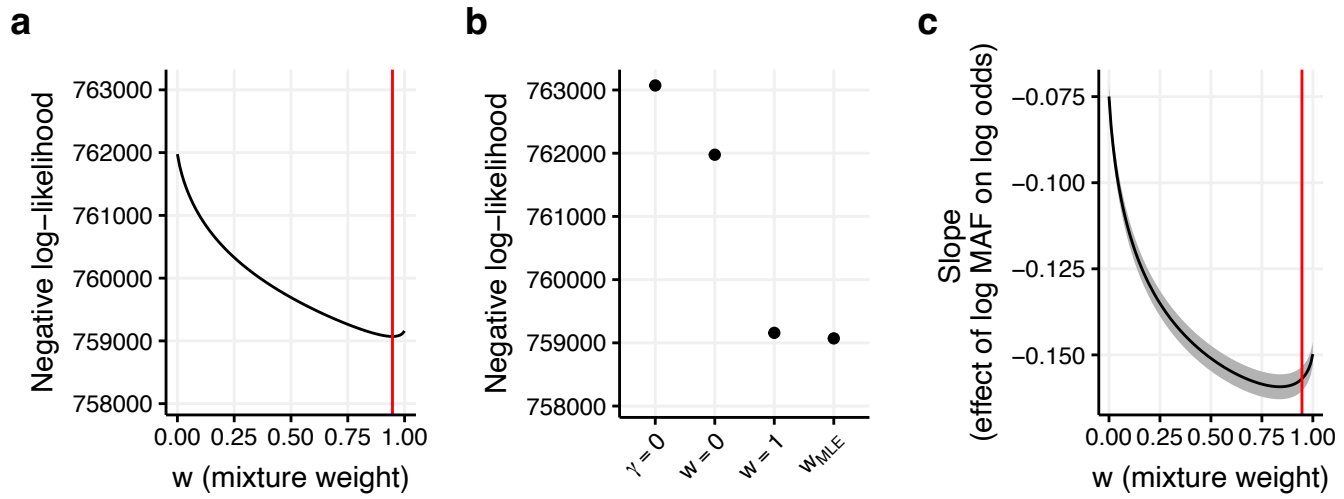


Figure 2: Best-fit parameters for a logistic regression of non-synonymous status on log(p_{mix}). We report **(a)** the negative profile log-likelihood for various models, including a model without MAF dependence (γ=0), a model with European MAF dependence (w=0), a model with African MAF dependence (w=1) 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 (γ) 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 γ. Numerical results are reported in Supplementary Table 7.

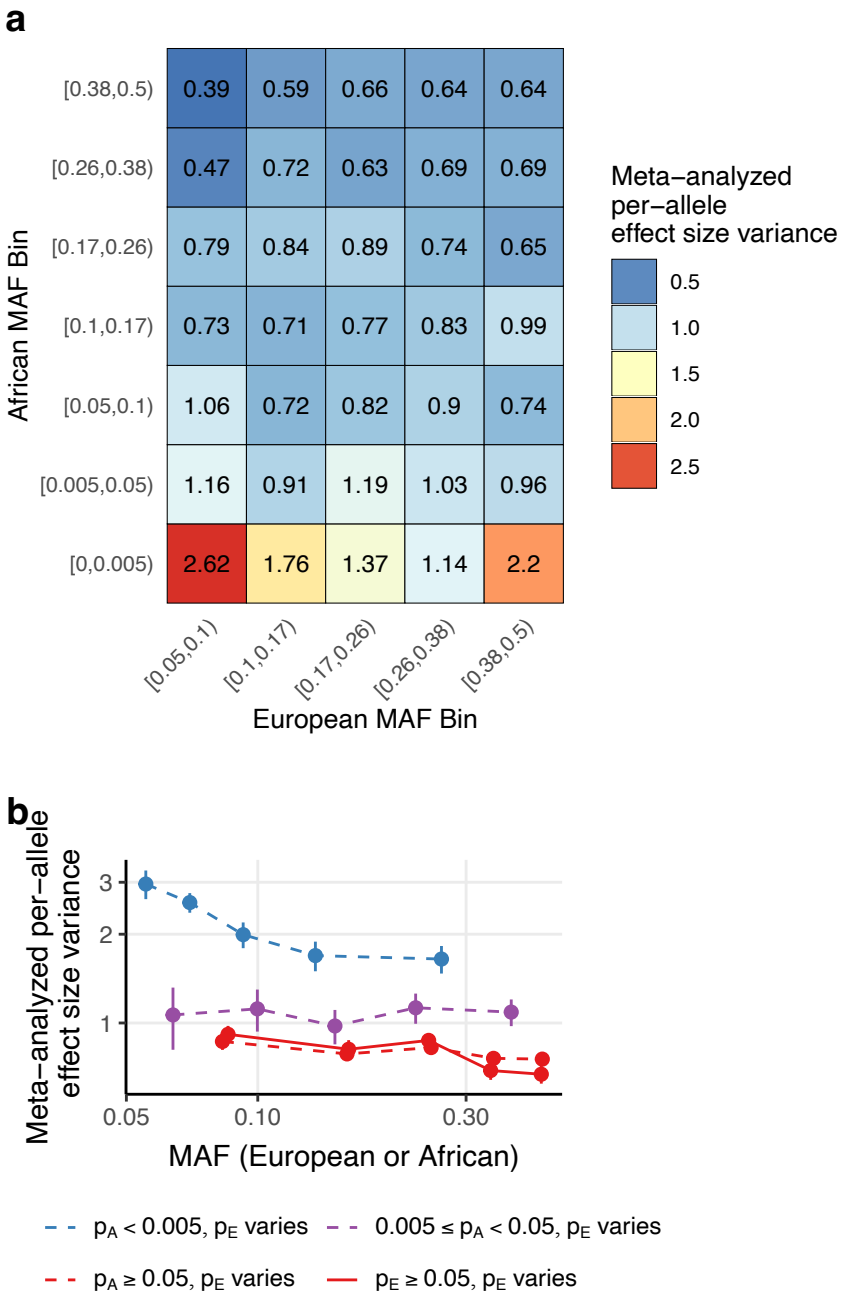


Figure 3: Per-allele effect size variance across 50 diseases/traits as a function of African and European minor allele frequencies. We report meta-analyzed scaled mean effect size variance as a function of **(a)** bivariate MAF bins defined by 7 African and 5 European MAF bins, and **(b)** African (solid 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.

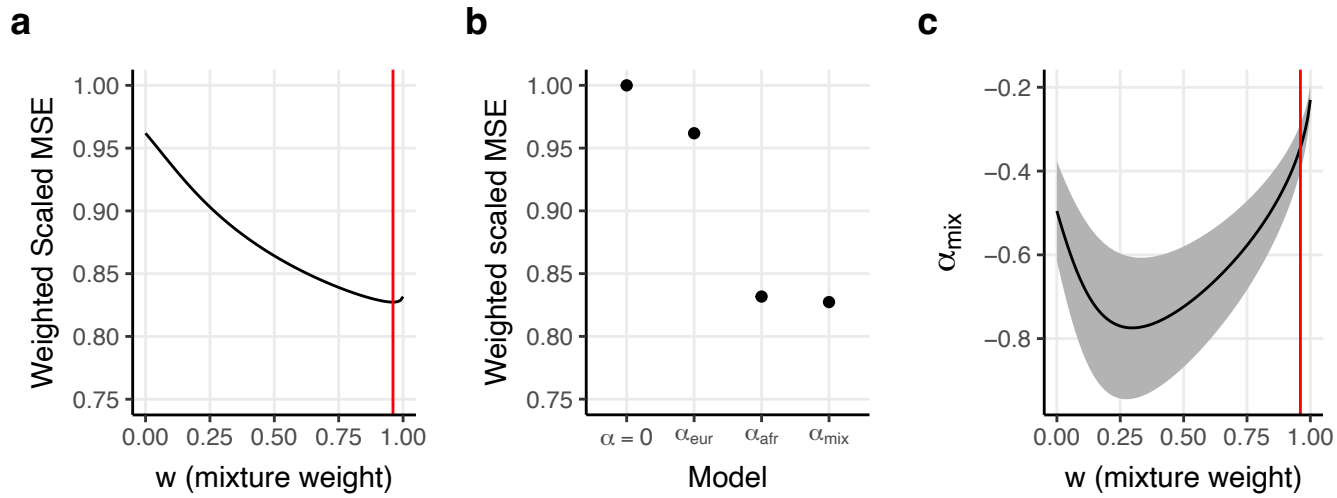


Figure 4: Results of fitting the α_{mix} model across 50 diseases/traits. We report **(a)** the weighted scaled mean squared error with respect to *w*, **(b)** the weighted scaled mean squared error for various models, including a model without MAF dependence (α=0), a model with European MAF dependence (α_{eur}), a model with African MAF dependence (α_{afr}), and the α_{mix} model, and **(c)** estimates of α_{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 α_{mix}. Numerical results are reported in Supplementary Table 10.

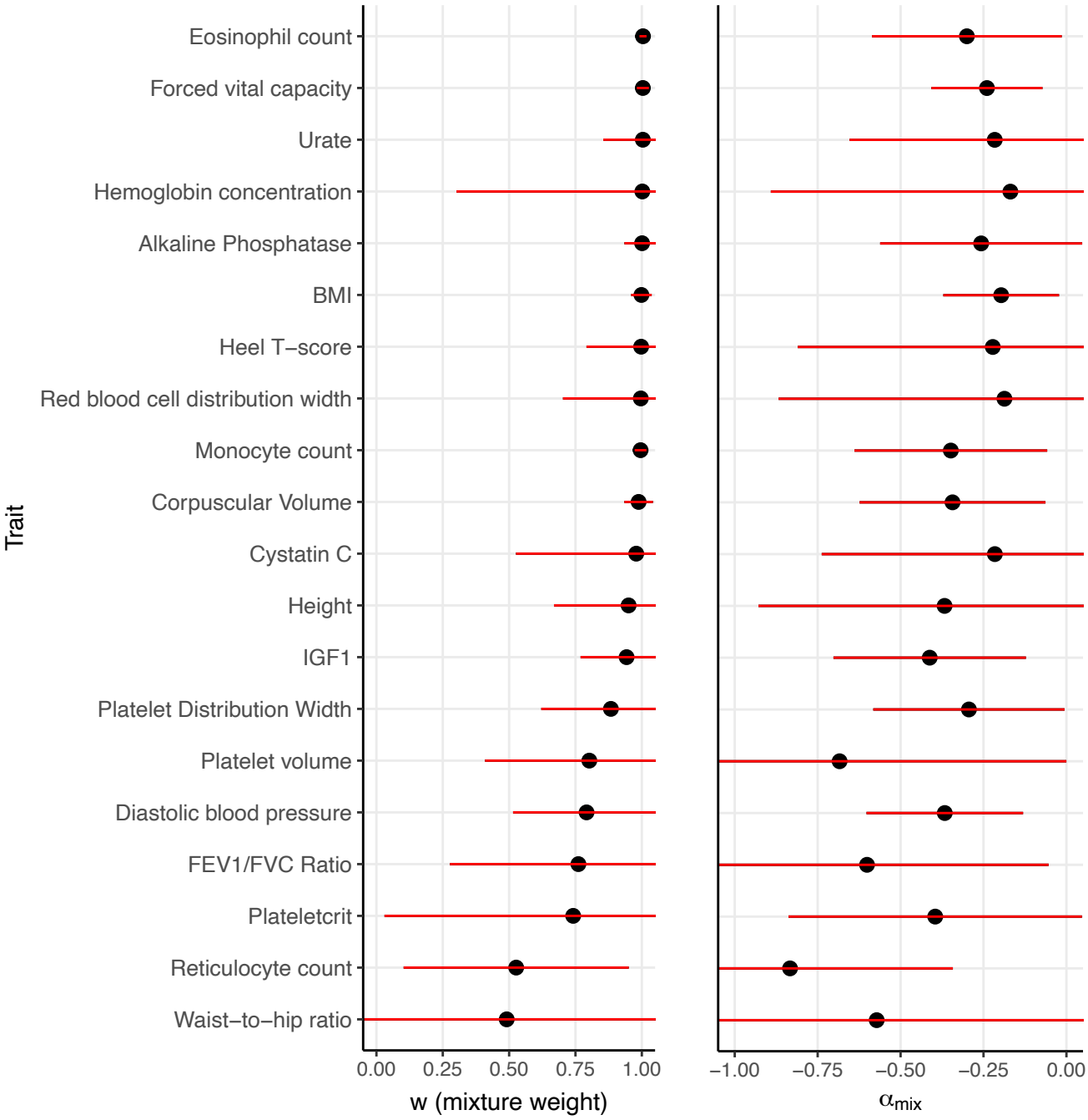


Figure 5: Results of fitting the α_{mix} model separately for 20 out of 50 diseases/traits. We report estimates of **(a)** *w* and **(b)** α_{mix}. Diseases/traits are ordered by point estimate of *w*. The 20 diseases/traits with highest $N*h^2$ point estimates are displayed. Supplementary Figure 11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

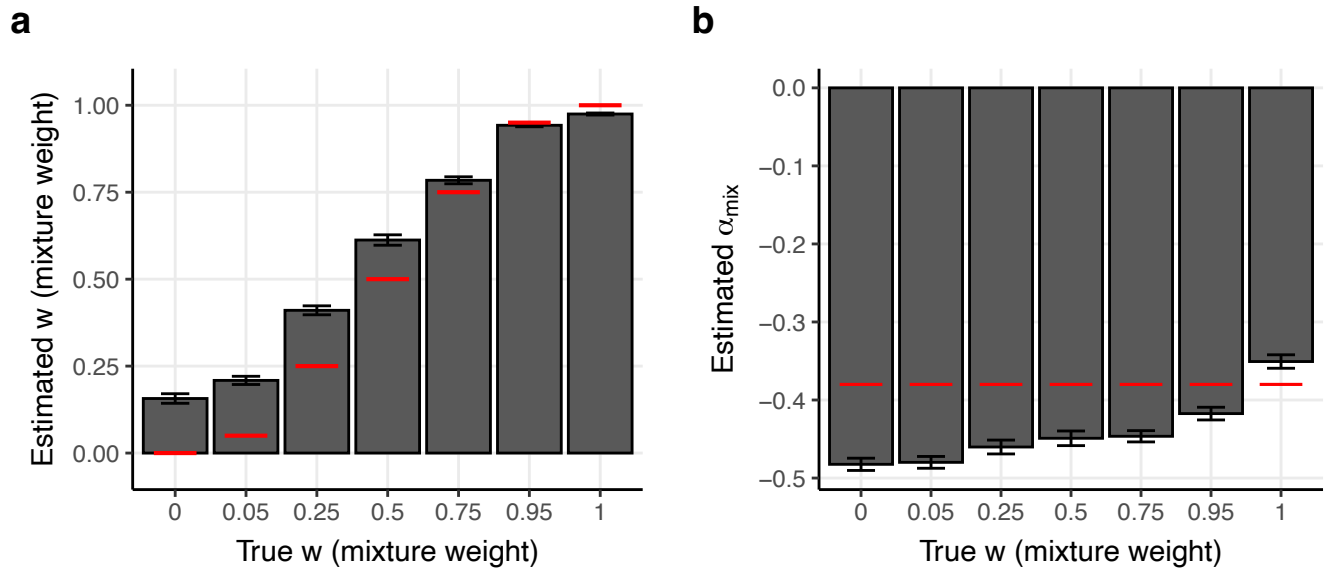


Figure 6: Results of fitting the α_{mix} model in simulations. We report **(a)** estimates of *w* and **(b)** α_{mix}, at different values of true *w*. Bars denote mean estimates across 40 simulation replicates (each analyzing 25 traits). Red lines denote true parameter values. Error bars denote 95% 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, γ estimate, γ standard error, and 95% confidence interval on γ 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)

