Supplementary material for "Leveraging fine-mapping and multi-population training data to improve cross-population polygenic risk scores"

Supplementary Table Captions

Supplementary Table 1: Detailed simulation results. For each combination of generative model parameters, ancestry, and trait we report **(Polygenicity):** trait polygenicity **(European h2):** Trait heritability in Europeans **(Cross-population rg):** Cross-population genetic correlation. **(Method):** The method name (see below). **(#Experiments):** Number of experiments performed. **(Ancestry):** Target ancestry. **(N):** Target ancestry sample size. **(Avg. h2):** Average h2 in the target ancestry. **(s.d. h2):** s.d. h2 in the target ancestry. **(Average R2):** Average R^2**. (R2 s.e.):** The standard errors of R^2. **(R2 (normalized vs. EUR)):** R^2 (normalized against the results of the same method in Europeans). **(R2 (normalized vs. EUR) s.e.):** Standard error of R^2 (normalized against the results of the same method in Europeans). **(R2 (normalized vs. BOLT-LMM-EUR)):** R^2 (normalized vs. BOLT-LMM in non-British Europeans). **(R2 (normalized vs. BOLT-LMM-EUR) s.e.):** The standard error of R^2 (normalized vs. BOLT-LMM in non-British Europeans). **(R2 (normalized vs. BOLT-LMM-EUR) P-value):** The p-value of the statistical test testing the null hypothesis that R^2 is the same as obtained by BOLT-LMM in non-British Europeans. **(R2 (normalized vs. BOLT-LMM)):** R^2 (normalized vs. BOLT-LMM in the target ancestry). **(R2 (normalized vs. BOLT-LMM) s.e.):** The standard error of R^2 (normalized vs. BOLT-LMM in the target ancestry). **(R2 (normalized vs. BOLT-LMM) P-value):** The p-value of the statistical test testing the null hypothesis that R^2 is the same as obtained by BOLT-LMM in the target ancestry. **(R2 diff vs. BOLT-LMM-EUR):** The difference between the R^2 obtained by the current method and BOLT-LMM in non-British Europeans. **(R2 diff vs. BOLT-LMM-EUR s.e.):** The standard error of the difference between the R^2 obtained by the current method and BOLT-LMM in non-British Europeans. **(R2 diff vs. BOLT-LMM-EUR P-value):** The p-value of the statistical test testing the null hypothesis that R^2 is the same as obtained by BOLT-LMM in non-British Europeans. **(R2 diff vs. BOLT-LMM):** The difference between the R^2 obtained by the current method and BOLT-LMM in the target ancestry. **(R2 diff vs. BOLT-LMM s.e.):** The standard error of the difference between the R^2 obtained by the current method and BOLT-LMM in the target ancestry. **(R2 diff vs. BOLT-LMM P-value):** The p-value of the statistical test testing the null hypothesis that R^2 is the same as obtained by BOLT-LMM in the target ancestry. The method names are as described in the main text, with the following rules. A suffix "-L1" indicates that PolyFun-pred was invoked assuming only one causal SNP per locus (and did not make use of LD information). A suffix "-N" indicates that the training sample size was reduced to the specified number. A suffix "-Nmix" indicates that the target-ancestry training sample size used for estimating mixing weights was modified from 500 to some other number. A suffix "-1000G" indicates that the LD reference panel used European data from the 1000 Genomes project. A suffix "-UK10K" indicates that the LD reference panel used UK10K. A suffix "-HM3" indicates that the SNP set consisted of only 1.2 million HapMap3 SNPs. A suffix "NoFun" indicates that the analysis did not use functional annotations to specify per-SNP prior causal probabilities.

Supplementary Table 2: Detailed simulation runtime analysis. For each combination of method and generative model parameters we report (**Method**): method name. **(#Experiments**): number of

experiments performed. (**Polygenicity**): trait polygenicity. (**European h2**): h2 in Europeans. (**Average run time (sec)**): average run time in seconds. (**SE (sec)**): the standard error of the runtime in seconds. (**Average run time (hr)**): average run time in hours. (**SE (hr)**): the standard error of the runtime in hours.

Supplementary Table 3: List of 49 diseases and complex traits. For each trait, we report its UK Biobank British sample size (used for training most methods), its UK Biobank British heritability estimate and its standard error (estimated using S-LDSC with the Baseline-LF v2.2.UKB model[1]), whether the trait was one of the 7 traits included in the meta-analysis, and whether the trait exists in Biobank Japan.

