

1 Leveraging genetic ancestry continuum information to 2 interpolate PRS for admixed populations

3

4 Yunfeng Ruan¹, Rohan Bhukar^{1,2}, Aniruddh Patel^{1,2,3}, Satoshi Koyama^{1,2,3,4}, Leland
5 Hull^{5,6}, Buu Truong^{1,2,3,8}, Whitney Hornsby^{1,2}, Haoyu Zhang⁷, Nilanjan Chatterjee^{9,10},
6 Pradeep Natarajan^{1,2,3,6}

7

8 Affiliations:

9

1. Program in Medical and Population, Genetics and Cardiovascular Disease

10

Initiative, Broad Institute of Harvard and MIT, Cambridge, MA, USA

11

2. Cardiovascular Research Center, Massachusetts General Hospital, Boston,

12

MA, USA

13

3. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA,

14

USA, USA

15

4. Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for

16

Integrative Medical Sciences, Yokohama, Japan

17

5. Division of General Internal Medicine, Massachusetts General Hospital,

18

Boston, MA, USA

19

6. Department of Medicine, Harvard Medical School, Boston, MA, USA

20

7. Division of Cancer Epidemiology and Genetics, National Cancer Institute,

21

Bethesda, MD, USA

22

8. Department of Genetic Epidemiology and Statistical Genetics, Harvard

23

T.H. School of Public Health, Cambridge, MA, US

24

9. Department of Biostatistics, Bloomberg School of Public Health, Johns

25

Hopkins University, Baltimore, MD, USA

26

10. Department of Oncology, School of Medicine, Johns Hopkins University,

27

Baltimore, MD, USA

28

29

Please address correspondence to:

30

Pradeep Natarajan, MD MMSc

31

185 Cambridge Street, CPZN 5.238

32

Boston, MA 02446

33

617-726-1843

34

pnatarajan@mgh.harvard.edu

35

36

Funding:

37

P.N. is supported by NHGRI U01HG011719 and NHLBI R01HL127564. N.C. is

38

supported by NHGRI U01HG011719 and R01HG010480.

39

40

41

Disclosures:

42

P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific,

43

Genentech / Roche, and Novartis, personal fees from Allelica, Apple, AstraZeneca,

44

Blackstone Life Sciences, Creative Education Concepts, CRISPR Therapeutics, Eli

45

Lilly & Co, Esperion Therapeutics, Foresite Capital, Foresite Labs, Genentech /

46

Roche, GV, HeartFlow, Magnet Biomedicine, Merck, Novartis, TenSixteen Bio, and

47

Tourmaline Bio, equity in Bolt, Candela, Mercury, MyOme, Parameter Health,

48

Preciseli, and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals,

49

all unrelated to the present work.

50 Abstract

51 Calculating optimal polygenic risk scores (PRS) across diverse ancestries,
52 particularly in admixed populations, is necessary to enable equitable genetic
53 research and clinical translation. However, the relatively low representation of
54 admixed populations in both discovery and fine-tuning individual-level datasets limits
55 PRS development for admixed populations. Under the assumption that the most
56 informative PRS weight for a homogeneous sample, which can be approximated by a
57 data point in the ancestry continuum space, varies linearly in that space, we
58 introduce a Genetic Distance-assisted PRS Combination Pipeline for Diverse Genetic
59 Ancestries (**DiscoDivas**) to interpolate a harmonized PRS for diverse, especially
60 admixed, ancestries, leveraging multiple PRS weights fine-tuned within single-
61 ancestry samples and the genetic ancestry continuum information. DiscoDivas treats
62 ancestry as a continuous variable and does not require shifting between different
63 models when calculating PRS for different ancestries. We generated PRS with
64 DiscoDivas and the current conventional method, i.e. fine-tuning multiple GWAS PRS
65 using the matched or similar ancestry sample, for simulated datasets and large-scale
66 biobank datasets (UK Biobank [UKBB] N=415,402, Mass General Brigham Biobank
67 N=53,306, *All of Us* N=245,394) and compared our method with the conventional
68 method with quantitative traits and complex disease traits. DiscoDivas generated a
69 harmonized PRS of the accuracy comparable to or higher than the conventional
70 approach, with the greatest advantage exhibited in admixed samples: DiscoDivas
71 PRS for admixed samples was more statistically accurate than the PRS fine-tuned in
72 matched or similar ancestry sample in 12 out of 16 simulated scenarios and was
73 statistically equivalent in the remaining four scenarios; when tested with quantitative
74 trait data in UKBB, DiscoDivas increased the PRS accuracy of admixed sample by
75 5% on average; yet no statistical difference was observed when tested for binary
76 traits in UKBB where ancestry-matched data was available. For the single ancestry
77 samples, the accuracy of DiscoDivas PRS and PRS fine-tuned in match samples
78 was similar. In summary, our method DiscoDivas yields a harmonized PRS of robust
79 accuracy for individuals across the genetic ancestry spectrum, including where
80 ancestry-matched training data may be incomplete.
81

82 Introduction/Main

83 Individuals not of European ancestry remain underrepresented in GWAS, which at
84 least partly explains why PRS performance is generally reduced among those of non-
85 European versus European ancestry¹. Within the constraints of existing data, the
86 current principal solution to increase the PRS accuracy among non-European
87 individuals is to fine-tune a combination of PRS derived from multiple populations or
88 multiple traits with the individual-level data of a validation sample²⁻⁶. However, PRS
89 accuracy decays as the genetic distance between the testing and validation samples
90 increases⁷. Relative to the vast diversity across the genetic ancestry continuum, the
91 existing and near-term individual-level datasets that can be used for fine-tuning PRS
92 combinations remains very sparse. Most existing individual-level genotype data are
93 mainly collected from single-ancestry populations and therefore admixed populations
94 are left underrepresented and largely unaddressed⁸⁻¹¹. Additionally, testing and
95 validation samples that are labeled as “from the same superpopulation” are often
96 truly genetically heterogeneous^{10,12-15}, leading to variable accuracy within such
97 samples.
98

99 PRS analysis across diverse ancestries may also be limited by inconsistency. The
100 raw PRS distributions of the same model varies by ancestry and therefore the raw
101 PRS values for individuals of different genetic ancestries should not be directly
102 compared without ancestry correction¹⁶⁻¹⁸. Although prior research^{16,18,19} has shown
103 that regressing out the top principal components of ancestry (PCA) from the PRS can
104 unify the PRS distributions of different ancestries (i.e., the mean and standard
105 deviation of corrected PRS sampled from different populations can become very
106 close), the inconsistency is only partially solved. In the application of PRS across
107 diverse ancestries, one would have to use one PRS model for all the individuals,
108 causing inconsistent PRS accuracy, or use several discrete PRS models for different
109 individuals approximating superpopulations causing inconsistent PRS modelling and
110 accuracy.

111
112 Given these issues and the increasing clinical use of PRS²⁰⁻²², PRS generation for
113 diverse genetic ancestries with more consistent accuracy and more unified PRS
114 distributions is critically needed. We devised a method, DiscoDivas, a Genetic
115 **Distance**-assisted PRS **Combination** Pipeline for **Diverse Genetic Ancestries**, to
116 generate PRS across the genetic ancestry continuum. This method is based on the
117 recent observation⁷ that the PRS accuracy in the testing data decays approximately
118 linearly as the genetic distance between the testing and validation samples
119 increases, and that the genetic distance can be approximated by Euclidian distance
120 of PCA based on the global ancestries⁷. Based on this observation, we assumed that
121 the most informative PRS weights for a sample can be linearly interpolated from the
122 currently available PRS weights that are fine-tuned in the ancestries surrounding it in
123 the global ancestry-based PCA space with the interpolation weights based on the
124 Euclidian distance of the PCA. In summary, DiscoDivas calculates PRS for diverse
125 genetic ancestries whose genetic data may not be sufficiently powerful to train the
126 PRS model by linearly interpolating the multiple PRS fine-tuned in ancestries whose
127 genetic data are more available. We evaluated its performance in simulated and
128 empiric data.
129

130 Results

131 Overview of DiscoDivas

132 DiscoDivas combines PRS fine-tuned in different validation samples - generally from
133 different single-ancestry populations - to linearly interpolate PRS for individuals of
134 diverse genetic ancestries, treating the ancestry information as a continuous variable.
135 The rationale for PRS combination is based on the observation that the correlation of
136 the most informative PRS weight for two samples of different ancestry drops as the
137 genetic distance, represented by Euclidean distance of global ancestry-based PCA,
138 increases⁷. Therefore, the best PRS weight for an ancestry representation can be
139 linearly interpolated from other PRS weights fine-tuned in other ancestries with the
140 additional consideration of the genetic distance between the samples.

141
142 Under the same principle of interpolating the PRS weight, the best PRS can be
143 interpolated from several PRS calculated using the weight fine-tuned in other
144 ancestries. Since generating individual-specific PRS weights in a testing dataset
145 causes redundant calculation and given the difficulty of normalizing information from
146 different datasets, we combine the PRS instead of the SNP weights. The PRS of

147 individuals in the testing sample is a linear combination of PRS based on the SNP
 148 weights fine-tuned in different validation samples:

149
$$PRS_i = \sum w_{i,k} PRS_{i,k}$$

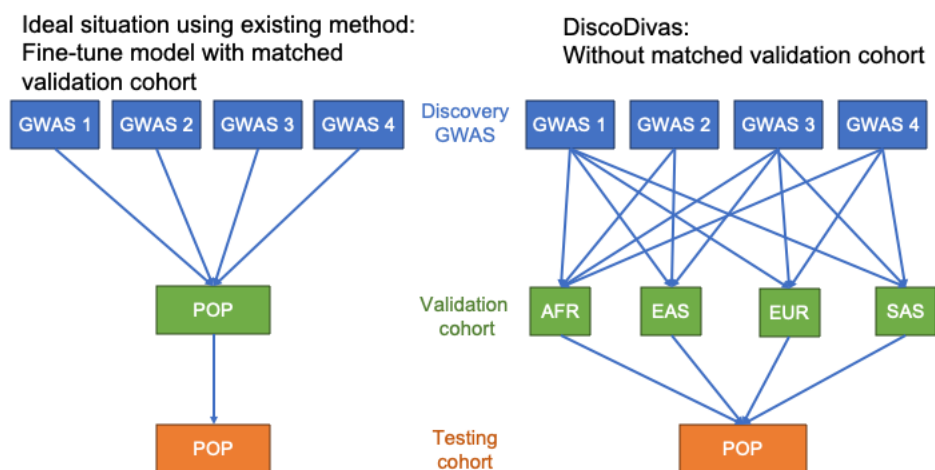
150 where $PRS_{i,k}$ is the PRS of testing individual i calculated using the weight fine-tuned
 151 in validation sample k ; $w_{i,k}$ is the combination coefficient mainly based on the
 152 reciprocal of the PCA Euclidean distance between the testing individual and median
 153 point of validation sample D_{i-k} . Note that the input PRS and PCA should be of the
 154 same scale: all the individuals are projected to the same PCA space based on a
 155 global ancestry reference panel and the PRS input $PRS_{i,k}$ is the raw PRS regressed
 156 out the top PCs and then standardized.

157
 158 In addition to the PCA distances, other factors are included in the model. First, since
 159 some fine-tuning samples are more correlated than others (e.g., EAS and SAS are
 160 more correlated than AFR and EUR), the combination coefficients should be further
 161 modified by these correlations, which can also be extracted from the PCA Euclidean
 162 distances. Second, since PRS fine-tuned in each of the validation samples may be of
 163 differing qualities (e.g., when the PRS model fine-tuned in different samples are
 164 based on GWAS of different sample sizes or populations), the quality of the PRS
 165 trained with each of the training data will vary and should be taken into account when
 166 combining the PRS. Thus, the combination coefficient $w_{i,k}$ in the previous formula is
 167 a function of multiple factors:

168
$$w_{i,k} = f\left(\frac{1}{D_{i-k}}, G, r_k\right)$$

169 where $\frac{1}{D_{i-k}}$ is the reciprocal of PCA Euclidean distance between the individual i and
 170 the validation sample k ; G is the matrix of PCA Euclidean distance between
 171 validation samples; r_k is the parameter describing the quality of validation samples. A
 172 more detailed description of implementing DiscoDivas is given in the Method session.
 173

174 Overview of multi-population GWAS PRS model



175
 176 **Figure 1: The workflow of comparing DiscoDivas with the existing method.**

177 Left: The ideal situation for the existing method is to fine-tune a PRS model that contains multiple
 178 GWAS with matched validation data, which is not currently available for many under-represented
 179 populations. Right: DiscoDivas first fine-tunes the PRS in the available ancestries, which are currently
 180 AFR, EAS, EUR, and SAS, and interpolates PRS for diverse ancestry groups based on these fine-tuned

181 PRS. In this plot, POP refers to any ancestry for which the PRS is to be calculated.

182

183 A common approach for constructing PRS is to include as much genome-wide
184 association study (GWAS) summary statistic data as possible in the discovery
185 data^{5,23,24}. The GWAS data is typically then processed by PRS methods that will
186 adjust the SNP effect size using a set of hyper-parameters. Individual-level data of an
187 independent validation sample is used to fine-tune the hyper-parameters across PRS
188 methods and the combination of the fine-tuned PRS. The resulting PRS is expected
189 to perform the best in samples of matched ancestry with the validation sample.

190

191 The current approach, as shown in the left panel of Figure 1, is to use the multi-
192 GWAS PRS fine-tuned in the matched sample or the closest approximation when the
193 matched sample is unavailable. The pipeline of adjusting SNP effect sizes and
194 combining information from different GWAS varies widely. Without loss of generality,
195 we built the following pipeline as a representation of the current conventional method:
196 we first adjusted the SNP effect size of each of the summary statistical GWAS
197 datasets by a Bayesian method and then chose the most predictive PRS from all the
198 PRS generated under different hyper-parameters. For simulated GWAS data, we
199 used PRS-CS²⁵ to adjust the SNP effect size and LDpred2²⁶ for real GWAS. Then
200 we used the validation data to first select the most predictive PRS based on each
201 GWAS and then to train the linear combination of the most predictive single-GWAS
202 PRS with a linear regression model. The final PRS model generated from each of the
203 validation datasets is a linear combination of PRS. For the empiric data set, the PRS
204 were fine-tuned controlling for the following covariates: top 20 PCA, sex, and age.
205 We used AFR, EAS, EUR, SAS, AMR, and admixed samples to fine-tune the PRS.

206

207 On top of the conventional method, DiscoDivas calculates PRS of any ancestry using
208 a linear combination of a group of PRS fine-tuned in currently available samples as
209 shown in the right panel of Figure 1. The PRS input for DiscoDivas in this study was
210 the multi-GWAS PRS fine-tuned in AFR, EAS, EUR, and SAS validation samples
211 with the conventional method pipeline as mentioned above. The interpolation of
212 these four PRS is based on the PCA calculated using the 1000 Genomes reference
213 panel. For most of the PRS analysis conducted in in the present study, the input PRS
214 of DiscoDivas are based on the same set of discovery GWAS and the validation
215 datasets are sufficiently large to generate a stable result. Therefore, we assumed
216 that all the input PRS can be viewed as of equal quality and their parameter for PRS
217 quality A_k can be viewed as a constant value in the present study.

218

219 Simulated data results

220 Summary-level GWAS used as discovery data were generated based on simulated
221 genotype data based on 1000 Genomes reference as described in the previous
222 publication provided by Zhang et al⁶. The phenotypes were based on 100, 300,
223 1,000, or 10,000 causal SNPs randomly selected from the 1.4 million Hapmap3
224 SNPs to represent traits of different polygenicity. The per-allele effect sizes of the
225 causal SNPs followed a normal distribution, and the heritability was set as 0.6.
226 Scenarios where both casual SNPs and effect sizes were constant across the
227 populations and where the casual SNPs were shared but the effect sizes varied
228 across the populations were simulated. We used up to 100,000 simulated individuals
229 from AFR, EAS, EUR, and SAS to generate the discovery summary statistic GWAS
230 dataset with PLINK2²⁷ and left the remaining samples out for other downstream
231 analyses.

