Very low depth whole genome sequencing in complex trait association studies

Arthur Gilly[1], Karoline Kuchenbaecker[1], Lorraine Southam[1,2], Daniel Suveges[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], Heather Elding[1,6], William J. Astle[7,8,9,10], Tao Jiang[10], Adam Butterworth[6,10,11], Nicole Soranzo[1,6,7,11], Emmanouil Tsafantakis[12], Maria Karaleftheri[13], George Dedoussis[4], Eleftheria Zeggini[1*]

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

[2] Wellcome Trust 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] The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge CB1 8RN, UK

[7] Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0PT, UK

[8] National Health Service (NHS) Blood and Transplant, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0PT, UK

[9] Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Cambridge Biomedical Campus, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK

[10] MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge CB1 8RN, UK

[11] British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK

[12] Anogia Medical Centre, Anogia, Greece

[13] Echinos Medical Centre, Echinos, Greece

* corresponding author

Abstract

Background

Very low depth sequencing is a cost-effective approach to capture low-frequency and rare variation in complex trait association studies. Here, we perform cohort-wide whole genome sequencing (WGS) at 1x depth coupled to genome-wide association analysis in 2,347 individuals from two isolated populations.

Results

We establish a robust pipeline for calling 1x WGS data, achieving an average minor allele concordance of 97% when compared to genotyping chip data. 9.5% of variants called using 1x WGS are variants with a high predicted quality not captured by genome-wide association study (GWAS) data in the same individuals imputed to a dense haplotype reference panel. Of the 54 association signals arising from genome-wide association analysis of 1x WGS variants with 25 haematological traits (at $p<5x10^{-7}$), only 57% are recapitulated by the imputed GWAS results in the same samples. Differences in strength of evidence for association are smaller for common than for low-frequency and rare variant signals. We further exemplify power gains by establishing robust evidence for a novel association between rs6489858, an intronic variant in *RPH3A* and increased lymphocyte count (beta=0.13, SE=0.11, $p=8x10^{-12}$), which replicates in an independent dataset comprising 173,480 samples.

Conclusions

We show that 1x WGS is an efficient alternative to imputed GWAS chip designs for empowering next-generation association studies in complex traits. We demonstrate that population-scale 1x WGS allows the interrogation of a large number of low-frequency and

rare variants missed by classical GWAS array imputation, resulting in potential higher association power.

Keywords

Whole-genome sequencing, association studies, population isolates

Introduction

Characterisation of the genetic determinants underpinning complex human traits of medical relevance can help improve our understanding of aetiopathology and point to biological processes amenable to intervention. Despite great progress in identifying common-frequency variants with small to modest effect sizes, the allelic architecture of low-frequency and rare variants for 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 genome-wide genotyping arrays. However, high-depth WGS costs do not yet allow application of the GWAS paradigm to large-scale sequencing of hundreds of thousands of individuals. As sample size and haplotype diversity are more important than sequencing depth in determining power for association studies [1], 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 reference haplotype panels [2, 3], and to the start of WGS-based association studies [4, 5]. Very low depth (<2x) sequencing has been

79 proposed as an efficient way to further improve the cost efficiency of sequencing-based
80 association studies. Simulations have shown that in whole-exome studies, extremely low
81 sequencing depths (0.1-0.5x) are effective in capturing single-nucleotide variants (SNVs) in
82 the common (MAF>5%) and low-frequency (MAF 1-5%) categories compared to imputed
83 GWAS arrays [6]. The CONVERGE consortium demonstrated the feasibility of such
84 approaches through the first successful case-control study of major depressive disorder in
85 4,509 cases and 5,337 controls [7].
86
87 Studying founder populations can further empower the search for association signals by
88 allowing the detection of population-specific variants, and of association signals at variants
89 that have drifted up in frequency compared to cosmopolitan populations, against the
90 backdrop of a homogeneous environment [8] [9] [10] [11]. Here, we perform very low depth
91 (1x), cohort-wide WGS in two isolated populations from Greece. We establish a robust 1x
92 WGS calling pipeline and compare the complement of variants captured to imputed GWAS
93 in the same samples. As a proof-of-principle, we perform association analysis across
94 medically-relevant haematological traits and identify a robustly-replicating novel locus
95 implicated in lymphocyte counts.
96

97 **Results**

98 As part of the Hellenic Isolated Cohorts (HELIC) study, we whole genome sequenced 990
99 individuals from the Minoan Isolates (HELIC-MANOLIS) cohort, and 1108 individuals from
100 the Pomak villages (HELIC-Pomak) at 1x depth, on the Illumina HiSeq2000 platform. In
101 addition, 249 samples from the MANOLIS cohort were sequenced at 4x depth [12].

102 Imputation-based genotype refinement was performed on the two cohort-wide datasets
103 using a combined reference panel of 10,422 haplotypes from MANOLIS 4x WGS, the 1000
104 Genomes [2] and UK10K [4] projects.
105
106 **Variant calling pipeline**
107 We established a variant calling, quality control (QC) and genotype refinement pipeline for
108 very low depth WGS by benchmarking nine pipelines that make use of state of the art
109 bioinformatics tools (Methods). Our optimised approach allowed the capture of 80% of true
110 low-frequency (MAF 1-5%) variants and 100% of true common-frequency (MAF>5%) SNVs
111 prior to imputation-based refinement, when compared to variants present on the Illumina
112 OmniExpress and HumanExome chips genotyped in the same samples. In order to assess
113 sensitivity and specificity of SNV calls pre-imputation, we estimate the false positive and
114 false negative rate by comparing 1x WGS variant calls with high-depth WGS data (see
115 Methods). We estimate that 12% of 1x sites with at least one heterozygote call are false
116 positives, whereas 24.6% of the sites called with high-depth WGS are not recapitulated
117 using 1x data.
118
119 In order to improve the false negative rate and genotype accuracy, we performed genotype
120 refinement and imputation using a large reference panel containing haplotypes from 4,873
121 cosmopolitan samples as well as the phased haplotypes from the 249 MANOLIS samples
122 sequenced at 4x depth. After imputation and QC, we captured 95% of rare, 99.7% of low-
123 frequency and 99.9% of common variants compared to the Illumina OmniExpress and
124 HumanExome GWAS chips, with an average minor allele concordance of 97% across the
125 allele frequency spectrum (Methods and Figure 1). By comparing 1x calls with those

126 produced by whole-exome sequencing in 10 individuals across both cohorts, we estimate a
127 false-positive rate of 2.4% post-imputation in the coding parts of the genome (Methods).
128
129
130 **Comparison of variant call sets with an imputed GWAS**
131 The genotype refinement and imputation step yielded 30,483,136 and 29,740,259 non-
132 monomorphic SNVs in 1,239 MANOLIS and 1,108 Pomak individuals, respectively. The
133 number of variants discovered using 1x WGS is nearly twice as high as that from array-based
134 approaches. In a subset of 982 MANOLIS individuals with 1x WGS, we called 25,673,116
135 non-monomorphic SNVs using 1x WGS data, compared to 13,078,518 non-monomorphic
136 SNVs in the same samples with OmniExpress and ExomeChip data imputed up to the same
137 panel [9]. The main differences are among rare variants (MAF<1%) (Figure 2): 13,671,225
138 (53.2%) variants called in the refined 1x WGS are absent from the imputed GWAS, 98% of
139 which are rare. 82% of these rare unique SNVs are singletons or doubletons, and therefore
140 9.5% of all variants called in the 1x WGS dataset were unique variants with MAC>2.
141
142 **Experimental validation of genotypes**
143 We performed experimental genotyping of 65 variants (23 common, 18 low-frequency and
144 24 rare) in a subset of 1087 and 859 samples in the MANOLIS and Pomak cohorts,
145 respectively, using the Agena Biosciences MassARRAY technology. On average, minor allele
146 concordance was 76% and positive predictive value was 82%. As expected, these values
147 differ between MAF categories (Additional File 1: Table S1). Minor allele and genotype
148 concordance between 1x calls and this set of directly assayed genotypes were in line with
149 those computed genome-wide between 1x calls and GWAS data (Figure 1).