Supplementary Table 4: Detailed results of analyses using UKB British training individuals applied to other UKB populations, compared vs. BOLT-LMM. For each combination of method, ancestry and trait (including meta-analyzed traits) we report (**Method**): The method name (see below). AUC ROC (normalized vs BOLT-LMM-EUR) The standard error of R^2, computed via genomic block-jackknife; (**R2 (normalized vs. BOLT-LMM-EUR)**): The R^2 value, divided by the R^2 of BOLT-LMM in non-British Europeans; (**R2 (normalized vs. BOLT-LMM-EUR) s.e.**): The standard error of the normalized R^2; (**R2 diff vs. BOLT-LMM**): The difference between R^2 and the R^2 obtained by BOLT-LMM; (**R2 diff vs. BOLT-LMM s.e.**) The standard error of the difference between R^2 and the R^2 obtained by BOLT-LMM, computed via genomic block-jackknife; (**R2 diff vs. BOLT-LMM (normalized vs. BOLT-LMM-EUR)**): The difference between normalized R^2 and the normalized R^2 obtained by BOLT-LMM; (**R2 diff vs. BOLT-LMM (normalized vs. BOLT-LMM-EUR) s.e.**): The standard error of the difference between normalized R^2 and the normalized R^2 obtained by BOLT-LMM; (**R2 vs. BOLT-LMM (normalized vs. BOLT-LMM-EUR) P-value**): The P-value of the difference between the normalized R^2 and the normalized R^2 obtained by BOLT-LMM; (**AUC ROC**): The area under the receiving operating characteristic (defined only for disease traits)**; (AUC ROC s.e.):** The standard error of the area under the receiving operating characteristic (defined only for disease traits); (**AUC ROC (normalized vs BOLT-LMM-EUR))**: The area under the receiving operating characteristic, divided by the area under the receiving operating characteristic obtained via BOLT-LMM when applied to non-British Europeans; (**AUC ROC (normalized vs BOLT-LMM-EUR) s.e.**): The standard error of the normalized area under the receiving operating characteristic; (**Regression slope**): Slope obtained when regressing the true phenotype on the PRS; (**Regression slope s.e.**): The standard error of the regression slope, computed via genomic block-jackknife; (**R2 ind-s.e.**): The standard error of R^2, computed via jackknife over individuals; (**Mixing weights**): The mixing weights of combined methods (blank for non-combined methods). The first value is the intercept, and the other values are PolyPred, BOLT-LMM (resp. SBayesR or PRS-CS), and BOLT-LMM-BBJ (when there are four numbers) (resp. SBayesR-BBJ or PRS-CS-BBJ); (**P vs. Europeans**): The p-value of the hypothesis that the R2 obtained in the target ancestry is the same as in non-British Europeans (under the same method and trait), as computed via a conservative Wald test (Methods). The method names that are not explicitly defined in the main text are the following: **Methods ending with -pX** (where X is a number) are methods using a fixed mixing weight X for PolyPred and its extensions; **Methods ending with with -100** use 100 individuals from the target cohort to estimate mixing weights (instead of 500 as used by most combined methods); **BOLT-LMM-727K**: BOLT-LMM using only genotyped SNPs; **LDpred-1000G-p**: LDpred using the 1000 genomes project Europeans as an LD reference panel, and assuming that proportion p of causal SNPs are causal; **LDpred-1000G-cheat**: LDpred using the 1000 genomes Europeans as an LD reference panel, and using the best value of p for each trait (as determined via R^2 in the test set); **LDpred-UK10K-p**: LDpred using the UK10K cohort as an LD reference panel, and assuming that proportion p of causal SNPs are causal; **LDpred-UK10K-cheat**: LDpred using the UK10 cohort as an LD reference panel, and using the best value of p for each trait (as determined via R^2 in

the test set) (we caution that standard errors of methods using only PIP>0.95 SNPs may not be accurate because of the small number of SNPs used); **PRS-CS-phi0.0001**: PRS-CS with –phi=0.0001; **PRS-CS-phi0.01**: PRS-CS with –phi=0.01; **PRS-CS**: PRS-CS using a Biobank reference panel LD, and without specifying –phi; **PRS-CS-cheat**: PRS-CS that uses the best value of –phi for each target ancestry; **PRS-CS-1000G**: PRS-CS, using *N*=489 1000 Genomes project Europeans as an LD reference panel; **PolyFun-pred-pipP**: PolyFun-pred restricted to SNPs with PIP greater than P; **PolyFun-pred-NoFun**: PolyFun-pred without using functional annotations; **PRS-CS-BBJ**: PRS-CS, trained on Biobank Japan individuals (using a UK Biobank East-Asian LD reference panel); **P+T-pX**: P+T that uses only SNPs with BOLT-LMM P-value <X; **P+T-cheat**: P+T that uses the best value of X for each target ancestry. **PolyPred+-Ext**: PolyPred+ with mixing weights estimated in Biobank Japan; **PolyPred-pipP**: PolyPred restricted to SNPs with PIP greater than P; **PolyPred-NoFun**: PolyPred without using functional annotations; **SBayesR-2.8M**: SBayesR using 2.8M common SNPs selected by the SBayesR authors; **SBayesR-UK10K:** SBayesR, using UK10K (*N*=3000) as an LD reference panel; **SBayesR-1000G**: SBayesR, using the 1000 Genomes project Europeans (N=489) as an LD reference panel; SBayesR-UK10K-489: SBayesR, using a subset of UK10K individuals matched to the 1000 Genomes project Europeans sample size (N=489) as an LD reference panel; **PolyFun-pred-UK10K**: PolyFun-pred, using UK10K (*N*=3000) as an LD reference panel; **BOLT-LMM-N-African**: BOLT-LMM, evaluated by subsampling the test set of each population to the African sample size of the corresponding trait (unless the sample size was smaller than the African sample size). To see the numerical results of the analyses reported in the main text, please filter the **Trait** column to show only the trait 'Meta-Analysis'.

Supplementary Table 5: Comparisons between pairs of methods in analyses of real UK Biobank and Biobank Japan traits. The table reports comparisons of selected pairs of methods reported in the main text. For each pair of methods we report its training data (UKB indicates individual-level data or summary statistics from 337K UK Biobank British individuals; ENGAGE indicates summary statistics from the European Network for Genetic and Genomic Epidemiology; UKB+BBJ indicates a combination of UK Biobank and Biobank Japan training data); the names of the two methods (Method1 and Method2); the target ancestry; the trait name; the target ancestry sample size; the accuracy (R^2) of method1; the difference in R^2 between the two methods (Method1 R^2 – Method2 R^2) and its standard error; and the p-value of the difference, as computed via a genomic block jackknife over 200 genomic blocks.

Supplementary Table 6: Detailed results of analyses using UKB British training individuals applied to other UKB populations, compared vs. PolyPred. The table is analogous to Table 4, but all results are normalized and compared with respect to PolyPred instead of BOLT-LMM.

Supplementary Table 7: Ancestry-specific SNP heritability estimates in the UK Biobank, across 7 independent complex traits. For each trait (including meta-analyzed traits) we report its sample size (n), its SNP heritability estimate (h2g) and its standard error (se). All estimates were performed using GCTA[2] with HapMap 3[3] SNPs due to the relatively small sample sizes. Non-British Europeans were down-sampled to 10,000 individuals to facilitate the analysis. Meta-analyzed h2g was computed via the average h2g, and the meta-analyzed standard error was computed via the square root of the average sampling variance, divided by the square root of the number of traits.

