TITLE

Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores

AUTHORS AND AFFILIATIONS

Julian R. Homburger,[1] Cynthia L. Neben,[1] Gilad Mishne,[1] Alicia Y. Zhou,[1] Sekar Kathiresan,[2-4] Amit V. Khera[2-4] *

[1]Color Genomics, 831 Mitten Road, Suite 100, Burlingame, CA, 94010 USA

[2]Center for Genomic Medicine and Cardiology Division, Department of Medicine, Massachusetts General Hospital, Boston, MA, 02114 USA

[3]Cardiovascular Disease Initiative of the Broad Institute of MIT and Harvard, Cambridge, MA, 02142 USA

[4]Harvard Medical School, Boston, MA, 02115 USA

CORRESPONDING AUTHOR

*Amit V. Khera, MD, MSc

Center for Genomic Medicine

Massachusetts General Hospital

Simches Research Building | CPZN 6.256

Boston, MA 02114 USA

Tel: 617.726.7876

Email: avkhera@mgh.harvard.edu

ABSTRACT

Background: The inherited susceptibility of common, complex diseases may be caused by rare, 'monogenic' pathogenic variants or by the cumulative effect of numerous common, 'polygenic' variants. As such, comprehensive genome interpretation could involve two distinct genetic testing technologies -- high coverage next generation sequencing for known genes to detect pathogenic variants and a genome-wide genotyping array followed by imputation to calculate genome-wide polygenic scores (GPSs). Here we assessed the feasibility and accuracy of using low coverage whole genome sequencing (lcWGS) as an alternative to genotyping arrays to calculate GPSs.

Methods: First, we performed downsampling and imputation of WGS data from ten individuals to assess concordance with known genotypes. Second, we assessed the correlation between GPSs for three common diseases -- coronary artery disease (CAD), breast cancer (BC), and atrial fibrillation (AF) -- calculated using lcWGS and genotyping array in 184 samples. Third, we assessed concordance of lcWGS-based genotype calls and GPS calculation in 120 individuals with known genotypes, selected to reflect diverse ancestral backgrounds. Fourth, we assessed the relationship between GPSs calculated using lcWGS and disease phenotypes in 11,502 European individuals seeking genetic testing.

Results: We found imputation accuracy r^2 values of greater than 0.90 for all ten samples -- including those of African and Ashkenazi Jewish ancestry -- with lcWGS data at 0.5X. GPSs calculated using both lcWGS and genotyping array followed by imputation in 184 individuals were highly correlated for each of the three common diseases (r^2 = 0.93 - 0.97) with similar score distributions. Using lcWGS data from 120 individuals of diverse ancestral backgrounds, including South Asian, East Asian, and Hispanic individuals, we found similar results with respect to imputation accuracy and GPS correlations. Finally, we calculated GPSs for CAD, BC,

and AF using lcWGS in 11,502 European individuals, confirming odds ratios per standard deviation increment in GPSs ranging 1.28 to 1.59, consistent with previous studies.

Conclusions: Here we show that lcWGS is an alternative approach to genotyping arrays for common genetic variant assessment and GPS calculation. lcWGS provides comparable imputation accuracy while also overcoming the ascertainment bias inherent to variant selection in genotyping array design.

KEYWORDS

Genome-wide polygenic score; low coverage whole genome sequencing; coronary artery disease; breast cancer; atrial fibrillation

BACKGROUND

Cardiovascular disease and cancer are common, complex diseases that remain leading causes of global mortality [1]. Long recognized to be heritable, recent advances in human genetics have led to consideration of DNA-based risk stratification to guide prevention or screening strategies. In some cases, such conditions can be caused by rare, 'monogenic' pathogenic variants that lead to a several-fold increased risk -- important examples are pathogenic variants in *LDLR* that cause familial hypercholesterolemia and pathogenic variants in *BRCA1* and *BRCA2* that underlie hereditary breast and ovarian cancer syndrome. However, the majority of individuals afflicted with these diseases do not harbor any such pathogenic variants. Rather, the inherited susceptibility of many complex traits and diseases is often 'polygenic,' driven by the cumulative effect of numerous common variants scattered across the genome [2].

Genome-wide polygenic scores (GPSs) provide a way to integrate information from numerous sites of common variation into a single metric of inherited susceptibility and are now able to identify individuals with a several-fold increased risk of common, complex diseases, including coronary artery disease (CAD), breast cancer (BC), and atrial fibrillation (AF) [3]. For example, for CAD, we noted that 8% of the population inherits more than triple the normal risk on the basis of polygenic variation, a prevalence more than 20-fold higher than monogenic familial hypercholesterolemia variants in *LDLR* that confer similar risk [3].

Comprehensive genome interpretation for common, complex disease therefore could involve both high-fidelity sequencing of important driver genes to identify potential monogenic risk pathogenic variants and a survey of all common variants across the genome to enable GPS calculation. High coverage whole genome sequencing (hcWGS; for example, 30X coverage) will likely emerge as a single genetic testing strategy, but current prices remain a barrier to large-scale adoption. Instead, the traditional approach has mandated use of two distinct genetic

88 testing technologies -- high coverage next generation sequencing (NGS) of important genes to
89 detect pathogenic variants and a genome-wide genotyping array followed by imputation to
90 calculate GPSs.
91
92 Low coverage whole genome sequencing (lcWGS; for example, 0.5X coverage) followed by
93 imputation is a potential alternative approach to genotyping arrays for assessing the common
94 genetic variants needed for GPS calculations. Several recent studies have demonstrated the
95 efficiency and accuracy of lcWGS for other applications of statistical genetics, including local
96 ancestry deconvolution, complex trait association studies, and detection of rare genetic variants
97 [4–7].
98
99 We developed a pipeline for common genetic variant imputation using lcWGS data on samples
100 from the 1000 Genomes Project (1KGP) and Genome in a Bottle (GIAB) Consortium and herein
101 demonstrate imputation accuracy for lcWGS similar to genotyping arrays. Using three recently
102 published GPSs for CAD [3], BC [8], and AF [3], we show high technical concordance in GPSs
103 calculated from lcWGS and genotyping arrays. Finally, using our pipeline in a large European
104 population seeking genetic testing, we observe similar GPS risk stratification performance as
105 previously published array-based results [3,8].
106
107 **METHODS**
108 **Study design**
109 The study design is summarized in Figure 1 and described in detail below. The pipeline
110 validation data set (n = 10) was used to assess imputation accuracy for common genetic
111 variants (Figure 1A). The technical concordance cohort (n = 184) was used to assess the
112 correlation between three previously published GPSs for CAD [3], BC [8], and AF [3] from
113 lcWGS and genotyping arrays (Figure 1B). The diverse ancestry data set (n = 120) was used to

114 assess imputation accuracy for common genetic variants and performance of GPS_{CAD}, GPS_{BC},
115 and GPS_{AF} (Figure 1B). The clinical cohort (n = 11,502) was used to assess performance of
116 GPS_{CAD}, GPS_{BC}, and GPS_{AF} in a large European population seeking genetic testing (Figure 1B).
117
118 **Data set and cohort selection**
119 The pipeline validation data set included seven globally representative samples from 1KGP
120 populations (HG02155, NA12878, HG00663, HG01485, NA21144, NA20510, and NA19420;
121 see Supplementary Table 1, Additional File 1) and a trio of Ashkenazi samples (NA24385,
122 NA24143, and NA24149) from the GIAB Consortium (Figure 1A).
123
124 The technical concordance cohort included DNA samples from 184 individuals whose
125 healthcare provider had ordered a Color multi-gene panel test (Figure 1B). All individuals 1) had
126 85% or greater European genetic ancestry calculated using fastNGSadmix [9] using 1KPG as
127 the reference panel, 2) self-identified as 'Caucasian', and 3) did not have pathogenic or likely
128 pathogenic variants in the multi-gene NGS panel test, as previously described [10] (see
129 Supplementary Methods, Additional File 2). Demographics are provided in Supplementary Table
130 2, Additional File 1. All phenotypic information was self-reported by the individual through an
131 online, interactive health history tool. Of the 184 individuals, 61 individuals reported having a
132 personal history of CAD (defined here as a myocardial infarction or coronary artery bypass
133 surgery), 62 individuals reported no personal history of CAD, and 61 individuals reported no
134 personal history of CAD but were suspected to have a high GPS_{CAD} based on preliminary
135 analysis. This preliminary analysis included imputation from multi-gene panel and off-target
136 sequencing data, which has been shown to have similar association statistics and effect sizes
137 compared to genotyping arrays [4]. These individuals were included in the technical
138 concordance cohort to artificially create a relatively uniform distribution of GPS_{CAD} in the data
139 set. Correlation coefficients between GPS_{CAD} from lcWGS and genotyping array were calculated