150

151 **Association analysis**

152 As a proof of principle, we performed single-point association analysis across 25
153 haematological traits with 14,948,665 and 15,564,905 variants with MAC>2 in MANOLIS and
154 Pomak, respectively (Methods). We used an empirical genetic relatedness matrix calculated
155 on high-confidence genotypes to account for relatedness within the two isolated cohorts.

156

157 Genome-wide significance was set at $p<1.0x10^{-9}$ based on the effective number of traits and
158 tested variants (see Methods). In the discovery sample, one association met this threshold
159 in the Pomak dataset. rs35004220, located in an intron of the haemoglobin B (*HBB*) gene,
160 was associated with six red blood cell traits (haemoglobin, mean corpuscular haemoglobin,
161 mean corpuscular volume, red cell distribution width in volume and percent, red blood
162 cells) (Additional File 1 and 2: Figure S1 and Table S2). A total of 5,090 variants with
163 association $p<1.0x10^{-5}$ in the discovery stage corresponding to 556 and 465 independent
164 signals in the MANOLIS and Pomak cohorts, respectively (Additional File 1: Table S2 and S3),
165 were carried forward to *in silico* replication using data from a large meta-analysis of 173,480
166 samples from the UK Biobank and INTERVAL studies [5]. Out of the 3,336 variants for which
167 replication data were available, 52.4% had a concordant direction of effect compared to the
168 discovery stage ($p=2.9x10^{-3}$, one-sided binomial test). Upon meta-analysis of the discovery
169 and replication data, we identify a previously unreported, genome-wide significantly
170 associated signal (Figure 3). The G-allele of rs6489858 (EAF=0.40) at 12q24.13 is associated
171 with increased lymphocyte count in MANOLIS (beta=0.022, SE=0.004, $p=2.29x10^{-9}$ in the
172 discovery and replication meta-analysis). We found evidence of heterogeneity (I^2=95.4%, Q-
173 statistic $p=2x10^{-6}$) and therefore applied a random effects meta-analysis model [13]

174 (beta=0.13, SE=0.11, $p=8x10^{-12}$). rs6489858 is located in an intronic region of the rabphilin
175 3A (*RPH3A*) gene, which encodes a peripheral membrane protein involved in protein
176 transport and synaptic vesicle traffic.
177
178 **Comparison of association summary statistics with imputed GWAS**
179 1x WGS calls a larger number of variants than imputed GWAS of the same samples. To
180 evaluate how this difference affects association study power, we compared the association
181 results of all independent suggestive signals at $p<5x10^{-7}$ from the 1x WGS with the imputed
182 GWAS results for the same variants (Figure 4). Among the 54 variants significantly
183 associated at this threshold in the 1x WGS, 17 (31%) were not observed in the imputed
184 GWAS study. Rare (MAF<1%) variant signals are more poorly captured by the imputed
185 GWAS, with 10 out of 16 signals (62.5%) being missed, however, for 12 (70%) of the missed
186 variants, a tagging SNV at $r^2>0.8$ was available in the imputed GWAS. For the signals where
187 imputed GWAS results are indeed present, the majority (62.2%) do not meet our
188 significance threshold, an effect which is more marked in the rare and low-frequency
189 (16/23, 69%) than in the common (7/14, 50%) signals. This observation persists when
190 considering tagging variants at $r^2>0.8$: twenty-seven (55%) out of the 49 taggable 1x WGS
191 signals have a p-value above the significance threshold. Generally, differences in p-values
192 between the two studies were smaller for common than for low-frequency and rare variant
193 signals (5.5 and 2.0 on the log-scale, $p=6x10^{-3}$, two-sample t-test).
194
195
196 **Discussion**

197 In this work, we empirically demonstrate the relative merits of very low depth WGS both in
198 terms of variant discovery and association study power for complex quantitative traits
199 compared to GWAS approaches. However, the advantages of 1x WGS have to be weighed
200 against compute and financial cost considerations. As of January 2017, 1x WGS on the HiSeq
201 4000 platform was approximately half of the cost of a dense GWAS array (e.g. Illumina
202 Infinium Omni 2.5Exome-8 array), 1.5 times the cost of a sparser chip such as the Illumina
203 HumanCoreExome array, and a third of the cost of WES at 50x depth. By comparison, 30x
204 WGS was 21 or 16 times more costly depending on the sequencing platform (Illumina HiSeq
205 4000 or HiSeqX, respectively). The number of variants called by 1x WGS is lower than high-
206 depth WGS, but is in the same order of magnitude, suggesting comparable disk storage
207 requirements for variant calls. However, storage of the reads required an average 650Mb
208 per sample for CRAMs, and 1.3Gb per sample for BAMs.
209
210 Genome-wide refinement and imputation of very low depth WGS generates close to 50
211 times more variants than a GWAS chip. The complexity of the imputation and phasing
212 algorithms used in this study is linear in the number of markers, linear in the number of
213 target samples and quadratic in the number of reference samples [14], which results in a 50-
214 fold increase in total processing time compared to an imputed GWAS study of equal sample
215 size. Therefore, parallelisation plays a crucial role in managing computational load. For
216 example, in MANOLIS the genome was divided in 13,276 chunks containing equal number of
217 SNVs, which took an average of 31 hours each to refine and impute. The total processing
218 time was 47 core-years (Methods and Additional File 2: Figure S2). Parallelisation allowed
219 processing the 1,239 MANOLIS samples in under a month.
220

221 As a proof of principle, we used 1x WGS in samples from isolated populations and identified
222 a novel association with lymphocyte count, previously missed by large-scale GWAS in
223 cosmopolitan populations [15-17]. The signal had a much larger effect size in the isolated
224 population discovery cohort (beta=0.25 standard deviation increase in the discovery
225 samples, beta=0.02 in the replication cohorts for rs6489858). In the subsample of 1,225
226 individuals with both 1x and GWAS data, minor allele concordance was 99.5% for
227 rs6489858, and the association p-value with LYM was in the same order of magnitude
228 ($p=7.5x10^{-7}$ in the imputed GWAS, $p=3.2x10^{-7}$ in the 1x WGS data). rs6489858 is located
229 260kb from *PTPN11*. In juvenile myelomonocytic leukemia, the RAS/MAPK pathway is
230 frequently deregulated due to somatic mutations in *PTPN11* [18]. *PTPN11* is also involved in
231 LEOPARD syndrome, metachondromatosis and Noonan syndrome. Animal models of this
232 gene show diverse and severe phenotypes including hematopoietic abnormalities such as
233 abnormal leukopoiesis [19]. *DTX1* is located 280kb away from the index variant. Deltex-1,
234 the cytoplasmic protein product of this gene, is a regulator of the Notch pathway and plays
235 an important role in the development of B and T lymphocytes [20]. Animal models of this
236 gene show various hematopoietic and immune related abnormalities [21] due to interfering
237 with T cell development.
238
239