Supplementary Table 8: Prediction accuracy using summary statistics from the from the European Network for Genetic and Genomic Epidemiology. The table is analogous to Supplementary Table 4, but reports results based on training data from the European Network for Genetic and Genomic Epidemiology

(ENGAGE) for four traits (BMI, waist-hip-ratio (adjusted for BMI), total cholesterol, and triglycerides) (average N=61K) instead of training data based on up to N=337K UK Biobank British individuals.

Supplementary Table 9: Detailed results of analyses applied to Biobank Japan and to Uganda-APCDR. The table is analogous to Table 4, but includes columns comparing each method to PolyPred in addition to columns comparing each method to BOLT-LMM.

Supplementary Table 10: Comparing prediction accuracy in UK Biobank Non-British Europeans and in Biobank Japan when using equal training set sample sizes. For each of 7 independent traits we report (**N**) its Biobank Japan training sample size (which was also used for the UK Biobank British training sample size in this analysis); (**h2g (UKB-EUR)**) its non-British European SNP heritability, as estimated by BOLT-REML; (**h2g (BBJ)**) its Biobank Japan SNP heritability, as estimated by BOLT-REML; (**R2-expected (UKB EUR)**) the expected R^2 in non-British Europeans as a function of training set sample size and SNP heritability, based on theory (see Supplementary Note); (**R2-expected (Biobank Japan)**) the expected R^2 in Biobank Japan as a function of training set sample size and SNP heritability, based on theory; (**R2 (UKB-EUR)**) the R^2 obtained in practice in non-British Europeans when training BOLT-LMM using a UK Biobank British training sample with the same sample size as the Biobank Japan training sample size; (**R2 (BBJ)**) the R^2 obtained in practice in 5K Biobank Japan individuals when training BOLT-LMM using a Biobank Japan training sample.

Supplementary Table 11: Description of 187 baseline-LF model annotations used by PolyFun-pred. For each annotation we report #SNPs in the annotation (unless it is a continuous-valued annotation), #common (MAF>0.05) SNPs in the annotation, whether it is binary or continuous-valued, and a literature reference.

Supplementary Note

Secondary analyses for simulations with in-sample LD

We performed 5 secondary analyses to investigate the sensitivity of the results to the simulation parameters. First, we performed simulations for much less polygenic (0.05%) and much more polygenic (0.5%) architectures. PolyPred remained the most accurate method, attaining the largest relative improvements vs. BOLT-LMM for the much less polygenic architecture, with slightly worse results for PolyPred-S and PolyPred-P (Supplementary Table 1); we conservatively restricted the remaining secondary analyses to the more polygenic (0.3%) architecture (for which PolyPred attains smaller relative improvements among the two main architectures simulated) and omitted PolyPred-S and PolyPred-P (due to their close similarity to PolyPred), unless otherwise indicated. Second, we performed simulations with lower (3%) or higher (7%) chromosome 22 heritability. PolyPred remained the most accurate method, with relative improvements vs. BOLT-LMM increasing with heritability (Supplementary Table 1). Third, we performed simulations with cross-population genetic correlations increased from 0.8 to 1.0. PolyPred remained the most accurate method, with relative improvements vs. BOLT-LMM remaining broadly similar (Supplementary Table 1). Fourth, we modified the number of training samples from the target population used to estimate mixing weights (N_{mix}) from 500 to various values from 100-1000. PolyPred remained the most accurate method in all these experiments, with relative improvements vs. BOLT-LMM increasing with N_{mix} but limited improvement above N_{mix}=500 (Supplementary Table 1). Fifth, we

decreased the number of British-ancestry training samples (*N*) from *N*=337K to *N*=100K or *N*=10K. Prediction accuracies decreased with decreasing training sample size for all methods, and the relative improvements of PolyPred vs. BOLT-LMM (and other methods) were substantially decreased for *N*=10K, though they remained statistically significant in Africans under 0.1% polygenicity (Supplementary Table 1).

We performed two secondary analyses to investigate the sensitivity of the results to the SNP set and functional annotations. First, we evaluated a modified version of PolyPred that uses only 1.2 million HapMap 3 SNPs (matching the SNP sets of BOLT-LMM, SBayesR, and PRS-CS) instead of 18 million SNPs. PolyPred suffered a substantial loss of accuracy in this setting, demonstrating the importance of using a dense SNP set for fine-mapping based PRS (Supplementary Table 1). Second, we evaluated a non-functionally informed method (PolyPred-NoFun) that linearly combines PolyNoFun-pred (a modification of PolyFun-pred that is not functionally-informed; see Methods) and BOLT-LMM, precluding the need for functional annotations. PolyPred-NoFun was slightly less accurate than PolyPred, but still more accurate than BOLT-LMM (Supplementary Table 1).

We performed two secondary analyses to evaluate the computational cost and memory cost of each method. First, we evaluated the computational cost of each method (for PolyPred, PolyPred-S, and PolyPred-P, we included the time cost of each constituent method); we focused on the time cost to compute SNP effect sizes used for prediction, as the time cost to compute predictions in target samples using these SNP effect sizes is approximately the same for each method. SBayesR was the fastest method (2.8 minutes), P+T was the second fastest method (7.4 minutes), PRS-CS was the third fastest method (113 minutes), BOLT-LMM was the fourth fastest method (224 minutes), PolyPred-S was the fifth fastest method (447 minutes), PolyPred-P was sixth fastest method (557 minutes), and PolyPred was the slowest method (668 minutes) (Supplementary Table 2). Second, we evaluated the memory cost of each method (for PolyPred, we computed the maximum memory cost of each constituent method). We performed this analysis using chromosome 1 instead of chromosome 22 because memory cost can increase with the number of SNPs in the analysis (but the memory cost of PolyFun-pred is fixed because it analyzes each 3Mb-locus separately). P+T used the least memory (1.5GB), PRS-CS used the second smallest amount of memory (1.8GB), SBayesR used the third smallest amount of memory (2.6GB), BOLT-LMM used the fourth smallest amount of memory (11GB), and PolyPred, PolyPred-S, and PolyPred-P all used the most memory (57GB) (Supplementary Table 2). The larger computational cost of PolyPred and its summary statistic-based analogues is dominated by the PolyFun-pred component, which is computationally intensive because (i) it performs fine-mapping and (ii) it analyses a large number of SNPs (see the Supplementary Note subsection Limitations of PolyPred and PolyPred+).