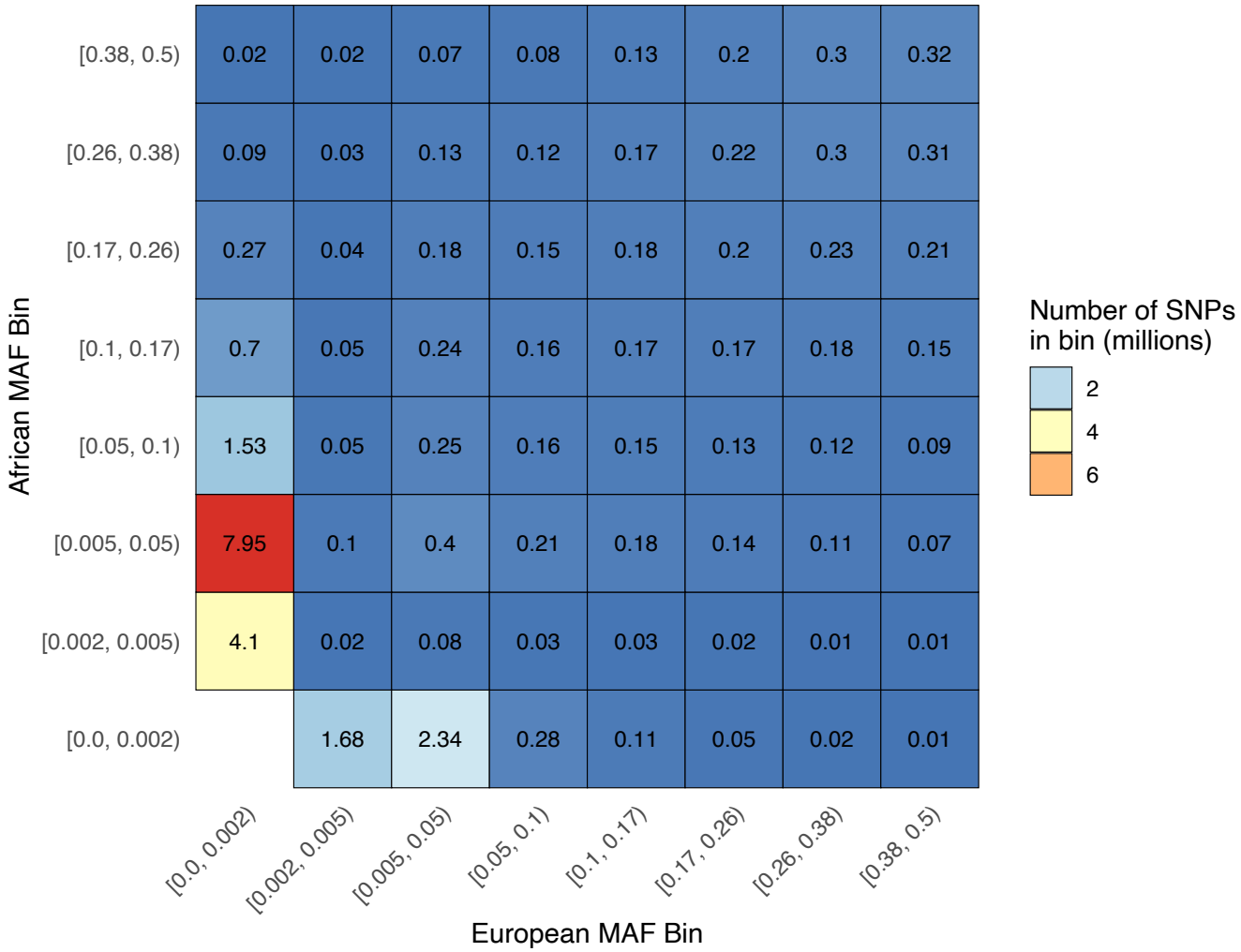
Supplementary Table 9: Per-allele effect size variance across 50 diseases/traits as a function of MAF quintiles, stratified by African MAF. We report, for each MAF quintile, the mean MAF, point estimate of the meta-analyzed scaled mean effect size variance, standard error of the meta-analyzed scaled 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 α_{mix} model across 50 diseases/traits. We report point estimates and standard errors on α_{mix} and point estimates and standard errors of the loss function across 101 values of *w* from 0 to 1.
(see .xlsx file)

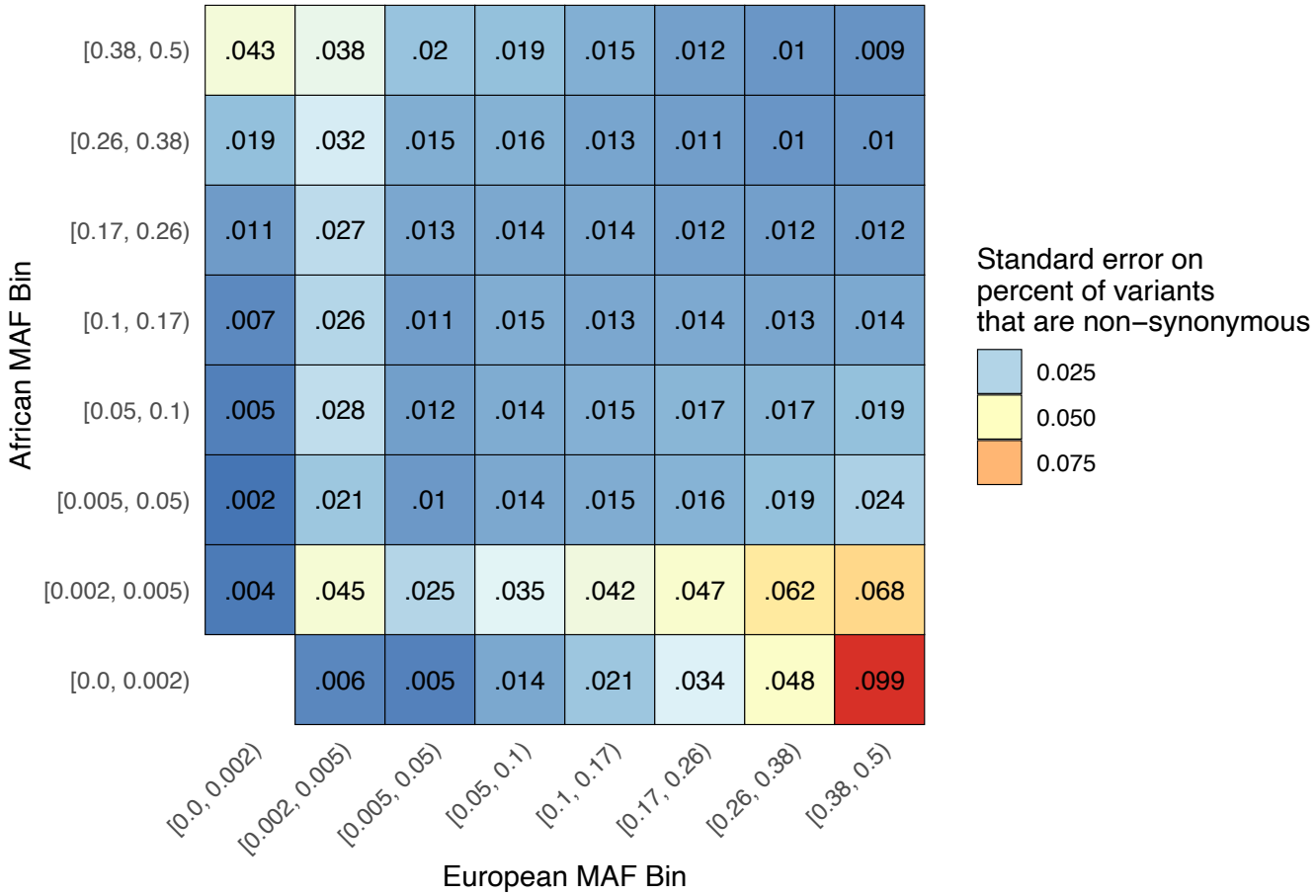
Supplementary Table 11: Results of fitting the α_{mix} model separately for 50 diseases/traits. We report point estimates and standard errors on *w*, α_{mix}, and the loss function for each of 50 diseases/traits.
(see .xlsx file)

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

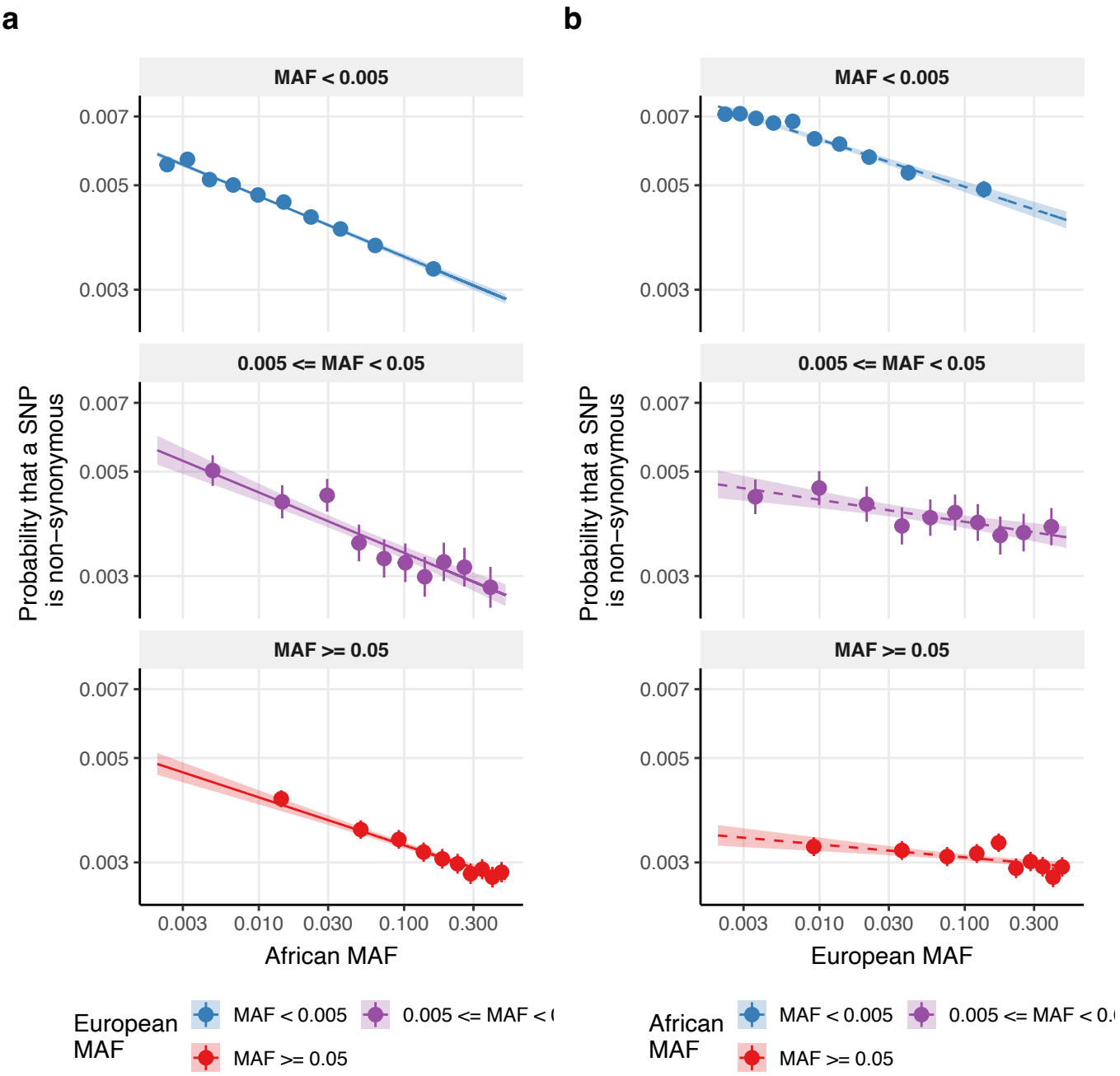
515 **Supplementary Figures:**



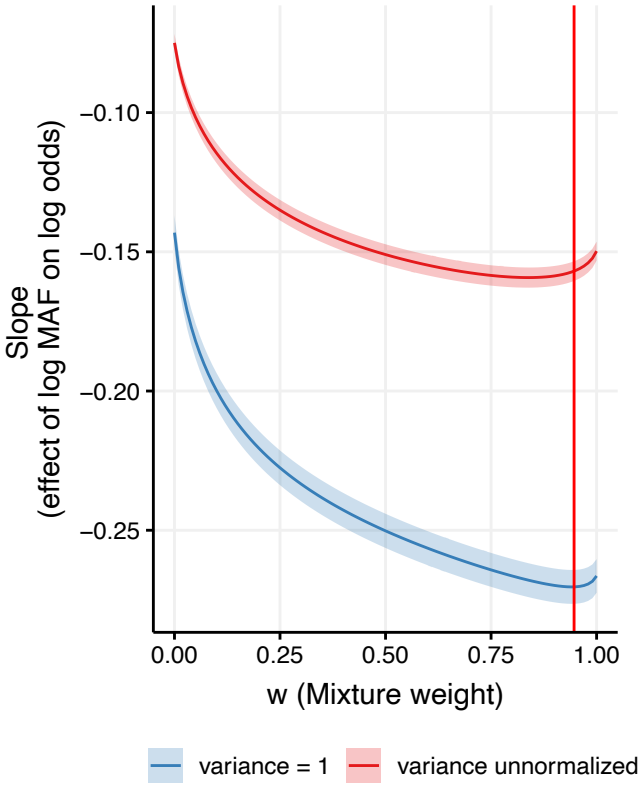
516
517 **Supplementary Figure 1: Number of SNPs in each AoU MAF stratum.** We report the number of
518 SNPs contained in each grid element given in Figure 1a. Each grid element contains SNPs with
519 unadmixed African and European MAF that fall in a pair of MAF intervals. A justification for MAF
520 interval boundaries is given in Methods.



