Causal variant capture in genotype discovery approaches drives polygenic prediction performance across traits and populations

Yi-Sian Lin[1], Taotao Tan[1,2], Ying Wang[3,4], Bogdan Pasaniuc[5], Alicia R. Martin[3,4], Elizabeth G. Atkinson[1,6*]

1. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
2. Department of Integrative Physiology, Baylor College of Medicine, Houston, TX 77030, USA
3. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
4. Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
5. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
6. Jan and Dan Duncan Neurological Research Institute, Houston, TX 77030, USA

***Corresponding author contact information**:

Baylor College of Medicine

One Baylor Plaza, Taub Research Building Room 621

Houston, TX, USA 77030

Email: elizabeth.atkinson@bcm.edu

Abstract

Polygenic scores (PGS) are widely used to estimate genetic predisposition to complex traits by aggregating the effects of common variants into a single measure. They hold promise in identifying individuals at increased risk for diseases, allowing earlier screening and interventions. Genotyping arrays, commonly used for PGS computation, are affordable and computationally efficient, while whole-genome sequencing (WGS) offers a more comprehensive view of genetic variation. In this study, we compared PGS derived from arrays and WGS across multiple traits to evaluate differences in predictive performance, portability across populations, and computational efficiency. We computed PGS for 10 traits, representing a range of heritability and polygenicity, in the three largest genetic ancestry groups in *All of Us* (European, African American, Admixed American), trained on multi-ancestry meta-analyses from the Pan-UK Biobank. Using the clumping and thresholding (C+T) method, we found that WGS-based PGS outperformed array-based PRS for highly polygenic traits but showed differentially reduced accuracy for sparse traits in certain populations. With the LD-informed PRS-CS method, we observed overall improved prediction performance compared to C+T, with WGS outperforming arrays across most non-cancer traits. The results obtained using PRS-CS closely align with those derived from pre-trained models in the PGS Catalog, with prediction achieving better performance using WGS than array genotypes for non-sparse traits. To further investigate factors influencing differential prediction performance between array and WGS, we ran simulations varying the proportions of causal SNPs directly captured by the technologies. These demonstrated that the proportion of causal variants genotyped dramatically affects prediction accuracy. Fine-mapping of empirical data supported this concept but also highlighted the importance of reducing non-informative variants for optimal prediction accuracy. In conclusion, while WGS-based PGS generally offer superior predictive power with PRS-CS, the advantage over arrays is context-dependent, varying by trait, population, and the PGS method. The ability to capture causal variants through these technologies largely

drives the prediction accuracy. This study provides insights into the complexities and potential advantages of using different genotype discovery approaches for polygenic predictions across populations and informs on strategies to enhance accuracy.

Keywords

polygenic score, genotyping array, whole-genome sequencing, biobanks, global data, causal variants

Introduction

Polygenic scores (PGS), also known as polygenic risk scores (PRS), estimate an individual's genetic predisposition to complex traits or diseases. As an emerging tool in precision medicine, PGS holds significant potential for informing early disease detection, risk stratification, prevention, and intervention[1]. PGS are calculated by aggregating the genotypes (i.e., the dosage of minor or risk alleles) from a target cohort, weighted by their corresponding effect sizes derived from relevant genome-wide association studies (GWAS) conducted in a discovery cohort[2]. The prediction accuracy of PGS is influenced by several factors, such as the sample size of the discovery cohort and the genetic similarity between the discovery and target population[3,4]. However, with up to 90% of GWAS for complex traits conducted in populations with predominantly European ancestry[5,6], PGS estimates are often less accurate for individuals of non-European ancestries, which can contribute to health disparities across populations[7].

Genotype data from both discovery and target cohorts are typically obtained using array technologies that detect known genetic polymorphisms. Since 2005, Affymetrix has developed high-throughput microarray genotyping platforms widely used in early GWAS and PGS studies[8,9], which initially assayed between 500,000 and nearly a million SNPs[10]. More recently, Illumina's Omni series arrays expanded this coverage to detect millions of genetic markers[11]. In addition to capturing genetic variation in European populations, arrays designed for global and diverse populations, such as the Global Screening Array[12], Global Diversity Array[13,14], Asian Screening Array[15], and H3 Africa Array[16,17], became available following the completion of the 1000 Genomes Project (TGP) Phase 3[18] and other diverse reference resources. To allow for more comprehensive genome-wide coverage, imputation is often employed to infer untyped variants, thereby enhancing the accuracy of GWAS and PGS predictions[19].

92 With the ongoing reduction in costs and advancements in high-throughput sequencing,
93 both whole exome sequencing (WES) and whole genome sequencing (WGS) have become
94 increasingly integrated into population-level genomic studies and clinical applications[20,21]. In
95 contrast to array-based genotyping, sequencing offers several distinct advantages. Unlike
96 genotyping arrays, which are typically designed to assay common variants and require imputation
97 to enhance coverage, deep sequencing captures all forms of genetic variants ranging from ultra-
98 rare to common alleles[22,23]. Additionally, the imputation accuracy for genotyping arrays can impact
99 the predictive performance of PGS[24], particularly in non-European populations. This is due to their
100 inability to reliably impute variants with low frequency, and variants enriched in populations that
101 are not well represented in existing reference panels. WGS enables the direct assessment of low-
102 frequency variants and loci that are missed or poorly imputed in arrays. Therefore, these
103 advantages could potentially lead to more accurate and robust PGS, especially for
104 underrepresented populations and traits influenced by rarer variants. However, WGS remains
105 more expensive, and generates vast amounts of data which requires computationally intensive
106 processing.

107 Previous studies have demonstrated comparable results between low-coverage
108 sequencing and array-based genotyping in gene discovery[25–27] and PGS for diseases like
109 Parkinson's disease[28]. However, few studies have assessed the impact of different genotype
110 discovery technologies on PGS performance at the biobank scale or considered high-coverage
111 WGS versus arrays. The *All of Us* Research Program[14] (v6), with 95,562 samples with both array
112 and WGS data, provides a unique opportunity to evaluate the predictive accuracy across
113 technologies in multiple ancestral groups and numerous phenotypes. In this study, we
114 benchmarked the performance of PGS derived from genotyping arrays, pre-curated models
115 released in the widely-used PGS catalog, and 30x high-coverage WGS across up to ten traits and
116 multiple populations in *All of Us* (Fig. 1a). Through these efforts, we aim to assess the conditions

117 under which WGS-based PGS outperforms array-based PGS and identify trade-offs that may
118 inform the choice of genomic data type for risk prediction. We also test the hypothesis that
119 differences in the proportion of causal variants captured explains the observed differences in
120 prediction accuracy across technologies, providing insights into improving polygenic prediction
121 (Fig. 1b).

122

123

124 **Results**

125 **Selection of traits across a range of genetic architectures**

126 The primary goal of this study is to benchmark the predictive performance from array-based
127 versus sequencing-based genotypes. We first selected 95,562 target samples from the *All of Us*
128 v6 dataset, each with both array and WGS genotype data. PGS are expected to exhibit higher
129 predictive accuracy when the ancestry of the discovery and target cohorts is matched[29]. However,
130 since this study aims to develop a generalized PGS model rather than focusing on a single
131 ancestry, we used effect sizes derived from multi-ancestry GWAS meta-analyses of the Pan-
132 UKBB cohort. For the target samples in *All of Us*, we focused on European (EUR), Latino/Admixed
133 American (AMR), and African/African American (AFR) cohorts based on predicted genetic
134 ancestry (Supplementary Fig. 1), as these groups had sample sizes sufficiently large to support
135 multi-trait comparative analyses. East Asian (EAS) and South Asian (SAS) populations were
136 included in some selected trait analyses if the sample size was sufficient.

137 We selected 10 complex traits with distinct genetic architectures (Supplementary Table 1) from
138 the Pan-UKBB and their equivalents in the *All of Us* dataset (Supplementary Table 2): standing
139 height (Height), diastolic blood pressure (DBP), white blood cell count (Leukocyte), red blood cell
140 count (RBC), high-density lipoprotein cholesterol (HDL), total cholesterol (TC), asthma, breast

141 cancer, colorectal cancer, and type 2 diabetes (T2D). Height exhibits the highest heritability and
142 polygenicity among the traits, followed by HDL, whereas TC and asthma are the least heritable,
143 and cancers are the sparsest traits (Supplementary Fig. 2, Supplementary Table 3).

144

145 **WGS improves prediction over arrays with LD-informed effect size shrinkage**

146 We first assessed the PGS using the classic clumping and thresholding (C+T) method
147 (Supplementary Table 4 and 5). Notably, the raw average PGS for array-based and WGS-based
148 genotypes were highly correlated across traits ($r > 0.9$ for all traits except for height and TC),
149 suggesting that both platforms captured similar genetic information, resulting in comparable
150 polygenic predictions. This also indicates that the genotype coverage from unimputed array data
151 is sufficient for constructing reliable PGS comparable to those from WGS in the *All of Us* dataset
152 for most of the traits considered. Furthermore, it is noteworthy that EUR populations tend to show
153 smaller and more homogenous raw PGS, whereas AFR populations tend to exhibit larger and
154 more widely spread PGS, particularly for colorectal cancer and breast cancer (Supplemental Fig.
155 3).