240 **Conclusions**

241 We show that very low depth whole-genome sequencing allows the accurate assessment of
242 most common and low-frequency variants captured by imputed GWAS designs and achieves
243 denser coverage of the low-frequency and rare end of the allelic spectrum, albeit at an
244 increased computational cost. This allows very low depth sequencing studies to identify

245 signals also discoverable by imputed chip-based efforts, and to discover significantly
246 associated variants missed by GWAS imputation [22]. As sequencing technologies continue
247 to evolve, higher sequencing depths will provide accurate genotyping across the full range
248 of the allelic spectrum, enabling comprehensive exploration of human phenotype
249 associations through rare variant aggregation tests.
250

251 Materials and methods

252 Cohort details

253 The HELIC (Hellenic Isolated Cohorts; www.helic.org) MANOLIS (Minoan Isolates) collection
254 focuses on Anogia and surrounding Mylopotamos villages on the Greek island of Crete. All
255 individuals were required to have at least one parent from the Mylopotamos area to enter
256 the study. The HELIC Pomak collection focuses on the Pomak villages, a set of isolated
257 mountainous villages in the North of Greece. Recruitment of both population-based
258 samples was primarily carried out at the village medical centres. The study includes
259 biological sample collection for DNA extraction and lab-based blood measurements, and
260 interview-based questionnaire filling. The phenotypes collected include anthropometric and
261 biometric measurements, clinical evaluation data, biochemical and haematological profiles,
262 self-reported medical history, demographic, socioeconomic and lifestyle information.
263

264 Sequencing

265 Sequencing and mapping for the 995 MANOLIS samples at 1x depth has been described
266 before [22], as well as for 250 MANOLIS samples at 4x [9]. 1166 samples from the Pomak
267 were sequenced at 1x depth using the same protocol using Illumina HiSeq 2000 and Illumina

268 HiSeq 2500 sequencers. For comparison, 5 samples from each cohort were whole-exome
269 sequenced at an average depth of 75x.
270
271 **Read mapping and variant calling**
272 Following generation of raw reads on the Illumina HiSeq 2000 and HiSeq 2500 sequencing
273 machines, reads were converted from BCL format to BAM format using the Illumina2BAM
274 (https://github.com/wtsi-npg/illumina2bam) software. Illumina2BAM was again used to de-
275 multiplex lanes that had been sequenced so that the tags were isolated from the body of
276 the read, decoded, and could be used to separate out each lane into lanelets containing
277 individual samples from the multiplex library and the PhiX control. The quality scores were
278 then recalibrated using the purity recalibration algorithm [23] using the PhiX data for
279 reference. Read mapping was then carried out using the BWA backtrack algorithm version
280 0.5.10 using the GRCh37 1000 Genomes phase III reference (also known as hs37d5). PCR
281 and optically duplicated reads were marked using Picard MarkDuplicates
282 (http://broadinstitute.github.io/picard).
283
284 In order to ensure the quality of the large quantity of BAMs produced for the project, an
285 automatic quality control system was used to reduce the number of data files that required
286 manual intervention. This system was derived from the one originally designed for the
287 UK10K project (http://www.uk10k.org) and used a series of empirically derived thresholds
288 to assess summary metrics calculated from the input BAMs. These thresholds included:
289 percentage of reads mapped; percentage of duplicate reads marked; various statistics
290 measuring INDEL distribution against read cycle and an insert size overlap percentage. Any
291 lane that fell below the “fail” threshold for any of the metrics were excluded; and any lane

292 that did not fall below these thresholds for any of the metrics was given a status of "pass"
293 and allowed to proceed into the later stages of the pipeline.
294
295 Passed lanelets were then merged into BAMs corresponding to the libraries for each sample
296 and duplicates were marked again with Picard MarkDuplicates after which they were then
297 merged into BAMs on a per sample basis. Finally sample level bam improvement was
298 carried out using GATK 1.6[24, 25] and samtools[26] from git commit
299 72d6457f7f361c323f62bd2d3170980132ba2113. This consisted of re-alignment of reads
300 around known and discovered INDELs followed by base quality score recalibration both
301 using the GATK, lastly samtools calmd was applied and indexes were created. Known
302 INDELs for realignment were taken from Mill Devine and 1000G Gold set and the 1000G
303 phase low coverage set both part of the Broad's GATK resource bundle version 2.2. Known
304 variants for BQSR were taken from dbSNP 137 also part of the Broad's resource bundle.
305
306 The input BAM files are fed into samtools mpileup to create all-sites BCF files, which are
307 piped into bcftools view to create variant-only VCF files containing genotype calls. We split
308 the genome into chunks of 100,000 base pairs, and separate these chunks into SNV and
309 INDEL files. We run GATK UnifiedGenotyper to calculate site-level annotations.
310
311 **Variant filtering**
312 Variant quality score recalibration was performed using GATK VQSR v.3.1.1. However, using
313 the default parameters for the VQSR mixture model yields poor filtering, with a Ti/Tv ratio
314 dropoff at 83% percent sensitivity and a Ti/Tv ratio of 1.8 for high-quality tranches
315 (Additional File 2: Figure S3.a). We therefore ran exploratory runs of VQSR across a range of

316 values for the model parameters, using the dropoff point of the transition/transversion
317 (Ti/Tv) ratio below 2.0 as an indicator of good fit (Additional File 2: Figure S4). A small
318 number of configurations outperformed all others, which allowed us to select an optimal set
319 of parameters. For the chosen set of parameters, false positive rate is estimated at 10%±5%
320 (Additional File 2: Figure S3.b). Indels were excluded from the dataset out of concerns for
321 genotype quality. We found that the version of VQSR, as well as the annotations used to
322 train the model, had a strong influence on the quality of the recalibration (Additional File 2:
323 Figure S4 and Supplementary Text).
324
325 **Comparison with Platinum genomes**
326 For quality control purposes, reads from 17 of the well-characterised Platinum Genomes
327 sequenced by Illumina at 50x depth [3], and downsampled to 1x depth using samtools [26]
328 were included in the merged BAM file. VQSR-filtered calls were then compared to the high-
329 confidence call sets made available by Illumina for those samples. 524,331 of the 4,348,092
330 non-monomorphic variant sites were not present in the high-confidence calls, whereas
331 1,246,403 of the 5,070,164 non-monomorphic high-confidence were not recapitulated in
332 the 1x data. This corresponds to an estimated false positive rate of 12% and false negative
333 rate of 24.6%. Both unique sets had a much higher proportion of singletons (corresponding
334 to MAF < 2.9%) than the entire sets (57.9% vs 19.9% of singletons among 1x calls and 51% vs
335 18.1% among high-confidence calls), which suggests that a large fraction of the erroneous
336 sites lies in the low-frequency and rare part of the allelic spectrum. However, genotype
337 accuracy is poor, to the point where it obscures peculiarities in the distribution of allele
338 counts (Additional File 2: Figure S5). Due to them being present in the 1000 genomes
339 reference panel, we remove the 17 Platinum Genomes prior to imputation.

340

341

342 **Genotype refinement and imputation**

343 *Reference Panel*

344 Phased haplotypes from 1092 samples from the 1000 Genomes Project Phase 1 study were
345 merged with 3781 7x WGS samples from the UK10K [4] TwinsUK [27] and ALSPAC [28]
346 studies, and with 249 MANOLIS samples sequenced at 4x depth [9] using SHAPEIT v2 [29]
347 and converted to VCF format. Alleles in the reference panel were flipped so as to
348 correspond to the reference allele in the called dataset. Positions where the alleles differed
349 between the called and reference datasets were removed from both sources. Indels were
350 filtered out due to poor calling quality.