Simulations with reference LD

The simulations described in the main text use in-sample LD (i.e., LD summary data based on the UK Biobank GWAS sample). However, researchers often do not have access to in-sample LD, necessitating external LD reference panels. We thus evaluated modified versions of PolyFun-pred, SBayesR and PRS-CS that use summary LD estimated from 1000 Genomes project Europeans[4] (*N*=489). We note that this LD reference panel is both smaller than the UK Biobank British LD reference panel (*N*=337K) and less well-matched to the GWAS sample, because it consists of pan-European ancestries rather than only British-ancestry individuals. We excluded BOLT-LMM from these analyses because it requires individual-level data.

The results of simulations with reference LD are reported in Supplementary Table 1. All methods became less accurate when using 1000 Genomes project Europeans LD summary data. The loss of accuracy was modest for SBayesR (-5% R^2 for non-British Europeans vs. using in-sample LD) but severe for PRS-CS (-42% R^2 for non-British Europeans vs. using in-sample LD) and PolyFun-pred (-90% for non-British Europeans vs. using in-sample LD). We caution that the differences observed in real trait analysis for SBayesR and PRS-CS were substantially different from those observed in our simulations (large loss of accuracy for SBayesR, no significant loss of accuracy for PRS-CS), suggesting that the effect of LD mismatch on PRS accuracy may be sensitive to the underlying genetic architecture.

We performed 3 secondary analyses. First, we evaluated a modified version of PolyFun-pred that uses summary LD from UK10K[5] (*N*=3,567). We observed only a moderate loss of accuracy in PolyFun-pred vs. using in-sample LD (-8% R^2 in non-British Europeans) (Supplementary Table 1). However, we caution that using UK10K led to substantial and statistically significant loss of accuracy in real trait analysis, suggesting that the results may be sensitive to the underlying genetic architecture. Second, we evaluated modified versions of PolyFun-pred using subsets of UK10K as an LD reference panel, ranging from *N*=3,000 to *N*=489 (matching the 1000 Genomes project Europeans reference LD sample size). The accuracy of PolyFun-pred decreased with the LD reference panel sample size, with the loss in accuracy vs. using in-sample LD (for non-British Europeans) ranging from -8% for *N*=3,000, to -90% for *N*=489 (Supplementary Table 1). Finally, we evaluated a modified version of PolyFun-pred (PolyFun-pred1) that assumes a single causal variant per locus, precluding the need for a reference LD panel (because fine-mapping under a single casual variant assumption does not require any LD information[1]). PolyFun-pred1 was substantially less accurate than all other methods (including P+T) and is thus not recommended for polygenic prediction (Supplementary Table 1).

We conclude that the accuracy of all methods increases with the size of the LD reference panel and its concordance with the GWAS sample population, but that the relationship may depend on the underlying genetic architecture. Hence, it may be best to assess the accuracy obtained under various LD reference panels using real trait analysis rather than simulations. Specifically, the simulation results do not support the use of PolyPred-S or PolyPred-P in the specific scenarios considered in these simulations. However, real data results with very large LD reference panels do support the use of PolyPred-S or PolyPred-P (Extended Data Figure 1). We did not perform simulations with very large unmatched LD (analogous to Extended Data Figure 1), as this would have required another very large individual-level data set in addition to UK Biobank.

Evaluating method calibration for PRS in 4 UK Biobank populations using British training data

We assessed the calibration of each prediction method. A predictor is correctly calibrated if a regression of the true phenotype vs. the predictor yields a slope of 1, and is miscalibrated otherwise[6]. Regression slopes are reported in Supplementary Table 4. In non-British Europeans, PolyPred was well-calibrated (regression slope = 1.01), BOLT-LMM and SBayesR were approximately well-calibrated (0.96-1.08), PRS-CS was slightly miscalibrated (1.26), and P+T was poorly calibrated (0.08). In non-European populations, PRS-CS was approximately well-calibrated (0.85-1.11), but BOLT-LMM and SBayesR suffered reduced

regression slopes (0.57-0.90), consistent with reduced prediction accuracy. In contrast, PolyPred and its summary statistic-based analogues remained well-calibrated (0.95-1.17), as expected due to their extra training step to estimate mixing weights in the target population.

Secondary analyses for PRS in 4 UK Biobank populations using British training data

We performed 5 secondary analyses to evaluate the impact of the LD reference panel and the SNP set on prediction accuracy (we note that analyses of summary statistics from a meta-analysis of many cohorts generally require using an LD reference panel instead of in-sample LD). First, we evaluated a modified version of PolyFun-pred using a reference panel based on UK10K (*N*=3,567) and observed a substantial and statistically significant reduction in accuracy, to a far greater degree that observed in simulations (Supplementary Tables 4-6). Second, we evaluated a modified version of PRS-CS that uses an LD reference panel from 1000 Genomes project Europeans (*N*=489) and observed statistically indistinguishable results from those obtained using in-sample LD (unlike in simulations, where we observed significantly reduced accuracy when using an LD reference panel from 1000 Genomes project Europeans) (Supplementary Tables 4-6). Third, we evaluated modified versions of SBayesR that use (i) an LD reference panel using UK10K (*N*=3,567); (ii) an LD reference panel using 1000 Genomes project Europeans (*N*=489); or (iii) an LD reference panel using a subset of UK10K (*N*=489). We observed (i) very similar and statistically indistinguishable accuracy when using UK10K, (ii) severely reduced accuracy ($P<4\times10^{-6}$) when using 1000 Genomes project Europeans, and (iii) moderately reduced accuracy (P=0.07 in East-Asians, $P<7\times10^{-6}$ in other target populations) when using a subset of UK10K, suggesting that the loss of accuracy primarily stems from LD mismatch rather than reduced sample size (Supplementary Tables 4-6). Fourth, we evaluated a modified version of SBayesR (SBayesR-2.8M) that uses 2.8M common SNPs specified by the authors of SBayesR[7] instead of 1.2 million HapMap 3 SNPs. SBayesR-2.8M was less accurate than SBayesR (significantly so for Africans) (Supplementary Tables 4-6). Thus, our use of SBayesR (using 1.2 million HapMap 3 SNPs) instead of SBayesR-2.8M in all primary comparisons is a conservative choice, since SBayesR outperforms SBayesR-2.8M (we note that naively scaling SBayesR and PRS-CS to use 18 million SNPs as in PolyFun-pred would be computationally infeasible[7,8]). Fifth, we evaluated a modified version of BOLT-LMM (BOLT-LMM-727K) that estimates effect sizes using only 727K genotyped SNPs (instead of 1.2 million imputed HapMap 3 SNPs). BOLT-LMM-727K was substantially and significantly less accurate than BOLT-LMM (Supplementary Table 4).

