Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits.

Luke M. Evans[1,7], Rasool Tahmasbi[1], Scott I. Vrieze[1], Gonçalo R. Abecasis[2], Sayantan Das[2], Doug W. Bjelland[1], Teresa R. deCandia[1], Haplotype Reference Consortium, Michael E. Goddard[3], Benjamin M. Neale[5], Jian Yang[4], Peter M. Visscher[4], Matthew C. Keller[1,6,7]

[1]Institute for Behavioral Genetics, University of Colorado, Boulder, CO 80309

[2]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109

[3]Faculty of Veterinary and Agricultural Science, University of Melbourne, Parkville, Victoria, Australia

[4]Institute for Molecular Bioscience and the Queensland Brain Institute, University of Queensland, Brisbane, 4072, Queensland, Australia

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

[6]Department of Psychology and Neuroscience, University of Colorado, Boulder, CO, 80309

[7]Corresponding authors luke.m.evans@colorado.edu and matthew.c.keller@gmail.com

ABSTRACT

Heritability, h^2, is a foundational concept in genetics, critical to understanding the genetic basis of complex traits. Recently-developed methods that estimate heritability from genotyped SNPs, h^2_{SNP}, explain substantially more genetic variance than genome-wide significant loci, but less than classical estimates from twins and families. However, h^2_{SNP} estimates have yet to be comprehensively compared under a range of genetic architectures, making it difficult to draw conclusions from sometimes conflicting published estimates. Here, we used thousands of real whole genome sequences to simulate realistic phenotypes under a variety of genetic architectures, including those from very rare causal variants. We compared the performance of ten methods across different types of genotypic data (commercial SNP array positions, whole genome sequence variants, and imputed variants) and under differing causal variant frequencies, levels of stratification, and relatedness thresholds. These results provide guidance in interpreting past results and choosing optimal approaches for future studies. We then chose two methods (GREML-MS and GREML-LDMS) that best estimated overall h^2_{SNP} and the causal variant frequency spectra to six phenotypes in the UK Biobank using imputed genome-wide variants. Our results suggest that as imputation reference panels become larger and more diverse, estimates of the frequency distribution of causal variants will become increasingly unbiased and the vast majority of trait narrow-sense heritability will be accounted for.

KEYWORDS

heritability, h^2; complex trait; genetic architecture; GREML

INTRODUCTION

Narrow-sense heritability, h^2, the proportion of the total phenotypic variance due to additive genetic variation, is a fundamental concept of medical and quantitative genetics. In addition to providing an understanding of the genetic basis of traits, h^2 determines the response to selection, the potential utility of individual genetic risk and trait prediction, and how much of the phenotypic variability could theoretically be accounted for in genome-wide association studies (GWAS)[1,2]. Importantly, while GWAS have now identified thousands of variants associated with complex traits[3–5], the loci identified by these studies have typically explained only a small fraction of traits' total heritability, with the remaining genetic variance termed "missing heritability." This remaining unaccounted for genetic variance may be attributable to a variety of causes, including the role of (typically rare) variants poorly tagged by arrays, small effect common variants that do not reach genome-wide significance due to insufficient sample sizes, or inflated family-based h^2 estimates[1,6–8].

While traditional family-based estimates of heritability, h^2_{FAM}, have provided valuable insights[9], the use of close relatives means that estimates of additive genetic variance can be biased by factors shared by close relatives—for example, the joint action of non-additive genetic and common environmental effects can inflate estimates of additive genetic variation[10,11]. Recently-developed approaches that utilize unrelated individuals to estimate the variance explained by all genotyped single nucleotide polymorphisms (SNPs), denoted as h^2_{SNP}, have the advantage of being unaffected by these sources of bias, and for many traits have found that a large proportion of the heritability is captured by common variants[6,12,13]. For certain complex traits, such as height, little unexplained additive genetic variance remains, as h^2_{SNP} approaches h^2_{FAM}[7,12]. Despite this, h^2_{SNP} estimates for most traits are still below h^2_{FAM}, with BMI a typical

68 example where h^2_{SNP} ~0.27 while h^2_{FAM} ~0.4-0.6 (ref. [12]). Thus, for many complex traits, including
69 disease traits, much of the heritability remains unaccounted for.

70 A second application of these approaches is to better understand the genetic architecture
71 of complex traits. Genetic architecture refers to the number, frequencies, effect sizes, and
72 locations of causal variants (CVs) underlying trait variation. Methods for estimating heritability
73 from SNPs have found that estimated genetic variance is proportional to chromosome length for
74 numerous complex traits, including height, BMI, schizophrenia, depression, and metabolic traits,
75 consistent with the hypothesis that these traits are influenced by hundreds to thousands of
76 variants with small effects spread throughout the genome[5,6,8,12–16]. More recently, these methods
77 have allowed insight into the frequency distribution and functional annotation of causal variants
78 by partitioning SNPs into MAF bins and annotation categories[17,18]. Such methods have allowed
79 insight into gene networks involved in complex traits[19], and helped determine optimal strategies
80 for large-scale genotyping, such as whether genotyped SNPs on commercial arrays with
81 subsequent imputation can capture the genetic variation from all frequency classes of causal
82 variants or if whole genome sequences instead are needed[12].

83 A variety of methods to estimate h^2_{SNP} and partition the genetic variance among sets of
84 markers have been developed for these purposes. Many of these methods use one or more
85 genetic relatedness matrices (GRMs) to estimate variances using restricted maximum likelihood
86 (GREML)[6,12,17,20]. Manipulations of the GRM via treelet covariance smoothing[21] or weighting by
87 linkage disequilibrium (LD) tagging of SNPs[13] have also been proposed. A much different
88 approach, LD-score regression, estimates h^2_{SNP} from GWAS summary statistics[22]. The
89 performance of these methods has typically been evaluated via simulation by assuming that
90 causal variants have the same properties, on average, as common SNPs found on commercial

91 genotyping arrays. However, such an approach is problematic because SNPs are specifically
92 selected because they are common, have unusually high LD with untyped SNPs, or have been
93 implicated in disease (e.g., the Affymetrix Axiom chip used in the UK Biobank[23]). SNPs on arrays
94 are therefore probably not reflective of typical CVs across the genome, and thus the ability of
95 these methods to estimate h^2_{SNP} or determine the genetic architecture of complex traits has not
96 yet been properly assessed, nor have these methods been directly compared across conditions,
97 such as levels of stratification or environmental confounding, that can cause biases. In particular,
98 how the various methods perform with traits derived from very rare CVs may be quite different
99 than how they perform on traits derived from common, well-tagged CVs, such as those used on
100 SNP arrays.

101 Here, we utilize thousands of recently-sequenced whole genomes to simulate complex
102 phenotypes to test the performance of the most widely used SNP heritability estimation methods.
103 We examine each method's ability to estimate h^2_{SNP} while varying the amount of population
104 stratification, the frequency distributions of causal variants, and the type of whole-genome data
105 analyzed (SNP array, imputed, and sequence). By using real sequence data to simulate
106 phenotypes, the genotypic data we use are highly realistic with respect to LD, allele frequency
107 distributions (with minor allele frequencies down to $3x10^{-4}$), variant density, and other genomic
108 properties found in real data. Finally, we use the best-performing methods to estimate h^2_{SNP} and
109 examine genetic architecture for six complex traits using the UK Biobank. While h^2_{SNP} estimation
110 following imputation can account for the majority of the heritability, larger sample sizes and
111 reference panels, or novel methods, will be needed to fully account for all the additive genetic
112 variance in complex traits involving very rare causal variants.
113

114

115 **MATERIALS AND METHODS**.

116 ***Samples and Population Structure***

117 We simulated continuous phenotypes derived from whole genome sequence (WGS) data
118 in the Haplotype Reference Consortium (HRC) dataset. Full details of the HRC can be found in
119 McCarthy et al.[24]. Briefly, this resource comprises roughly 32,500 individual whole genome
120 sequences from multiple whole-genome sequencing studies, with phased genotype calls
121 available at all sites with a minor allele count of at least 5. The HRC contains world-wide
122 populations, but the majority are of European (EUR) origin. This large collection allowed us to
123 simulate phenotypes with differing genomic architectures under realistic patters of LD structure,
124 stratification, and relatedness with the whole genomes. We obtained permission to access the
125 following HRC cohorts (recruitment region & sample size): AMD (Europe & worldwide; 3,189),
126 BIPOLAR (European ancestry; 2,487), GECCO (European ancestry; 1,112), GOT2D (Europe,
127 2,709), HUNT (Norway; 1,023), SARDINIA (Sardinia; 3,445), TWINS (Minnesota; 1,325), 1000
128 Genomes (worldwide; 2,495), UK10K (UK; 3,715) (see web resources for HRC information
129 including specific cohorts). The subset of the HRC data we accessed totaled 21,500 whole
130 genome sequences comprising 38,913,048 biallelic SNPs.

131 Our goal was to assess the bias and precision of various h^2_{SNP} estimation methods using
132 data similar to that typically used in GWAS and h^2_{SNP} analyses. In order to mimic this kind of
133 data, we first extracted variant positions corresponding to a widely-used commercially available
134 genotyping array, the UKBiobank Affymetrix Axiom array. We performed principal components
135 analysis using flashpca[25] on 133,603 SNPs after LD and MAF pruning (plink2[26] commands –maf
136 0.05 --indep-pairwise 1000 400 0.2), extracting the first ten PCs, and performing K-means

clustering in R[27]. We used the 1000 Genomes individuals in the HRC as anchor points for ancestry and identified 19,478 individuals of European descent, including individuals of Finnish and Sardinian ancestry (Figure S1).

To identify subsets of these 19,478 individuals spanning different levels of genetic heterogeneity, we reran PCA with only these individuals, then proceeded to identify four increasingly homogenous subgroups within them using K-means clustering (Fig. 1). The most stratified group contained all EUR samples (N=19,478). The somewhat stratified group excluded Sardinian and Finnish samples (N=14,424). The low stratification group contained only northern/western European samples (N=11,243), and the least stratified (homogeneous) group was a subset of British ancestry samples (N=8,506). We used GCTA[20] to estimate relatedness and remove samples so that the maximum relatedness was 0.1 within each of the four samples. In the most homogeneous (smallest) sample, this left 8,201 individuals. To avoid confounding sample size with degree of stratification, we randomly chose 8,201 of the unrelated individuals from within each of the other three more stratified subsamples. Our purpose in identifying these groups was to vary the amount of genetic heterogeneity within a sample, similar to what might be found across a range of different GWAS samples, rather than formal population assignment or classification of individuals. We also identified individuals with relatedness less than 0.05 within each group, and used both subsets to examine how a 0.1 or 0.05 relatedness cutoff influences h^2_{SNP} estimates. Sample sizes when using the 0.05 relatedness cutoff were 7792, 8115, 8129, and 8186 for the four genetic structure subsamples.

Simulated Phenotypes Using Whole Genome Sequencing Data

159 To assess how methods performed on a range of genetic architectures, we simulated
160 phenotypes from CVs drawn randomly from five MAF ranges from the whole genome sequence
161 data: common (MAF≥0.05), uncommon (0.01≤MAF<0.05), rare (0.0025≤MAF<0.01), very rare
162 (0.0003≤MAF<0.0025), and all variants that had a minor allele count (MAC) of at least 5
163 (MAF≥0.0003) (Fig. S2). Phenotypes were generated from 1,000 CVs from the model $y_i = g_i + e_i$,
164 where $g_i = \sum w_{ik}\beta_k$, w_{ik} is the genotype (coded as 0, 1, or 2) of individual i at the kth CV, and β_k is
165 the kth allelic effect size, drawn from ~N(0,1/[2p_k(1-p_k)]), where p_k is the MAF of allele k within a
166 population subset. This model therefore assumes larger average additive effect sizes for rarer
167 variants. The g_i's were standardized and added to residual error drawn from ~N(0, (1- h^2)/h^2) for
168 a h^2 of 0.5 for simulated phenotypes. A total of 100 repetitions were simulated for phenotypes
169 derived from each CV MAF range and for each of the four population stratification subsets. It is
170 important to note that we did not simulate any phenotypic effects as a function of ancestry within
171 any of the subsamples, and thus biases related to stratification in our results were due to the
172 genotypic (e.g., long-range LD), not phenotypic, effects of stratification.
173

174 ***SNPs, WGS, and Imputed Variants***

175 Most marker heritability studies utilize commonly available commercial arrays, and
176 estimates of h^2_{SNP} reflect how well SNPs on these arrays tag CVs. In particular, CVs with low
177 MAF or that exist in regions of low LD are typically tagged poorly by SNP arrays[6,13] and h^2_{SNP} <
178 h^2 in these situations. Alternatively, as large WGS reference panels (e.g., 1KG, UK10K, HRC)
179 become increasingly available, imputing genome-wide variants based on SNP arrays is an
180 attractive option for capturing more and rarer genetic variants than possible on arrays, although
181 imputation accuracy declines with MAF[12]. Finally, using WGS data to estimate GRMs should

182 reflect relatedness at all CVs, including those that are rare or in low LD with other SNPs.
183 Although WGS data in phenotyped samples is not yet widely available at the sample sizes
184 required for precise estimation of h^2_{SNP}, we include it as a benchmark for results based on array
185 and imputed data and because large WGS samples are likely to become increasingly available
186 in the future. We therefore tested each of these data types (array, imputed, and WGS variants)
187 using each of the methods described below to determine how much of the heritability can be
188 captured from each data type, and how closely results from imputed data mimic those from WGS
189 data.