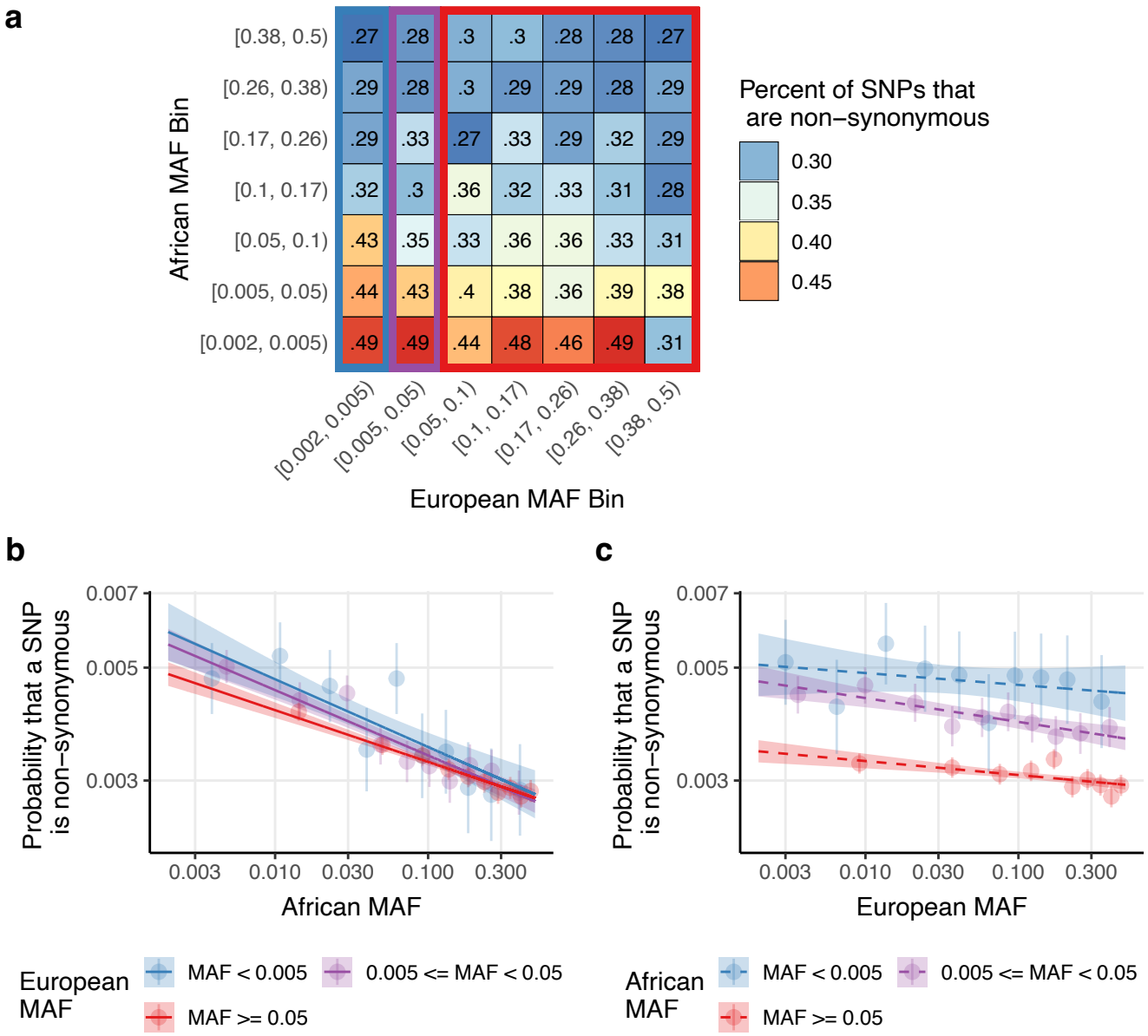
Supplementary Figure 2: Standard error on the proportion of non-synonymous SNPs in each MAF stratum. We report the binomial proportion standard error for each grid element given in Figure 1a. Each grid element contains SNPs with unadmixed African and European MAF that fall in a pair of MAF intervals.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



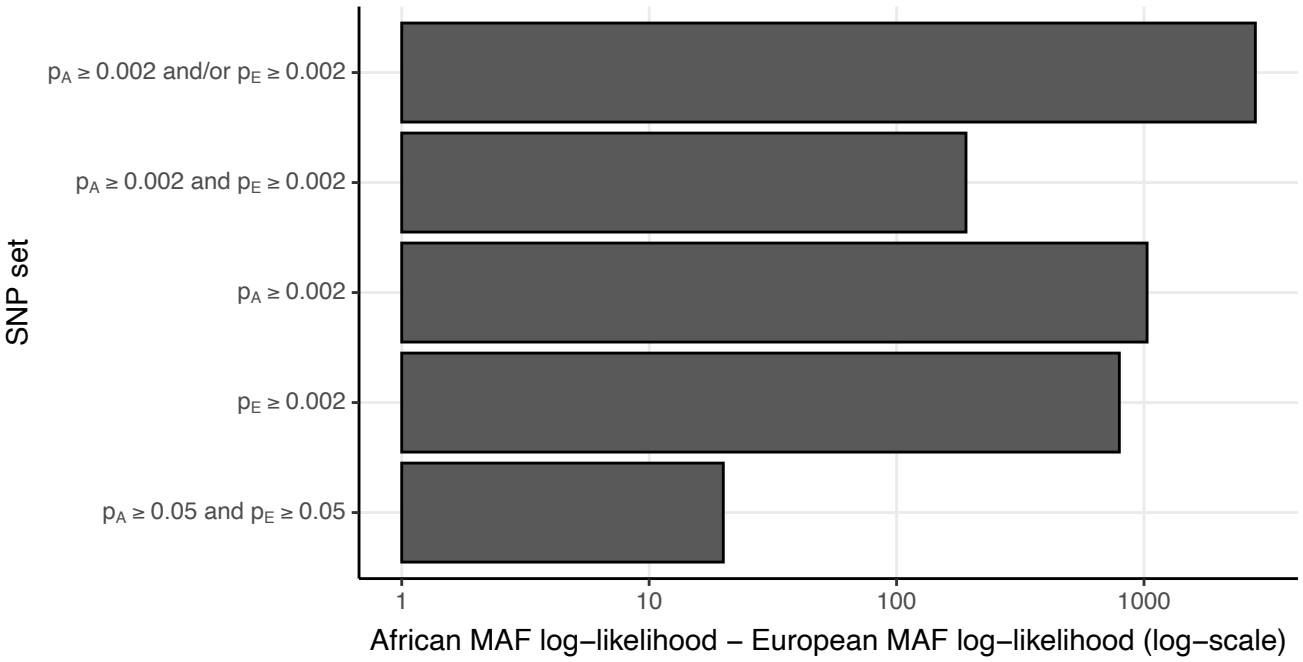
Supplementary Figure 3: Logistic regression for non-synonymous status as a function of unadmixed African and European minor allele frequencies separated by MAF category. We report **(a)** the probability that a SNP is non-synonymous as a function of African MAF after stratifying by European MAF categories (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.



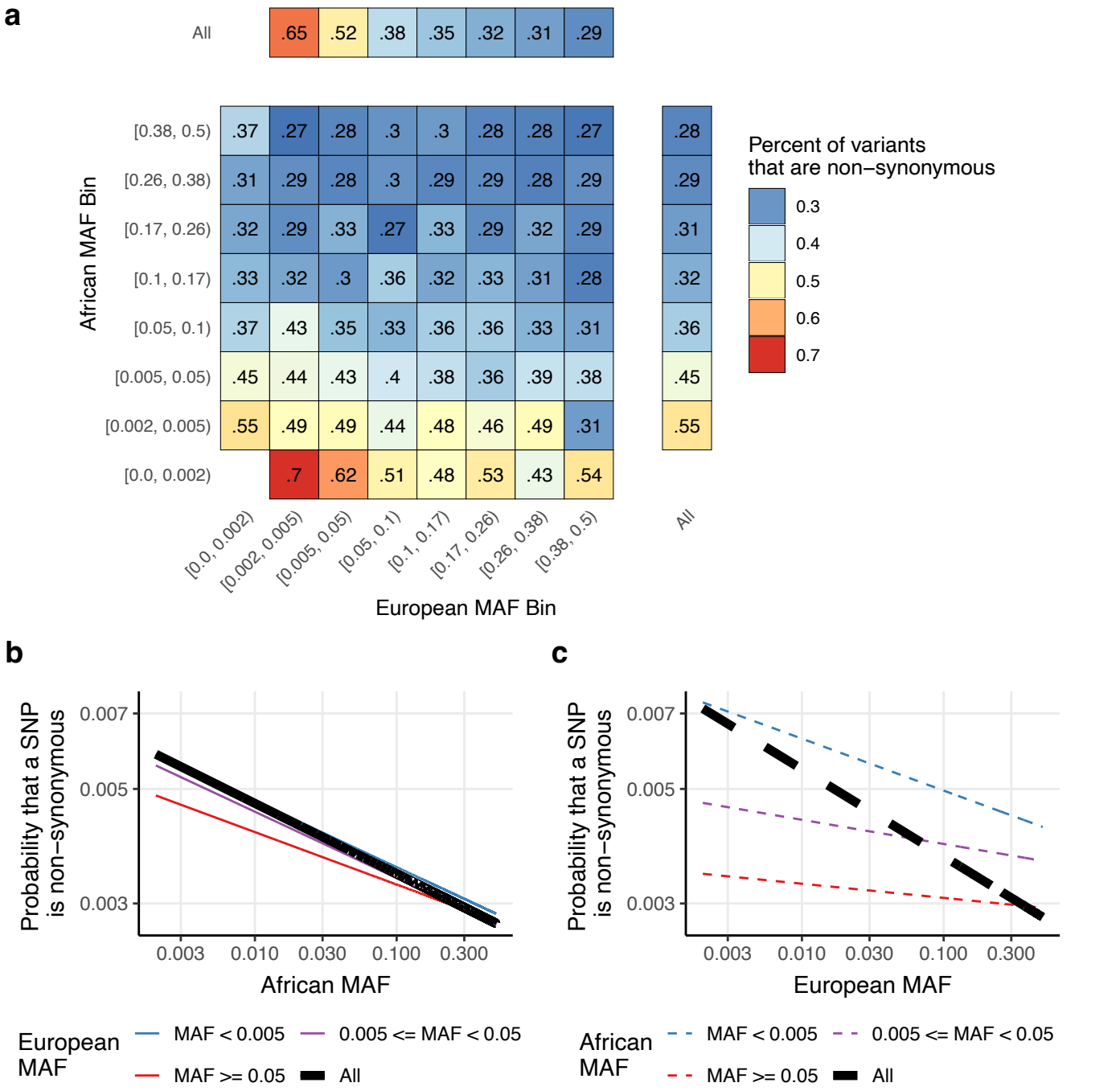
Supplementary Figure 4: Best fit parameter for a logistic regression of non-synonymous status on log(p_{mix}) while scaling log(p_{mix}) to have unit variance. We report the regression slope (γ) for values of w between 0 and 1. The blue curve denote results for a logistic regression on log(p_{mix}) while 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.



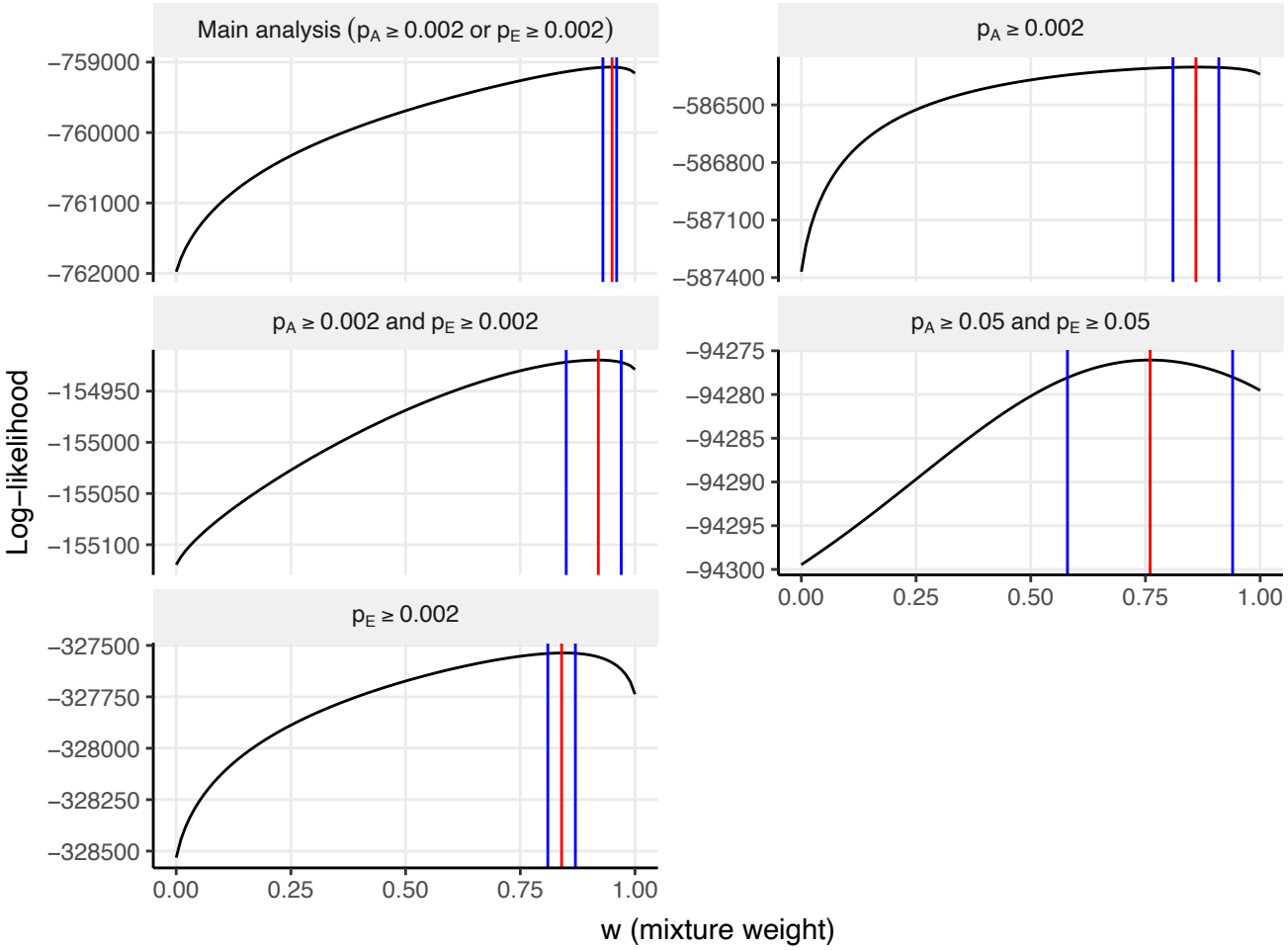
Supplementary Figure 5: The proportion of non-synonymous SNPs as function of unadmixed African and European minor allele frequencies, restricted to variants with $p_A \geq 0.002$ and $p_E \geq 0.002$: We report **(a)** the percentage of SNPs that are non-synonymous in bivariate bins defined by unadmixed 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.



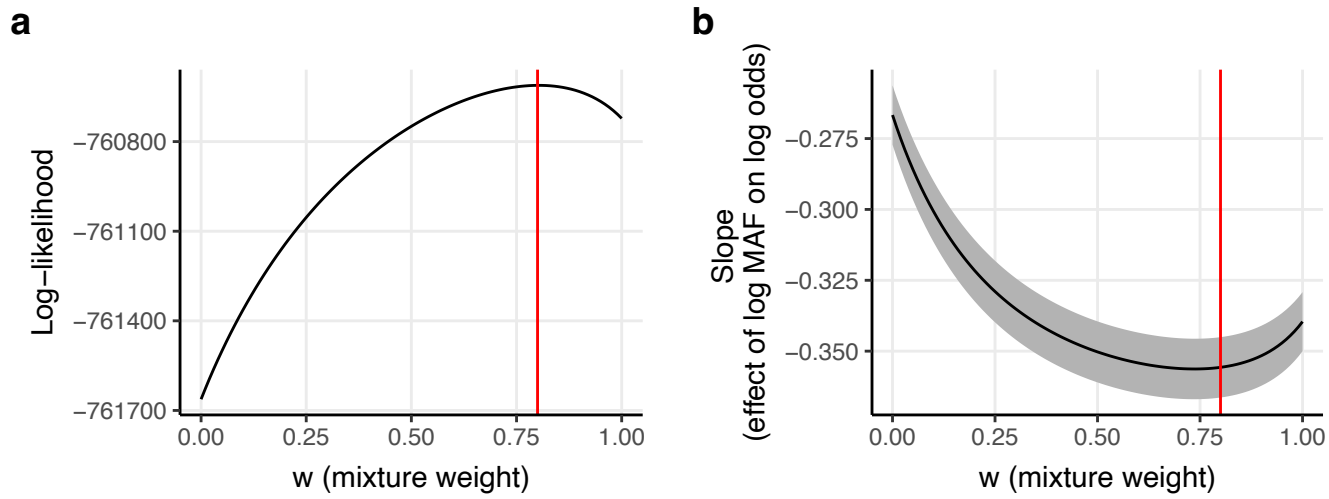
Supplementary Figure 6: Difference in log-likelihood for a logistic regression of non-synonymous status on African MAF vs. a logistic regression of non-synonymous status on European MAF. We report the log-likelihood for a logistic regression of PNS on log(p_A) minus the log-likelihood for a logistic 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).