351

352 *Pipeline*

353 As described previously [22], we used Beagle v.4 [30] to perform a first round of imputation-
354 based genotype refinement on 1,239 HELIC MANOLIS variant callsets, using a previously
355 described [9] reference panel composed of 10,244 haplotypes from the 1000 Genomes,
356 UK10K and MANOLIS 4x reference sequences. This was followed by a second round of
357 reference-free imputation, using the same software. The same pipeline was applied to 1166
358 individuals from the Pomak cohort.

359

360 *Evaluation of pipelines*

361 The authors of SHAPEIT [29] advise to phase whole chromosome when performing pre-
362 phasing in order to preserve downstream imputation quality. This approach is

363 computationally intractable for the 1x datasets, where the smallest chromosomes contain
364 almost 7 times more variants than the largest chromosomes in a GWAS dataset.
365
366 For benchmarking purposes, we tested 13 genotype refinement pipelines involving Beagle
367 v4.0 [30] and SHAPEIT2 [29] using a 1000 Genomes phase 1 reference panel, which we
368 evaluated against minor allele concordance. All pipelines were run using the vr-runner
369 scripts (https://github.com/VertebrateResequencing/vr-runner). Pipelines involving Beagle
370 with the use of a reference panel ranked consistently better (Additional File 2: Figure S6),
371 with a single run of reference-based refinement using Beagle outperforming all other runs.
372 IMPUTE2 performed worst on its own, whether with or without reference panel; in fact the
373 addition of a reference panel did not improve genotype quality massively. Phasing with
374 Beagle without an imputation panel improved genotype quality, before or after IMPUTE2.
375
376 Halving the number of SNVs per refinement chunk to 2,000 (including 500 flanking
377 positions) resulted in only a modest loss of genotype quality in the rare part of the allelic
378 spectrum (Additional File 2: Figure S7), while allowing for a twofold increase in refinement
379 speed. Genotype quality dropped noticeably for rare variants when imputation was turned
380 on (Additional File 2: Figure S7), but remained high for low-frequency and common ones. A
381 reference-free run of Beagle allowed to phase all positions and remove genotype
382 missingness with no major impact on quality and a low computational cost. We also tested
383 thunderVCF [31] for phasing sites, however, the program took more than 2 days to run on
384 5,000 SNV chunks and was abandoned.
385
386 *Variant-level QC*

387 Beagle provides two position level imputation metrics, allelic R-squared and dosage R-
388 squared. Both measures are highly correlated (Additional File 2: Figure S8.a). Values
389 between 0.3 and 0.8 are typically used for filtering [32]. In both 1x datasets 59% and 91% of
390 imputed variants lie below those two thresholds, respectively. The distribution of scores
391 does not provide an obvious filtering threshold (Additional File 2: Figure S8.b) due to its
392 concavity. Since most imputed variants are rare and R-squared measures are highly
393 correlated with MAF, filtering by AR2 and DR2 would be similar to imposing a MAF
394 threshold (Additional File 2: Figure S8.c and d.). Moreover, due to a technical limitation of
395 the vr-runner pipelines, imputation quality measures were not available for refined
396 positions at the time, only imputed ones. Therefore, we did not apply any prior filter in
397 downstream analyses, but used imputation metrics as well as variant quality scores to
398 prioritise variants post-association.
399
400 **Sample QC**
401 Due to the sparseness of the 1x datasets, sample-level QC was performed after imputation.
402 58 individuals were removed from the Pomak cohort due to contamination and sample
403 swap issues. 5 samples were excluded from the MANOLIS 1x cohort and 1 sample from the
404 4x cohort following PCA-based ethnicity checks.
405
406 **Comparison with WES**
407 A set of high confidence genotypes was generated for the 5 exomes in MANOLIS using filters
408 for variant quality (>200), call rate (AN=10, 100%) and depth (250). These filters were
409 derived from the respective distributions of quality metrics (Additional File 2: Figure S9).

410 When compared to 5 whole-exome sequences from each cohort, imputed 1x calls
411 recapitulated 77.2% of non-monomorphic, high-quality exome sequencing calls.
412 Concordance was high, with only 3.5% of the overlapping positions exhibiting some form of
413 allelic mismatch. When restricting the analysis to singletons, 9105 (58%) of the 15,626 high-
414 quality singletons in the 10 exomes were captured, with 21% of the captured positions
415 exhibiting false positive genotypes (AC>1). To assess false positive call rate, we extracted 1x
416 variants falling within the 71,627 regions targeted by the Agilent design file for WES in
417 overlapping samples, and compared them to those present in the unfiltered WES dataset.
418 103,717 variants were called in these regions from WES sequences, compared to 58,666
419 non-monomorphic positions in the 1x calls. 1,419 (2.4%) of these positions were unique to
420 the 1x dataset, indicating a low false-positive rate in exonic regions post-imputation.
421
422 **Genetic relatedness matrix**
423 In order to correct for genetics relatedness within the two isolated cohorts, we calculated a
424 genetic relatedness matrix using GEMMA [33]. Given the isolated nature of the population
425 and the specificities of the sequencing dataset, we used different variant sets to calculate
426 kinship coefficients. Using the unfiltered 1x variant dataset produced the lowest coefficients
427 (Figure 10.a), whereas well-behaved set of common SNVs [34] produced the highest, with
428 an average difference of 3.67×10^{-3}. Filtering for MAF lowered the inferred kinship
429 coefficients. Generally, the more a variant set was sparse and enriched in common variants,
430 the higher the coefficients were. However, these differences only had a marginal impact on
431 association statistics, as evidenced by a lambda median statistic difference of 0.02 between
432 the two most extreme estimates of relatedness when used for a genome-wide association
433 of triglycerides in Pomak (Additional File 2: Figure S9.b). For our association study, we used

434 LD-pruned 1x variants filtered for MAF<1% and Hardy Weinberg equilibrium $p<1x10^{-5}$ to
435 calculate the relatedness matrix.
436
437 **Phenotype preparation**
438 Twenty-five haematological phenotypes were prepared, with certain traits measured only in
439 one cohort or the other (Additional File 2: Table S4), and high levels of correlation for some
440 traits(Additional File 2: Figure S11). Full details of the trait transformation, filters and
441 exclusions are described in Additional File 2: Table S4. The 'transformPhenotype'
442 (https://github.com/gmelloni/transformPhenotype) R script was used to apply a
443 standardised preparation for all phenotypes. If gender differences were significant
444 (Wilcoxon rank sum $P < 0.05$), the phenotype was stratified accordingly. Following trait-
445 specific exclusions and adjustments, outliers were filtered out based on 3 standard
446 deviations (SD) away from the mean where necessary. Traits not normally distributed were
447 transformed to normality using an inverse normal transformation, after testing for a
448 number of power transformation including logarithmic. For all traits age and age^2 were
449 added as covariates as necessary and standardised residuals were used. If male and female
450 phenotypes were prepared separately these were standardised before combining the
451 residuals. Summary statistics for all of the traits are provided in Additional File 2: Table S5.
452
453 **Single-point association**
454 *Pipeline*
455 Association analysis was performed on each cohort separately using the linear mixed model
456 implemented in GEMMA [33] on all variants with minor allele count (MAC) greater than 2
457 (14,948,665 out of 30,483,158 variants in MANOLIS and 15,564,905 out of 29,740,281 for

