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 proposed as an efficient way to further improve the cost efficiency of sequencing-based association studies. Simulations have shown that in whole-exome studies, 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 [6]. 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 [7].

Studying founder populations can further empower the search for association signals by allowing the detection of population-specific variants, and of association signals at variants that have drifted up in frequency compared to cosmopolitan populations, against the backdrop of a homogeneous environment [8] [9] [10] [11]. Here, we perform very low depth (1x), cohort-wide WGS in two isolated populations from Greece. We establish a robust 1x WGS calling pipeline and compare the complement of variants captured to imputed GWAS in the same samples. As a proof-of-principle, we perform association analysis across medically-relevant haematological traits and identify a robustly-replicating novel locus implicated in lymphocyte counts.

Results

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

Imputation-based genotype refinement was performed on the two cohort-wide datasets using a combined reference panel of 10,422 haplotypes from MANOLIS 4x WGS, the 1000 Genomes [2] and UK10K [4] projects.

Variant calling pipeline

We established a variant calling, quality control (QC) and genotype refinement pipeline for very low depth WGS by benchmarking nine pipelines that make use of state of the art bioinformatics tools (Methods). Our optimised approach allowed the capture of 80% of true low-frequency (MAF 1-5%) variants and 100% of true common-frequency (MAF>5%) SNVs prior to imputation-based refinement, when compared to variants present on the Illumina OmniExpress and HumanExome chips genotyped in the same samples. In order to assess sensitivity and specificity of SNV calls pre-imputation, we estimate the false positive and false negative rate by comparing 1x WGS variant calls with high-depth WGS data (see Methods). We estimate that 12% of 1x sites with at least one heterozygote call are false positives, whereas 24.6% of the sites called with high-depth WGS are not recapitulated using 1x data.

In order to improve the false negative rate and genotype accuracy, we performed genotype refinement and imputation using a large reference panel containing haplotypes from 4,873 cosmopolitan samples as well as the phased haplotypes from the 249 MANOLIS samples sequenced at 4x depth. After imputation and QC, we captured 95% of rare, 99.7% of low-frequency and 99.9% of common variants compared to the Illumina OmniExpress and HumanExome GWAS chips, with an average minor allele concordance of 97% across the allele frequency spectrum (Methods and Figure 1). By comparing 1x calls with those produced by whole-exome sequencing in 10 individuals across both cohorts, we estimate a false-positive rate of 2.4% post-imputation in the coding parts of the genome (Methods).

Comparison of variant call sets with an imputed GWAS

The genotype refinement and imputation step yielded 30,483,136 and 29,740,259 non-monomorphic SNVs in 1,239 MANOLIS and 1,108 Pomak individuals, respectively. The number of variants discovered using 1x WGS is nearly twice as high as that from array-based approaches. In a subset of 982 MANOLIS individuals with 1x WGS, we called 25,673,116 non-monomorphic SNVs using 1x WGS data, compared to 13,078,518 non-monomorphic SNVs in the same samples with OmniExpress and ExomeChip data imputed up to the same panel [9]. The main differences are among rare variants (MAF<1%) (Figure 2): 13,671,225 (53.2%) variants called in the refined 1x WGS are absent from the imputed GWAS, 98% of which are rare. 82% of these rare unique SNVs are singletons or doubletons, and therefore 9.5% of all variants called in the 1x WGS dataset were unique variants with MAC>2.

Experimental validation of genotypes

We performed experimental genotyping of 65 variants (23 common, 18 low-frequency and 24 rare) in a subset of 1087 and 859 samples in the MANOLIS and Pomak cohorts, respectively, using the Agena Biosciences MassARRAY technology. On average, minor allele concordance was 76% and positive predictive value was 82%. As expected, these values differ between MAF categories (Additional File 1: Table S1). Minor allele and genotype concordance between 1x calls and this set of directly assayed genotypes were in line with those computed genome-wide between 1x calls and GWAS data (Figure 1).

Association analysis

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

Genome-wide significance was set at $p<1.0x10^{-9}$ based on the effective number of traits and tested variants (see Methods). In the discovery sample, one association met this threshold in the Pomak dataset. rs35004220, located in an intron of the haemoglobin B (*HBB*) gene, was associated with six red blood cell traits (haemoglobin, mean corpuscular haemoglobin, mean corpuscular volume, red cell distribution width in volume and percent, red blood cells) (Additional File 1 and 2: Figure S1 and Table S2). A total of 5,090 variants with association $p<1.0x10^{-5}$ in the discovery stage corresponding to 556 and 465 independent signals in the MANOLIS and Pomak cohorts, respectively (Additional File 1: Table S2 and S3), were carried forward to *in silico* replication using data from a large meta-analysis of 173,480 samples from the UK Biobank and INTERVAL studies [5]. Out of the 3,336 variants for which replication data were available, 52.4% had a concordant direction of effect compared to the discovery stage ($p=2.9x10^{-3}$, one-sided binomial test). Upon meta-analysis of the discovery and replication data, we identify a previously unreported, genome-wide significantly associated signal (Figure 3). The G-allele of rs6489858 (EAF=0.40) at 12q24.13 is associated with increased lymphocyte count in MANOLIS (beta=0.022, SE=0.004, $p=2.29x10^{-9}$ in the discovery and replication meta-analysis). We found evidence of heterogeneity ($I^2=95.4\%$, Q-statistic $p=2x10^{-6}$) and therefore applied a random effects meta-analysis model [13] (beta=0.13, SE=0.11, p=8x10^{-12}). rs6489858 is located in an intronic region of the rabphilin 3A (*RPH3A*) gene, which encodes a peripheral membrane protein involved in protein transport and synaptic vesicle traffic.