We performed 9 additional secondary analyses. First, we evaluated LDpred[6] using 1000 Genomes project Europeans[4] or UK10K[5] as the LD reference panel (Methods). Both versions of LDpred were consistently less accurate than BOLT-LMM (Supplementary Table 4). Second, we evaluated modified versions of PolyPred that specify fixed mixing weights instead of estimating mixing weights in the target populations. We considered mixing weights for PolyFun-pred/BOLT-LMM equal to 0%/100%, 25%/75%, 50%/50%, 75%/25%, and 100%/0%. The 25%/75% and 50%/50% methods performed very similarly to PolyPred, with no statistically significant differences (Supplementary Table 6). Third, we restricted the PolyFun-pred component of PolyPred to only include SNPs with posterior causal probability greater than a fixed threshold (0.05, 0.50 or 0.95). This restriction decreased prediction accuracy (Supplementary Table 4,6), implying that estimating causal effect sizes is beneficial for prediction even at loci that cannot be confidently fine-mapped. Fourth, we evaluated a non-functionally informed method (PolyPred-NoFun) that linearly combines PolyNoFun-pred (a modification of PolyFun-pred that is not functionally-informed; see Methods) and BOLT-LMM. PolyPred-NoFun was slightly less accurate than PolyPred, but still more

accurate than BOLT-LMM (Supplementary Tables 4,6). The difference between PolyPred-NoFun vs. PolyPred was not statistically significant, in contrast to previous studies reporting a large and statistically significant increase in prediction accuracy from incorporating functional annotations[9–11]. Fifth, we reduced the number of training samples from the target population used to estimate mixing weights (N_{mix}) from 500 to 100. PolyPred suffered slightly reduced accuracy but remained the most accurate method, although relative improvements vs. BOLT-LMM were no longer statistically significant due to larger standard errors (Supplementary Table 4). Sixth, we computed standard errors of relative-R^2 using a jackknife over individuals[9] (instead of a genomic block-jackknife over SNPs; see Methods). Standard errors computed using a jackknife over individuals were generally smaller, increasing the statistical significance of relative improvements of PolyPred vs. BOLT-LMM (Supplementary Table 4). Seventh, we observed very similar results when down-sampling the non-British European target sample size to match the African target sample size, demonstrating that the reduced accuracy in Africans vs. Europeans is not due to the lower target sample size (Supplementary Table 4).Eighth, we evaluated two versions of PRS-CS that use pre-specified values of its global shrinkage parameter (0.01 and 0.001, following the recommendations of the authors of PRS-CS[8]). Both versions were less accurate than the default version of PRS-CS (which automatically adjusts the value of this parameter), justifying the use of the default version of PRS-CS in this work (Supplementary Tables 4-5). Finally, we assessed the potential contribution of ancestry-specific heritability to reductions in cross-population prediction accuracy[12], by applying GCTA[2] to estimate the SNP-heritability explained by HapMap 3 SNPs[3,13] in each target population. SNP-heritabilities were largest in non-British Europeans and smallest in Africans (Supplementary Table 7) (these differences could be due to SNP ascertainment[14], sample ascertainment, and/or ancestry-specific architectures[15]), likely contributing to reductions in cross-population prediction accuracy.

Secondary analyses for PRS in Biobank Japan and Uganda-APCDR cohorts

We performed 6 secondary analyses. First, we assessed the calibration of each method by computing regression slopes, which are reported in Supplementary Table 9. Similar to our analyses of non-European UK Biobank target populations, PolyPred and its summary statistic-based analogues were the only approximately well-calibrated methods, as expected due to their extra training step to estimate mixing weights in the target population. We restricted the remaining secondary analyses to PolyPred (as PolyPred-S and PolyPred-P are analogous to PolyPred with respect to these analyses). Second, we evaluated a modification of PolyPred that estimates mixing weights using 500 UK Biobank individuals from the genetically closest target population (UK Biobank East Asians for Biobank Japan, UK Biobank Africans for Uganda-APCDR) instead of 500 individuals from the target cohort. The differences between the original and modified versions of PolyPred were small and not statistically significant (Supplementary Table 9), indicating that PolyPred mixing weights can be estimated using 500 individuals from any cohort with the same continental ancestry as the target population. Third, we evaluated modified versions of PolyPred that specify fixed mixing weights instead of estimating mixing weights in the target populations. We considered mixing weights for PolyFun-pred/BOLT-LMM equal to 0%/100%, 25%/75%, 50%/50%, 75%/25%, and 100%/0%. The 25%/75% and 50%/50% methods performed very similarly to PolyPred, with no statistically significant differences (Supplementary Table 9). Fourth, we reduced the number of training samples from the target population used to estimate mixing weights (N_{mix}) from 500 to 100. PolyPred suffered slightly reduced accuracy but remained the most accurate method, with the improvement vs. BOLT-LMM in Biobank Japan remaining statistically significant (Supplementary Table 9). Fifth, we

computed standard errors of relative-R^2 using a jackknife over individuals[9] (instead of a genomic block-jackknife over SNPs). We obtained standard errors that were almost identical to those obtained using a genomic block-jackknife (unlike the above results for UK Biobank), suggesting that Biobank Japan may be more heterogeneous across samples, possibly due to its hospital-based recruitment (Supplementary Table 9). Finally, we meta-analyzed the results of each method across three independent diseases in Biobank Japan: type 2 diabetes, asthma, and all autoimmune disease. Similar to our UK Biobank analyses above, PolyPred attained the highest prediction accuracy in each disease, though relative improvements were not statistically significant due to lower power (Supplementary Table 9).