458 Pomak). We used the aforementioned centered kinship matrix. GC-corrected p-values from
459 the likelihood ratio test (p_lrt) are reported. Singletons and doubletons are removed due to
460 overall low minor allele concordance.
461
462 *Estimating the significance threshold*
463 We determine the significance threshold by calculating $\alpha_{adj} = \frac{0.05}{N_{eff} \times k_{eff}}$, where N_{eff} is the
464 effective number of SNVs after correcting for LD and k_{eff} is the effective number of traits
465 tested after correcting for correlation. We estimated k_{eff} using two different methods. The
466 first method selects the number of principal components (PCs) in a principal component
467 analysis (PCA) of standardised, normalised traits that explain 95% of total trait variance. This
468 yielded $k_{eff} = 8$ for both MANOLIS and Pomak ($M = 20$ and $M = 18$, respectively). The
469 second method uses the Kaiser method on the eigenvalues of the trait correlation matrix to
470 calculate k_{eff} [35], and gives $k_{eff} = 9$ for MANOLIS and $k_{eff} = 8$ for Pomak.
471 For N_{eff}, we extrapolate the number of SNVs based on calibration curves[36] that provide
472 the number of independent SNVs given the total number of tested SNVs (assuming
473 MAF>0.5%). This gives $N_{eff} = 5{,}145{,}236$ for MANOLIS ($\alpha_{adj} = 1.08 \times 10^{-9}$) and $N_{eff} =$
474 $5{,}361{,}759$ for Pomak ($\alpha_{adj} = 1.16 \times 10^{-9}$). Performing LD-pruning using PLINK [37] yields
475 8,123,367 variants with MAC>2 for Pomak ($\alpha_{adj} = 7.7 \times 10^{-10}$) and 6,833,823 variants for
476 MANOLIS ($\alpha_{adj} = 8.12 \times 10^{-10}$). We define genome-wide significance at $\alpha_{adj} = 1.0 \times 10^{-9}$,
477 which is reasonably close to these estimates.
478
479 *Signal prioritisation*

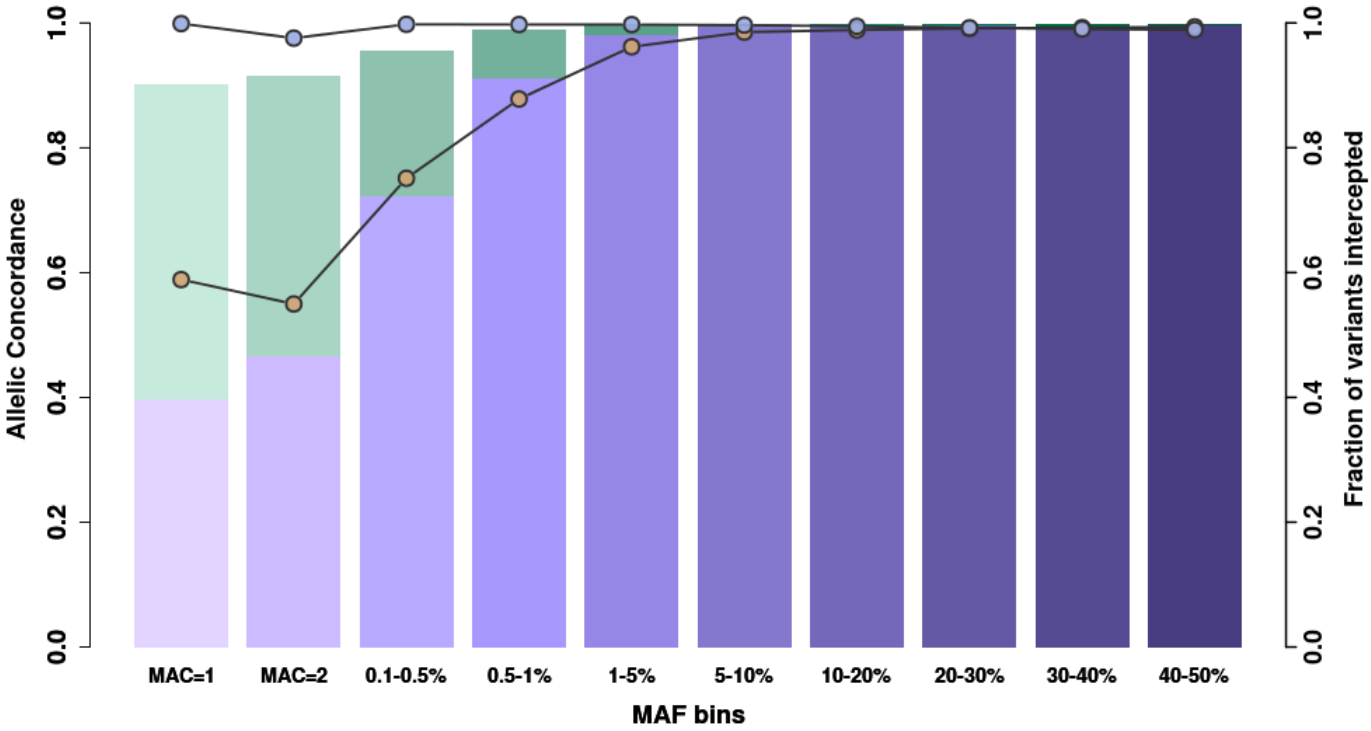
480 Signals were extracted using the peakplotter software (https://github.com/wtsi-
481 team144/peakplotter) using a window size of 1Mb.
482
483 **Replication**
484 The discovery and validation studies were conducted in different populations. This can
485 affect the strength of associations of genetic variants and lead to heterogeneity in effect
486 sizes [38, 39]. Therefore, we assessed heterogeneity and carried out a random effects meta-
487 analysis when there was evidence of heterogeneity at $p<0.05$. We estimated heterogeneity
488 using I2 and Q statistics. We used the method described in [13] for the random effects
489 meta-analysis, as it was shown to have higher power to detect associations than the
490 conventional random effects method.
491
492 ## Declarations
493
494 **Ethics approval and consent to participate**
495 The study was approved by the Harokopio University Bioethics Committee and informed
496 consent was obtained from every participant.
497
498 **Consent for publication**
499 Not applicable.
500
501 **Availability of data and materials**
502 The following HELIC genotype and WGS datasets have been deposited to the European
503 Genome-phenome Archive (https://www.ebi.ac.uk/ega/home): EGAD00010000518;

504 EGAD00010000522; EGAD00010000610; EGAD00001001636, EGAD00001001637. The
505 peakplotter software is available at https://github.com/wtsi-team144/peakplotter, the
506 transformPhenotype app can be downloaded at https://github.com/wtsi-
507 team144/transformPhenotype.
508
509
510 **Competing interests**
511 The authors declare that they have no competing interests.
512
513 **Funding**
514 This work was funded by the Wellcome Trust [098051] and the European Research Council
515 [ERC-2011-StG 280559-SEPI]. Funding for UK10K was provided by the Wellcome Trust under
516 award WT091310.
517
518 **Author's contributions**
519 AG performed variant call set quality control, evaluated and ran imputation pipelines,
520 performed single-point association with the help of KH, DS and XG and was a major
521 contributor in writing the manuscript. KK performed follow-up on the single-point signals
522 and contributed to writing the manuscript. LS analysed the genotype data and ran signal
523 discovery scripts. RM performed sample quality control on the Pomak dataset. GEMM
524 contributed to the transformPhenotypes tool. AEF under the supervision of GD and with the
525 help of ET and MK, coordinated the sample collection. GR performed bioinformatics
526 analyses on selected variants. JS performed quality control on the 4x sequences. PD
527 designed the runner pipelines used for imputation. BK maintained the phenotype database.

