Very low depth whole genome sequencing in complex trait association studies

Arthur Gilly[1], Lorraine Southam[1,2], Daniel Suveges[1], Karoline Kuchenbaecker[1], Rachel Moore[1], Giorgio E.M. Melloni[1,3], Konstantinos Hatzikotoulas[1], Aliki-Eleni Farmaki[4], Graham Ritchie[1,5], Jeremy Schwartzentruber[1], Petr Danecek[1], Britt Kilian[1], Martin O. Pollard[1], Xiangyu Ge[1], Emmanouil Tsafantakis[6], George Dedoussis[4], Eleftheria Zeggini[1*]

[1] Department of Human Genetics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1HH, UK

[2] Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK

[3] Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139, Milan, Italy

[4] Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Greece

[5] European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SH, UK.

[6] Anogia Medical Centre, Anogia, Greece

* to whom correspondence should be addressed

Abstract

Motivation: Very low depth sequencing has been proposed as a cost-effective approach to capture low-frequency and rare variation in complex trait association studies. However, a full characterisation of the genotype quality and association power for very low depth sequencing designs is still lacking.

Results: We perform cohort-wide whole genome sequencing (WGS) at low depth in 1,239 individuals (990 at 1x depth and 249 at 4x depth) from an isolated population, and establish a robust pipeline for calling and imputing very low depth WGS genotypes from standard bioinformatics tools. Using genotyping chip, whole-exome sequencing (WES, 75x depth) and high-depth (22x) WGS data in the same samples, we examine in detail the sensitivity of this approach, and show that imputed 1x WGS recapitulates 95.2% of variants found by imputed GWAS with an average minor allele concordance of 97% for common and low-frequency variants. In our study, 1x further allowed the discovery of 140,844 true low-frequency variants with 73% genotype concordance when compared to high-depth WGS data. Finally, using association results for 57 quantitative traits, we show that very low depth WGS is an efficient alternative to imputed GWAS chip designs, allowing the discovery of up to twice as many true association signals than the classical imputed GWAS design.

Supplementary Data: Supplementary Data are appended to this manuscript.

Introduction

The contribution of low-frequency and rare variants to the allelic architecture of complex traits remains largely unchartered. Power to detect association is central to genetic studies examining sequence variants across the full allele frequency spectrum. Whole genome sequencing (WGS)-based association studies hold the promise of probing a larger proportion of sequence variation compared to imputed genome-wide genotyping arrays. However, although large-scale high-depth WGS efforts are now underway (Brody, et al., 2017), comparatively high costs do not yet allow for the generalised transposition of the GWAS paradigm to high-depth sequencing. As sample size and haplotype diversity are more important than sequencing depth in determining power for association studies (Alex Buerkle and Gompert, 2013; Le and Durbin, 2011), low-depth WGS has emerged as an alternative, cost-efficient approach to capture low-frequency variation in large studies. Improvements in calling algorithms have enabled robust genotyping using WGS at low depth (4x-8x), leading to the creation of large haplotype reference panels (1000 Genomes Project Consortium, et al., 2015; McCarthy, et al., 2016), and to several low-depth WGS-based association studies (Astle, et al., 2016; Tachmazidou, et al., 2017; UK10K Consortium, et al., 2015). Very low depth (<2x) sequencing has been proposed as an efficient way to further improve the cost efficiency of sequencing-based association studies. Simulations have shown that in WES designs, extremely low sequencing depths (0.1-0.5x) are effective in capturing single-nucleotide variants (SNVs) in the common (MAF>5%) and low-frequency (MAF 1-5%) categories compared to imputed GWAS arrays (Pasaniuc, et al., 2012). The CONVERGE consortium demonstrated the feasibility of such approaches through the first successful case-control study of major depressive disorder in 4,509 cases and 5,337 controls (Converge Consortium, 2015), and we previously showed that 1x WGS allowed the discovery of replicating burdens

66 of low-frequency and rare variants (Gilly, et al., 2016). However, a systematic examination of
67 genotyping quality from 1x WGS and its implications for power in association studies is
68 lacking, posing the question of the generalisability of such results in the wider context of next-
69 generation association studies. Here, we perform very low depth (1x), cohort-wide WGS in an
70 isolated population from Greece, show that imputation tools commonly used with chip data
71 perform well using 1x WGS, and establish a detailed quality profile of called variants. We then
72 demonstrate the advantages of 1x WGS compared to the more traditional imputed GWAS
73 design both in terms of genotype accuracy and power to detect association signals.
74

75 **Results**

76 As part of the Hellenic Isolated Cohorts (HELIC) study, we whole genome sequenced 990
77 individuals from the Minoan Isolates (HELIC-MANOLIS) cohort at 1x depth, on the Illumina
78 HiSeq2000 platform. In addition, 249 samples from the MANOLIS cohort were sequenced at
79 4x depth (Southam, et al., 2017). Imputation-based genotype refinement was performed on
80 the cohort-wide dataset using a combined reference panel of 10,244 haplotypes from
81 MANOLIS 4x WGS, the 1000 Genomes (1000 Genomes Project Consortium, et al., 2015) and
82 UK10K (UK10K Consortium, et al., 2015) projects (Figure 1).

