nature COMMUNICATIONS

Corrected: Author correction

ARTICLE

DOI: 10.1038/s41467-018-07070-8 OPEN

Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits

Arthur Gilly[1], Daniel Suveges[1], Karoline Kuchenbaecker[1,2,3], Martin Pollard[1,4], Lorraine Southam[1,5], Konstantinos Hatzikotoulas[1,6], Aliki-Eleni Farmaki[7,8], Thea Bjornland[9], Ryan Waples[10], Emil V.R. Appel[11], Elisabetta Casalone[12], Giorgio Melloni[13], Britt Kilian[1], Nigel W. Rayner[1,5,14], Ioanna Ntalla[15], Kousik Kundu[1,16], Klaudia Walter[1], John Danesh[1,17,18], Adam Butterworth[17,18,19], Inês Barroso[1], Emmanouil Tsafantakis[20], George Dedoussis[8], Ida Moltke[10] & Eleftheria Zeggini[1,6]

The role of rare variants in complex traits remains uncharted. Here, we conduct deep whole genome sequencing of 1457 individuals from an isolated population, and test for rare variant burdens across six cardiometabolic traits. We identify a role for rare regulatory variation, which has hitherto been missed. We find evidence of rare variant burdens that are independent of established common variant signals (*ADIPOQ* and adiponectin, $P = 4.2 \times 10^{-8}$; *APOC3* and triglyceride levels, $P = 1.5 \times 10^{-26}$), and identify replicating evidence for a burden associated with triglyceride levels in *FAM189B* ($P = 2.2 \times 10^{-8}$), indicating a role for this gene in lipid metabolism.

[1] Department of Human Genetics, Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom. [2] Division of Psychiatry, University College of London, London W1T 7NF, United Kingdom. [3] UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom. [4] Department of Medicine, Addenbrooke's Hospital, University of Cambridge, Hills Road, Cambridge CB2 0QQ, United Kingdom. [5] Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom. [6] Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg D-85764, Germany. [7] Department of Health Sciences, College of Life Sciences, University of Leicester, Leicester LE1 6TP, United Kingdom. [8] Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens 176-71, Greece. [9] Department of Mathematical Sciences, Norwegian Institute of Science and Technology, Trondheim 7491, Norway. [10] The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark. [11] Section for Metabolic Genetics, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen 2200, Denmark. [12] Human Genetics Foundation, University of Torino, Torino IT-10126, Italy. [13] Department of Biomedical Informatics, Harvard Medical School, Boston 02115 MA, USA. [14] Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Old Road, Headington, Oxford OX3 7LE, United Kingdom. [15] William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, United Kingdom. [16] Department of Haematology, Cambridge Biomedical Campus, University of Cambridge, Long Road, Cambridge CB2 0PT, United Kingdom. [17] The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, University of Cambridge, Cambridge CB1 8RN, United Kingdom. [18] MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Wort's Causeway, University of Cambridge, Strangeways Research Laboratory, Cambridge CB1 8RN, United Kingdom. [19] British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, United Kingdom. [20] Anogia Medical Centre, Anogia 740 51, Greece. These authors contributed equally: Arthur Gilly, Daniel Suveges, Karoline Kuchenbaecker. Correspondence and requests for materials should be addressed to E.Z. (email: Eleftheria@sanger.ac.uk)

Genome-wide association studies have gleaned substantial insights into the genetic architecture of complex traits. The contribution of common-frequency variants to complex traits has been well-documented, and progress in understanding the role of low frequency variation has also gained considerable traction. However, the role of rare variants in the genetic architecture of medically-relevant complex traits remains less well-understood, and the allelic architecture of complex trait association signals has not yet been fully resolved. Rare variant association studies have so far mainly focussed on exonic regions[1], and in whole-genome sequencing studies the optimal analytical approach for rare regulatory variants remains an open question[2]. Population-scale deep whole genome sequencing can capture genetic variation across the entire allele frequency spectrum traversing the coding and non-coding genome. In addition, population isolates offer increased power gains in detecting associations involving rare and low-frequency variants[3].

Here, to improve our understanding of the role of rare variants, we perform cohort-wide deep whole genome sequencing of 1457 individuals from a deeply-phenotyped, isolated population from Crete, Greece (the HELIC-MANOLIS cohort[4–6]) at an average depth of 22.5× (Supplementary Fig. 1), capturing 98% of true single nucleotide variants (SNVs) (Methods and Supplementary Fig. 2). The population genetics characteristics of HELIC-MANOLIS have been studied, and indicate an effective population size of $N_e = 6242$ and an approximate time of divergence of 1100 years from the general Greek population[4,7]. We address open questions on whole genome sequencing study design, analysis and interpretation, and identify burdens of coding and regulatory rare variants associated with cardiometabolic traits.

Results

Effect of sequencing depth. Comparing whole genome sequencing at various depths ranging from 15× to 30× (Methods), we find that 96.4% of singletons, 97.9% of doubletons and 97.6% of variants called using 30× sequencing are recapitulated at 22.5x depth. Genotype accuracy (as measured by r^2) is 99.7% for 22.5× depth and 98.5% for 15× depth, suggesting that increases between 15× and 30× translate into marginal improvements in both call rate and quality of very rare SNVs (Fig. 1, Supplementary Fig. 3 and Methods). We find that false discovery rates and genotype accuracy are substantially more dependent on sequencing depth for INDELs than for SNVs (Fig. 1).

Landscape of sequence variation. Following quality control (QC), we call 24,163,896 non-monomorphic SNVs and INDELs, 97.9% of which are biallelic. 14,281,180 (60.31%) of the biallelic SNVs are rare (minor allele frequency [MAF] < 0.01); 3,103,273 (13.1%) are low-frequency (MAF 0.01–0.05); and 6,292,726 (26.57%) are common (MAF > 0.05). We call 8,294 non-monomorphic variants annotated as loss-of-function (LoF) with low-confidence (LC)[8], and 438 variants annotated as LoF with high-confidence (HC) (Supplementary Fig. 4). On average, each individual carries 405 (standard deviation $\sigma = 19$) LC LoF variants and 31 ($\sigma = 6$) HC LoF variants, compared to 149 LoF variants per sample in a whole genome sequencing study of 2636 Icelanders[9]. 0.6 and 1% of HC and LC LoF carrier genotypes are homozygous, respectively. INDELs are significantly more frequent among LoF variants, with 53.2 and 76% in the low-confidence and high-confidence sets, respectively, compared to 13.5% genome-wide. We observe an enrichment of rare variants among the coding and splice variant categories in MANOLIS (one-sided exact binomial $P = 9.5 \times 10^{-16}$), and we recapitulate this in an independent dataset of 3724 individuals with whole genome sequencing from the UK-based INTERVAL cohort[10](Fig. 2). We also observe a lower rate of singletons compared to the general Greek population and the INTERVAL cohort ($P \approx 10^{-167}$ and $P < 10^{-200}$, respectively, one-sided empirical P-value) (Methods and Supplementary Fig. 5), in keeping with the isolated nature of this Cretan population. Among the 5,102,175 novel biallelic variants (not present in gnomAD[11] or Ensembl release 84[12]), 4,394,678 are SNVs, and the majority are rare (Supplementary Fig. 6).