Comparison of association summary statistics with imputed GWAS

1x WGS calls a larger number of variants than imputed GWAS of the same samples. To evaluate how this difference affects association study power, we compared the association results of all independent suggestive signals at p<5x10^{-7} from the 1x WGS with the imputed GWAS results for the same variants (Figure 4). Among the 54 variants significantly associated at this threshold in the 1x WGS, 17 (31%) were not observed in the imputed GWAS study. Rare (MAF<1%) variant signals are more poorly captured by the imputed GWAS, with 10 out of 16 signals (62.5%) being missed, however, for 12 (70%) of the missed variants, a tagging SNV at r^2>0.8 was available in the imputed GWAS. For the signals where imputed GWAS results are indeed present, the majority (62.2%) do not meet our significance threshold, an effect which is more marked in the rare and low-frequency (16/23, 69%) than in the common (7/14, 50%) signals. This observation persists when considering tagging variants at r^2>0.8: twenty-seven (55%) out of the 49 taggable 1x WGS signals have a p-value above the significance threshold. Generally, differences in p-values between the two studies were smaller for common than for low-frequency and rare variant signals (5.5 and 2.0 on the log-scale, p=6x10^{-3}, two-sample t-test).

Discussion

In this work, we empirically demonstrate the relative merits of very low depth WGS both in terms of variant discovery and association study power for complex quantitative traits compared to GWAS approaches. However, the advantages of 1x WGS have to be weighed against compute and financial cost considerations. As of January 2017, 1x WGS on the HiSeq 4000 platform was approximately half of the cost of a dense GWAS array (e.g. Illumina Infinium Omni 2.5Exome-8 array), 1.5 times the cost of a sparser chip such as the Illumina HumanCoreExome array, and a third of the cost of WES at 50x depth. By comparison, 30x WGS was 21 or 16 times more costly depending on the sequencing platform (Illumina HiSeq 4000 or HiSeqX, respectively). The number of variants called by 1x WGS is lower than high-depth WGS, but is in the same order of magnitude, suggesting comparable disk storage requirements for variant calls. However, storage of the reads required an average 650Mb per sample for CRAMs, and 1.3Gb per sample for BAMs.

Genome-wide refinement and imputation of very low depth WGS generates close to 50 times more variants than a GWAS chip. The complexity of the imputation and phasing algorithms used in this study is linear in the number of markers, linear in the number of target samples and quadratic in the number of reference samples [14], which results in a 50-fold increase in total processing time compared to an imputed GWAS study of equal sample size. Therefore, parallelisation plays a crucial role in managing computational load. For example, in MANOLIS the genome was divided in 13,276 chunks containing equal number of SNVs, which took an average of 31 hours each to refine and impute. The total processing time was 47 core-years (Methods and Additional File 2: Figure S2). Parallelisation allowed processing the 1,239 MANOLIS samples in under a month.

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

Conclusions

We show that very low depth whole-genome sequencing allows the accurate assessment of most common and low-frequency variants captured by imputed GWAS designs and achieves denser coverage of the low-frequency and rare end of the allelic spectrum, albeit at an increased computational cost. This allows very low depth sequencing studies to identify signals also discoverable by imputed chip-based efforts, and to discover significantly associated variants missed by GWAS imputation [22]. As sequencing technologies continue to evolve, higher sequencing depths will provide accurate genotyping across the full range of the allelic spectrum, enabling comprehensive exploration of human phenotype associations through rare variant aggregation tests.

Materials and methods

Cohort details

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

Sequencing

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

HiSeq 2500 sequencers. For comparison, 5 samples from each cohort were whole-exome sequenced at an average depth of 75x.

Read mapping and variant calling

Following generation of raw reads on the Illumina HiSeq 2000 and HiSeq 2500 sequencing machines, reads were converted from BCL format to BAM format using the Illumina2BAM (https://github.com/wtsi-npg/illumina2bam) software. Illumina2BAM was again used to de-multiplex lanes that had been sequenced so that the tags were isolated from the body of the read, decoded, and could be used to separate out each lane into lanelets containing individual samples from the multiplex library and the PhiX control. The quality scores were then recalibrated using the purity recalibration algorithm [23] using the PhiX data for reference. Read mapping was then carried out using the BWA backtrack algorithm version 0.5.10 using the GRCh37 1000 Genomes phase III reference (also known as hs37d5). PCR and optically duplicated reads were marked using Picard MarkDuplicates (http://broadinstitute.github.io/picard).

In order to ensure the quality of the large quantity of BAMs produced for the project, an automatic quality control system was used to reduce the number of data files that required manual intervention. This system was derived from the one originally designed for the UK10K project (http://www.uk10k.org) and used a series of empirically derived thresholds to assess summary metrics calculated from the input BAMs. These thresholds included: percentage of reads mapped; percentage of duplicate reads marked; various statistics measuring INDEL distribution against read cycle and an insert size overlap percentage. Any lane that fell below the “fail” threshold for any of the metrics were excluded; and any lane that did not fall below these thresholds for any of the metrics was given a status of “pass” and allowed to proceed into the later stages of the pipeline.

Passed lanelets were then merged into BAMs corresponding to the libraries for each sample and duplicates were marked again with Picard MarkDuplicates after which they were then merged into BAMs on a per sample basis. Finally sample level bam improvement was carried out using GATK 1.6[24, 25] and samtools[26] from git commit 72d6457f7f361c323f62bd2d3170980132ba2113. This consisted of re-alignment of reads around known and discovered INDELs followed by base quality score recalibration both using the GATK, lastly samtools calmd was applied and indexes were created. Known INDELs for realignment were taken from Mill Devine and 1000G Gold set and the 1000G phase low coverage set both part of the Broad’s GATK resource bundle version 2.2. Known variants for BQSR were taken from dbSNP 137 also part of the Broad’s resource bundle.