528 MP performed variant calling. HE, WJA, TJ, AB and NZ provided replication results in the UK
529 Biobank meta-analysis. EZ supervised the project. All authors read and approved the final
530 manuscript.
531
532 **Acknowledgements**
533 The authors thank the residents of the Mylopotamos and Pomak villages for taking part. The
534 MANOLIS study is dedicated to the memory of Manolis Giannakakis, 1978–2010. We also
535 thank the Exome Aggregation Consortium and the groups that provided exome variant data
536 for comparison. A full list of contributing groups can be found at
537 http://exac.broadinstitute.org/about. This study makes use of data generated by the UK10K
538 Consortium, derived from samples from UK10K_COHORTS_TWINSUK (The TwinsUK Cohort)
539 and UK10K_COHORT_ALSPAC (the Avon Longitudinal Study of Parents and Children). A full
540 list of the investigators who contributed to the generation of the data is available from
541 www.UK10K.org, last accessed April 29, 2016. The GATK3 program was made available
542 through the generosity of Medical and Population Genetics program at the Broad Institute,
543 Inc. This research has been conducted using the UK Biobank Resource using Application
544 Number 13745.
545

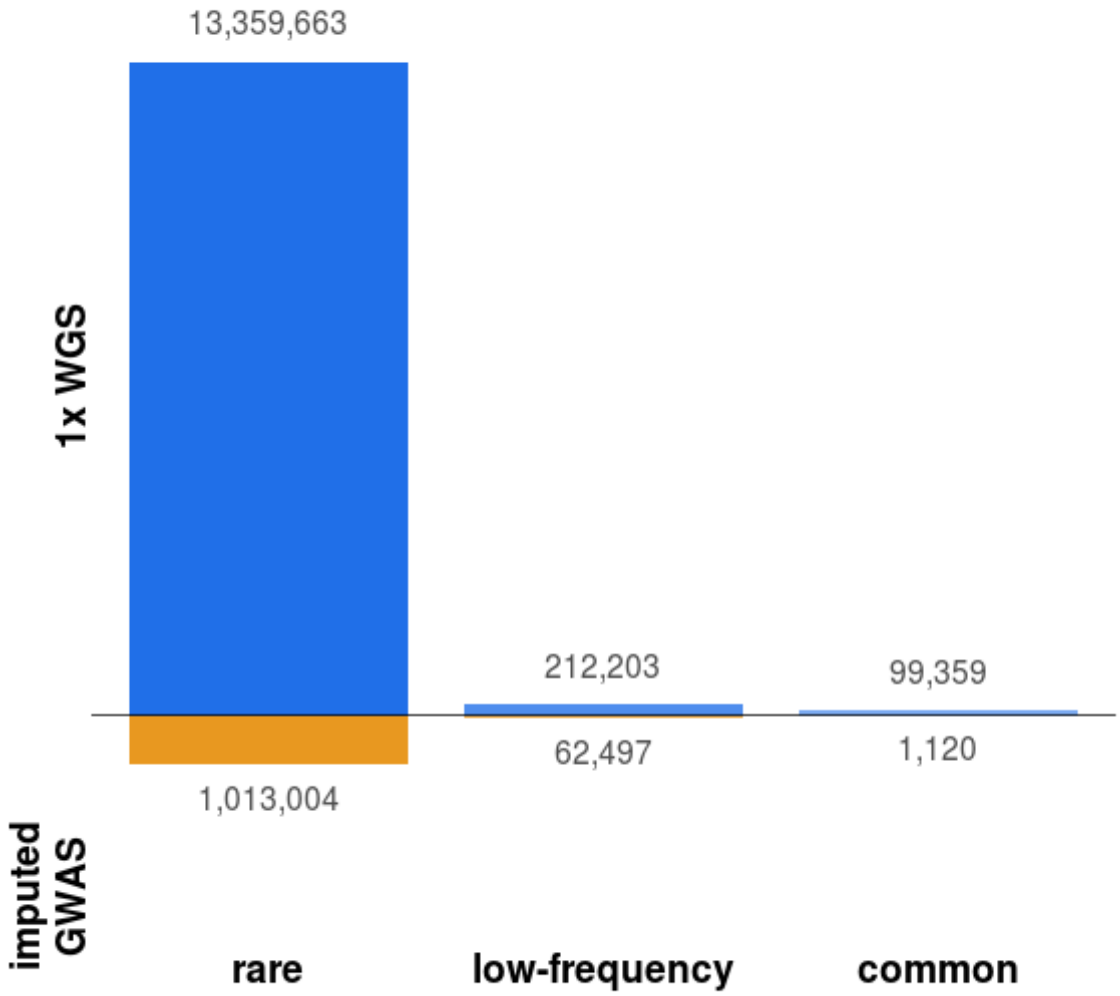
546 **Figures**

547 **Figure 1: Concordance and call rate for 1x WGS genotypes.** Genotype (blue circles) and
548 minor allele (yellow circles) concordance is computed for 1239 samples in MANOLIS against
549 merged OmniExpress and ExomeChip data. Call rate is assessed for the refined (purple) and
550 refined plus imputed (green) datasets.