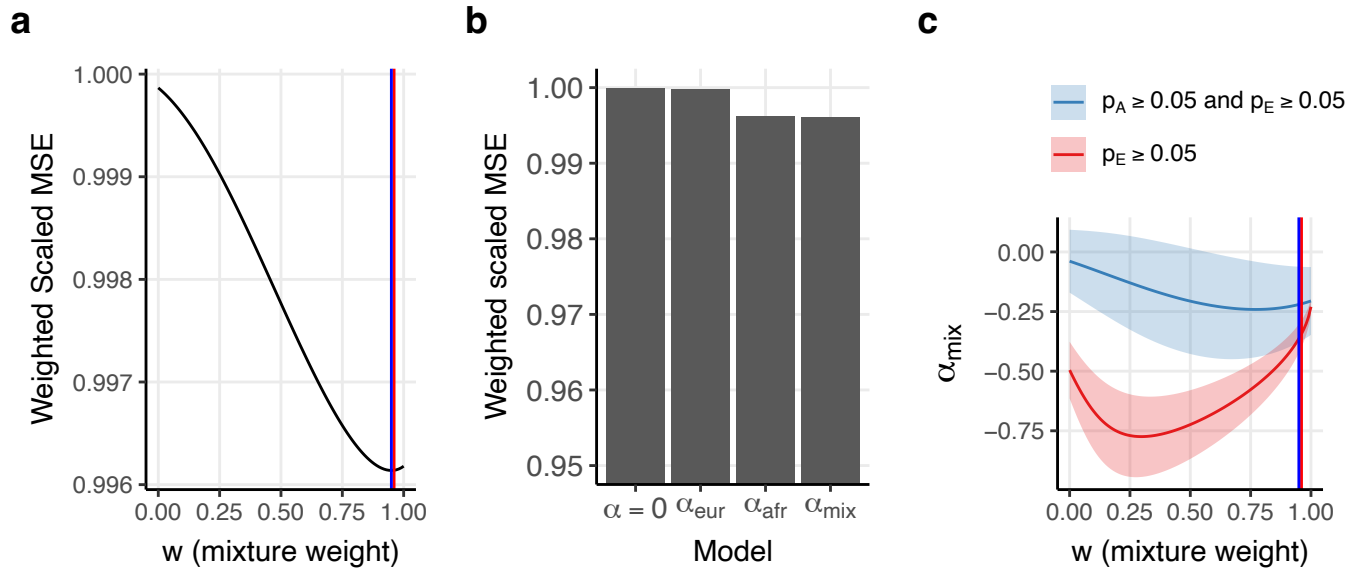
Supplementary Figure 7: The proportion of non-synonymous SNPs as function of unadmixed African and European minor allele frequencies, highlighting the marginal effects of African MAF for SNPs with $p_A \geq 0.002$ and European SNPs with $p_E \geq 0.002$: We report **(a)** the percentage of SNPs that 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.



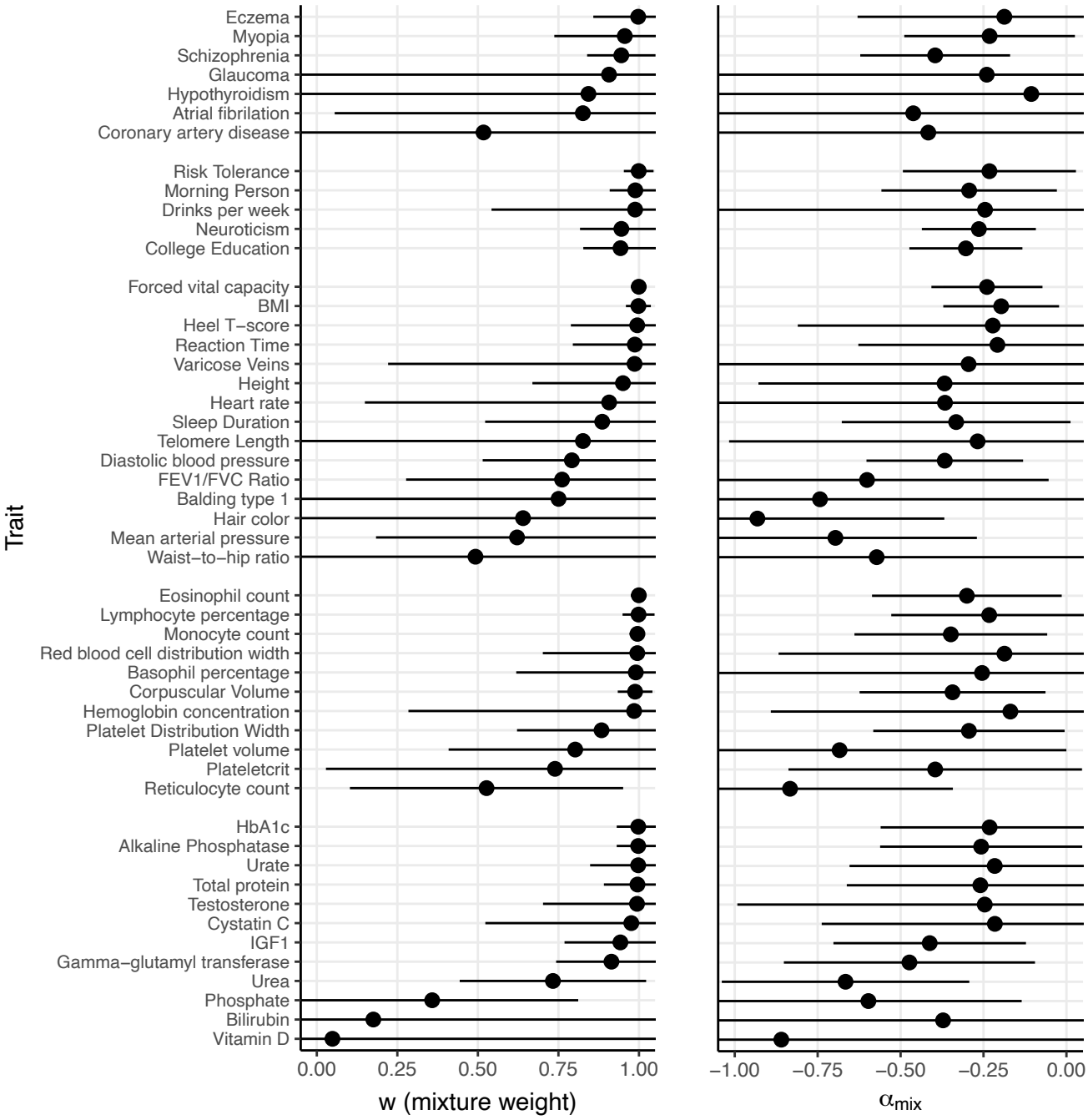
Supplementary Figure 8: Log-likelihood for a logistic regression of non-synonymous status on log(p_{mix}), varying the set of SNPs analyzed. We report the profile log-likelihood with respect to *w*, treating the regression intercept and slope as nuisance parameters. Each curve corresponds to a distinct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



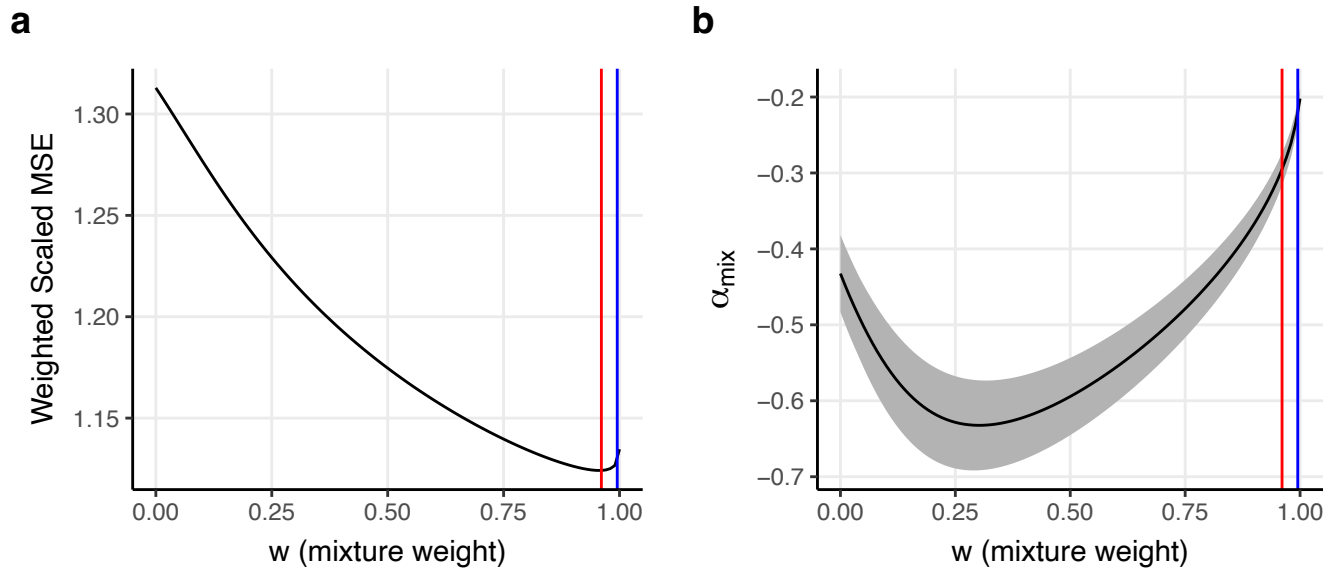
586
587 **Supplementary Figure 9: Best-fit parameters for a logistic regression of non-synonymous**
588 **status on log(p_{mix}), thresholding MAF at an alternate value of 0.05.** We report **(a)** the profile
589 log-likelihood with respect to *w*, treating the regression intercept and slope as nuisance
590 parameters, thresholding p_A and p_E at 0.05, and **(b)** the regression slope (γ) for values of *w*
591 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
592 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
593 analyses, we threshold all MAF values at 0.002 (after filtering). Red lines in (a) and (b) denote the
594 MLE estimate of *w*. The shaded region in (b) denotes 95% confidence intervals for γ.



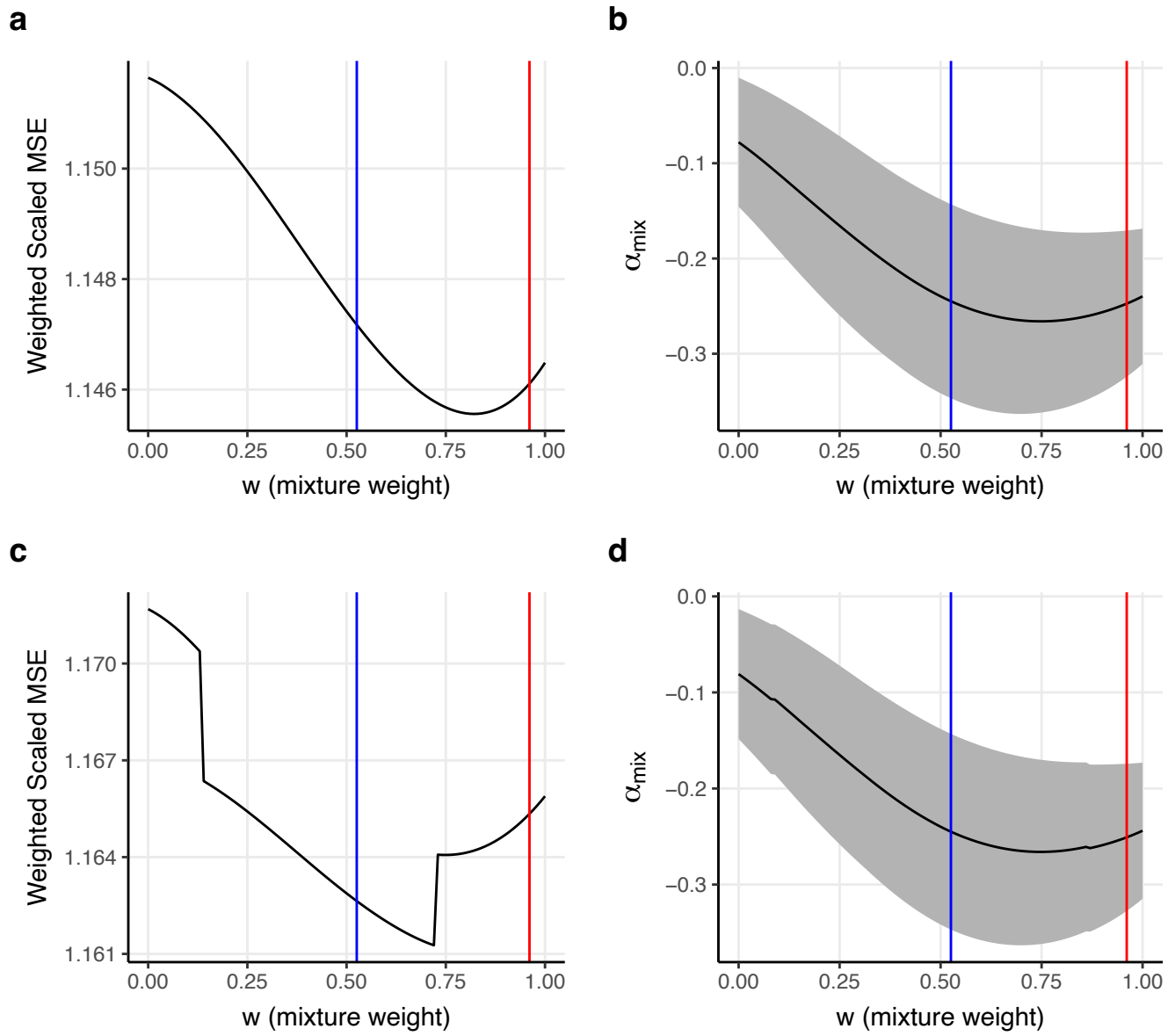
Supplementary Figure 10: Results of fitting the α_{mix} model across 50 diseases/traits restricted to variants with $p_A \geq 0.05$ and $p_E \geq 0.05$. We report **(a)** the weighted scaled mean squared error with respect to *w*, **(b)** the weighted scaled mean squared error for various models, including a model without MAF dependence (α=0), a model with European MAF dependence (α_{eur}), a model with African MAF dependence (α_{afr}), and the α_{mix} model, and **(c)** estimates of α_{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.