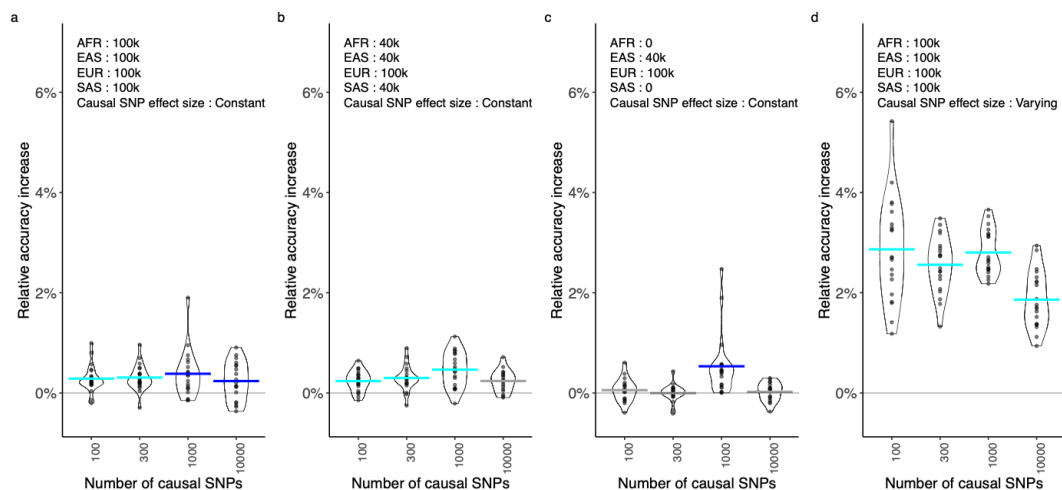
232

233 Validation and testing samples were also simulated based on UKBB genotype data
 234 with the phenotypes simulated using the same pipeline and parameters as described
 235 above. UKBB participants were divided by genetic ancestry using 1000 Genomes as
 236 a reference. In addition to the groups of AFR, EAS, EUR, SAS, and AMR whose PCA
 237 information were matched with the 1000 Genomes reference, we identified the group
 238 of “to-be-determined” (tbd) for admixed individuals whose PCA information was not
 239 matched with any of the five ancestries by definition. From each ancestry group, 1.3k
 240 individuals were used as the validation datasets (See section entitled ‘Simulated
 241 data’ in Methods). The SNP effect sizes from the discovery GWAS data were
 242 adjusted using PRS-CS with the default parameters, fine-tuned, then combined using
 243 the validation data based on 1.3k UKBB-based individuals per population and tested
 244 in the rest of the UKBB-based individuals. The process of selecting causal SNPs,
 245 assigning effect size, simulating phenotype data, and the downstream GWAS and
 246 PRS analysis was repeated 20-fold.

247

248 We primarily focused on the PRS performance in the admixed testing sample.
 249 DiscoDivas, which is based on PRS fine-tuned in AFR, EAS, EUR, and SAS, was
 250 compared with the conventional PRS fine-tuned in the matched admixed validation
 251 sample in scenarios of different causal SNP numbers, different discovery GWAS
 252 sample sizes, and different causal SNP distribution across ancestry (See Figure 2)

253



254

255 **Figure 2 Relative R^2 increase of DiscoDivas over the conventional PRS fine-tuned in a matched**
 256 **sample when tested in admixed individuals. The x-axis shows the simulated number of causal**
 257 **SNPs. The horizontal bar shows the mean relative R^2 increase and the color of the horizontal bar**
 258 **indicates the p -value of the paired t-test of DiscoDivas PRS R^2 and conventional PRS R^2 , with**
 259 **cyan being p -value<0.0005, dark blue being p -value<0.05 and grey being p -value>0.05. In panels**
 260 **a, b, and c, the causal SNP effect sizes are constant across different populations. The annotation**
 261 **texts on the top of each panel shows the sample size of discovery GWAS of different**
 262 **populations and the distribution of causal SNP effect sizes.**

263

264 Although the comparison between DiscoDivas and the conventional method of fine-
 265 tuning PRS with matched ancestry sample in a single test iteration usually showed
 266 no statistical significance due to the small numeric differences, the paired t-test of
 267 DiscoDivas R^2 and the conventional PRS R^2 over the 20 iterations better clarified
 268 significant differences. When effect sizes of causal SNPs were constant across
 269 different ancestries (Figure 2 panel a, b, and c), the PRS generated by DiscoDivas
 270 had comparable accuracy with the PRS fine-tuned using matched data. We noticed
 271 that when the sample size of non-European discovery GWAS dropped and the
 272 dataset was relatively more Eurocentric, the advantage of DiscoDivas became less

273 statistically significant. In Figure 2 panel d, we compared DiscoDivas and the
274 conventional PRS method of fine-tuning the PRS with matched ancestry in the
275 scenario where causal SNPs were shared across all populations, but the effect sizes
276 varied linearly in the PCA space. The advantage of DiscoDivas over conventional
277 PRS method was more obvious in this scenario than when the effect sizes were
278 constant across populations (Figure 2 panel a and d), presumably because
279 personalized PRS combination with DiscoDivas better captured the changing effect
280 sizes for the admixed testing sample. In all the scenarios tested, the advantage of
281 DiscoDivas was least statistically significant when the number of causal SNPs was
282 10,000 but still significant when the number of causal SNPs was 1,000. Notably, the
283 accuracy of both DiscoDivas and the conventional PRS method was the lowest when
284 the number of causal SNPs was 10,000 (Supplementary Figure 1), indicating that the
285 difference of the two PRS methods became less obvious when the input data
286 became increasingly underpowered.

287

288 When predicting the individuals that are usually classified as single ancestries, i.e.
289 AFR, EAS, EUR, and SAS, DiscoDivas showed no statistically significant difference
290 or a slight advantage over the conventional PRS method (Supplementary Figure 2).
291 When predicting AMR individuals, we used admixed validation data (tbd) to fine-tune
292 the conventional PRS due to the small sample size of the AMR dataset. The PRS
293 performance when testing in the AMR dataset was similar as in admixed data but the
294 statistical significance was weaker, potentially due to the small sample size and the
295 high heterogeneity of the AMR dataset. In general, DiscoDivas showed its clearest
296 advantage over the conventional method of fine-tuning PRS with matched PRS when
297 the testing data and the validation data for the conventional method were of different
298 ancestries.

299

300 Sensitivity tests

301 Since the quality of validation data is essential to the performance of DiscoDivas, we
302 evaluated the influence of minor missing information or alternative choices of
303 validation data with the following tests:

304

305 First, we considered the possible scenario where PRS weights for different
306 ancestries are provided from a publication but key detailed information about the
307 validation data was not fully available, especially the PCA information of the
308 validation datasets. A convenient approximation of the PCA of validation datasets is
309 the median PCA value of the 1000 Genome²⁸ individuals of a certain ancestry or
310 “superpopulation.” Based on the simulation test as mentioned above, we tested the
311 influence of replacing the actual PCA of the UKBB validation datasets with the 1000
312 Genome approximate on the 1) PCA Euclidean distance, 2) combination coefficient
313 for interpolation, and 3) the PCA accuracy.

314

315 Since the UKBB individuals were selected to be included in validation datasets based
316 on the PCA information of the 1000 Genomes, the PCA distribution of UKBB
317 validation datasets closely aligned with that of the 1000 Genomes reference
318 (Supplementary Figure 3). PCA distances between the testing individuals and the
319 median point of validation datasets based on the actual UKBB validation data were
320 highly correlated with the PCA distances based on the 1000 Genome approximation
321 for all the testing ancestry groups AFR, AMR, EAS, EUR, SAS, and admixed (“tbd”),
322 with the data residing within the highly overlapped intervals and the correlation of
323 individual datapoints close to 1 (Supplementary Figure 4). The combination

324 coefficients were calculated assuming that the PRS weights fine-tuned in the four
325 validation datasets, i.e. AFR, EAS, EUR, and SAS, were of equal quality. According
326 to the formula in Method session, $w_{i,k}$, the combination coefficient of weighting the
327 PRS fine-tuned in validation sample k for individual i , should be $w_{i,k} \equiv \frac{1}{D_{i-k}} d_k$,
328 with the D_{i-k} being the PCA distance between individual i and the median point of
329 sample k ; d_k being the adjustment coefficient for PRS fine-tuned in sample k based
330 on the distance between the validation samples. $w_{i,k}$ was compared between the two
331 scenarios of using the actual UKBB validation datasets versus using the 1000
332 Genomes approximate. The correlation of the combination coefficients was lower
333 than the correlation of the PCA distance, especially for SAS testing individuals.
334 However, each combination coefficient remained in almost the same range and the
335 PRS fine-tuned in the SAS sample still had the highest weights (Supplementary
336 Figure 5). When testing with the simulated data, the PRS R^2 had almost identical
337 distribution and correlation > 0.99 with the R^2 of PRS based on the actual PCA
338 information in all the simulated scenarios including when the causal SNP effect size
339 varied with the ancestries (Supplementary Figure 6). The high similarity of the PRS
340 accuracy despite the difference in combination coefficient might partly result from the
341 correlation of the PRS fine-tuned in different samples and the constant combination
342 coefficient range.

343

344 In addition, we tested if the results of DiscoDivas would remain robust for admixed
345 individuals when using a different set of validation datasets. In addition to the primary
346 simulation test where the validation datasets were simulated data based on UKBB
347 genotype, PRS were fine-tuned with the left-out simulated datasets that were
348 independent from the discovery GWAS while other parts of the analysis pipeline
349 remained the same. The simulated AMR dataset was used to fine-tune PRS for AMR
350 and admixed (tbd) testing samples for the conventional method. The PRS R^2 based
351 on the two sets of validation datasets were compared in the scenarios where the
352 discovery GWAS was based on 100k AFR, EAS, EUR, and SAS, and where the
353 causal SNP effect sizes were constant across different ancestries. The correlation of
354 PRS R^2 based on UKBB-based validation datasets and purely simulated validation
355 datasets of DiscoDivase was larger than 0.99 in at scenarios, much higher than that
356 of conventional PRS method, which ranged from 0.73 to 0.98 (Supplementary Figure
357 7). The advantage of DiscoDivas over the conventional PRS method showed a
358 similar pattern (Supplementary Figure 8)

359

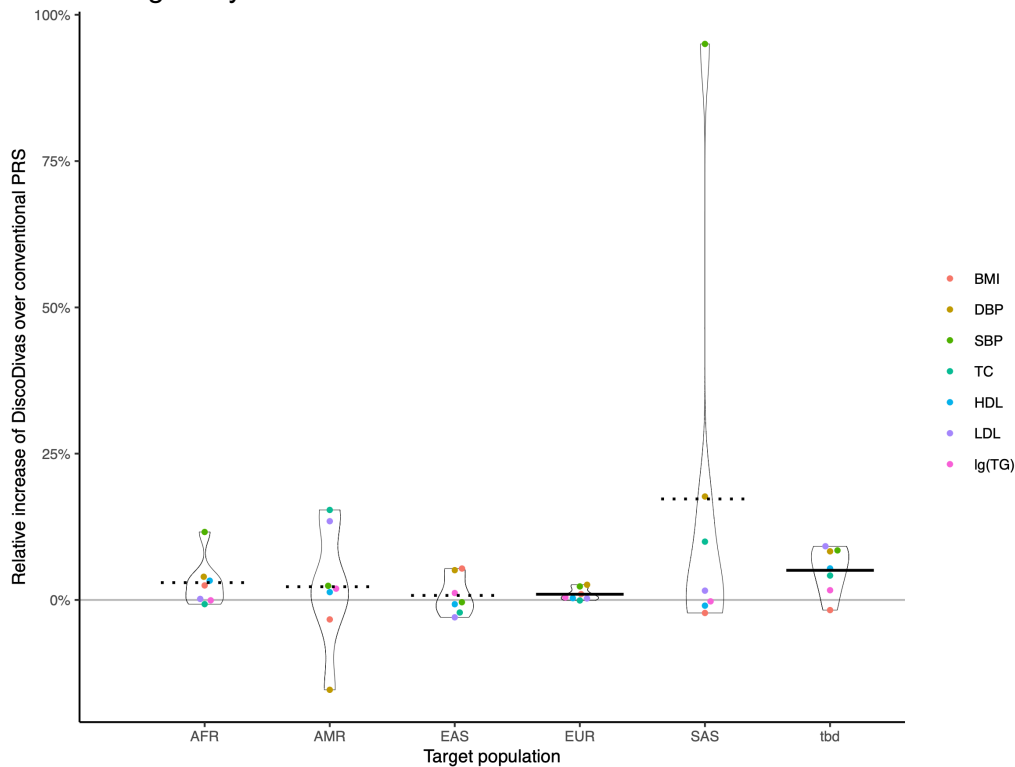
360 Biobank data results

361 We downloaded publicly available summary statistical data of body-mass index
362 (BMI), high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol
363 (LDL), total cholesterol (TC), triglycerides (TG), systolic blood pressure (SBP),
364 diastolic blood pressure (DBP), coronary artery disease (CAD), and diabetes mellitus
365 (DM2) and adjusted the SNP effect size using LDpred2 as described previously⁵.

366

367 For the quantitative traits, we used the validation samples of AFR, EAS, EUR, SAS,
368 and admixed (“tbd”) ancestry to fine-tune the model. The remaining UKBB samples
369 were used as the testing data. The results for empiric quantitative trait data were
370 highly aligned with the simulation results (Figure 4): DiscoDivas showed a robust
371 advantage over the conventional PRS method of fine-tuning PRS with matched or
372 similar ancestry samples when compared across the 7 traits in the admixed testing
373 dataset. When predicting AFR, EAS, EUR, and SAS, DiscoDivas and the
374 conventional PRS method had similar performance. The results of both methods in

375 AMR testing dataset had large deviations due to the small sample size and greater
 376 genetic heterogeneity of the AMR data.



377
 378
 379 **Figure 3 Relative R^2 increase of DiscoDivas over the conventional PRS fine-tuned in a matched**
 380 **sample. The x-axis shows the population in which the PRS was tested. We used “tbd” as the**
 381 **fine-tuning dataset for the test in both tbd and AMR due to the absence of matched AMR**
 382 **validation data. The horizontal bar shows the mean of relative increase and the line-type of the**
 383 **bar indicates the p -value of paired t-test of DiscoDivas PRS R^2 and conventional PRS R^2 , with the**
 384 **solid bar being p -value <0.05 and dotted bar being p -value >0.05**

385 For the binary traits coronary artery disease (CAD) and type 2 diabetes (DM2)
 386 (Figure 4), we used the AFR, EAS, EUR, SAS, AMR, and OTH (i.e., unclassified)
 387 samples from AoU as the fine-tuning data and tested in AFR, EAS, EUR, SAS, and
 388 tbd individuals in UKBB and AFR, EAS, EUR, SAS, and AMR individuals in MGBB.
 389 The DiscoDivas PRS were based on the PRS fine-tuned in AFR, EAS, EUR, and
 390 SAS and used the default assumption that the PRS fine-tuned from all the samples
 391 were of similar quality even though the sample sizes of both discovery GWAS and
 392 the fine-tuning samples were not balanced across different ancestries. AMR in UKBB
 393 was excluded because of the small sample size (N=669).

394
 395 The PRS fine-tuned in different single samples and the DiscoDivas PRS had similar
 396 performances. It also appeared that some of the validation sample could be
 397 underpowered: generally, we expect the PRS fine-tuned in the matched sample to
 398 perform the best in the testing samples, but PRS fine-tuned in larger validation data
 399 performed better than PRS fine-tuned in smaller validation data in general. For
 400 example, the PRS fine-tuned in EAS AoU data performed worse than other PRS in
 401 both MGBB and UKBB EAS data and had low accuracy in other testing data as well;
 402 the CAD PRS fine-tuned in EUR performed better than all the other PRS in all the
 403 testing data and the effective sample size of EUR CAD validation data was much
 404 larger than all the other validation data.

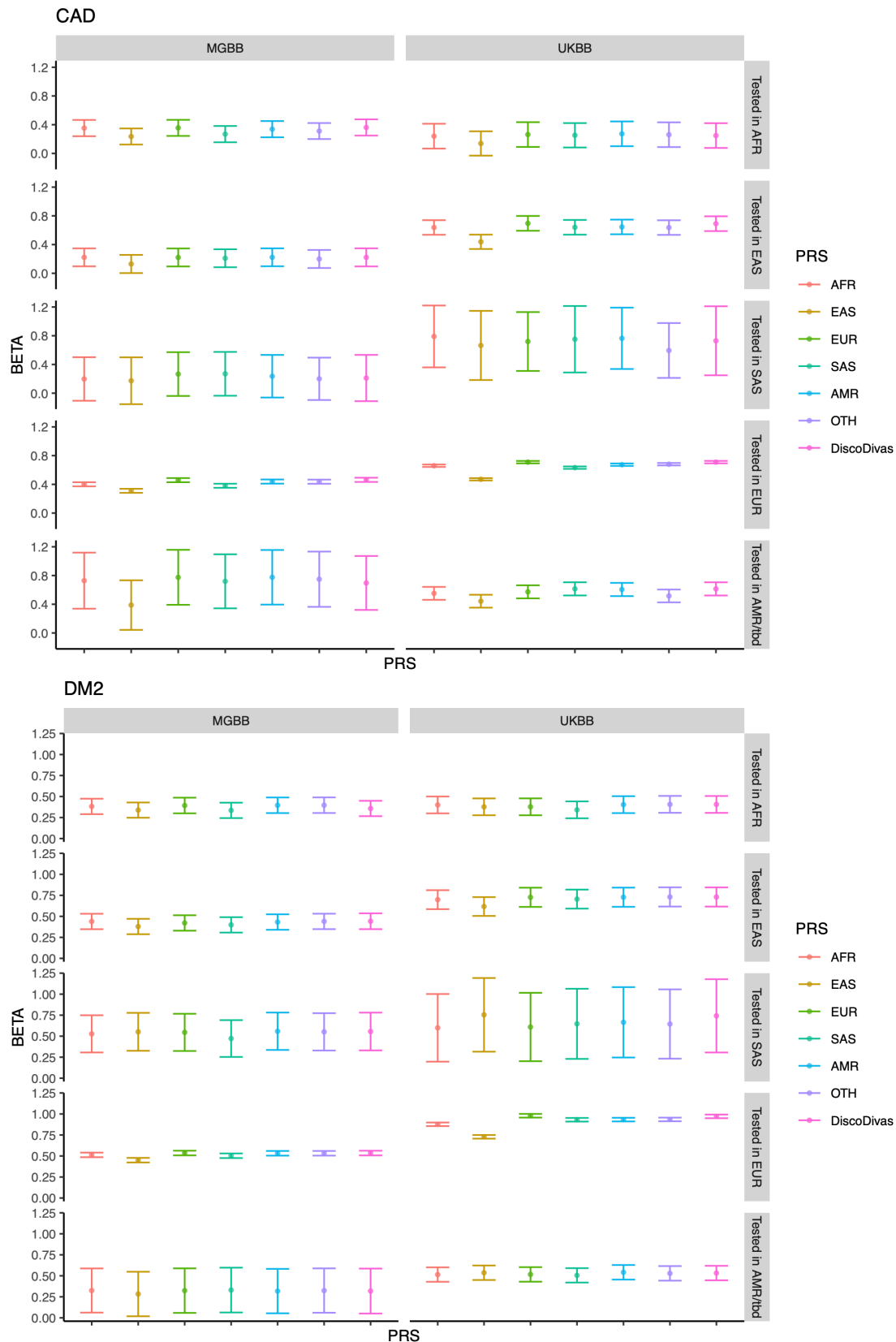


Figure 4 PRS performance for coronary artery disease (CAD) and type 2 diabetes (DM2) tested in UKBB and MGBB. The plot shows OR per SD with the error bar showing 95% CI. The sub-panels show that population of the testing sample and the different colors show the method for generating the PRS, either fine-tuning in a single sample or combining the PRS using DiscoDivas.