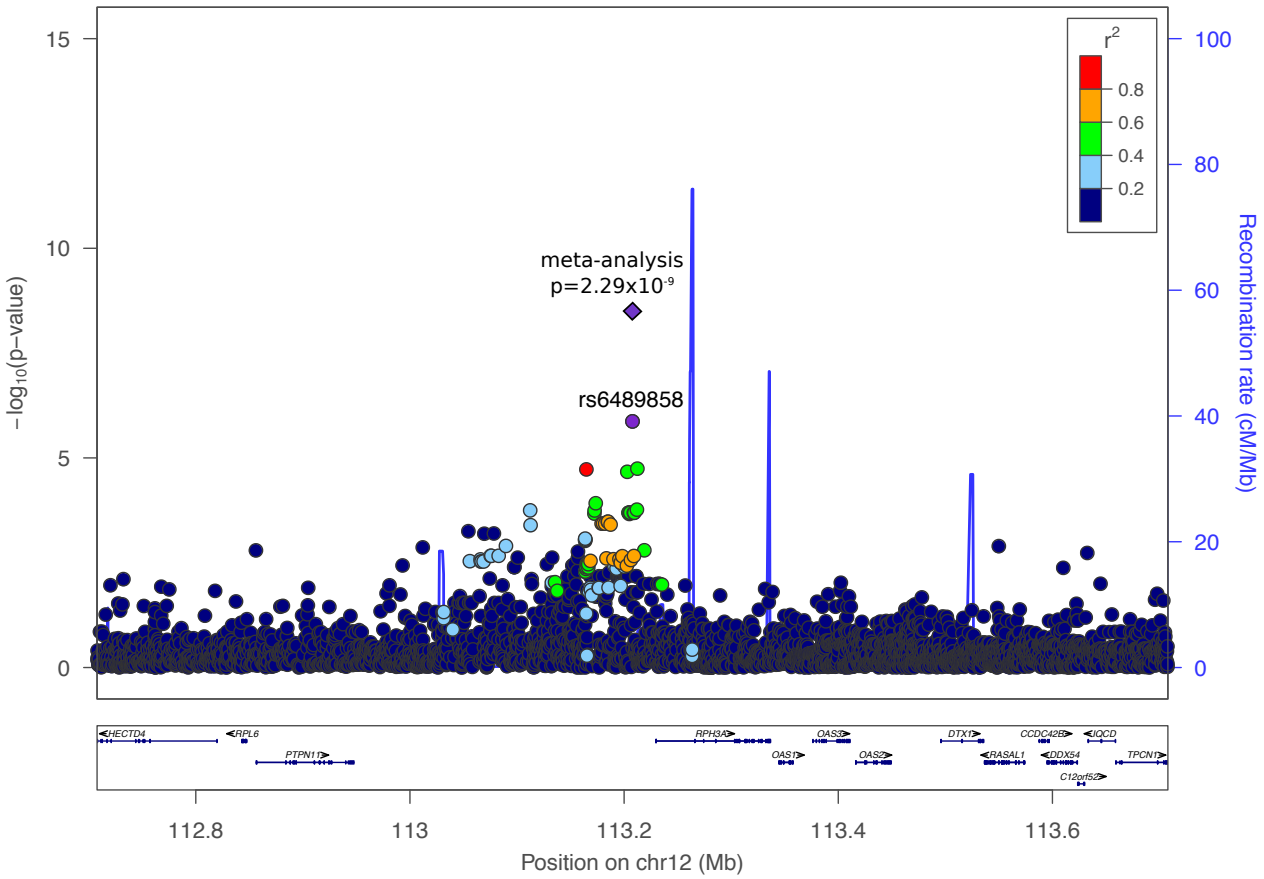
551

552 **Figure 2: Unique variants called by sequencing and imputed GWAS**. Variants unique to
553 either dataset, arranged by MAF bin. Both datasets are unfiltered apart from
554 monomorphics, which are excluded. MAF categories: rare (MAF<1%), low-frequency (MAF
555 1-5%), common (MAF>5%).



556

557 **Figure 3: Association between rs6489858 and lymphocyte count in MANOLIS.**



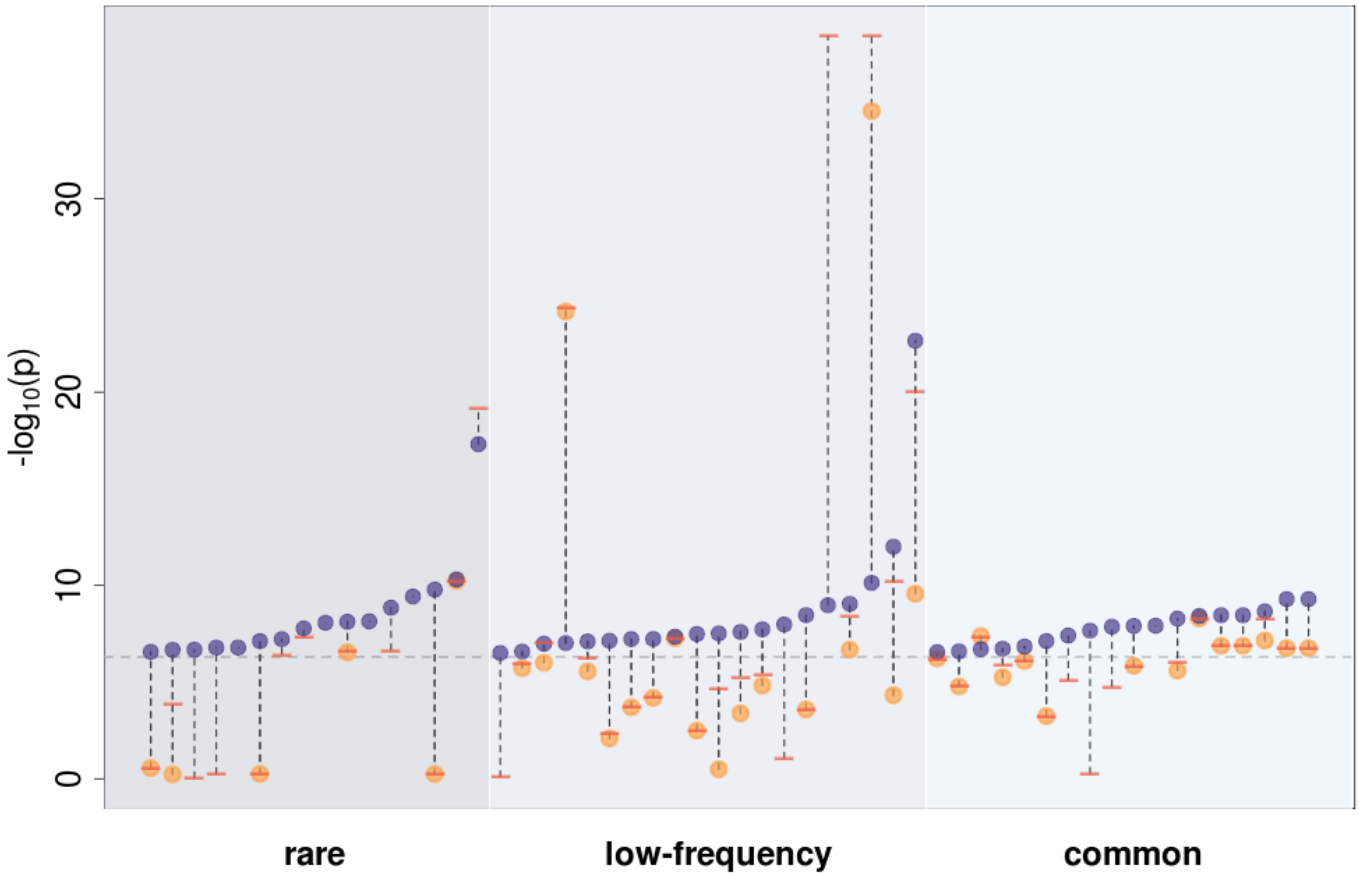
558 **Figure 4: Imputed GWAS results for association signals found in the 1x WGS at $p<5x10^{-7}$.**

559 Purple dots represent significant results in the 1x analysis. Orange dots, if present, denote

560 the p-value of the same SNP in the imputed GWAS study. Absence of a dot indicates the

561 variant was not found in the imputed GWAS dataset. Red dashes indicate the minimum p-

562 value among all tagging SNPs in the imputed GWAS (r^2>0.8).



563 ## Additional Files

File name	File Format	Title	Description
Additional File 1.xlsx	Excel (xlsx)	Supplementary Tables	Tables S1 to S5
Additional File 2.pdf	PDF	Supplementary Figures and Text	Figures S1 to S11, Supplementary Text

564

565 ## References

566 1. Le SQ, Durbin R: **SNP detection and genotyping from low-coverage sequencing data**
567 **on multiple diploid samples.** *Genome Res* 2011, **21:**952-960.
568 2. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel
569 JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human**
570 **genetic variation.** *Nature* 2015, **526:**68-74.
571 3. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM,
572 Fuchsberger C, Danecek P, Sharp K, et al: **A reference panel of 64,976 haplotypes for**
573 **genotype imputation.** *Nat Genet* 2016, **48:**1279-1283.
574 4. Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR,
575 Xu C, Futema M, et al: **The UK10K project identifies rare variants in health and**
576 **disease.** *Nature* 2015, **526:**82-90.
577 5. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-
578 Mckay F, Kostadima MA, et al: **The Allelic Landscape of Human Blood Cell Trait**
579 **Variation and Links to Common Complex Disease.** *Cell* 2016, **167:**1415-1429 e1419.
580 6. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM,
581 Daly MJ, Sklar P, et al: **Extremely low-coverage sequencing and imputation**
582 **increases power for genome-wide association studies.** *Nat Genet* 2012, **44:**631-635.