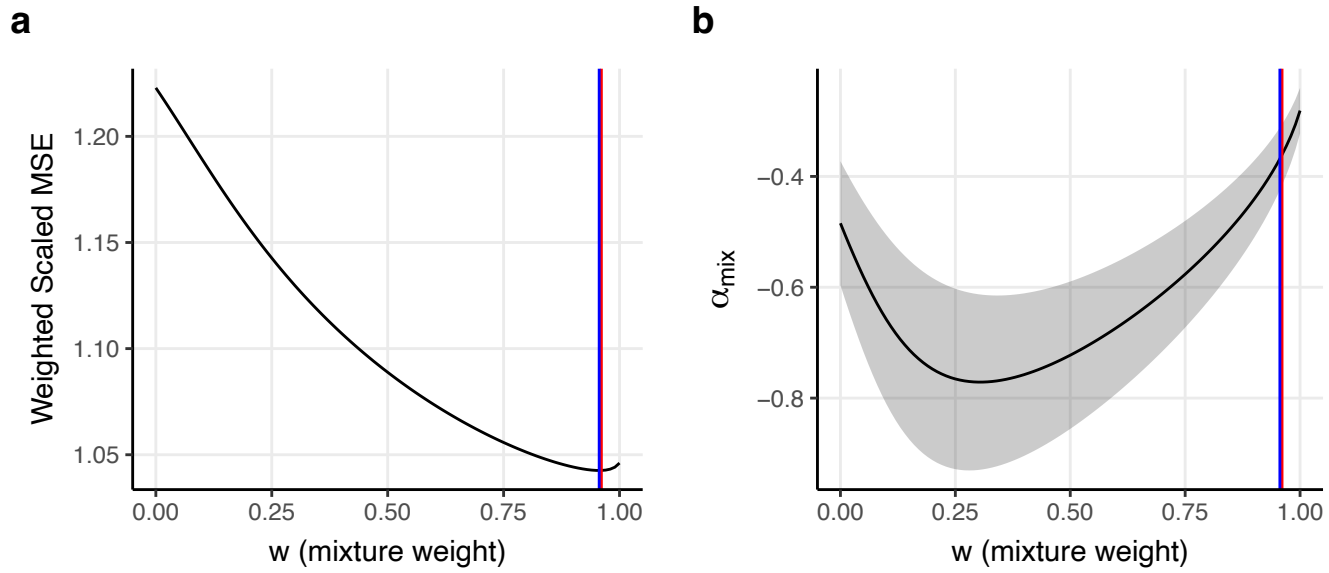
Supplementary Figure 11: Results of fitting the α_{mix} model separately for 50 diseases/traits. We report estimates of **(a)** *w* and **(b)** α_{mix}. Diseases/traits are ordered by point estimate of *w* within disease/trait categories. Error bars denote 95% confidence intervals. Numerical results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



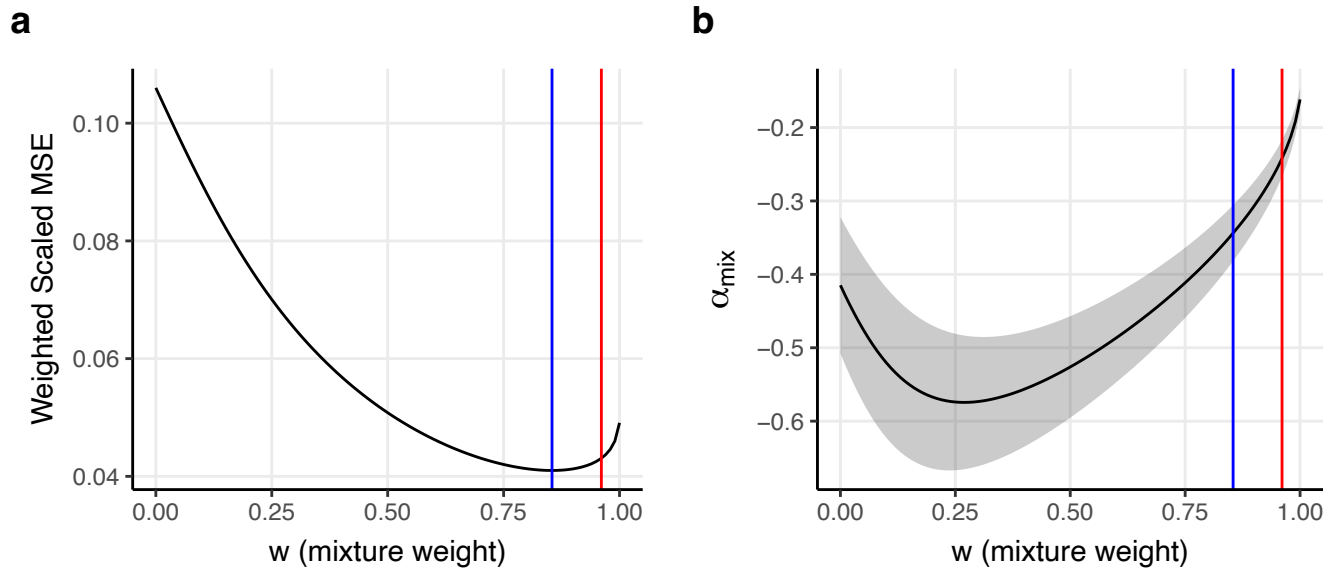
Supplementary Figure 12: **Results of fitting the α_{mix} model across 50 diseases/traits via a random effects meta-analysis of disease/trait-specific alpha-mix parameters.** We report **(a)** the weighted scaled mean squared error, summed across traits, with respect to *w*, and **(b)** estimates of α_{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 α_{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.



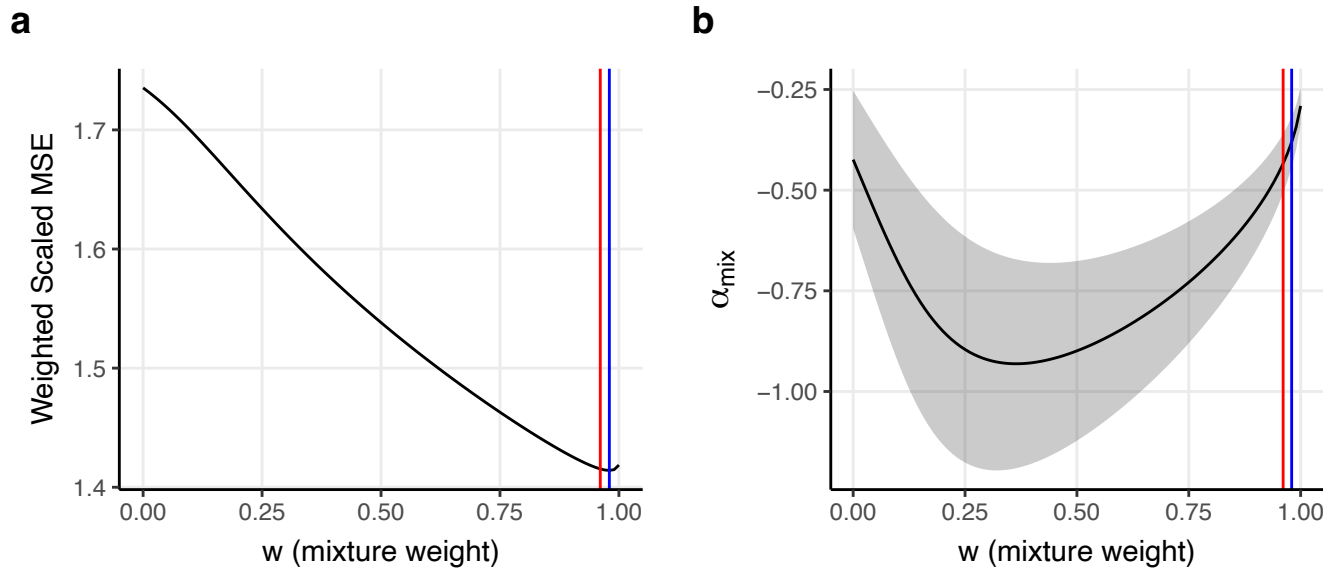
Supplementary Figure 13: **Results of fitting the α_{mix} model across 50 diseases/traits via a random effects meta-analysis of disease/trait-specific alpha-mix parameters restricted to SNPs with $p_A \geq 0.05$ and $p_E \geq 0.05$.** We report **(a)** the weighted scaled mean squared error with respect 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)** α_{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 α_{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 α_{mix} (difference of .0040).
Weighted scaled MSE was scaled by an arbitrary constant.



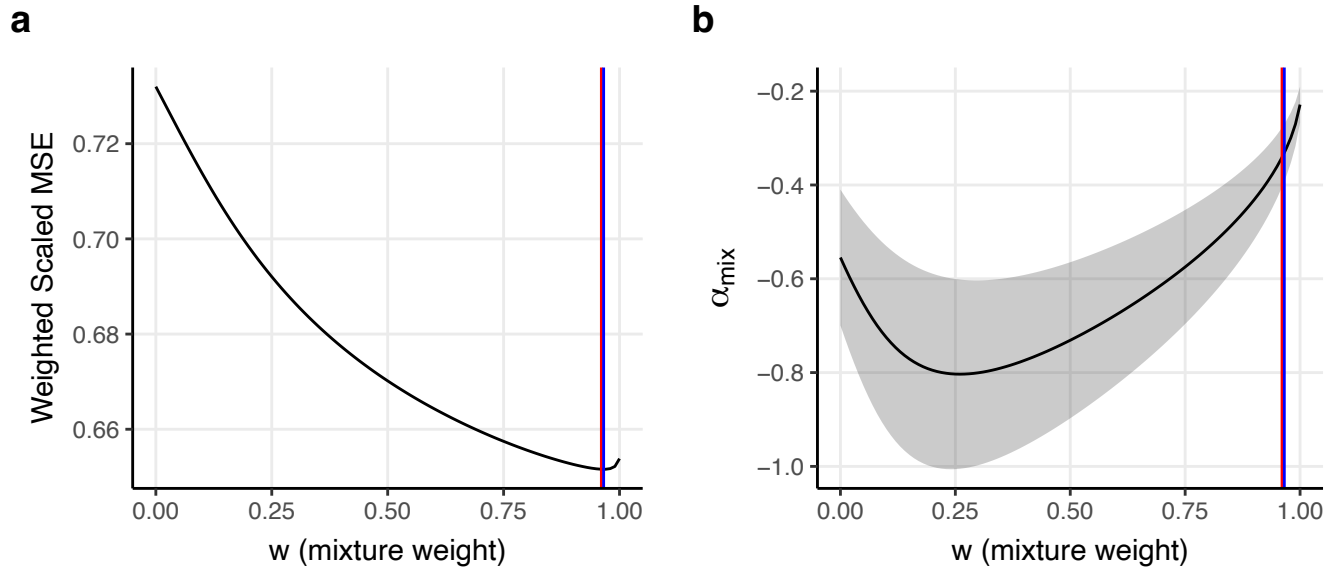
Supplementary Figure 14: Results of fitting the α_{mix} model across 50 diseases/traits while estimating African MAF using *N*=107 Yoruban genomes. We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of α_{mix} as a function of *w*. Blue lines denote our 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 α_{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 α_{mix} model.
Weighted scaled MSE was scaled by an arbitrary constant.



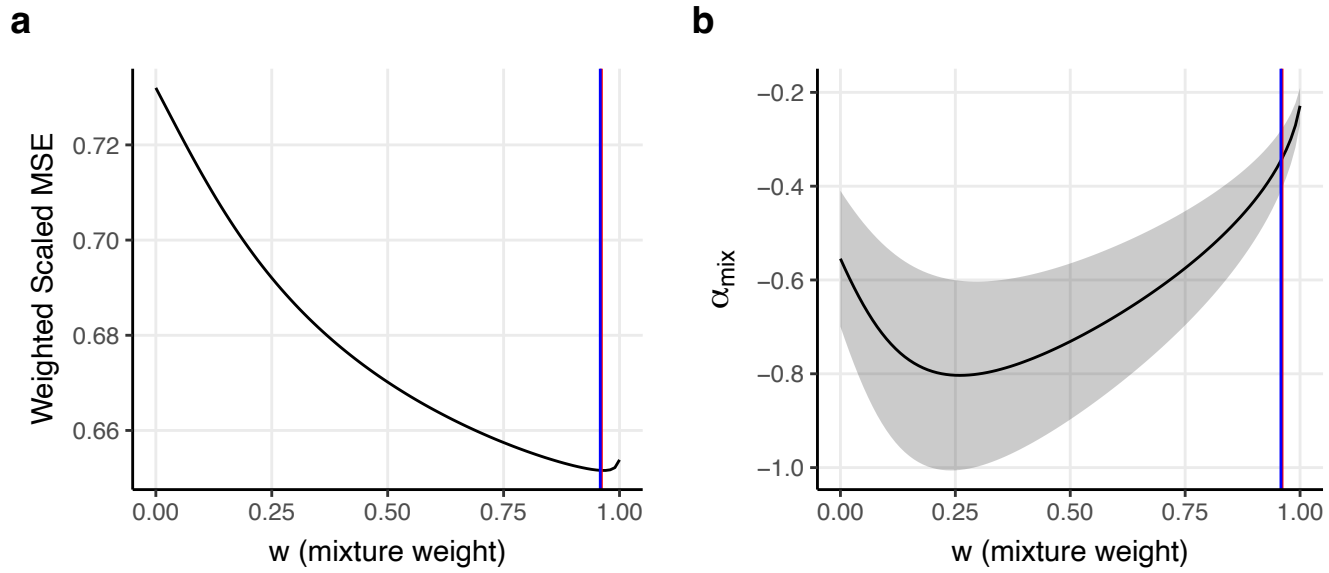
Supplementary Figure 15: Results of fitting the α_{mix} model across 50 diseases/traits using the standard Baseline-LD (v2.2) model without annotations for African MAF. We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of α_{mix} as a function of *w*. 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 α_{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.



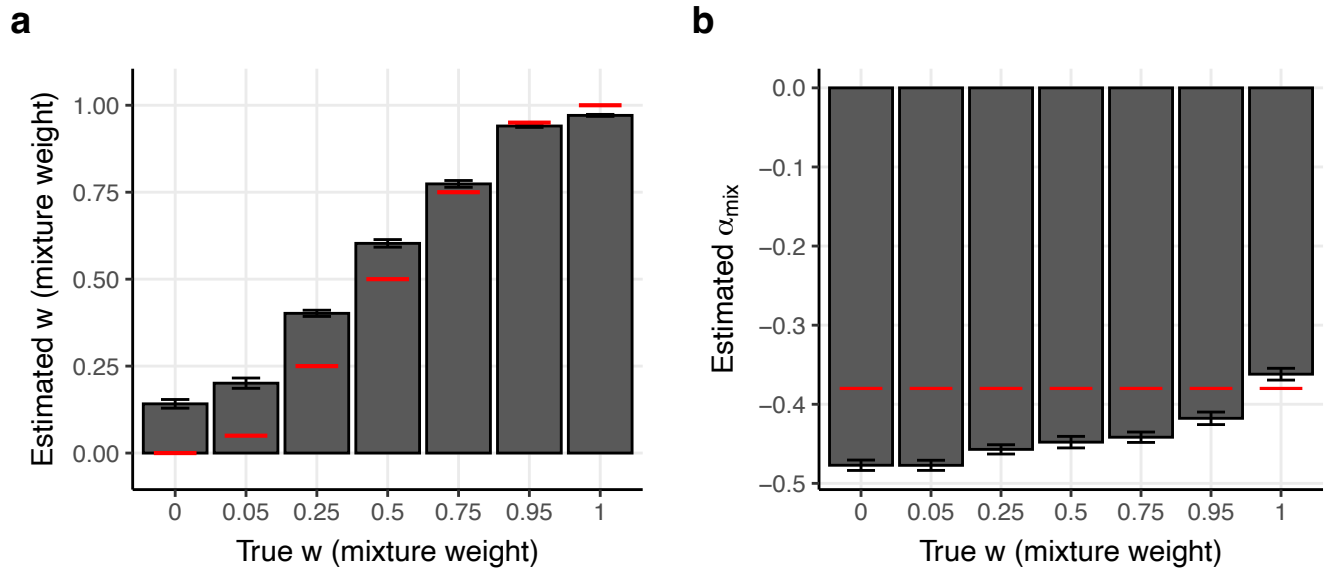
Supplementary Figure 16: Results of fitting the α_{mix} model across 50 diseases/traits using 2-dimensional MAF bin annotations without including the BaselineLD-v2.2 model. We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of α_{mix} as a function of *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 α_{mix}. Weighted scaled MSE was scaled by an arbitrary constant.