The input BAM files are fed into samtools mpileup to create all-sites BCF files, which are piped into bcftools view to create variant-only VCF files containing genotype calls. We split the genome into chunks of 100,000 base pairs, and separate these chunks into SNV and INDEL files. We run GATK UnifiedGenotyper to calculate site-level annotations.

Variant filtering

Variant quality score recalibration was performed using GATK VQSR v.3.1.1. However, using the default parameters for the VQSR mixture model yields poor filtering, with a Ti/Tv ratio dropoff at 83% percent sensitivity and a Ti/Tv ratio of 1.8 for high-quality tranches (Additional File 2: Figure S3.a). We therefore ran exploratory runs of VQSR across a range of values for the model parameters, using the dropoff point of the transition/transversion (Ti/Tv) ratio below 2.0 as an indicator of good fit (Additional File 2: Figure S4). A small number of configurations outperformed all others, which allowed us to select an optimal set of parameters. For the chosen set of parameters, false positive rate is estimated at 10%±5% (Additional File 2: Figure S3.b). Indels were excluded from the dataset out of concerns for genotype quality. We found that the version of VQSR, as well as the annotations used to train the model, had a strong influence on the quality of the recalibration (Additional File 2: Figure S4 and Supplementary Text).

Comparison with Platinum genomes

For quality control purposes, reads from 17 of the well-characterised Platinum Genomes sequenced by Illumina at 50x depth [3], and downsampled to 1x depth using samtools [26] were included in the merged BAM file. VQSR-filtered calls were then compared to the high-confidence call sets made available by Illumina for those samples. 524,331 of the 4,348,092 non-monomorphic variant sites were not present in the high-confidence calls, whereas 1,246,403 of the 5,070,164 non-monomorphic high-confidence were not recapitulated in the 1x data. This corresponds to an estimated false positive rate of 12% and false negative rate of 24.6%. Both unique sets had a much higher proportion of singletons (corresponding to MAF < 2.9%) than the entire sets (57.9% vs 19.9% of singletons among 1x calls and 51% vs 18.1% among high-confidence calls), which suggests that a large fraction of the erroneous sites lies in the low-frequency and rare part of the allelic spectrum. However, genotype accuracy is poor, to the point where it obscures peculiarities in the distribution of allele counts (Additional File 2: Figure S5). Due to them being present in the 1000 genomes reference panel, we remove the 17 Platinum Genomes prior to imputation.

Genotype refinement and imputation

Reference Panel

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

Pipeline

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

Evaluation of pipelines

The authors of SHAPEIT [29] advise to phase whole chromosome when performing pre-phasing in order to preserve downstream imputation quality. This approach is computationally intractable for the 1x datasets, where the smallest chromosomes contain almost 7 times more variants than the largest chromosomes in a GWAS dataset.

For benchmarking purposes, we tested 13 genotype refinement pipelines involving Beagle v4.0 [30] and SHAPEIT2 [29] using a 1000 Genomes phase 1 reference panel, which we evaluated against minor allele concordance. All pipelines were run using the vr-runner scripts (https://github.com/VertebrateResequencing/vr-runner). Pipelines involving Beagle with the use of a reference panel ranked consistently better (Additional File 2: Figure S6), with a single run of reference-based refinement using Beagle outperforming all other runs. IMPUTE2 performed worst on its own, whether with or without reference panel; in fact the addition of a reference panel did not improve genotype quality massively. Phasing with Beagle without an imputation panel improved genotype quality, before or after IMPUTE2.

Halving the number of SNVs per refinement chunk to 2,000 (including 500 flanking positions) resulted in only a modest loss of genotype quality in the rare part of the allelic spectrum (Additional File 2: Figure S7), while allowing for a twofold increase in refinement speed. Genotype quality dropped noticeably for rare variants when imputation was turned on (Additional File 2: Figure S7), but remained high for low-frequency and common ones. A reference-free run of Beagle allowed to phase all positions and remove genotype missingness with no major impact on quality and a low computational cost. We also tested thunderVCF [31] for phasing sites, however, the program took more than 2 days to run on 5,000 SNV chunks and was abandoned.

Variant-level QC

Beagle provides two position level imputation metrics, allelic R-squared and dosage R-squared. Both measures are highly correlated (Additional File 2: Figure S8.a). Values between 0.3 and 0.8 are typically used for filtering [32]. In both 1x datasets 59% and 91% of imputed variants lie below those two thresholds, respectively. The distribution of scores does not provide an obvious filtering threshold (Additional File 2: Figure S8.b) due to its concavity. Since most imputed variants are rare and R-squared measures are highly correlated with MAF, filtering by AR2 and DR2 would be similar to imposing a MAF threshold (Additional File 2: Figure S8.c and d.). Moreover, due to a technical limitation of the vr-runner pipelines, imputation quality measures were not available for refined positions at the time, only imputed ones. Therefore, we did not apply any prior filter in downstream analyses, but used imputation metrics as well as variant quality scores to prioritise variants post-association.

Sample QC

Due to the sparseness of the 1x datasets, sample-level QC was performed after imputation. 58 individuals were removed from the Pomak cohort due to contamination and sample swap issues. 5 samples were excluded from the MANOLIS 1x cohort and 1 sample from the 4x cohort following PCA-based ethnicity checks.

Comparison with WES

A set of high confidence genotypes was generated for the 5 exomes in MANOLIS using filters for variant quality (>200), call rate (AN=10, 100%) and depth (250). These filters were derived from the respective distributions of quality metrics (Additional File 2: Figure S9).