140 after removing the 61 individuals who were suspected to have a high GPS_{CAD} based on multi-
141 gene panel and off-target sequencing data to avoid artificial inflation of the correlation
142 coefficient. Two individuals who reported no personal history of CAD but were suspected to
143 have a high GPS_{CAD} failed genotyping (quality control call rate of < 97%) and lcWGS (overall
144 coverage of < 0.5X), leaving a total of 182 individuals for analyses.
145
146 The diverse ancestry data set included a total of 120 samples from the following populations
147 from 1KGP: Han Chinese in Beijing, China (CHB); Yoruba in Ibadan, Nigeria (YRI); Gujarati
148 Indian from Houston, Texas (GIH); Americans of African Ancestry in Southwest USA (ASW);
149 Mexican Ancestry from Los Angeles, USA (MXL); and Puerto Ricans from Puerto Rico (PUR)
150 (see Supplementary Table 3, Additional File 1; Figure 1B). Four samples, including NA18917
151 and NA19147 from the YRI population and NA19729 and NA19785 from the MXL population,
152 were below the target 0.5X coverage and removed from analyses.
153
154 The clinical cohort included DNA samples from 11,502 individuals whose healthcare provider
155 had ordered a Color multi-gene panel test (Figure 1B). All individuals 1) had 90% or greater
156 European genetic ancestry calculated using fastNGSadmix [9] using 1KPG as the reference
157 panel, 2) self-identified as 'Caucasian', 3) provided history of whether they had a clinical
158 diagnosis of CAD, BC, or AF, and 4) did not have pathogenic or likely pathogenic variants
159 detected in the multi-gene NGS panel test, as previously described [10] (see Supplementary
160 Methods, Additional File 2). Demographics are provided in Supplementary Table 2, Additional
161 File 1. All phenotypic information was self-reported by the individual through an online,
162 interactive health history tool.
163
164 **Whole genome sequencing**

165 DNA was extracted from blood or saliva samples and purified using the Perkin Elmer Chemagic
166 DNA Extraction Kit (Perkin Elmer, Waltham, MA) automated on the Hamilton STAR (Hamilton,
167 Reno, NV) and the Chemagic Liquid Handler (Perkin Elmer, Waltham, MA). The quality and
168 quantity of the extracted DNA were assessed by UV spectroscopy (BioTek, Winooski, VT). High
169 molecular weight genomic DNA was enzymatically fragmented and prepared using the Kapa
170 HyperPlus Library Preparation Kit (Roche Sequencing, Pleasanton, CA) automated on the
171 Hamilton Star liquid handler and uniquely tagged with 10 bp dual-unique barcodes (IDT,
172 Coralville, IA). Libraries were pooled together and loaded onto the NovaSeq 6000 (Illumina, San
173 Diego, CA) for 2 x 150 bp sequencing.
174
175 For the pipeline validation data set, all samples underwent WGS with mean coverage of 13.22X
176 (range 7.82X to 17.30X); downsampling was then performed using SAMtools to simulate
177 lcWGS. For the technical concordance cohort, all samples underwent lcWGS with mean
178 coverage of 1.24X (range 0.54X to 1.76X). Imputed genotypes were compared with published,
179 high-confidence known genotypes from 1KGP and the GIAB Consortium. For the diverse
180 ancestry data set, all samples underwent lcWGS with mean coverage of 0.89X (range 0.68X to
181 1.24X). For the clinical cohort, all samples underwent lcWGS with mean coverage of 0.95X
182 (range 0.51X to 2.57X).
183
184 **Downsampling**
185 For the pipeline validation data set, aligned reads were downsampled using SAMtools [11] to
186 2.0X, 1.0X, 0.75X, 0.5X, 0.4X, 0.25X, and 0.1X coverage. For the technical concordance cohort,
187 aligned reads were downsampled to 1.0X, 0.75X, 0.5X, 0.4X, 0.25X, and 0.1X coverage. In a
188 few cases in the technical concordance cohort, the primary samples had fewer reads than the
189 target downsample. In those situations, all of the reads were retained. For example, if the
190 primary sample only had 0.8X coverage, when downsampled to 1.0X, all reads were retained.

Downsampling was repeated using two independent seeds in SAMtools. Once the downsampled data was generated, the imputation was repeated to generate imputed genotypes using only the downsampled reads.

Imputation site selection

All data sets and cohorts were imputed to a set of autosomal SNP and insertion-deletion (indel) sites from 1KGP with greater than 1% allele frequency in any of the five 1KGP super populations (African, American, East Asian, European, and South Asian), for a total of 21,770,397 sites. This is hereafter referred to as the 'imputation SNP loci.' Multi-allelic SNPs and indels were represented as two biallelic markers for imputation.

Genotype likelihood calculations and imputation

Genotype likelihood calculations and imputation were performed independently for each sample. Sequence reads were aligned with the human genome reference GRCh37.p12 using the Burrows-Wheeler Aligner (BWA) [12], and duplicate and low quality reads were removed. Genotype likelihoods were then calculated at each of the biallelic SNP loci in the imputation SNP loci that were covered by one or more sequencing reads called using the mpileup command implemented in bcftools version 1.8 [13]. Indels or multi-allelic sites were not included in this first genotype likelihood calculation. Reads with a minimum mapping alignment quality of 10 or greater and bases with a minimum base quality of 10 or greater were included. Genotype likelihoods at each observed site were then calculated using the bcftools call command with allele information corresponding to the imputation SNP loci. This procedure discarded calls with indels or calls where the observed base did not match either the reference or expected alternate allele for the SNP locus.

216 Imputation was performed using the genotype likelihood imputation option implemented in
217 BEAGLE 4.1 [14]. This imputation used default parameters except with a model scale
218 parameter of 2 and the number of phasing iterations to 0. A custom reference panel was
219 constructed for each sample being imputed by selecting the 250 most similar samples to that
220 sample from 1KGP Phase 3 release using Identity-by-State (IBS) comparison. A reference
221 panel size of 250 was selected to best balance imputation run time and accuracy (see
222 Supplementary Figure 1, Additional File 2). To ensure that IBS values were comparable across
223 samples, a set of regions consistently sequenced at high depth (> 20X) across all samples was
224 utilized. When imputation was performed on samples included in 1KGP Phase 3 release, that
225 sample and any related samples were excluded from the custom reference panel.
226
227 To generate genotypes at all of the remaining untyped sites, a second round of imputation was
228 performed using BEAGLE 5.0 [15]. This imputation used default settings and included the full
229 1KGP as the imputation reference panel. To note, when performing analysis using 1KGP
230 samples, any related individuals were removed. Each sample then had imputed genotype calls
231 at each of the imputation SNP loci. Indels and multiallelic sites were included in this second
232 genotype likelihood calculation.
233
234 **Genotyping array**
235 DNA was extracted from blood or saliva samples and purified using the Perkin Elmer Chemagic
236 DNA Extraction Kit (Perkin Elmer, Waltham, MA) automated on the Hamilton STAR (Hamilton,
237 Reno, NV) and the Chemagic Liquid Handler (Perkin Elmer, Waltham, MA). The quality and
238 quantity of the extracted DNA were assessed by UV spectroscopy (BioTek, Winooski, VT).
239
240 DNA was genotyped on the Axiom UK Biobank Array by Affymetrix (Santa Clara, CA).
241 Genotypes were filtered according to the manufacturer's recommendations, removing loci with