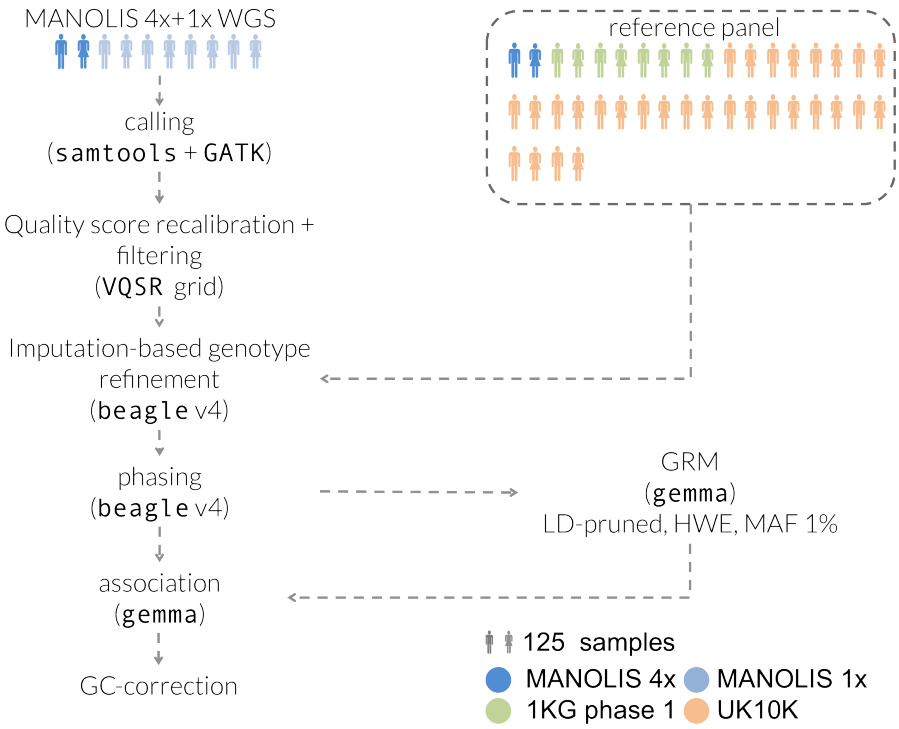


Figure 1: Processing pipeline for the MANOLIS 1x data. Tools and parameters for the genotype refinement and phasing steps were selected after benchmarking nine pipelines involving four different tools (See Methods).

Variant calling pipeline

Prior to any imputation-based refinement, our approach allowed the capture of 80% and 100% of low-frequency (MAF 1-5%) and common (MAF>5%) SNVs, respectively, when compared to variants present on the Illumina OmniExpress and HumanExome chips genotyped in the same samples. In 10 control samples with high-depth WGS data downsampled to 1x, joint calling with MANOLIS resulted in pre-imputation false positive and false negative rates of 12% and 24.6%, respectively (See Methods).

In order to improve sensitivity and genotype accuracy, we compared nine genotype refinement and imputation pipelines using tools commonly used for genotyping chip imputation, using directly typed OmniExpress and ExomeChip genotypes as a benchmark (See Methods). We used a reference panel containing haplotypes from 4,873 cosmopolitan

samples from the 1000 Genomes and UK10K projects, as well as the phased haplotypes from 249 MANOLIS samples sequenced at 4x depth. The best-performing pipeline, described in Figure 1, captures 95% of rare, 99.7% of low-frequency and 99.9% of common variants present in chip data, with an average minor allele concordance of 97% across the allele frequency spectrum (see Methods, Figure 2a., Supplementary Figure 1). 79.7% of 1x WGS variants were found using high-depth WGS at 22x in a subset of the MANOLIS samples (n=1,225), although this positive predictive value varied across the MAF spectrum, from 8.9% for singletons to 95.1% for common variants (Figure 2b.). Genotype concordance was similar, although slightly lower, when compared to the chip variants. Due to the 22x data being aligned to a different build, we were unable to compute genome-wide false positive rates, however by comparing 1x calls with those produced by whole-exome sequencing in 5 individuals from the MANOLIS cohort, we estimate a false-positive rate of 2.4% post-imputation in the coding parts of the genome (see Methods).