Secondary analyses for PRS in East Asians using British and Japanese training data

We performed 6 secondary analyses. We restricted these secondary analyses to PolyPred+ (as PolyPred-S+ and PolyPred-P+ are analogous to PolyPred+ with respect to these analyses). First, we verified that PolyPred+ using European and East Asian training data does not outperform PolyPred in UK Biobank populations other than East Asians; differences between PolyPred+ and PolyPred were very small and not statistically significant (Supplementary Table 6). Second, we verified that PolyPred+ was well-calibrated (Supplementary Table 4; results for other methods are described above), as expected due to its extra training step to estimate mixing weights in the target population. Third, we evaluated a modified version of PolyPred+ that estimates mixing weights using 500 Biobank Japan individuals instead of 500 UK Biobank East Asians. The modified version of PolyPred+ was far less accurate than the original version (52% lower relative-R^2; Supplementary Table 6). The mixing weights estimated in Biobank Japan assign much higher weight to the Biobank Japan training data (Supplementary Table 6), perhaps due to cohort effects; thus, it may be important to estimate PolyPred+ mixing weights using the target cohort (as opposed to the training cohort) if cohort effects are present. Fourth, we reduced the number of training samples from the target population used to estimate mixing weights (N_{mix}) from 500 to 100. PolyPred+ suffered slightly reduced accuracy, though the difference was not statistically significant (Supplementary Table 6). Fifth, we evaluated a prediction method using only the N=124K Biobank Japan individuals to train effect sizes (BOLT-LMM-BBJ). BOLT-LMM-BBJ substantially underperformed methods that use UK Biobank British training data (−27% vs. BOLT-LMM, −34% vs. PolyPred, −41% vs. PolyPred+; Supplementary Table 4). Finally, we computed standard errors of relative-R^2 using a jackknife over individuals[9] (instead of a genomic block-jackknife over SNPs). Standard errors computed using a jackknife over individuals were smaller, increasing the statistical significance of relative improvements of PolyPred+ vs. other methods (Supplementary Table 6).

Loss of PRS accuracy under an infinite European training sample

Under an infinite European training sample, the ratio between R^2_{Eur} and $R^2_{\text{non-Eur}}$, which denote R^2 in a European sample and in a non-European sample, respectively, is approximately given by:

$$\rho_g^2 \times \frac{h^2_{\text{non-Eur}}}{h^2_{\text{Eur}}} \times \left(\sum_k \sqrt{\frac{p_{\text{k,non-EUR}}(1-p_{\text{k,non-EUR}})}{p_{\text{k,EUR}}(1-p_{\text{k,EUR}})}} \right)^2 \times \frac{\text{var}(\text{PGS}_{\text{EUR}})}{\text{var}(\text{PGS}_{\text{non-EUR}})}.$$

Here, ρ_g is the cross-population genetic correlation, $h^2_{\text{non-Eur}}$, h^2_{Eur} are the heritabilities in the non-European and the European populations, respectively, k iterates over causal SNPs, $p_{\text{k,non-EUR}}$, $p_{\text{k,EUR}}$ are

minor allele frequencies in the non-European and the European population, respectively, and $\mathrm{var}(\mathrm{PGS}_{\mathrm{EUR}}), \mathrm{var}(\mathrm{PGS}_{\mathrm{non-EUR}})$ are the variances of the polygenic risk scores in the non-European and the European populations, respectively. This equation is directly derived from Equation 1 in ref.[12], after assuming that causal SNPs are approximately not in LD with each other, and that the predictor SNPs are the causal SNPs under an infinite sample size.

Limitations of PolyPred and PolyPred+

Polypred training time is slower than alternative PRS methods

PolyPred and its summary statistic-based analogues are slower than alternative PRS methods, requiring over 1,000 hours of computation time for training, vs. less than 100 hours for BOLT-LMM (Supplementary Note). This is dominated by the PolyFun-pred component, which is computationally intensive because (i) PolyFun-pred performs fine-mapping, which is a more computationally intensive task than other approaches to computing PRS coefficients (e.g. computing posterior mean tagging effect sizes, as in SBayesR); and (ii) PolyFun-pred analyzes a large number of SNPs, e.g. 18 million SNPs in UK Biobank training data and 8.1 million SNPs in ENGAGE training data (vs. 1.2 million SNPs for SBayesR). We do not foresee the larger computation time for training as a major limitation in real-world settings, because training only needs to be performed once, can be parallelized, and provides genome-wide fine-mapping results of direct interest[1].

PolyPred cannot use data from a fixed-effects meta-analysis of GWAS data from different ancestry groups

One of the main conclusions of our work is that leveraging training data from different ancestry groups (e.g. different continental ancestries) improves PRS in diverse populations. However, we recommend against using training data consisting of a traditional fixed-effect meta-analysis of GWAS data from different ancestry groups, for two reasons: (i) fixed-effect meta-analysis implies that European training samples and training samples from the non-European target population would receive equal weight, whereas our work shows that the latter should receive higher weight in order to maximize PRS accuracy; and (ii) it may be challenging to construct an LD reference panel whose ancestry matches the ancestry of the meta-analysis of different ancestry groups. When possible, it would be preferable to separately incorporate European training data and training data from the non-European target population, with appropriate LD reference panels. Although there is no single optimal way to choose a training cohort, training sample size should be a primary consideration, as it is a critical factor impacting PRS accuracy.

PolyPred requires a small training sample from the target cohort to maintain calibrated predictions

PolyPred ideally requires a small training sample from the target cohort to estimate mixing weights. Our results suggest that it is possible to improve cross-population PRS accuracy even without such a training sample, by linearly combining PolyFun-pred and BOLT using mixing weights of either 25%/75% or 50%/50%, respectively. However, we caution that PRS linearly combined using fixed mixing weights may not always be well-calibrated.