156 The predictive accuracy, measured by R^2 or Nagelkerke's R^2, was assessed across different
157 genotype discovery approaches and populations (Fig. 2, Supplementary Table 6 and 7). Given
158 that EUR individuals constitute a large proportion of both the discovery cohort (95%) and the
159 target cohort (52%), predictive accuracy in EUR was generally higher than in other populations,
160 as expected, regardless of the genotype discovery technology (ranging from $1.58 \times 10^{-3} \pm 1.28$
161 $\times 10^{-3}$ to $1.07 \times 10^{-1} \pm 6.12 \times 10^{-4}$). AFR populations generally showed lower performance
162 compared to other groups (ranging from $7.33 \times 10^{-4} \pm 7.53 \times 10^{-4}$ to $3.37 \times 10^{-2} \pm 1.07 \times 10^{-3}$,
163 Supplementary Table 8). Surprisingly, PGS performance from WGS genotypes was suboptimal

164 relative to array genotypes for multiple traits, and only outperformed arrays in all three populations
165 for one trait - height. This discrepancy may be due to height's highly polygenic nature, suggesting
166 that the higher number of variants detected by WGS provides more informative variants for PGS
167 prediction than array genotypes. Additionally, there are a few phenotypes with notable differences
168 in R^2 across genotype discovery approaches. In addition to height, prediction from WGS also
169 performed better for TC, RBC, and leukocyte count in EUR, and for leukocyte count and asthma
170 in AFR. In contrast, we observed that PGS derived from array genotypes performed better for
171 several other traits, particularly HDL in AMR (average array-based R^2=0.061; average WGS-
172 based R^2=0.048; paired t-test $p = 3.43 \times 10^{-4}$) and AFR populations (average array-based
173 R^2=0.027; average WGS-based R^2=0.009; paired t-test $p = 7.7 \times 10^{-7}$), as well as RBC in AFR,
174 T2D in AMR, and breast cancer in EUR (Supplementary Table 8 and 9). Several findings from
175 previous studies may provide insights for interpreting these observations, including aspects of
176 population-enriched architecture, as detailed in Supplementary Table 10.

177 C+T retains only quasi-independent variants within each LD block after LD clumping; however,
178 this approach may remove informative variants important for PGS calculation. In contrast, PRS-
179 CS does not rely on clumping but instead continuously shrinks effect size estimates in an LD-
180 informed manner to retain as many informative variants as possible[30]. We benchmarked C+T and
181 PRS-CS to determine whether prediction accuracy would differ between the two methods across
182 genotype discovery technologies. For raw PGS using PRS-CS, we again observed a high
183 correlation between PGS derived from array-based genotypes and WGS genotypes across traits
184 ($r > 0.9$ for continuous traits) (Supplementary Fig. 4). When comparing the predictive accuracy of
185 PGS generated using PRS-CS to those generated using the C+T method, we found that PRS-CS
186 generally outperformed C+T, with performance being approximately two times better for most
187 traits, except for cancers. Additionally, performance using WGS variants was higher than that
188 using array variants for all traits except for cancers and TC (Fig. 2).

189

190 **HapMap3 variants are sufficient for PRS-CS prediction compared with full WGS panels**

191 Note that the variants selected for PGS computation were restricted to approximately 1.1 million
192 HapMap3 variants, as the default LD panel used in PRS-CS was based on HapMap3 variants.
193 This is reflective of the general implementation of the PRS-CS software. To fully assess the
194 advantage of using WGS genotypes, however, which include around 9 million variants, we
195 additionally computed an EUR-derived LD matrix using samples from the Thousand Genomes
196 Project (TGP) and the Human Genome Diversity Project (HGDP) joint call sets, resulting in
197 approximately 7.3 million reference variants. (Supplementary Table 11). Switching from the default
198 LD panel to this new panel, we observed similar levels of predictive accuracies for traits using
199 array genotypes, but, surprisingly, lower accuracy using WGS variants in AMR and AFR for
200 multiple traits (Supplementary Fig. 5, Supplementary Table 12). This could be attributable to the
201 higher number of variants in the EUR LD panel, which does not tag well for non-EUR population.
202 However, taken together, LD information from HapMap3 variants is generally sufficient to perform
203 PRS-CS and achieve optimal performance without the computationally demanding need to
204 construct fuller LD resources.

205

206 **Pre-trained PGS show performance gains from imputation and WGS**

207 PGS methods can substantially influence predictive accuracy. We next evaluated the
208 performance of pre-trained PGS models on genotypes called from different technologies.
209 Specifically, we computed PGS using (posterior) effect size estimates reported in PGS Catalog[31]
210 studies for four traits: height[32,33], HDL[34,35], T2D[36], and breast cancer[37]. This analysis enabled
211 assessment of the generalizability of prediction performance across genotype discovery

212 approaches using a third PGS framework (Supplementary Table 13). In addition to evaluation in
213 the *AoU* cohort, we applied the same models to the UK Biobank (UKBB) cohort, allowing
214 comparison of array-based predictions both before and after genotype imputation.

215 Several patterns emerge. First, within the same discovery studies, prediction derived from WGS
216 generally outperformed those from array in the *AoU* cohort across most all traits and populations,
217 with exception of breast cancer in AMR, EAS, and SAS populations. Similarly, in UKBB, imputed
218 array genotypes consistently yielded higher predicative accuracy than non-imputed arrays across
219 all traits and populations (Figure 3 hashed bars). Second, the performance gain from non-imputed
220 to imputed array in UKBB was generally larger than the gain observed from array to WGS in *AoU*.
221 Although WGS is expected to provide a less biased and more accurate representation of genetic
222 variation, the apparent advantage of imputation may reflect the introduction of dense LD structure
223 that inflate prediction, and these results should therefore be interpretated with caution. Third, we
224 compared height PGS from two independent studies. Scores from Yengo *et al.* showed higher
225 predictive performance in UKBB than in *AoU*, whereas scores from Gunn *et al.* performed better
226 in *AoU* than in UKBB. Notably, the original PGS from Yengo *et al.* was evaluated in UKBB, while
227 the PGS from Gunn *et al.* was originally evaluated in *AoU*. This demonstrates that prediction
228 performance is higher in the cohort where tuning occurred.

229

230 **Causal variant capture shapes polygenic prediction**

231 We next investigated factors influencing differential predictive performance across genotype
232 discovery approaches. We hypothesized that prediction accuracy increases with the proportion
233 of causal variants captured by a given genotype discovery technology (Fig. 1b). To test this
234 hypothesis, we simulated polygenic scenarios in which array and WGS data captured different

235 proportions of causal variants (total number of causal variants M = 500, ~0.1% of the studied
236 genomic variants, Fig. 4a). Using *AoU* genotypes for simulation, we observed a decrease in
237 prediction R^2 as the proportion of captured causal variants decreased for both WGS (100%, 50%,
238 or 25%) and array data (100%, 75%, 50%, 25%, or 0%). This dose-response pattern remained
239 consistent under a less polygenic setting (M = 50, Fig. 4b). Surprisingly, when using UKBB
240 genotypes for simulation, reducing the proportion of causal variants used by imputed array did
241 not result in a corresponding decrease in predictive performance (Supplementary Fig. 6). This
242 discrepancy may reflect inflated performance due to introduction of dense LD structure during
243 imputation, consistent with observations from real data analyses (Fig. 3). Overall, these results
244 indicate that the proportion of causal variants captured is a key determinant of polygenic
245 prediction accuracy.

246 To test this concept in real data, we performed statistical fine-mapping using SuSiE to infer likely
247 causal variants for four traits - height, HDL, T2D, and breast cancer – using the Pan-UKBB
248 summary statistics for EUR. In total, we identified 1,185 likely causal SNPs for height, 225 for
249 HDL, 8 for T2D, and 3 for breast cancer (Supplementary Table 14-17). We then quantified the
250 proportion of those inferred causal SNPs included in the default PRS-CS framework (restricted to
251 HapMap3 variants) and in PRS-CS using the extended LD reference panel. For T2D, the
252 proportion of inferred causal variants captured by array increased from 12% using the HapMap3
253 SNP set to 38% using the full variant set, with no noticeable increase for other traits
254 (Supplementary Fig. 7a). In contrast, for WGS data, the proportion of inferred causal SNPs used
255 increased substantially when extending the variant set beyond HapMap3, indicating that WGS
256 captures a large number of likely causal variants outside the HapMap3 subset (Supplementary
257 Fig. 7b). However, capturing a higher proportion of causal variants alone did not translate into
258 improved prediction accuracy (Fig. 5), potentially due to the inclusion of additional noise from
259 uninformative SNPs in the model. These results indicate that the signal-to-noise ratio, rather than

260 the absolute number of causal variants captured, might be a key determinant of relative
261 performance between array- and WGS-based PGS. To further investigate this hypothesis, we ran
262 PRS-CS using only inferred causal variants while excluding all other variants. Unexpectedly,
263 prediction performance decreased substantially, suggesting that either many true causal variants
264 are missed by statistical fine-mapping or that additional or more common variants correlated with
265 casual SNPs (for example, through LD) are required for improved polygenic prediction.