When compared to 5 whole-exome sequences from each cohort, imputed 1x calls recapitulated 77.2% of non-monomorphic, high-quality exome sequencing calls. Concordance was high, with only 3.5% of the overlapping positions exhibiting some form of allelic mismatch. When restricting the analysis to singletons, 9105 (58%) of the 15,626 high-quality singletons in the 10 exomes were captured, with 21% of the captured positions exhibiting false positive genotypes (AC>1). To assess false positive call rate, we extracted 1x variants falling within the 71,627 regions targeted by the Agilent design file for WES in overlapping samples, and compared them to those present in the unfiltered WES dataset. 103,717 variants were called in these regions from WES sequences, compared to 58,666 non-monomorphic positions in the 1x calls. 1,419 (2.4%) of these positions were unique to the 1x dataset, indicating a low false-positive rate in exonic regions post-imputation.

Genetic relatedness matrix

In order to correct for genetics relatedness within the two isolated cohorts, we calculated a genetic relatedness matrix using GEMMA [33]. Given the isolated nature of the population and the specificities of the sequencing dataset, we used different variant sets to calculate kinship coefficients. Using the unfiltered 1x variant dataset produced the lowest coefficients (Figure 10.a), whereas well-behaved set of common SNVs [34] produced the highest, with an average difference of 3.67×10^{-3}. Filtering for MAF lowered the inferred kinship coefficients. Generally, the more a variant set was sparse and enriched in common variants, the higher the coefficients were. However, these differences only had a marginal impact on association statistics, as evidenced by a lambda median statistic difference of 0.02 between the two most extreme estimates of relatedness when used for a genome-wide association of triglycerides in Pomak (Additional File 2: Figure S9.b). For our association study, we used

LD-pruned 1x variants filtered for MAF<1% and Hardy Weinberg equilibrium p<1x10^{-5} to calculate the relatedness matrix.

Phenotype preparation

Twenty-five haematological phenotypes were prepared, with certain traits measured only in one cohort or the other (Additional File 2: Table S4), and high levels of correlation for some traits(Additional File 2: Figure S11). Full details of the trait transformation, filters and exclusions are described in Additional File 2: Table S4. The 'transformPhenotype' (https://github.com/gmelloni/transformPhenotype) R script was used to apply a standardised preparation for all phenotypes. If gender differences were significant (Wilcoxon rank sum P < 0.05), the phenotype was stratified accordingly. Following trait-specific exclusions and adjustments, outliers were filtered out based on 3 standard deviations (SD) away from the mean where necessary. Traits not normally distributed were transformed to normality using an inverse normal transformation, after testing for a number of power transformation including logarithmic. For all traits age and age^2 were added as covariates as necessary and standardised residuals were used. If male and female phenotypes were prepared separately these were standardised before combining the residuals. Summary statistics for all of the traits are provided in Additional File 2: Table S5.

Single-point association

Pipeline

Association analysis was performed on each cohort separately using the linear mixed model implemented in GEMMA [33] on all variants with minor allele count (MAC) greater than 2 (14,948,665 out of 30,483,158 variants in MANOLIS and 15,564,905 out of 29,740,281 for

Pomak). We used the aforementioned centered kinship matrix. GC-corrected p-values from the likelihood ratio test (p_lrt) are reported. Singletons and doubletons are removed due to overall low minor allele concordance.

Estimating the significance threshold

We determine the significance threshold by calculating $\alpha_{adj} = \frac{0.05}{N_{eff} \times k_{eff}}$, where N_{eff} is the effective number of SNVs after correcting for LD and k_{eff} is the effective number of traits tested after correcting for correlation. We estimated k_{eff} using two different methods. The first method selects the number of principal components (PCs) in a principal component analysis (PCA) of standardised, normalised traits that explain 95% of total trait variance. This yielded $k_{eff} = 8$ for both MANOLIS and Pomak ($M = 20$ and $M = 18$, respectively). The second method uses the Kaiser method on the eigenvalues of the trait correlation matrix to calculate k_{eff} [35], and gives $k_{eff} = 9$ for MANOLIS and $k_{eff} = 8$ for Pomak. For N_{eff}, we extrapolate the number of SNVs based on calibration curves[36] that provide the number of independent SNVs given the total number of tested SNVs (assuming MAF>0.5%). This gives $N_{eff} = 5{,}145{,}236$ for MANOLIS ($\alpha_{adj} = 1.08 \times 10^{-9}$) and $N_{eff} = 5{,}361{,}759$ for Pomak ($\alpha_{adj} = 1.16 \times 10^{-9}$). Performing LD-pruning using PLINK [37] yields 8,123,367 variants with MAC>2 for Pomak ($\alpha_{adj} = 7.7 \times 10^{-10}$) and 6,833,823 variants for MANOLIS ($\alpha_{adj} = 8.12 \times 10^{-10}$). We define genome-wide significance at $\alpha_{adj} = 1.0 \times 10^{-9}$, which is reasonably close to these estimates.

Signal prioritisation

Signals were extracted using the peakplotter software (https://github.com/wtsi-team144/peakplotter) using a window size of 1Mb.

Replication

The discovery and validation studies were conducted in different populations. This can affect the strength of associations of genetic variants and lead to heterogeneity in effect sizes [38, 39]. Therefore, we assessed heterogeneity and carried out a random effects meta-analysis when there was evidence of heterogeneity at p<0.05. We estimated heterogeneity using I2 and Q statistics. We used the method described in [13] for the random effects meta-analysis, as it was shown to have higher power to detect associations than the conventional random effects method.

Declarations

Ethics approval and consent to participate

The study was approved by the Harokopio University Bioethics Committee and informed consent was obtained from every participant.

Consent for publication

Not applicable.

Availability of data and materials

The following HELIC genotype and WGS datasets have been deposited to the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/home): EGAD00010000518;

EGAD00010000522; EGAD00010000610; EGAD00001001636, EGAD00001001637. The peakplotter software is available at https://github.com/wtsi-team144/peakplotter, the transformPhenotype app can be downloaded at https://github.com/wtsi-team144/transformPhenotype.