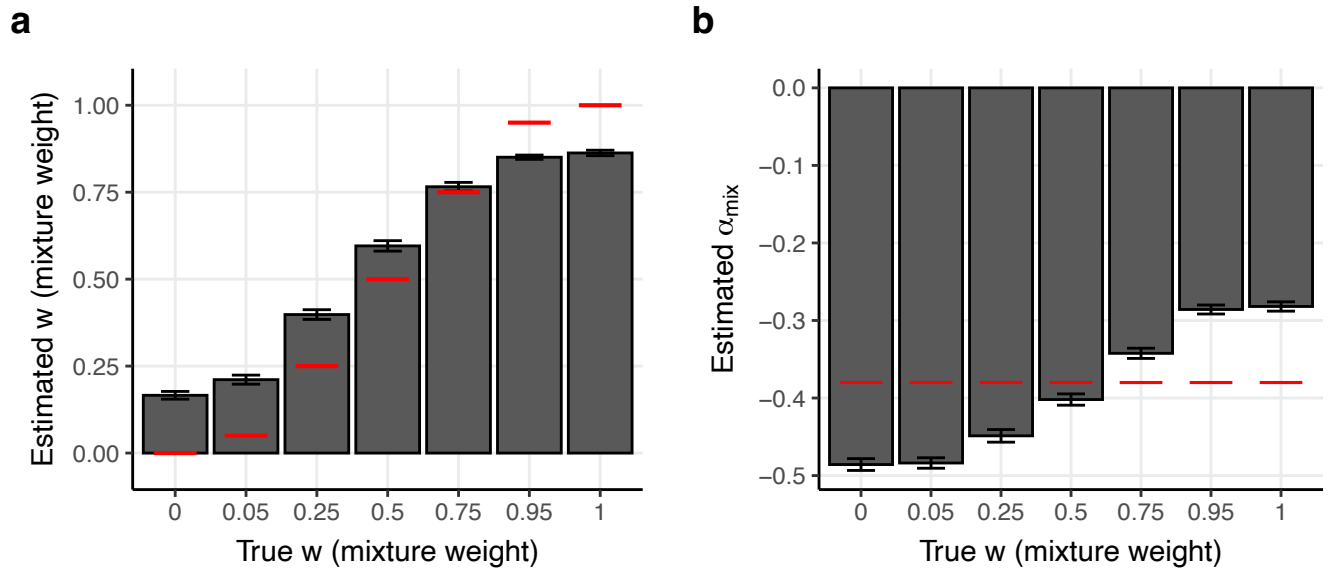
Supplementary Figure 17: Results of fitting the α_{mix} model across 29 diseases/traits from UK Biobank. We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of α_{mix} as a function of *w*. Blue lines denote our point estimate of *w* using 29 UK Biobank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



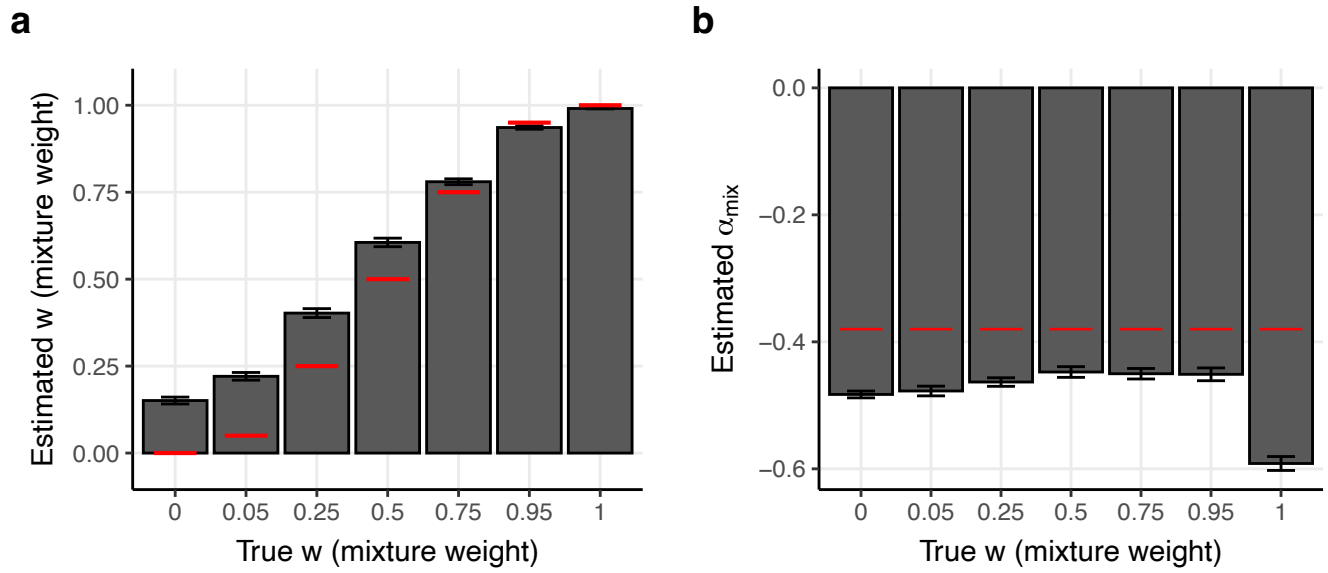
Supplementary Figure 18: Results of fitting the α_{mix} model across 21 diseases/traits from publicly available summary statistics. We report **(a)** the weighted scaled mean squared error with respect to *w*, and **(b)** estimates of α_{mix} as a function of *w*. Blue lines denote our point estimate 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 α_{mix}. Weighted scaled MSE was scaled by an arbitrary constant.



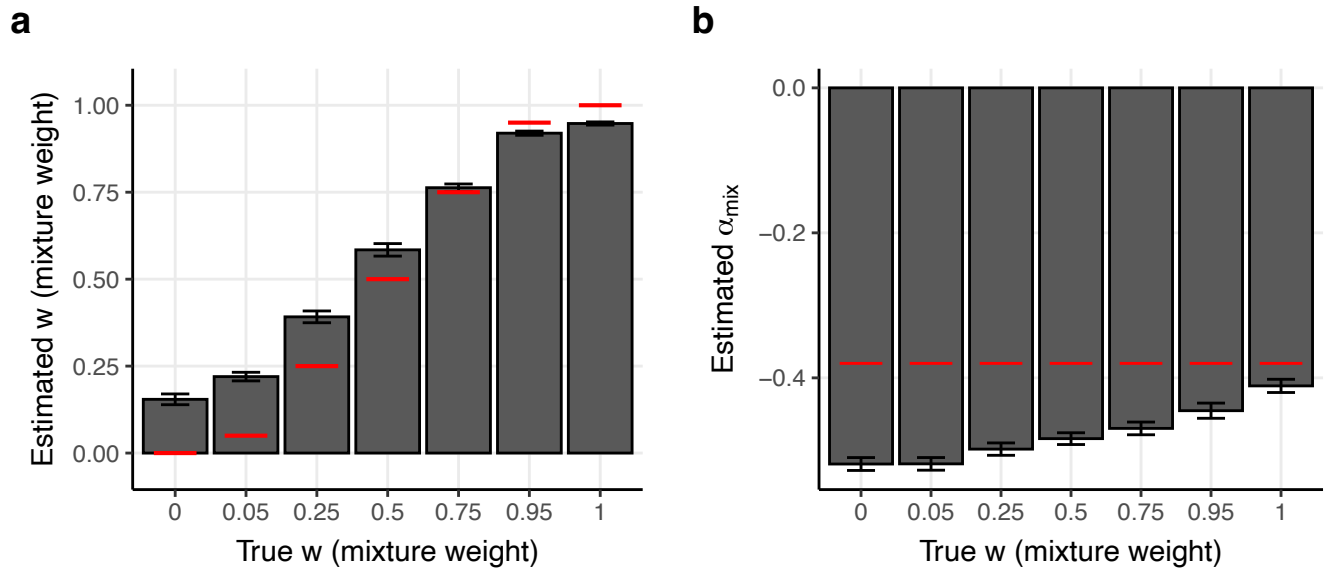
Supplementary Figure 19: Results of fitting the α_{mix} model across groups of 25 traits in simulations, while using European MAF from UK Biobank for the generative model, and European MAF from 1000 Genomes for inference. We report estimates of *w* **(a)** and α_{mix} **(b)**. Bars denote the mean estimate 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 α_{mix} inference.



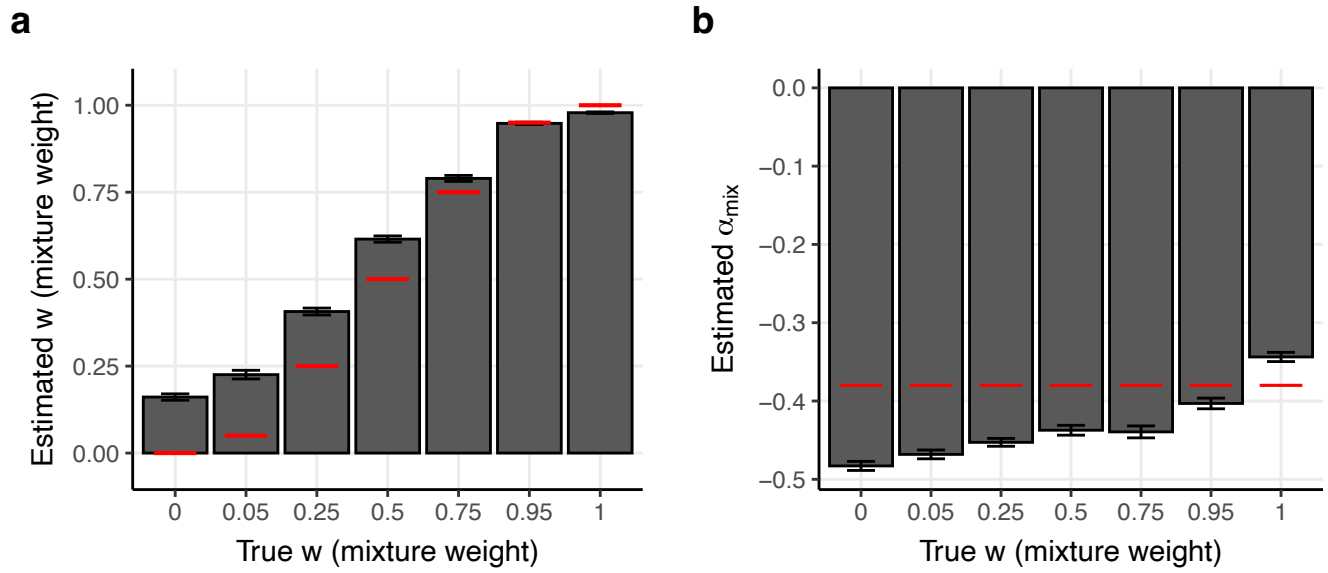
Supplementary Figure 20: Results of fitting the α_{mix} model across groups of 25 traits in simulations, while thresholding MAF at *T*=0.05. We report estimates of *w* **(a)** and α_{mix} **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value of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



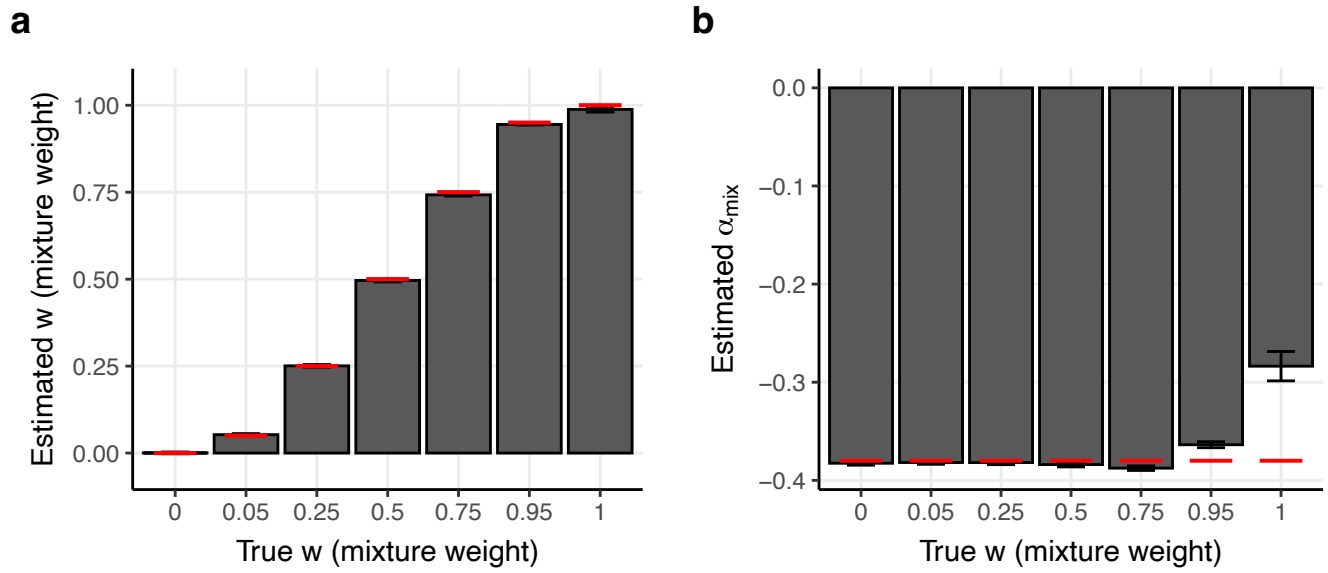
Supplementary Figure 21: Results of fitting the α_{mix} model across groups of 25 traits in simulations, while thresholding MAF at *T*=1.5*10^{-5} (equivalent to a minor allele count of 1 in our sample of genomes with local African ancestry from All of Us). We report estimates of *w* **(a)** and α_{mix} **(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 α_{mix} inference.