266

267 **Trade-off between computational efficiency and prediction accuracy**

268 From a practical perspective, the choice of genotype discovery approach involves important trade-
269 offs between predictive performance, data generation cost, and computational burden. The cost
270 of genotyping arrays is around $100 per sample, whereas 30X WGS is about 5-6 times more
271 expensive. Computationally, PGS construction using array genotypes was much more efficient
272 than using WGS data. Focusing on post-QC computation with C+T, WGS required approximately
273 twice the CPU time per trait compared with array data in the *AoU* dataset (~54 vs. ~27 CPU hours)
274 for LD clumping and selection of the optimal PGS via 10-fold cross-validation. For PRS-CS,
275 inferring posterior effect sizes using WGS required 2.2-fold more CPU time than array (175.5 vs.
276 81 CPU hours) on a local server outside of *AoU*, and up to an 18-fold increase in CPU time using
277 a full-sized LD matrix (WGS: 5,616 CPU hours; array: 306 CPU hours; Supplementary Table 18).
278 Subsequent PGS computation within *AoU* environment was also more computationally expensive
279 for WGS than for array data. Together, these results highlight data generation and computational
280 cost as key considerations when selecting a genotype discovery approach for PGS analyses.

281 In conclusion, polygenic prediction performance varies across genotype discovery approaches,
282 PGS methods, traits, and populations. WGS data, when used with LD-informed methods such as

PRS-CS, generally yields higher predictive power than array data, particularly for traits with higher heritability. However, array-derived PGS remain more cost-effective and computationally efficient, making them a practical choice for rapid PGS generation and more accurate for traits with sparser genetic architectures, such as cancers. Consistent with our primary analyses, evaluation using pre-trained models from the PGS Catalog recapitulate similar performance patterns across genotype discovery technologies, while providing a computationally efficient alternative for rapid PGS assessment. However, the accuracy of pre-trained models is generally lower than that of self-trained models. Finally, our results underscore that capturing causal variants is critical for polygenic prediction, whereas the inclusion of noninformative variants negatively influence performance across genotype discovery approaches.

Discussion

In this study, we systematically benchmarked the prediction accuracy of PGS generated from array and WGS genotypes across multiple traits and populations in the *All of Us* cohort. We show that the relative performance of these approaches depends on the interplay between genotype discovery technology, PGS method, trait architecture, and ancestry. By comparing our results, particularly PRS-CS (Fig. 2), with pre-trained scores from the PGS Catalog, we demonstrate the consistency that WGS-based PGS generally outperforms array-based scores for non-sparse traits when using Bayesian methods. This suggests that these results are generalize across more advanced modeling strategies. Furthermore, through direct comparison with the paired UKBB array data before and after imputation array, we illustrate that imputation can achieve or even exceed the predictive gains of WGS relative to unimputed arrays (Fig. 3).

Using the C+T method, WGS generally did not outperform arrays. (Fig. 2). This could potentially be due to the LD clumping, which drastically reduced the number of WGS variants utilized from approximately 9 million to ~467K (~94.5% reduction), while array variants clumped down from ~1 million to ~262K (a more modest ~71% reduction) (Supplementary Table 5). The large reduction in WGS variants may have removed important variants for polygenic prediction, leading to comparable or even lower prediction accuracy than PGS derived from array variants. Given that sequencing can better capture additional important genetic variations such as INDELs and CNVs, incorporating more complex genetic variation could boost PGS performance for WGS and recover some missing heritability. Finally, sequencing error rates range from 0.1% to 1% across platforms[38,39]. However, even low error rates can introduce significant errors when large numbers of nucleotides are sequenced[40]. These errors may contribute noise to the PGS, thereby reducing predictive accuracy.

316 Using PRS-CS, however, WGS-based PGS generally performed better than array-based PGS
317 (Fig. 2). Interestingly, computing our own full LD matrix (rather than utilizing the pre-computed
318 PRS-CS LD matrix that is restricted to HapMap3 SNPs) with 6.2 M additional sites did not improve
319 WGS based results compared to the more limited variant set (Supplementary Fig. 10). These
320 results suggest that fewer, reliably captured, more common, and well-tagging variants appear to
321 suffice for PGS computation for most traits, particularly those that are less polygenic. This can be
322 illustrated with the theoretical accuracy of PGS, where the expected $R^2 \approx h^2/[1 + M/(h^2 N)]$[41].
323 With the same sample size N, using WGS over array variants might increase the SNP heritability
324 h^2, but it does not necessarily provide a benefit when a very large number of variants M are
325 involved.

326 Traditional SNP arrays have predominantly been based on genomes of European ancestry due
327 to the higher availability of EUR samples in reference databases[42]. In contrast, sequencing does
328 not rely on pre-selected markers and is expected to provide greater benefits for underrepresented
329 groups in genomic studies. Consequently, we anticipated a differentially greater improvement in
330 prediction accuracy using variants from sequencing compared to those from arrays in non-
331 European populations. However, this trend was not universally observed, suggesting that the
332 Global Diversity Array used in *All of Us* is not as biased toward European populations as some
333 other arrays are. Ongoing efforts to generate array data for diverse populations would do well to
334 similarly consider capturing variation represented in the target samples.

335 As has been demonstrated previously[3,43,44], prediction performance was generally better in EUR
336 than in other populations, likely due to the high representation of EUR in discovery GWAS.
337 Although a low proportion of Pan-UKBB, AMR achieved similar performance to EUR for some
338 traits, while AFR always had low predictive power (Fig. 2). This may be attributed to the larger
339 proportion of European ancestry in the genomes of AMR[45], and the degree of genetic divergence

340 from EUR, as has been previously explored[4]. We specifically observed a four-fold reduction in
341 performance for T2D in AFR compared to EUR in this study (Fig. 2), highlighting the ongoing
342 performance disparities of PGS for diverse populations across genotype discovery approaches
343 and PGS methods. Another consideration is that AMR and AFR are admixed populations, with
344 genomes composed of genetic segments from multiple ancestries[46]. Ancestry-informed GWAS
345 methods tailored to admixed populations, such as *Tractor* GWAS[47], may produce summary
346 statistics for polygenic prediction that function better for admixed individuals[29].

347 In the PGS Catalog analysis, we observed that imputed genotypes in UKBB confer a huge
348 advantage for prediction performance, particularly in the EUR population (Fig. 3). This is likely
349 because the UKBB imputation relied on reference panels heavily enriched for European samples
350 (e.g., HRC, 1000 Genomes, and UK10K), resulting in more accurate imputation and,
351 consequently, higher prediction accuracy in EUR individuals. Additionally, the imputation process
352 can introduce “artificial” LD, which may inflate apparent predictive gain not only in EUR but in
353 other populations.

354 In the simulation, we show a clear pattern where a higher proportion of causal variants captured
355 improved prediction performance (Fig. 4). However, in real data, increasing the number of causal
356 variants inferred by statistical fine-mapping did not necessarily lead to better performance,
357 especially when a large number of total variants were included (Fig. 5). When we included only
358 the likely causal SNPs in the PGS, in fact, performance dropped significantly. These findings may
359 suggest that many potential causal variants remain unidentified, and that reducing the number of
360 non-informative variants while increasing informative variants could improve performance.

361 It is important to note that the current analysis relied on QC’d SNPs, which were common and bi-
362 allelic. Including rare variants and other types of variants could potentially increase the number of
363 likely causal variants identified, provided that appropriate fine-mapping methodologies for these

364 variants are available. The types of variants utilized in both PGS computation and fine-mapping
365 are inherently limited by the variants available in the summary statistics. In our current analysis,
366 the full potential of WGS was restricted by the overlap with summary statistics generated from
367 imputed arrays, which typically represent common variants. If sequencing-based GWAS summary
368 statistics become more prevalent in the future, this is expected to enable broader coverage of
369 variant types in genomic analyses and further improve WGS-based PGS prediction.

370 The ongoing advancements in sequencing technology may help advance relevant method
371 development. As of 2025, the cost for genotyping arrays ranges from $90 for a screening array to
372 $150 for a global diversity array, while the cost for 30X short-read WGS is about $600. However,
373 the price of WGS has dramatically declined since next-generation sequencing technology came
374 into use in 2007[48]. With the rapid advancement of sequencing technology, WGS is expected to
375 become more affordable and potentially replace arrays as the standard platform in large-scale
376 genomic studies and personal health screening. This development is anticipated to not only
377 accelerate genomic discovery for historically underrepresented populations but also advance
378 PGS method development. The larger and more representative GWAS and reference panels
379 generated from WGS for populations with diverse genetic architectures will enhance prediction
380 accuracy for traits and diseases that have lacked statistical power. This will create many
381 opportunities for genetic prediction models tailored to specific health conditions and populations,
382 facilitating the progress of precision medicine.