190 From the HRC sequence data (the WGS dataset), we extracted positions corresponding
191 to the Axiom array as noted above (the array SNP dataset) with MAF>0.01. To impute, we used
192 the 8,201 unrelated individuals in each population stratification set and added their close
193 relatives (relatedness > 0.1) back into the sample as described below in the GREML-SC method
194 description. We added these close relatives back in to the target imputation set in order to a)
195 remove close relatives from the reference panel which would artificially increase imputation
196 accuracy, and b) because some of the methods described below require the use of closely
197 related individuals. We phased these individuals using SHAPEIT2[28], imputed using minimac3[29],
198 and retained variants with imputation $R^2 \geq 0.3$ (ref. [12]). We used the HRC sequence data as our
199 imputation reference panel after removing all target (8201 unrelated + relatives) individuals,
200 thereby assuring ~independence (no relatedness) between the target and reference panels.
201 Final reference panel sizes for the four structure subsamples were 11,584; 12,799; 12,785; and
202 12,994. Reducing the sample size of the reference panel likely resulted in poorer imputation than
203 had we used the full HRC panel but was nevertheless substantially larger than reference panels
204 used in most past imputation procedures (e.g., 1,000 Genomes). Moreover, because the target

205 and reference samples were from the same populations and the same cohorts, the imputation
206 quality is likely higher than most GWAS samples would obtain. However, given that the HRC has
207 become a widely-used imputation reference panel, our imputation quality is probably roughly
208 reflective of imputation quality using modern procedures.

209 The amount of tagging throughout the genome differs between the various commercial
210 arrays[12], and these differences may lead to differing h^2_{SNP} estimates. To assess this, for the
211 GREML-SC and GREML-MS methods (see below) using array positions data, we compared
212 results from the Axiom array to those from the Illumina Omni2.5 array. For reference, MAF
213 distributions of the different data types for two of the structure subsamples are shown in Figure
214 S2.

215

216 ***Heritability Estimation Methods Tested***

217 Numerous methods have recently been developed to estimate h^2_{SNP} and partition genetic
218 variance using genomic data. Among these, we compared the most widely used, including the
219 various single and multiple component GREML approaches implemented in the GCTA
220 software[6,12,17], approaches that specifically take into account how LD influences the tagging of
221 nearby sites by SNPs[13], those that use related and unrelated samples to account for rare and
222 common variant effects[8], those that denoise the GRM using treelet covariance smoothing[21],
223 those that relate the effect sizes of SNPS from a GWAS to their degree of LD tagging[19,22], and
224 computationally efficient mixed model approaches[18]. Here, we briefly describe our
225 implementation of each of these methods; for additional information on the methods themselves,
226 see the above references. For all methods except LD-Score Regression and BOLT-REML
227 (described below), we generated GRMs following the procedures of each method, and estimated

228 h^2_{SNP} using GCTA[20]. In all models, variance component estimates were unconstrained (e.g., by
229 using the –reml-no-constrain option of GCTA), and included 20 PCs (10 from worldwide PCA
230 and 10 from the specific subsample PCA) as continuous covariates and sequencing cohort as a
231 categorical covariate.
232
233 **Single Component GREML (GREML-SC)**
234 Yang et al.[6] introduced the single component GRM approach using a mixed-effects
235 model, with GRM entries:

236 $$A_{ij} = \frac{1}{m}\sum_k^m \frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)} \quad (1)$$

237 where m is the number of SNPs, x_{jk} is the genotype (coded as 0, 1, or 2) of individual j at the kth
238 locus, and p_k is the MAF of the kth locus. The variance of the phenotypes is

239 $$var(\boldsymbol{y}) = \boldsymbol{A}\sigma_v^2 + \boldsymbol{I}\sigma_e^2 \quad (2)$$

240 where the variance explained by the SNPs (σ^2_v) and error variance (σ^2_e) are estimated using
241 restricted maximum likelihood (REML) implemented in the GCTA package[20]. The proportion of
242 the total variance explained by all SNPs is then a measure of heritability ($h^2_{SNP} = \sigma^2_v / (\sigma^2_v +$
243 $\sigma^2_e)$). Typically, the set of m SNPs used to build the GRM is the set of SNPs with MAF≥0.01
244 (hereafter "common SNPs") and unrelated individuals (relatedness ≤ 0.05). Because the Axiom
245 array contains some rare markers, we compared this approach to one using all SNPs with
246 MAC≥5 (hereafter "all SNPs") in each particular stratification subsample, as well as to an
247 approach using less stringent relatedness thresholds (relatedness < 0.10 and no relatedness
248 threshold). For analyses that used no relatedness threshold, inclusion of close relatives

249 increased our sample sizes to 9916, 8701, 8715, and 8506 for the samples with most, some,
250 low, and least stratification, respectively (Fig. 1).
251
252 **MAF-Stratified GREML (GREML-MS)**
253 Biased estimates of h^2_{SNP} are expected when using the GREML-SC method if the MAF
254 distribution of the CVs does not match the MAF distribution of SNPs used to generate the
255 GRM[17]. Stratifying variants into MAF classes and using a multiple GRM GREML approach can
256 mitigate this bias and can also partition the genetic variance into that explained by different MAF
257 categories of SNPs, lending insight into the genetic architecture of complex traits[12,30]. We applied
258 this approach using 4 MAF categories, matching the CV MAF categories used for phenotype
259 simulation.
260
261 **LD- and MAF-Stratified GREML (GREML-LDMS)**
262 Extending the GREML-MS method to account for different levels of LD throughout the
263 genome, Yang et al.[12] introduced an LD score-stratified method to the GREML-MS approach.
264 GREML-LDMS stratifies variants according to both MAF categories as well as an LD-score,
265 defined as the sum of r^2 between the focal variant and all other variants in a window. We
266 estimated LD scores using the default settings in GCTA (10Mb block size with a 5Mb overlap),
267 and stratified variants into LD score quartiles. Combined with the four MAF categories above, we
268 used 16 GRMs for this approach.
269
270 **Single Component and MAF-Stratified LD-Adjusted Kinships (LDAK-SC and LDAK-MS)**

271 Speed et al.[13] noted that because LD varies across the genome, CVs in regions of high
272 LD are given disproportionate weight by eqn. (1) above. They proposed a method to weight
273 SNPs according to local LD, which potentially corrects for the bias introduced when there is
274 variation in how well CVs are tagged by SNPs. We used LDAK5[13] to estimate these LD-weighted
275 GRMs. This approach thins SNPs in very high LD first to reduce redundant tagging, then
276 estimates SNP weights that are inversely proportional to their average LD with other SNPs. We
277 also applied the MAF-stratified approach described above with the LDAK method (LDAK-MS).
278 For the single component model (LDAK-SC), we used all SNPs (MAC≥5) as well as only
279 common SNPs (MAF≥0.01) to build the GRM. For the MAF-stratified approach, following
280 recommendations in the LDAK documentation, we estimated variant weights over the union of all
281 variants (MAC≥5), then computed GRMs for each MAF class separately. We then applied the
282 multiple GRM method with these LDAK-weighted GRMs to estimate h^2_{SNP} using GCTA.
283

284 **Extended Genealogy with Thresholded GRMs**

285 Zaitlen et al.[8] introduced a method to simultaneously estimate the full narrow-sense
286 heritability (incorporating the effects of poorly tagged SNPs) and h^2_{SNP} using two GRMs in a
287 sample containing close relatives. The first GRM contains relatedness from SNPs for all
288 individuals while relatedness estimates below a threshold, *t*, are set to 0 in the second GRM. The
289 first GRM, therefore, contains information on allele sharing of (mostly common) variants in
290 unrelated and related individuals and is used to estimate h^2_{SNP}, while the second only contains
291 information from closely related individuals, presumably reflecting sharing of both common and
292 rare CVs, and provides an estimate of what we call $h^2_{IBS>t}$, (following Zaitlen et al.[8]). The sum of
293 $h^2_{IBS>t}$ and h^2_{SNP} should therefore provide an estimate of total h^2, similar to $h^2_{FAM.}$, with all the

294 same potential biases that exist in h^2_{FAM} estimates from designs that use close relatives. We
295 tested two relatedness thresholds ($t \leq 0.05$ and 0.1) for the second GRM. By necessity, all
296 analyses using the relatedness thresholded GRM approach included close relatives.
297

298 **Treelet Covariance Smoothing (TCS)**

299 Crossett et al.[21] noted that the GRM estimates (particularly for unrelated individuals) are
300 inherently noisy. They proposed a method to smooth the estimates using treelet covariance
301 smoothing (TCS) to obtain more accurate estimates of relatedness. Their method takes
302 advantage of the hierarchical nature of relatedness in samples to obtain better estimates of A_{ij}
303 among unrelated individuals. We replicated their methods, using common SNPs (MAF≥0.01) and
304 including related individuals, and implemented the TCS method in the *treelet* R package[31]. TCS
305 requires identifying a smoothing parameter, λ (distinct from the genomic control inflation factor
306 λ_{GC}). Crossett et al.[21] propose two methods to optimize λ, one based on minimizing the GREML
307 likelihood and one based on minimizing a loss function ($H(\lambda)$) at different levels of λ based on
308 subsamples of the SNPs. With the large number of simulations across stratification subsamples
309 and genetic architectures, minimizing the GREML likelihood for each simulated phenotype was
310 not feasible. Minimizing $H(\lambda)$ using the second approach requires estimating the GRM and
311 applying the TCS method to over 50 subsets of data, also impractical computationally with over
312 8,000 individuals. We therefore used a modification of the 2nd approach. We built GRMs from
313 2000 randomly chosen individuals from each stratification subsample and optimized λ for each
314 subsample following the published methodology (Fig. S3), then applied the optimal λ to the full
315 GRM of over 8,000 individuals.
316

317 **LD-Score Regression**

318 LD-score regression uses a different approach to estimating h^2_{SNP}. Rather than estimating
319 relatedness within a sample for use in mixed-model GREML analysis, LD-score regression
320 regresses GWAS test statistics (χ^2) on SNPs' LD scores, which reflect the degree to which each
321 SNP is correlated with surrounding SNPs[19,22]. For a polygenic model, the expected GWAS test
322 statistic of variant *j*, χ^2_j, is

323

324 $$E[\chi^2_j \mid l_j] = N(h^2_{SNP})l_j / M + Na + 1 \quad (3)$$

325

326 where *N* is the sample size, *M* is the number of SNPs, l_j is the LD score ($= \Sigma_k r^2_{jk}$) measuring the
327 tagging of surrounding variants by SNP *j*, and *a* is a measure of confounding biases arising from
328 stratification and cryptic relatedness. Thus, regressing GWAS test statistics on per-variant LD
329 scores allows for both estimation of h^2_{SNP} and assessing the degree of confounding or
330 polygenicity of a trait[22]. Bulik-Sullivan et al.[22] argue that LD-score regression provides unbiased
331 estimates of h^2_{SNP} regardless of whether GWAS test statistics are estimated with or without
332 controlling for ancestry or environmental covariates or relatedness. Here, we estimated GWAS
333 test statistics using plink2 without controlling for ancestry covariates, controlling for ancestry
334 covariates (20 PCs and sequencing cohort as above), and controlling for ancestry covariates as
335 fixed effects in a mixed model that included a kinship matrix. For the latter, we applied the GCTA
336 leave-one-chromosome-out (LOCO) approach[32]; because the GCTA-LOCO approach is
337 computationally intensive, we ran only 20 repetitions of each phenotype rather than 100, and did
338 so only for the array SNP dataset. We used the *ldsc* package with default parameters (see
339 URLs) to perform LD score regression. We calculated LD scores for all variants using the whole

340 genome sequence data, including common and rare variants. As recommended by Bulik-Sullivan
341 et al.[22], we used unrelated individuals (relatedness ≤ 0.05) and only common variants to perform
342 the LD score regression itself, because the relationship between the GWAS χ^2 and LD-score is
343 unclear for rare (MAF<.01) SNPs.

344 LD score regression can also be used to partition heritability among annotations[19]. We
345 applied this approach using the four MAF categories described above. Because our MAF
346 categories included very rare variants, for this MAF-stratified LD score regression, we used
347 GWAS test statistics from all variants (MAF≥0.0003, using the --not-5-50 flag in the ldsc
348 package) while controlling for covariates as above.

