doi:10.1038/nature14962

Table of Contents

Methods ... 3
Sample collections ... 3
Whole genome - Cohorts ... 3
Whole exome - Rare Disease ... 3
Whole exome - Extreme Obesity ... 3
Whole exome - Neurodevelopmental disorders ... 4
Sequence data production ... 5
Low read-depth whole genome sequencing (UK10K Cohorts dataset) ... 5
Alignment and BAM processing ... 5
Alignment ... 5
BAM improvement and sample file production ... 5
Variant calling ... 6
Filtering ... 6
INDEL pre-filtering ... 6
Site filtering ... 7
Post-genotyping sample QC ... 8
Genotype refinement ... 8
Post-refinement sample QC ... 8
Re-phasing ... 8
Removal of sequencing centre batch effects ... 9
Removal of related and non-European ancestry samples for association analyses ... 9
Data quality evaluation ... 10
Determination of sequencing accuracy against high read-depth exomes ... 10
Determination of sequencing accuracy using MZ twin pairs ... 10
Determination of site overlap with 1000 Genomes Project and GoNL ... 10
Copy-number variation ... 10
Loss of function annotations ... 11
High read-depth whole exome sequencing (rare, neurodevelopmental and obesity disease sets) ... 11
Alignment and BAM processing ... 12
Variant calling ... 12
Detection of cross-sample contamination ... 12
SNV quality control ... 12
Site filtering ... 12
Genotype call filtering ... 13
Further site filtering ... 13
Filtering around INDELs ... 13
INDEL quality control ... 14
Site filtering ... 14
Genotype call filtering ... 14
Further site filtering ... 14
Filtering around INDELs ... 14
Sample selection ... 14
Data analyses in the UK10K Cohorts data set ... 15
Association analyses ... 15
Imputation from the UK10K reference panel ... 15
Genome-wide SNP array data ... 15
The UK10K haplotype reference panel ... 16
Imputation of UK10K data into SNP arrays ... 16
Trait preparation protocols ... 16
Single-marker tests ... 17

WGS data 17
GWA data 17
Large deletions 18
Genome-wide significance threshold in single-marker and genome-wide rare variant tests 18
Follow-up strategies for single-marker associations 18
Rare variant tests 19
Exome-wide analyses 20
Genome-wide analyses 20
Technical validation of rare variant associations through bespoke genotyping 20
Variance explained of WGS statistics 21
Power calculations 21
Relative power for single marker tests 21
Estimation of genotype error rates 22
Relative power for region based tests 22
Analyses of genomic architecture 23
Functional annotation of WGS statistics 23
Genome functional annotation maps 23
GWAS data 24
Population genetic analyses 25
Allele sharing by distance and WTCCC1-defined regions 25
Genotype-phenotype structure within the UK 25
Fine structure of genetic variation 26
Haplotype ages 27
Meta-analysis of single marker association studies in lipid levels 27
Samples 27
In silico GWAS samples from non-UK10K cohorts 27
Association testing 33
Discovery and replication of novel lipid loci using single-marker association tests 33
Locus selection from single-variant tests 33
Data analyses in the UK10K Exome data set 34
Comparison with results from the NHLBI Exome Sequencing Project (ESP) 34
Examination of SNV functional consequences using the Illumina Body Map dataset 35
Assessment of incidental findings in exome data 37
Supplementary References 39

Methods

Sample collections

Whole genome - Cohorts

ALSPAC. The Avon Longitudinal Study of Parents and Children (ALSPAC) is a long-term health research project. More than 14,000 mothers enrolled during pregnancy in 1991 and 1992, and the health and development of their children has been followed in great detail ever since[1 2]. The ALSPAC families have provided a large amount of genetic and environmental information during the course of this longitudinal study. Please note that the study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/).

TwinsUK. The Department of Twin Research and Genetic Epidemiology (DTR) is the UK's only twin registry of 12,000 identical and non-identical twins between the ages of 16 and 85 years [3]. The database is used to study the genetic and environmental aetiology of age-related complex traits and diseases.

For both cohorts, study participants were selected to maximise phenotypic coverage, previous genome-wide array genotyping, coverage with other “-omic” datasets (transcriptomic, metabolomic) and consent to whole genome sequencing, but were otherwise representative of the original population samples.

Whole exome - Rare Disease

Eight different clinical collections were used to provide a total of 1,000 samples to the rare diseases arm of the project; a Severe Insulin Resistance (SIR) sample set, samples for genetic neuromuscular diseases (including congenital muscular dystrophies and congenital myopathies, neurogenic conditions, mitochondrial disorders and periodic paralysis), ocular coloboma samples, and samples from patients with congenital heart disease, ciliopathies and familial hypercholesterolaemia, Samples were also recruited from the Familial Intellectual Disability (FIND) study, and a thyroid disorders cohort of patients with congenital hypothyroidism or resistance to thyroid hormone.

Whole exome - Extreme Obesity

The inclusion criteria for the obesity arm of the project was a body mass index (BMI) ≥ 40 or BMI standard deviations > 3. Participants were drawn from three different collections. The Severe Childhood Onset Obesity Project (SCOOP) is a sub-cohort of the Genetics Of Obesity Study (GOOS) cohort and represents patients with severe early onset obesity in whom known monogenic causes of obesity have been excluded. Samples from obese individuals were also provided by The Generation Scotland: Scottish Family Health Study (GS:SFHS); a family-based genetic study with more than 24,000 volunteers across Scotland, consisting of DNA, clinical and socio-demographic data. Finally, obese individuals belonging to the TwinsUK cohort were also whole-exome sequenced.