242 greater than 5% global missingness and those that significantly deviated from Hardy-Weinberg
243 equilibrium. In addition, all A/T and G/C SNPs were removed due to potential strand
244 inconsistencies. Each of the remaining SNPs were aligned with the hg19 reference sequence to
245 correctly code the reference alleles as allele 1, matching the sequencing data.
246
247 To generate genotypes at all of the remaining untyped sites, imputation was performed using
248 BEAGLE 5.0 [15]. This imputation used default settings and included the full 1KGP as the
249 imputation reference panel. To note, when performing analysis using 1KGP samples, any
250 related individuals were removed. Each sample then had imputed genotype calls at each of the
251 imputation SNP loci.
252
253 **Imputation accuracy and quality assessment**
254 Imputation accuracy for 1KGP and GIAB samples was calculated by comparing imputation
255 results with previously released genotypes, excluding regions marked as low confidence by
256 GIAB.
257
258 Imputation accuracy on the genotyped samples was assessed on 470,363 sites that were
259 included on the genotyping array at different allele frequency buckets: 257,362 sites with greater
260 than 5% allele frequency, 119,978 sites between 1-5% allele frequency, and 93,022 sites with
261 less than 1% allele frequency. Imputation quality was assessed through site-specific dosage r^2
262 comparing with genotype values from the genotyping array.
263
264 **GPS selection**
265 The GPSs for CAD [3], BC [8], and AF [3] were previously published and selected based on
266 their demonstrated ability to accurately predict and stratify disease risk as well as identify
267 individuals at risk comparable to monogenic disease. GPS_{CAD} contained 6,630,150

268 polymorphisms, GPS_{BC} contained 3,820 polymorphisms, and GPS_{AF} contained 6,730,541
269 polymorphisms. All loci included in these scores were included in the imputation SNP loci.
270
271 **GPS normalization**
272 In the clinical cohort, raw GPSs were normalized by taking the standardized residual of the
273 predicted score after correction for the first 10 principal components (PC) of ancestry [16]. PCs
274 were calculated by projecting lcWGS samples into 10 dimensional PC analysis (PCA) space
275 using the LASER program [17]. A combination of samples from 1KGP and the Human Origins
276 [18] project were used as a reference for the projection.
277
278 **RESULTS**
279 **Development and validation of imputation pipeline for lcWGS**
280 Previous studies have evaluated the potential use of lcWGS in local ancestry deconvolution,
281 complex trait association studies, and detection of rare genetic variants [4–6]. To assess the
282 feasibility and accuracy of this approach for GPSs, we first developed an imputation pipeline
283 that reads raw fastq sequence data and generates a vcf with imputed site information at 21.7
284 million sites (imputation SNP loci) (Figure 1A, B). Briefly, reads are aligned to the reference
285 genome and filtered for duplicates and low quality. Using this BAM file, we then calculate
286 genotype likelihoods and impute expected genotypes using 1KGP as the imputation reference
287 panel.
288
289 To validate this imputation pipeline, we performed hcWGS and downsampling on seven
290 samples from different 1KGP populations and a trio of Ashkenazi Jewish GIAB samples
291 (pipeline validation data set) to varying depths of coverage from 2.0X to 0.1X (See
292 Supplementary Table 1, Additional File 2). We used the published genotype calls for each of
293 these samples as truth data and found that imputation accuracy was above 0.90 r^2 for all

294 samples at 0.5X and higher (Figure 2). As expected, this was correlated with sequencing depth,
295 with diminishing gains observed at coverages above 1.0X. While imputation accuracy was
296 similar across diverse populations, it was slightly reduced in the Colombian sample (HG01485),
297 likely due to complex local ancestry related to admixture, and in the Yoruban sample
298 (NA19240), likely due to the shorter blocks of linkage disequilibrium and higher genetic diversity
299 in Africa [19]. Taken together, these data suggest that at sequencing depth at or above 0.5X,
300 our pipeline has similar imputation accuracy to genotyping array-based imputation across
301 individuals from multiple populations. As such, we set 0.5X as a quality control for success and
302 removed samples with coverage below this threshold in subsequent analyses.
303
304 **Technical concordance between GPSs calculated from lcWGS and genotyping array**
305 To assess the technical concordance of using lcWGS to calculate GPSs, we performed low
306 coverage sequencing and used genotyping arrays on DNA from 184 individuals (technical
307 concordance cohort) (Figure 1B). This concordance assessment was restricted to individuals of
308 European ancestry to most closely align with the populations used for GPS training and
309 validation.
310
311 We first compared the lcWGS genotype dosages with a subset of variants directly genotyped (n
312 = 470,362) on the genotyping array to assess imputation performance. Assuming the typed loci
313 called on the genotyping array as 'true', we observed an average imputation $r^2 > 0.90$ at 0.5X
314 depth for variants with global minor allele frequency (MAF) greater than 5% (see Supplementary
315 Figure 2, Additional File 3). As expected, imputation accuracy was highest for variants with
316 higher MAF. For lower frequency variants, we saw a reduction in imputation accuracy, as
317 expected, with $r^2 > 0.85$ for variants at 1% to 5% MAF and $r^2 > 0.80$ for variants less than 1%
318 global MAF. Taken together, this demonstrates that lcWGS has high accuracy in this test
319 setting.

320

321 We then calculated previously published GPSs for CAD [3], BC [8], and AF [3] on each sample
322 using genotyping array data or lcWGS data. We found that GPS_{CAD}, GPS_{BC}, and GPS_{AF} were
323 highly correlated (Figure 3A-C), with the score mean (Student t-test p = 0.17) and variance (F
324 test p = 0.91) equivalent between lcWGS and the genotyping array. The correlations of GPS_{CAD}
325 and GPS_{AF} (r^2 = 0.98 and r^2 = 0.97, respectively) were slightly higher than that of GPS_{BC} (r^2 =
326 0.93), which could be due to 1) the smaller number of loci in GPS_{BC} (6.6 million compared to
327 3820 SNPs), 2) differences in allele frequencies between SNPs with high weights, and/or 3) the
328 fact that GPS_{BC} was trained and validated on a different genotyping array, the OncoArray, than
329 the Axiom UK Biobank Array used in this study [8].

330

331 The technical concordance cohort ranged in coverage from 0.54X to 1.76X with a mean
332 coverage of 1.24X, and we have shown that depth can impact imputation performance -- depth
333 increases above 0.5X have a smaller but measurable effect on imputation performance (Figure
334 2; see Supplementary Figure 2, Additional File 3). To determine the low coverage sequencing
335 depth required for GPS accuracy, we used SAMtools to downsample the lcWGS data in this
336 cohort to 1.0X, 0.75X, 0.5X, 0.4X, 0.25X, and 0.1X. We found that GPS_{CAD}, GPS_{BC}, and GPS_{AF}
337 are robust to lcWGS sequencing depth 0.5X and that coverages do not systematically bias GPS
338 calculations in a specific direction (see Supplementary Figure 3 and Supplementary Figure 4,
339 Additional File 3), indicating that samples above 0.5X with small changes in coverage variation
340 can be combined for downstream analysis. In addition, the correlation increases logarithmically
341 as coverage increases (see Supplementary Figure 5, Additional File 3). These data
342 demonstrate high correlation between GPSs from lcWGS data and genotyping array in a
343 randomly selected sample. Interestingly, correlation at 0.1X was still high enough that GPSs at
344 this coverage may have research utility, suggesting that significant amounts of data regarding
345 common genetic variation could be recovered from off-target reads in exome and multi-gene

346 panel sequencing studies to allow for GPS calculation. Taken together, these data demonstrate
347 that lcWGS provides equivalent accuracy for calculation of GPSs, with sequencing coverage as
348 low as 0.5X.
349

350 **Assessment of imputation performance and technical concordance across diverse**
351 **populations**

352 To further assess the performance of our imputation pipeline across diverse populations, we
353 performed lcWGS on 120 additional samples from six 1KGP populations (CHB, GIH, YRI, ASW,
354 MXL, and PUR; see Supplementary Table 3, Additional File 1) that represent the range of
355 ancestry observed in admixed populations (diverse ancestry data set). We compared genotypes
356 imputed using our lcWGS pipeline to known 1KGP WGS data and found that imputation
357 accuracy was above 0.90 r^2 for all samples (range 0.94 - 0.97) (Figure 4A). In addition, we
358 found that GPS calculated from lcWGS data and GPS calculated from the Phase 3 1KGP WGS
359 data release have a high correlation, with an r^2 value of 0.98, 0.91, and 0.98 for CAD, BC, and
360 AF, respectively (Figure 4B-D). These results suggest that lcWGS can enable accurate
361 imputation and calculation of GPSs in diverse populations.
362