Competing interests

The authors declare that they have no competing interests.

Funding

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.

Author's contributions

AG performed variant call set quality control, evaluated and ran imputation pipelines, performed single-point association with the help of KH, DS and XG and was a major contributor in writing the manuscript. KK performed follow-up on the single-point signals and contributed to writing the manuscript. LS analysed the genotype data and ran signal discovery scripts. RM performed sample quality control on the Pomak dataset. GEMM contributed to the transformPhenotypes tool. AEF under the supervision of GD and with the help of ET and MK, coordinated the sample collection. GR performed bioinformatics analyses on selected variants. JS performed quality control on the 4x sequences. PD designed the runner pipelines used for imputation. BK maintained the phenotype database.

MP performed variant calling. HE, WJA, TJ, AB and NZ provided replication results in the UK Biobank meta-analysis. EZ supervised the project. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the residents of the Mylopotamos and Pomak villages for taking part. The MANOLIS study is dedicated to the memory of Manolis Giannakakis, 1978–2010. We also thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about. This study makes use of data generated by the UK10K Consortium, derived from samples from UK10K_COHORTS_TWINSUK (The TwinsUK Cohort) and UK10K_COHORT_ALSPAC (the Avon Longitudinal Study of Parents and Children). 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 GATK3 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.

Figures

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

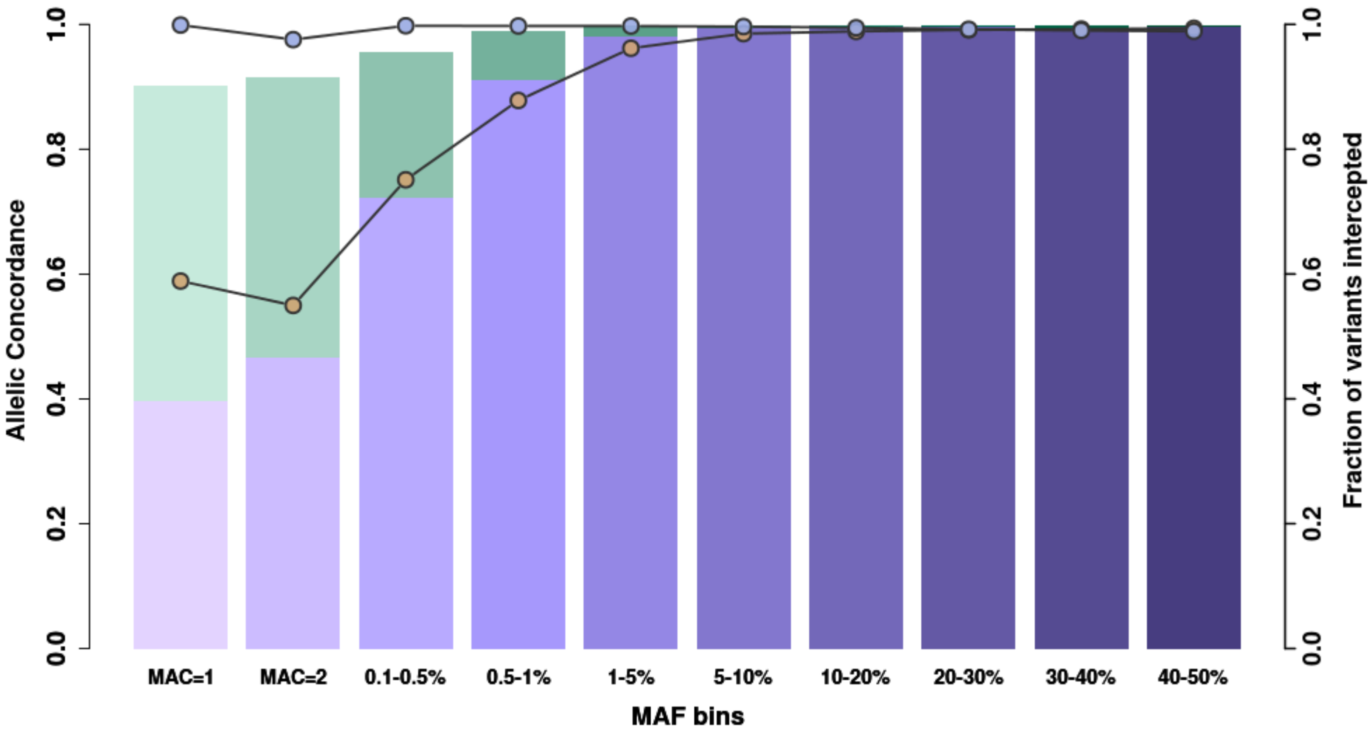
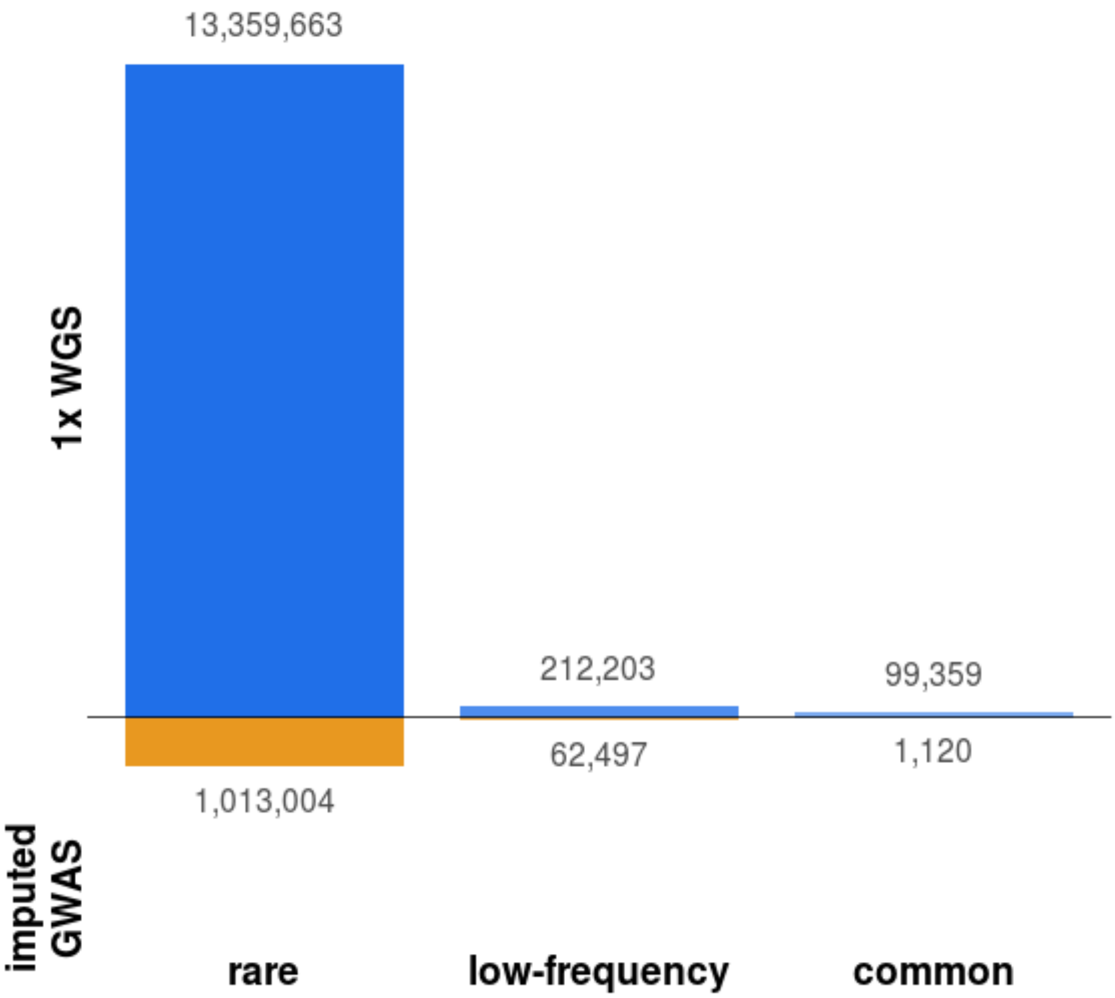