Causal vs. tagging effects

We consider a linear model $y = \sum_i x_i \beta_i + \epsilon$ where y is a trait, x_i is the number of minor alleles at SNP i, β_i is the (true) causal effect size of SNP i, and ϵ is a residual term sampled from a normal distribution. We consider a method (such as PolyFun-pred) that estimates β_i. If the generative model holds and all SNPs i are considered in the estimation procedure, then any consistent estimator $\hat{\beta}_i$ of β_i represents a causal effect. In contrast, if only a subset of the SNPs, such as HapMap3 SNPs, are considered in the estimation procedure (i.e. if we incorrectly assume the generative model $y = \sum_{i \in S} x_i \beta_i + \epsilon$, where S is a subset of SNPs) then the estimated value $\hat{\beta}_i$ represents a linear combination of β_i and of the effect sizes of other SNPs.

The exact value estimated by $\hat{\beta}_i$ depends on the estimation procedure. For example, assuming an ordinary least squares estimator for simplicity, the vector $\widehat{\boldsymbol{\beta}}_S$ of estimated coefficients is a consistent estimator of $\left[I_{m-k}\ R_{S\bar{S}}^{-1} R_{S\bar{S}}\right]\boldsymbol{\beta}$, where m is the total number of SNPs, k is the number of SNPs in the set S, R_{SS} is the LD matrix of the SNPs in the set S, $R_{S\bar{S}}$ is a matrix wherein each entry i, j is the correlation between SNP i in the set S and SNP j in the set of SNPs that are not in S, and $\boldsymbol{\beta}$ is the vector of true effect sizes, assuming without loss of generality that the set S includes the first k SNPs (out of m SNPs considered). It is easy to derive this quantity by writing down the conditional expectation of $\widehat{\boldsymbol{\beta}}_S$ under an ordinary least squares estimator, given by $\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_S \mid \boldsymbol{\beta}\right] = \mathrm{E}[(\boldsymbol{X}_S^T \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^T \boldsymbol{y} \mid \boldsymbol{\beta}]$, where $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a vector of observed phenotypes and $\boldsymbol{X}$ is the corresponding matrix of SNPs, $\boldsymbol{X}_S$ is the submatrix of $\mathbf{X}$ consisting of columns of SNPs in the set S, and we assume that ϵ is independent of $\boldsymbol{X}$.

Investigating if off-cohort loss of accuracy is driven by SNP heritability differences

We investigated if lower prediction accuracies in Biobank Japan vs. the UK Biobank can be largely explained by SNP heritability differences. We began by comparing trait heritabilities across the UK Biobank and Biobank Japan, using BOLT-REML[16] applied to UK Biobank British-ancestry individuals (average *N*=325K) and to Biobank Japan (average *N*=124K), restricting to HapMap 3 SNPs. The average heritability in the UK Biobank was 67% larger (Supplementary Table 10), indicating differences in either trait measurement, cohort ascertainment, the ability of HapMap 3 SNPs to tag East Asian causal SNPs[14], or in the true underlying heritabilities (we could not perform a similar analysis with UK Biobank East Asian individuals due to small sample sizes leading to large standard errors). We next asked if the observed differences in PRS accuracy between Biobank Japan and the UK Biobank can be explained by the 67% increased average SNP heritability in the UK Biobank. To this end, we computed the expected R^2 within each cohort as function of SNP heritability, sample size, and the effective number of independent SNPs[17,18]:

$$\mathrm{E}[R^2] = h^2 \frac{h^2}{h^2 + \frac{m}{n}}.$$

Here, h^2 is SNP heritability, n is sample size, and m is the effective number of independent SNPs (which we specified as 55,000, determined by dividing the number of HapMap 3 SNPs by their average within-HapMap 3 LD-score). We used the smaller Biobank Japan sample size in both cohorts to eliminate differences due to sample size differences (by choosing a random subset of UK Biobank British individuals as a training set). The average expected R^2 in the UK Biobank was 104% larger than in Biobank Japan (Supplementary Table 10). We then trained BOLT-LMM using subsets of the UK Biobank British sample (matching the Biobank Japan sample size for each trait) and applied the predictions to UK Biobank non-British Europeans. The average R^2 in UK Biobank non-British Europeans (when training BOLT-LMM using the reduced British training sample) was 108% larger than the average R^2 in Biobank Japan (when training BOLT-LMM using the Biobank Japan training sample) (Supplementary Table 10), strongly consistent with

the 104% increase expected from theory. Finally, we determined that when training BOLT-LMM using the full UK Biobank British training set (average *N*=325K), the average R^2 in UK Biobank East Asians across the 7 independent traits is 93% larger than in Biobank Japan (Supplementary Tables 4 and 9), broadly consistent with the previous results. Assuming that the main factor differentiating the UK Biobank East Asian sample from the Biobank Japan sample is SNP heritability differences (rather than differences in MAF, LD, or causal effect sizes), these findings suggest that the main factor leading to lower prediction accuracies in Biobank Japan vs. the UK Biobank is SNP heritability differences.

To further investigate if off-cohort loss of accuracy is driven by SNP heritability differences, we compared prediction accuracies in UK Biobank East Asians and in Biobank Japan, when training BOLT-LMM using the Biobank Japan training sample. The average relative-R^2 in UK Biobank East Asians across the 7 independent traits was 9.0% larger (Supplementary Tables 4,10), though the difference was not statistically significant (P=0.18), possibly owing to the small UK Biobank East Asian sample size.

Although these results are not conclusive, they suggest that heritability differences drive most of the differences in prediction accuracies observed between the UK Biobank and Biobank Japan. Surprisingly, these results are consistent with a model in which HapMap 3 SNPs in Biobank Japan tag approximately 50% of the causal SNPs that they tag in the UK Biobank, rather than a model in which SNP heritabilities in Biobank Japan are smaller due to smaller causal effect sizes. This is because under the second model, we would expect to see large increase in prediction accuracy in UK Biobank East Asians vs. Biobank Japan when training BOLT-LMM using Biobank Japan (compared with only a 9.0% increase observed in practice). A partial explanation is that the HapMap 3 SNP set consists of a combination of two genotyping chips, one of which is explicitly designed to optimize tagging in Europeans[19].