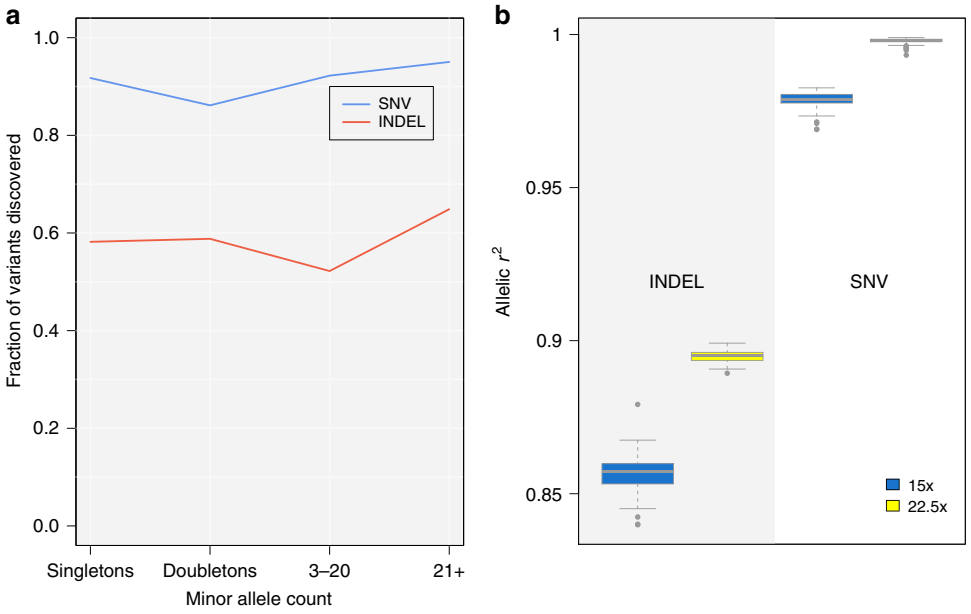

Fig. 1 Variant discovery and quality in WGS data from 100 samples. **a** variant discovery rate in 22.5×; **b** allelic r^2 for SNVs and INDELs in both 15× and 22.5× calls. Depth is downsampled randomly from 30×. INDEL: insertion/deletion. SNV: single nucleotide variant. Boxes represent the interquartile range. Bold horizontal lines in boxplots represent the median, the whiskers extend to 1.5 times the interquartile range, and grey dots represent outliers outside the whisker range

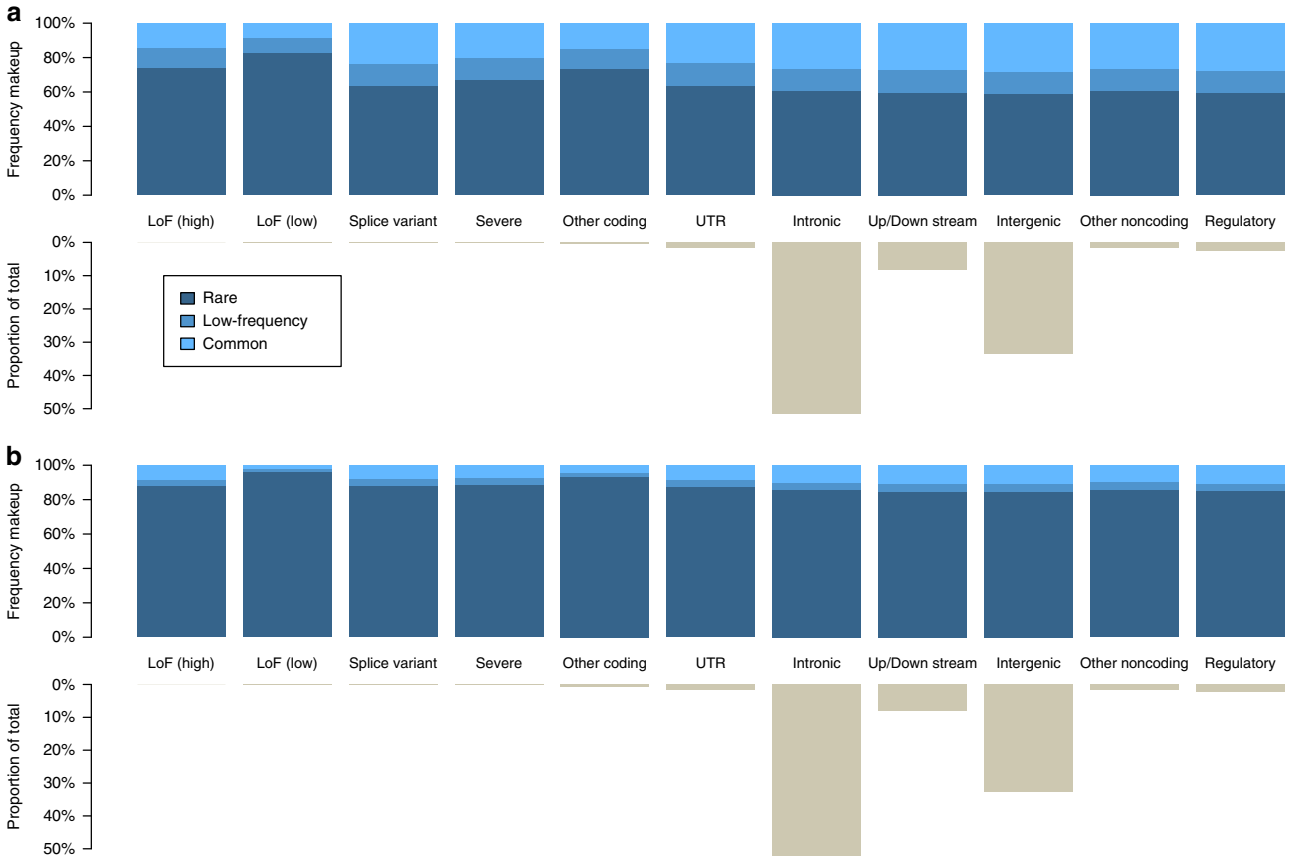

Fig. 2 Variant count proportions and minor allele frequency bin by functional class. **a**, **b** Data is shown for MANOLIS (**a**) and INTERVAL (**b**). Functional classes are derived from the Ensembl VEP consequences as detailed in Supplementary Table 6. The number of intergenic variants is likely to be an underestimate due to Ensembl's most severe consequence annotation. For each panel, the bottom half represents the proportion of variants in each class relative to the total number of variants, the upper half represents the frequency makeup of variants in each class

Refinement of parameters for rare variant burden testing. We carried out genome-wide rare variant burden analyses for six medically-relevant traits: serum adiponectin, bilirubin, gamma-glutamyltransferase, low- and high-density lipoprotein, and triglyceride levels. As choice of genomic region, variant selection and weighting remain open questions for rare variant analysis, we benchmark 10 approaches using different regions of interest (exonic, exonic and regulatory, and regulatory only), variant inclusion and weighting methods (Methods; Supplementary Table 1). Overall, association statistics correlate highly within three distinct clusters (Supplementary Fig. 7). Among exonic-only analyses, rare variant tests that only include unweighted high-consequence variants cluster separately from those in which variants are weighted according to their functionality scores. The third cluster encompasses all tests that include regulatory variants. Neither the variant weighting scheme nor the transformation used for adjusting the weights has a notable influence on the results.