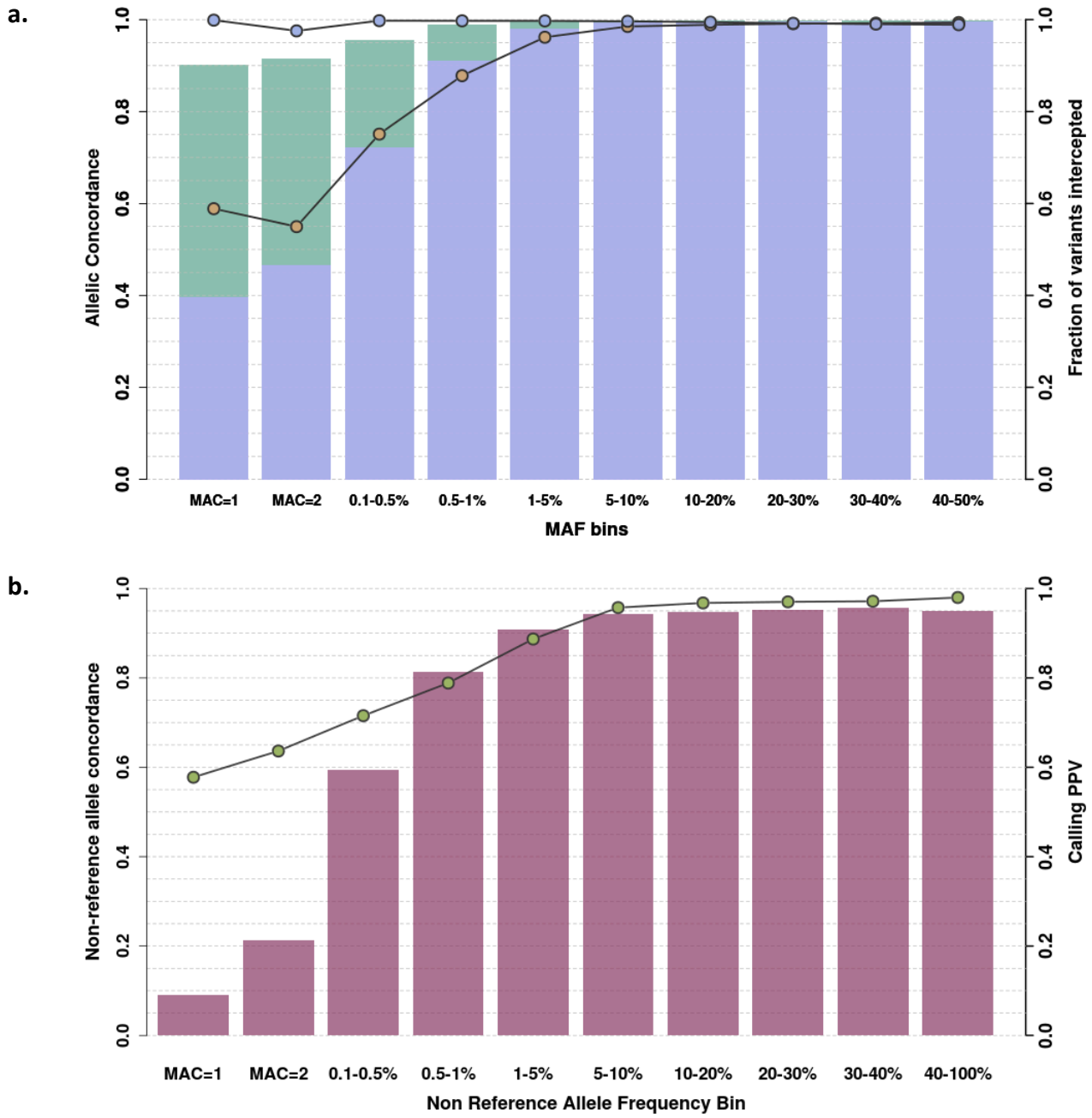


Figure 2: Concordance and call rate for very low depth WGS genotypes. a. Genotype (blue circles) and minor allele (yellow circles) concordance is computed for 1,239 samples in MANOLIS against merged OmniExpress and ExomeChip data. Call rate is assessed for the refined (purple) and refined plus imputed (green) datasets. **b.** Non-reference allele concordance (green circles) and positive predictive value (PPV) (fuchsia bars) is computed for 1,225 MANOLIS samples with both 22x WGS and low-depth calls.

Comparison of variant call sets with an imputed GWAS

132 The genotype refinement and imputation step yielded 30,483,136 non-monomorphic SNVs in
133 1,239 MANOLIS individuals. The number of variants discovered using 1x WGS is nearly twice
134 as high as that from array-based approaches. In a subset of 982 MANOLIS individuals with
135 both 1x WGS, OmniExpress and ExomeChip data, we called 25,673,116 non-monomorphic
136 SNVs using 1x WGS data, compared to 13,078,518 non-monomorphic SNVs in the same
137 samples with chip data imputed up to the same panel (Southam, et al., 2017) without any
138 imputation INFO score filtering. The main differences are among rare variants (MAF<1%)
139 (Figure 3): 13,671,225 (53.2%) variants called in the refined 1x WGS are absent from the
140 imputed GWAS, 98% of which are rare. 82% of these rare unique SNVs are singletons or
141 doubletons, and therefore 9.5% of all variants called in the 1x WGS dataset were unique
142 variants with MAC>2.

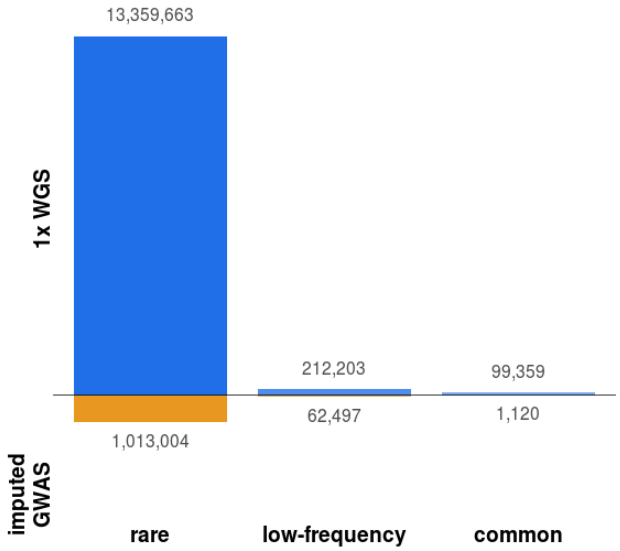
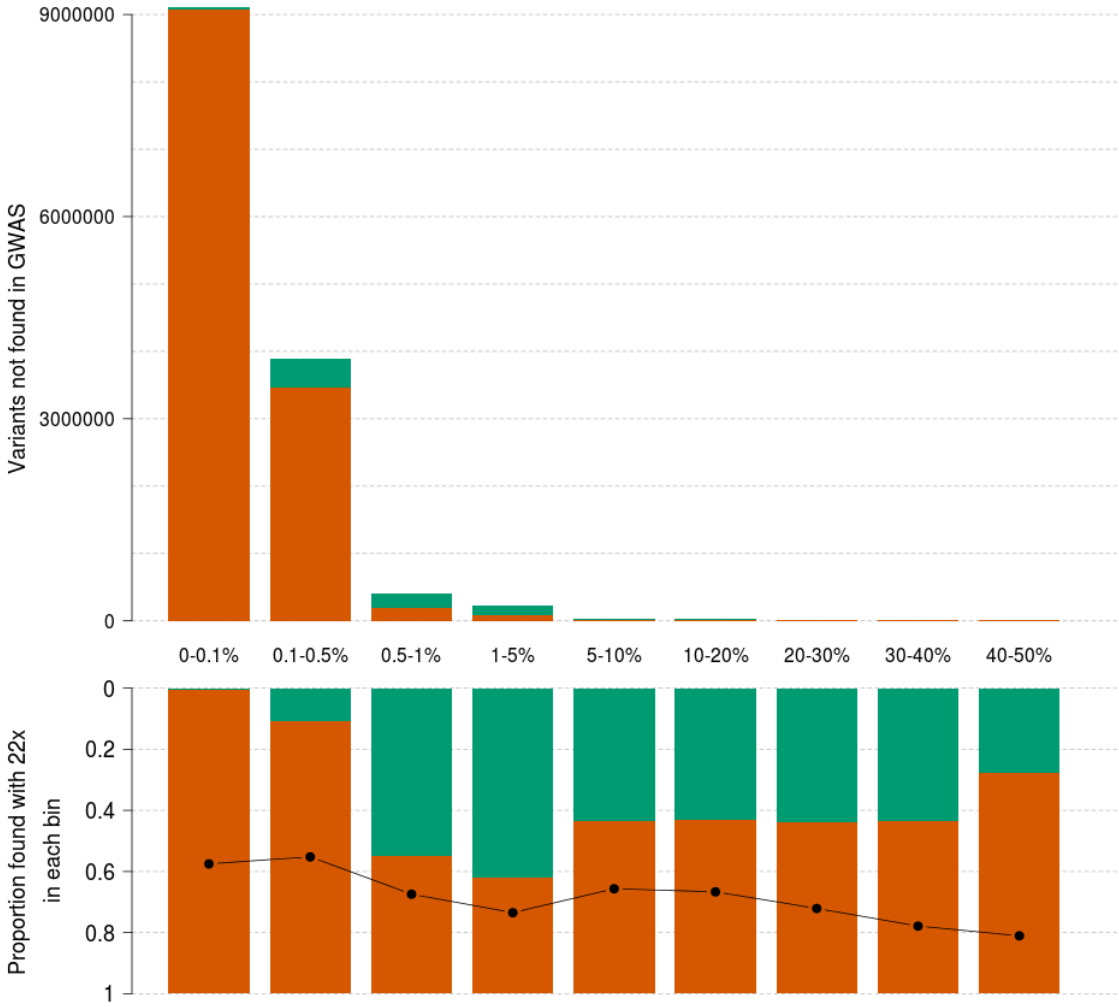