413

414 Discussion

415 We propose a new method, DiscoDivas, to interpolate the PRS for diverse, especially
416 admixed, ancestries with a generalized framework that does not requiring binning
417 into discrete ancestries. Our results shows that the accuracy of DiscoDivas was
418 comparable to or greater than the conventional method, i.e. fine-tuning using the
419 matched population sample when available. In addition, when generating PRS for a
420 wide range of ancestries, DiscoDivas did not require shifting from several sets of
421 PRS weight fine-tuned in discrete samples while remaining matched with the
422 ancestry information. Our method provides a new solution to generate PRS for
423 underrepresented, generally admixed, populations and as well as generate a
424 harmonized PRS model across different ancestries.

425

426 The performance of our method depends on the quality of both the discovery GWAS
427 data and the validation data. As shown in the simulation test, discovery GWAS
428 datasets that represent diverse ancestries with sufficient sample size will increase
429 the accuracy of interpolated PRS generated by DiscoDivas. On the contrary,
430 Eurocentric and underpowered discovery GWAS datasets would limit the advantage
431 of DiscoDivas over the conventional PRS method. This might partly explain the
432 limited advantage of DiscoDivas when predicting binary traits: the discovery GWAS
433 datasets were highly Eurocentric and the GWAS, especially the non-European ones,
434 could be more underpowered than quantitative trait GWAS. Furthermore, the PRS
435 fine-tuned in validation datasets of insufficient sample size will be overfitted and
436 cannot be used to fairly evaluate the performance of either the conventional PRS
437 method or DiscoDivas. We aimed to address this issue by only using traits that 1)
438 had effective sample sizes larger than 200 in all the validation samples, and 2) had
439 high-quality phenotyping data in both the validation datasets and the testing datasets,
440 However, Asian populations were largely under-represented in the current public
441 biobanks: the effective sample size of many binary traits in EAS or SAS can be as
442 small as <200 even in AoU, the most diverse and large-scale largely publicly-
443 available biobank we had access to. This limited our choice for binary traits to only
444 CAD and DM2. One additional limitation of our method is that DiscoDivas does not
445 consider the local ancestry information, which improve PRS predictions in various
446 research^{24,29,30}, especially PRS prediction of newly admixed populations³¹.

447

448 Our research underscores the notion that non-European populations, both admixed
449 and single-ancestry populations, remain largely under-represented in the existing
450 genetic data. Furthermore, some potential extensions of our method will not become
451 possible until we collect more diverse and larger datasets. First, our method has not
452 been designed nor tested for extrapolating data, e.g. generating PRS for continental
453 African samples based on African American, European, and Asian samples. Even
454 though it is mathematically plausible to alter our method to extrapolate the PRS, we
455 lack data such as continental African samples to test the method. Secondly, we
456 currently only consider the assumption that the most informative genome-wide PRS
457 weight shifts linearly in the PCA space. Although more complicated PRS
458 interpolation, e.g. interpolation guided by local ancestry information^{24,29,30}, pathway-
459 specific^{32,33} and annotation-guided³⁴ PRS weights and polynomial interpolation^{35,36},
460 can possibly further improve the PRS accuracy, training such complicated models
461 would require collecting much larger and more diverse datasets than the existing
462 data. Finally, additional biological insights could be revealed by interpolating PRS if
463 genetic data of all the involved diverse ancestries are of sufficient power. In this case,

464 the differences between interpolated PRS and the PRS trained using the matched
465 ancestry would indicate the population- or sample- specific factors absent in the
466 interpolation model, e.g. population-specific genetic variance³⁷, complicated
467 population stratification involving confounding factors^{38,39}, sample/ancestry-specific
468 modifiers like local adaptation³⁸, gene x environment interactions⁴⁰ or other factors
469 that contribute to the genetic variant frequency or effect size in these samples/
470 ancestries.

471
472 In conclusion, our method provides a new option to treat the ancestry information as
473 a continuous variable and interpolate a harmonized PRS for diverse ancestries.
474 Notably, although our method was developed primarily to calculate PRS when the
475 matched validation datasets were unavailable, our research showed that successfully
476 interpolating PRS required sufficient input data and highlighted the need to collect
477 genetic data for underrepresented populations. We believe that more diverse and
478 larger data collected in the coming future will enable the development of new
479 methods of interpolating PRS and the elucidation of the genetic basis of complex
480 traits.
481

482 **Methods**

483 **DiscoDivas**

484 DiscoDivas interpolates PRS of testing individuals of diverse ancestry according to
485 the testing individual's PRS calculated using the PRS weight fine-tuned in other
486 validation datasets and genetic distance information. The pipeline consists of two
487 parts: harmonizing the input PRS data and interpolating the PRS.
488

489 **Data harmonization**

490 To reduce the bias in the interpolation, the PRS and PCA information should be in a
491 unified and harmonized scale. First, all the individuals in the validation datasets and
492 the testing dataset are projected in the same PCA space based on balanced
493 reference data covering the global genetic ancestry continuum. The reference data is
494 essential to avoid a skewed correlation between the similarity of the genotype and
495 the genetic distance and ensure that the PCA Euclidian distance can present
496 consistent genetic distance. The results presented in this study were based on PCA
497 calculated using pruned SNPs of 1000 Genomes²⁸ samples. A detailed description of
498 calculating PCA and Euclidian distance in this research is provided in the section
499 entitled 'Calculation of 1000 Genomes-based PCA and Euclidean distance.'

500
501 Second, all the PRS input should be transferred to a comparable scale. We
502 regressed out the top 10 PCs from the PRS and then standardized the PRS residuals
503 to mean = 0 and standard deviation = 1.
504

505 **PRS interpolation**

506 The overall mathematical model of DiscoDivas is a linear combination of PRS based
507 on the weight fine-tuned in different validation datasets:

508
$$PRS_i = \sum a_i w_{i,k} PRS_{i,k}$$

509 where $PRS_{i,k}$ is the normalized PRS of individual i trained in validation dataset k ; $w_{i,k}$
510 is the combination weight which is a function of genetic distance between the
511 individual i and dataset k and other factors. a_i is a constant for individual i so that
512 $\sum a_i w_{i,k} = 1$.

513
514 The essential factor contributing to $w_{i,k}$ is the reciprocal of D_{i-k} , the genetic distance
515 between individual i and dataset k , so that the interpolation is basically linear. In the
516 ideal situation where all the validation samples are independent and generate PRS of
517 the same level of accuracy, the genetic distance between the testing individuals and
518 validation datasets is the only contribution factor and we define:

519
$$w_{i,k} \equiv \frac{1}{D_{i-k}}.$$

520 However, considering the more realistic scenarios where the validation samples can
521 be correlated and the PRS trained from different validation datasets can be of
522 different accuracy, we introduce two parameters:

- 523 1) d_k : tuning parameters based on the genetic distance / correlation between
524 the training datasets.
525 2) r_k : tuning parameter that represent the accuracy of the PRS fine-tuned in
526 sample k ;

527 so that the overall combination weight is

528
$$w_{i,k} \equiv \frac{1}{D_{i-k}} d_k r_k$$

529 The final model of DiscoDivas is

530
$$PRS_i = \sum a_i \frac{1}{D_{i-k}} d_k r_k PRS_{i,k}$$

531

532 Here we propose the default method for calculating d_k and r_k :

533 d_k is based on the genetic distance matrix G in which each row and column
534 represent a validation sample, and the element is genetic distance between the
535 samples, with diagonal ones being zero. The shrinkage follows the similar principle of
536 correcting marginal SNP effect size by inverting the LD matrix, except that the
537 correlation between the shrinkage in this step d_k and other factors like the genetic
538 distance between individuals in the testing data and the validation sample D_{i-k} and
539 the accuracy of the PRS fine-tuned in the validation sample, the vector of shrinkage
540 parameter d_k is only derived from G^{-1} :

541
$$\vec{d} = G^{-1} \vec{1}$$

542 , where $\vec{1}$ is a vector of the same length as \vec{d} and with all the elements being 1.

543 r_k is based on the accuracy of the PRS fine-tuned in the sample k . Theoretically, a
544 PRS based on common SNPs can explain all the heritability contributed by common
545 SNPs under the additive assumption. Under this scenario, the PRS R^2 is close to the
546 heritability h^2 , and the PRS is saturated (namely, adding more samples in the
547 Discovery GWAS would further increase the PRS accuracy if other conditions remain
548 the same). Previous research used percentage of heritability explained to present the
549 accuracy of the PRS so that PRS predicting traits of different heritability and binary
550 traits of different prevalence can be compared directly. Here we recommend using
551 common SNP heritability explained as the first choice of r_k if the heritability of the
552 target trait in the validation samples or a larger sample but is homogenous with the
553 validation sample is available:

554
$$r_k = \sqrt{\frac{R_{PRS}^2}{h^2}}$$

555 If h^2 is unknown, we provided an alternative approach that approximates the PRS

556 given the fact that most of the PRS is far from saturation and the accuracy of the
557 PRS is roughly linearly correlated with the sample size of the discovery GWAS N_k ,
558 we combine the information from different samples in a similar way by combining Z
559 scores in fixed-effect size meta-analysis:

$$560 \quad r_k = \sqrt{N_{k \text{ discovery}}}$$

561 If the trait used in the discovery GWAS is not the same for all the validation data, the
562 accuracy of the PRS for predicting the target traits also depends on the heritability of
563 the trait in the discovery GWAS $h_{k \text{ discovery}}^2$ and the genetic correlation between the
564 discovery GWAS trait and the target trait $r_{g_{k \text{ discovery-target}}}$

$$565 \quad r_k = \sqrt{N_{k \text{ discovery}} * h_{k \text{ discovery}}^2 * r_{g_{k \text{ discovery-target}}}}$$

566 In a common scenario where discovery data come from multiple GWAS from
567 different ancestries and of decent statistical power for each ancestry yet the
568 heritability of the target trait in the validation sample is unknown or cannot be
569 accurately estimated, the accuracy of the PRS fine-tuned in each single-ancestry
570 validation sample is hard to estimate but is likely of similar accuracy. Therefore r_k
571 can be omitted, or equivalently set to a default constant value of 1.

572
573 In the DiscoDivas script provided, the r_k is set to be the default value 1 unless
574 defined by the user otherwise, a_i, D_{i-k}, d_k is automatically calculated from the PCA
575 information provided by the user.

576

577 Constructing PRS with single fine-tuning sample

578 The PRS were derived from multiple GWAS conducted in different populations.
579 GWAS data were first processed with PRS-CS or LDpreds2 to generate adjusted
580 SNP weight: Hapmap3 SNP were first extracted from each GWAS as the input for
581 the PRS method. The PRS methods were performed using default parameters: For
582 PRS-CS, parameters of the prior distribution were set to $\phi = 1, 10^{-2}, 10^{-4}, 10^{-6}$, $a =$
583 $1, b = 0.5$ and the parameters of Markov Chain Monte Carlo (MCMC) were total
584 number of MCMC iterations = 1000, number of burn-in = 500, thinning factor of the
585 Markov chain = 5. For LDpred2, the parameter were the default set as in previous
586 research⁵: proportion of variants assumed to be causal was $1.0 \times 10^{-4}, 1.8 \times 10^{-4},$
587 $3.2 \times 10^{-4}, 5.6 \times 10^{-4}, 1.0 \times 10^{-3}, 1.8 \times 10^{-3}, 3.2 \times 10^{-3}, 5.6 \times 10^{-3}, 1.0 \times 10^{-2},$
588 $1.8 \times 10^{-2}, 3.2 \times 10^{-2}, 5.6 \times 10^{-2}, 1.0 \times 10^{-1}, 1.8 \times 10^{-1}, 3.2 \times 10^{-1}, 5.6 \times 10^{-1}$ and 1, the
589 scale of heritability was 0.7, 1 and 1.4 times of the estimated heritability, with options
590 of whether allowing a sparse output or not.

591

592 Each set of parameters generated a corresponding set of adjusted SNP weight,
593 which were then used to calculate PRS in the fine-tuning samples. The most
594 predictive PRS for each GWAS was selected based on a linear or logistic regression
595 model predicting the phenotype using the PRS and adjusting for top 10 PCs, age,
596 sex information and genotyping batch for biobank empirical analyses, and adjusting
597 for only top 10 PCs for simulated analyses. For All of Us data, sex information
598 combined assigned sex and self-reported gender to capture inclusiveness in data
599 collection.

600

601 To generate the final PRS weight, multiple top-performing PRS based on each
602 GWAS were combined through a linear or logistic regression in the fine-tuning
603 sample. Final adjusted PRS weights were a linear combination of the top SNP
604 weights from each GWAS, weighted by the regression coefficients. These combined

605 SNP weights were subsequently used to calculate PRS in the testing sample.

606

607 Conventionally, when fine-tuning polygenic risk scores (PRS) for a testing sample, it
608 is ideal to use a sample from a matched or similar population. If such a sample is
609 unavailable, the PRS can be fine-tuned using any available sample, which is often
610 from individuals of European ancestry.

611

612 Calculation of 1000 Genomes-based PCA and Euclidean

613 distance

614 We use 1000 Genomes as the reference panel for PCA calculation. The PCA should
615 be based on SNPs that are constantly included in as many samples as possible to
616 enable the use of wide-ranging training and validation datasets. We started with the
617 Hapmap3 SNPs for this set of SNPs, which has been widely used as a subset of
618 SNPs that approximates the feature of genome-wide common SNPs in many recent
619 studies that involve multi-ancestry prediction^{6,25,26,41}. We further filtered for the SNPs
620 likely to be frequently genotyped or imputed with relatively high quality by most
621 samples based on the 1000 Genome data: Hapmap3 SNPs were first extracted from
622 the five super-populations, Africans (AFR), Admixed Americans (AMR), East Asians
623 (EAS), Europeans (EUR) and South Asians (SAS) of the 1000 Genomes. Secondly,
624 SNPs described as the following were excluded: 1) of minor allele frequency lower
625 than 1% in any of the super-population, 2) of minor allele frequency lower than 5% in
626 the combined 1000 Genomes data, and 3) in the long-range LD region (25Mb –
627 35Mb by hg19 assembly on chromosome 6 and 7Mb – 13Mb on chromosome 8). To
628 calculating the PCA loading, the QC'ed SNPs of the five super-populations were
629 merged then pruned using the PLINK2 function “indep-pairwise” with the parameter
630 “200 100 0.1” - namely the pruning was performed using window size = 200kb, step
631 size = 100, and phased-hardcall- r^2 = 0.1. The principal components and the SNP
632 loadings are calculated using PLINK2 function “pca” with the parameter “allele-wts”
633 based on the pruned SNPs.

634

635 Based on the protocol suggested on the PLINK2 website ([https://www.cog-](https://www.cog-genomics.org/plink/2.0/score#pca_project)
636 [genomics.org/plink/2.0/score#pca_project](https://www.cog-genomics.org/plink/2.0/score#pca_project)), we projected samples for fine-tuning and
637 PRS testing into the PCA space as describe above by calculation the linear score,
638 i.e. the sum of alternative alleles weighted by the SNP effect size, using the PLINK2
639 function “score” with the SNP loadings as effect size. The original online protocol
640 suggested linear score should be first scaled to standard variation and then rescaled
641 by multiplying the square root of eigenvalue. However, the actual standard deviation
642 of a sample in the same PCA space varies with the homogeneity and the ancestry of
643 the sample. Forcing the PCA of all the samples to have the same standard deviation
644 will cause inconsistent scaling when the samples can be of different ancestries.
645 Therefore, we directly calculated the PCA from sum basic linear score based on the
646 SNP loadings as generated above without any further scaling. The PCA in this study
647 was the sum basic linear score calculated using the PLINK2 function “score” with the
648 parameter “cols=+scoresums”. For large samples whose genotype data were divided
649 into per-chromosome files, the same commands were used to calculate per-
650 chromosome score and the genome-wide score was the sum of the score of all the
651 autosomes.