Rare variant burden discovery. In total, twenty burden signals exceed the study-wide significance threshold of 2.0×10^{-7} (Supplementary Fig. 8), arising from four independent genes. Providing proof-of-principle, we identify association of a burden of loss-of-function variants with blood triglyceride and high-density lipoprotein levels in the *APOC3* gene (Fig. 3.a, Supplementary Data File 1)[5,13]. The strongest signal arises when the splice-donor variant rs138326449 (minor allele count (MAC) = 38, minor allele frequency (MAF) = 0.013) and the stop-gained variant rs76353203 (MAC = 62, MAF = 0.022) are included in the analysis ($P = 1.6 \times 10^{-26}$). We replicate the association of a burden of rare coding *APOC3* variants with triglyceride levels in INTERVAL, in which we identify a burden of 25 exonic variants ($P = 3.1 \times 10^{-6}$) (Supplementary Data File 2). This is driven by rs138326449 and rs187628630, a rare 3' UTR variant (MAF = 0.008), with a two-variant burden $P = 9.0 \times 10^{-7}$. rs138326449 is the only loss-of-function variant in *APOC3* present in this cohort, and is four times rarer than in MANOLIS ($MAF_{INTERVAL} = 0.003$ vs $MAF_{MANOLIS} = 0.013$).

We detect a new association of triglyceride levels with rare variants in the *FAM189B* gene (Fig. 3.b, Supplementary Data File 1). The burden association ($P = 1.5 \times 10^{-7}$) is driven by two independent novel splice variants: chr1:155251911 G/A (human genome build 38, MAC = 3, $P = 8.2 \times 10^{-6}$) and chr1:155254079 C/G (MAC = 2, $P = 6.04 \times 10^{-4}$). In both cases, the minor allele is associated with increased triglyceride levels (effect size $\beta = 2.59$ units of standard deviation, $\sigma = 0.57$ and $\beta = 2.40$ $\sigma = 0.69$, respectively). Both variants exhibit high quality scores (VQSLOD > 19), high sequencing read depth (24× and 26.5×, respectively) and no missingness. A further novel splice region variant (chr1:155251496 T/C) and a stop gained variant (rs145265828), both singletons, were also included in the analysis; however their contribution to the burden is insignificant (burden $P = 2.2 \times 10^{-8}$ when excluding them). We replicate evidence for a burden signal at *FAM189B* in the INTERVAL cohort ($P = 9.3 \times 10^{-3}$) (Supplementary Data File 2), which includes two stop gained variants with one driving the association: chr1:155250417 (rs749626426, MAC = 2, $\beta = 1.96$ $\sigma = 0.70$, $P = 5.4 \times 10^{-3}$). In

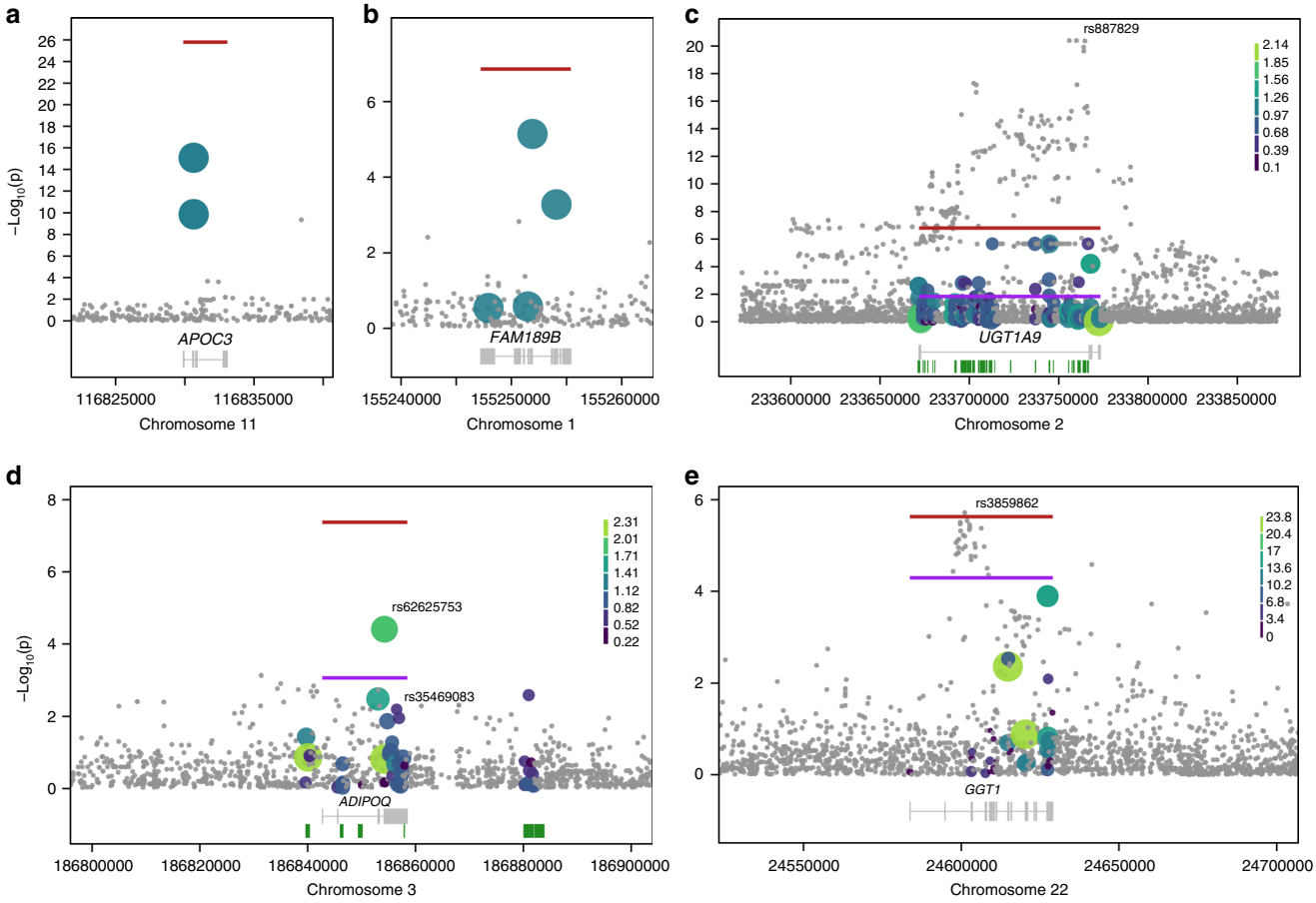

Fig. 3 Regional association plots for burdens in *APOC3, FAM189B, UGT1A9, ADIPOQ,* and *GGT1*. Red lines denote the burden *P*-value and extend over the tested gene. Purple lines indicate the conditioned *P*-value for the variant described in the text (variants rs887829, rs62625753, and rs3859862 in **c**, **d**, **e**. respectively). Small grey dots indicate single-point *P*-value for variants in the region not included in the test. Larger coloured dots represent variants included in the test, with size and colour proportional to the score used in the most significant test (Supplementary Table 1 and Supplementary Data File 1). When no weights are applied (**a** and **b**), included variants are coloured blue-green. In the gene track below the regional plots, green bars below the gene, if present, denote regulatory regions associated with the gene which were used to include variants in the burden. These are present only for genes where regulatory regions were included in the burden

keeping with the discovery dataset, the disruptive minor allele is associated with increased triglyceride levels. The two novel splice-region variants discovered in MANOLIS are not present in either the INTERVAL study or in a compendium of 123,136 exomes and 15,496 whole genomes assembled as part of the gnomAD project[11]. *FAM189B* has not been previously associated with blood lipid levels.