349

350 **BOLT-REML**

351 Unlike other GREML approaches, BOLT-REML uses a Monte Carlo approximation of the
352 gradient for the likelihood function to reduce computation time and memory requirements in
353 variance component estimation[18]. When using whole genome sequence and imputed variant
354 data with >14M variants (see below), time required by BOLT-REML, even when highly
355 parallelized, was prohibitive for 100 repetitions of each combination of variables we tested, as it
356 scales with $MN^{1.5}$, where M is the number of markers and N is the number of samples (see
357 Supplementary Table 1 of Loh et al.[18] for computational performance). Note that GREML takes
358 longer for a single sample due to the length of time to create the GRM; in our simulations with
359 GCTA-style approaches, the GRM computation was done only once, and therefore was much
360 faster when estimating heritability for many repetitions created from randomly-drawn CVs with a
361 single GRM. We therefore only applied BOLT-REML to the array dataset. We applied the
362 method with a single component using either all array positions or only common markers

363 (MAF>0.01) as well as a MAF-stratified approach with the same four MAF partitions and same
364 covariates described above.
365

366 **Confounding between relatedness and shared environments**

367 Many of the methods we tested use unrelated individuals to avoid the assumption of no
368 shared environmental effect among near relatives[6]. However, several, such as the extended
369 genealogy with thresholding, require the use of near relatives. This could lead to confounding
370 between relatedness estimates and shared environmental effects within families or closely
371 related individuals if shared environmental effects are not modeled[7,33]. Indeed, Zaitlen et al.[8]
372 argue that such shared environmental effects were the likely cause of higher h^2_{FAM} estimates
373 among relatives who shared an environment through cohabitation (e.g., half-siblings) compared
374 to equally related relatives that did not share a cohabitation environment (e.g., grand-parents
375 and grand-children). We therefore assessed whether h^2_{SNP} and h^2_{FAM} estimates are biased for
376 methods that use closely related individuals when extended shared environmental effects are
377 present but unmodeled.

378 We first identified all groups of individuals connected by at least one pairwise relatedness
379 value > 0.2 ("extended families"). Note that many of the pairwise relationships within these
380 extended families were below 0.2. For example, spouses are typically unrelated but are
381 nevertheless defined as being in the same family if their offspring are present, and cousins would
382 be defined as being in the same family if their parents were present in the sample. We then
383 simulated phenotypes with a shared extended family environmental effect that accounted for
384 10% of the variance (c^2=0.1). Simulations were similar to those described above, with genotypic
385 values exactly the same as above, but with shared effects for each family drawn from ~N(0, V_c),

386 where $V_c = c^2 * V_g / h^2$, V_g is the variance of genetic values, and c^2 is the proportion of the
387 phenotypic variance due to shared environments, and residual error added as ~N(0, (1- h^2-
388 c^2)*V_g/h^2), for a simulated h^2=0.5, c^2=0.1, and e^2=0.4. We applied GREML-SC, LD score
389 regression, and extended genealogy with thresholded GRMs using common variants from array
390 SNPs controlling for the same covariates as above and without modeling the shared
391 environmental effect. This tested whether methods are robust to violations of the assumption of
392 no shared environmental effects on the phenotype.
393

394 ***Heritability of Complex Traits in the UK Biobank***

395 We estimated h^2_{SNP} for six continuous phenotypes in the UK Biobank using the methods
396 (GREML-MS and GREML-LDMS) that produced consistently unbiased estimates of h^2 and
397 partitioned the genetic variance most accurately in the simulations above. The UK Biobank is a
398 large, publicly available resource of ~500,000 UK adults, with deep phenotyping, family history,
399 and genotype data[23]. The current release includes ~150,000 individuals, primarily of European
400 ancestry, genotyped on the Affymetrix Axiom platform, phased using SHAPEIT2 and imputed to
401 a combined 1000 Genomes and UK10K reference panel (N=6,285 individuals). The details of the
402 official UK Biobank genotyping and imputation methods in the released data can be found at
403 http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf and
404 http://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf (accessed 29 Feb. 2016). We
405 excluded individuals with no genetic data and those whose self-reported and genetic sex
406 conflicted (data fields f.31.0.0 and f.22001.0.0). Poor quality samples identified by the UK
407 Biobank and Affymetrix were also removed (f.220010.0.0) as were UKBiLEVE poor-quality
408 samples (f.22051.0.0), leaving a total of 151,661 individuals. To reduce population stratification,

409 we included only individuals of European ancestry in our analyses. The UK Biobank identified
410 self-reported “British” individuals as “Caucasian” based on grouping of individuals with CEU
411 individuals in PCA (see UK Biobank documentation). To these individuals (f.22006.0.0), we
412 added those who self-identified as “White,” “Irish,” or “Any other white background” whose PC
413 scores on the first four axes (f.22009.0.1-4) were within the range of the UK Biobank-identified
414 “Caucasian” individuals, resulting in 126,338 individuals. We projected the UK Biobank samples
415 onto the HRC PCA axes using the loadings from the HRC EUR individuals, demonstrating that
416 the UK Biobank individuals we used in the analyses below are similar to the least stratified or
417 unstratified subsamples of the HRC we used (Fig. 1). To estimate the GRMs, we separately
418 used directly genotyped Axiom array positions as well as imputed genome-wide variants with
419 IMPUTE info score ≥0.3.

420 We estimated h^2_{SNP} for the following traits in the UK Biobank (field ID number): height
421 (f.50.0.0), body mass index (BMI; f.21001.0.0), whole-body impedance (f.23127.0.0), trunk fat
422 percentage (f.23127.0.0), fluid intelligence (f.20016.0.0), and neuroticism (f.20127.0.0). We
423 normalized phenotypes and removed observations greater than 5 standard deviations away from
424 the mean. We included sex (f.31.0.0), UK Biobank assessment centre (f.54.0.0), genotype
425 measurement batch (f.22000.0.0), and educational attainment (“qualification”, f.6138.0.0) as
426 categorical covariates, and the Townsend deprivation index (f.189.0.0), age at assessment
427 (f.21003.0.0), age at assessment squared, and the 15 PC scores from the UK Biobank
428 (f.22009.0.1-15) as quantitative covariates.

429 For GREML-MS, we binned variants into eight MAF-categories: MAC≥5 & MAF<0.0001,
430 0.0001-0.001, 0.001-0.01, 0.01-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, & 0.4-0.5. For GREML-LMDS, we
431 were limited in the number of predictor GRMs to use due to computational constraints (1Tb of

432 RAM); we therefore, used 4 MAF bins (common: MAF>0.05, uncommon: 0.01<MAF<0.05, rare:
433 0.0001<MAF<0.01, and very rare: MAC>5 & MAF<0.0001) and 2 LD-score bins (above and
434 below the median LD-score).
435
436 **RESULTS**
437 ***Simulation Results***
438 We found clear differences across methods, degree of stratification, and data types (array
439 SNP, WGS, or imputed variants) in their ability to estimate the simulated h^2 for different CV MAF
440 architectures (Figs. 2-3 and S4-S6, Tables S1-S3). Below, we describe results for each method
441 in detail. Please refer to Figures 2-4, Figures S4-S6, and Tables S1-S5 for estimates of
442 heritability, and Figures S7-S9 for estimates of the heritability standard errors.
443
444 **Single Component GREML (GREML-SC)**
445 Estimates of h^2_{SNP} using GREML-SC were highly sensitive to the CV allele frequencies,
446 dataset type (SNPs, WGS, or imputed variants), level of stratification, and MAF cutoff for SNPs
447 used to build the GRM. Using only Axiom array positions, h^2_{SNP} was overestimated by ~20% for
448 common CV phenotypes, and progressively underestimated with rarer CVs, regardless of
449 whether all or just common (MAF>0.01) SNPs were used to build the GRM. The underestimation
450 of h^2_{SNP} when the GRM is built from SNPs that are more common on average than the CVs is
451 well known[6]. It is due to a more general principle: when the average LD between CVs and the
452 markers used to build the GRM is lower than the average LD among the markers themselves,
453 h^2_{SNP} is underestimated[12]. Thus, h^2_{SNP} was underestimated for rare CVs because they tend to
454 have lower LD with common markers than the common markers have with each other.

455 The overestimation of h^2_{SNP} for common CVs in our results is explained by the same
456 principle–the average LD between CVs and markers is, in this case, higher than the average LD
457 among markers used to build the GRM. This occurs for two reasons. First, the common CVs
458 (MAF≥0.05) have higher MAF on average than the markers on the array (using either an
459 MAF≥0.01 or MAC≥5 cutoff for the Axiom-based GRM computation). Second, markers on arrays
460 are not chosen at random, but are typically chosen to minimally tag one another (to reduce
461 redundancy) and to maximally tag variants not on the array, leading to lower average LD among
462 markers than between markers and common variants not on the array (see also ref. [13]). To
463 understand if the overestimation of h^2_{SNP} for common CV phenotypes was unique to the Axiom
464 array positions, we reran the analysis with SNPs on the Illumina Omni2.5 array and observed
465 similar h^2_{SNP} inflations for common CVs on the Illumina array as well, although the impact of
466 sample stratification appeared to more strongly influence the Illumina chip, perhaps due to the
467 incorporation of a larger number of rare (MAF<0.01) variants on the Illumina array (Figs. S2 and
468 S10).

469 Utilizing imputed or WGS data to build the GRMs resulted in complex patterns of h^2_{SNP}
470 estimates depending on CV MAF class and stratification. Using a MAF>0.01 cutoff for imputed
471 SNPs in building the GRM resulted in patterns similar to array-based estimates above (Fig. S5-
472 S6), although the overestimates for common CVs were not as large, probably because the
473 imputed markers used to build the GRM included all common SNPs rather than an
474 overrepresentation of tag SNPs. On the other hand, when all imputed markers were used to
475 build the GRM, GREML-SC estimates depended strongly on stratification level and the CV MAF.
476 GREML-SC produced large overestimates for common CV phenotypes but underestimates for
477 rarer CV phenotypes in unstratified samples (Fig. S6), as previously noted in Yang et al.[12]. The

478 pattern was reversed for stratified samples: estimates of h^2_{SNP} were approximately unbiased for
479 common CV phenotypes but underestimated for uncommon-to-rare CV phenotypes and
480 overestimated for very rare CV phenotypes. Finally, when the frequency distribution of the CVs
481 matched that of the WGS (e.g., randomly drawn from all WGS variants), the estimates were
482 unbiased regardless of stratification when using WGS data to build the GRM (Fig. S5), but were
483 slightly underestimated when using imputed data (Fig. S6), presumably due to imperfect
484 imputation. The reason for this complex pattern of h^2_{SNP} estimates, where the effect of CV MAF
485 depended on stratification, was likely due to changes in CV-marker and marker-marker LD as a
486 function of stratification. The pattern of h^2_{SNP} estimates in unstratified samples is predictable
487 based on the logic outlined above: when CVs are more common than the markers used to build
488 the GRM, h^2_{SNP} is over-estimated, and vice-versa when CVs are less common than SNPs used
489 to build the GRM. In highly stratified samples, however, very rare variants tend to be ancestry-
490 specific and therefore weak proxies for variants elsewhere in the genome that predict ancestry
491 (long-range LD). This makes the LD between very rare CVs and markers that predict ancestry
492 elsewhere in the genome higher on average than the LD among the markers used to build the
493 GRM, thereby inflating h^2_{SNP} estimates for very rare CV phenotypes in stratified samples.
494 These results underscore that h^2_{SNP} estimates from GREML-SC, the typical approach
495 used, are sensitive to differences in average CV-marker LD vs. marker-marker LD (Fig. S11).
496 This difference itself depends on complex interplays between the CV MAF distribution, the
497 frequency distribution of markers used to build the GRM, and the level of stratification in the
498 sample. Thus, h^2_{SNP} estimates using single-component GREML are highly context dependent,
499 which may help explain the variation in estimates sometimes observed across studies for the

same traits. Fortunately, stratifying SNPs based on MAF and LD, to which we turn next, largely ameliorates these issues.

GREML using MAF-Stratified (GREML-MS) and LD- and MAF-Stratified (GREML-LDMS) GRMs

Genome partitioning using GREML-MS and GREML-LDMS produced h^2_{SNP} estimates that were substantially less biased and less sensitive to stratification than those from GREML-SC. GREML-MS h^2_{SNP} from array-based GRMs were underestimated for rarer CV phenotypes, as expected given the lack of LD between common array SNPs and rarer CVs, and were very slightly overestimated for common CV phenotypes, probably because of the LD properties of the SNPs chosen to be on the array (e.g., Illumina vs. Axiom positions, Fig. S10), as described in the previous section. GREML-MS using imputed variants slightly underestimated h^2_{SNP} for common to rare CV phenotypes. For very rare CV phenotypes, h^2_{SNP} was underestimated by ~18% in unstratified samples, likely due to poorer imputation quality for very rare SNPs, but underestimated by only ~7% in stratified samples. The higher estimates in stratified samples for very rare CVs is probably a lingering overestimation effect of long-range tagging of such variants in stratified samples. WGS-based estimates appeared unbiased for all combinations of CV MAF, relatedness, and stratification, with estimates all ~0.5.