583 7. consortium C: **Sparse whole-genome sequencing identifies two loci for major**
584 **depressive disorder.** *Nature* 2015, **523:**588-591.
585 8. Hatzikotoulas K, Gilly A, Zeggini E: **Using population isolates in genetic association**
586 **studies.** *Brief Funct Genomics* 2014, **13:**371-377.
587 9. Southam L, Gilly A, Suveges D, Farmaki AE, Schwartzentruber J, Tachmazidou I,
588 Matchan A, Rayner NW, Tsafantakis E, Karaleftheri M, et al: **Whole genome**
589 **sequencing and imputation in two Greek isolated populations identifies**
590 **associations with complex traits of medical importance.** *Nat Comms* 2017, **in**
591 **review**.
592 10. Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, Mulas A, Pistis
593 G, Steri M, Danjou F, et al: **Genome sequencing elucidates Sardinian genetic**
594 **architecture and augments association analyses for lipid and blood inflammatory**
595 **markers.** *Nat Genet* 2015, **47:**1272-1281.
596 11. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A,
597 Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, et al: **Large-scale whole-**
598 **genome sequencing of the Icelandic population.** *Nat Genet* 2015, **47:**435-444.
599 12. Southam L, Gilly A, Suveges D, Farmaki AE, Schwartzentruber J, Tachmazidou I,
600 Matchan A, Rayner NW, Tsafantakis E, Karaleftheri M, et al: **Whole genome**
601 **sequencing and imputation in isolated populations identify genetic associations**
602 **with medically-relevant complex traits.** *Nat Comms* 2017.
603 13. Han B, Eskin E: **Random-effects model aimed at discovering associations in meta-**
604 **analysis of genome-wide association studies.** *Am J Hum Genet* 2011, **88:**586-598.
605 14. Browning BL, Browning SR: **Genotype Imputation with Millions of Reference**
606 **Samples.** *Am J Hum Genet* 2016, **98:**116-126.

607 15. Okada Y, Hirota T, Kamatani Y, Takahashi A, Ohmiya H, Kumasaka N, Higasa K,
608 Yamaguchi-Kabata Y, Hosono N, Nalls MA, et al: **Identification of nine novel loci**
609 **associated with white blood cell subtypes in a Japanese population.** *PLoS Genet*
610 2011, **7:**e1002067.
611 16. Nalls MA, Couper DJ, Tanaka T, van Rooij FJ, Chen MH, Smith AV, Toniolo D, Zakai
612 NA, Yang Q, Greinacher A, et al: **Multiple loci are associated with white blood cell**
613 **phenotypes.** *PLoS Genet* 2011, **7:**e1002113.
614 17. Lo KS, Wilson JG, Lange LA, Folsom AR, Galarneau G, Ganesh SK, Grant SF, Keating BJ,
615 McCarroll SA, Mohler ER, 3rd, et al: **Genetic association analysis highlights new loci**
616 **that modulate hematological trait variation in Caucasians and African Americans.**
617 *Hum Genet* 2011, **129:**307-317.
618 18. Loh ML, Sakai DS, Flotho C, Kang M, Fliegauf M, Archambeault S, Mullighan CG, Chen
619 L, Bergstraesser E, Bueso-Ramos CE, et al: **Mutations in CBL occur frequently in**
620 **juvenile myelomonocytic leukemia.** *Blood* 2009, **114:**1859-1863.
621 19. Chan G, Kalaitzidis D, Usenko T, Kutok JL, Yang W, Mohi MG, Neel BG: **Leukemogenic**
622 **Ptpn11 causes fatal myeloproliferative disorder via cell-autonomous effects on**
623 **multiple stages of hematopoiesis.** *Blood* 2009, **113:**4414-4424.
624 20. Izon DJ, Aster JC, He Y, Weng A, Karnell FG, Patriub V, Xu L, Bakkour S, Rodriguez C,
625 Allman D, Pear WS: **Deltex1 redirects lymphoid progenitors to the B cell lineage by**
626 **antagonizing Notch1.** *Immunity* 2002, **16:**231-243.
627 21. Hsiao HW, Liu WH, Wang CJ, Lo YH, Wu YH, Jiang ST, Lai MZ: **Deltex1 is a target of**
628 **the transcription factor NFAT that promotes T cell anergy.** *Immunity* 2009, **31:**72-
629 83.

630 22. Gilly A, Ritchie GR, Southam L, Farmaki AE, Tsafantakis E, Dedoussis G, Zeggini E:
631 **Very low-depth sequencing in a founder population identifies a cardioprotective**
632 **APOC3 signal missed by genome-wide imputation.** *Hum Mol Genet* 2016, **25:**2360-
633 2365.
634 23. Abnizova I, Skelly T, Naumenko F, Whiteford N, Brown C, Cox T: **Statistical**
635 **comparison of methods to estimate the error probability in short-read Illumina**
636 **sequencing.** *J Bioinform Comput Biol* 2010, **8:**579-591.
637 24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
638 Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a**
639 **MapReduce framework for analyzing next-generation DNA sequencing data.**
640 *Genome Res* 2010, **20:**1297-1303.
641 25. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del
642 Angel G, Rivas MA, Hanna M, et al: **A framework for variation discovery and**
643 **genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43:**491-
644 498.
645 26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
646 Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format**
647 **and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
648 27. Moayyeri A, Hammond CJ, Hart DJ, Spector TD: **The UK Adult Twin Registry**
649 **(TwinsUK Resource).** *Twin Res Hum Genet* 2013, **16:**144-149.
650 28. Golding J, Pembrey M, Jones R, Team AS: **ALSPAC--the Avon Longitudinal Study of**
651 **Parents and Children. I. Study methodology.** *Paediatr Perinat Epidemiol* 2001,
652 **15:**74-87.

653 29. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J: **Haplotype estimation using**
654 **sequencing reads.** *Am J Hum Genet* 2013, **93:**687-696.
655 30. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data**
656 **inference for whole-genome association studies by use of localized haplotype**
657 **clustering.** *Am J Hum Genet* 2007, **81:**1084-1097.
658 31. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR: **Low-coverage sequencing:**
659 **implications for design of complex trait association studies.** *Genome Res* 2011,
660 **21:**940-951.
661 32. Browning BL: **Private communication.** 2014.
662 33. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association**
663 **studies.** *Nat Genet* 2012, **44:**821-824.
664 34. Arthur R, Schulz-Trieglaff O, Cox AJ, O'Connell J: **AKT: ancestry and kinship toolkit.**
665 *Bioinformatics* 2017, **33:**142-144.
666 35. Li MX, Yeung JM, Cherny SS, Sham PC: **Evaluating the effective numbers of**
667 **independent tests and significant p-value thresholds in commercial genotyping**
668 **arrays and public imputation reference datasets.** *Hum Genet* 2012, **131:**747-756.
669 36. Xu C, Tachmazidou I, Walter K, Ciampi A, Zeggini E, Greenwood CM, Consortium UK:
670 **Estimating genome-wide significance for whole-genome sequencing studies.** *Genet*
671 *Epidemiol* 2014, **38:**281-290.
672 37. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation**
673 **PLINK: rising to the challenge of larger and richer datasets.** *Gigascience* 2015, **4:**7.
674 38. Tang H: **Confronting ethnicity-specific disease risk.** *Nat Genet* 2006, **38:**13-15.
675 39. Barroso I, Luan J, Wheeler E, Whittaker P, Wasson J, Zeggini E, Weedon MN, Hunt S,
676 Venkatesh R, Frayling TM, et al: **Population-specific risk of type 2 diabetes**

677 **conferred by HNF4A P2 promoter variants: a lesson for replication studies.**
678 *Diabetes* 2008, **57:**3161-3165.
679
680