383 To conclude, this study evaluates the performance of PGS using variants identified from different
384 genotype discovery technologies. The advantage of either genotype discovery approach depends
385 on several factors. Array-PGS exhibit higher predictive accuracy than WGS-PGS for sparse traits
386 when variants were clumped. In contrast, WGS and imputation outperformed array-PGS for most
387 traits with higher polygenicity when all variants were used for prediction and effect size estimates
388 were adjusted based on LD. While causal variants drive the prediction power, the inclusion of

389 non-informative variants can negatively impact performance. We hope this study provides
390 researchers with a comprehensive foundation for selecting an optimal genotype calling platform
391 and polygenic scoring method, across diverse populations and trait architectures.

392

393 ## Methods

394 **GWAS summary statistics QC**

395 The multi-ancestry meta-analyzed GWAS summary statistics for six continuous traits (height,
396 DBP, leukocyte, RBC, HDL, and TC) and four binary traits (asthma, breast cancer, colorectal
397 cancer, and T2D) were downloaded from Pan-UKBB. In brief, 500,000 UK Biobank samples
398 were genotyped and densely imputed (~97 million variants)[49]. After filtering for INFO scores >
399 0.8, approximately 30 million variants were retained and included in the meta-analysis, resulting
400 in a total of 28,987,534 variants in the final summary statistics. A summary table of the selected
401 phenotypes, including sample sizes by population, is provided in Supplementary Table 1. Each
402 summary statistic was lifted over from hg19 to hg38 to harmonize with the genome build of the
403 target genotypes in *All of Us*. Variants in all summary statistics were filtered to retain only
404 biallelic SNPs. After annotating the summary statistics to the Hail table in *All of Us*, mismatches
405 between the summary statistics and target variants were resolved by swapping reference and
406 alternate alleles and/or flipping strands, with corresponding reversal of effect size signs.
407 Ambiguous variants were excluded from further analysis. For quantitative traits, only high-quality
408 samples (denoted with the suffix "hq"), effect sizes, and *P* values were used.

409

410 **Heritability and polygenicity estimation**

411 Heritability for the 10 Pan-UKBB phenotypes was estimated using LD Score Regression

412 (LDSC)[50]. Given that EUR constitutes 95% of the total samples in the Pan UKBB GWAS meta-
413 analysis, we downloaded TGP EUR LD scores as a representative reference from LDSC. To
414 match the SNP IDs to the rsIDs in the EUR LD scores, the GWAS summary statistics were
415 annotated with dbSNP rsIDs v155 using the R package *SNPlocs.Hsapiens.dbSNP155.GRCh37*.
416 The formatted summary statistics were used to compute observed-scale heritability for
417 continuous traits and liability-scale heritability for binary traits. Sample prevalence for the four
418 binary traits was estimated using phenotype data from the UK Biobank. Polygenicity for the ten
419 traits was inferred using LDPred2-auto[51]. We used the precomputed EUR LD correlation matrix
420 provided by LDPred2 and ran the Markov Chain Monte Carlo process with default
421 hyperparameter settings.

422

423 ***All of Us* phenotype curation**

424 For the 10 phenotypes with GWAS summary statistics available from Pan-UKBB, we curated
425 the corresponding phenotypes in the *All of Us* following the methodology outlined in Tsuo *et al*[52].
426 We curated the quantitative traits based on concept IDs, and binary traits using PheCode map
427 v1.2, which facilitate the mapping of ICD codes into their corresponding phecodes. For
428 quantitative phenotypes, we harmonized their units by adopting the most frequently observed
429 unit as the reference standard. We further removed those individuals with values exceeding 5
430 standard deviations from the mean. Subsequently, we performed inverse-rank normal
431 transformation to ensure normalization.

432

433 **Target sample selection**

434 Samples with both genotyped array and WGS data were selected. In total, 95,562 individuals

435 had both types of data in *All of Us* v6. Samples were excluded if more than 10% of genotypes
436 were missing. Related samples were removed based on a kinship score > 0.1. Samples
437 predicted to be of EUR (European, composed of Finnish and non-Finnish European), AMR
438 (Latino/admixed American), or AFR (African) ancestry were selected based on an auxiliary
439 genetically predicted ancestry file provided in *All of Us*. Briefly, the ancestry labels were inferred
440 using the prediction probabilities from a random forest classifier, which was trained on the 16-
441 dimenional PCA space of reference HGDP and TGP samples and autosomal variants, which
442 obtained from gnomAD[14]. A summary table of sample sizes for the 10 curated phenotypes is
443 provided in Supplementary Table 2.

444

445 **Target variant QC**

446 The QC criteria were identical for both array and WGS variants in *All of Us* v6. First, only
447 autosomal variants were retained. Variants with a call rate lower than 90% were excluded.
448 Indels were removed, and only common biallelic SNPs with a MAF > 0.01 were retained. QC of
449 the 1,824,517 array variants resulted in 975,876 SNPs, while QC of the 702,574,937 WGS
450 variants resulted in 8,996,707 SNPs (Supplementary Table 4). We did not impute array variants
451 in *All of Us* due to the high computational cost associated with cloud-based analysis. *All of Us*
452 genotype data is only released unimputed, such that this is expected to be standard practice for
453 users of this particular dataset. *All of Us* array data is genotyped with the Global Diversity Array,
454 which was designed to integrate highly optimized multiethnic variants with a genome-wide
455 scaffold, incorporating WGS data from over 1,000 individuals of African and/or American
456 ancestry, reflecting the genetic diversity of participants in the *All of Us* Research Program. As
457 such, comparing polygenic prediction performance across ancestral groups using array and
458 WGS data from *All of Us* is particularly relevant.

459

460 **C+T construction**

461 For C+T, QC'd array and WGS variants were first restricted to those overlapping with QC'd
462 GWAS summary statistics to maximize variant overlap. The counts of intersecting array or WGS
463 variants with the discovery GWAS are presented in Supplementary Table 4. LD was estimated
464 using 3,202 unrelated samples from five continental populations in the TGP Phase 3 as the
465 reference panel; first- and second-degree relatives were excluded. Reference variants were
466 filtered to retain common biallelic SNPs (MAF > 0.01). LD clumping was performed using PLINK
467 to retain approximately independent variants (--clump-p1 1, --clump-r2 0.1, and --clump-kb 250).
468 For each trait, clumped variants were subsequently filtered across a range of GWAS *p*-value
469 thresholds (1, 0.5, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 1e-8), constituting the
470 thresholding step of the C+T procedure. The number of retained variants after clumping is
471 reported in Supplementary Table 5. PGS were calculated for each individual as:

472 $$PGS_i = \sum_{j=1}^{M} x_{ij}\beta_j$$

473 where X represents the dosage of alternate alleles {0, 1, 2} from the genotype matrix, β denotes
474 the effect size estimates, i indexes the individual, j indexes the variant, and M is the total number
475 of variants used for the calculation (i.e. number of clumped SNPs). Optimal *p*-value thresholds
476 were selected separately for each population using 10-fold cross-validation (1 fold for validation
477 and 9 folds for testing). Model performance was assessed using R^2 for continuous traits and
478 Nagelkerke's R^2 for binary traits. R^2 is the proportion of variance explained by the full model
479 relative to the null model, calculated using linear regression. Nagelkerke's R^2 was computed via
480 logistic regression by comparing the likelihoods of the null and full models, adjusting for the overall
481 fit to obtain a value indicating the proportion of variance explained by the PGS in the context of

482 the full model. The null model included age, male sex, female sex, and the first 16 principal
483 components as predictors, while the full model included all variables from the null model in
484 addition to PGS.

485

486 **PRS-CS construction**

487 We prepared the LD reference panel from two sources: (1) A precomputed EUR LD panel
488 constructed using UKBB data, restricted to 1,120,696 HapMap3 variants, which was
489 downloaded from the PRS-CS GitHub repository. (2) A genome-wide EUR LD panel
490 constructed using the TGP and HGDP joint call set[53]. To generate this LD panel, we
491 downloaded precomputed EUR LD blocks based on the GRCh38 genome from
492 LDblocks_GRCh38[54]. We then computed the LD matrices for each LD block using PLINK, with
493 a total of 7,330,169 post-QC (bi-allelic SNPs with MAF >0.5%) genome-wide variants from the
494 VCF of 664 unrelated EUR samples in the TGP and HGDP joint call set. The LD matrices
495 across 22 chromosomes were saved in HDF5 format, which is compatible with PRS-CS. To run
496 PRS-CS, we did not set the global shrinkage parameter phi and allowed the program to
497 automatically search for the optimal value. After obtaining posterior effect sizes generated from
498 PRS-CS-auto, we computed PGS for each target individual by aggregating the product of their
499 genotypes weighted by the updated effect sizes, as shown in the equation below:

500 $$PGS_i = \sum_{j=1}^{M} x_{ij}\beta_{j,posterior}$$

501 PGS performance was then evaluated using R^2 for continuous traits and Nagelkerke's R^2 for
502 binary traits, as described in the C+T section.