We find evidence of a low frequency and rare variant burden association with bilirubin levels in the *UGT1A9* gene (Fig. 3.c, Supplementary Data File 1). This association arises from the analyses including exonic and regulatory variants ($P = 1.9 \times 10^{-8}$), and from the analyses including regulatory variants only ($P = 7.2 \times 10^{-8}$). We find evidence for association in the exon plus regulatory region burden analysis in the INTERVAL replication cohort ($P = 1.7 \times 10^{-45}$, Supplementary Data File 2). A common variant in the first intron of *UGT1A9* (rs887829, MAF = 0.28, $\beta = 0.426$ $\sigma = 0.04$, two-sided score test $P = 4.0 \times 10^{-21}$ in the MANOLIS cohort) has previously been associated with bilirubin levels[14,15]. As expected, genotype correlation between rs887829 and each of the low-frequency and rare variants included in the burden is low ($r_{max}^2 = 0.1$). The rs887829 signal is not attenuated when conditioning on carrier status for the two main drivers of the burden (single-point score test $P_{conditional} = 4.5 \times 10^{-21}$), or when conditioning on the number of rare alleles carried per individual ($P_{conditional} = 4.0 \times 10^{-21}$). The evidence for association with the rare variant burden in *UGT1A9* is substantially reduced when conditioned on rs887829 (burden $P_{conditional} = 0.0146$). Conversely, the two-variant signal for the two main burden drivers is attenuated from $P = 1.4 \times 10^{-7}$ to $P_{conditional} = 7.0 \times 10^{-3}$ when conditioning on rs887829, indicating that it likely recapitulates part of a signal driven by a known common-variant association in the region.

We identify an association of adiponectin levels with low-frequency and rare variants in the *ADIPOQ* gene (Fig. 3.d, Supplementary Data File 1). The evidence for association is stronger for exonic and regulatory variants combined ($P = 4.2 \times 10^{-8}$) than in either the regulatory-only ($P = 0.19$) or exon-only ($P = 2.0 \times 10^{-6}$) analyses, suggesting a genuine contribution of both classes of variants to the burden. The missense variant rs62625753 (MAF = 0.031, two-sided score test $P = 4.0 \times 10^{-5}$) contributes to the burden signal and is predicted to be damaging. The strength of association for the burden is reduced, but not entirely attenuated, when conditioned on rs62625753 ($P_{conditional} = 8.9 \times 10^{-4}$), indicating that it is not singly driven by this variant. rs35469083 (MAF = 0.044) also contributes to the burden, and is an expression quantitative trait locus (eQTL) for *ADIPOQ* in visceral adipose tissue (minor allele associated with decreased gene expression). rs62625753 and rs35469083 have

consistent directions of effect, with the minor alleles associated with reduced adiponectin levels, in keeping with their functional consequences on the gene (two-variant burden $P = 4.8 \times 10^{-7}$). No common-variant signal for adiponectin levels is present in this region in our dataset. The burden signal remains significant upon conditioning on the genotypes of all variants with previous associations for adiponectin, type 2 diabetes or obesity that are polymorphic in MANOLIS (Supplementary Table 2).

In addition to the four genes that meet study-wide significance, we find gamma-glutamyltransferase levels to be suggestively associated with a burden of low frequency and rare exonic variants in the gamma-glutamyltransferase 1 (*GGT1*) gene ($P = 2.3 \times 10^{-6}$) (Fig. 3.e, Supplementary Data File 1). A previously-reported, common-variant association is also present in an intron of this gene (rs3859862, MAF = 0.46, two-sided score test $P = 1.9 \times 10^{-6}$). The burden signal in *GGT1* is maintained when conditioning on rs3859862 ($P_{\text{conditional}} = 5.1 \times 10^{-5}$), suggesting that rare variants be independently contributing to this established association. Similarly, the single-point association at rs3859862 conditioned on carrier status for all rare variants included in the burden is not attenuated ($P_{\text{conditional}} = 2.8 \times 10^{-5}$), a result recapitulated by conditioning the same variant on the number of rare alleles carried per individual ($P_{\text{conditional}} = 1.8 \times 10^{-5}$), providing evidence for an independent rare variant signal at this locus.

Signatures of selection. We surveyed the genomic loci with evidence of rare variant burden signals for signatures of recent or ongoing positive selection in the MANOLIS cohort, using integrated haplotype scores (iHS)[16]. Previous studies have shown that an elevated fraction of SNVs with $|\text{iHS}| > 2$ in a genomic region is a signature of recent or ongoing selection and notably, we find that 32% of the SNVs in *FAM189B* have an iHS score above 2, placing it in the top 5% of all genes analysed (96.7th percentile). This result is robust across several definitions of the genomic region representing the genes (95.6th-98.3th percentile) and to conditioning on gene length (94.6th percentile) (Supplementary Table 3). To further investigate this potential signature of selection in *FAM189B*, we examined the extent to which the allele frequencies in *FAM189B* differ between the MANOLIS cohort and the 1000 Genomes CEU population sample using weighted mean F_{ST}. Like with iHS, *FAM189B* lies in the top 5% of all genes analysed across several definitions of the genomic regions (Supplementary Table 4).

Discussion

In this work, we have whole genome sequenced 1457 individuals from the HELIC-MANOLIS cohort at an average depth of 22.5×. We describe the genomic variation landscape in this special population, discover 5.1 million novel variants, and perform rare variant burden testing across the entire genome for medically-relevant biochemical traits.

We empirically address several open whole genome sequencing study design and analysis questions. Through a downsampling approach, we demonstrate that it is possible to achieve near-perfect sensitivity and quality for rare SNV calling and genotyping with half the depth, and at substantially lower cost, compared to 30× sequencing. This observation does not extend to INDELs, for which depth increases above 15× can result in a 15% increase in genotype quality and a 40% increase in true positive rate.

Defining the genomic regions in which to select variants, filtering strategies and variant weighting schemes constitute unresolved challenges in whole genome sequence-based studies. We find that association signal profiles of tests including regulatory region variants differ markedly from other scenarios, with some signals being driven by this variant class. Further, signal strength differs substantially between analyses that include high-severity consequence exonic variants only, and those in which all exonic variants are weighted according to their predicted consequence. We find that, as a rule, variant and functional unit selection, rather than weighting scheme, plays the largest role in association testing.

We identify a role for rare regulatory variants in the allelic architecture of complex traits. It is therefore important to leverage the whole genome sequence nature of the study data, and not to restrict analyses to coding variation only. We observe congruent directions of effect among regulatory and coding rare variants in burden signals that combine both classes of variation, for example across eQTL and damaging missense variants in the *ADIPOQ* gene that are together associated with adiponectin levels.