Figure 3: Unique variants called by sequencing and imputed GWAS. Variants unique to either dataset, arranged by MAF bin. Both datasets are unfiltered apart from monomorphics, which are excluded. MAF categories: rare (MAF<1%), low-frequency (MAF 1-5%), common (MAF>5%).

154

155 A crucial question is the proportion of true positives among these additional SNVs not found
156 by GWAS and imputation. By comparing their positions and alleles with high-depth WGS in
157 the same samples, we find that the PPV profile for these variants is much lower compared to
158 when all variants are examined (Figure 4 and Figure 2.b). As expected, PPV is almost zero for
159 additional singletons and doubletons, and just above 40% for the few additional common

160 variants. 62% of low-frequency variants unique to the 1x are true positives, which
161 corresponds to 140,844 low-frequency variants with high genotyping quality that are missed
162 by the imputed GWAS. Minor allele concordance is lower than for all variants, with a lower
163 bound at 55% for rare variants and reaching 73% for novel low-frequency variants.



164 **Figure 4: Positive predictive value of additional variants called in 1x sequencing**. 1x variants not found in
165 the GWAS data, arranged by MAF bin, in raw numbers (top). Green bars count variants recapitulated in the
166 22x (true positives). The proportion of these over the total (positive predictive value) is displayed in each
167 bin in the bottom panel. The black line indicates minor allele concordance for true positive variants. The
168 first category (0-0.1%) contains singletons and doubletons only.

169

170

171

172 **Comparison of association summary statistics with imputed GWAS**

173 1x WGS calls a larger number of variants and is noisier than imputed GWAS in the same
174 samples. To evaluate how this difference affects association study power, we performed
175 genome-wide association of 57 quantitative traits in 1,225 overlapping samples with both
176 imputed OmniExome and 1x WGS using both sources of genotype data. We then compared
177 independent suggestively associated signals at $p<5x10^{-7}$ (Supplementary Table 1). These
178 signals were then cross-referenced with a larger (n=1,457) study based on 22x WGS on the
179 same traits in the same cohort(Gilly, et al., 2018). We only considered signals to be true if
180 they displayed evidence for association with at most a two order of magnitude attenuation
181 compared to our suggestive significance threshold ($P<5x10^{-5}$). According to this metric, 52 of
182 182 independent signals (28.5%) were true in the imputed GWAS, in contrast to 108 of 462
183 (23.4%) in the 1x study (Figure 5). With an equal sample size and identically transformed
184 traits, 1x therefore allowed to discover twice as many independent GWAS signals with almost
185 identical truth sensitivity. Seven rare and three suggestive low-frequency variant associations
186 in the 1x WGS data (9.2% of all signals) were driven by a variant not present and without a
187 tagging SNP at $r^2>0.8$ in the imputed GWAS, whereas the converse is true for only two rare
188 variants in the imputed GWAS. Among variants called or tagged in the imputed GWAS, 4 rare,
189 11 low-frequency and 5 common SNV associations detected in the 1x (19% of total) are not
190 seen associated below that threshold in the imputed GWAS. As expected, there are
191 significantly fewer (3.8%, P=0.01, one-sided chi-square proportion test) true associations in
192 the imputed GWAS not recapitulated by the 1x study.