Partitioning of the variance among the four MAF-stratified GRMs using GREML-MS allowed examination of the CV frequency distributions (Fig. S12, Table S4). GREML-MS estimated from GRMs built from array markers correctly apportioned the variation for common CV phenotypes, but as expected progressively underestimated h^2_{SNP} due to poor tagging of rare CVs with common SNPs (Fig. S12). Imputed variant GREML-MS provided more accurate

523 estimates of the CV frequency distributions, but still underestimated the effects of rare and very
524 rare CVs by as much as ~20% in unstratified samples (Fig. 3). Using WGS, the appropriate
525 proportion of the variance explained by each MAF-stratified GRM in the model was recovered
526 (Fig. S13, Table S4). Thus, the use of multiple GRMs based on MAF using imputed or WGS data
527 produces generally accurate GREML estimates of both h^2_{SNP} and the CV frequency distribution,
528 with only modest downward biases for very rare CVs when using imputed data.

529 The patterns of h^2_{SNP} estimates (Fig. 2-3, Figs. S12, S13) from GREML-LDMS were
530 almost identical to those from GREML-MS, which might be expected because the CVs in our
531 simulation were drawn at random within frequency bins and without regard to their LD. There
532 were, however, two minor differences between the GREML-LDMS and GREML-MS results. First,
533 for array-based GRMs (Fig. 2), estimates from GREML-LDMS for common CV phenotypes were
534 unbiased, whereas those for GREML-MS were slightly overestimated. As noted above, array
535 markers are more likely to tag common SNPs not on the array better than those on the array,
536 leading to higher CV-marker than marker-marker LD and creating a slight upward bias. By
537 binning by LD in addition to MAF, GREML-LDMS removes this source of bias, leading to
538 unbiased h^2_{SNP} estimates for common CVs. Second, for unknown reasons, GREML-LDMS using
539 whole genome sequence data gave slight (~3%) underestimates of h^2_{SNP} in highly stratified
540 samples for rare to common CVs, but not for very rare CVs (Fig. 2 and S5). This effect was not
541 apparent in imputed data, and may be simply sampling variance.

542 In summary, our findings suggest that using GREML-MS or GREML-LDMS on imputed
543 data generally leads to accurate estimates of h^2_{SNP} and the CV allele frequency distributions,
544 with only modest underestimation of variance due to rare and very rare CVs. Moreover, once
545 large enough WGS datasets become available, the underestimation of rarer CVs should be

546 largely ameliorated, although these methods can never estimate variance due to CVs that are so
547 rare as to be unshared in a given sample.
548
549 **Single Component and MAF-Stratified LD-Adjusted Kinships (LDAK-SC and LDAK-MS)**
550 Single component LD-adjusted estimates of the kinship matrix (LDAK-SC) downweights
551 markers that better tag other SNPs, thereby correcting for the overestimation of h^2_{SNP} observed
552 in GREML-SC for common CV phenotypes in array-based data due to redundant tagging (Fig.
553 2). As with other methods using GRMs based on array SNPs, LDAK-SC produced downwardly
554 biased h^2_{SNP} estimates for rarer CV phenotypes. Using the MAF-stratified approach (LDAK-MS)
555 resulted in similar patterns.
556 As with GREML-SC, using LDAK-SC on imputed data resulted in a complex set of biases
557 that depended on CV MAF, data type, and stratification, although the patterns of bias were
558 different. LDAK-SC h^2_{SNP} estimates using only common (MAF > 0.01) imputed variants were
559 similar to those using only array SNP positions. LDAK-SC using all imputed variants led to
560 roughly unbiased h^2_{SNP} estimates in unstratified samples, but led to h^2_{SNP} estimates that varied
561 wildly depending on the CV MAF in the stratified samples (Fig. 2). LDAK-MS on imputed variants
562 produced h^2_{SNP} estimates that were less biased that LDAK-SC, but nevertheless more biased
563 and more sensitive to stratification compared to those produced by GREML-MS on imputed data
564 (Fig. 2).
565 Using LDAK-SC on WGS data also resulted in biases. With only common variants, results
566 mirrored those found using array and imputed variants (Fig. S5). However, when all WGS
567 variants were used, h^2_{SNP} for very rare CV phenotypes was overestimated, especially in highly
568 stratified samples, but underestimated for all other phenotypes. When using LDAK-MS on WGS

569 data, the biases were less extreme. However, LDAK-MS resulted in over-estimated h^2_{SNP} for
570 common CVs and underestimated h^2_{SNP} for all other CV phenotypes (Fig. 2). Similar to LDAK-
571 MS estimates of total h^2_{SNP}, using LDAK-MS to partition genetic variance among MAF ranges,
572 produced estimates that were less precise and more biased than either GREML-MS or GREML-
573 LDMS for the array, imputed, or WGS based GRMs (Figs. 3, S7-S9, S12-S13).
574 Much of the observed patterns was likely due to the relationship between MAF and LDAK
575 weights (Fig. S14; ref.[12]) and differences in MAF distributions of array, imputed, and WGS
576 variants (Fig. S2). More very rare variants were observed and given higher weightings in the
577 WGS data than in either the imputed or array datasets. Similarly, in stratified datasets more very
578 rare variants were imputed (Fig. S2) and this likely contributed to stratification effects and
579 differences among imputed and WGS datasets.
580

581 **Extended Genealogy with Thresholded GRMs**

582 Patterns in the biases of h^2_{SNP} estimates were similar to those found using GREML-SC
583 (Fig. 2) when using the extended genealogy method, demonstrating that h^2_{SNP} estimates are
584 unaffected by the inclusion of close relatives so long as the model includes a second
585 (thresholded) GRM that contains only information on genomic sharing among close relatives.
586 However, the relative amount of variance attributable to the unthresholded GRM (estimating
587 h^2_{SNP}) versus the thresholded GRM (estimating $h^2_{IBS>t}$) varied considerably, and depended on
588 whether common (MAF>0.01; Fig. S15) or all (Fig. S16) markers were used to estimate the
589 GRMs. Using GRMs built from common (MAF>.01) array markers (Fig. S15), the estimate of
590 $h^2_{IBS>t}$ was negative, while h^2_{SNP} was overestimated for common CV phenotypes. As CVs
591 became rarer, $h^2_{IBS>t}$ grew while h^2_{SNP} shrunk, consistent with Zaitlen et al.'s interpretation that

592 $h^2_{IBS>t}$ would estimate variance due to rarer CVs. This pattern was more pronounced when all
593 markers were used (Fig. S16). Using imputed or WGS data, the pattern of negative variances
594 estimated for some of the GRMs remained. Nevertheless, estimates of total heritability, similar
595 to h^2_{FAM}, the sum of $h^2_{IBS>t}$ and h^2_{SNP}, were nearly unbiased or slightly downwardly biased in most
596 datasets and stratification subsamples (Fig. S15-S16). Even the total heritability of very rare CV
597 phenotypes was underestimated by less than 5%, regardless of the dataset used (SNP, WGS, or
598 imputed variant). It is important to note, however, that shared environmental effects can inflate
599 estimates of total h^2 using this method (see *Confounding between relatedness and shared*
600 *environments* below).
601

602 **Treelet Covariance Smoothing (TCS)**

603 Estimates of h^2_{SNP} from the TCS approach were highly unstable. Using samples of
604 unrelated individuals, the TCS method produced widely varying estimates of h^2_{SNP} depending on
605 the CV MAF, level of stratification, and type of data used to build the GRM (Fig. 2). We note that
606 the original implementation[21] used related individuals for h^2_{SNP} estimation; however, performance
607 did not improve when using samples of related individuals (Figs. S4-S6). The estimated and
608 empirical standard errors were substantially higher than any other estimation method (Fig. S7-
609 S9). Moreover, the pattern of results was complex and depended strongly on the simulation
610 condition; for estimates from GRMs built from array (Fig. S4) or imputed (Fig. S6) markers, h^2_{SNP}
611 was typically underestimated for all CV MAF frequencies irrespective of inclusion of close
612 relatives. However, h^2_{SNP} estimates were too high when WGS data was used for certain
613 combinations of CV MAF frequencies and stratification levels, and too low for others. It is
614 possible the TCS method would work better in samples that included more close relatives, but it

615 should be noted that other approaches (e.g., the thresholded GRM approach above) that rely
616 upon inclusion of close relatives produced unbiased total estimates with our sample sizes.
617

618 **LD Score Regression**

619 Estimates of h^2_{SNP} from LD Score Regression were similar when utilizing either Axiom
620 SNPs, imputed, or WGS data (Figs. 2 and S4-S6), as were estimates of the intercept (which
621 reflect the contribution of stratification and cryptic relatedness to the GWAS test statistics; Figs.
622 S17-S19). Across data types, h^2_{SNP} was generally slightly underestimated (5-10%) for common
623 CV phenotypes. This downward bias was slightly reduced in simulations using 10,000 causal
624 variants, but remained (Fig. S20); it is possible that this bias would be eliminated under the truly
625 infinitesimal model assumed by the model. h^2_{SNP} was increasingly underestimated for
626 phenotypes caused by increasingly rare CVs (Fig. 2), regardless of data type. This
627 underestimate of rare CV variation occurs because h^2_{SNP} is estimated only from common marker
628 (MAF>0.01) GWAS statistics[22], which are typically unaffected by rarer CVs. Interestingly, in the
629 highly stratified subsample, common CV phenotype h^2_{SNP} was overestimated with no covariate
630 correction with array SNPs, but controlling for PCs and sequencing cohorts using regression
631 (Figs. 2 and S4-S6) or a mixed-model approach (GCTA-LOCO; Fig. S4) removed this bias,
632 suggesting that h^2_{SNP} estimates from LD score regression are not immune to biases due to
633 stratification.

634 Estimates of h^2_{SNP} using MAF-partitioned LD score regression were highly variable, but in
635 many cases biased upwards (Fig. S4-S6). For common CV phenotypes, the estimates were less
636 biased than the standard LD score regression estimates described above. However, with rarer
637 CV phenotypes, regardless of the data used (array positions, imputed variants, or WGS data),

638 h^2_{SNP} was severely overestimated, expected when including very rare SNPs in the
639 regression[19,22,34].

640 The genomic control inflation factor, λ_{GC}, was greater in more stratified subsamples
641 without covariate correction, demonstrating the bias in GWAS with structure even in the absence
642 of confounding environmental effects (Figs. S17-S19), consistent with previous work that shows
643 structure alone can inflate GWAS test statistics[32,35,36] due to chance CV allele frequency
644 differences. We confirmed this using simulated data for two populations spanning a range of
645 structure (F_{ST}) and polygenicity without confounding environmental effects (Fig. S21). After
646 controlling for PC covariates using regression or by inclusion of a kinship matrix (using GCTA-
647 LOCO; Axiom SNPs only), there was limited effect of stratification, but λ_{GC} was still greater than
648 one for phenotypes derived from common, uncommon and rare CVs (Figs. S17-S19). That λ_{GC}
649 was not inflated for very rare CVs probably only reflects low statistical power for testing low MAF
650 markers.

651 The LD score regression intercept, which reflects the amount of confounding by
652 stratification and polygenicity[22], was greater than one when no covariate control was applied
653 across all stratification subsamples for all but the common CV traits (Figs. S17-S19). This was
654 stronger for the more stratified subsamples, as expected. The intercept was ~1 when the
655 covariates (and relatedness using Axiom SNPs) were accounted for, with the exception of
656 uncommon and rare CV phenotypes, which were slightly >1, suggesting that the control of
657 covariates was sufficient to account for the majority of the inflation in test statistics due to
658 stratification. We note that these simulations included no confounding environmental effects,
659 which may covary with stratification, and lead to inflation of GWAS statistics independent of the
660 inflation of the intercept observed here[22]. Nevertheless, such inflated GWAS statistics generally

561 should not be associated with the degree of LD-tagging of the markers, and thus should not
562 inflate estimates of h^2_{SNP}.
563
564 **Confounding between relatedness and shared environments**
565 We tested the effect of confounding between relatedness and shared environment
566 (simulated $c^2 = 0.1$) for GREML-SC, LD score regression, and thresholded GRMs, using
567 common array positions only. With unmodeled shared environmental effects, including all
568 relatives and using GREML-SC resulted in overestimates of h^2_{SNP}, especially for rare CV
569 phenotypes and for stratified samples (Fig. 4). However, when close relatives were removed at
570 thresholds of 0.05 or 0.1, shared environmental effects produced no additional upward bias (Fig.
571 4) over those observed when no shared environmental effects existed (Fig. 2) Thus, as
572 previously argued[6] removing close relatives appears to correct for this type of shared
573 environmental effect. Also as argued in ref.[22], h^2_{SNP} from LD score regression was not biased
574 upward due to unmodeled shared environmental effects, even when close relatives were
575 included. Finally, using the thresholded GRM method with environmental confounding, h^2_{SNP} was
576 biased slightly upward, particularly with a 0.1 relatedness threshold, but total heritability
577 overestimation reached 20%, consistent with all or almost all shared environmental variance
578 being estimated as additive genetic variance. Thus, care must be taken in interpreting results
579 from methods that use SNP GRMs to estimate heritability when related individuals are included;
580 shared environmental variance can masquerade as genetic variance.
581
582 ***Heritability of Complex Traits in the UK Biobank***

683 We applied the GREML-MS and GREML-LDMS approaches to six complex traits in the
684 UK Biobank data, and partitioned estimates of the heritability by marker MAF using either directly
685 genotyped Axiom SNPs or imputed genome-wide variants (Fig. 5, S23, Tables S6-S7). Total
686 h^2_{SNP} was on average 12% lower using imputed data rather than the directly genotyped Axiom
687 positions; our simulation results suggest this may be due to overestimation of variation due to
688 common CVs for array markers (Fig. S4) and slight underestimation of variation due to common
689 CVs for imputed markers (Fig. S6) using this method. The difference between array and imputed
690 data was most apparent in the estimates of h^2_{SNP} per MAF bin, where h^2_{SNP} was lower using
691 imputed data for common variant bins (MAF>0.01), but higher for rarer MAF bins. For example,
692 the rare MAF bins (MAF<0.01) accounted for 8.8% of the phenotypic variance of height using
693 imputed markers but only 0.6% using genotyped SNPs, whereas common MAF bins accounted
694 for 48% and 59%, respectively. Fluid intelligence was even more striking, with rarer SNPs
695 accounting for 11% and 3.4% using imputed and directly genotyped markers, respectively, while
696 common markers accounted for 14% and 20%. Our simulations results suggest the h^2_{SNP}
697 estimates from imputed data are more trustworthy.