We discover replicating evidence for association of a rare variant burden with triglyceride levels at a locus not previously linked with the trait. *FAM189B* (Family With Sequence Similarity 189 Member B), also known as *COTE1* or *C1orf2*, codes for a membrane protein that is widely expressed, including in adult liver tissue[17]. Expression of *FAM189B* has been found to be correlated with endogenous SREBP-1 activation in vitro[18]. Sterol-regulatory element binding proteins (SREBPs) control the expression of genes involved in fatty acid and cholesterol biosynthesis, therefore indicating a mechanism by which *FAM189B* could be involved in lipid metabolism. We found *FAM189B* to contain an elevated fraction of SNVs with $|\text{iHS}|>2$, a potential signature of recent positive selection. Furthermore, *FAM189B* is in the top 5% of all genes in terms of population differentiation (F_{ST}) between the MANOLIS cohort and the 1000 Genomes Project CEU sample, which is consistent with selection having happened in MANOLIS. This is particularly interesting in the context of this population, which has a high animal fat content diet[6], and for which loss of function variants in *APOC3* have risen in frequency compared to the general population and confer a cardioprotective effect[19,20]. For the same reason, it is interesting to note that *FAM189B* has not previously been reported to be under selection in other populations[21]. However, we caution that although *FAM189B* is in the top 5% of all genes for both iHS and F_{ST}, it is not an extreme outlier for either, suggesting that it could be a false positive or that the selection has not acted strongly enough or for long enough to leave more than subtle signatures in the haplotype structure and allele frequencies in the gene. It is also possible that selection has acted on several rare alleles making the signature more complex than simple directional selection.

We replicate the *FAM189B* association in an independent dataset with deep whole genome sequence data, in which the disruptive rare alleles are also associated with the same trait in the same direction. Across the board, we replicate all burden signals for which replication cohort trait measurements are available. We find that allelic heterogeneity is prevalent, partly due to the rare nature of the variants contributing to the burdens, and partly due to the distinct population genetics characteristics of the discovery and replication sets. Perhaps as a consequence, the outcome of variant filtering and weighting was quite sensitive to the study population, and all the burdens reported here replicated strongest in slightly different testing conditions from the discovery, although in the same broad functional class. The association in *FAM189B* was discovered when including exonic variants with a relaxed severity threshold, whereas it replicated in the LoF-only analysis. Similarly, the *APOC3* signal was discovered in the LoF-only analysis but replicated in the CADD-weighted exonic analysis. These findings have important consequences for defining replication in sequence-based studies of rare variants, and

highlight the importance of defining replication at the locus level rather than the variant level for burden signals.

We demonstrate pervasive allelic heterogeneity at complex trait loci, and identify exonic and regulatory rare variant associations at established signals. We find multiple instances of burden signals that remain independent of colocalising common variant signals, and one instance of burden signal attenuation when conditioning on the established common variant association. Within the power constraints of the study, we do not find evidence for synthetic association at established signals, i.e., there is no evidence for multiple rare variants at a locus accounting for a common variant association.

The discovery of rare variant burden associations with a modest sample size has been made possible due to the special population genetics characteristics of the isolated cohort under study. Rare variant signals, such as the ones discovered in *APOC3* and *FAM189B* in MANOLIS, are driven by variants with severe consequences that are rarer or absent in cosmopolitan populations. This demonstrates that the well-rehearsed power gains conferred by isolated cohorts in genome-wide association studies[3] extend to whole genome sequence-based rare variant association designs.

Our findings indicate that deep whole genome sequencing at scale will be required to enable exhaustive description of the rare variant burden landscape in a population. For example, in the case of the *FAM189B* signal, low-depth sequencing (1× depth) of 1239 MANOLIS samples[22] misses one of the two burden-driving variants (chr1:155251911, MAC = 3). Similarly, genome-wide genotyping coupled to dense imputation of the same samples does not capture the variants driving the burden signal identified here through deep whole genome sequencing[23].

Our findings provide evidence for a role of low-frequency and rare, regulatory and coding variants in complex traits, and highlight the complex nature of locus-specific architecture at established and newly emerging signals. We anticipate that larger-scale, cohort-wide, deep whole genome sequencing initiatives will substantially further contribute to our understanding of the genetic underpinning of complex traits.

Methods

Ethics and informed consent statement. In the TEENAGE study, prior to recruitment all study participants gave their verbal assent along with their parents'/guardians' written consent forms. The study was approved by the Institutional Review Board of Harokopio University and the Greek Ministry of Education, Lifelong Learning and Religious Affairs. The MANOLIS study was approved by the Harokopio University Bioethics Committee and informed consent was obtained from every participant. The INTERVAL study was approved by the Cambridge South Research Ethics Committee and informed consent was obtained from every participant.

Sequencing. For MANOLIS, genomic DNA (500 ng) from 1482 samples was sheared to a median insert size of 500 bp and subjected to standard Illumina paired-end DNA library construction. Adapter-ligated libraries were amplified by 6 cycles of PCR and subjected to DNA sequencing using the HiSeqX platform (Illumina) according to manufacturer's instructions. For TEENAGE, one hundred samples from the general Greek population were sequenced, as well as the Genome in a Bottle NA12878 sample. Sample identity checks were performed using Fluidigm and aliquots prepared. These aliquots underwent library preparation using the standard HiSeqX method. Size selection was performed to target 350 base pairs. Sequencing was performed on the Sanger Institute's Illumina HiSeqX plat-form with a target depth of 30x and PhiX spike-in.

Evaluation of sequencing accuracy at various depths. Reads from the NA12878 were downsampled to several read depths (from 5× to 30×) using the -s option of samtools view, aligned and processed through GATK Variant Quality Score Recalibrator. They were then compared to Genome in a Bottle (GIAB) 0.2 calls to extract the true positive rate (Supplementary Fig. 2). At 22.5×, true positive rates are 98% for SNVs and 76% for INDELs.

Comparison with the general Greek population. We compared variant callsets in MANOLIS to a dataset of 100 samples from the Greek general population (TEENAGE study), for which an identical sequencing protocol was used. The average depth in the TEENAGE study was 32.1×. We downsampled the individual BAMs to 22.5× and 15× using the -s option of samtools based on the average depth of the TEENAGE dataset, then performed variant calling using GATK HaplotypeCaller v3.3 (https://github.com/mp15/af_analysis) and filtering using GATK Variant Quality Score Recalibrator. The downsampled and original datasets were then compared using bcftools stats to extract allelic r-squared (Fig. 1.b.). For the 22.5× dataset, we compared variant overlap with bcftools isec (Fig. 1.a. and Supplementary Fig. 3).

Rare variant counts in MANOLIS, TEENAGE and INTERVAL. Since sample sizes differ between the three datasets, we randomly subsampled the larger dataset to a matching size for each pairwise comparison. We used these resampled datasets to build empirical distributions for rare variant counts in the larger dataset, and compared it to counts in the smaller dataset. TEENAGE ($n = 100$) was smaller compared to MANOLIS ($n = 1482$), so we drew 1000 sets of 100 samples from the MANOLIS study for the comparison. We counted 270,916 singletons and 61,690 doubletons in TEENAGE, compared with a median of 179,100 (one-sided $P = 1.4 \times 10^{-94}$ from a fitted normal distribution) and 75,280 (one-sided $P = 3.0 \times 10^{-19}$ from a fitted normal distribution), respectively, in MANOLIS (Supplementary Fig. 5a,b.). For n = 100, singletons correspond to MAF < 0.005 and doubletons to 0.005 < MAF < 0.1.