652

653 In DiscoDivas' default setting, the genetic distance between two individuals is defined
654 as the Euclidian distance between the PCA of the two individuals. When the genetic

655 distance calculation involves a sample, we use the median point to present the whole
656 sample.

657

658 We also explored the relationship between number of PCs included in the calculation
659 and the Euclidean distance calculated (Supplementary Figure 9) and the distance
660 calculated converged when the number of PCs was larger than 6 in our tests. In our
661 analysis we use the top 10 PCs to calculate the PCs.

662

663 Genetic ancestry reference

664 We noticed that the protocol of generating top PCs for ancestry references varied in
665 previous publications. In our pilot test (see supplementary text section entitled 'Pilot
666 test of generating PCA based on less QC'ed SNPs'), we compared the ancestry
667 reference based on Hapmap3 SNP without any QC and found the result to be highly
668 correlated. We used the same set of PCs based on QC'ed SNP as described in
669 section 'Calculation of 1000 Genomes-based PCA and Euclidean distance' for both
670 genetic ancestry reference and Euclidean distance calculation for data consistency.

671

672 Random forest model of 100 trees was trained based on the 1000 Genome data. The
673 out-of-bag estimate of error rate stabilize at the level of 0.28% after the number of
674 PCs passed 5. We used the model using the top 6 PCs to infer the genetic ancestry
675 of UK Biobank individuals and the Mass General Brigham Biobank individuals. The
676 genetic ancestry of an individual was assigned to any of the five ancestries
677 represented in the 1000 Genomes reference data, i.e. AFR, AMR, EAS, EUR and
678 SAS, if the highest probability of an individual belonging to that ancestry passed a
679 threshold. If none of the ancestries had a probability above the threshold, the
680 individuals were assigned as "to be decided" (tbd), which indicated that the individual
681 was of admixed ancestries. With the consideration of the sample size and confirmed
682 by visual inspection, the threshold of probability for UK Biobank and the Mass
683 General Brigham Biobank was 0.9 and 0.8 respectively.

684

685 Data

686 UK Biobank

687 The UK Biobank (UKBB) is a volunteer sample of approximately 500,000 adults aged
688 40-69 upon enrollment living in the United Kingdom recruited since 2006⁴². UKBB
689 data used in this research were first QC'ed with the following process: Remove the
690 individuals meeting the criteria that indicate low genotype quality or contamination: 1)
691 have missing genotype rate larger than 0.02; 2) have genotype-phenotype sex
692 discordance; 3) are identified as having excess heterozygosity and missing rates; 4)
693 are identified as putatively carrying sex chromosome configurations that are not
694 either XX or XY; 5) appeared to have unreasonably large numbers of relatives. From
695 the remaining samples, individuals from a group of multiple individuals that are closer
696 than 3rd-degree relatives were retained. 415,402 individuals were left after the QC.
697 390,037 were self-identified as EUR, 7,039 AFR, 8,652 non-Chinese Asian (ASN),
698 1430 Chinese (CHN) and 6572 unknown or not answered ("tbd"), and 1672 as
699 admixed (MIX). The genetic ancestry referred from PC was largely correlated with
700 the self-reported race, with 385,038 EUR, 7,450 AFR, 8,298 SAS, 2,163 EAS, 669
701 AMR and 11,784 admixed, or to-be-decided ("tbd").

702

703 In the PRS test, UKBB samples were grouped by their genetic ancestry (see section
704 'Genetic ancestry reference'). The validation datasets for the single-ancestry
705 populations (AFR, EAS, EUR and SAS) were based on 1.3k randomly selected
706 samples whose self-report ancestry matched with their genetic ancestry and the
707 probability of random forest = 1. The validation dataset for admixed ancestry ("tbd") is
708 1.3k randomly selected samples of individuals of "tbd" genetic ancestry (see
709 Supplementary Figure 3). AMR didn't have its corresponding validation dataset due
710 to its small sample size and we used "tbd" validation datasets as a proxy since the
711 two genetic ancestries had similar PCA. The remaining individuals of UKBB were
712 used as testing data.

713
714 The quantitative trait of the UKBB samples was the measurement collected after the
715 participants enrolled. The lipid trait measurement was adjusted for cholesterol-
716 lowering medication by dividing TC by 0.8 and LDL by 0.7 as before⁴³. Cases of
717 coronary artery diseases (CAD) are defined using the definition described
718 previously²⁴; Cases of diabetes are defined as ever report the following code: E10X,
719 E11X, E12X, E13X, and E14X where X can be any integer between 0 to 9 in the
720 ICD10 diagnosis code.

721
722 UKBB participants provided consent in accordance with the primary IRB protocol,
723 and the Massachusetts General Hospital IRB approved the present secondary data
724 analysis of the UKBB data under UKBB application 7089.

725

726 **Mass General Brigham Biobank**

727 The Mass General Brigham Biobank (MGBB) is a volunteer sample of approximately
728 142,000 participants receiving medical care in the Mass General Brigham health care
729 system recruited starting 2010. 53,306 MGBB participants underwent genotyping via
730 Illumina Global Screening Array (Illumina, CA). MGBB genotype data was quality
731 controlled, imputed and assigned one of the populations AFR, AMR, EAS, EUR, SAS
732 using K-nearest neighbor model as described previously⁴⁴. The phenotype data of
733 CAD and diabetes are drawn from PheCodes based on International Classification of
734 Diseases codes, Ninth (ICD9)110 and Tenth (ICD10) revisions, from the EHR as
735 described previously³². MGBB participants provided consent in accordance with the
736 primary IRB protocol, and the Massachusetts General Hospital IRB approved the
737 present secondary data analysis.

738

739 **All of Us Research Program**

740 The *All of Us* (AoU) Research Program is a volunteer sample of more than one
741 million United States residents recruited starting 2016. AoU aims to engage
742 communities previously underrepresented in biomedical research in the United
743 States and beyond⁴⁵. In the present analysis, genetic data from the v7 245,394
744 participants who were genotyped using short read whole genome sequencing
745 (srWGS) data. Hapmap3 SNPs were extracted for the callset with the threshold of
746 (AF) > 1% or population-specific allele count (AC) > 100. Related individuals were
747 pruned according to the information provided by AoU. Due to the inclusive data
748 collection, we didn't excluded individuals whose self-report gender were different with
749 their assigned sex at birth and used the combination of self-report gender and
750 assigned sex as one of the covariates. The predicted ancestry information was
751 provided by AoU⁴⁶. The phenotypes were defined as described in previous research
752 by Buu *et al*⁴⁷.

753

754 Simulated data

755 To generate the simulated GWAS summary statistics, the genotype data was
756 generated by Zhang et al⁶ and downloaded from
757 <https://dataverse.harvard.edu/dataverse/multiancestry>. Only Hapmap3 SNPs were
758 included in the simulation. Causal SNPs were randomly selected from the Hapmap3
759 SNPs and simulated per allele effect size following normal distribution. The ladder of
760 causal SNP number was 100, 1000, 3000, 10000 and the heritability in each of the
761 population was 0.6.
762 When simulated trait whose causal SNP effect size varied linearly in the PCA space,
763 we first assumed that individuals whose PCA was the median point of the 1000
764 Genomes followed multivariate normal distribution with the covariance matrix being:
765

	EUR	SAS and EAS	AFR
EUR	1	0.7	0.4
SAS and EAS	0.7	1	0.7
AFR	0.4	0.7	1

766
767 Similar to the principle that PRS weight can be interpolated is equivalent to PRS can
768 be interpolated, causal SNP effect size varies linearly in the PCA space and,
769 therefore, can be interpolated is equivalent to genetic burden can be interpolated.
770 The genetic burden of an individual is the weighted sum of what the genetic burden
771 could be based on the simulated SNP effect size of the median point of each
772 validation sample, with the weight proportion to the reciprocal of the PCA distance.
773 We assumed that the non-genetic factor of individuals across different ancestries
774 could be summed up as a quantitative variable independently drawn from the same
775 normal distribution. The phenotype is the sum of genetic burden and non-genetic
776 factor:

$$777 \text{phenotype}_i = \sum \beta_j x_{j,i} + E_i$$

778 where the phenotype_i and E_i were the phenotype and non-genetic factor of individual
779 i ; β_j was the effect size of causal SNP j , and $x_{j,i}$ was the number of risk alleles of
780 individual i in SNP j .

781
782 We used the PLINK2²⁸ to calculate the genetic burden based on the simulated causal
783 SNPs and effect size and used R to simulate the non-genetic factors, scale the
784 genetic burden and non-genetic factor, and generate a phenotype of heritability set to
785 be 0.6. We used up to 100k individuals per population to generate the summary
786 statistical GWAS as the discovery data for the PRS test. The rest simulated data
787 were left out for the validation and testing datasets
788

789 In addition to the completely simulated data, we generated more realistic validation
790 and testing datasets of a wider ancestry range by using the QC'ed genotype data
791 from UKBB described in the section 'Biobank data.' While we used all the non-
792 European testing data, the EUR testing dataset was down-sampled to 10,000 for the
793 simulation test to reduce the computation burden. We simulated the genetic burden,
794 non-genetic factor, and phenotype based on the real-life UKBB genotype data with
795 the same pipeline and parameters. The simulated data based on UKBB genotype
796 data were used as validation and testing data in the main test and left-out completely
797 simulated data were used in the sensitivity test.
798

799 Acknowledgement

800 We thank the participants and staff of the UK Biobank (application 7089), Mass
801 General Brigham Biobank, and *All of Us* Research Program. We thank Ying Wang,
802 Paul O'Reilly, Raymond Walters for helpful discussion and advice.
803

804 Data and Code availability

805 The access to biobank data (UK Biobank, Mass General Brigham Biobank, and All of
806 US Research Program) were gained upon application. The simulated genotype data
807 based on 1000 Genomes were downloaded from
808 <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/COXHAP>
809 ; The resource of summary statistics GWAS data used to generate PRS were given
810 in the supplementary file. The scripts of running DiscoDivas and other supporting
811 files can be found at <https://github.com/YunfengRuan/DiscoDivas>
812
813

814 Reference

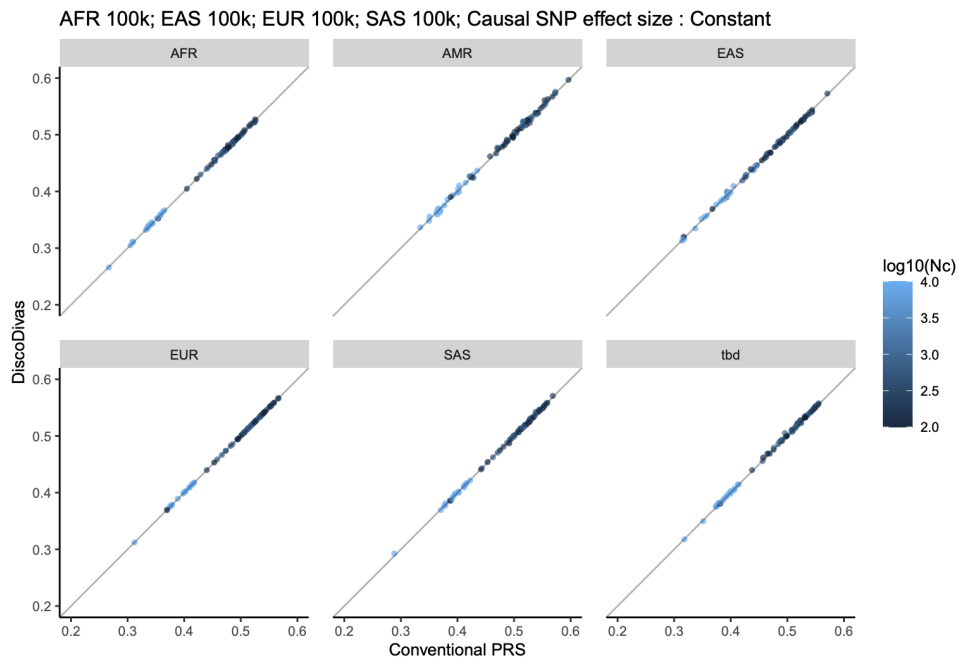
- 815 1. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may
816 exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019).
- 817 2. Miao, J. *et al.* Quantifying portable genetic effects and improving cross-
818 ancestry genetic prediction with GWAS summary statistics. *Nat Commun* **14**,
819 832 (2023).
- 820 3. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse
821 populations. *Nat Genet* **54**, 573–580 (2022).
- 822 4. Jin, J. *et al.* MUSSEL: Enhanced Bayesian Polygenic Risk Prediction
823 Leveraging Information across Multiple Ancestry Groups. *bioRxiv*
824 2023.04.12.536510 (2023) doi:10.1101/2023.04.12.536510.
- 825 5. Patel, A. P. *et al.* A multi-ancestry polygenic risk score improves risk prediction
826 for coronary artery disease. *Nat Med* **29**, 1793–1803 (2023).
- 827 6. Zhang, H. *et al.* A new method for multi-ancestry polygenic prediction improves
828 performance across diverse populations. *Nat Genet* **55**, 1757–1768 (2023).
- 829 7. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry
830 continuum. *Nature* **618**, 774–781 (2023).
- 831 8. Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L. & Bonham, V. L. Lack Of
832 Diversity In Genomic Databases Is A Barrier To Translating Precision
833 Medicine Research Into Practice. *Health Aff* **37**, 780–785 (2018).
- 834 9. The All of Us Research Program Genomics Investigators. Genomic data in the
835 All of Us Research Program. *Nature* **627**, 340–346 (2024).
- 836 10. Fatumo, S. *et al.* Polygenic risk scores for disease risk prediction in Africa:
837 current challenges and future directions. *Genome Medicine* vol. 15 Preprint at
838 <https://doi.org/10.1186/s13073-023-01245-9> (2023).
- 839 11. Dokuru, D. R., Horwitz, T. B., Freis, S. M., Stallings, M. C. & Ehringer, M. A.
840 South Asia: The Missing Diverse in Diversity. *Behav Genet* **54**, 51–62 (2024).
- 841 12. Stefflova, K. *et al.* Dissecting the Within-Africa ancestry of populations of
842 African descent in the Americas. *PLoS One* **6**, (2011).

- 843 13. Harlemon, M. *et al.* A custom genotyping array reveals population-level
844 heterogeneity for the genetic risks of prostate cancer and other cancers in
845 Africa. *Cancer Res* **80**, 2956–2966 (2020).
- 846 14. Anagnostou, P. *et al.* Inter-individual genomic heterogeneity within European
847 population isolates. *PLoS One* **14**, (2019).
- 848 15. Pan, Z. & Xu, S. Population genomics of East Asian ethnic groups. *Hereditas*
849 vol. 157 Preprint at <https://doi.org/10.1186/s41065-020-00162-w> (2020).
- 850 16. Khera, A. V *et al.* Whole-Genome Sequencing to Characterize Monogenic and
851 Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial
852 Infarction. *Circulation* **139**, 1593–1602 (2019).
- 853 17. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in
854 diverse human populations. *Nat Commun* **10**, 3328 (2019).
- 855 18. Wang, M. *et al.* Validation of a Genome-Wide Polygenic Score for Coronary
856 Artery Disease in South Asians. *J Am Coll Cardiol* **76**, 703–714 (2020).
- 857 19. Ge, T. *et al.* Development and validation of a trans-ancestry polygenic risk
858 score for type 2 diabetes in diverse populations. *Genome Med* **14**, 70 (2022).
- 859 20. O'Sullivan, J. W. *et al.* Polygenic Risk Scores for Cardiovascular Disease: A
860 Scientific Statement From the American Heart Association. *Circulation* vol. 146
861 E93–E118 Preprint at <https://doi.org/10.1161/CIR.0000000000001077> (2022).
- 862 21. Lewis, C. M. & Vassos, E. Polygenic risk scores: From research tools to
863 clinical instruments. *Genome Medicine* vol. 12 Preprint at
864 <https://doi.org/10.1186/s13073-020-00742-5> (2020).
- 865 22. Xiang, R. *et al.* Recent advances in polygenic scores: translation, equitability,
866 methods and FAIR tools. *Genome Medicine* vol. 16 Preprint at
867 <https://doi.org/10.1186/s13073-024-01304-9> (2024).
- 868 23. Truong, B. *et al.* Integrative polygenic risk score improves the prediction
869 accuracy of complex traits and diseases. *Cell Genomics* **4**, (2024).
- 870 24. Wang, Y. *et al.* Polygenic prediction across populations is influenced by
871 ancestry, genetic architecture, and methodology. *Cell Genomics* **3**, (2023).
- 872 25. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic
873 prediction via Bayesian regression and continuous shrinkage priors. *Nat*
874 *Commun* **10**, 1776 (2019).
- 875 26. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger.
876 *Bioinformatics* **36**, 5424–5431 (2021).
- 877 27. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger
878 and richer datasets. *Gigascience* **4**, (2015).
- 879 28. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**,
880 68–74 (2015).
- 881 29. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of
882 admixed individuals in GWAS and to boost power. *Nat Genet* **53**, 195–204
883 (2021).
- 884 30. Sun, Q. *et al.* Improving polygenic risk prediction in admixed populations by
885 explicitly modeling ancestral-differential effects via GAUDI. *Nat Commun* **15**,
886 (2024).
- 887 31. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can
888 improve susceptibility predictions in recently admixed individuals. *Nat*
889 *Commun* **11**, (2020).
- 890 32. Xu, Y. *et al.* Effect of Pathway-Specific Polygenic Risk Scores for Alzheimer's
891 Disease (AD) on Rate of Change in Cognitive Function and AD-Related
892 Biomarkers Among Asymptomatic Individuals. *Journal of Alzheimer's Disease*
893 **94**, 1587–1605 (2023).
- 894 33. Tubbs, J. D. *et al.* Pathway-Specific Polygenic Scores Improve Cross-Ancestry
895 Prediction of Psychosis and Clinical Outcomes. Preprint at