503

504 **Evaluation of prediction performances using PGS Catalog**

505 To evaluate the PGS performance in the *AoU* and UKBB cohorts, we downloaded publicly
506 available scoring files from the PGS Catalog for 4 traits: Height, HDL cholesterol, Type 2 Diabetes,
507 and breast cancer. Each scoring file included SNP ID, effect allele, and corresponding effect size
508 estimates derived from the published PGS models. The PGS studies selected were prioritized to
509 those trained the model on multi-ancestry datasets and evaluated across multiple populations.
510 Details of the selected scoring files are provided in Supplementary Table 13. Variant QC criteria
511 for *AoU* data have been described previously. For UKBB array and imputed genotype, variants
512 were restricted to bi-allelic SNPs with MAF > 0.005. Duplicate and ambiguous SNPs were
513 removed, and alleles were harmonized between scoring files and the target genotypes, including
514 strand and sign flipping where necessary. Phenotype curation for *AoU* was described previously.
515 For UKBB, continuous traits were curated using the first data instance and values were converted
516 with inverse rank normal transformation for each population groups (EUR, AMR, AFR, East Asian
517 (EAS), and Central South Asian (CSA)). Binary disease traits were defined based on primary and
518 secondary ICD-10 diagnosis codes, with individuals having at least one relevant code classified
519 as cases and those without any diagnosis classified as controls. Polygenic scores were computed
520 using PLINK to aggregate the weighted genotype dosage, where weights corresponded to effect
521 size estimates from the PGS Catalog and genotypes were from *AoU*'s array and WGS data and
522 from UKBB's array and imputation data. Prediction performance was assessed with either R^2 or
523 Negelkerke R^2, adjusting for age, sex, and 16 PCs.

524

525 **Simulation**

526 We simulated phenotypes based on real genotype data from the *AoU* (array and WGS) and UKBB
527 (array and imputation). For *AoU*, we randomly selected 50,000 unrelated EUR samples. To
528 ensure comparable genotypes between array and WGS, we extracted QC'd chromosome 22

529 variants from the short-read WGS (srWGS) common Allele Count/Allele Frequency (ACAF)
530 threshold callset (*AoU* release v7) to created two datasets: a full WGS variant set and an array-
531 matched subset obtained by restricting WGS variants to those overlapping array SNPs.
532 Phenotypes and effect sizes were simulated using GCTA[55]. Continuous phenotypes were
533 simulated under 10 scenarios, each with 10 replicates, where each scenario differed in the
534 proportion of causal variants present in array vs. WGS data (ranging from 100% to 0%) (Fig. 1b).
535 Causal SNPs were randomly assigned to variants with allele frequencies between 0.005 and 0.05
536 across all scenarios. Two levels of genetic architecture were considered: 500 causal SNPs to
537 represent polygenic traits (~0.1% of the studied genomic variants) and 50 causal variants to
538 represent sparse traits (~0.01% of the studied genomic variants). Effect sizes were estimated by
539 conducting GWAS on 80% of the samples with PGS computed on the remaining 20%. Ten PCs
540 were generated for the discovery and target sets and included as covariates in GWAS and PGS
541 analyses. For UKBB, we similarly selected 50,000 unrelated EUR samples and extracted
542 chromosome 22 variants from the QC'd imputation data, subsequently generating an array-
543 matched subset by restricting variants to those overlapping with *AoU* array SNPs to facilitate
544 comparison. All other simulation and analysis procedures were identical to those applied to *AoU*
545 data. PGS were constructed using C+T, since PRS-CS is limited to HapMap3 variants, which
546 reduces the total number of SNPs. Predictive performance was evaluated using R^2.
547

548 **Statistical fine-mapping**

549 To identify likely causal variants in real data, we performed statistical fine-mapping using
550 SuSiE[56](susieR package v0.15.2), a Bayesian regression framework that models multiple causal
551 variants within a locus. Pan-UKBB summary statistics (height, HDL, T2D, and breast cancer) and
552 a precomputed UKBB EUR LD matrices were used as inputs. For each genome-wide significant
553 locus (p < 5e-8), SuSiE was applied assuming up to 10 causal variants. Variants within 95%

554 credible sets were prioritized, and those with posterior inclusion probabilities (PIPs) > 0.9 were
555 considered likely causal for downstream analyses.
556

557 **Declaration of interests**

558 A.R.M. has received speaker fees from Novartis. Other authors declare no competing interests.

559

560 **Acknowledgements**

561 We gratefully acknowledge *All of Us* participants for their contributions, without whom this
562 research would not have been possible. We also thank the National Institutes of Health's *All of*
563 *Us* Research Program for making available the participant genotype and phenotype data
564 examined in this study. We thank Jessica Honorato Mauer for her assistance with QC of the UKBB
565 imputed data and Aishi Ayyanathan for her help in generating the QC'd sample list used in this
566 study. This research has been conducted using the UK Biobank Resource under Application
567 Number 95179. This work was supported by NIH grants K01MH121659 and R01HG012869 to
568 E.G.A.

569

570 **Author contributions**

571 Y-S.L. analyzed the data, interpreted the results, and drafted the manuscript. T.T. analyzed the
572 data and interpreted the results. Y.W. curated *All of Us* phenotypes. B.P. and A.R.M. provided
573 critical comments. E.G.A designed and supervised the study and revised the manuscript. All
574 authors approved the final manuscript.

575

576 **Web resources**

Pan-UKBB Downloads page, https://pan.ukbb.broadinstitute.org/downloads

EUR LD scores, https://zenodo.org/records/8182036

TGP Phase 3 reference panels, https://www.cog-genomics.org/plink/2.0/resources

PRS-CS, https://github.com/getian107/PRScs

PGS Catalog, https://www.pgscatalog.org/

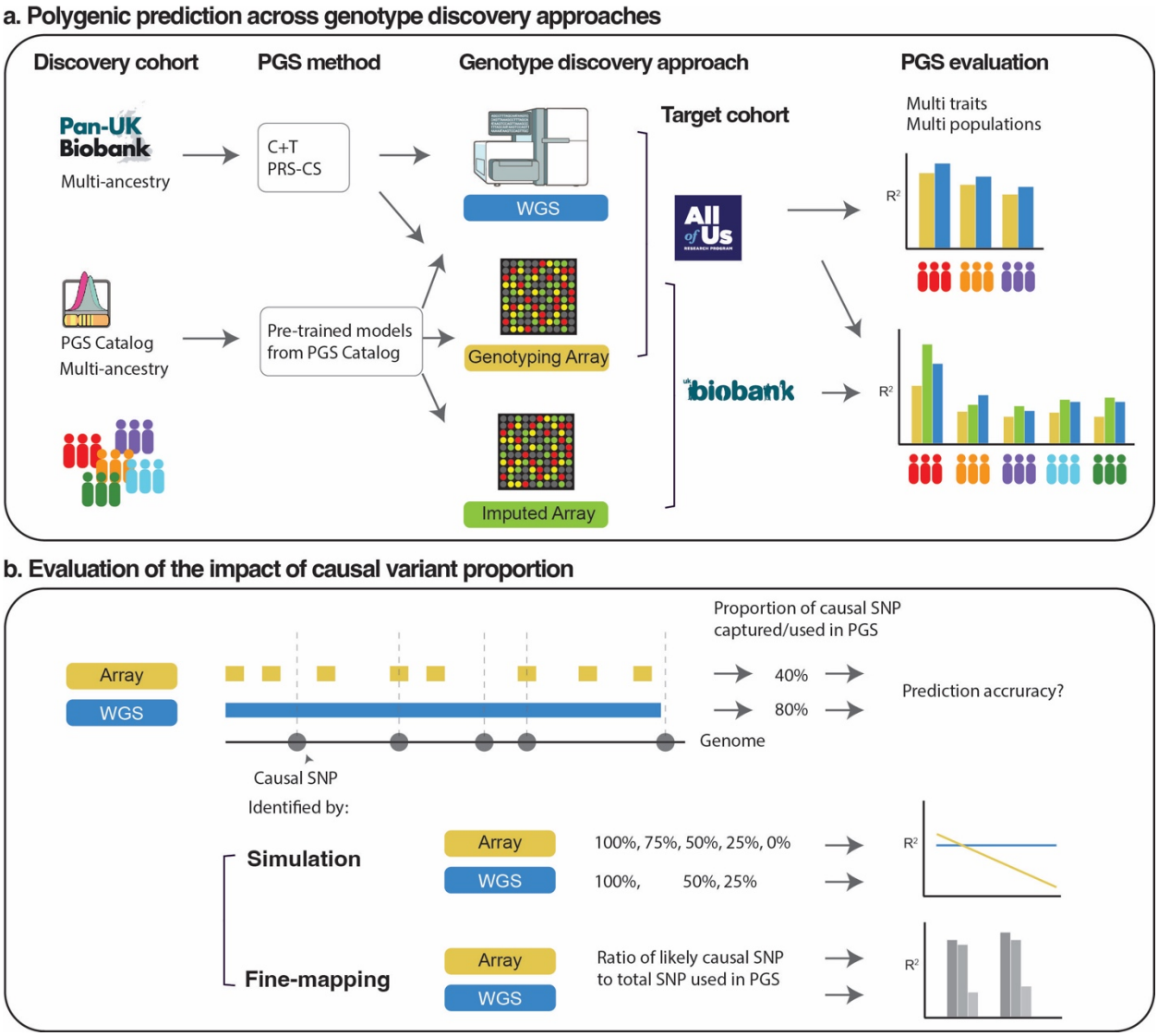
susieR, https://stephenslab.github.io/susieR/