193

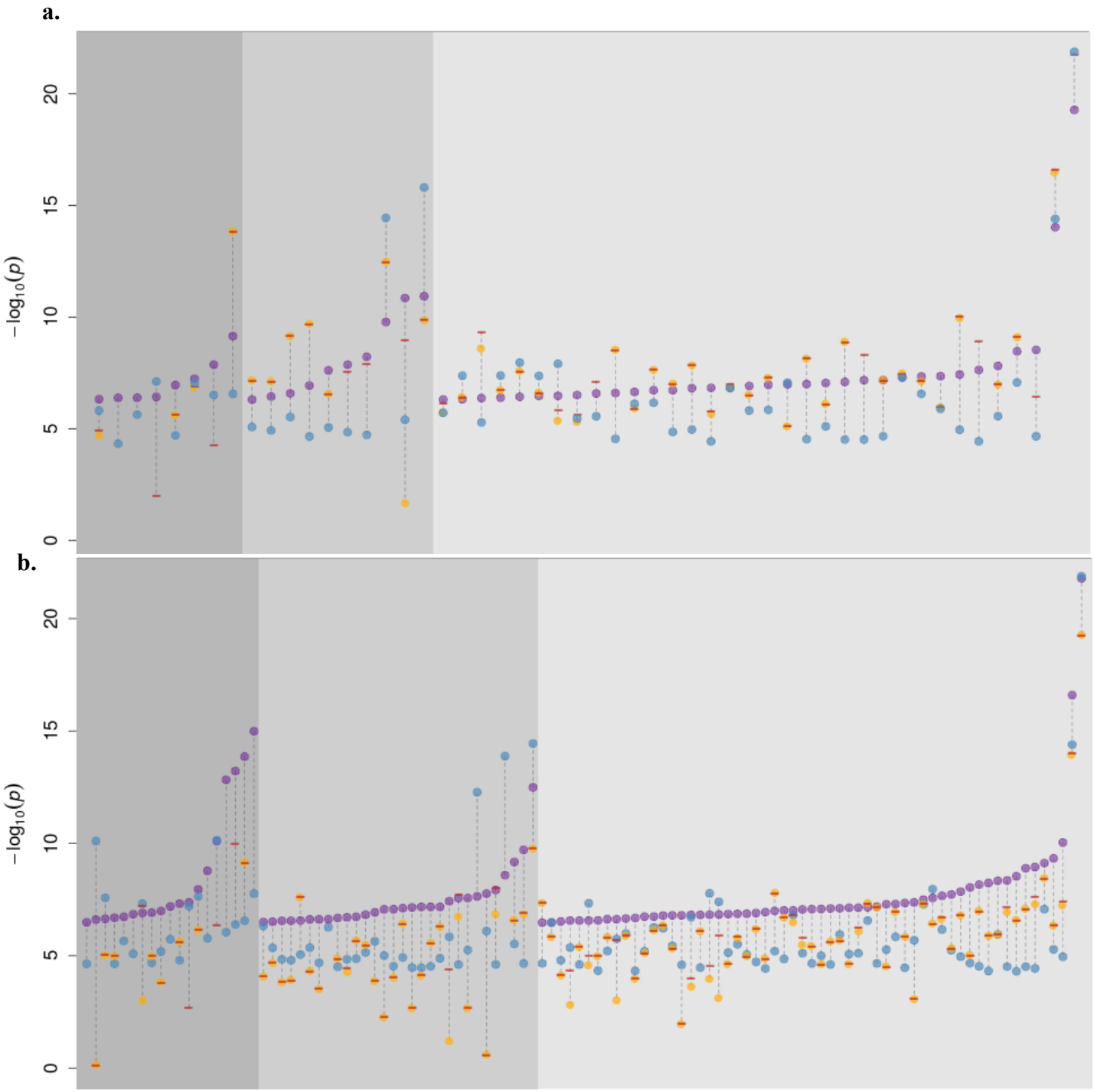


Figure 5: Association signals in the 1x WGS and imputed GWAS at $p<5x10^{-7}$ for 57 quantitative traits in 1,225 samples. Purple dots represent significant results in the 1x WGS (a.) and imputed GWAS (b.) analysis. Orange dots, if present, denote the p-value of the same SNP in the other study. Blue dots represent the association p-value in a larger (n=1,457) association study based on 22x WGS. Signals with a 22x WGS p-value above $5x10^{-5}$ were considered as false positives in both studies and excluded from the plot. Red dashes indicate the minimum p-value among all tagging SNVs in the other dataset (r^2>0.8). Absence of an orange dot and/or a red dash means that the variant was not present and/or no tagging variant could be found for that signal in the other study.

Discussion

205 In this work, we empirically demonstrate the relative merits of very low depth WGS both in
206 terms of variant discovery and association study power for complex quantitative traits
207 compared to GWAS approaches. However, the advantages of 1x WGS have to be weighed
208 against compute and financial cost considerations. As of summer 2018, 1x WGS on the HiSeq
209 4000 platform was approximately half of the cost of a dense GWAS array (e.g. Illumina
210 Infinium Omni 2.5Exome-8 array), the same cost as a sparser chip such as the Illumina
211 HumanCoreExome array, and half of the cost of WES at 50x depth. By comparison, 30x WGS
212 was 23 or 15 times more costly depending on the sequencing platform (Illumina HiSeq 4000
213 or HiSeqX, respectively). The number of variants called by 1x WGS is lower than high-depth
214 WGS, but is in the same order of magnitude, suggesting comparable disk storage
215 requirements for variant calls. However, storage of the reads required an average 650Mb per
216 sample for CRAMs, and 1.3Gb per sample for BAMs.
217
218 Genome-wide refinement and imputation of very low depth WGS generates close to 50 times
219 more variants than a GWAS chip. The complexity of the imputation and phasing algorithms
220 used in this study is linear in the number of markers, linear in the number of target samples
221 and quadratic in the number of reference samples (Browning and Browning, 2016), which
222 results in a 50-fold increase in total processing time compared to an imputed GWAS study of
223 equal sample size. In MANOLIS the genome was divided in 13,276 chunks containing equal
224 number of SNVs, which took an average of 31 hours each to refine and impute. The total
225 processing time was 47 core-years (see Methods and Supplementary Figure 2). This
226 parallelisation allowed processing the 1,239 MANOLIS samples in under a month, and as
227 imputation software continue to grow more efficient (Bycroft, et al., 2017), future pipelines
228 should greatly simplify postprocessing of very low depth sequencing data.

229

230 As a proof of principle, we used imputed GWAS, 1x and 22x WGS in overlapping samples from
231 an isolated population to assess how genotyping quality influences power in association
232 studies. As we only wanted to study the implications of varying genotype qualities afforded
233 by different designs on association p-values in a discovery setting, we considered only
234 suggestively associated signals and did not seek replication in a larger cohorts for the
235 discovered signals. In our study of 57 quantitative traits, we show that an 1x-based design
236 allows the discovery of twice as many of the signals suggestively associated in the more
237 accurate 22x WGS study, compared to the imputed GWAS design. Almost 10% of the
238 suggestive signals arising in the 1x data are not discoverable in the imputed GWAS, but the
239 great majority (96%) of imputed GWAS signals is found using the 1x.

240

241 The 1x-based study seems to discover more signals than the imputed GWAS across the MAF
242 spectrum, and this remains true whether or not the signals are filtered for suggestive
243 association p-value in the more accurate 22x based study (Supplementary Table 2). At first
244 glance this suggests 1x WGS has better detection power than the imputed GWAS across the
245 MAF spectrum, however it is unlikely that this is true for common variants, which are reliably
246 imputed using chip data. Instead, this phenomenon is likely due to a slightly less accurate
247 imputation than in the GWAS dataset caused by a noisier raw genotype input (Supplementary
248 Text). This effect is marginal, as evidenced by genome-wide concordance measures (Figure 2)
249 which are very high at the common end of the MAF spectrum. However, it is important to
250 note that this slightly less accurate imputation can attenuate some signals as well as boosting
251 others. For this reason, we would recommend relaxing the discovery significance threshold in