363 **Association of lcWGS-calculated GPSs with disease phenotypes in a clinical cohort**

364 Previous studies have demonstrated the association of GPSs with prevalent disease using
365 genotyping arrays [3,8,20–22] and hcWGS [16]. To observe the performance of lcWGS-
366 calculated GPSs in a large population, we performed low coverage sequencing on 11,502
367 European individuals (clinical cohort) (See Supplementary Table 2, Additional File 1) and
368 calculated GPS_{CAD}, GPS_{BC}, and GPS_{AF} for each individual. Raw GPSs were normalized by
369 taking the standardized residual of the predicted score after correction for the first 10 PCAs (see
370 Supplementary Figure 6, Additional File 3) [16,23]. First, we note that there are no major outliers
371 (defined as a z-score greater than 5) in GPS_{CAD}, GPS_{BC}, and GPS_{AF} and that the normalized

372 scores formed an approximately normal distribution for each (see Supplementary Figure 7,
373 Additional File 3). Each of the GPSs were strongly associated with self-reported history of
374 disease, with effect estimates comparable to prior reports using genotyping arrays to calculate
375 GPS -- GPS_{CAD} (OR per standard deviation = 1.59 (1.32 - 1.92), n = 11,010), GPS_{BC} (OR per
376 standard deviation = 1.56 (1.45 - 1.68); n = 8722), and GPS_{AF} (OR per standard deviation =
377 1.28 (1.12 - 1.46); n = 10,303) (Figure 5).
378
379 Previous studies have noted significantly increased disease prevalence among individuals in the
380 extreme tails of the GPS distribution when compared to the remainder of the population [3,8].
381 We replicated this observation by assessing the prevalence of disease in the highest 5% of the
382 GPS distribution for each of the three diseases, noting odds ratios of 4.5 (2.62 - 7.74), 2.62
383 (2.04 - 3.36), and 1.96 (1.24 - 3.11) for GPS_{CAD}, GPS_{BC}, and $GPS_{AF,}$ respectively.
384
385 Area under the curve (AUC) is an additional metric used to assess the ability of a given risk
386 factor to discriminate between affected cases and disease-free controls. When only the GPS
387 was included in the prediction model, GPS_{CAD} had an AUC of 0.60, GPS_{BC} had an AUC of 0.63,
388 and GPS_{AF} had an AUC of 0.57. The additional inclusion of age and sex increased the AUCs to
389 0.86 for GPS_{CAD}, 0.78 for GPS_{BC}, and 0.78 for GPS_{AF}. For each of these three common,
390 complex diseases, the magnitude of associations with clinical disease and AUC metrics were
391 consistent with previous publications [3,8]. Taken together, these results suggest that lcWGS-
392 calculated GPSs can accurately stratify risk with comparable accuracy to previously published
393 GPS-disease associations calculated on the basis of genotyping array data.
394
395 **DISCUSSION**
396 For the past two decades, genotyping array-based GWAS and imputation have been the driving
397 force in our discovery of genetic loci predictive of disease and derivation and calculation of

398 GPSs. In this study, we developed and validated an imputation pipeline to calculate GPSs from
399 variably downsampled hcWGS and lcWGS data sets. While the efficiency of lcWGS has been
400 reported for other applications of statistical genetics [4–6], we demonstrate that lcWGS achieves
401 similar technical concordance as the Axiom UK Biobank Array by Affymetrix for determining
402 GPSs. Furthermore, the imputation r^2 from lcWGS was greater than 90%, which is similar to the
403 imputation accuracy reported from other commercially-available genotyping arrays [24]. Taken
404 together, these data suggest that lcWGS has comparable accuracy to genotyping arrays for
405 assessment of common variants and subsequent calculation of GPSs.
406
407 Our finding that lcWGS can be used for accurate genotyping and imputation of common genetic
408 variants has implications for the future of genomic research and medicine. Currently, disease
409 GWAS are performed using a variety of genotyping arrays that are designed to target specific
410 sets of genes or features, reducing imputation quality in regions that are not targeted [25].
411 lcWGS enables less biased imputation than genotyping arrays by not pre-specifying the genetic
412 content that is included for assessment, as is necessary for genotyping arrays. Because initial
413 GWAS focused on populations with high homogeneity to reduce noise and increase fit of risk
414 stratification, many genotyping arrays were designed to capture common genetic variants based
415 on the linkage disequilibrium structure in European populations [26]. However, this
416 ascertainment bias reduces the imputation performance from genotyping array data in diverse
417 populations [27–29]. Imputation from lcWGS data reduces this bias by including all SNPs
418 observed in 1KGP populations as potential predictors. The effects of SNP selection bias are
419 also not equivalent across genotyping arrays, and therefore variants included in a GPS trained
420 and validated on one genotyping array may not be as predictive on another genotyping array
421 [30]. lcWGS systematically surveys variants independent of SNP selection bias and thus
422 provides one approach to overcome this issue. Our findings here demonstrate that GPSs
423 trained and validated on different genotyping arrays are transferable to lcWGS-calculated GPS.

Furthermore, as new genetic associations are discovered, lcWGS can be re-analyzed with ever more inclusive sets of known SNPs, further reducing SNP selection bias and advancing the study and understanding of the genetic contributions to disease. In contrast, genotyping arrays are static and cannot be easily updated or changed without designing a *de novo* platform.

lcWGS also has the potential to easily integrate into current clinical sequencing pipelines. In contrast to genotyping arrays, which require investment in separate laboratory technology, lcWGS can be performed on the same platform as current hcWGS or targeted multi-gene panel clinical testing. The ease of combining these two pathways could help to drive GPS adoption into clinical practice and can likely be achieved at a cost comparable to genotyping arrays [4]. As the cost of next generation sequencing continues to decrease, the cost of lcWGS will also continue to decrease.

This study should be interpreted in the context of potential limitations. First, the imputation accuracy observed in our analysis may have been limited by the reference panel size. Future efforts using an even larger reference panel may lead to further improved imputation accuracy, particularly for variants with allele frequency less than 1% [24]. Second, while lcWGS may ultimately enable derivation of GPSs with improved predictive accuracy or ethnic transferability, this was not explicitly explored here. Rather, we demonstrate the feasibility and accuracy of using lcWGS of calculating GPSs published in previous studies. Third, disease phenotypes in our clinical cohort were based on individual self-report rather than review of health records. However, several studies have shown that self-reported personal history data have high concordance with data reported by a healthcare provider or electronic health records [31–34], and any inaccuracies would be expected to bias GPS-disease associations to the null.

CONCLUSIONS

450 In conclusion, this work establishes lcWGS as an alternative approach to genotyping arrays for
451 common genetic variant assessment and GPS calculation -- providing comparable accuracy at
452 similar cost while also overcoming the ascertainment bias inherent to variant selection in
453 genotyping array design.
454
455 **LIST OF ABBREVIATIONS**
456 GPS, genome-wide polygenic score
457 lcWGS, low coverage whole genome sequencing
458 CAD, coronary artery disease
459 BC, breast cancer
460 AF, atrial fibrillation
461 1KGP, 1000 Genomes Project
462 GIAB, Genome in a Bottle
463 Indel, insertion-deletion
464 BWA, Burrows-Wheeler Aligner
465 IBS, Identity-by-State
466 PC, principal components
467 PCA, PC analysis
468 MAF, minor allele frequency
469 AUC, area under the curve
470
471 **DECLARATIONS**
472 **Ethics approval and consent to participate**
473 All individuals in the technical concordance cohort and clinical cohort gave electronic informed
474 consent to have their de-identified information and sample used in anonymized studies
475 (Western Institutional Review Board, #20150716).

476

477 **Consent for publication**

478 All individuals in the technical concordance cohort and clinical cohort gave electronic informed
479 consent that Color may author publications using non-aggregated, de-identified information,
480 either on its own or in collaboration with academic or commercial third parties.