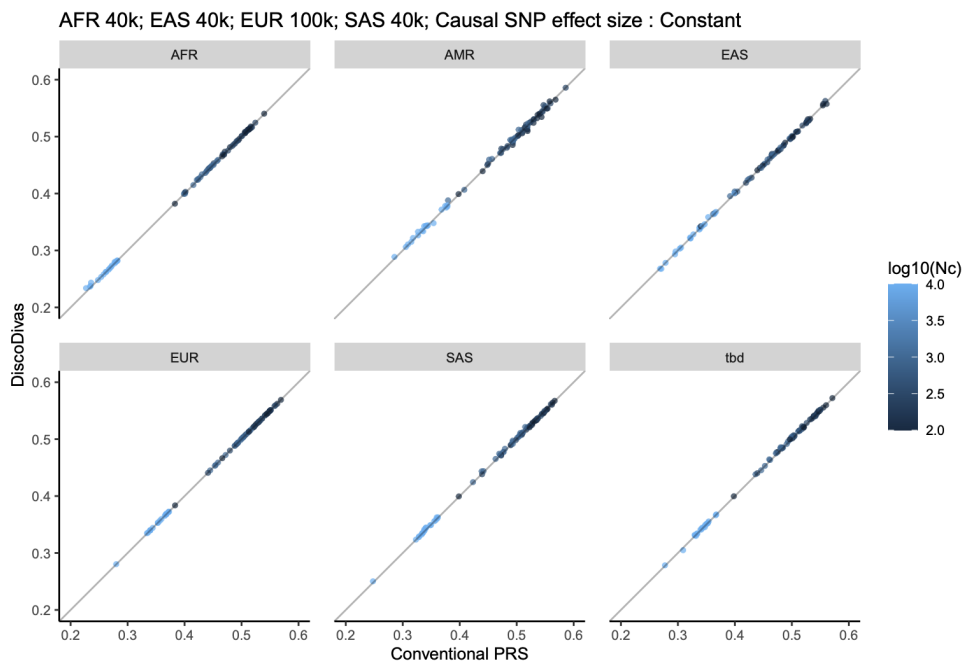
- 896 <https://doi.org/10.1101/2023.09.01.23294957> (2023).
- 897 34. Miao, J. *et al.* Quantifying portable genetic effects and improving cross-
898 ancestry genetic prediction with GWAS summary statistics. *Nat Commun* **14**,
899 (2023).
- 900 35. Kumar, R., Bhattacharya, S. & Murmu, G. Exploring Optimality of Piecewise
901 Polynomial Interpolation Functions for Lung Field Modeling in 2D Chest X-Ray
902 Images. *Front Phys* **9**, (2021).
- 903 36. Womersley, R. S. & Sloan, I. H. *How Good Can Polynomial Interpolation on*
904 *the Sphere Be? Advances in Computational Mathematics* vol. 14 (2001).
- 905 37. Choudhury, A. *et al.* Population-specific common SNPs reflect demographic
906 histories and highlight regions of genomic plasticity with functional relevance.
907 *BMC Genomics* **15**, (2014).
- 908 38. Rees, J. S., Castellano, S. & Andrés, A. M. The Genomics of Human Local
909 Adaptation. *Trends in Genetics* vol. 36 415–428 Preprint at
910 <https://doi.org/10.1016/j.tig.2020.03.006> (2020).
- 911 39. Mas-Sandoval, A., Mathieson, S. & Fumagalli, M. The genomic footprint of
912 social stratification in admixing American populations. **12**, 84429 (2023).
- 913 40. Patel, R. A. *et al.* Genetic interactions drive heterogeneity in causal variant
914 effect sizes for gene expression and complex traits. *Am J Hum Genet* **109**,
915 1286–1297 (2022).
- 916 41. Yengo, L. *et al.* A saturated map of common genetic variants associated with
917 human height. *Nature* **610**, 704–712 (2022).
- 918 42. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and
919 genomic data. *Nature* **562**, 203–209 (2018).
- 920 43. Graham, S. E. *et al.* The power of genetic diversity in genome-wide
921 association studies of lipids. *Nature* **600**, 675–679 (2021).
- 922 44. Koyama, S. *et al.* Decoding Genetics, Ancestry, and Geospatial Context for
923 Precision Health. *medRxiv* (2023) doi:10.1101/2023.10.24.23297096.
- 924 45. Kathiresan, N. *et al.* Representation of Race and Ethnicity in the
925 Contemporary US Health Cohort All of Us Research Program. *JAMA Cardiol*
926 **8**, 859–864 (2023).
- 927 46. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole
928 genomes. *Nature* **586**, 763–768 (2020).
- 929 47. Truong, B. *et al.* Integrative polygenic risk score improves the prediction
930 accuracy of complex traits and diseases. *medRxiv* (2023)
931 doi:10.1101/2023.02.21.23286110.
- 932
- 933

934 Supplementary Figures

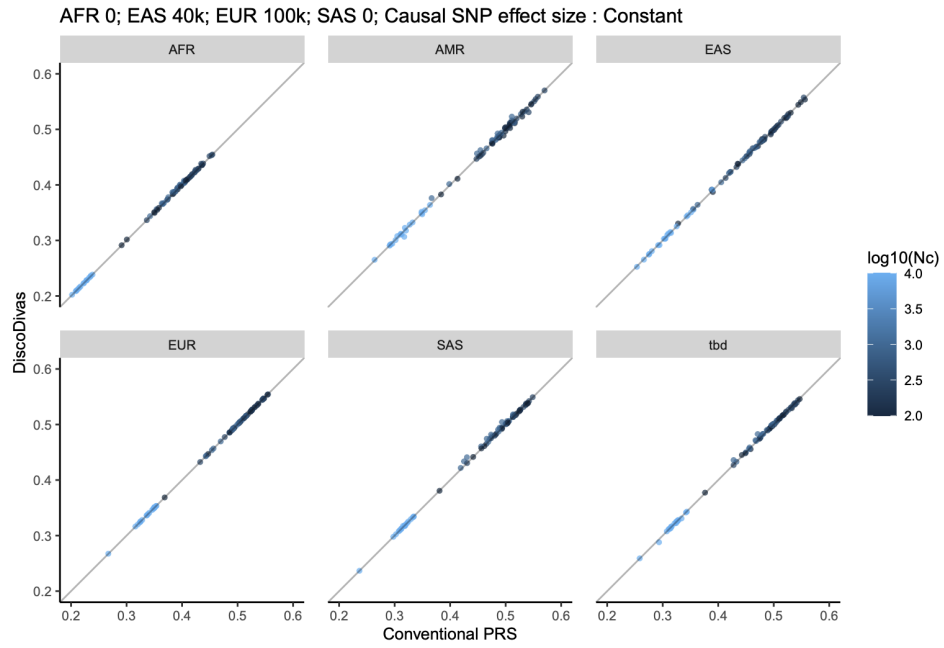
935 Supplementary Figure 1: Comparison of PRS R^2 of conventional PRS method and
936 DiscoDivas when the validation dataset for the conventional method were of the
937 matched ancestry with the testing dataset. The four subplots correspond to the four
938 simulated scenarios of different discovery GWAS sample size and causal SNP effect
939 size shown in the 4 panels in Figure 2. Within each subplot, each panel shows the
940 performance of the two methods in each testing sample; the color of the datapoints
941 showed the number of causal SNPs simulated.
942