Overall, these results suggest that differences in SNP-heritability due to ancestry differences (e.g. SNP ascertainment[14], sample ascertainment, and/or ancestry-specific architectures[15]) or due to cohort differences (e.g. differences in phenotype definitions[20], different recruiting strategies[20], or assay artifacts) may explain most of the differences in prediction accuracies observed between the UK Biobank and Biobank Japan. Our results are consistent with recent results showing almost no loss of accuracy when applying PRS based on UK Biobank training data to other European-ancestry cohorts[7]. Importantly, our results suggest that factors that inflate within-cohort PRS accuracy[21] (such as cohort-specific GxE, cohort-specific indirect effects[22], cohort-specific population structure, or cohort-specific assortative mating) are unlikely explanations for the observed accuracy differences between the UK Biobank and Biobank Japan.

Decomposing the PolyFun-pred and BOLT-LMM predictors into shared and non-shared components

A linear combination of PRS predictors is not necessarily suboptimal, even if the methods are correlated. (As an extreme example, a linear combination of two perfectly correlated predictors is optimal.) However, a linear combination could be suboptimal if the correlation between the (effect sizes underlying the) two predictors varies across the genome. As an extreme example, consider a scenario where one predictor is perfectly accurate across the first half of a chromosome but uninformative across the second half, whereas the second predictor is uninformative across the first half but perfectly accurate across the second half. Clearly, the optimal combination would use only the (effect sizes of the) first predictor for

the first half of the chromosome, and only the (effect sizes of the) second predictor for the second half of the chromosome. However, a simple linear combination assigns only a single mixing weight to each predictor, and will thus assign equal weights to both predictors, resulting in a suboptimal predictor.

We performed several attempts to improve upon a simple linear combination of PRS predictors by partitioning the genome into segments and estimating different linear mixing weights in different segments. However, this more complex approach did not outperform the simple approach of assigning a simple mixing weight to each predictor (results not shown), and we thus did not pursue it further.

Generating data for UK Biobank simulations

To simulate data, we first computed the variance of per-standardized-genotype effect η_i for every SNP i with annotations $\boldsymbol{a}_i$ using the baseline-LF (version 2.2.UKB) model, $\mathrm{var}[\eta_i|\boldsymbol{a}_i] = \sum_c \tau^c a_i^c$, where c are annotations and τ^c estimates are taken from a fixed-effects meta-analysis of 16 well-powered genetically uncorrelated ($|r_g|$<0.2) UK Biobank traits (age of menarche, BMI, balding, bone mineral density, eosinophil count, FEV1/FVC ratio, forced vital capacity, hair color, height, platelet count, red blood cell distribution width, red blood cell count, systolic blood pressure, tanning, waist-hip ratio adjusted for BMI, white blood count), scaled such that $\sum_i \mathrm{var}[\eta_i|\boldsymbol{a}_i]$ is the same across all traits (as detailed in ref.[1]). Each SNP was specified to be causal with probability proportional to $\mathrm{var}[\eta_i|\boldsymbol{a}_i]$, such that the average causal probability was equal to the desired proportion of causal SNPs (0.1% or 0.3% in most simulations).

We generated ancestry-specific effect sizes as follows. First, we generated a British per-allele causal effect size for each SNP i via $\beta_i^{\mathrm{British}} = \gamma_i / \sqrt{2f_i(1-f_i)}$, where $\gamma_i \sim \mathcal{N}(0, h^2/m)$, m is the number of causal SNPs, and f_i is the maximal MAF of SNP i among British, non-British European, South Asian, East Asian, or African UK Biobank individuals. Afterwards, for each of the main UK Biobank non-European ancestries (South Asian, East Asian, and African) a we generated an ancestry-specific per-allele effect size β_i^a via $\beta_i^a = r_g \cdot \beta_i^{\mathrm{British}} + \sqrt{1-r_g^2}\, z_i^a$, where r_g is the cross-population genetic correlation (set to 0.8 by default, following previous works[15,23,24]), and $z_i^a \sim \mathcal{N}(0,1)$. The use of f_i bounds the per-allele causal effect sizes by the MAF of the ancestry in which the SNP is most common, which guarantees that SNPs that are infrequent in Europeans but are common in other ancestries do not explain a very large proportion of heritability.

After generating ancestry-specific per-allele causal effect sizes, we generated a phenotype y for every UK Biobank individual in each ancestry a via $y = \sum_i x_i \beta_i^a + \epsilon$, where x_i is the number of minor alleles of SNP i carried by that individuals, β_i^a is the ancestry-specific per-allele causal effect size of SNP i, and $\epsilon \sim \mathcal{N}(0, 1-h^2)$ is the environmental variance of the generated trait. We generated phenotypes based on dosage data from imputed genotypes, using Plink 2.0[25,26]. We used self-reported ancestry based on UK Biobank data field 21000 (Ethnic background). We considered Irish-ancestry as a non-British European ancestry.

References

1. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* 1–9 (2020).
2. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
3. HapMap3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).
4. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
5. The UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82 (2015).
6. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
7. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 1–11 (2019).
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**, 1–10 (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. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.* **13**, e1005589 (2017).
11. Marquez-Luna, C. *et al.* LDpred-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv* 375337 (2020).

12. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).

13. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

14. Bhangale, T. R., Rieder, M. J. & Nickerson, D. A. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat. Genet.* **40**, 841–843 (2008).

15. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1–15 (2021).

16. 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).

17. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS One* **3**, e3395 (2008).

18. Visscher, P. M. & Hill, W. G. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet* **5**, e1000628 (2009).

19. Duan, S., Zhang, W., Cox, N. J. & Dolan, M. E. FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformation* **3**, 139 (2008).

20. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584 (2019).

21. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).

22. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).

23. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

24. Galinsky, K. J. *et al.* Estimating cross-population genetic correlations of causal effect sizes. *Genet. Epidemiol.* **43**, 180–188 (2019).

25. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

26. Purcell, S & Chang, C. *PLINK v2.00a3LM*.