481

482 **Availability of data and material**

483 The technical concordance and clinical cohort data are not publicly available given the potential
484 to compromise research participant privacy or consent.

485

486 1KGP, http://www.internationalgenome.org/
487 GIAB, ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/
488 Samtools/Bcftools, http://www.htslib.org/
489 BEAGLE, https://faculty.washington.edu/browning/beagle/beagle.html
490 FastNGSAdmix, http://www.popgen.dk/software/index.php/FastNGSadmix

491

492 **Competing interests**

493 JRH, CLN, and AYZ are currently employed by and have equity interest in Color Genomics.
494 JRH has previously consulted for Twist Bioscience and Etalon Diagnostics. GM was previously
495 employed at Color Genomics and Operator. JRH and GM report a patent application related to
496 low coverage whole genome sequencing. SK is an employee of Verve Therapeutics and holds
497 equity in Verve Therapeutics, Maze Therapeutics, Catabasis, and San Therapeutics. He is a
498 member of the scientific advisory boards for Regeneron Genetics Center and Corvidia
499 Therapeutics; he has served as a consultant for Acceleron, Eli Lilly, Novartis, Merck, Novo
500 Nordisk, Novo Ventures, Ionis, Alnylam, Aegerion, Haug Partners, Noble Insights, Leerink
501 Partners, Bayer Healthcare, Illumina, Color Genomics, MedGenome, Quest, and Medscape; he

502 reports patents related to a method of identifying and treating a person having a predisposition
503 to or afflicted with cardiometabolic disease (20180010185) and a genetic risk predictor
504 (20190017119). AVK has served as a consultant for Color Genomics and reports a patent
505 related to a genetic risk predictor (20190017119).
506
507 **Funding**
508 This work was supported by Color Genomics.
509
510 **Authors' contributions**
511 JRH, GM, AYZ, and AVK designed the overall study. JRH, CLN, GM, AYZ, SK, and AVK
512 contributed to data acquisition and analysis. JRH, CLN, GM, AYZ, SK, and AVK drafted or
513 critically revised the manuscript for important intellectual content. AYZ and AVK are the
514 guarantors of this work and, as such, have full access to all of the data in the study and take
515 responsibility for the integrity of the data and the accuracy of the data analysis.
516
517 **Acknowledgements**
518 We would like to thank Will Stedden, Carmen Lai, and Anjali D. Zimmer for helpful discussions
519 and Justin Lock, Alok Sabnis, and Valerie Ngo for laboratory support and sample processing.

520 **REFERENCES**

521 1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional
522 mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis
523 for the Global Burden of Disease Study 2010. Lancet. 2012;380:2095–128.

524 2. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to
525 Omnigenic. Cell. 2017;169:1177–86.

526 3. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide
527 polygenic scores for common diseases identify individuals with risk equivalent to monogenic
528 mutations. Nat Genet [Internet]. 2018; Available from: http://dx.doi.org/10.1038/s41588-018-
529 0183-z

530 4. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-
531 coverage sequencing and imputation increases power for genome-wide association studies. Nat
532 Genet. 2012;44:631–5.

533 5. Gilly A, Southam L, Suveges D, Kuchenbaecker K, Moore R, Melloni GEM, et al. Very low
534 depth whole genome sequencing in complex trait association studies. Bioinformatics [Internet].
535 2018; Available from: http://dx.doi.org/10.1093/bioinformatics/bty1032

536 6. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, et al. Genomic Analyses from Non-
537 invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and
538 Chinese Population History. Cell. 2018;175:347–59.e14.

539 7. Navon O, Sul JH, Han B, Conde L, Bracci PM, Riby J, et al. Rare variant association testing
540 under low-coverage sequencing. Genetics. 2013;194:769–79.

541 8. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores

542 for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019;104:21–
543 34.

544 9. Jørsboe E, Hanghøj K, Albrechtsen A. fastNGSadmix: admixture proportions and principal
545 component analysis of a single NGS sample. Bioinformatics. 2017;33:3148–50.

546 10. Neben CL, Zimmer AD, Stedden W, van den Akker J, O'Connor R, Chan RC, et al. Multi-
547 Gene Panel Testing of 23,179 Individuals for Hereditary Cancer Risk Identifies Pathogenic
548 Variant Carriers Missed by Current Genetic Testing Guidelines. J Mol Diagn [Internet]. Elsevier;
549 2019 [cited 2019 Jun 11];0. Available from: https://jmd.amjpathol.org/article/S1525-
550 1578(18)30334-9/fulltext

551 11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
552 Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

553 12. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM
554 [Internet]. arXiv [q-bio.GN]. 2013. Available from: http://arxiv.org/abs/1303.3997

555 13. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and
556 population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–
557 93.

558 14. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J
559 Hum Genet. 2016;98:116–26.

560 15. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation
561 Reference Panels. Am J Hum Genet. 2018;103:338–48.

562 16. Khera AV, Chaffin M, Zekavat SM, Collins RL, Roselli C, Natarajan P, et al. Whole Genome
563 Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized

564 with Early-Onset Myocardial Infarction. Circulation [Internet]. American Heart Association
565 Bethesda, MD; 2018 [cited 2018 Nov 27]; Available from:
566 https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.118.035658

567 17. Wang C, Zhan X, Liang L, Abecasis GR, Lin X. Improved ancestry estimation for both
568 genotyping and sequencing data using projection procrustes analysis and genotype imputation.
569 Am J Hum Genet. 2015;96:926–37.

570 18. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights
571 into the origin of farming in the ancient Near East. Nature. 2016;536:419–24.

572 19. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang
573 HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

574 20. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic
575 Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary
576 Prevention. J Am Coll Cardiol. 2018;72:1883–93.

577 21. Richardson TG, Harrison S, Hemani G, Smith GD. An atlas of polygenic risk score
578 associations to highlight putative causal relationships across the human phenome [Internet].
579 bioRxiv. 2018 [cited 2018 Nov 27]. p. 467910. Available from:
580 https://www.biorxiv.org/content/early/2018/11/11/467910

581 22. Mavaddat N, Pharoah PDP, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of
582 breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst [Internet].
583 2015;107. Available from: http://dx.doi.org/10.1093/jnci/djv036

584 23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal
585 components analysis corrects for stratification in genome-wide association studies. Nat Genet.
586 2006;38:904–9.

587 24. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference
588 panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48:1279–83.

589 25. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a
590 custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric
591 traits. PLoS Genet. 2012;8:e1002793.

592 26. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is
593 important, and how to correct it. Bioessays. 2013;35:780–6.

594 27. Wojcik GL, Fuchsberger C, Taliun D, Welch R, Martin AR, Shringarpure S, et al. Imputation-
595 Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies.
596 G3 . 2018;8:3255–67.

597 28. Nelson SC, Doheny KF, Pugh EW, Romm JM, Ling H, Laurie CA, et al. Imputation-Based
598 Genomic Coverage Assessments of Current Human Genotyping Arrays [Internet]. G3:
599 Genes|Genomes|Genetics. 2013. p. 1795–807. Available from:
600 http://dx.doi.org/10.1534/g3.113.007161

601 29. Carlson CS, Matise TC, North KE, Haiman CA, Fesinmeyer MD, Buyske S, et al.
602 Generalization and dilution of association results from European GWAS in populations of non-
603 European ancestry: the PAGE study. PLoS Biol. 2013;11:e1001661.

604 30. Johnson EO, Hancock DB, Levy JL, Gaddis NC, Saccone NL, Bierut LJ, et al. Imputation
605 across genotyping arrays for genome-wide association studies: assessment of bias and a
606 correction strategy. Hum Genet. 2013;132:509–22.

607 31. Gentry-Maharaj A, Fourkala E-O, Burnell M, Ryan A, Apostolidou S, Habib M, et al.
608 Concordance of National Cancer Registration with self-reported breast, bowel and lung cancer
609 in England and Wales: a prospective cohort study within the UK Collaborative Trial of Ovarian

610 Cancer Screening. Br J Cancer. 2013;109:2875–9.

611 32. D'Aloisio AA, Nichols HB, Hodgson ME, Deming-Halverson SL, Sandler DP. Validity of self-
612 reported breast cancer characteristics in a nationwide cohort of women with a family history of
613 breast cancer. BMC Cancer. 2017;17:692.