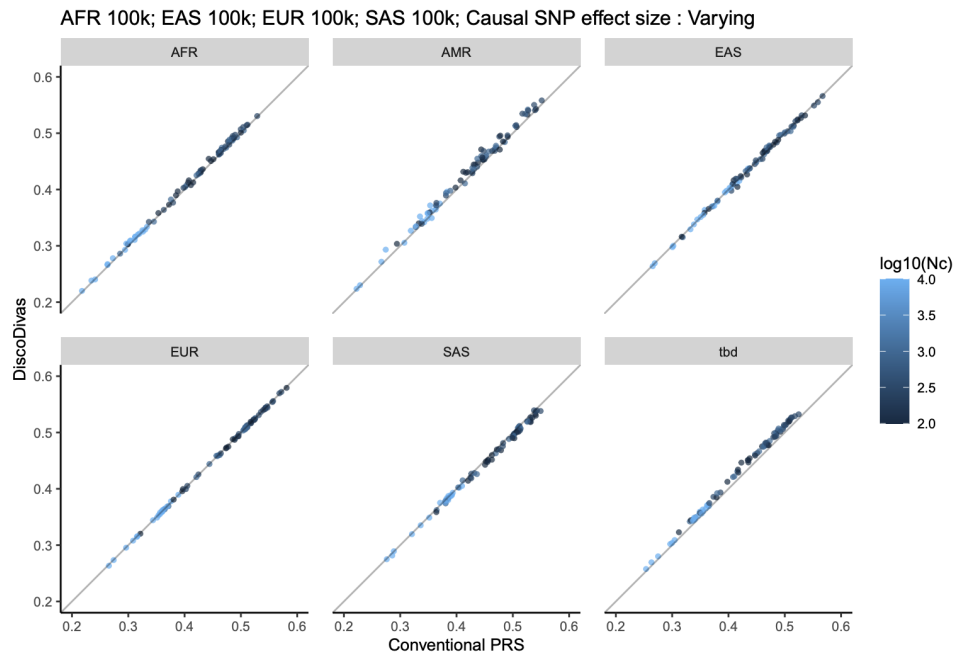
943



944

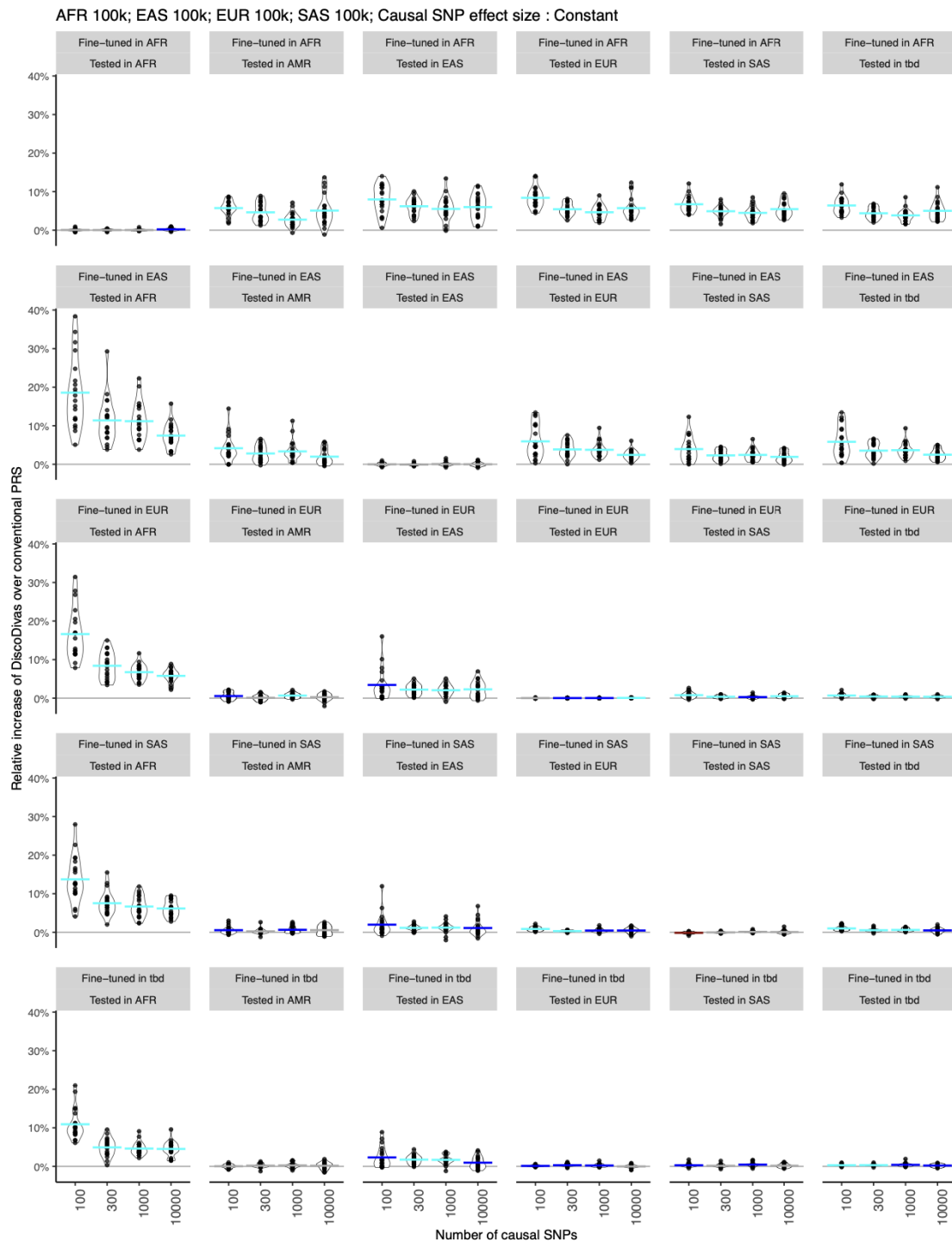


945

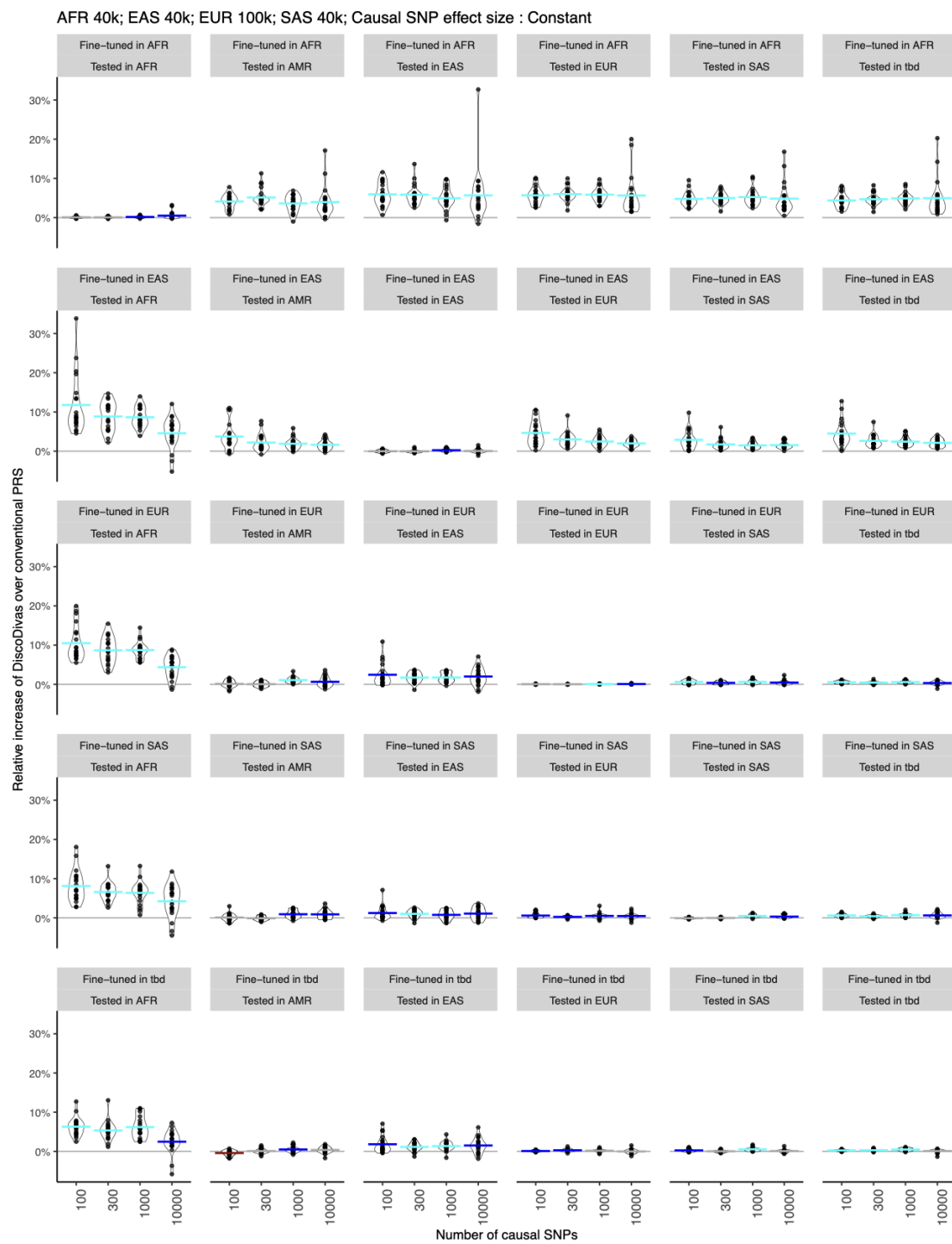


946

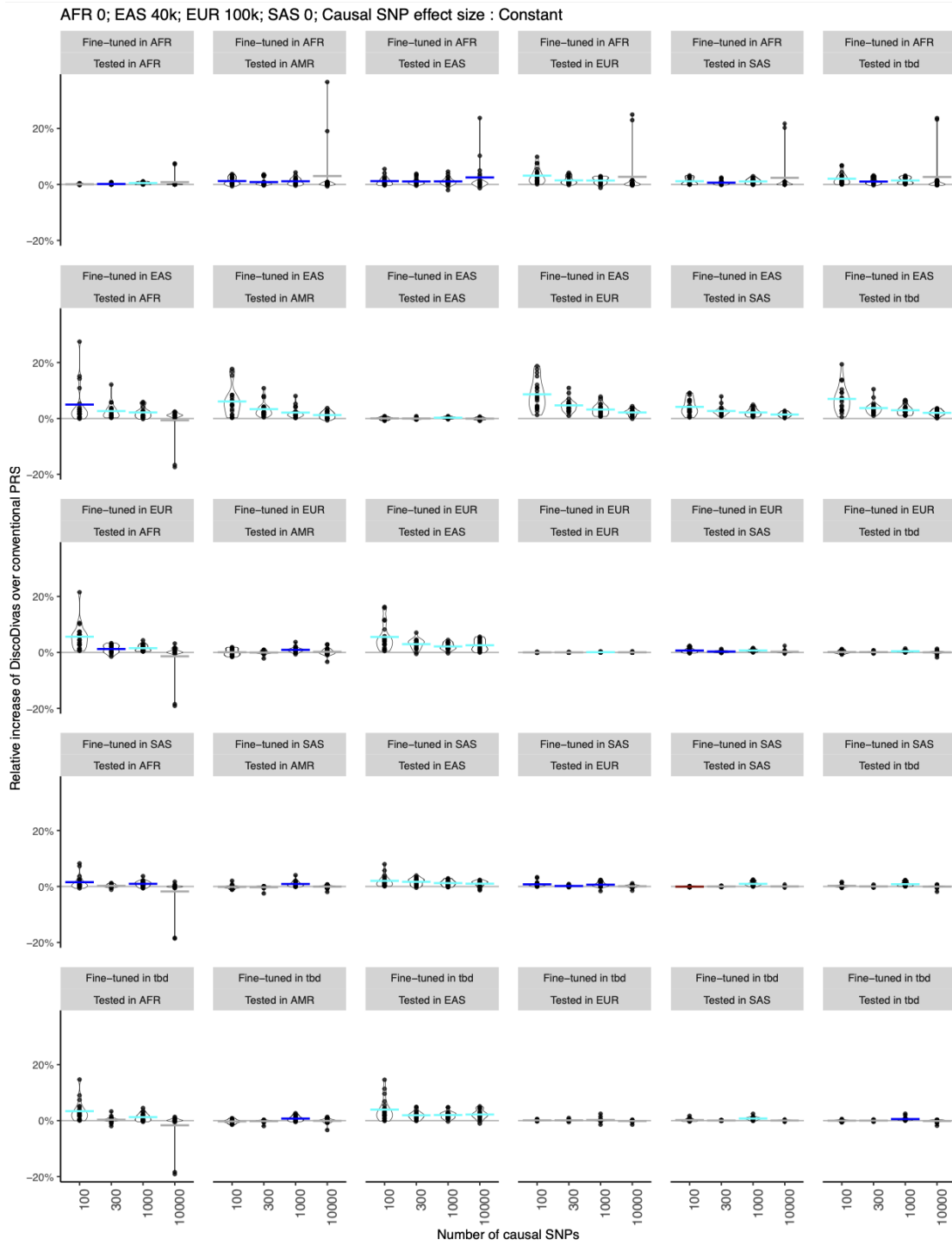
947 Supplementary Figure 2: The relative increase PRS R^2 of DiscoDivas over
 948 conventional PRS method. The four subplots correspond to the four simulated
 949 scenarios of different discovery GWAS sample size and causal SNP effect size shown
 950 in the 4 panels in Figure 2. Within each subplot, each panel shows the performance of
 951 the two methods in each combination of validation sample for the conventional PRS
 952 method and the testing sample; the horizontal bar show the mean value of the relative
 953 increase; the color of the horizontal bar indicating mean value of relative increase and
 954 p-value of the paired t-test of DiscoDivas PRS R^2 and conventional PRS R^2 , with cyan
 955 being the mean increase >0 and p-value <0.0005 , dark blue being mean increase >0
 956 and p-value <0.05 , dark red being mean increase <0 and p-value <0.05 , and grey being
 957 p-value >0.05
 958



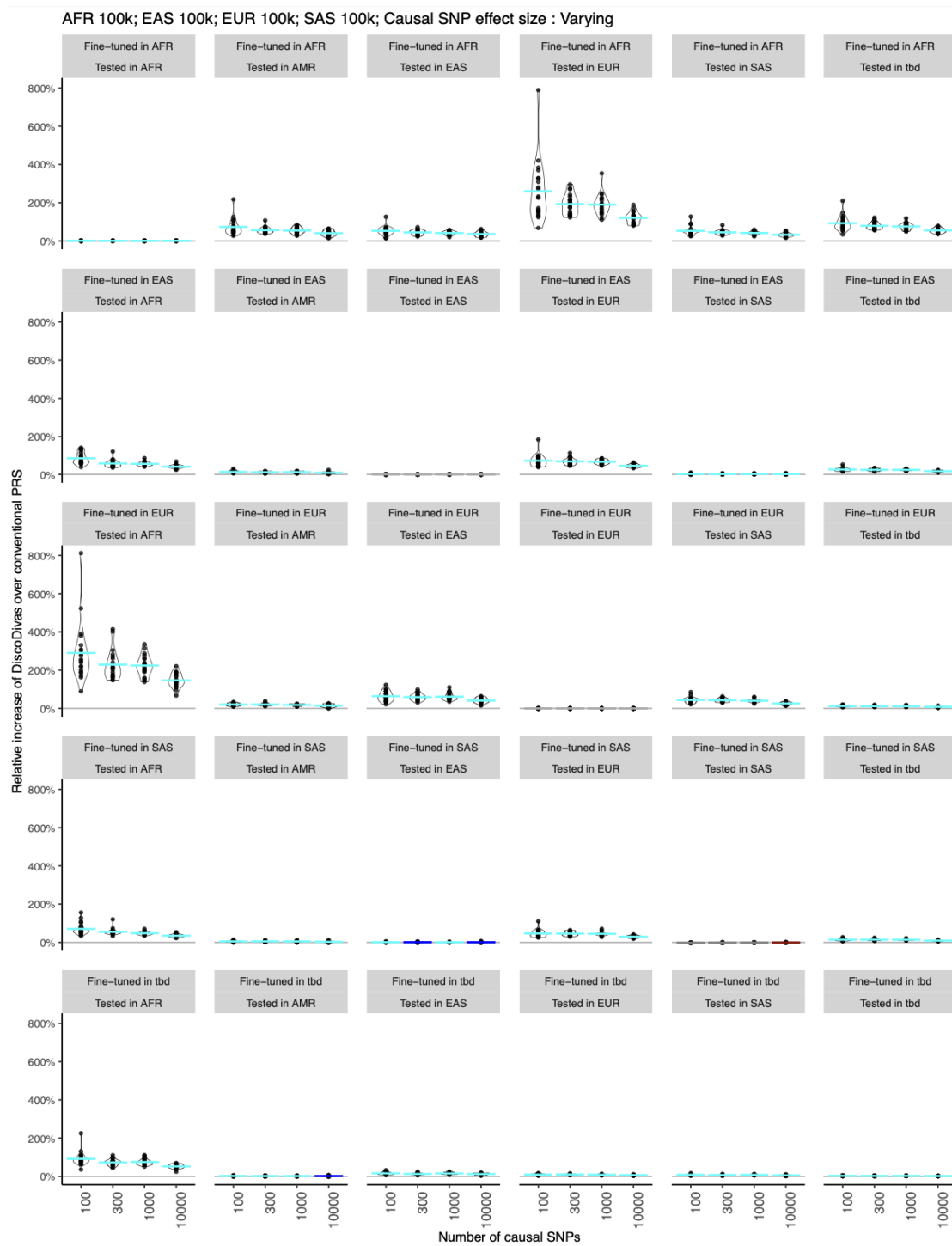
959



960
961

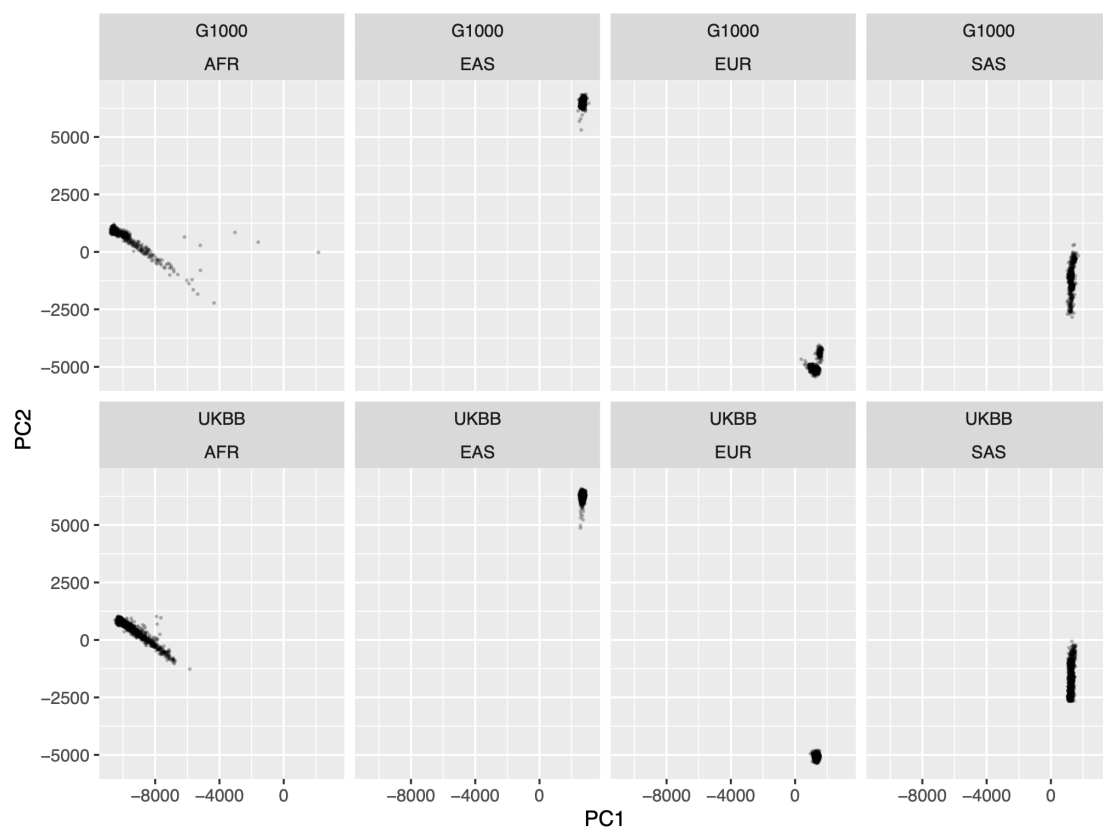


962
963

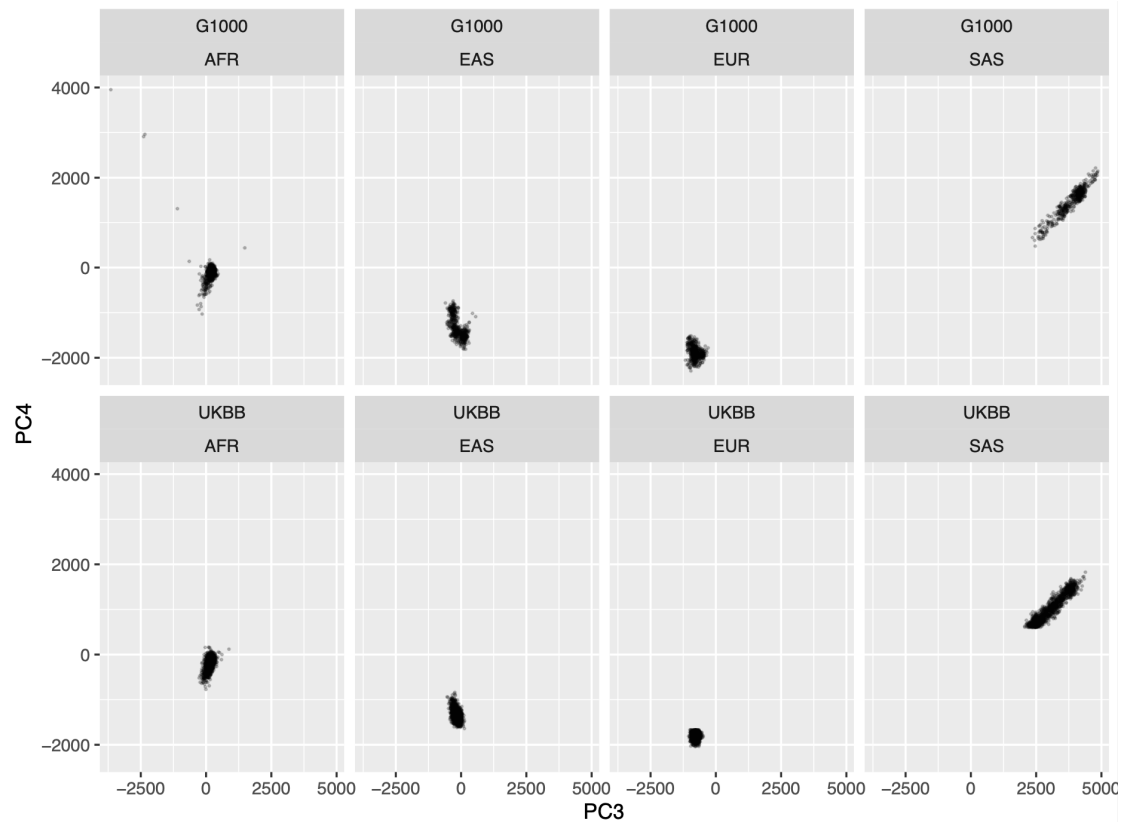


964
965
966

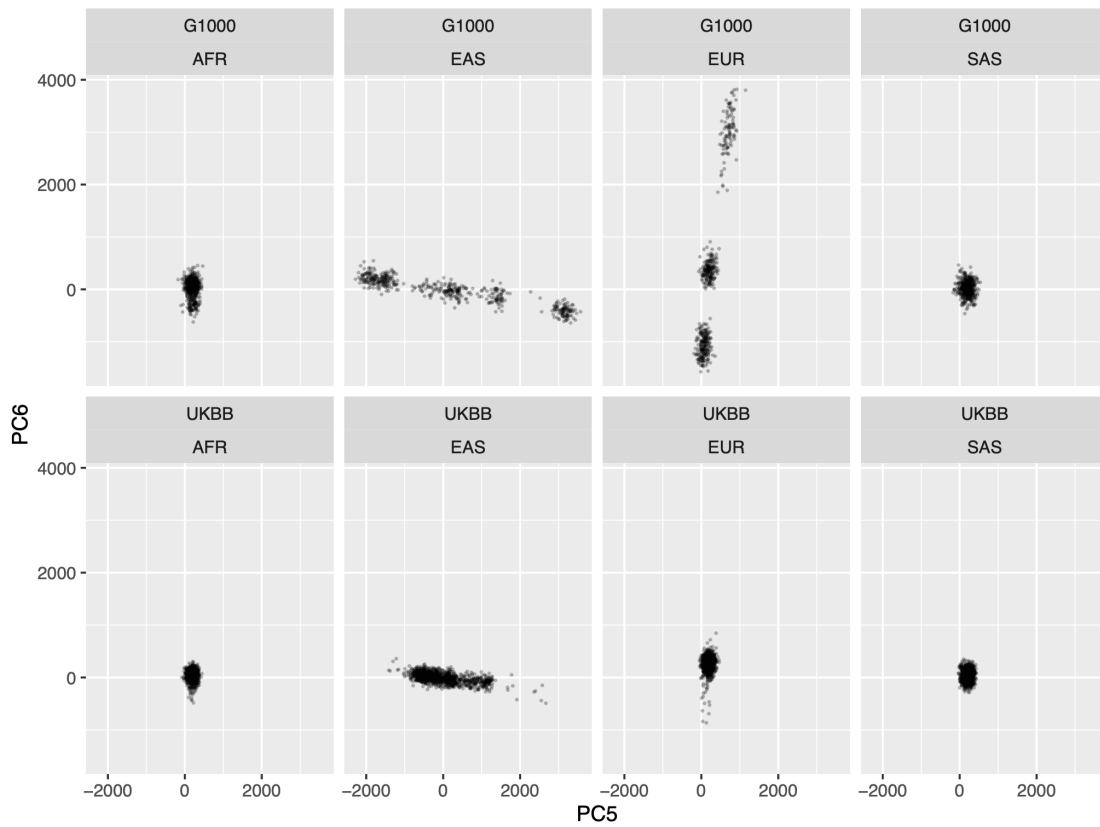
967 Supplementary Figure 3: The comparison of PCA of UKBB validation sample and the
968 1000 Genomes reference.
969