Figure 2: Unique variants called by sequencing and imputed GWAS. Variants unique to either dataset, arranged by MAF bin. Both datasets are unfiltered apart from monomorphics, which are excluded. MAF categories: rare (MAF<1%), low-frequency (MAF 1-5%), common (MAF>5%).



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

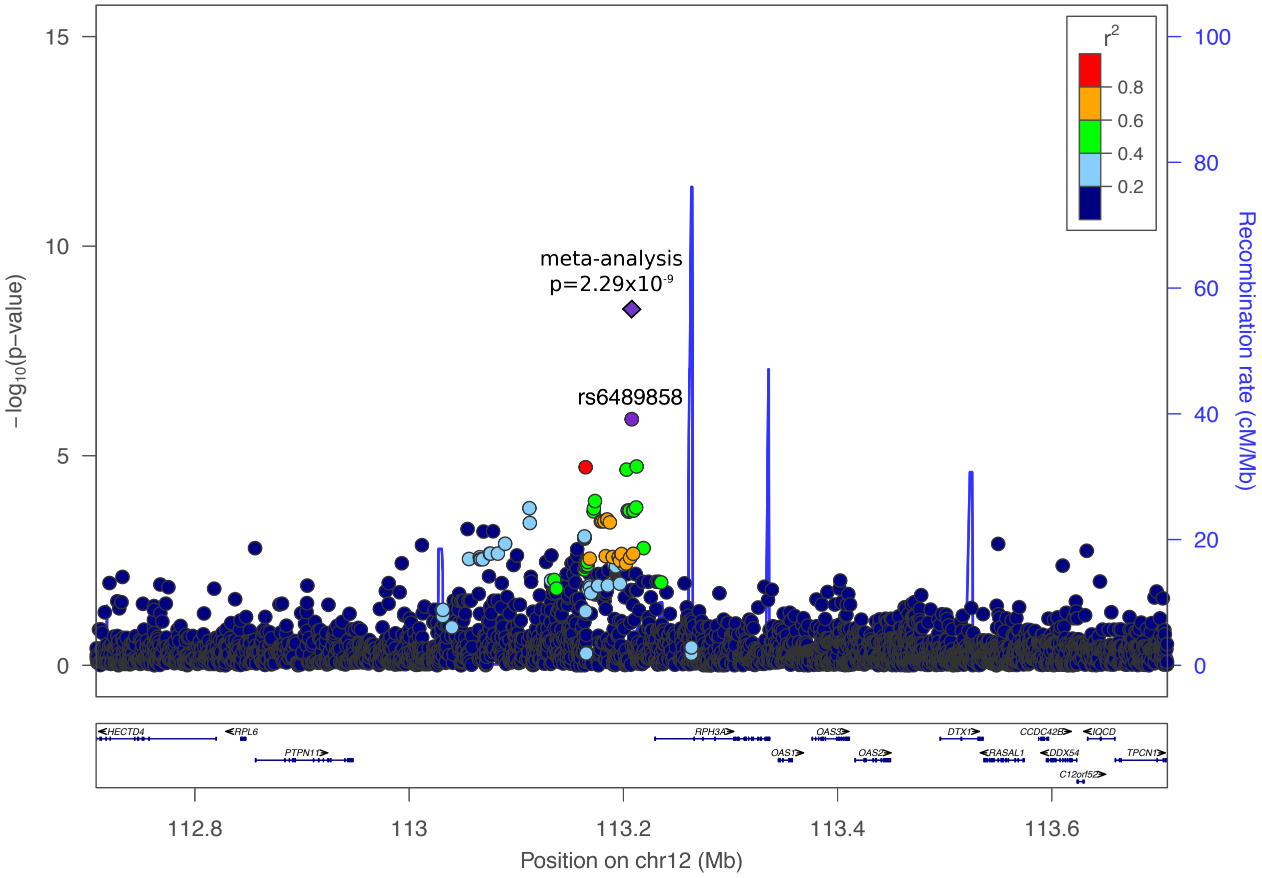
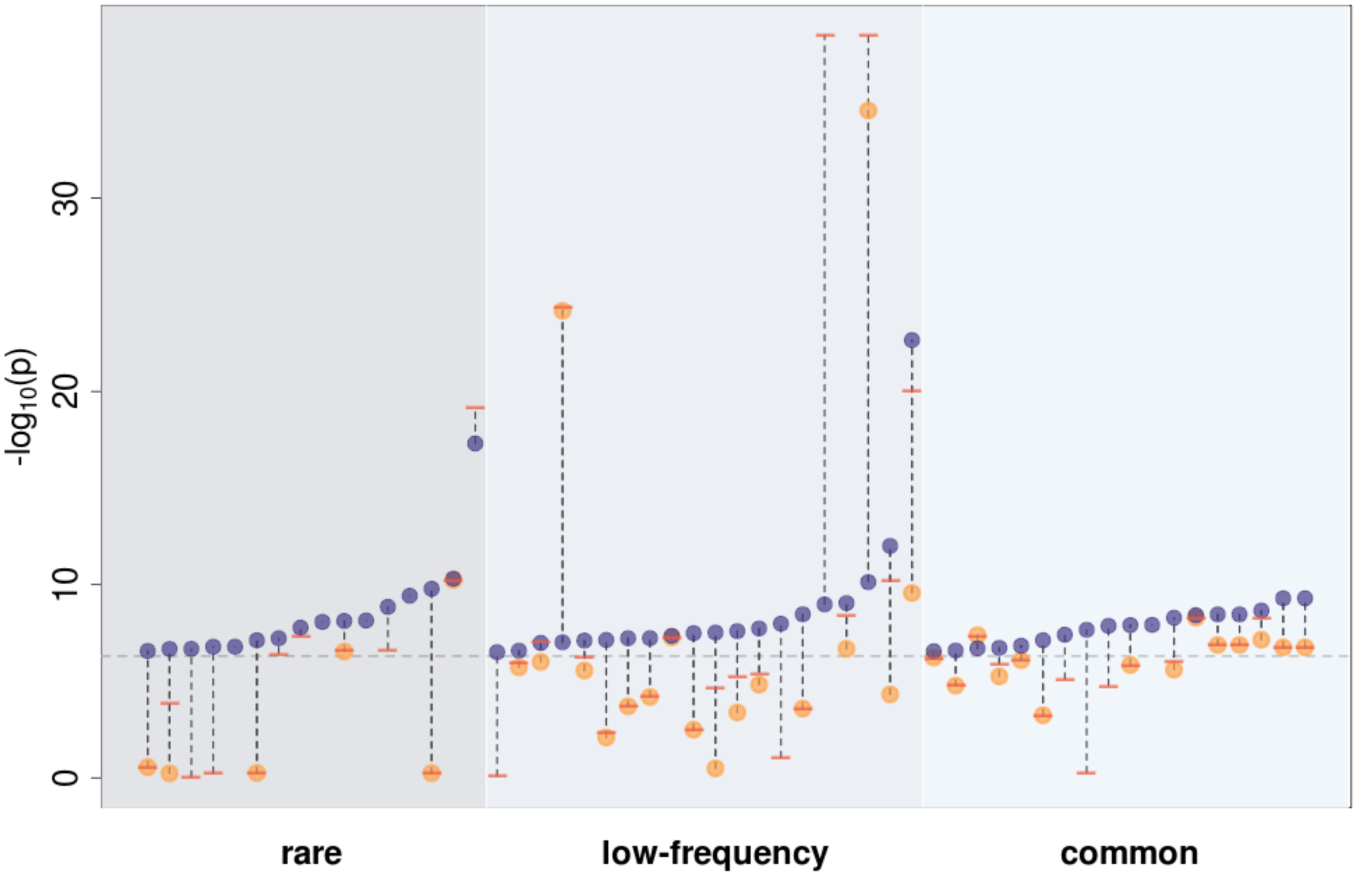


Figure 4: Imputed GWAS results for association signals found in the 1x WGS at $p<5x10^{-7}$. Purple dots represent significant results in the 1x analysis. Orange dots, if present, denote the p-value of the same SNP in the imputed GWAS study. Absence of a dot indicates the variant was not found in the imputed GWAS dataset. Red dashes indicate the minimum p-value among all tagging SNPs in the imputed GWAS (r^2>0.8).



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

References

1. Le SQ, Durbin R: **SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples.** *Genome Res* 2011, **21:**952-960.

2. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526:**68-74.

3. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al: **A reference panel of 64,976 haplotypes for genotype imputation.** *Nat Genet* 2016, **48:**1279-1283.

4. Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, et al: **The UK10K project identifies rare variants in health and disease.** *Nature* 2015, **526:**82-90.

5. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, et al: **The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.** *Cell* 2016, **167:**1415-1429 e1419.

6. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM, Daly MJ, Sklar P, et al: **Extremely low-coverage sequencing and imputation increases power for genome-wide association studies.** *Nat Genet* 2012, **44:**631-635.