252 1x studies in order to capture those less well imputed, signal-harbouring variants, followed
253 by rigorous replication in larger cohorts and direct validation of genotypes.
254
255 Our study's intent was to focus on the performance on commonly used general-purpose tools
256 for low-depth sequencing data in isolates, both for genotype calling (GATK) and imputation
257 (BEAGLE, IMPUTE). There are ongoing efforts to leverage the specificities of both low-depth
258 sequencing (Davies, et al., 2016)(https://www.gencove.com) and of isolated populations
259 (Livne, et al., 2015). The popularity and long-term support of established generic methods is
260 an advantage when running complex study designs, as has been shown in other isolate studies
261 (Herzig, et al., 2018).
262
263 We show that very low depth whole-genome sequencing allows the accurate assessment of
264 most common and low-frequency variants captured by imputed GWAS designs and achieves
265 denser coverage of the low-frequency and rare end of the allelic spectrum, albeit at an
266 increased computational cost. This allows very low depth sequencing studies to recapitulate
267 signals discovered by imputed chip-based efforts, and to discover significantly associated
268 variants missed by GWAS imputation (Gilly, et al., 2016). Although cohort-wide high-depth
269 WGS remains the gold standard for the study of rare and low-frequency variation, very low-
270 depth WGS designs using population-specific haplotypes for imputation remain a viable
271 alternative when studying populations poorly represented in existing large reference panels.
272

273 # Methods

274 **Cohort details**

275 The HELIC (Hellenic Isolated Cohorts; www.helic.org) MANOLIS (Minoan Isolates) collection
276 focuses on Anogia and surrounding Mylopotamos villages on the Greek island of Crete. All
277 individuals were required to have at least one parent from the Mylopotamos area to enter
278 the study. Recruitment was primarily carried out at the village medical centres. The study
279 includes biological sample collection for DNA extraction and lab-based blood measurements,
280 and interview-based questionnaire filling. The phenotypes collected include anthropometric
281 and biometric measurements, clinical evaluation data, biochemical and haematological
282 profiles, self-reported medical history, demographic, socioeconomic and lifestyle
283 information. The study was approved by the Harokopio University Bioethics Committee and
284 informed consent was obtained from every participant.

285

286 **Sequencing**

287 Sequencing and mapping for the 990 MANOLIS samples at 1x depth has been described
288 previously (Gilly, et al., 2016), as well as for 249 MANOLIS samples at 4x (Southam, et al.,
289 2017), and for 1,457 samples at 22x (Gilly, et al., 2018). For comparison, 5 samples from the
290 cohort were also whole-exome sequenced at an average depth of 75x. We use a standard
291 read alignment and variant calling pipeline using samtools(Li, et al., 2009) and
292 GATK(McKenna, et al., 2010), which is described in detail in the Supplementary Text.

293

294 **Variant filtering**

295 Variant quality score recalibration was performed using GATK VQSR v.3.1.1. However, using
296 the default parameters for the VQSR mixture model yields poor filtering, with a Ti/Tv ratio

297 dropoff at 83% percent sensitivity and a Ti/Tv ratio of 1.8 for high-quality tranches
298 (Supplementary Figure 3.a). We therefore ran exploratory runs of VQSR across a range of
299 values for the model parameters, using the dropoff point of the transition/transversion
300 (Ti/Tv) ratio below 2.0 as an indicator of good fit (Supplementary Figure 4). A small number
301 of configurations outperformed all others, which allowed us to select an optimal set of
302 parameters. For the chosen set of parameters, false positive rate is estimated at 10%±5%
303 (Supplementary Figure 3.b). Indels were excluded from the dataset out of concerns for
304 genotype quality. We found that the version of VQSR, as well as the annotations used to train
305 the model, had a strong influence on the quality of the recalibration (Supplementary Figure 4
306 and Supplementary Text).
307
308 **Comparison with downsampled whole genomes**
309 For quality control purposes, reads from 17 of the well-characterised Platinum Genomes
310 sequenced by Illumina at 50x depth (McCarthy, et al., 2016), and downsampled to 1x depth
311 using samtools (Christopoulos, 1997) were included in the merged BAM file. VQSR-filtered
312 calls were then compared to the high-confidence call sets made available by Illumina for those
313 samples. 524,331 of the 4,348,092 non-monomorphic variant sites were not present in the
314 high-confidence calls, whereas 1,246,403 of the 5,070,164 non-monomorphic high-
315 confidence were not recapitulated in the 1x data. This corresponds to an estimated false
316 positive rate of 12% and false negative rate of 24.6%. Both unique sets had a much higher
317 proportion of singletons (corresponding to MAF < 2.9%) than the entire sets (57.9% vs 19.9%
318 of singletons among 1x calls and 51% vs 18.1% among high-confidence calls), which suggests
319 that a large fraction of the erroneous sites lies in the low-frequency and rare part of the allelic
320 spectrum. However, genotype accuracy is poor, to the point where it obscures peculiarities

321 in the distribution of allele counts (Supplementary Figure 5). Due to these being present in
322 the 1000 genomes reference panel, we remove the 17 Platinum Genomes prior to imputation.
323
324 **Genotype refinement and imputation**
325 *Evaluation of pipelines*
326 The authors of SHAPEIT (Delaneau, et al., 2013) advise to phase whole chromosome when
327 performing pre-phasing in order to preserve downstream imputation quality. This approach
328 is computationally intractable for the 1x datasets, where the smallest chromosomes contain
329 almost 7 times more variants than the largest chromosomes in a GWAS dataset.
330
331 For benchmarking purposes, we designed 13 genotype refinement pipelines involving Beagle
332 v4.0 (Browning and Browning, 2007) and SHAPEIT2 (Delaneau, et al., 2013) using a 1000
333 Genomes phase 1 reference panel, which we evaluated against minor allele concordance. All
334 pipelines were run using the vr-runner scripts
335 (https://github.com/VertebrateResequencing/vr-runner). Pipelines involving Beagle with the
336 use of a reference panel ranked consistently better (Supplementary Figure 1), with a single
337 run of reference-based refinement using Beagle outperforming all other runs. IMPUTE2
338 performed worst on its own, whether with or without reference panel; in fact the addition of
339 a reference panel did not improve genotype quality massively. Phasing with Beagle without
340 an imputation panel improved genotype quality, before or after IMPUTE2.
341
342 Halving the number of SNVs per refinement chunk (including 500 flanking positions) from the
343 4,000 recommended by the vr pipelines resulted in only a modest loss of genotype quality in
344 the rare part of the allelic spectrum (Supplementary Figure 7), while allowing for a twofold