For the INTERVAL ($n = 3742$) comparison, MANOLIS was the smaller dataset, so we resampled 500 sets of 1482 samples from the INTERVAL cohort and counted variants up to MAC = 29 (MAF = 0.01). The increased resolution provided by this larger sample size shows that rare variant counts are greater in the cosmopolitan population below MAC = 4 (MAF = 0.0013), but greater in the isolate for 0.0013 < MAF < 0.1, consistent with our coarser observation in TEENAGE (Supplementary Fig. 5.c).

For p-values of singleton counts, empirical quantiles cannot be computed for such large deviations from the mean. We fitted a normal distribution to singleton counts, and computed the theoretical quantile corresponding to the observed count in the smaller cohort.

Variant calling. Basecall files for each lane were transformed into unmapped BAMs using Illumina2BAM, marking adaptor contamination and decoding barcodes for removal into BAM tags. PhiX control reads were mapped using BWA Backtrack and were used to remove spatial artefacts. Reads were converted to FASTQ and aligned using BWA MEM 0.7.8 to the 1000 Genomes hs37d5 (for NA12878) and hg38 (GRCh38) with decoys (HS38DH) (for TEENAGE) references. The alignment was then merged into the master sample BAM file using Illumina2BAM MergeAlign. PCR and optical duplicates are marked using biobambam markduplicates and the files were archived in CRAM format.

Per-lane CRAMs were retrieved and reads pooled on a per-sample basis across all lanes to produce library CRAMs; these were each divided in 200 chunks for parallelism. GVCFs were generated using HaplotypeCaller v.3.5 from the Genome Analysis Toolkit (GATK) for each chunk. All chunks were then merged at sample level, samples were then further combined in batches of 150 samples using GATK CombineGVCFs v.3.5. Variant calling was then performed on each batch using GATK GenotypeGVCFs v.3.5. The resulting variant callsets were then merged across all batches into a cohort-wide VCF file using bcftools concat.

Quality control. Variant-level QC was performed using the Variant Quality Score Recalibration tool (VQSR) from the Genome Analysis Toolkit (GATK) v. 3.5-0-g36282e4[24], using a tranche threshold of 99.4% for SNPs, which provided an estimate false positive rate of 6%, and a true positive rate of 95%. For INDELs, we used the recommended threshold of 1%. For sample-level QC, we made extensive use of a previously described[23] GWAS dataset in 1175 overlapping samples. Four individuals failed sex checks, 8 samples had low concordance ($\hat{\pi}$<0.8) with chip data, 11 samples were duplicates, and 12 samples displayed traces of contamination (Freemix score from the verifyBamID suite[25]>5%). In case of sample duplicates, the sample with highest quality metrics (depth, freemix and chipmix score) was kept. As contamination and sex mismatches were correlated, a total of 25 individuals were excluded ($n = 1457$). No further samples were excluded based on depth, heterozygosity, transition/transversion (Ti/Tv) rate, missingness or ethnicity. No rare or low-frequency variant (MAF < 5%) was excluded based on the Hardy-Weinberg equilibrium test at $P = 1.0 \times 10^{-5}$. We filtered out 14% of variants with call rates < 99%.

Genetic relatedness matrix. Several methods are available to estimate the genetic relatedness present in isolated cohorts such as HELIC-MANOLIS[26]. We compared methods proposed in GEMMA[27], EMMAX[28], KING[29] and PLINK[30]., and found that the kinship coefficients reported by each method were highly correlated, but on a different scale from each other (Supplementary Fig. 9). For consistency with previous studies performed on the same samples, we calculated a genetic relatedness matrix using GEMMA[27] after filtering for MAF < 0.05, missingness <1% and LD-based pruning. In addition, MONSTER requires self-kinship coefficients on the

diagonal of the relatedness matrix, which we calculated using the $\widehat{F_1}$ metric from PLINK 1.9. The matrix was then converted to the long format using the reshape2 R package.

Association testing. Burden testing was performed using MONSTER[31], a method that extends the SKAT-O[32] model to account for relatedness and/or structure present in cohorts such as population isolates when testing for association. We ran burden testing across all genes defined in GENCODE v25 using 10 different conditions, i.e., combinations of regions of interest (coding regions only, coding and regulatory regions and regulatory regions only), variant filters (inclusion criteria based on severity of predicted consequence) and weighting schemes (Supplementary Table 1). QQ-plots for all testing conditions and traits are presented in Supplementary Fig. 10.

First, we extracted exonic coordinates for all protein-coding genes, which defines the region of interest for strictly exonic variants. These regions of interest were used in combination with 5 different variant filtering and weighting schemes. First, we included only variants predicted as high-confidence (HC) loss-of-function (LoF) by LOFTEE[8] that reside in the exons of protein-coding genes (Supplementary Table 1: LOFTEE HC). As only 460 variants in 85 genes passed this inclusion criterion, we performed an additional analysis including 8,570 low-confidence (LC) loss-of-function variants spread across 1,727 genes (Supplementary Table 1: LOFTEE LC). Stop-gained and frameshift mutations were the largest contributors to both the LC and HC sets. However, the LC set also includes a large number of splice donor and splice acceptor variants (Supplementary Fig. 4). We further performed an analysis with more relaxed inclusion criteria, including all exonic variants for which the Ensembl most severe consequence was more damaging than missense as predicted by the Variant Effect Predictor[33] (Supplementary Table 1: Exon severe). We also employed Combined Annotation Dependent Depletion (CADD)[34] scores, either to weigh all exonic variants (Supplementary Table 1: Exon CADD) or to filter out variants with CADD scores below the genome-wide median (Supplementary Table 1: Exon CADD median). Finally, we extended exon boundaries as defined above with 50 base pairs either side, to account for cases where potentially damaging variants occur on the edges of exons, as has been shown to happen for previously identified rare variant burdens[5]. These regions of interest were used in combination with one variant weighting scheme only (Exon + 50 CADD).

We extracted regulatory regions (promoters, enhancers and transcription-factor binding sites) from Ensembl build 84[12]. We assigned regulatory regions to genes if they directly overlapped or if the regulatory region overlapped with an eQTL for the gene based on the GTEx database[35]. If an eQTL was reported for several genes, overlapping variants were assigned to all of them. We did not take tissue specificity into account. For selecting variants, we either used the coordinates of the regulatory features alone, or regulatory features plus the extended exons. We used Eigen, an aggregate score that combines information from multiple regulatory annotation tracks[36], to weigh variants in all tests that include regulatory variants. In addition to raw Eigen scores, the authors also proposed EigenPC, a score derived from the first eigenvector of the correlation matrix of annotations. Both scores were available as is, or transformed using Phred-scaling, which maps a distribution's support to] 0,+∞[, thereby guaranteeing inclusion and relative up-weighting of all variants. In the regulatory regions plus exon analyses we used both the raw Eigen scores, shifted by 1 unit to the right, with negative scores set to $0 + \epsilon$ (Supplementary Table 1: Exon and regulatory Eigen), and the Phred-transformed Eigen and EigenPC scores (Supplementary Table 1: Exon and regulatory EigenPhred and EigenPCPhred). This transformation was a technical requirement as MONSTER could only read weights belonging to]0,+∞[. In the analyses containing the regulatory regions only, variants were weighted using the Phred-scaled Eigen scores (Supplementary Table 1: regulatory only EigenPhred) only.