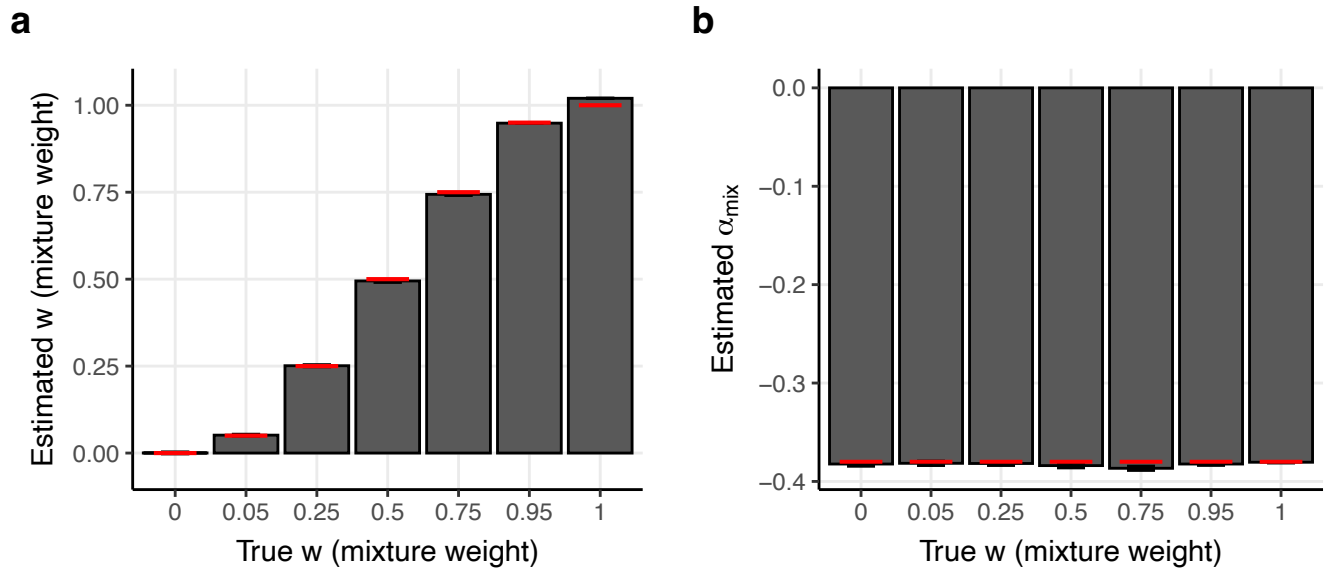
Supplementary Figure 22: Results of fitting the α_{mix} model across groups of 25 traits in simulations for traits with 5,000 causal variants. We report estimates of *w* **(a)** and α_{mix} **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value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



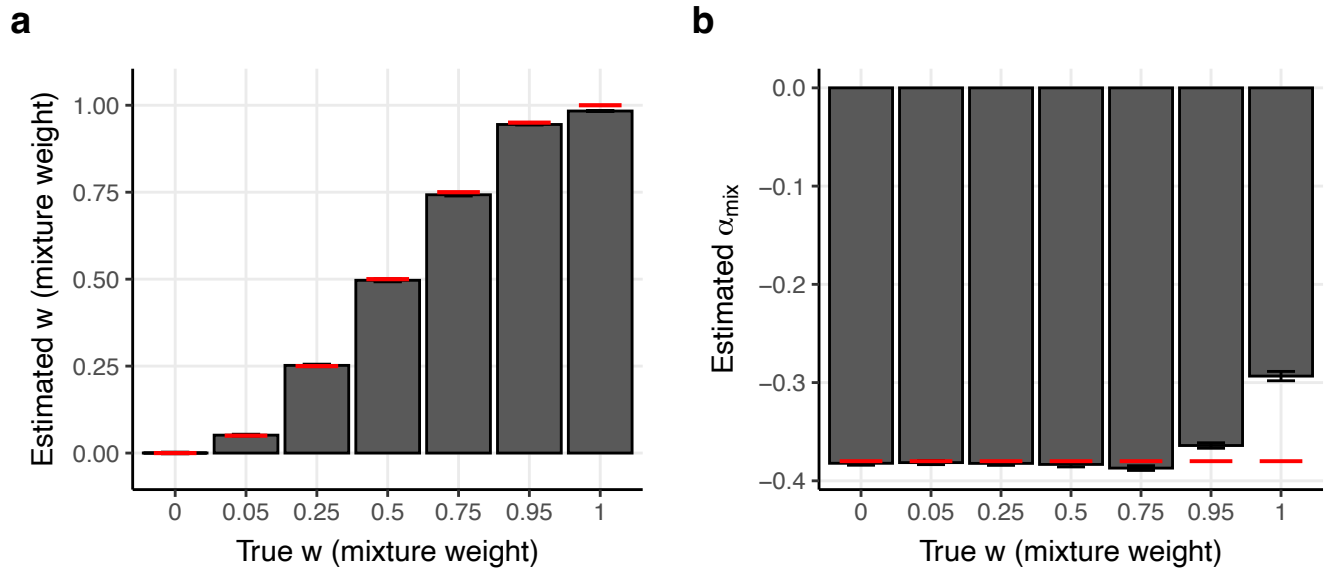
Supplementary Figure 23: Results of fitting the α_{mix} model across groups of 25 traits in simulations for traits with 20,000 causal variants. We report estimates of *w* **(a)** and α_{mix} **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value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



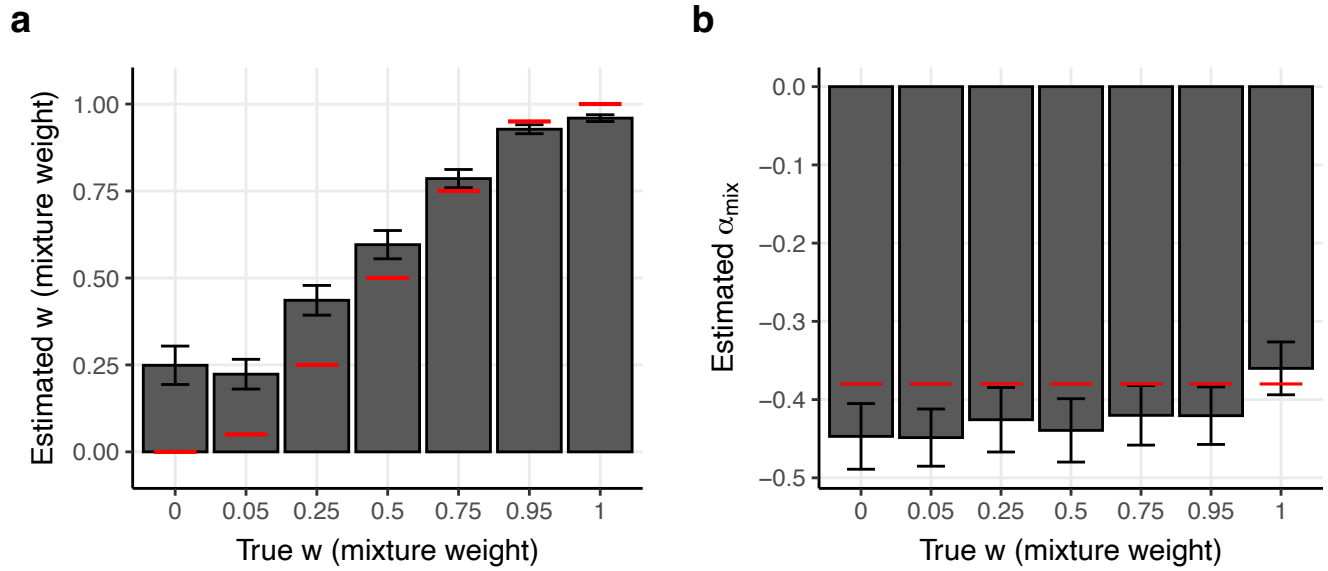
Supplementary Figure 24: Results of fitting the α_{mix} model across groups of 25 traits in simulations using exact values of β^2 during inference, rather than estimates from S-LDSC. We report estimates of *w* **(a)** and α_{mix} **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines denote the true value of each parameter. Error bars denote 95% confidence intervals.



Supplementary Figure 25: Results of fitting the α_{mix} model across groups of 25 traits in simulations while using exact values of β^2 (rather than estimates from S-LDSC), values of p_A and p_E derived from causal variants (rather than all variants), and thresholding p_{mix} at T=0.005. We report estimates of w **(a)** and α_{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. p_{mix} was computed separately for each causal SNP, thresholded at T=0.005, and then averaged over all causal SNPs in each two-dimensional MAF bin.



727
728 **Supplementary Figure 26: Results of fitting the α_{mix} model across groups of 25 traits in**
729 **simulations while using exact values of β^2 (rather than estimates from S-LDSC) and values**
730 **of p_A and p_E derived from causal variants (rather than all variants).** We report estimates of *w*
731 **(a)** and α_{mix} **(b)**. Bars denote the mean estimate of each parameter across 40 replicates. Red lines
732 denote the true value of each parameter. Error bars denote 95% confidence intervals. p_{mix} was
733 computed separately for each causal SNP, thresholded at $1.5*10^{-5}$ (equivalent to a minor allele
734 count of 1 in our sample of genomes with local African ancestry from All of Us), and then averaged
735 over all causal SNPs in each two-dimensional MAF bin.



Supplementary Figure 27: Results of fitting the α_{mix} model separately for each trait in simulations. We report estimates of *w* **(a)** and α_{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.

742 **Methods**

743 *Minor allele frequency estimation*

744 Throughout this manuscript we refer to unadmixed African MAF as African MAF. We used
745 continental ancestry group assignments provided by All of Us[42] to identify individuals residing in the
746 United States with predominantly African ancestry; we note that segments of African ancestry in
747 these individuals are primarily of West African origin[48], such that our African MAF estimates
748 primarily represent West African ancestry. We defined the MAF of a SNP in a specific ancestry to
749 be the frequency of the less common (minor) allele *in that ancestry*. As a result, for a single SNP, it
750 is possible for the minor allele in Africans to be different from the minor allele in Europeans.

751 We estimated unadmixed African MAF using All of Us[42] (see Data Availability) whole
752 genome-sequencing data for *N*=46,672 unrelated individuals labeled as African ancestry by All of
753 Us (AoU). This cohort has moderate levels of European admixture on average, so estimating
754 unadmixed allele frequencies (AF) is more complex than in an unadmixed cohort[74]. We restricted
755 to 30,906,411 single-nucleotide polymorphisms with either European MAF ≥ 0.001 or admixed
756 African MAF ≥ 0.001 (as reported by the AoU Variant Annotation Table) in the All of Us v8 short
757 read whole-genome sequencing Allele Count/Allele Frequency threshold Plink dataset. We first
758 estimated the local ancestry of each individual using RFMix2 [49] with 1000 Genomes[44] YRI, CEU,
759 PEL, EAS (concatenation of CHS, CDX, KHV, CHB, JPT), and SAS (concatenation of PJL, BEB,
760 STU, ITU, GIH) as the reference populations, restricting to HapMap3 variants. The local ancestry
761 of SNPs that were not included in HapMap3 was interpolated between that of the closest HapMap3
762 SNPs. Next, we computed the unadmixed African allele frequency of each SNP using Plinkv1.9[75]
763 while restricting to individuals with two African haplotypes at that SNP.

764 We estimated unadmixed European MAF in 70,346 unrelated European ancestry individuals
765 from the All of Us v8 short read whole-genome sequencing Allele Count/Allele Frequency threshold
766 Plink dataset, restricting to the same set of variants used for African MAF estimation. These

767 European MAF estimates were used for analyses of non-synonymous variation (Figures 1-2), but
768 not analyses for analyses of diseases/traits (Figures 3-6).

769 For S-LDSC analyses of diseases/traits, unadmixed African AF was estimated as described
770 above, but using the All of Us v7.1 short read whole-genome sequencing Allele Count/Allele
771 Frequency threshold Plink dataset. AF were converted from GRCh38 to GRCh37 using
772 BCFtools/liftover[76]. The African MAF of SNPs in 1000 Genomes that could not be matched to any
773 of the SNPs from AoU, was calculated using *N*=107 individuals from Yoruba in 1000 Genomes
774 restricting to unrelated, QC'ed individuals from TGP2261 from ref.[77]. For S-LDSC analyses of
775 diseases/traits, European AF was calculated using *N*=489 European ancestry individuals from
776 1000 Genomes restricting to unrelated, QC'ed individuals from TGP2261 from ref.[77]. We used
777 1000 Genomes rather than the All of Us to ensure consistency between our MAF estimates and
778 those used in the Baseline-LD (v2.2) model, where MAF and LD scores are calculated using the
779 same 489 individuals from 1000 Genomes.

780 Throughout this manuscript, we stratified variants by African and European allele frequency
781 using a common set of bin boundaries, which are justified as follows. SNPs were first stratified by
782 common (MAF ≥ 0.05), low-frequency (0.005 ≤ MAF < 0.05), and rare (MAF < 0.005) categories. In
783 some analyses, we further stratified common variants into 5 or 10 bins with boundaries
784 corresponding to 1000 Genomes European ancestry common MAF quintiles or deciles as defined
785 by the Baseline-LD (v2.2) model. We used 5 common variant bins for visualization (e.g. Figures
786 1a, 3a) and in-text reporting of non-parametric MAF-dependence statistics (paragraph 1 of the *Per-*
787 *allele disease and complex trait effect sizes are predominantly African MAF-dependent in*
788 *European populations* subsection of Results) to improve interpretability. We used 10 common
789 variant bins when fitting the α_{mix} model (Figures 4-5) because using more bins decreases the
790 absolute difference between the MAF of each SNP in a bin and the mean MAF of the bin. In All of
791 Us analyses, rare SNPs were sometimes further stratified using a boundary of 0.002 to create a

792 rare SNP bin that avoids ascertainment bias (e.g. a SNP with unadmixed African MAF=0.0011 has
793 African American MAF around 0.0008 < 0.001 and thus would typically be excluded by our initial
794 MAF filter; however, a SNP with unadmixed African MAF=0.002 has African American MAF around
795 0.0016 > 0.001 and thus would typically be included in our dataset).
796
797 *Modeling non-synonymous status as a function of MAF*
798 We defined a set of non-synonymous SNPs in AoU by grouping the following values for the
799 "consequence" field in the All of Us v8 Variant Annotation Table: frameshift_variant,
800 inframe_deletion, inframe_insertion, missense_variant, start_lost, stop_gained, stop_lost. We
801 computed the proportion of SNPs that are non-synonymous (PNS) in a stratum with respect to all
802 SNPs, including SNPs in non-coding regions.
803 We modeled the probability that a SNP is non-synonymous as a function of p_{mix} via the
804 following logistic regression model:

805 $$\log\left(\frac{P(Y)}{1-P(Y)}\right) = \mu + \gamma \log(p_{mix}), \quad (2)$$

806 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,
807 this is equivalent to fitting a logistic regression using European or African, respectively. To estimate
808 w, we computed a profile likelihood, treating μ and γ as nuisance parameters, and used a
809 numerical optimization method to maximize the profile likelihood with respect to w. The profile
810 likelihood is given by:

811 $$L(w) = \max_{\gamma,\mu} L(w,\gamma,\mu), \quad (3)$$

812 where $L(w,\gamma,\mu)$ is the likelihood from equation 2. We calculated $\max_{\gamma,\mu} L(w,\gamma,\mu)$ for a specific choice
813 of w using discrete.discrete_model.Logit from the statsmodels[78] Python package and numerically
814 identified the maximum of $L(w)$ with respect to w using optimize.minimize_scalar from the SciPy

815 Python package[79]. While fitting logistic regression models, values of p_A and p_E < 0.002 were set to
816 0.002, except for in Supplementary Figure 8, where a threshold of 0.05 is used.
817 Significance testing related to logistic regressions was performed in four ways depending on
818 the specific analysis. We estimated P-values for γ in p_A-only and p_E-only univariate logistic
819 regressions using a two-tailed t-test. When comparing model fit for a p_A-only model to a p_E-only
820 model on a shared set of SNPs, we computed p-values using Vuong's test because both models
821 had the same number of degrees of freedom (precluding the use of a standard likelihood ratio
822 test). When comparing r^2 and regression coefficients for a p_A-only model to a p_E-only model which
823 were fit on different sets of SNPs, we computed p-values using a genomic block jackknife with 200
824 approximately equally sized blocks. All p_{mix} logistic regression *P*-values and confidence intervals
825 were computed by using a likelihood ratio test statistic with 1 degree of freedom. Throughout the
826 manuscript, *P*-values less than 10^{-15} were reported as $P<10^{-15}$.
827
828 *Disease/trait selection*
829 We first collected a set of 364 GWAS summary statistics in European ancestry cohorts
830 across a wide range of complex traits and diseases. For each complex trait or disease, we
831 estimated heritability using S-LDSC with baseline-LD (v2.2) LD scores. For each pair of complex
832 traits and diseases, we estimated genetic correlation using S-LDSC[80] without partitioning based on
833 functional annotations. Next, we filtered to traits with heritability z-score greater than 6 and
834 $N*h^2$>20,000. Finally, we greedily selected traits based on $N*h^2$, subject to the constraint that for
835 each pair of selected traits, the squared genetic correlation < 0.1. The output of this procedure was
836 50 diseases/complex traits (average *N*=483K).
837
838 *Estimation of per-allele effect size variance via stratified linkage disequilibrium score regression*

839 Stratified linkage disequilibrium score regression[3] (S-LDSC) estimates the common SNP
840 heritability of a GWAS complex trait or disease partitioned across predefined functional
841 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
842 *d* indexes a set of *D* functional annotations, $a_d(i)$ is the value of annotation *d* for SNP *i*, and τ_d
843 represents the increase in h_i^2 associated with a unit increase in $a_d(i)$. S-LDSC requires the
844 specification of a set of *D* functional annotations that explain how heritability varies across the
845 genome.