7. consortium C: **Sparse whole-genome sequencing identifies two loci for major depressive disorder.** *Nature* 2015, **523:**588-591.

8. Hatzikotoulas K, Gilly A, Zeggini E: **Using population isolates in genetic association studies.** *Brief Funct Genomics* 2014, **13:**371-377.

9. Southam L, Gilly A, Suveges D, Farmaki AE, Schwartzentruber J, Tachmazidou I, Matchan A, Rayner NW, Tsafantakis E, Karaleftheri M, et al: **Whole genome sequencing and imputation in two Greek isolated populations identifies associations with complex traits of medical importance.** *Nat Comms* 2017, **in review**.

10. Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, Mulas A, Pistis G, Steri M, Danjou F, et al: **Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers.** *Nat Genet* 2015, **47:**1272-1281.

11. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, et al: **Large-scale whole-genome sequencing of the Icelandic population.** *Nat Genet* 2015, **47:**435-444.

12. Southam L, Gilly A, Suveges D, Farmaki AE, Schwartzentruber J, Tachmazidou I, Matchan A, Rayner NW, Tsafantakis E, Karaleftheri M, et al: **Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits.** *Nat Comms* 2017.

13. Han B, Eskin E: **Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies.** *Am J Hum Genet* 2011, **88:**586-598.

14. Browning BL, Browning SR: **Genotype Imputation with Millions of Reference Samples.** *Am J Hum Genet* 2016, **98:**116-126.

15. Okada Y, Hirota T, Kamatani Y, Takahashi A, Ohmiya H, Kumasaka N, Higasa K, Yamaguchi-Kabata Y, Hosono N, Nalls MA, et al: **Identification of nine novel loci associated with white blood cell subtypes in a Japanese population.** *PLoS Genet* 2011, **7:**e1002067.

16. Nalls MA, Couper DJ, Tanaka T, van Rooij FJ, Chen MH, Smith AV, Toniolo D, Zakai NA, Yang Q, Greinacher A, et al: **Multiple loci are associated with white blood cell phenotypes.** *PLoS Genet* 2011, **7:**e1002113.

17. Lo KS, Wilson JG, Lange LA, Folsom AR, Galarneau G, Ganesh SK, Grant SF, Keating BJ, McCarroll SA, Mohler ER, 3rd, et al: **Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans.** *Hum Genet* 2011, **129:**307-317.

18. Loh ML, Sakai DS, Flotho C, Kang M, Fliegauf M, Archambeault S, Mullighan CG, Chen L, Bergstraesser E, Bueso-Ramos CE, et al: **Mutations in CBL occur frequently in juvenile myelomonocytic leukemia.** *Blood* 2009, **114:**1859-1863.

19. Chan G, Kalaitzidis D, Usenko T, Kutok JL, Yang W, Mohi MG, Neel BG: **Leukemogenic Ptpn11 causes fatal myeloproliferative disorder via cell-autonomous effects on multiple stages of hematopoiesis.** *Blood* 2009, **113:**4414-4424.

20. Izon DJ, Aster JC, He Y, Weng A, Karnell FG, Patriub V, Xu L, Bakkour S, Rodriguez C, Allman D, Pear WS: **Deltex1 redirects lymphoid progenitors to the B cell lineage by antagonizing Notch1.** *Immunity* 2002, **16:**231-243.

21. Hsiao HW, Liu WH, Wang CJ, Lo YH, Wu YH, Jiang ST, Lai MZ: **Deltex1 is a target of the transcription factor NFAT that promotes T cell anergy.** *Immunity* 2009, **31:**72-83.

22. Gilly A, Ritchie GR, Southam L, Farmaki AE, Tsafantakis E, Dedoussis G, Zeggini E: **Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation.** *Hum Mol Genet* 2016, **25:**2360-2365.

23. Abnizova I, Skelly T, Naumenko F, Whiteford N, Brown C, Cox T: **Statistical comparison of methods to estimate the error probability in short-read Illumina sequencing.** *J Bioinform Comput Biol* 2010, **8:**579-591.

24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.

25. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43:**491-498.

26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

27. Moayyeri A, Hammond CJ, Hart DJ, Spector TD: **The UK Adult Twin Registry (TwinsUK Resource).** *Twin Res Hum Genet* 2013, **16:**144-149.

28. Golding J, Pembrey M, Jones R, Team AS: **ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology.** *Paediatr Perinat Epidemiol* 2001, **15:**74-87.

29. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J: **Haplotype estimation using sequencing reads.** *Am J Hum Genet* 2013, **93:**687-696.

30. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet* 2007, **81:**1084-1097.

31. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR: **Low-coverage sequencing: implications for design of complex trait association studies.** *Genome Res* 2011, **21:**940-951.

32. Browning BL: **Private communication.** 2014.

33. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nat Genet* 2012, **44:**821-824.

34. Arthur R, Schulz-Trieglaff O, Cox AJ, O'Connell J: **AKT: ancestry and kinship toolkit.** *Bioinformatics* 2017, **33:**142-144.

35. Li MX, Yeung JM, Cherny SS, Sham PC: **Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets.** *Hum Genet* 2012, **131:**747-756.

36. Xu C, Tachmazidou I, Walter K, Ciampi A, Zeggini E, Greenwood CM, Consortium UK: **Estimating genome-wide significance for whole-genome sequencing studies.** *Genet Epidemiol* 2014, **38:**281-290.

37. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *Gigascience* 2015, **4:**7.

38. Tang H: **Confronting ethnicity-specific disease risk.** *Nat Genet* 2006, **38:**13-15.

39. Barroso I, Luan J, Wheeler E, Whittaker P, Wasson J, Zeggini E, Weedon MN, Hunt S, Venkatesh R, Frayling TM, et al: **Population-specific risk of type 2 diabetes**

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