698 Our simulation results also suggest that frequency distribution of CVs is best estimated
699 using imputed data. The h^2_{SNP} across MAF bins from a GREML-MS model in the UK Biobank
700 imputed data suggest real differences in genetic architectures across the six traits (Fig. 5) . For
701 example, height and adiposity phenotypes (BMI, impedance, and trunk fat) appear to be
702 influenced mostly be common CVs, whereas fluid intelligence appears to have an important
703 contribution from rare (MAF < 0.01) CVs. Results from GREML-MS (Fig. 5) were similar to those
704 from GREML-LDMS (Fig. S23, Table S7), although GREML-LDMS suggested that more trait

variance, even that attributable to common SNPs, was due to variants in the lower half of LD scores.

Our results suggest that most of the genetic variance of these traits is attributable to relatively common (MAF>0.01) variants. However, the contribution of increasingly rare CVs is likely to be underestimated for a few reasons. First, our simulations suggest that variation due to very rare CVs (0.0003<MAF<0.0025) is underestimated by ~ 20% due to low imputation quality of rarer variants. Second, this under-estimate was probably more severe in these results given the imputation reference panel used in the UK Biobank data was half the size of the reference panel used in our simulations. The variation due to CVs not present in the imputation reference panel used for the UK Biobank (UK10K and 1,000 Genomes) were missed in our results.

DISCUSSION

Performance of h^2_{SNP} Methods in Simulated Data

We have demonstrated that estimates of genetic variation using SNP data can be biased in a number of sometimes difficult to foresee ways, and depend strongly on a complex interplay between method used, the frequency distribution of CVs, the type of data used in the analysis, the degree of sample stratification, whether relatives are included or excluded, and the importance of shared environmental effects. Approaches that are able to explore genetic architecture of complex traits also differ in their ability to correctly estimate the CV frequency distributions. Understanding how the different methods behave under different contexts is crucial for proper interpretation of SNP-heritability estimates and for optimal design of future studies. There has been much debate surrounding the relative importance of common vs. rare variants

727 and the degree to which heritability remains unexplained (e.g., ref. [7,12,13]), and the findings
728 presented here offer context for how results from these methods inform these debates.
729 Through simulations, we have provided evidence that the use of WGS data, in
730 combination with genome partitioning methods such as GREML-MS or GREML-LDMS, results in
731 roughly unbiased h^2 estimates in unrelated samples, regardless of trait genetic architecture or
732 population stratification in the sample, although variation due to extremely rare variants (e.g., de
733 novo mutations) that are unshared between individuals in the sample will still be missed. Even
734 with the most comprehensive imputation reference panel available, using imputed genome-wide
735 markers still results in downwardly biased h^2_{SNP} estimates to the degree that rare variants are
736 important to trait variation, but not nearly to the degree observed when using array markers. This
737 is important, because it implies that the narrow-sense heritability remains underestimated in
738 current studies using imputed data. Even with datasets using large reference panels, such as the
739 UK Biobank data presented here, h^2_{SNP} from very rare CVs is likely underestimated due to poor
740 imputation of rare SNPs. As imputation reference panels, such as the HRC and the forthcoming
741 TOPMed panel[37], continue to grow in size and diversity, accurate imputation of increasingly rarer
742 variants will allow for increasingly accurate estimation of not the full narrow sense heritability, as
743 well as for increasingly accurate estimation of the frequency distribution of CVs. Alternatively,
744 novel methods, such as those that rely on sharing at identical-by-descent haplotypes rather than
745 allele sharing at measured SNPs[38], may better-capture effects of rare and poorly-tagged
746 variants, and is a potential future direction for estimating the variation due to rare CVs.
747 Linkage disequilibrium (LD) between CVs and markers is central to the methods reviewed
748 here. The observed patterns of over- and underestimation can be partly understood through the
749 effect of LD among causal variants and markers (Fig. S11). As Yang et al.[12] demonstrated, using

750 GREML-SC, h^2_{SNP} estimates should be unbiased when the average LD between markers and
751 CVs ($\overline{r^2}_{QM}$) is the same as the average LD among all markers ($\overline{r^2}_{MM}$), which occurs when
752 markers and CVs are sampled from the same allele frequency distribution. This explains the
753 underestimate of h^2_{SNP} using array genotypes when the CVs are rare, because common markers
754 on an array typically have lower LD with rare CVs than with other markers, leading to $\overline{r^2}_{QM} / \overline{r^2}_{MM}$
755 << 1 and $h^2_{SNP} << h^2$. On the other hand, when the CVs are a random sample of markers, this
756 ratio is ~1 and the estimated $h^2_{SNP} \approx h^2$. Finally, when the CVs are more common than markers
757 used to create the GRM, LD between common CVs and markers will typically be higher than LD
758 among markers, leading to $\overline{r^2}_{QM} / \overline{r^2}_{MM} > 1$ and $h^2_{SNP} > h^2$.

759 The bias arising from a mismatch in CV and marker frequency distributions is not
760 alleviated by weighting of markers by LD. Speed et al.[13] showed that redundant marker tagging
761 of CVs can bias h^2_{SNP} upward, and proposed weighting markers inversely to their LD score,
762 which partially mitigates this bias in sparse genotype data. However, using such weights in
763 dense whole genome sequence or imputed data leads to near 0 weights for most common
764 markers, typically leading to underestimates of heritability arising from common CVs and,
765 potentially, to overestimates of heritability from very rare CVs. What does appear to alleviate
766 both the bias arising from a mismatch in CV and marker frequency distributions as well as the
767 bias due to differential LD is binning markers by different MAF and LD bins[12]. When used on
768 imputed or sequence data, GREML-MS and GREML-LDMS provide the most accurate
769 partitioning of the variance and least biased total h^2_{SNP} estimates across genomic data types, CV
770 frequency distributions, and levels of stratification. Although we showed that WGS is the ideal
771 data source for creating GRMs, imputation will, for the time being, remain a cost-effective way to

772 capture most of the trait variation, and will only improve as sequencing initiatives continue to
773 amass larger, publicly available reference panels.

774 Our simulation results highlighted both limitations and advantages to LD score regression.
775 Although it uses a much different approach than GREML, LD score regression suffers from many
776 of the same problems as single-component GREML approaches. LD regression leverages the
777 fact that for common variants under an infinitesimal model, the effect size of a marker is related
778 to how well it tags the surrounding variants (and therefore how likely it is to tag a CV)[19,22].
779 Because LD is strongly related to MAF, the method increasingly underestimates variation as CVs
780 become rarer. Moreover, unlike GREML-MS, it provides unreliable estimates if used on rare
781 variants (MAF < .01), meaning that it cannot be used to accurately estimate CV frequency
782 distributions, or variation due to rare CVs, even if GWAS statistics from imputed or WGS data
783 are available. Nevertheless, LD score regression has several important advantages. Foremost
784 among them, it can be used on summary statistics alone, bypassing the need for raw genotype
785 data and allowing analyses based on sample sizes that would otherwise be impossible.
786 Furthermore, as argued by its originators and as we have shown, it is generally robust to
787 confounding biases due to stratification or shared family environmental effects, even when
788 relatives are included in the sample. Finally, it is readily applied to various marker annotations in
789 order to understand, for example, the relative importance of gene networks and functional
790 categories[19].

791 In our LD score regression simulation results, the contribution of common CVs to
792 phenotypic variance were slightly underestimated, regardless of the data type used (array SNPs,
793 imputed variants, or sequence data), a pattern previously reported[39]. This underestimate was not
794 seen in the simulations performed by Bulik-Sullivan et al.[22]. This difference may stem from the

795 fact that Bulik-Sullivan et al.[22] simulated phenotypes caused by a much larger proportion of
796 markers whereas we simulated phenotypes with only 1,000 or 10,000 CVs. Consistent with this
797 possibility, when we increased the number of CVs to 10,000, our estimates were somewhat less
798 biased. Nevertheless, it seems unlikely that the infinitesimal model truly holds for any phenotype,
799 and thus h^2_{SNP} estimates from LD-regression are likely to be biased downward, especially as
800 CVs become rarer.

801 There are several limitations to the findings presented here. First, although a subset of our
802 simulations included shared environmental effects among close relatives, we did not model more
803 complicated ways that environmental and genetic similarity can be confounded. For example, we
804 did not simulate "vertical transmission" models in which distant ancestry can lead to low levels of
805 environmental similarity, nor situations where environmental effects are confounded with
806 ancestry. Previous studies have investigated this latter issue[40,41], and fitting ancestry PCs
807 removes much of the bias.

808 Second, other than varying CV MAF frequency distributions, we did not simulate
809 situations where the LD of CVs differed systematically from the LD of markers used to estimate
810 the GRM. As Speed et al.[13] demonstrated, if CVs come from regions of low LD (e.g., DNase I-
811 hypersensitivity sites[42]), h^2_{SNP} will be underestimated and vice-versa when CVs come from
812 regions of high LD. Yang et al.[12] have shown that GREML-LDMS accounts for LD differences
813 between CVs and markers and provides unbiased estimates. However, as we shown (Fig. S7-
814 S9), standard errors for GREML-LDMS results are higher than GREML-MS. Given this tradeoff,
815 we recommend that investigators report results from both approaches, and trust those from
816 GREML-LDMS if there is a difference.

817 Third, we simulated CV effect sizes that were proportional to their minor allele frequencies
818 ($\propto [p(1-p)]^{-\alpha}$, where α = -1 in nomenclature of ref. [13]), so that the per-variant contribution to
819 heritability remained constant across MAF, similar to other studies[6,43,44]. The validity of this
820 assumption has been the subject of recent debate (e.g., ref. [13,45,46]) and it is clear that if this
821 assumption is unmet in real data, using a single component model will bias estimates, as several
822 well-designed evaluations of GREML-SC and LDAK-SC have shown[13]. However, two relevant
823 findings from those studies bear mentioning. First, the scaling we applied (α = -1) is the most
824 robust to violations of the model assumptions, and in sensitivity analyses of real data, scaling
825 with various approaches often led to qualitatively and quantitatively similar conclusions[12].
826 Second, the GREML-MS and GREML-LDMS stratified approaches allow variances to differ
827 across MAF partitions, effectively achieving the same goal as varying the scaling factor and
828 allowing a greater exploration of CV frequency distributions. An interesting avenue of future work
829 could be exploring possible values of α among functional annotations for evidence of purifying or
830 positive selection.
831

832 **h^2_{SNP} Estimates in the UK Biobank**

833 Using over 120,000 individuals with imputed genome-wide variants, we obtained
834 estimates of h^2_{SNP} for complex traits similar to those previously published using directly
835 genotyped markers and imputed genome wide markers for height and BMI (e.g., ref. [12]).
836 Estimates of h^2_{SNP} for measures of adiposity (impedance, trunk fat, and BMI) were similar to
837 each other, as expected given the relationship between these traits. Accounting for imperfect
838 imputation and using our simulation results as guidance, our results suggest that the true
839 narrow-sense heritability of height is 60-70%, and that of BMI is 20-30%, with some additional

840 variation possibly from very rare and poorly-imputed CVs. Furthermore, the majority (~80%) of
841 the additive genetic variance in these complex traits is explained by common variants with small
842 additive effects, with a smaller proportion attributable to rarer variants. This finding has been
843 discussed elsewhere[6,12,13]. This indicates that larger sample sizes will be required to identify
844 common variants of very small effects in GWAS, but that little still-missing additive genetic
845 variation remains.