846 We applied S-LDSC using annotations from a new extension of the Baseline-LD (v2.2)
847 model[5]. The Baseline-LD (v2.2) model includes standard functional annotations in addition to MAF
848 and linkage disequilibrium (LD) annotations that reduce MAF and LD-related bias. Our extension of
849 the Baseline-LD (v2.2) model incorporated new annotations based on bivariate MAF bins, where
850 each bivariate MAF bin is defined by an African MAF bin and a European MAF bin. Specifically, we
851 defined 12 univariate unadmixed African MAF bins: 10 for common African variants ($p_A \geq 0.05$), 1
852 for low-frequency African variants ($0.005 \leq p_A < 0.05$), and 1 for rare African variants ($p_A < 0.005$).
853 The boundaries of the 10 common variant bins were the same as those used for univariate
854 European common MAF in the Baseline-LD (v2.2) model. We then created 120 bivariate MAF bins
855 by stratifying SNPs based on the interaction between the 12 univariate African MAF bins and 10
856 univariate European MAF bins. We removed the 10 univariate European MAF bins from the
857 combined model to avoid collinearity of functional annotations. We calculated partitioned LD scores
858 using 489 European ancestry individuals from 1000 Genomes.

859 To estimate the *per-allele* effect size variance of variant *i*, we divided $\widehat{h^2}$ (as estimated by S-
860 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
861 the European MAF for SNP *i*. This procedure has been previously applied in ref.[81]. We calculated
862 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

863 the number of SNPs contained in *C*. When meta-analyzing across traits, we normalized $\widehat{\beta_t^2}$ within

864 each trait before taking the mean ($\frac{\sum_{i \in C} \widehat{\beta_t^2}/n_C}{\sum_i \widehat{\beta_t^2}/n}$, where *n* is the number of SNPs with $p_E > 0.05$), and

865 meta-analyzed these normalized quantities using random-effect meta-analysis as implemented in

866 stats.meta_analysis.combine_effects from the statsmodels Python package.

867

868 *Estimation of alpha-mix parameters*

869 We propose the α_{mix} model which relates per-allele effect size variance to unadmixed

870 African and European MAF:

871 $$\mathrm{E}[\beta_t^2 | p_A, p_E] = \sigma_{g,t}^2 [2p_{mix}(1 - p_{mix})]^{\alpha_{mix}} \quad (4)$$

872 where β_t^2 is the squared *per-allele* effect size of a single-nucleotide polymorphism (SNP) in trait *t*,

873 $\sigma_{g,t}^2$ is a variance parameter, $p_{mix}=w*p_A + (1-w)*p_E$, *w* specifies the relative contributions of p_A and

874 p_E, and α_{mix} specifies the level of MAF dependence. Equation 4 is identical to equation 1 with a

875 trait-specific subscript.

876 We estimated the parameters of the α_{mix} model via a generalized method of moments

877 estimator, with population moments from equation 4 and conditional sample moments estimated

878 via S-LDSC. We first partitioned variants with European MAF ≥ 0.05 into 120 two-dimensional MAF

879 bins based on the interaction between 12 African MAF bins and 10 European MAF bins. We then

880 defined one moment condition for each trait-bin pair: $E\left[\sigma_{g,t}^2 \left[2\overline{p_{mix,k}}\left(1 - \overline{p_{mix,k}}\right)\right]^{\alpha_{mix}} - \widehat{\beta_{t,k}^2}\right] = 0$,

881 where $\widehat{\beta_{t,k}^2}$, is the S-LDSC estimate of the per-allele effect size variance (described in *Estimation of*

882 *per-allele effect size variance via stratified linkage disequilibrium score regression*) and $\overline{p_{mix,k}}$ is the

883 mean of p_{mix} across all SNPs in bin *k*.

884 We next applied a numerical optimization procedure to minimize the sum of squared

885 moment conditions, weighted by the number of SNPs in each bin. When estimating alpha-mix

886 parameters for a single trait, we optimized the following loss function: $L_t\left(w, \alpha_{mix}, \sigma_{g,t}^2\right) =$

887 $\sum_k^{120} n_k\left[\sigma_{g,t}^2\left[2\overline{p_{mix,k}}\left(1-\overline{p_{mix,k}}\right)\right]^{\alpha_{mix}}-\widehat{\beta_{t,k}^2}\right]^2$, using the differential evolution optimization method

888 with default parameters as implemented by the SciPy package to search over α_{mix} ∈ [-1.5, 1.5]

889 and w ∈ [-0.5,1.5], while solving for $\sigma_{g,t}^2$ analytically (see below).

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

891 loss function:

892 $$L\left(w,\alpha_{mix},\sigma_{g,1}^2,\ldots,\sigma_{g,T}^2\right)=\sum_t^T L_t\left(w,\alpha_{mix},\sigma_{g,t}^2\right)=\sum_t^T\sum_k^{120} n_k\left[\sigma_{g,t}^2\left[2\overline{p_{mix,k}}\left(1-\overline{p_{mix,k}}\right)\right]^{\alpha_{mix}}-\beta_{t,k}^2\right]^2 \quad (5)$$

893 To avoid optimizing this function numerically with respect to *T* + 2 unknown parameters,

894 note that for a fixed choice of *w* and α_{mix},

895 $$\mathrm{argmin}_{\sigma_{g,t}^2} L\left(w,\alpha_{mix},\sigma_{g,1}^2,\ldots,\sigma_{g,T}^2\right)=\frac{\sum_{k=1}^{120} n_k\widehat{\beta_{t,k}^2}\left[\overline{p_{mix,k}}\left(1-\overline{p_{mix,k}}\right)\right]^{\alpha_{mix}}}{2\sum_{k=1}^{120} n_k\left[\overline{p_{mix,k}}\left(1-\overline{p_{mix,k}}\right)\right]^{2\alpha_{mix}}} \quad (6)$$

896 so we can treat this as an optimization problem with two unknown parameters, *w* and α_{mix}, and

897 use equation 6 to identify $\sigma_{g,1}^2,\ldots,\sigma_{g,T}^2$ analytically, in turn, for each trait. We minimized *w* and

898 α_{mix} using the SciPy package implementation of the differential evolution method with default

899 parameters, restricting α_{mix} to [-1.5, 1.5], and w to [-0.5, 1.5].

900 To estimate uncertainty of parameter estimates and perform significance testing, we

901 wrapped the entire procedure in a genomic block-jackknife with 200 blocks (including both S-LDSC

902 parameter estimation and GMM estimation). When estimating α_{mix} and *w* jointly across traits, we

903 used a modified version of S-LDSC which forces jackknife block boundaries to be identical across

904 summary statistics that include distinct sets of SNPs.

905

906 *Simulations*

907 We simulated quantitative traits using imputed genotypes for 337,448 unrelated British

908 individuals from UK Biobank (version 3 of the Haplotype Reference Consortium imputed data) and

909 5,907,305 autosomal SNPs with p_E ≥ 0.05 in 1000 Genomes and INFO score ≥ 0.60. Quantitative

910 traits were simulated under a standard additive model: $Y = X\beta + \epsilon$, where Y is a phenotype vector,
911 X is a genotype matrix, β is a sparse vector with n_{causal} non-zero entries representing per-allele
912 causal effects, and ϵ is normally distributed residual variance. The non-zero elements of β were
913 sampled separately for each trait. The effects of non-zero elements of β were sampled from a
914 normal distribution with mean 0 and variance $\sigma_g^2 * [2 * \max(T, p_{mix}) * (1 - \max(T, p_{mix}))]^{\alpha_{mix}}$,
915 where T represents a threshold below which varying MAF does not change effect size (see left
916 side of Figure 2 from ref.[19]). When simulating effects, we used 1000 Genomes for European MAF,
917 and All of Us and 1000 Genomes allele frequencies for African MAF (see the *Minor allele*
918 *frequency estimation* subsection of Methods), except when stated otherwise. β was scaled such
919 that $var(X\beta) = var(Y)/2$ to set h^2=0.5. GWAS and computation of $X\beta$ were performed using Plink
920 v2.0[82].
921

922 **Code Availability**

923 We have publicly released custom code used to estimate α_{mix} and perform all analyses described
924 in this manuscript at https://github.com/jordanero/alphamix.
925

926 **Data Availability**

927 GWAS summary statistics and S-LDSC results for all diseases/traits are publicly available at
928 https://zenodo.org/records/18292998. Unadmixed African MAF estimates are available at
929 https://zenodo.org/records/18292998 (MAF corresponding to a minor allele count less than 40 has
930 been censored in accordance with the All of Us Data and Statistics Dissemination Policy). AoU
931 short read individual-level WGS data are available to authorized users on the All of Us Researcher
932 Workbench. The UKBB resource is publicly available by application (http://www.ukbiobank.ac.uk).

933 **Acknowledgements**

We are grateful to Xihong Lin, Po-Ru Loh, David Reich, Xilin Jiang, Mashaal Sohail, Bogdan Pasaniuc, Sasha Gusev, and Benedikt Geiger 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, 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.

Competing interests

J.R. is an employee of Genentech. The remaining authors declare no competing interests.

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

016 35. Speed, D. & Evans, D. M. Estimating disease heritability from complex pedigrees allowing for
017 ascertainment and covariates. *Am. J. Hum. Genet.* **111**, 680–690 (2024).
018 36. Henches, L. *et al.* Polygenic risk score prediction accuracy convergence. *Hum. Genet. Genomics Adv.* **6**,
019 100457 (2025).
020 37. Hof, J. P. & Speed, D. LDAK-KVIK performs fast and powerful mixed-model association analysis of
021 quantitative and binary phenotypes. *Nat. Genet.* https://doi.org/10.1101/2024.07.25.24311005 (2025)
022 doi:10.1101/2024.07.25.24311005.
023 38. Wainschtein, P. *et al.* Estimation and mapping of the missing heritability of human phenotypes. *Nature*
024 https://doi.org/10.1038/s41586-025-09720-6 (2025) doi:10.1038/s41586-025-09720-6.
025 39. Patel, R. A. *et al.* Characterizing selection on complex traits through conditional frequency spectra.
026 *GENETICS* **229**, (2025).
027 40. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-Wide Inference of Ancestral
028 Recombination Graphs. *PLoS Genet.* **10**, e1004342 (2014).
029 41. Nait Saada, J., Tsangalidou, Z., Stricker, M. & Palamara, P. F. Inference of Coalescence Times and Variant
030 Ages Using Convolutional Neural Networks. *Mol. Biol. Evol.* **40**, msad211 (2023).
031 42. The All of Us Research Program Genomics Investigators *et al.* Genomic data in the All of Us Research
032 Program. *Nature* **627**, 340–346 (2024).
033 43. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach
034 for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
035 44. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature*
036 **526**, 68–74 (2015).
037 45. Tishkoff, S. A. *et al.* The Genetic Structure and History of Africans and African Americans. *Science* **324**,
038 1035–1044 (2009).

039 46. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**,
040 327–332 (2015).
041 47. Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741–
042 748 (2020).
043 48. Price, A. L. *et al.* Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed
044 Populations. *PLoS Genet.* **5**, e1000519 (2009).
045 49. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most Rare Missense Alleles Are Deleterious in Humans:
046 Implications for Complex Disease and Association Studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
047 50. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous
048 coding variants. *Nat. Genet.* **42**, 969–972 (2010).
049 51. Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing
050 of Human Exomes. *Science* **337**, 64–69 (2012).
051 52. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants.
052 *Nature* **493**, 216–220 (2013).
053 53. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291
054 (2016).
055 54. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**,
056 173–178 (2017).
057 55. Kanai, M. *et al. Insights from Complex Trait Fine-Mapping across Diverse Populations*.
058 http://medrxiv.org/lookup/doi/10.1101/2021.09.03.21262975 (2021)
059 doi:10.1101/2021.09.03.21262975.
060 56. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*
061 **507**, 354–357 (2014).
062 57. Harris, K. & Nielsen, R. The Genetic Cost of Neanderthal Introgression. *Genetics* **203**, 881–891 (2016).

063 58. Juric, I., Aeschbacher, S. & Coop, G. The Strength of Selection against Neanderthal Introgression. *PLOS*
064 *Genet.* **12**, e1006340 (2016).
065 59. Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-
066 genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
067 60. Koch, E. *et al.* Genetic association data are broadly consistent with stabilizing selection shaping human
068 common diseases and traits. Preprint at https://doi.org/10.1101/2024.06.19.599789 (2024).
069 61. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-
070 population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
071 62. Fatumo, S. *et al.* Promoting the genomic revolution in Africa through the Nigerian 100K Genome
072 Project. *Nat. Genet.* **54**, 531–536 (2022).
073 63. Feng, Y.-C. A. *et al.* Taiwan Biobank: A rich biomedical research database of the Taiwanese population.
074 *Cell Genomics* **2**, 100197 (2022).
075 64. Ziyatdinov, A. *et al.* Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature*
076 **622**, 784–793 (2023).
077 65. Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. *Cell*
078 *Genomics* **3**, 100361 (2023).
079 66. Fatumo, S. *et al.* KidneyGenAfrica: A multi-cohort Genome-wide association study and polygenic
080 prediction of kidney function in 110,000 continental and diasporan Africans. Preprint at
081 https://doi.org/10.21203/rs.3.rs-5555692/v1 (2024).
082 67. Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran
083 Program. *Science* **385**, eadj1182 (2024).
084 68. Lee, H. A. *et al.* Data Resource Profile: The Statistics of the Korea National Health and Nutrition
085 Examination Survey (KNHANES) Biobank Project. *J. Korean Med. Sci.* **40**, e189 (2025).

69. Fan, S. *et al.* Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**, 923-939.e14 (2023).

70. Gouveia, M. H. *et al.* Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies. *Nat. Commun.* **14**, 6802 (2023).

71. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

72. Haller, B. C. & Messer, P. W. SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Mol. Biol. Evol.* **34**, 230–240 (2017).

73. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet.* **9**, e1003264 (2013).

74. Sharma, S. *et al.* Genetic ancestry and population structure in the All of Us Research Program cohort. Preprint at https://doi.org/10.1101/2024.12.21.629909 (2024).

75. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

76. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics* **40**, (2024).

77. Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E. & Leutenegger, A.-L. High level of inbreeding in final phase of 1000 Genomes Project. *Sci. Rep.* **5**, 17453 (2015).

78. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in *Proceedings of the Python in Science Conference* 92–96 (SciPy, Austin, Texas, 2010). doi:10.25080/majora-92bf1922-011.

79. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

80. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

81. Sullivan, P. F. *et al.* Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023).

82. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).