614 33. Kehoe R, Wu SY, Leske MC, Chylack LT Jr. Comparing self-reported and physician-
615 reported medical history. Am J Epidemiol. 1994;139:813–8.

616 34. Malmo V, Langhammer A, Bønaa KH, Loennechen JP, Ellekjaer H. Validation of self-
617 reported and hospital-diagnosed atrial fibrillation: the HUNT study. Clin Epidemiol. 2016;8:185–
618 93.

619

620 **ADDITIONAL FILES**

621 Additional File 1

622 Homburger et al Additional File 1, PDF

623 Supplementary tables and legends

624

625 Additional File 2

626 Homburger et al Additional File 2, PDF

627 Supplementary Methods

628

629 Additional File 3

630 Homburger et al Additional File 3, PDF

631 Supplementary figures and legends

632

633 **FIGURE TITLES AND LEGENDS**

634 Figure 1. Study design and imputation pipelines. The study design has four groups: (A) pipeline
635 validation data set and (B) technical concordance cohort, diverse ancestry data set, and clinical
636 cohort. The imputation pipeline for each group is depicted. hcWGS, high coverage whole
637 genome sequencing. lcWGS, low coverage whole genome sequencing. HWE, Hardy–Weinberg
638 equilibrium. GPS, genome-wide polygenic score. CAD, coronary artery disease. BC, breast
639 cancer. AF, atrial fibrillation.

640

641 Figure 2. Assessment of imputation performance in the pipeline validation data set.
642 Downsampling from 30X to 0.1X showed that lcWGS accuracy was above 0.90 r^2 for all
643 samples at 0.5X (n = 4 independent random seeds for each sample and coverage value; error
644 bars are 95% confidence intervals). The thick brown dashed line is a smoothed trendline of the
645 average imputation quality while the thin grey dashed line demonstrates previously reported
646 imputation quality from a genotyping array (r^2 = 0.90) [4]. AJ, Ashkenazi Jewish. CDX, Chinese
647 Dai in Xishuangbanna, China. CEU, Utah Residents with Northern and Western European
648 Ancestry. CHB, Han Chinese in Beijing, China. CLM, Colombians from Medellin, Colombia.
649 GIH, Gujarati Indian from Houston, Texas. TSI, Toscani in Italia. YRI, Yoruba in Ibadan, Nigeria.

650

651 Figure 3. Correlation of GPSs between genotyping array and lcWGS in the technical
652 concordance cohort. (A) GPS_{CAD} calculated using lcWGS was highly correlated (r^2 = 0.98) with
653 those calculated using genotyping array (n = 182). (B) GPS_{BC} calculated using lcWGS was
654 highly correlated (r^2 = 0.93) with those calculated using genotyping array (n = 182). (C) GPS_{AF}
655 was highly correlated (r^2 = 0.97) with those calculated using genotyping arrays (n = 182). x-axis
656 is the raw GPS calculated from the genotyping array, and y-axis is the raw GPS calculated from
657 the lcWGS data; raw GPS values are unitless. lcWGS, low coverage whole genome

658 sequencing. GPS, genome-wide polygenic score. CAD, coronary artery disease. BC, breast
659 cancer. AF, atrial fibrillation.
660
661 Figure 4. Assessment of imputation performance and technical concordance across diverse
662 populations. (A) GPS_{CAD} calculated using lcWGS data was highly correlated with those
663 calculated using known 1KGP data (n = 116), with all samples having a correlation coefficient
664 above 0.90. The thin grey dashed line demonstrates previously reported imputation quality from
665 a genotyping array (r^2 = 0.90) [4]. (B) GPS_{CAD} calculated using lcWGS data was highly
666 correlated (r^2 = 0.98) with those calculated using known 1KGP data (n = 116). (C) GPS_{BC}
667 calculated using lcWGS data was highly correlated (r^2 = 0.91) with those calculated using
668 known 1KGP data (n = 116). (D) GPS_{AF} was highly correlated (r^2 = 0.98) with those calculated
669 using known 1KGP data (n = 116). 1KGP, 1000 Genomes Project. lcWGS, low coverage whole
670 genome sequencing. GPS, genome-wide polygenic score. CAD, coronary artery disease. BC,
671 breast cancer. AF, atrial fibrillation.
672
673 Figure 5. Association of lcWGS-calculated GPSs with disease phenotypes in the clinical cohort.
674 lcWGS-calculated GPS_{CAD} was associated with personal history of CAD (OR = 1.589 (1.32 -
675 1.92), n = 11,010, p = 1.32 x 10^{-6}). GPS_{CAD} was adjusted for age and sex. lcWGS-calculated
676 GPS_{BC} was associated with personal history of BC (OR = 1.56 (1.45 - 1.68); n = 8,722, p = 1.0 x
677 10^{-16}). GPS_{BC} was calculated only for females and adjusted for age at menarche. lcWGS-
678 calculated GPS_{AF} was associated with personal history of AF (OR = 1.277 (1.12 - 1.46); n =
679 10,303, p = 0.000292). GPS_{AF} was adjusted for age and sex. lcWGS, low coverage whole
680 genome sequencing. GPS, genome-wide polygenic score. CAD, coronary artery disease. BC,
681 breast cancer. AF, atrial fibrillation.