UKBB LD reference, https://uchicago.app.box.com/s/jqocacd2fulskmhoqnasrknbt59x3xkn

Data and code availability

The GWAS summary statistics used in this study were downloaded from the Pan UKBB Downloads page. Individual genotype data were obtained from the *All of Us* Research Program Controlled Tier Dataset v6. EUR LD scores from TGP, used for heritability estimation, were downloaded from the zenodo web resources describe above. The TGP Phase 3 reference panels used for C+T were downloaded from PLINK resources. The EUR reference LD panel, constructed using UKBB data for PRS-CS, was downloaded from the PRS-CS GitHub repository. PGS Catalog scoring files were downloaded from the link provided in Supplementary Table 13. The UKBB LD reference were downloaded from the UChicagoBox. All code used for this analysis is available at https://github.com/Atkinson-Lab/PGS_Array_WGS.

602 **Figures**



603

604 **Fig. 1: Study design.**

605 In the phase (a), PGS were calculated for each *All of Us* individual in the EUR, AMR, or AFR
606 populations using genotypes derived from either array or WGS approaches. C+T and PRS-CS
607 methods were benchmarked. Pre-trained models from the PGS Catalog were used to evaluate
608 performance in *All of Us* and UKBB individuals across multiple populations, using genotypes
609 derived from direct genotyping, imputation, or WGS. In the phase (b), prediction performance was

evaluated based on different proportions of causal variants captured by the genotype discovery approach, from simulation and fine-mapping.

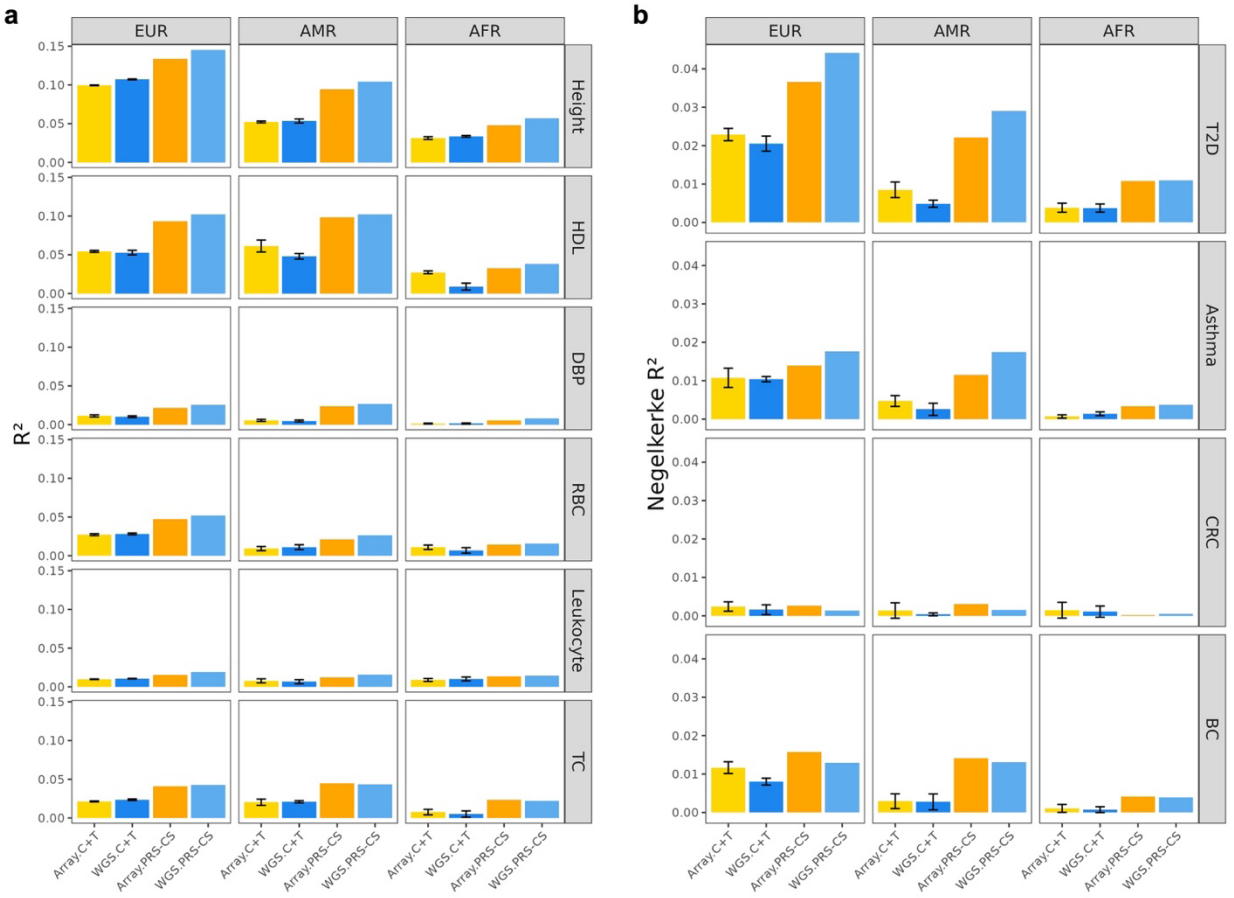


Fig. 2: Comparison of PGS performance across genotype discovery approaches.

PGS were computed for samples with genotypes derived from arrays (yellow, orange) or WGS (blues) in *All of Us*. Continuous traits (a) and binary traits (b) are ordered by polygenicity. R^2 or Nagelkerke's R^2 were determined by adjusting for age, sex, and the first 16 genetic PCs. Black bars indicate 95% confidence intervals. HDL: high-density lipoprotein cholesterol. DBP: diastolic blood pressure. RBC: red blood cell count. TC: total cholesterol. T2D: type 2 diabetes. CRC: colorectal cancer. BC: breast cancer.