846 The two behavioral traits we examined appear to have qualitatively different genetic
847 architectures. Little of the additive genetic variance in neuroticism was explained by rare
848 variants, but roughly half of fluid intelligence h^2_{SNP} was explained by rare variants with MAF <
849 0.01. Family- and twin- based estimates of heritability of intelligence are ~50%, while recent
850 studies using common SNPs have estimated h^2_{SNP} ~ 0.25[47,48]. Our estimates, using an
851 independent sample, are not dissimilar from these, and accounting for the downward bias in
852 h^2_{SNP} using imputed data, heritability is likely ~30%, with roughly half of that from rare variants,
853 and some additional variance caused by very rare and poorly-imputed CVs. However, given that
854 we know that variation due to increasingly rare CVs is increasingly underestimated, it is possible
855 that a larger proportion of the additive genetic variation in fluid intelligence is due to extremely
856 rare CVs. Nevertheless, 30% is substantially lower than the ~50% estimates from family-based
857 studies. However, it is also possible that these twin- and family-based estimates are
858 overestimated, and that little remaining heritability will be explained by increasingly rare CVs.
859 Our estimates of neuroticism heritability suggest that little of the variance is due to rare SNPs. In
860 the UK Biobank data, our estimate of h^2_{SNP} (0.09) is slightly higher than some published
861 estimates (h^2_{SNP} = 0.06[ref. [49]]), but lower than a recent study using the same UK Biobank data
862 (h^2_{SNP} = 0.14-16[ref. [50]]). This may be due to our use of MAF-stratified GREML, rather than

863 single component GREML with array data as in Smith et al.[50], which we have showed here leads
864 to overestimation of variance due to common CVs. Extended-twin family studies, which can
865 provide estimates of narrow-sense heritability while addressing concerns of shared
866 environmental and non-additive genetic influences, suggest that the narrow-sense heritability of
867 neuroticism is ~30%[10], which still leaves much of the additive genetic variance unexplained and
868 presents a puzzle to be solved by future investigation.
869

870 **Conclusions**

871 Heritability is a fundamental concept of genetics and its unbiased estimation is critical for
872 understanding complex trait genetics as well as for designing better studies and obtaining a
873 clearer picture of the possible explanatory power of GWAS. Below we provide our recommended
874 best practices for studies aiming to estimate h^2_{SNP} and CV frequency distributions for complex
875 traits. Even when applying these best approaches, heritability is still likely underestimated, but
876 will improve as larger sample sizes, larger imputation panels, and better methods to account for
877 rare variants are developed.

878 *Recommended Practices*

879 - Careful quality control in genetic data, for instance based on missingness and Hardy-
880 Weinberg equilibrium, is critical, particularly for case-control data and/or when the sample
881 is comprised of multiple cohorts[44].
882 - Include appropriate covariates, such as principal components, cohorts, and other potential
883 confounders as fixed effects in GREML models and in the GWAS models for LD score
884 regression.

885 • MAF- and/or LD-stratified GREML approaches[12] on WGS or imputed data provide the
886 most accurate estimates of h^2_{SNP} and CV frequency distributions. Even if CV frequency
887 distributions are not of interest, these methods provide the most accurate estimates of
888 h^2_{SNP} and are also the most robust to biases caused by stratification and differences
889 between the CV and marker allele frequency distributions. However, there is a bias-
890 precision tradeoff: more GRMs lead to larger standard errors, necessitating larger sample
891 sizes for these methods. We recommend to report results from both GREML-LDMS and
892 GREML-MS, and to trust the results of GREML-LDMS if there is a meaningful difference.
893 • If possible, run GREML models on WGS data if available, and otherwise data imputed
894 using the largest and most diverse reference panel possible. Currently, this is the HRC[24].
895 • If raw genomic data is not available, use LD score regression on summary statistics, but
896 calculate LD scores using a large sequence reference panel. Estimates from LD score
897 regression are typically lower than those produced by GREML-SC on array data.
898 • Related individuals may share common environmental and non-additive genetic effects
899 that can inflate estimates of h^2_{SNP}. Removing related individuals provides estimates that
900 are less likely to be inflated by such environmental and non-additive genetic factors.
901 • Most reports of h^2_{SNP} in the literature have used the GREML-SC approach. However, as
902 we have demonstrated, these estimates are subject to a number of sometimes conflicting
903 biases, making interpretation of GREML-SC results challenging. Most crucially, GREML-
904 SC is especially sensitive to the similarity between the frequency distributions of the CVs
905 and the markers used to create the GRM, which can differ across genomic data types and
906 array types. Moreover, GREML-SC can be sensitive to stratification effects, even when
907 ancestry covariates are included in the model.

CONFLICTS OF INTEREST

The authors declare no competing financial interests.

SUPPLEMENTAL DATA DESCRIPTION

The supplemental data includes 22 additional figures and 7 additional tables.

CONSORTIA

Haplotype Reference Consortium:

1.

Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, Yang Luo, Carlo Sidore, Alan Kwong, Nicholas Timpson, Seppo Koskinen, Scott Vrieze, Laura J Scott, He Zhang, Anubha Mahajan, Jan Veldink, Ulrike Peters, Carlos Pato, Cornelia M van Duijn, Christopher E Gillies, Ilaria Gandin, Massimo Mezzavilla, Arthur Gilly, Massimiliano Cocca, Michela Traglia, Andrea Angius, Jeffrey C Barrett, Dorrett Boomsma, Kari Branham, Gerome Breen, Chad M Brummett, Fabio Busonero, Harry Campbell, Andrew Chan, Sai Chen, Emily Chew, Francis S Collins, Laura J Corbin, George Davey Smith, George Dedoussis, Marcus Dorr, Aliki-Eleni Farmaki, Luigi Ferrucci, Lukas Forer, Ross M Fraser, Stacey Gabriel, Shawn Levy, Leif Groop, Tabitha Harrison, Andrew Hattersley, Oddgeir L Holmen, Kristian Hveem, Matthias Kretzler, James C Lee, Matt McGue, Thomas Meitinger, David Melzer, Josine L Min, Karen L Mohlke, John B Vincent, Matthias Nauck, Deborah Nickerson, Aarno Palotie, Michele Pato,

Nicola Pirastu, Melvin McInnis, J Brent Richards, Cinzia Sala, Veikko Salomaa, David Schlessinger, Sebastian Schoenherr, P Eline Slagboom, Kerrin Small, Timothy Spector, Dwight Stambolian, Marcus Tuke, Jaakko Tuomilehto, Leonard H Van den Berg, Wouter Van Rheenen, Uwe Volker, Cisca Wijmenga, Daniela Toniolo, Eleftheria Zeggini, Paolo Gasparini, Matthew G Sampson, James F Wilson, Timothy Frayling, Paul I W de Bakker, Morris A Swertz, Steven McCarroll, Charles Kooperberg, Annelot Dekker, David Altshuler, Cristen Willer, William Iacono, Samuli Ripatti, Nicole Soranzo, Klaudia Walter, Anand Swaroop, Francesco Cucca, Carl A Anderson, Richard M Myers, Michael Boehnke, Mark I McCarthy, Richard Durbin, Gonçalo Abecasis, & Jonathan Marchini

ACKNOWLEDGMENTS

This work was supported by NIH grant R01MH100141 (to MCK), NHMRC grants 1078037 (PMV) and 1113400 (PMV and JY), and Sylvia & Charles Viertel Charitable Foundation Senior Medical Research Fellowship (JY). We thank the participants of the individual HRC cohorts. This research has been conducted using the UK Biobank Resource. We thank Doug Speed for providing LDAK5. We thank the Keller and Vrieze lab groups, the Institute for Behavioral Genetics, the CU Research Computing facility, and Sean Caron.

WEB RESOURCES

BOLT-REML: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/

954 GCTA: http://cnsgenomics.com/software/gcta/index.html

955 Haplotype Reference Consortium: http://www.haplotype-reference-consortium.org/home

956 LD score regression: github.com/bulik/ldsc/wiki

957 LDAK: http://dougspeed.com/ldak/

958 UK Biobank: http://www.ukbiobank.ac.uk/

959

960

961

962 **REFERENCES**

963

964 1. Tenesa, A., and Haley, C.S. (2013). The heritability of human disease: estimation, uses and

965 abuses. Nat. Rev. Genet. *14*, 139–149.

966 2. Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era--concepts

967 and misconceptions. Nat. Rev. Genet. *9*, 255–266.

968 3. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada,

969 K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and

970 biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

971 4. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS

972 discovery. Am. J. Hum. Genet. *90*, 7–24.

973 5. Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-H., Holmans, P.A., Lee, P., Bulik-

974 Sullivan, B., Collier, D.A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-

975 associated genetic loci. Nature *511*, 421–427.

976 6. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P. a,

977 Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large
978 proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
979 7. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing
980 heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. U. S. A. *109*,
981 1193–1198.
982 8. Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L.
983 (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and
984 dichotomous traits. PLoS Genet. *9*,.
985 9. Polderman, T.J.C., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher,
986 P.M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty
987 years of twin studies. Nat. Genet. *47*, 702–709.
988 10. Keller, M.C., and Coventry, W.L. (2005). Quantifying and addressing parameter
989 indeterminacy in the classical twin design. Twin Res. Hum. Genet. *8*, 201–213.
990 11. Eaves, L.J., Last, K.A., Young, P.A., and Martin, N.G. (1978). Model-fitting approaches to the
991 analysis of human behaviour. Heredity (Edinb). *41*, 249–320.
992 12. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R.,
993 Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J. V, et al. (2015). Genetic variance estimation
994 with imputed variants finds negligible missing heritability for human height and body mass index.
995 Nat. Genet. *47*, 1114–1120.
996 13. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability
997 estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.
998 14. Maier, R., Moser, G., Chen, G.-B., Ripke, S., Coryell, W., Potash, J.B., Scheftner, W.A., Shi,
999 J., Weissman, M.M., Hultman, C.M., et al. (2015). Joint analysis of psychiatric disorders

000 increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive
001 disorder. Am. J. Hum. Genet. *96*, 283–294.
002 15. Hyde, C.L., Nagle, M.W., Tian, C., Chen, X., Paciga, S.A., Wendland, J.R., Tung, J.Y.,
003 Hinds, D.A., Perlis, R.H., and Winslow, A.R. (2016). Identification of 15 genetic loci associated
004 with risk of major depression in individuals of European descent. Nat. Genet. *48*, 1031–1036.
005 16. Okbay, A., Baselmans, B.M.L., De Neve, J.-E., Turley, P., Nivard, M.G., Fontana, M.A.,
006 Meddens, S.F.W., Linnér, R.K., Rietveld, C.A., Derringer, J., et al. (2016). Genetic variants
007 associated with subjective well-being, depressive symptoms, and neuroticism identified through
008 genome-wide analyses. Nat. Genet. *633*, 1–13.
009 17. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de
010 Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of
011 genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.
012 18. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J.,
013 Psychiatric Genomics Consortium, S.W.G., de Candia, T.R., Lee, S.H., Wray, N.R., et al. (2015).
014 Contrasting regional architectures of schizophrenia and other complex diseases using fast
015 variance components analysis. Nat. Genet. *47*, 1385–1392.
016 19. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V.,
017 Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using
018 genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
019 20. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: A tool for genome-
020 wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
021 21. Crossett, A., Lee, A.B., Klei, L., Devlin, B., and Roeder, K. (2013). Refining genetically
022 inferred relationships using Treelet Covariance Smoothing. Ann. Appl. Stat. *7*, 669–690.

023 22. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Consotrium, S.W.G. of
024 the P.G., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression
025 distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*,
026 291–295.
027 23. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P.,
028 Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the
029 Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. *12*, 1–10.
030 24. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M.,
031 Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes
032 for genotype imputation. Nat. Genet. *48*, 1279–1283.
033 25. Abraham, G., and Inouye, M. (2014). Fast principal component analysis of large-scale
034 genome-wide data. PLoS One *9*, e92766.
035 26. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., Purcell, S.,
036 Neale, B., Todd-Brown, K., Thomas, L., et al. (2015). Second-generation PLINK: rising to the
037 challenge of larger and richer datasets. Gigascience *4*, 7.
038 27. Team, R.C. (2015). R: A language and environment for statistical computing. R Foundation
039 for Statistical Computing, Vienna, Austria.
040 28. Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing
041 for disease and population genetic studies. Nat. Methods *10*, 5–6.
042 29. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew,
043 E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and
044 methods. Nat. Genet. *48*, 1284–1287.
045 30. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C.,

046 Visscher, P.M., and Wray, N.R. (2012). Estimating the proportion of variation in susceptibility to
047 schizophrenia captured by common SNPs. Nat. Genet. *44*, 247–250.
048 31. Liu, D., and Gaugler, T. (2015). treelet: An adaptive multi-scale basis for high-dimensional,
049 sparse and unordered data. R package version 1.1.
050 32. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages
051 and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106.
052 33. Xia, C., Amador, C., Huffman, J., Trochet, H., Campbell, A., Porteous, D., Hastie, N.D.,
053 Hayward, C., Vitart, V., Navarro, P., et al. (2016). Pedigree- and SNP-Associated Genetics and
054 Recent Environment are the Major Contributors to Anthropometric and Cardiometabolic Trait
055 Variation. PLoS Genet. *12*, e1005804.
056 34. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Consortium, R.,
057 Genomics Consortium, P., of the Wellcome Trust Consortium, G.C. for A., Perry, J.R.B.,
058 Patterson, N., et al. (2015). An Atlas of Genetic Correlations across Human Diseases and Traits.
059 Nat. Genet. *47*, 1236–1241.
060 35. Marchini, J., Cardon, L.R., Phillips, M.S., and Donnelly, P. (2004). The effects of human
061 population structure on large genetic association studies. Nat. Genet. *36*, 512–517.
062 36. Price, A.L., Zaitlen, N. a, Reich, D., and Patterson, N. (2010). New approaches to population
063 stratification in genome-wide association studies. Nat. Rev. Genet. *11*, 459–463.
064 37. TOPMed NHLBI.
065 38. Browning, S.R., and Browning, B.L. (2013). Identity-by-descent-based heritability analysis in
066 the Northern Finland Birth Cohort. Hum. Genet. *132*, 129–138.
067 39. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Nolte, I.M., van Vliet-
068 Ostaptchouk, J. V., Snieder, H., Esko, T., Milani, L., et al. (2015). Genome-wide genetic