345 increase in refinement speed. Genotype quality dropped noticeably for rare variants when
346 imputation was turned on (Supplementary Figure 7), but remained high for low-frequency
347 and common ones. A reference-free run of Beagle allowed to phase all positions and remove
348 genotype missingness with no major impact on quality and a low computational cost. We also
349 tested thunderVCF (Pollin, et al., 2008) for phasing sites, however, the program took more
350 than 2 days to run on 5,000 SNV chunks and was abandoned.
351
352 *Production pipeline for the MANOLIS cohort*
353 For production, we used a reference panel composed of 10,244 haplotypes from the 1000
354 Genomes Project Phase 1 (n=1,092), UK10K (UK10K Consortium, et al., 2015) TwinsUK
355 (Moayyeri, et al., 2013) and ALSPAC (Golding, et al., 2001) (n=3,781, 7x WGS), and 249
356 MANOLIS samples sequenced at 4x depth, which has been described before (Southam, et al.,
357 2017). Alleles in the reference panel were matched to the reference allele in the called
358 dataset. Positions where the alleles differed between the called and reference datasets were
359 removed from both sources. Indels were filtered out due to poor calling quality.
360
361 The pipeline with best minor allele concordance across the board used Beagle v.4 (Browning
362 and Browning, 2007) to perform a first round of imputation-based genotype refinement on
363 1,239 HELIC MANOLIS variant callsets, using the aforementioned reference panel. This was
364 followed by a second round of reference-free imputation, using the same software.
365
366
367
368

369 *Variant-level QC*

370 Beagle provides two position level imputation metrics, allelic R-squared (AR2) and dosage R-
371 squared (DR2). Both measures are highly correlated (Supplementary Figure 8.a). Values
372 between 0.3 and 0.8 are typically used for filtering (Browning, 2014). In both 1x datasets 59%
373 and 91% of imputed variants lie below those two thresholds, respectively. The distribution of
374 scores does not provide an obvious filtering threshold (Supplementary Figure 8.b) due to its
375 concavity. Since most imputed variants are rare and R-squared measures are highly correlated
376 with MAF, filtering by AR2 and DR2 would be similar to imposing a MAF threshold
377 (Supplementary Figure 8.c and d.). Moreover, due to a technical limitation of the vr-runner
378 pipelines, imputation quality measures were not available for refined positions at the time of
379 analysis, only imputed ones. Therefore, we did not apply any prior filter in downstream
380 analyses.

381

382 **Sample QC**

383 Due to the sparseness of the 1x datasets, sample-level QC was performed after imputation. 5
384 samples were excluded from the MANOLIS 1x cohort following PCA-based ethnicity checks.

385

386 **Comparison with WES**

387 A set of high confidence genotypes was generated for the 5 exomes in MANOLIS using filters
388 for variant quality (QUAL>200), call rate (AN=10, 100%) and depth (250x). These filters were
389 derived from the respective distributions of quality metrics (Supplementary Figure 9).

390 When compared to 5 whole-exome sequences from each cohort, imputed 1x calls
391 recapitulated 77.2% of non-monomorphic, high-quality exome sequencing calls. Concordance
392 was high, with only 3.5% of the overlapping positions exhibiting some form of allelic

393 mismatch. When restricting the analysis to singletons, 9105 (58%) of the 15,626 high-quality
394 singletons in the 10 exomes were captured, with 21% of the captured positions exhibiting
395 false positive genotypes (AC>1). To assess false positive call rate, we extracted 1x variants
396 falling within the 71,627 regions targeted by the Agilent design file for WES in overlapping
397 samples, and compared them to those present in the unfiltered WES dataset. 103,717
398 variants were called in these regions from WES sequences, compared to 58,666 non-
399 monomorphic positions in the 1x calls. 1,419 (2.4%) of these positions were unique to the 1x
400 dataset, indicating a low false-positive rate in exonic regions post-imputation.
401
402 **Genetic relatedness matrix**
403 Relatedness was present at high levels in our cohort, with 99.5% of samples having at least
404 one close relative (estimated $\hat{\pi} > 0.1$) and an average number of close relatives of 7.8. In
405 order to correct for this close kinship typical of isolated cohorts, we calculated a genetic
406 relatedness matrix using GEMMA (Zhou and Stephens, 2012). Given the isolated nature of the
407 population and the specificities of the sequencing dataset, we used different variant sets to
408 calculate kinship coefficients. Using the unfiltered 1x variant dataset produced the lowest
409 coefficients (Figure 10.a), whereas well-behaved set of common SNVs (Arthur, et al., 2017)
410 produced the highest, with an average difference of 3.67×10^{-3}. Filtering for MAF lowered the
411 inferred kinship coefficients. Generally, the more a variant set was sparse and enriched in
412 common variants, the higher the coefficients were. However, these differences only had a
413 marginal impact on association statistics, as evidenced by a lambda median statistic
414 difference of 0.02 between the two most extreme estimates of relatedness when used for a
415 genome-wide association of triglycerides (Supplementary Figure 10.b). For our association
416 study, we used LD-pruned 1x variants filtered for MAF<1% and Hardy Weinberg equilibrium

417 $p<1x10^{-5}$ to calculate the relatedness matrix, which translated into 2,848,245 variants for
418 MANOLIS.
419
420 **Single-point association**
421 *Pipeline*
422 For association, fifty-seven phenotypes were prepared, with full details of the trait
423 transformation, filters and exclusions described in Supplementary Table 3. The
424 'transformPhenotype' (https://github.com/wtsi-team144/transformPhenotype) R script was
425 used to apply a standardised preparation for all phenotypes. Association analysis was
426 performed on each cohort separately using the linear mixed model implemented in GEMMA
427 (Zhou and Stephens, 2012) on all variants with minor allele count (MAC) greater than 2
428 (14,948,665 out of 30,483,158 variants in MANOLIS). We used the aforementioned centered
429 kinship matrix. GC-corrected p-values from the likelihood ratio test (p_lrt) are reported.
430 Singletons and doubletons are removed due to overall low minor allele concordance. Signals
431 were extracted using the peakplotter software (https://github.com/wtsi-
432 team144/peakplotter) using a window size of 1Mb.
433
434 **Data Access**
435 The HELIC genotype and WGS datasets have been deposited to the European Genome-
436 phenome Archive (https://www.ebi.ac.uk/ega/home): EGAD00010000518;
437 EGAD00010000522; EGAD00010000610; EGAD00001001636, EGAD00001001637. The
438 peakplotter software is available at https://github.com/wtsi-team144/peakplotter, the
439 transformPhenotype app can be downloaded at https://github.com/wtsi-
440 team144/transformPhenotype.