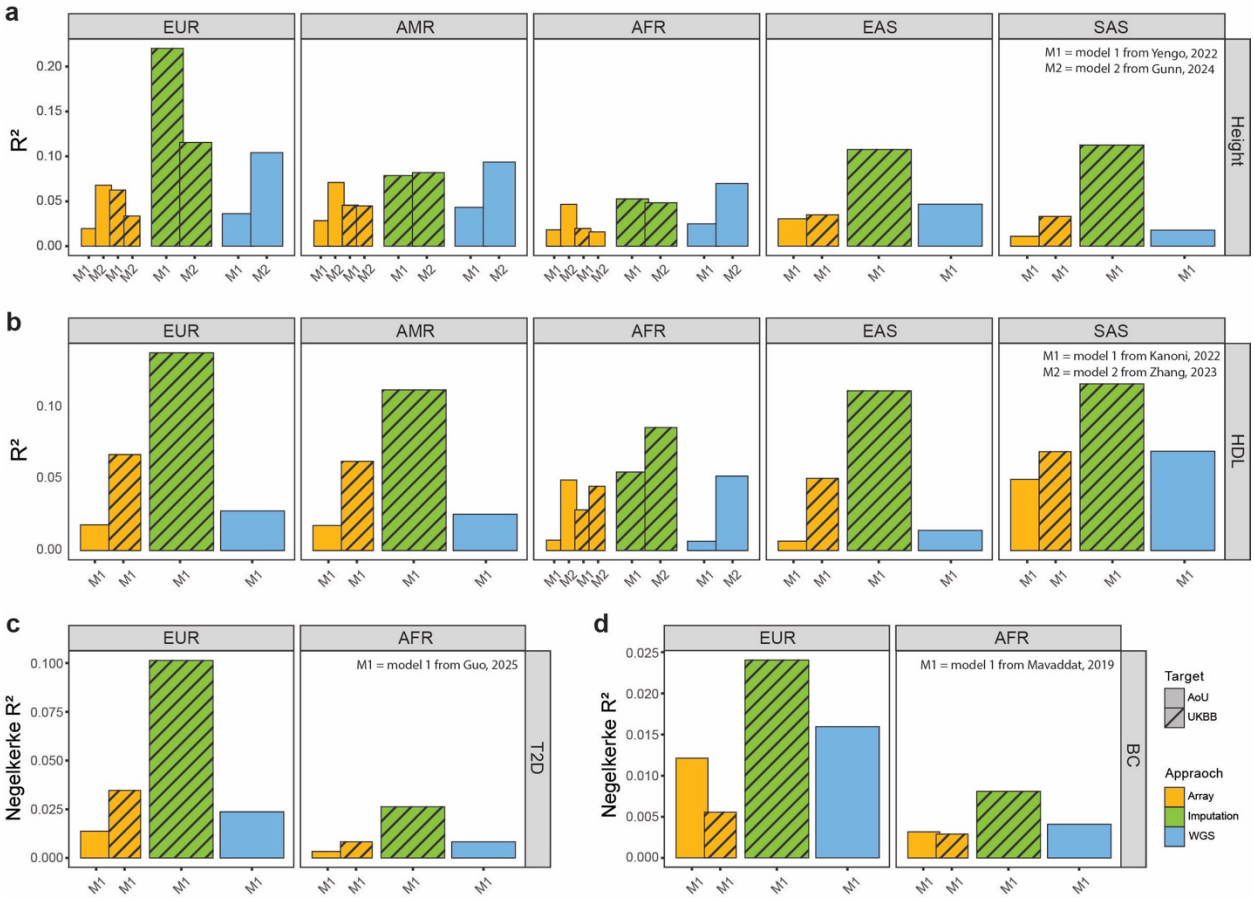


Fig. 3: Evaluation of PGS performance from the PGS Catalog.

Polygenic score data for four traits, (a) height, (b) HDL, (c) T2D, and (d) breast cancer, were extracted from the PGS Catalog. The studies for each trait, where PGS models were trained, are labeled in the upper right corner. Performance was evaluated using genotype discovery approaches from two sources: array and WGS data from *AoU*, and direct genotyping array and imputation data from UKBB. The central South Asian (CSA) group in UKBB was renamed to the SAS group for comparison with *AoU*. For T2D and breast cancer, results are shown only for EUR and AFR populations due to insufficient sample sizes (n < 100) in AMR within UKBB and in EAS and SAS within *AoU*. BC: breast cancer.

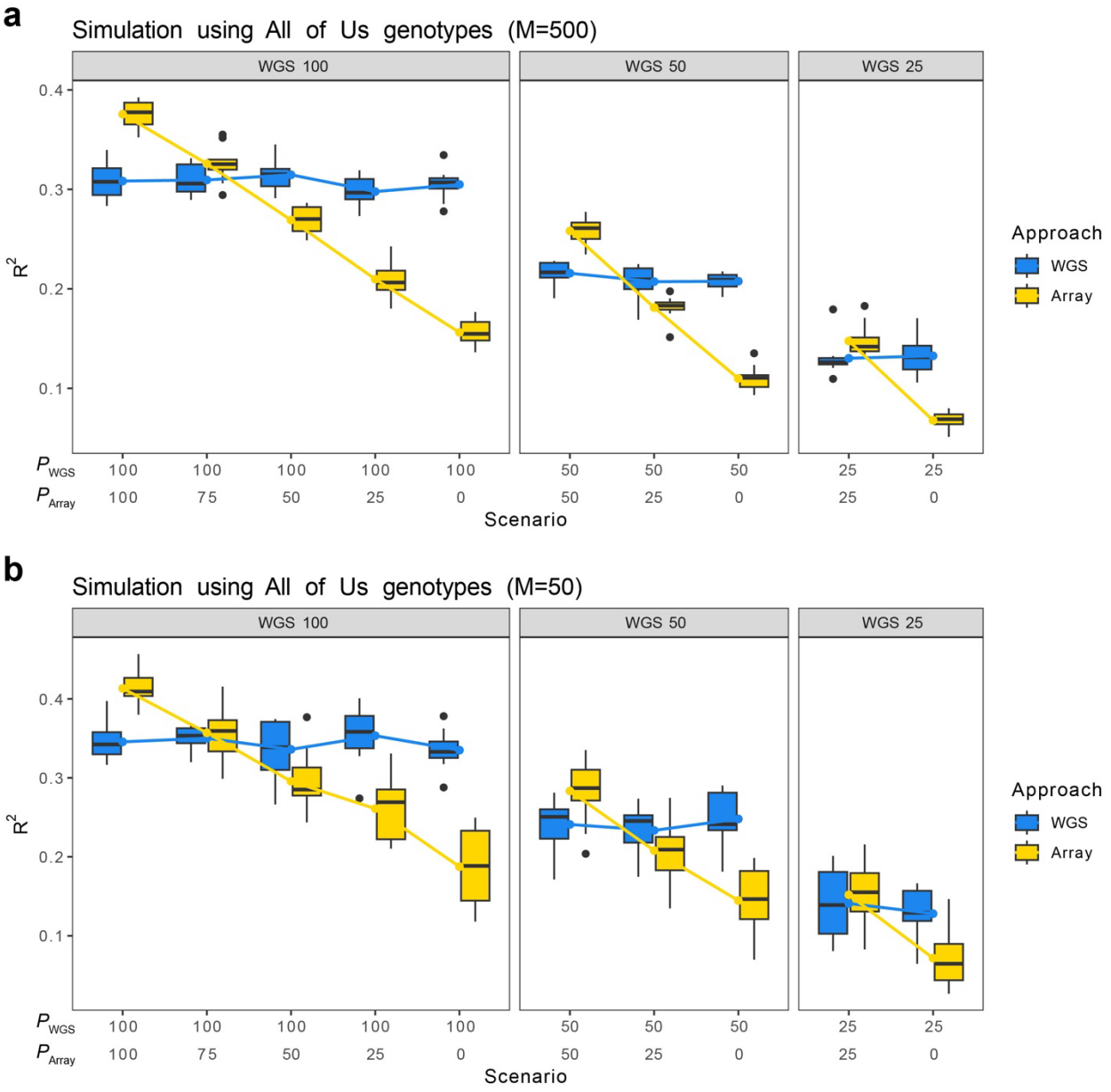


Fig. 4: Simulation settings of causal variants captured.

Traits were simulated using *AoU* WGS genotypes under polygenic settings (**a**, with $M = 500$ causal variants) and sparse settings (**b**, with $M = 50$ causal variants), both with heritability = 0.5. Ten scenarios were simulated. P_{WGS} represent the percentage of causal variants captured by WGS, and P_{Array} denote the percentage of causal variants captured by array.

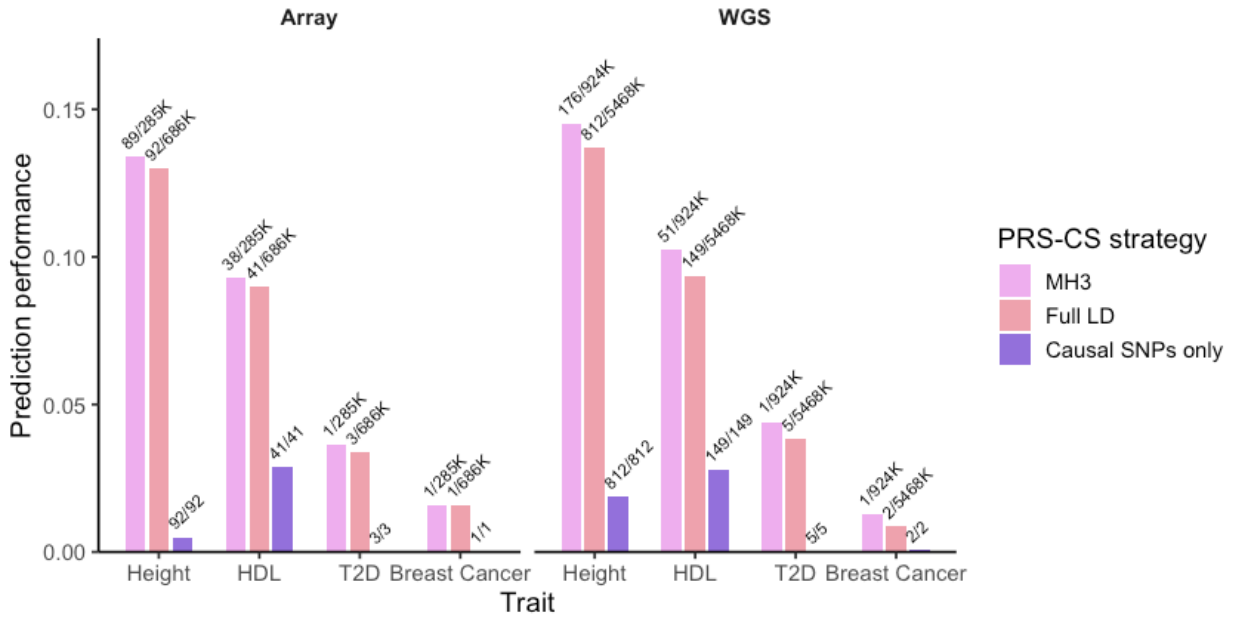


Fig. 5: Prediction performance of PGS including different numbers of causal variants.