Figure 1

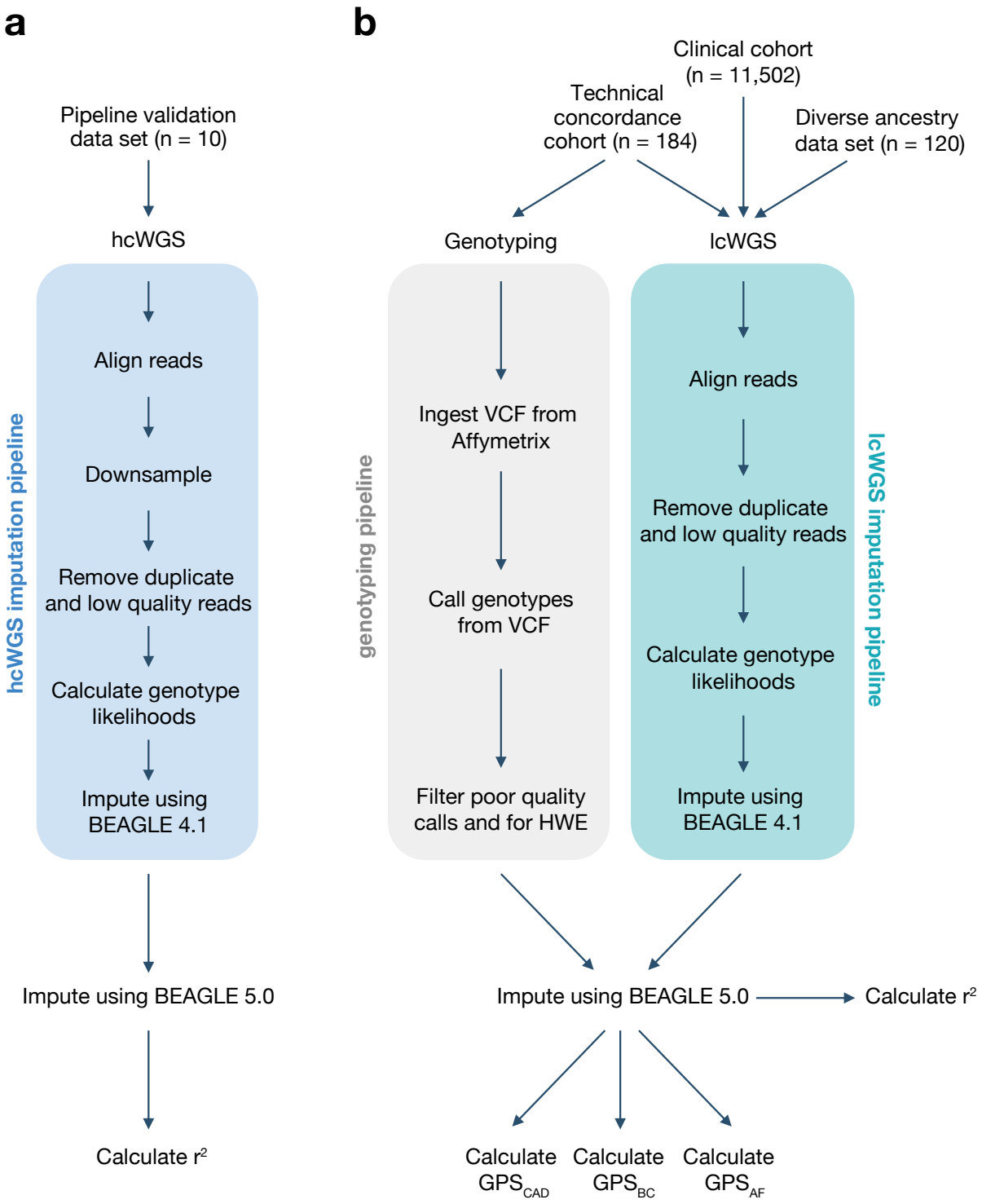

a
Pipeline validation data set (n = 10)
hcWGS
hcWGS imputation pipeline
Align reads
Downsample
Remove duplicate and low quality reads
Calculate genotype likelihoods
Impute using BEAGLE 4.1
Impute using BEAGLE 5.0
Calculate r2
b
Clinical cohort (n = 11,502)
Technical concordance cohort (n = 184)
Diverse ancestry data set (n = 120)
Genotyping
lcWGS
genotyping pipeline
Ingest VCF from Affymetrix
Call genotypes from VCF
Filter poor quality calls and for HWE
lcWGS imputation pipeline
Align reads
Remove duplicate and low quality reads
Calculate genotype likelihoods
Impute using BEAGLE 4.1
Impute using BEAGLE 5.0
Calculate r2
Calculate GPSCAD
Calculate GPSBC
Calculate GPSAF

Figure 2

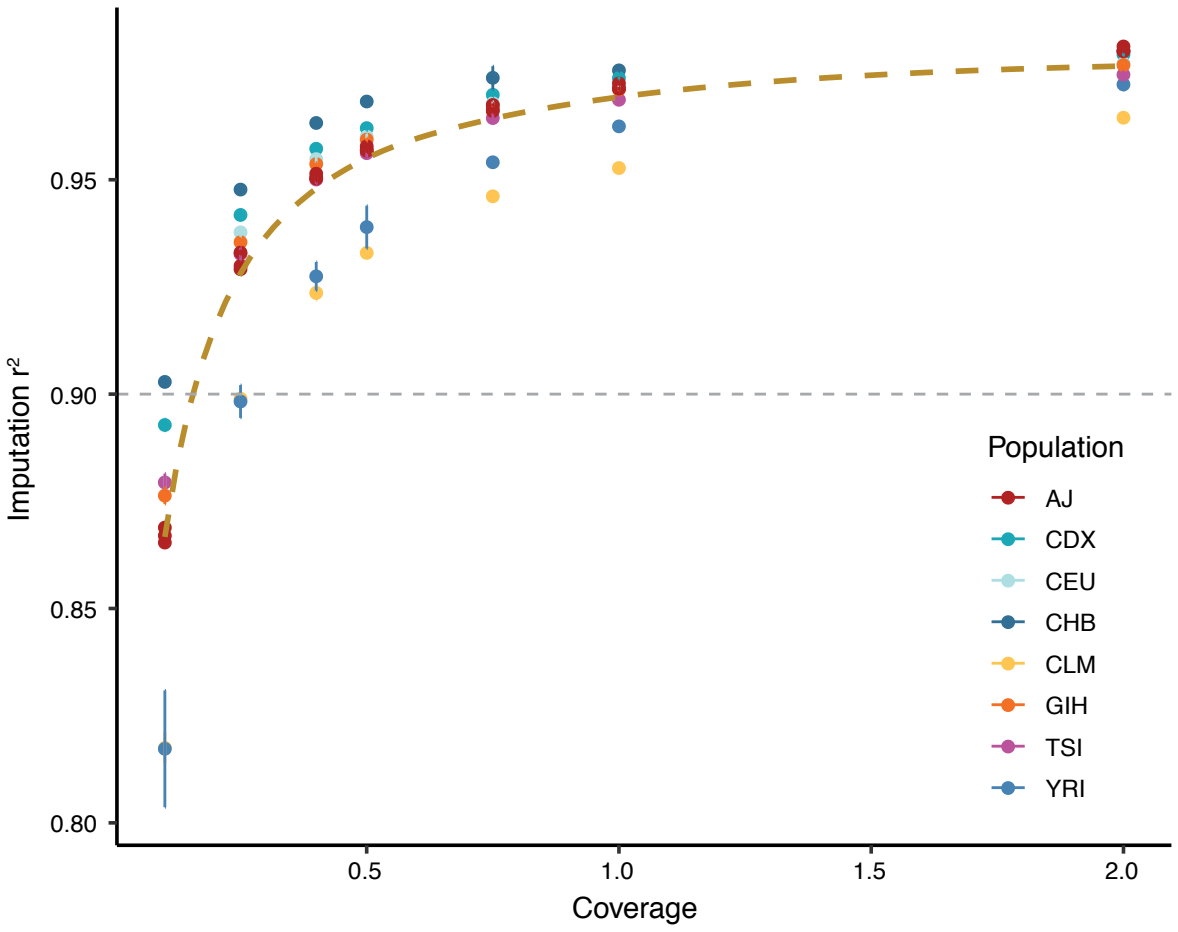

Imputation r²
0.95
0.90
0.85
0.80
0.5
1.0
1.5
2.0
Coverage
Population
AJ
CDX
CEU
CHB
CLM
GIH
TSI
YRI

Figure 3

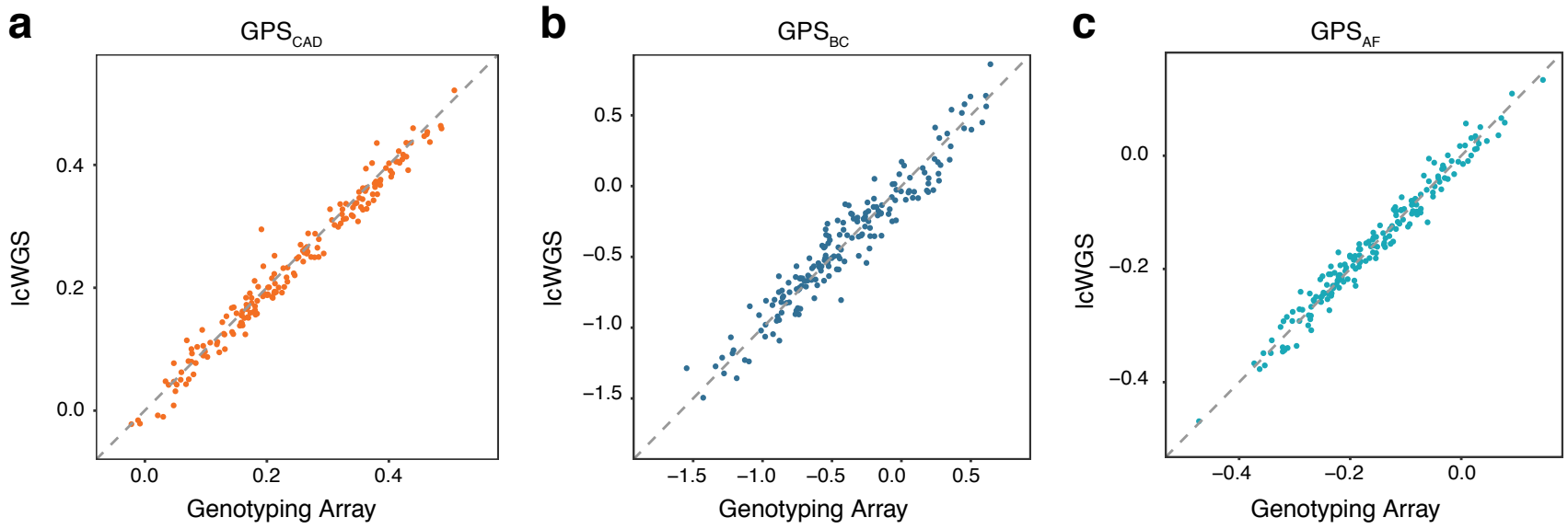

a
GPS$_{CAD}$
lcWGS
Genotyping Array
0.0
0.2
0.4
b
GPS$_{BC}$
lcWGS
Genotyping Array
−1.5
−1.0
−0.5
0.0
0.5
c
GPS$_{AF}$
lcWGS
Genotyping Array
−0.4
−0.2
0.0