Acknowledgements

The authors thank the residents of the Mylopotamos villages for taking part in the study. The MANOLIS study is dedicated to the memory of Manolis Giannakakis, 1978–2010. This study makes use of data generated by the UK10K Consortium, derived from samples from the TwinsUK Cohort and the Avon Longitudinal Study of Parents and Children (ALSPAC). A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org, last accessed April 29, 2016. The GATK program was made available through the generosity of Medical and Population Genetics program at the Broad Institute, Inc. This research has been conducted using the UK Biobank Resource using Application Number 13745. This work was funded by the Wellcome Trust [098051] and the European Research Council [ERC-2011-StG 280559-SEPI]. Funding for UK10K was provided by the Wellcome Trust under award WT091310. The authors thank Sophie Hackinger and Bram Prins for proofreading the article. The authors also thank Heather Elding, William Astle, Tao Jiang Adam Butterworth and Nicole Soranzo for their contributions to a previous version of the manuscript.

Disclosure Declaration

The authors declare that they have no competing interests.

463 **References**

464

465 1000 Genomes Project Consortium*, et al.* A global reference for human genetic variation.
466 *Nature* 2015;526(7571):68-74.

467 Alex Buerkle, C. and Gompert, Z. Population genomics based on low coverage sequencing:
468 how low should we go? *Mol Ecol* 2013;22(11):3028-3035.

469 Arthur, R.*, et al.* AKT: ancestry and kinship toolkit. *Bioinformatics* 2017;33(1):142-144.

470 Astle, W.J.*, et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to
471 Common Complex Disease. *Cell* 2016;167(5):1415-1429 e1419.

472 Brody, J.A.*, et al.* Analysis commons, a team approach to discovery in a big-data environment
473 for genetic epidemiology. *Nat Genet* 2017;49(11):1560-1563.

474 Browning, B.L. Private communication. 2014.

475 Browning, B.L. and Browning, S.R. Genotype Imputation with Millions of Reference Samples.
476 *Am J Hum Genet* 2016;98(1):116-126.

477 Browning, S.R. and Browning, B.L. Rapid and accurate haplotype phasing and missing-data
478 inference for whole-genome association studies by use of localized haplotype clustering. *Am*
479 *J Hum Genet* 2007;81(5):1084-1097.

480 Bycroft, C.*, et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*
481 2017.

482 Christopoulos, K.T.D. Minorities in Greece. Kritiki; 1997.

483 Converge Consortium. Sparse whole-genome sequencing identifies two loci for major
484 depressive disorder. *Nature* 2015;523(7562):588-591.

485 Davies, R.W.*, et al.* Rapid genotype imputation from sequence without reference panels. *Nat*
486 *Genet* 2016;48(8):965-969.

487 Delaneau, O.*, et al.* Haplotype estimation using sequencing reads. *Am J Hum Genet*
488 2013;93(4):687-696.

489 Gilly, A.*, et al.* Very low-depth sequencing in a founder population identifies a
490 cardioprotective APOC3 signal missed by genome-wide imputation. *Hum Mol Genet*
491 2016;25(11):2360-2365.

492 Gilly, A.*, et al.* Cohort-wide deep whole genome sequencing and the allelic architecture of
493 complex traits. *bioRxiv* 2018.

494 Golding, J.*, et al.* ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study
495 methodology. *Paediatr Perinat Epidemiol* 2001;15(1):74-87.

496 Herzig, A.F., *et al.* Strategies for phasing and imputation in a population isolate. *Genet*
497 *Epidemiol* 2018;42(2):201-213.

498 Le, S.Q. and Durbin, R. SNP detection and genotyping from low-coverage sequencing data on
499 multiple diploid samples. *Genome Res* 2011;21(6):952-960.

500 Li, H., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics*
501 2009;25(16):2078-2079.

502 Livne, O.E., *et al.* PRIMAL: Fast and accurate pedigree-based imputation from sequence data
503 in a founder population. *PLoS Comput Biol* 2015;11(3):e1004139.

504 McCarthy, S., *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat*
505 *Genet* 2016;48(10):1279-1283.

506 McKenna, A., *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
507 generation DNA sequencing data. *Genome Res* 2010;20(9):1297-1303.

508 Moayyeri, A., *et al.* The UK Adult Twin Registry (TwinsUK Resource). *Twin Res Hum Genet*
509 2013;16(1):144-149.

510 Pasaniuc, B., *et al.* Extremely low-coverage sequencing and imputation increases power for
511 genome-wide association studies. *Nat Genet* 2012;44(6):631-635.

512 Pollin, T.I., *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and
513 apparent cardioprotection. *Science* 2008;322(5908):1702-1705.

514 Southam, L., *et al.* Whole genome sequencing and imputation in isolated populations identify
515 genetic associations with medically-relevant complex traits. *Nat Comms* 2017(in press).

516 Southam, L., *et al.* Whole genome sequencing and imputation in two Greek isolated
517 populations identifies associations with complex traits of medical importance. *Nat Comms*
518 2017;in review.

519 Tachmazidou, I., *et al.* Whole-Genome Sequencing Coupled to Imputation Discovers Genetic
520 Signals for Anthropometric Traits. *Am J Hum Genet* 2017;100(6):865-884.

521 UK10K Consortium, *et al.* The UK10K project identifies rare variants in health and disease.
522 *Nature* 2015;526(7571):82-90.

523 Zhou, X. and Stephens, M. Genome-wide efficient mixed-model analysis for association
524 studies. *Nat Genet* 2012;44(7):821-824.

525