The performance for selected traits was assessed with three PRS-CS strategies for variant selection: (HM3) variants intersected with the LD panel constructed using HapMap3 variants (default in PRS-CS), (Full LD) variants intersected with a full LD reference panel, and (Causal SNPs only) only inferred causal SNPs were used with the full LD panel. The numbers on top of each bar indicate the number of causal variants used, out of (/) the total variants included in the PRS-CS construction.

656 **References**

657 1. Konuma, T. & Okada, Y. Statistical genetics and polygenic risk score for precision
658 medicine. *Inflamm. Regen.* **41**, 18 (2021).

659 2. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic
660 risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

661 3. Wang, Y., Tsuo, K., Kanai, M., Neale, B. M. & Martin, A. R. Challenges and
662 opportunities for developing more generalizable polygenic risk scores. *Annu. Rev.*
663 *Biomed. Data Sci.* **5**, 293–320 (2022).

664 4. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry
665 continuum. *Nature* **618**, 774–781 (2023).

666 5. Fitipaldi, H. & Franks, P. W. Ethnic, gender and other sociodemographic biases in
667 genome-wide association studies for the most burdensome non-communicable
668 diseases: 2005-2022. *Hum. Mol. Genet.* **32**, 520–532 (2023).

669 6. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic
670 studies. *Cell* **177**, 26–31 (2019).

671 7. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate
672 health disparities. *Nat. Genet.* **51**, 584–591 (2019).

673 8. Ikegawa, S. A short history of the genome-wide association study: where we were
674 and where we are going. *Genomics Inform.* **10**, 220–225 (2012).

675 9. Cross, B., Turner, R. & Pirmohamed, M. Polygenic risk scores: An overview from
676 bench to bedside for personalised medicine. *Front. Genet.* **13**, 1000667 (2022).

677 10. Yeung, J. M., Sham, P. C., Chan, A. S. & Cherny, S. S. OpenADAM: an open source
678 genome-wide association data management system for Affymetrix SNP arrays. *BMC*
679 *Genomics* **9**, 636 (2008).

680 11. Xing, C. *et al.* Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to
681 detect risk variants for human complex diseases. *Eur. J. Hum. Genet.* **24**, 1029–1034
682 (2016).

683 12. Infinium Global Screening Array-24 Kit | Population-scale genetics.
684 https://www.illumina.com/products/by-type/microarray-kits/infinium-global-
685 screening.html.

686 13. Infinium Global Diversity Array-8 Kit | Multiethnic human genotyping.
687 https://www.illumina.com/products/by-type/microarray-kits/infinium-global-
688 diversity.html.

689 14. All of Us Research Program Genomics Investigators. Genomic data in the All of Us
690 Research Program. *Nature* **627**, 340–346 (2024).

691 15. Infinium Asian Screening Array | Genotyping of East Asian populations.
692 https://www.illumina.com/products/by-type/microarray-kits/infinium-asian-
693 screening.html.

694 16. content, C. A powerful array for genetic studies focused on African populations.

695 17. H3Africa Consortium *et al.* Enabling the genomic revolution in Africa. *Science* **344**,
696 1346–1348 (2014).

697 18. 1000 Genomes Project Consortium *et al.* A global reference for human genetic
698 variation. *Nature* **526**, 68–74 (2015).

699 19. Naito, T. & Okada, Y. Genotype imputation methods for whole and complex
700 genomic regions utilizing deep learning technology. *J. Hum. Genet.* **69**, 481–486
701 (2024).

702 20. McMahon, A. *et al.* Sequencing-based genome-wide association studies reporting
703 standards. *Cell Genomics* **1**, (2021).

704 21. Brlek, P. *et al.* Implementing whole genome sequencing (WGS) in clinical practice:
705 advantages, challenges, and future perspectives. *Cells* **13**, (2024).

706 22. Wendl, M. C. & Wilson, R. K. The theory of discovering rare variants via DNA
707 sequencing. *BMC Genomics* **10**, 485 (2009).

708 23. Wheeler, M. M. *et al.* Whole genome sequencing identifies structural variants
709 contributing to hematologic traits in the NHLBI TOPMed program. *Nat. Commun.*
710 **13**, 7592 (2022).

711 24. Appadurai, V. *et al*. Accuracy of haplotype estimation and whole genome
712 imputation affects complex trait analyses in complex biobanks. *Commun. Biol.* **6**,
713 101 (2023).

714 25. Martin, A. R. *et al*. Low-coverage sequencing cost-effectively detects known and
715 novel variation in underrepresented populations. *Am. J. Hum. Genet.* **108**, 656–668
716 (2021).

717 26. Chat, V., Ferguson, R., Morales, L. & Kirchhoff, T. Ultra Low-Coverage Whole-
718 Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide
719 Association Studies. *Front. Genet.* **12**, 790445 (2021).

720 27. Petter, E. *et al*. Genotype error due to low-coverage sequencing induces uncertainty
721 in polygenic scoring. *Am. J. Hum. Genet.* **110**, 1319–1329 (2023).

722 28. Kim, S. *et al*. Evaluation of low-pass genome sequencing in polygenic risk score
723 calculation for Parkinson's disease. *Hum Genomics* **15**, 58 (2021).

724 29. Wang, Y. *et al*. Polygenic prediction across populations is influenced by ancestry,
725 genetic architecture, and methodology. *Cell Genomics* **3**, 100408 (2023).

726 30. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via
727 Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776
728 (2019).

729 31. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for
730 reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).

731 32. Yengo, L. *et al.* A saturated map of common genetic variants associated with
732 human height. *Nature* **610**, 704–712 (2022).

733 33. Gunn, S. *et al.* Comparison of methods for building polygenic scores for diverse
734 populations. *HGG Adv.* **6**, 100355 (2025).

735 34. Zhang, H. *et al.* A new method for multiancestry polygenic prediction improves
736 performance across diverse populations. *Nat. Genet.* **55**, 1757–1768 (2023).

737 35. Kanoni, S. *et al.* Implicating genes, pleiotropy, and sexual dimorphism at blood lipid
738 loci through multi-ancestry meta-analysis. *Genome Biol.* **23**, 268 (2022).

739 36. Guo, B. *et al.* Polygenic risk score for type 2 diabetes shows context-dependent
740 effects across populations. *Nat. Commun.* **16**, 8632 (2025).

741 37. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast
742 cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).

743 38. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing
744 instruments. *NAR Genom. Bioinform.* **3**, lqab019 (2021).

745 39. Wall, J. D. *et al.* Estimating genotype error rates from high-coverage next-generation
746 sequence data. *Genome Res.* **24**, 1734–1739 (2014).

747 40. Sleep, J. A., Schreiber, A. W. & Baumann, U. Sequencing error correction without a
748 reference genome. *BMC Bioinformatics* **14**, 367 (2013).

749 41. Visscher, P. M., Yang, J. & Goddard, M. E. A commentary on "common SNPs explain
750 a large proportion of the heritability for human height" by Yang et al. (2010). *Twin*
751 *Res. Hum. Genet.* **13**, 517–524 (2010).

752 42. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**,
753 243–250 (2022).

754 43. Kachuri, L. *et al.* Principles and methods for transferring polygenic risk scores
755 across global populations. *Nat. Rev. Genet.* **25**, 8–25 (2024).

756 44. Kullo, I. J. *et al.* Polygenic scores in biomedical research. *Nat. Rev. Genet.* **23**, 524–
757 532 (2022).

758 45. Montinaro, F. *et al.* Unravelling the hidden ancestry of American admixed
759 populations. *Nat. Commun.* **6**, 6596 (2015).

760 46. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic
761 ancestry of African Americans, Latinos, and European Americans across the United
762 States. *Am. J. Hum. Genet.* **96**, 37–53 (2015).

763 47. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed
764 individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195–204 (2021).

765 48. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**,
766 31–46 (2010).

767 49. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic
768 data. *Nature* **562**, 203–209 (2018).

769 50. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from
770 polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

771 51. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger.
772 *Bioinformatics* **36**, 5424–5431 (2021).

773 52. Tsuo, K. *et al.* All of Us diversity and scale improve polygenic prediction contextually
774 with greatest improvements for under-represented populations. *BioRxiv* (2025)
775 doi:10.1101/2024.08.06.606846.

776 53. Koenig, Z. *et al.* A harmonized public resource of deeply sequenced diverse human
777 genomes. *Genome Res.* **34**, 796–809 (2024).

778 54. MacDonald, J., Harrison, T., Bammler, T., Mancuso, N. & Lindstroem, S. An updated
779 map of GRCh38 linkage disequilibrium blocks based on European ancestry data.
780 *BioRxiv* (2022) doi:10.1101/2022.03.04.483057.

781 55. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide
782 complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

56. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).