069 homogeneity between sexes and populations for human height and body mass index. Hum. Mol.
070 Genet. *24*, 7445–7449.
071 40. Browning, S.R., and Browning, B.L. (2011). Population structure can inflate SNP-based
072 heritability estimates. Am. J. Hum. Genet. *89*, 191–193.
073 41. Goddard, M.E., Lee, S.H., Yang, J., Wray, N.R., and Visscher, P.M. (2011). Response to
074 Browning and Browning. Am. J. Hum. Genet. *89*, 193–195.
075 42. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhj??lmsson, B.J., Xu, H., Zang, C., Ripke,
076 S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning heritability of regulatory and cell-type-
077 specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
078 43. Zaitlen, N., Pasaniuc, B., Sankararaman, S., Bhatia, G., Zhang, J., Gusev, A., Young, T.,
079 Tandon, A., Pollack, S., Vilhjálmsson, B.J., et al. (2014). Leveraging population admixture to
080 characterize the heritability of complex traits. Nat. Genet. *46*, 1356–1362.
081 44. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing
082 heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
083 45. Lee, S.H., Yang, J., Chen, G.B., Ripke, S., Stahl, E.A., Hultman, C.M., Sklar, P., Visscher,
084 P.M., Sullivan, P.F., Goddard, M.E., et al. (2013). Estimation of SNP heritability from dense
085 genotype data. Am. J. Hum. Genet. *93*, 1151–1155.
086 46. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2013). Response to Lee et al.:
087 SNP-based heritability analysis with dense data. Am. J. Hum. Genet. *93*, 1155–1157.
088 47. Plomin, R., and Deary, I. (2014). Genetics and intelligence differences: five special findings.
089 Mol. Psychiatry *20*, 98–108.
090 48. Davies, G., Armstrong, N., Bis, J.C., Bressler, J., Chouraki, V., Giddaluru, S., Hofer, E.,
091 Ibrahim-Verbaas, C.A., Kirin, M., Lahti, J., et al. (2015). Genetic contributions to variation in

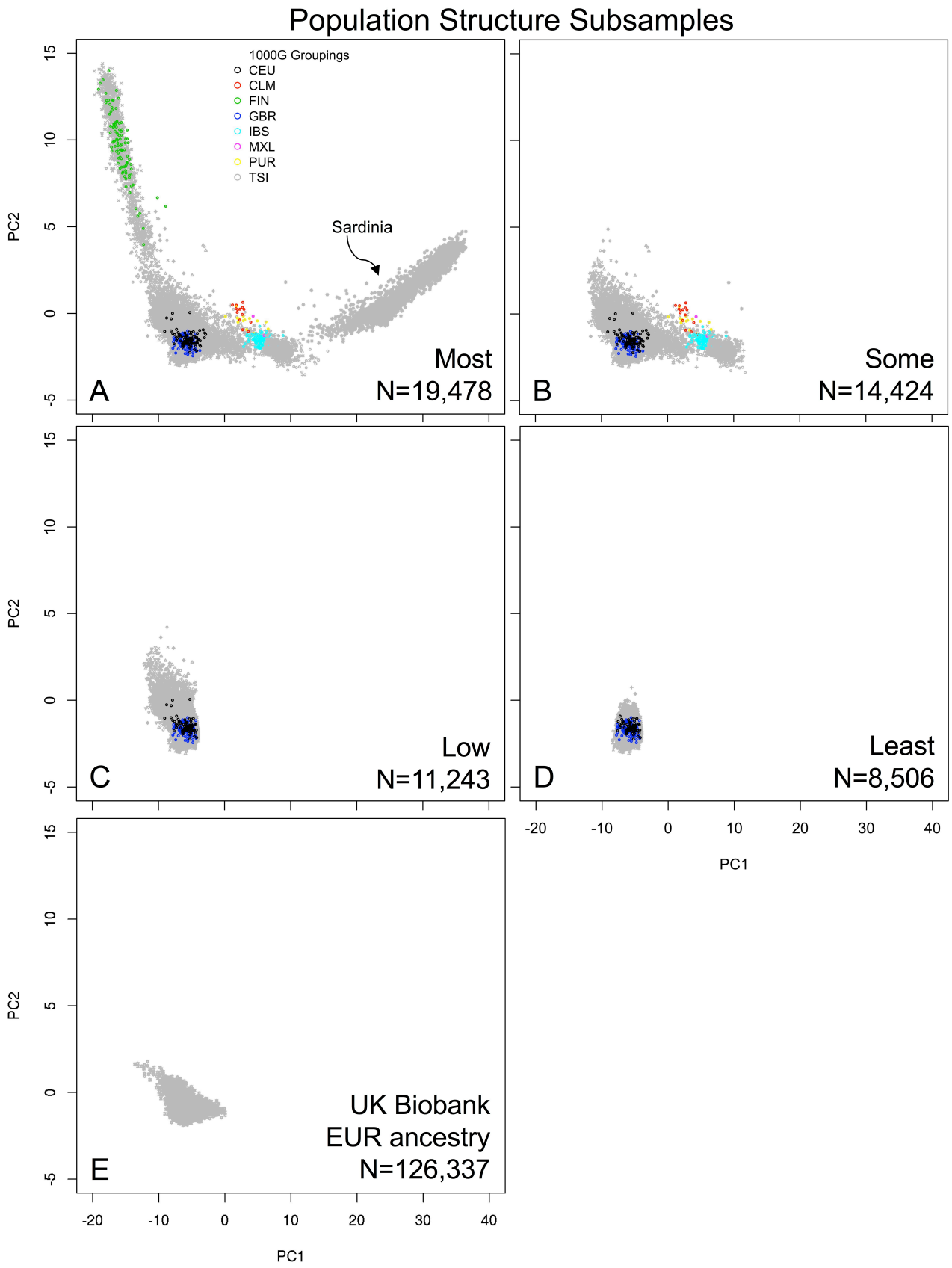
092 general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE
093 consortium (N=53949). Mol. Psychiatry *20*, 183–192.
094 49. Vinkhuyzen, A., Pedersen, N.L., Yang, J., Lee, S.H., Magnusson, P.K.E., Iacono, W.G.,
095 McGue, M., Madden, P., Heath, a C., Luciano, M., et al. (2012). Common SNPs explain some of
096 the variation in the personality dimensions of neuroticism and extraversion. Transl. Psychiatry *2*,
097 e102.
098 50. Smith, D.J., Escott-Price, V., Davies, G., Bailey, M.E.S., Colodro-Conde, L., Ward, J.,
099 Vedernikov, A., Marioni, R., Cullen, B., Lyall, D., et al. (2016). Genome-wide analysis of over 106
100 000 individuals identifies 9 neuroticism-associated loci. Mol. Psychiatry *21*, 749–757.
101
102
103 **FIGURE TITLES AND LEGENDS**
104
105 **Figure 1.** Population structure subsamples of European ancestry individuals in the HRC (A-D).
106 and UK Biobank individuals projected onto these axes (E). Total sample sizes are shown in
107 each panel. To keep sample size constant across stratification level, we randomly sampled
108 8,201 individuals with relatedness < 0.1 (the number of unrelated individuals in the most
109 homogeneous and smallest set in panel D) from each subsample to create the subsamples used
110 in the simulations.
111
112 **Figure 2.** Average h^2_{SNP} estimates across 100 replicates (± SEM) from GRMs built from Axiom
113 array positions (left), whole genome sequence data (center), or imputed genome-wide variants
114 (right). Horizontal panels show MAF ranges (specified in insert) of 1,000 randomly chosen causal

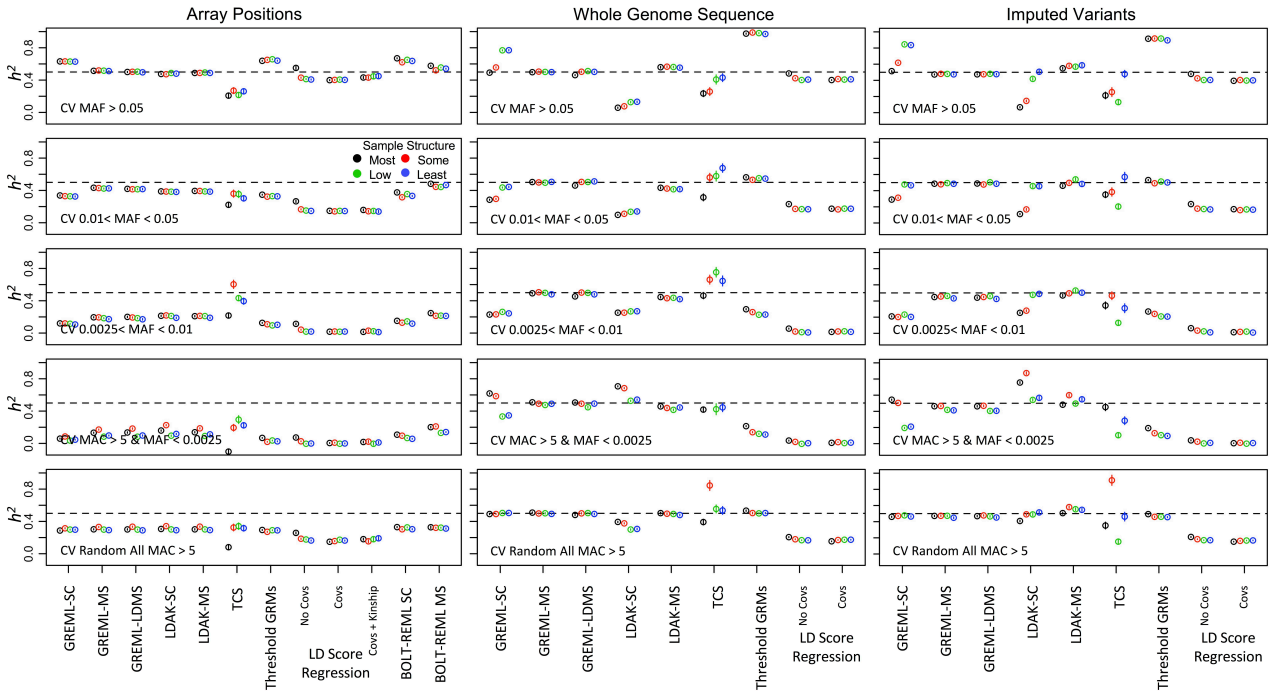
variants (CVs). Methods are listed on the X-axis as follows: Single component GREML (GREML-SC); MAF-stratified GREML (GREML-MS); LD- & MAF-stratified GREML (GREML-LDMS); Single-component Linkage Disequilibrium-Adjusted Kinships (LDAK-SC); MAF-stratified LDAK (LDAK-MS); Treelet Covariance Smoothing (TCS); Extended Genealogy with Thresholded GRMs; LD Score Regression using no PCs as covariates in GWAS, using PCs as covariates, or using both PCs and the kinship matrix; and Single Component and MAF-stratified BOLT-REML. Estimates are from samples of unrelated individuals (relatedness <0.05) except for samples used in the Threshold GRM method, which included all individuals. For the Threshold GRM method we plot h^2_{SNP} rather than total h^2 (h^2_{SNP} + $h^2_{ibs>t}$) from models where t = .05. Dotted line is the simulated (true) h^2 = 0.5. Colors represent the 4 subsamples varying in genetic structure. See Figs. S4-6 for estimates using different relatedness thresholds.