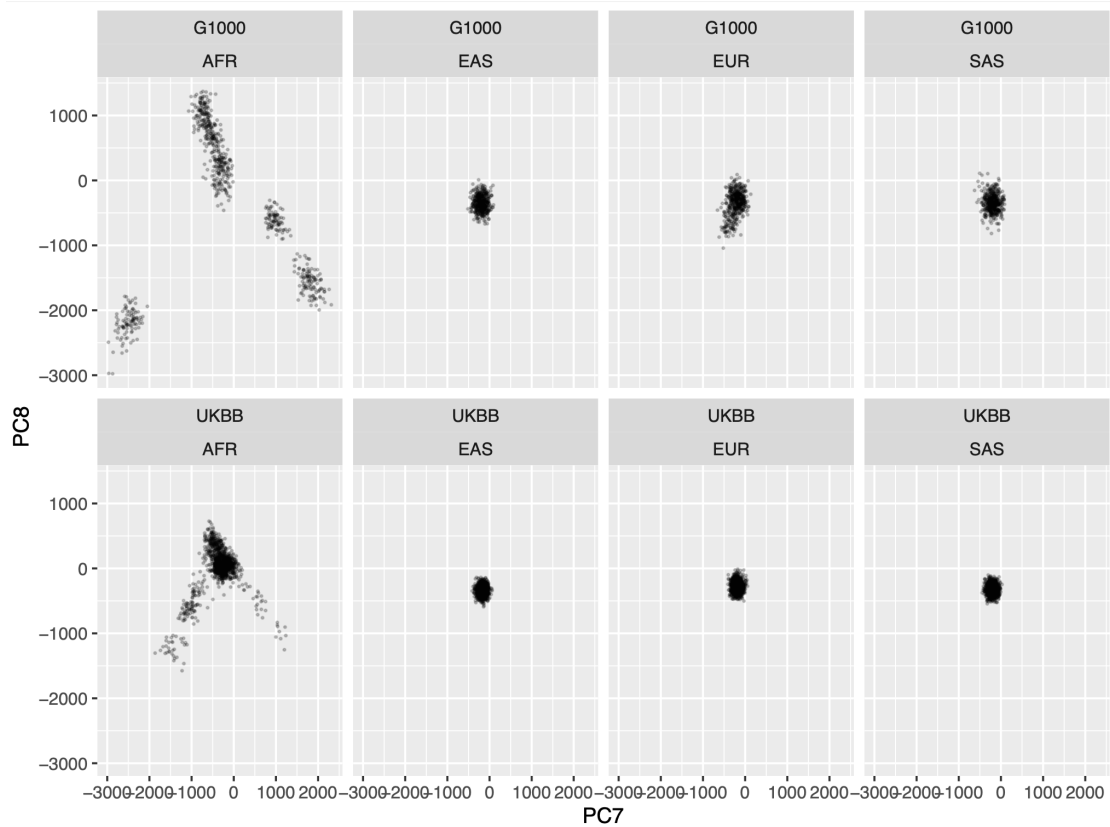
970



971

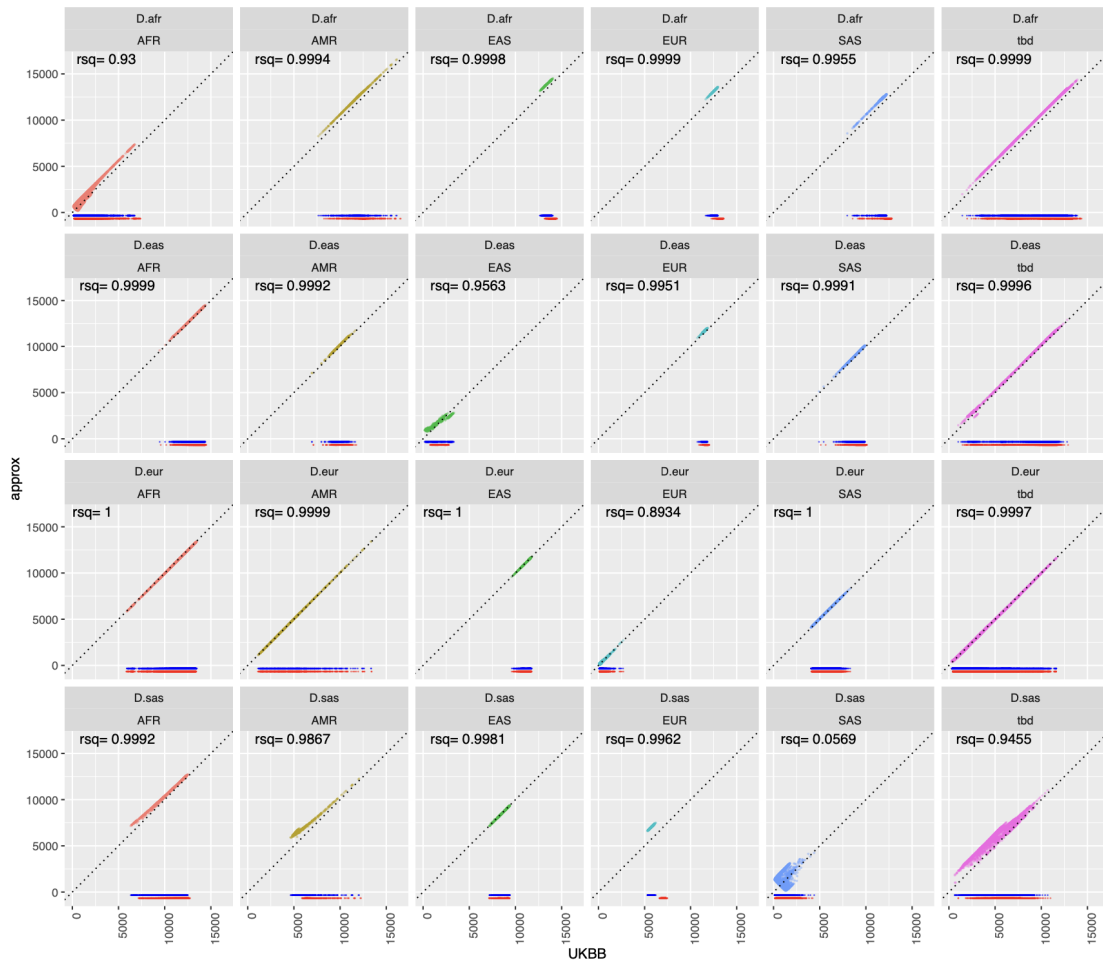


972



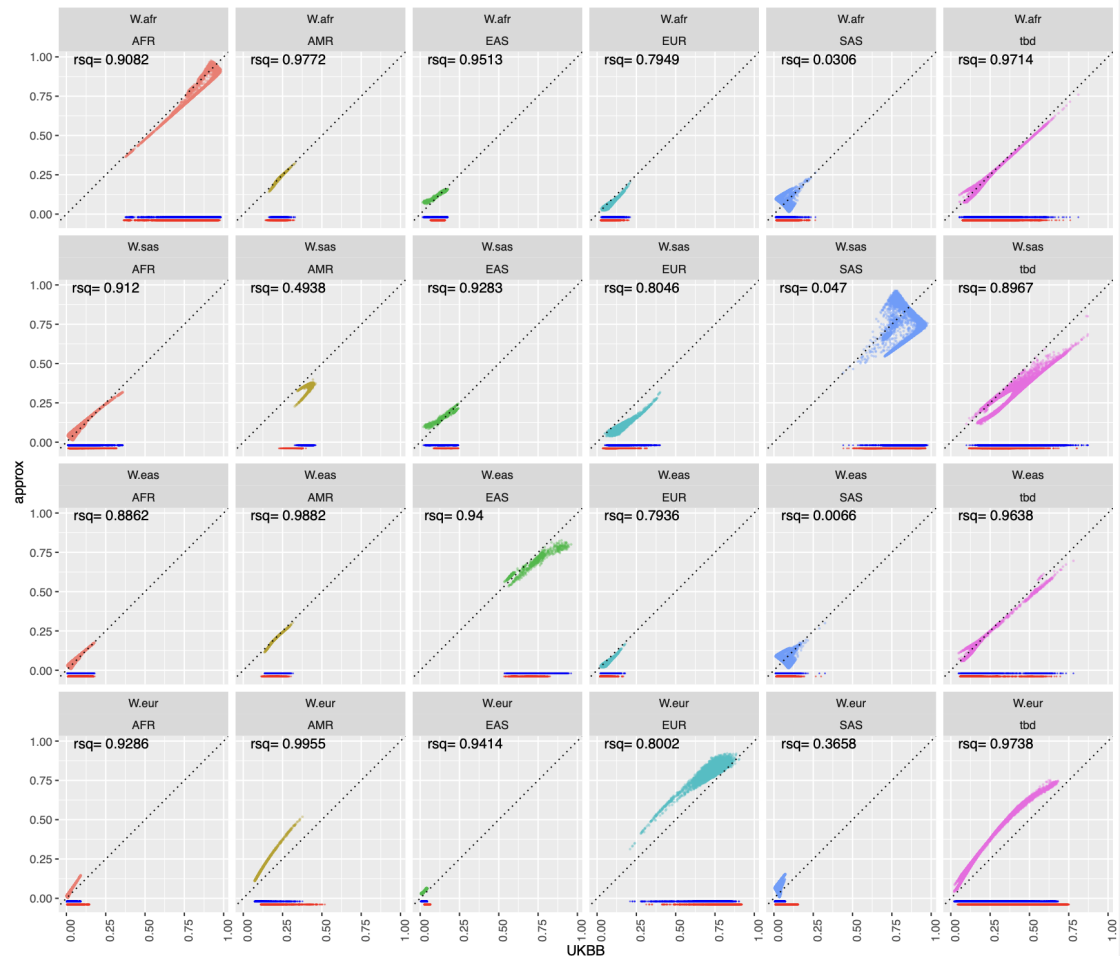
973
974

975 Supplementary Figure 4: The comparison of PCA distance between the testing
 976 individual and the median point of validation samples when using the actual value of
 977 UKBB validation samples and the approximated value of using the 1000 Genomes
 978 sample. Each column of the panels shows the ancestry of testing individuals, and each
 979 row of the panels shows the distance to the validation samples. The range of the
 980 distance to median point of the validation samples shows in the lower edge of the panel,
 981 with the blue color indicating the range of distance based on actual UKBB validation
 982 sample and the red color indicating the range of distance based on 1000 Genomes
 983 approximate. The correlation of the two sets of calculated distance is shown in each
 984 panel.



985
 986

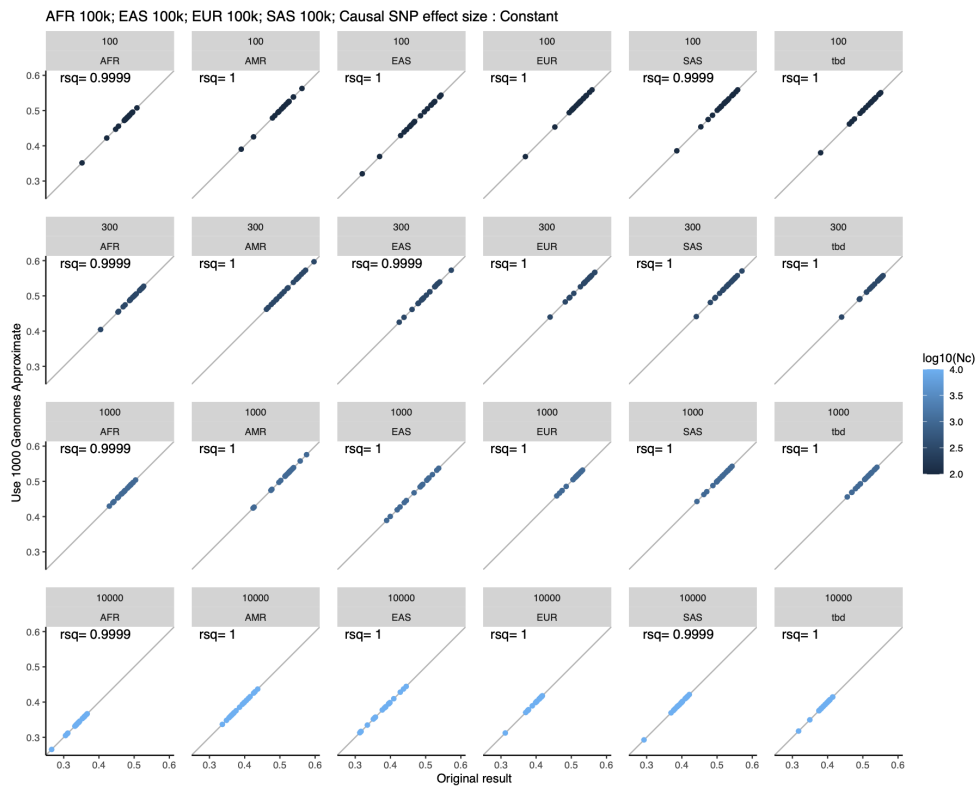
987 Supplementary Figure 5: The comparison of the interpolation combination coefficient
 988 when using the actual value of UKBB validation samples and the approximated value
 989 of using the 1000 Genomes sample. Each column of the panels shows the ancestry of
 990 testing individuals, and each row of the panels shows validation sample for which the
 991 interpolation combination coefficient is for. The range of interpolation combination
 992 coefficients shows in the lower edge of the panel, with the blue color indicating the
 993 range of combination coefficients based on actual UKBB validation sample and the red
 994 color indicating the range of combination coefficients based on 1000 Genomes
 995 approximate. The correlation of the two sets of combination coefficients is shown in
 996 each panel.
 997