Finally, we applied a MAF threshold of 0.05, a missingness threshold of 1% and a Hardy–Weinberg filter using a mid-p adjusted P-value[37] threshold of 1.0×10^{-5} to all variants prior to testing. We only performed a test if at least two SNVs passed the inclusion criteria for a given condition.

Establishing the significance threshold. We calculated $\alpha_{\text{eff}} = \frac{0.05}{N \times n_{\text{cond}} \times M}$, where N is the number of genes tested, n_{cond} is the effective number of inclusion and weighting criteria tested and M=6 is the number of traits. For n_{cond}, we plotted the correlation matrix of z-scores for all 10 analyses, and determined that the analyses using similar region definitions (exonic loss-of-function, exonic, exonic and regulatory variants) cluster together, reducing the effective number of analyses to 3 (Supplementary Fig. 7). Although $N = 18{,}997$ protein-coding genes are available in GENCODE V25, not all genes were tested in every condition. For example, for many genes only one variant might pass inclusion criteria in a high-confidence loss-of-function run, thereby excluding those genes from the analysis. A summary of the number of genes included in every analysis is presented in Supplementary Table 5. On average, $N = 13{,}854$ genes are included, hence we define study-wide significance at $P = 2.0 \times 10^{-7}$.

Burden prioritisation and novelty. We applied stringent checks to test the validity of rare variant burden association signals. Every suggestively associated burden (arbitrarily defined as $P \leq 5 \times 10^{-5}$) was conditioned on the genotypes of the variant included in the burden set with the lowest single-point P-value. If the P-value dropped more than two orders of magnitude below the suggestive significance threshold (i.e., $P \leq 5.0 \times 10^{-3}$), the burden was excluded from downstream analyses. We examined burden signals using the plotburden software (https://github.com/wtsi-team144/plotburden) to assess variant functionality, single-point association P-values, LD structure, as well as prior associations in the region. When a prior association was found in the region, we considered a signal known when the P-value dropped below $P = 1.0 \times 10^{-4}$ when conditioning on the genotypes of the existing signal. We examined rare variant burden associations with suggestive significance ($P < 5 \times 10^{-5}$) across the six traits under investigation, and do not find evidence of further rare variant signals at established loci.

Replication. The INTERVAL randomised controlled trial is a large-scale study focusing on healthy blood donors[10]. Sequencing, variant calling and quality control was performed for 3762 INTERVAL participants using the same protocol and pipeline as for the MANOLIS sequences. 38 samples were excluded on the basis of ethnicity, excessive relatedness ($\hat{\pi} > 0.125$), excess heterozygosity and contamination. VQSR thresholds of 99 and 90% for SNVs and INDELs, respectively, were applied to variant calls. Gamma-glutamyltransferase and adiponectin levels were not available in the INTERVAL replication cohort.

Selection analyses. For the selection analyses we used the haplotype-based iHS statistic[16]. We used this statistic because we were mainly interested in recent or ongoing selection, i.e., selective sweeps where the advantageous allele has not yet reached a high frequency (>80%), and iHS has been shown to be more powerful than other commonly used statistics like Tajima's D[38] and XP-EHH[39] for detecting such sweeps[16,40]. Briefly, the iHS value of an SNV becomes elevated when one of the alleles of that SNV reside on haplotypes that are longer than expected under neutrality, given the frequency of the allele. This is considered a signature of positive selection because positive selection will cause haplotypes carrying an advantageous allele to increase in frequency faster than if the allele had been neutral which leaves less time for recombination to shorten them.

To investigate if any of the four genes with study-wide significant burden association signals have undergone recent or ongoing positive selection, we calculated the fraction of SNVs with |iHS|>2 for these four genes and assessed if these fractions were elevated by comparing them to the empirical distribution for all genes. We focused on the fraction of SNVs with |iHS|>2, because previous studies have compared different methods of summarising iHS for a genomic region of interest, and found that the fraction of SNVs |iHS|>2 is often the most powerful iHS summary for detecting selection[16,40].

The primary data used for the selection analyses is the MANOLIS genotype data described above, which we phased using Beagle v.4.1[41]. We also used the ancestral allele annotations for each site in hg38 from ENSEMBL, and the recombination map from UCSC, which was built on hg37 and lifted-over to hg38. From this map, we excluded 3140 sites to achieve a recombination map with monotonically increasing cM positions. Linear interpolation was subsequently used to produce cM positions for all sites not found on the map. For quality control, we combined the MANOLIS genotype data with genotype data from the 1000 Genomes phase 3[42].

Because close relatives can complicate and potentially bias analyses of signals of selection, we removed close relatives within the MANOLIS data after phasing, as well as one admixed individual. We used the same criteria as in a previous study of this population[4], i.e., we used the --genome option in PLINK 1.9 to estimate PI-HAT and randomly excluded one individual from each pair of individuals with $\hat{\pi}$>0.2. These exclusions left 810 unrelated individuals from MANOLIS, on which we based the selection analysis.

We restricted our iHS analysis to known, common SNVs with ancestral allele annotations. Specifically, we excluded sites: not on the autosomes, with more than 2 alleles, with alleles that were not length 1 (INDEL-like), with MAF < 0.05, without ancestral allele annotations, with HWE midp-value $< 1 \times 10^{-30}$, not present in 1000 Genomes phase 3 vcf files, or that were outside a mappable region of the hg38 reference genome, defined as a GEM 100-mer score below 0.8[43]. These filtering steps resulted in 5,126,987 SNVs as input for iHS calculations.

The iHS statistic was calculated with the hapbin program[44], using default parameters. The raw iHS statistic is sensitive to allele frequency, so SNVs were subsequently binned by derived allele count (82 equally-spaced bins) and the iHS statistic was normalised within each bin to have a mean of zero and a standard deviation of one, as suggested in[16]. Finally, we examine the absolute value of the normalised iHS statistic to capture selection signals associated with both derived and ancestral alleles. Due to edge effects at chromosome ends and other gaps, we examine iHS values for 5,116,861 SNVs (99.8% of input sites).

For each gene, we considered four distinct ways to define the genomic region representing the gene: (1) sites within exons, (2) sites within exons extended by 50 bp or in regulatory elements, (3) sites within the region spanned by connecting all exons, and (4) sites within the region spanned by connecting all exons extended by 50 bp and regulatory elements. For each gene and each of the four genic region definitions, we extracted SNVs with an iHS value and calculated the fraction of SNVs with normalised |iHS| above 2. When interpreting our results, we mainly focused on the results for the most inclusive definition, definition 4, as selection signatures tend to span fairly large genomic regions, but included the other

definitions to be able to assess if this choice of definition markedly affected our results.