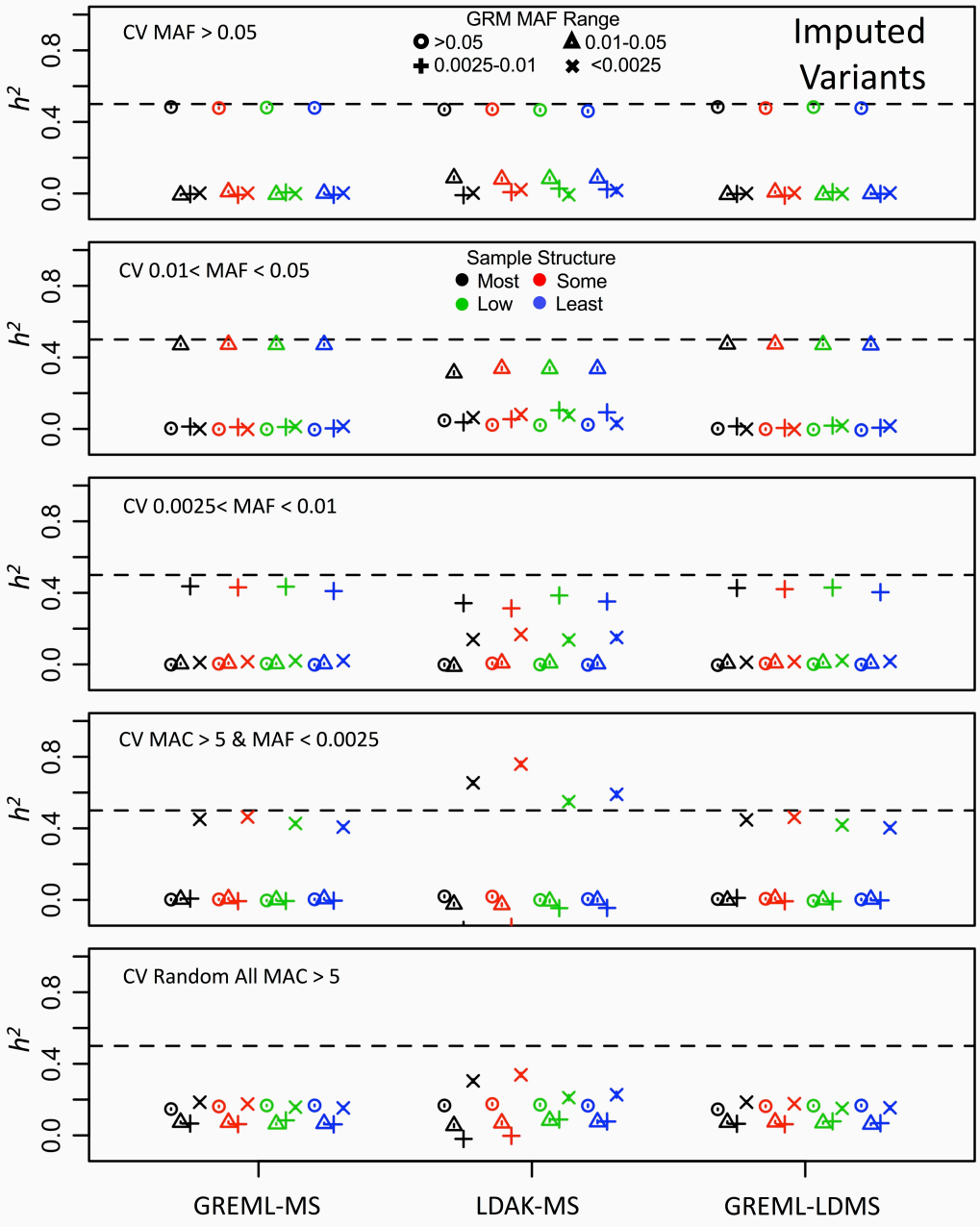
Figure 3. Average of 100 h^2_{SNP} estimates (± SEM) from GRMs constructed from imputed genome-wide variants of different MAF ranges (different symbols) in samples of unrelated (<0.05) individuals. Horizontal panels show MAF ranges (specified in insert) of 1,000 randomly chosen CVs and colors represent the 4 subsamples varying in genetic structure. GREML-MS & GREML-LDMS partition the phenotypic variance to the correct MAF-range GRM, while LDAK-MS often attributed genetic variance to incorrect GRMs.

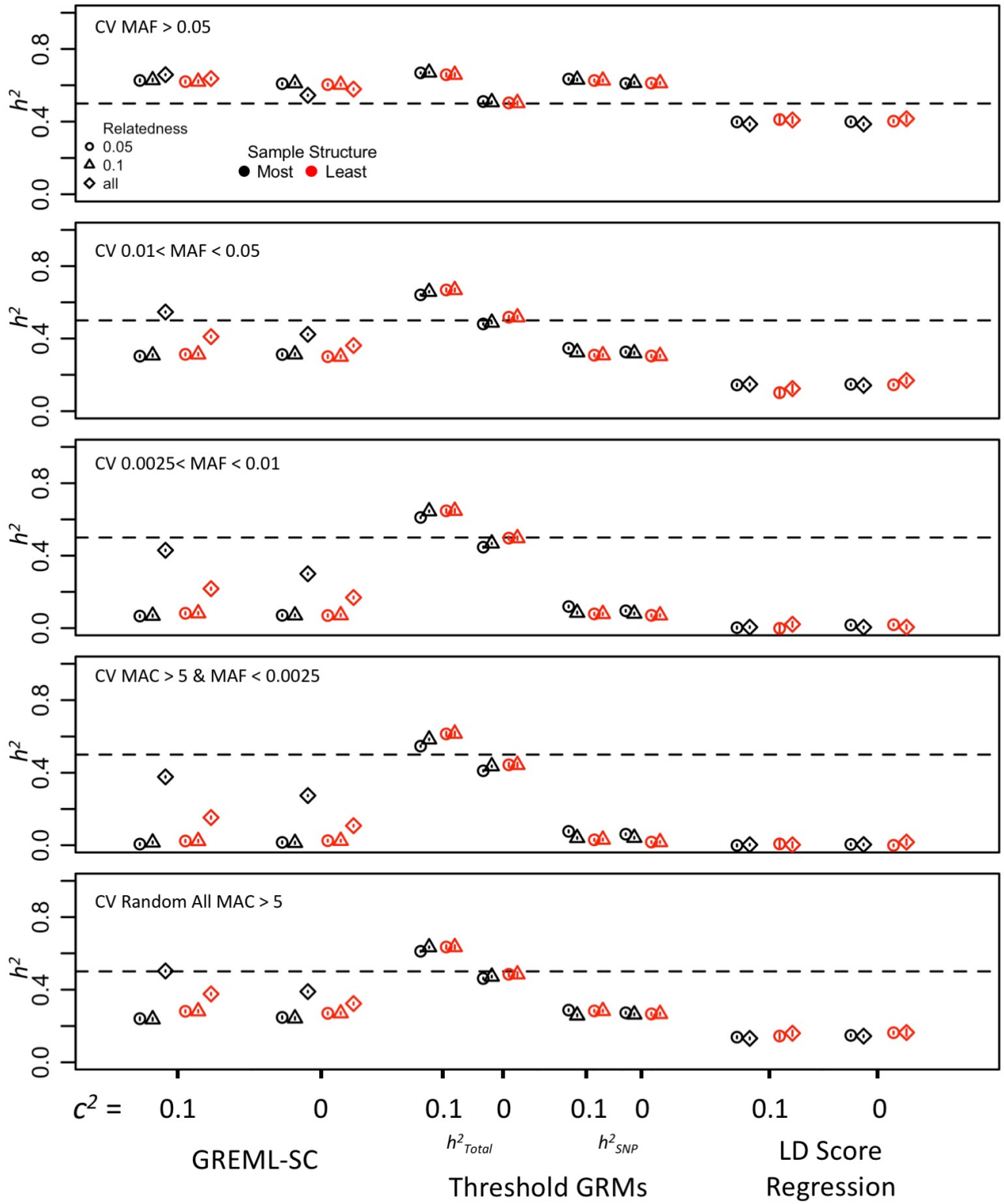
Figure 4. Mean heritability estimates (± SEM) from 100 replicates of phenotypes simulated with or without confounding shared environmental effects among families for three different methods (x axis) for different genetic architectures. GRMs were estimated using common (MAF>0.01) array SNP positions for the most structured and most homogeneous stratification subsamples

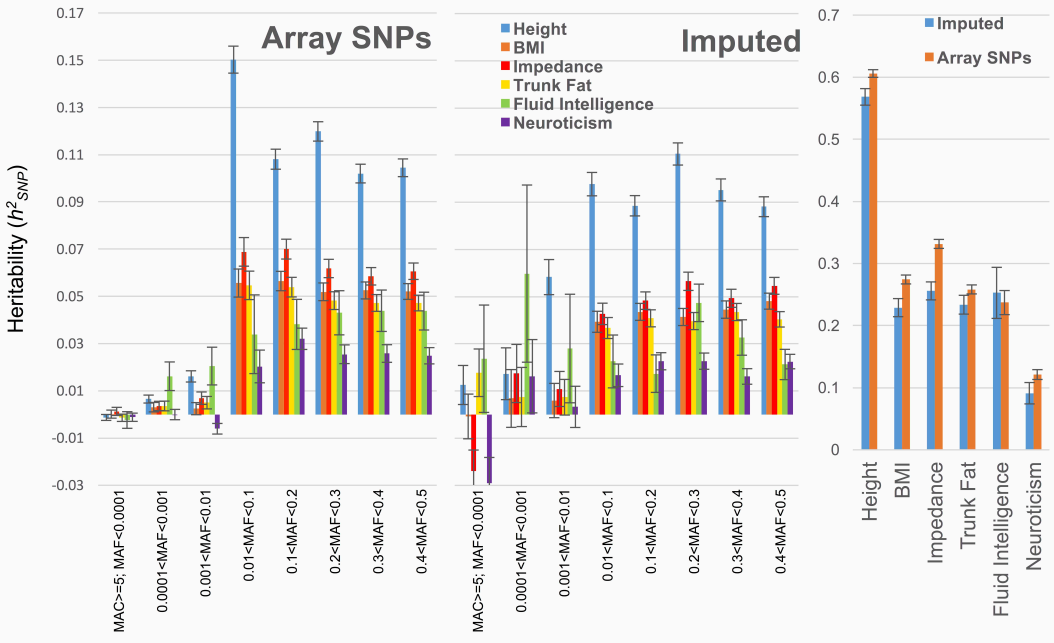
138 only. Different symbols indicate the relatedness cutoffs used. For GREML-SC, we used three
139 thresholds, including no relatedness cutoff (all individuals included). For LD Score Regression,
140 we did not apply a 0.1 relatedness cutoff, as most studies will use a 0.05 or lower threshold for
141 individuals included in GWAS. The threshold GRM approach requires all individuals, and the
142 different symbols indicates the relatedness threshold (*t*) below which the thresholded GRM was
143 set to 0. h^2_{Total} is the sum of both variance components, h^2_{SNP} is the variance component of the
144 unthresholded GRM. Each horizontal panel indicates the minor allele frequency (MAF) range of
145 the 1,000 randomly chosen causal variants (CV), with the range specified in the inset.
146
147 **Figure 5.** Estimates of MAF partitioned h^2_{SNP} using GREML-MS on Axiom array SNPs (left) and
148 imputed genome-wide variants (center) for six complex traits in the UK Biobank. Total h^2_{SNP}
149 shown on right.
150
151
152
153


Population Structure Subsamples
1000G Groupings
CEU
CLM
FIN
GBR
IBS
MXL
PUR
TSI
Sardinia
Most
N=19,478
A
Some
N=14,424
B
Low
N=11,243
C
Least
N=8,506
D
UK Biobank
EUR ancestry
N=126,337
E
PC1
PC2


Array Positions
Whole Genome Sequence
Imputed Variants
h^2
0.0
0.4
0.8
CV MAF > 0.05
CV 0.01< MAF < 0.05
CV 0.0025< MAF < 0.01
CV MAC > 5 & MAF < 0.0025
CV Random All MAC > 5
Sample Structure
Most
Some
Low
Least
GREML-SC
GREML-MS
GREML-LDMS
LDAK-SC
LDAK-MS
TCS
Threshold GRMs
No Covs
Covs
Covs + Kinship
BOLT-REML SC
BOLT-REML MS
LD Score Regression


Imputed Variants
GRM MAF Range
>0.05
0.01-0.05
0.0025-0.01
<0.0025
Sample Structure
Most
Some
Low
Least
CV MAF > 0.05
CV 0.01< MAF < 0.05
CV 0.0025< MAF < 0.01
CV MAC > 5 & MAF < 0.0025
CV Random All MAC > 5
h2
0.0
0.4
0.8
GREML-MS
LDAK-MS
GREML-LDMS


CV MAF > 0.05
CV 0.01< MAF < 0.05
CV 0.0025< MAF < 0.01
CV MAC > 5 & MAF < 0.0025
CV Random All MAC > 5
h^2
0.0
0.4
0.8
Relatedness
0.05
0.1
all
Sample Structure
Most
Least
c^2 =
0.1
0
0.1
0
0.1
0
0.1
0
h^2_{Total}
h^2_{SNP}
GREML-SC
Threshold GRMs
LD Score Regression


Array SNPs
Imputed
Height
BMI
Impedance
Trunk Fat
Fluid Intelligence
Neuroticism
Heritability (h^2_{SNP})
0.17
0.15
0.13
0.11
0.09
0.07
0.05
0.03
0.01
-0.01
-0.03
MAC>=5; MAF<0.0001
0.0001<MAF<0.001
0.001<MAF<0.01
0.01<MAF<0.1
0.1<MAF<0.2
0.2<MAF<0.3
0.3<MAF<0.4
0.4<MAF<0.5
Imputed
Array SNPs
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Height
BMI
Impedance
Trunk Fat
Fluid Intelligence
Neuroticism