Figure 4

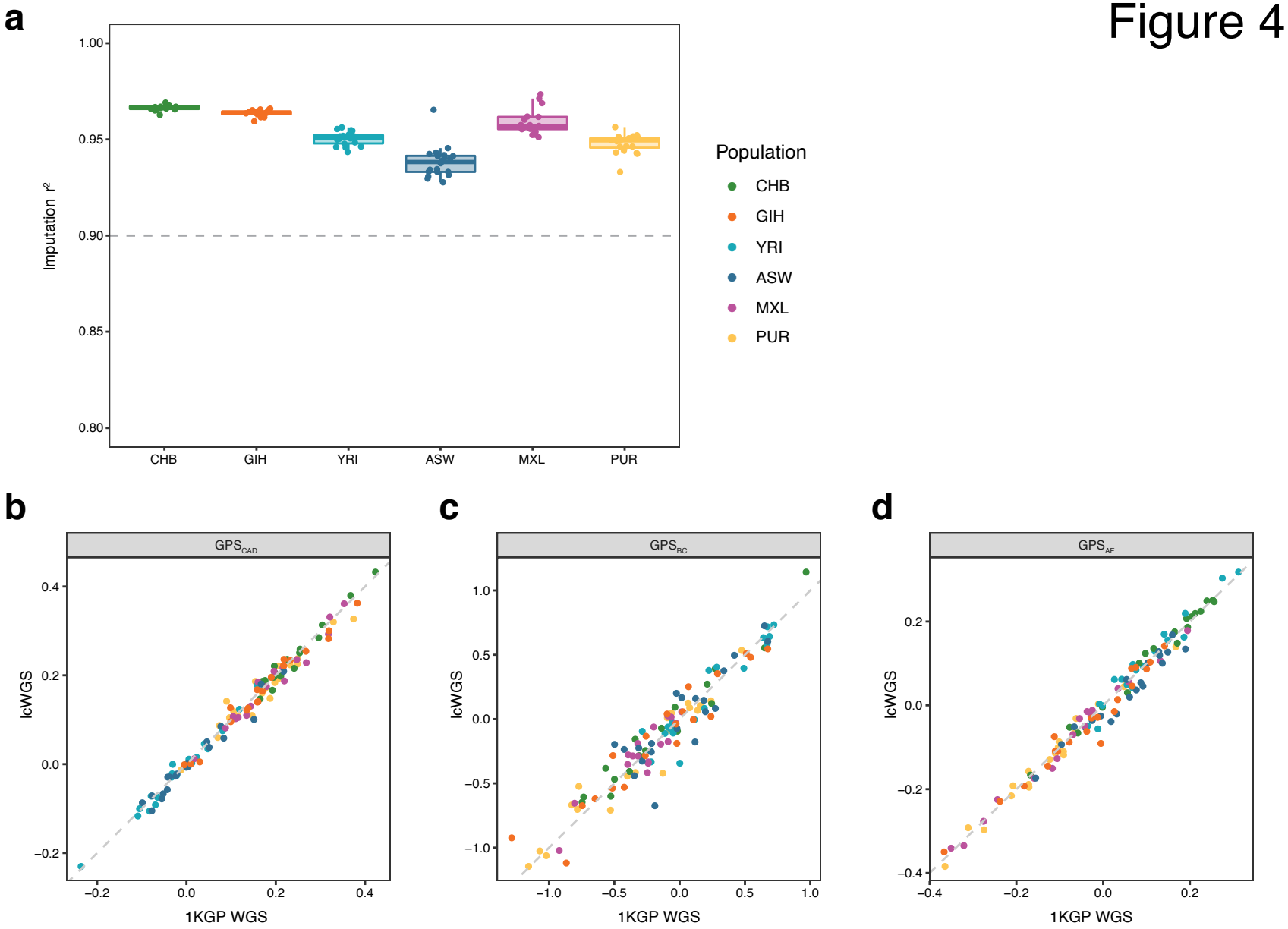
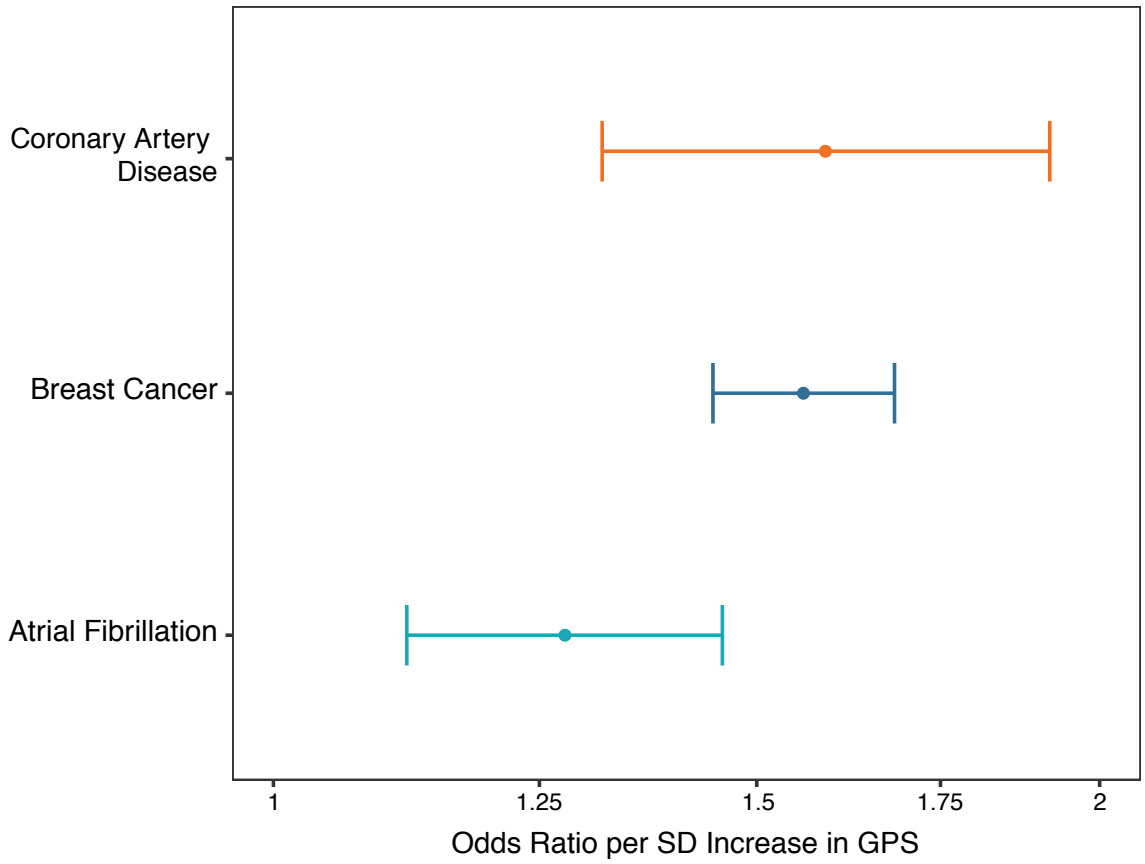

a
Imputation r²
1.00
0.95
0.90
0.85
0.80
CHB
GIH
YRI
ASW
MXL
PUR
Population
CHB
GIH
YRI
ASW
MXL
PUR
b
GPS_CAD
lcWGS
1KGP WGS
c
GPS_BC
lcWGS
1KGP WGS
d
GPS_AF
lcWGS
1KGP WGS

Figure 5


Coronary Artery Disease
Breast Cancer
Atrial Fibrillation
1
1.25
1.5
1.75
2
Odds Ratio per SD Increase in GPS