For each lead gene with a rare variant study-wide significant burden signal (*APOC3, UGT1A9, ADIPOQ*, and *FAM189B*), we compared its fraction of SNVs with |iHS|>2 to all other genes with at least 1 iHS value-bearing SNV, using each of the four different gene region definitions. Each comparison was quantified by the percentile of genes with a higher fraction of SNVs with |iHS|>2. *FAM189B* was the only of the four burden genes with a fraction |iHS|>2 above zero. For this gene, we also performed a comparison to the subset of genes with a similar number of SNVs with iHS values as *FAM189B* (defined as +/− 10% of the number of SNVs with iHS in *FAM189B*) to ensure the varying number of SNVs in the genes we compared *FAM189B* to did not drastically affect the percentiles. Note that with the gene definitions used some SNVs will be included in several genes and thus the data points in the empirical distribution used for comparison are not entirely independent.

F_{ST} between two populations is a measure of population differentiation and is expected to increase in a region harbouring an allele which has been under positive selection mainly in one of the two populations.

To further investigate the *FAM189B* gene we calculated F_{ST} between the MANOLIS cohort and the European 1000 genomes CEU population for this gene and compared it to that of all other genes. The comparison was performed like for the iHS values, i.e., using quantiles and by performing several different comparisons to check for robustness of the results. For this analysis, we used the same genetic data from MANOLIS as for the iHS analyses combined with data from the 1000 genomes CEU population sample, except we did not filter away SNVs with MAF < 0.05, without ancestral allele annotations or with HWE midp-value $< 1 \times 10^{-30}$. Following published recommendations[45], all F_{ST} estimates were performed using the Hudson estimator and per SNP estimates were combined using the ratio of the average numerator and the average denominator (also referred to as weighted mean F_{ST}). However, we note that similar results were obtained using the Weir and Cockerham F_{ST} estimator[46].

Code availability. MUMMY, the script used to run burden tests genome wide using MONSTER, is available at https://github.com/wtsi-team144/burden_testing. The plotburden script, which builds interactive visualisations of burden signals, is available at https://github.com/wtsi-team144/plotburden.

Data availability

Sequencing data are available at the European Genome-Phenome Archive under accession numbers EGAS00001001207 for MANOLIS, EGAS00001000988 for TEENAGE, and EGAS00001001355, EGAS00001002461, and EGAS00001002787 for INTERVAL.

Received: 16 March 2018 Accepted: 8 October 2018
Published online: 07 November 2018

References

1. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, (2016).
2. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).
3. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief. Funct. Genom.* **13**, 371–377 (2014).
4. Panoutsopoulou, K. et al. Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat. Commun.* **5**, 5345 (2014).
5. Gilly, A. et al. Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. *Hum. Mol. Genet.* **25**, 2360–2365 (2016).
6. Farmaki, A. E. et al. The mountainous Cretan dietary patterns and their relationship with cardiovascular risk factors: the Hellenic Isolated Cohorts MANOLIS study. *Public Health Nutr.* **20**, 1063–1074 (2017).
7. Xue, Y. et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat. Commun.* **8**, 15927 (2017).
8. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
9. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
10. Moore, C. et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).
11. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
12. Zerbino, D. R. et al. Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
13. TG and HDL Working Group of the Exome Sequencing Project, N.H.L. et al. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N. Engl. J. Med.* **371**, 22–31 (2014).
14. van Es, H. H. et al. Assignment of the human UDP glucuronosyltransferase gene (UGT1A1) to chromosome region 2q37. *Cytogenet. Cell Genet.* **63**, 114–116 (1993).
15. Sanna, S. et al. Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum. Mol. Genet.* **18**, 2711–2718 (2009).
16. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
17. Zhang, H. et al. Ectopic overexpression of COTE1 promotes cellular invasion of hepatocellular carcinoma. *Asian Pac. J. Cancer Prev.* **13**, 5799–5804 (2012).
18. Kallin, A. et al. SREBP-1 regulates the expression of heme oxygenase 1 and the phosphatidylinositol-3 kinase regulatory subunit p55 gamma. *J. Lipid Res.* **48**, 1628–1636 (2007).
19. Tachmazidou, I. et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat. Commun.* **4**, 2872 (2013).
20. Pollin, T. I. et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702–1705 (2008).
21. Li, M. J. et al. dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res.* **42**, D910–D916 (2014).
22. Gilly, A. et al. Very low depth whole genome sequencing in complex trait association studies. *bioRxiv* (2017).
23. Southam, L. et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat. Commun.* **8**, 15606 (2017).
24. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
25. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
26. Eu-Ahsunthornwattana, J. et al. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet.* **10**, e1004445 (2014).
27. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
28. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
29. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
30. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
31. Jiang, D. & McPeek, M. S. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet. Epidemiol.* **38**, 10–20 (2014).
32. Lee, S. et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
33. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
34. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
35. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
36. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
37. Graffelman, J. & Moreno, V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Stat. Appl. Genet. Mol. Biol.* **12**, 433–448 (2013).
38. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
39. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
40. Pickrell, J. K. et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).
41. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
42. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
43. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS ONE* **7**, e30377 (2012).
44. Maclean, C. A., Chue Hong, N. P. & Prendergast, J. G. hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. *Mol. Biol. Evol.* **32**, 3027–3029 (2015).

45. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
46. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

Acknowledgements

HELIC-MANOLIS study: We thank the residents of the Mylopotamos villages for taking part. The MANOLIS study is dedicated to the memory of Manolis Giannakakis, 1978–2010. This work was funded by the Wellcome Trust [098051] and the European Research Council [ERC-2011-StG 280559-SEPI]. INTERVAL study: Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping was funded by the National Institute of Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk/) and the NIHR Cambridge Biomedical Research Centre (www.cambridge-brc.org.uk). The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), and NIHR Research Cambridge Biomedical Research Centre. This report is independent research by the National Institute for Health Research. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. This work was undertaken by Cambridge who received funding from the NHSBT; the views expressed in this publication are those of the authors and not necessarily those of the NHSBT. TEENAGE study: The TEENAGE study has been supported by the Wellcome Trust (098051), European Union (European Social Fund—ESF) and Greek national funds through the Education and Lifelong Learning Operational Program of the National Strategic Reference Framework (NSRF)—Research Funding Program: Heracleitus II, Investing in knowledge society through the European Social Fund. The GATK3 program was made available through the generosity of the Medical and Population Genetics program at the Broad Institute, Inc. We acknowledge Giuseppe Matullo's contribution as EC's PhD supervisor.

Author contribution

Sample collection and phenotyping: A.E.F., I.N., E.T., J.D., G.D., E.Z. Sequencing Quality Control: A.G., D.S., K.H., E.C. Study design: A.G., D.S., K.K., T.B., E.V.R.A., E.Z. Association analyses: A.G., D.S., L.S. Software development: A.G., D.S. Bioinformatics: A.G., D.S. Selection analysis: R.W., I.M. Phenotype data management: K.K., G.M., B.K., N.W.R., A.B. Downsampling/depth analysis: A.G., M.P. Replication cohort analyses: A.G., KousikK, K.W. Manuscript writing: A.G., R.W., I.M., I.B., E.Z. Project supervision: E.Z.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07070-8.

Competing interests: The authors declare no competing interests.