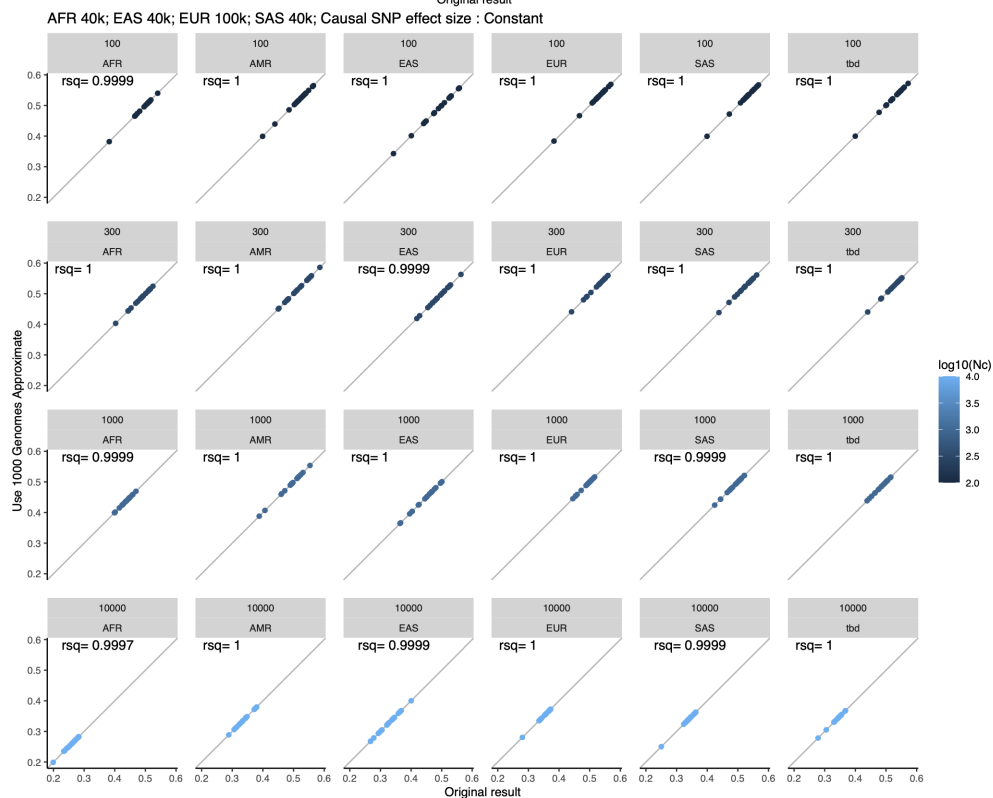
998
 999

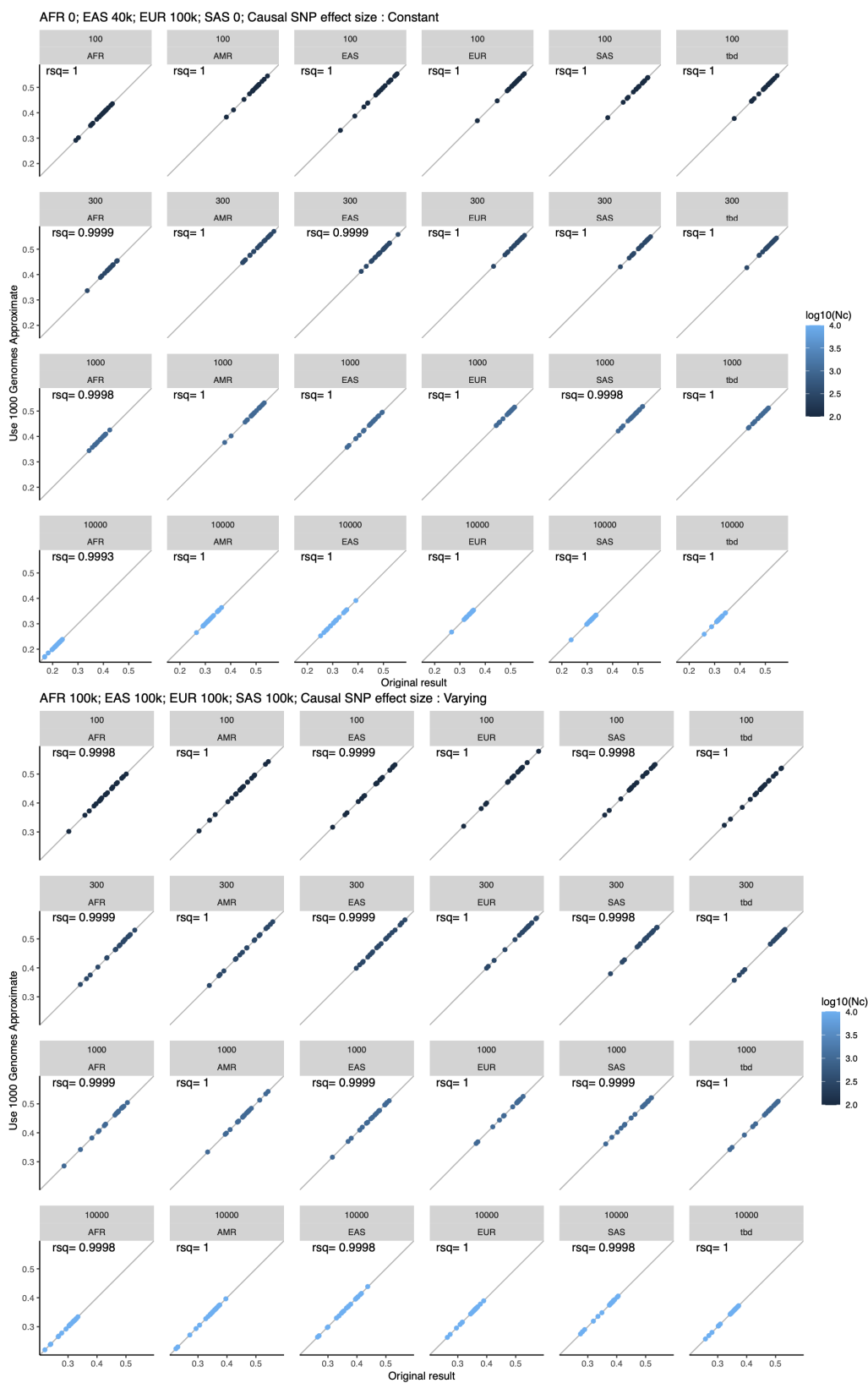
1000 Supplementary Figure 6: The comparison of DiscoDivas PRS R^2 when using the actual value of UKBB validation samples and the approximated value of using the 1000 Genomes sample. Each column of the panels shows the ancestry of testing individuals, and each row of the panels shows the simulated number of causal SNPs. The four subplots correspond to the four simulated scenarios of different discovery GWAS sample size and causal SNP effect size shown in the 4 panels in Figure 2.

1007



1008

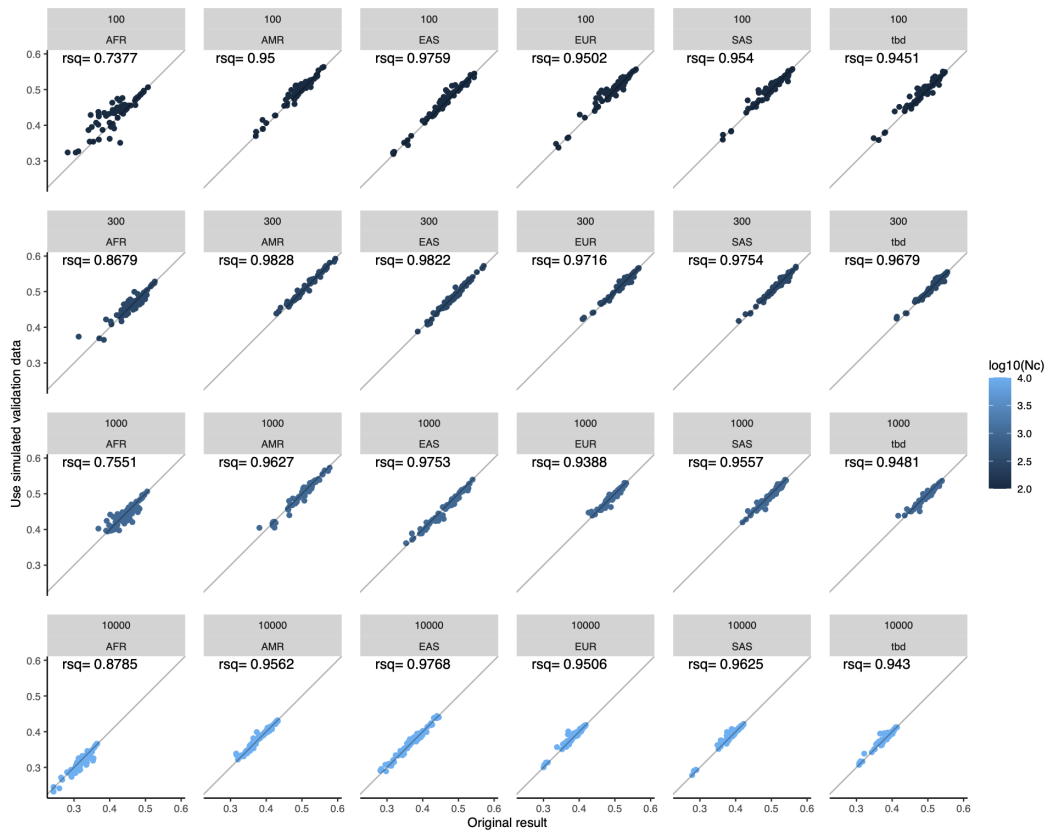




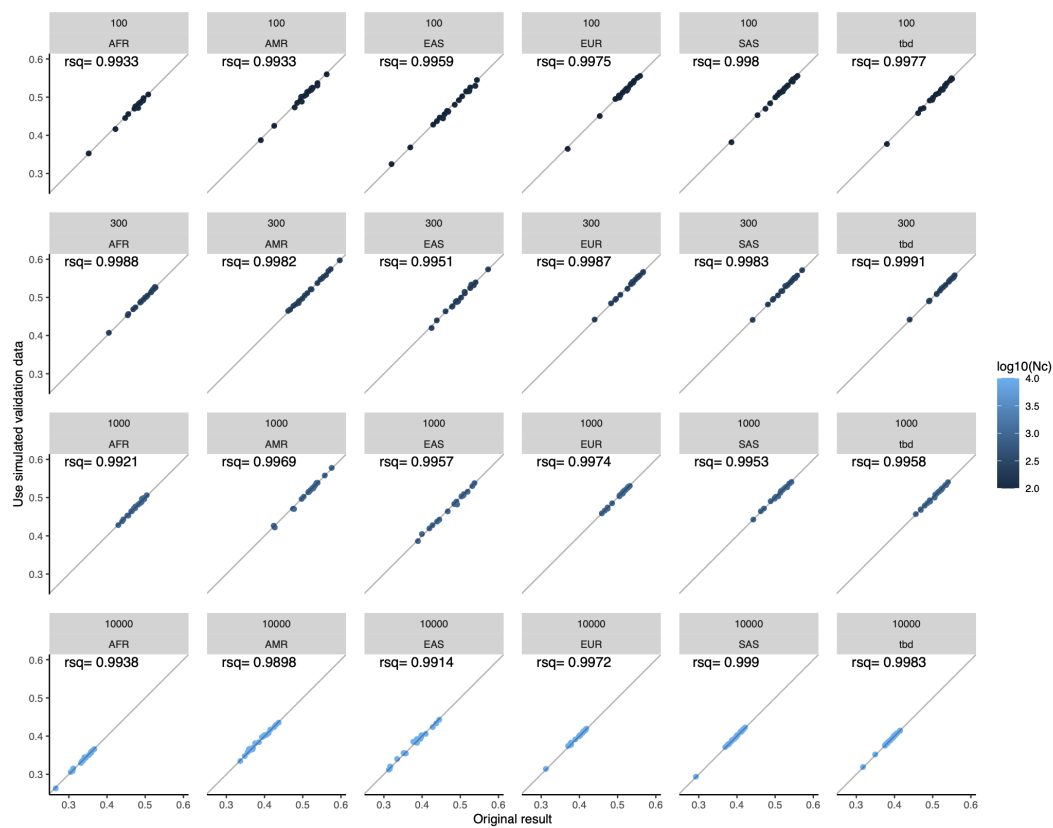
1009

1010
1011

1012 Supplementary Figure 7 Comparison of PRS R^2 of using UKBB-based validation
1013 sample and using the purely simulated validation sample. The upper subplot shows
1014 the results of conventional PRS method and the lower shows the result of
1015 DiscoDivase
1016 Within each subplot, the column of panels and the color of the datapoints shows the
1017 testing sample and the row of the panels shows the simulated number of causal
1018 SNPs;
1019

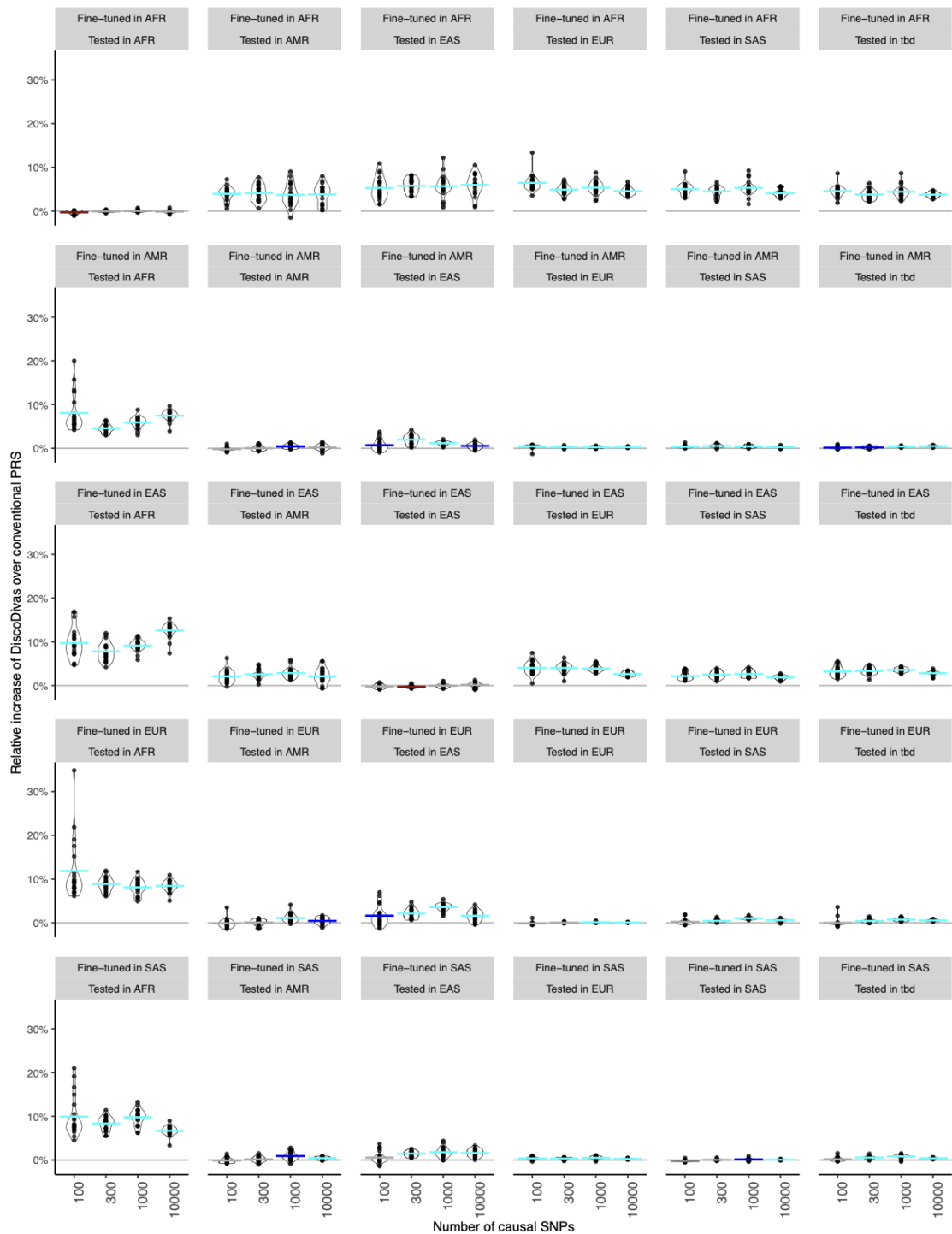


1020
1021



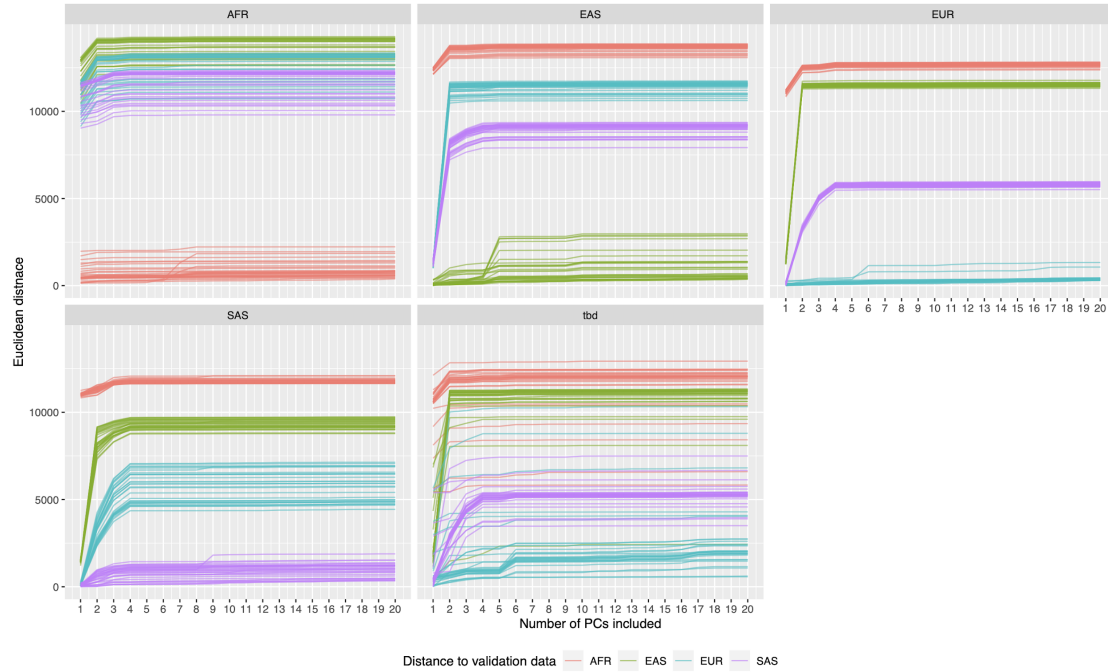
1022
1023

1024 Supplementary Figure 8: The relative increase PRS R^2 of DiscoDivas over
 1025 conventional PRS method when using the purely simulated validation data. Each panel
 1026 shows the performance of the two methods in each combination of validation sample
 1027 for the conventional PRS method and the testing sample; the horizontal bar show the
 1028 mean value of the relative increase; the color of the horizontal bar indicating mean
 1029 value of relative increase and p-value of the paired t-test of DiscoDivas PRS R^2 and
 1030 conventional PRS R^2 , with cyan being the mean increase >0 and p-value <0.0005 , dark
 1031 blue being mean increase >0 and p-value <0.05 , dark red being mean increase <0 and
 1032 p-value <0.05 , and grey being p-value >0.05
 1033



1034
 1035

1036 Supplementary Figure 9: The PRS distance between individuals in testing samples to
1037 the median point of the validation samples when including different numbers of PCA
1038 in the distance calculation. Each panel shows the ancestry of the testing samples and
1039 the color of the line shows the validation samples to which the distance is calculated.
1040 The plot is based on 100 randomly selected individuals from each UKBB testing
1041 sample.



1042
1043
1044