Whole exome - Neurodevelopmental disorders

A total of 15 different collections studying either autism or schizophrenia were recruited to the neurodevelopmental arm of the UK10K project, and unless specified were of UK origin. The MUIR collection comprises subjects with schizophrenia, autism, or other psychoses in conjunction with mental retardation (learning disability) and appeared to have a higher than typical rate of familial schizophrenia. The SKUSE sample set consists of clinically identified subjects with Autism Spectrum Disorders (ASD), mostly without intellectual disability (*i.e.* Verbal IQ > 70), and includes both children and adults with Autism, Asperger syndrome or Atypical Autism. The TAMPERE autism sample set is Finnish subjects with ASD and IQ > 70, that were recruited from a clinical centre for the diagnosis and treatment of children with ASD. The BioNED (Biomarkers for Childhood onset neuropsychiatric disorders) collection is children with ASD that have had detailed phenotypic assessments. The MGAS (Molecular Genetics of Autism Study) samples are clinical cases seen by specialists at the Maudsley hospital with detailed phenotypic assessments by ADI-R and ADOS, which measure cognition/adaptive function. There are two schizophrenia datasets from The Finnish Schizophrenia Family Study that were recruited from a population cohort using national registers. The Sub-isolate Schizophrenia Sample (Northeastern sub-isolate) set is from the Kuusamo region where there is a three-fold increased life-time risk for schizophrenia and the Whole Finland Schizophrenia Sample set is from families elsewhere that had at least two affected siblings. The Finnish ASD samples were collected from Central Hospitals across Finland in collaboration with the University of Helsinki, and consist of individuals with a diagnosis of autistic disorder or Asperger syndrome from 36 families with at least two affected individuals. The COLLIER collection is composed of three different studies: The Genetics and Psychosis (GAP) sample set are subjects with new onset schizophrenia; The Maudsley twin series are probands from the Maudsley Twin Register, defined as patients from multiple births who have suffered psychotic symptoms; and finally The Maudsley family study (MFS) contains > 250 families with a history of schizophrenia or bipolar disorder. The UKSCZ cohort contains samples collected from throughout the UK and Ireland, and includes patients either with a positive family history of schizophrenia (collected as sib-pairs or from multiplex kindreds), or samples that were systematically collected in South Wales and had a full diagnostic work up and detailed cognitive testing. The IMGSAC dataset is an international collection of families with children that have ASD. The ABERDEEN sample set comprises cases of schizophrenia with additional cognitive measurements, collected in Aberdeen, Scotland. The GALLAGHER dataset is Irish individuals with autism (approximately 50% with comorbid intellectual disability). Individuals in this cohort have been diagnosed with ADI/ADOS and represent a more severe, narrowly defined cohort of ASD subjects. The EDINBURGH cohort consists of subjects with schizophrenia and IQ > 70, recruited from psychiatric in-patient and outpatient facilities in Scotland. Finally, the GURLING collection consists of multiply affected schizophrenia families all of which are uni-lineal for transmission of schizophrenia, i.e. they have only one affected parent with schizophrenia, or a relative of only one transmitting or obligate carrier parent with schizophrenia.

Sequence data production

Low read-depth whole genome sequencing (UK10K Cohorts dataset)

Low read-depth whole-genome sequencing (WGS) was performed at both the Wellcome Trust Sanger Institute (WTSI) and the Beijing Genomics Institute (BGI). DNA (1-3µg) from lymphoblastoid cell lines (ALSPAC) or PBMCs (TwinsUK) was sheared to 100–1000 bp using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA). Sheared DNA was subjected to Illumina paired-end DNA library preparation. Following size selection (300-500 bp insert size), DNA libraries were sequenced using the Illumina HiSeq platform as paired-end 100 base reads according to manufacturer's protocol.

Alignment and BAM processing

Data generated at the WTSI and BGI were aligned to the human reference separately by the respective centres. The BAM files [4] produced from these alignments were submitted to the European Genome-phenome Archive (EGA). The Vertebrate Resequencing Group at the WTSI then performed further processing.

Alignment

Sequencing reads that failed QC were removed using the Illumina GA Pipeline, and the rest were aligned to the GRCh37 human reference, specifically the reference used in Phase 1 of the 1000 Genomes Project (1000GP) [5] (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz). Reads were aligned using BWA (v0.5.9-r16) [4]. This involved the following steps:

1. Index the reference fasta file:

```
bwa index -a bwtsw <reference_fasta>
```

2. For each fastq file:

```
bwa aln -q 15 -f <sai_file> <reference_fasta> <fastq_file>
```

3. Create SAM files [sam] using bwa sampe for paired-end reads:

```
bwa sampe -f <sam_file> <reference_fasta> <sai_files> <fastq_files>
```

4. Create sorted BAM from SAM. For alignments created at the WTSI this was done using Picard (v1.36) (http://picard.sourceforge.net/) SamFormatConverter and samtools (v0.1.11) sort. For alignments created at the BGI, this was done using samtools (v0.1.8) view and samtools sort.

5. PCR duplicates reads in the WTSI alignments were marked as duplicate using the Picard MarkDuplicates, while in the BGI alignments they were removed using samtools rmdup.

BAM improvement and sample file production

Further processing to improve SNV and INDEL calling, including realignment around known INDELs, base quality score recalibration, addition of BAQ tags, merging and duplicate marking follows that used for Illumina low-coverage data in 1000GP. Software versions used for UK10K for the steps described in that section were GATK version 1.1-5-g6f43284, Picard version 1.64 and samtools version 0.1.16.

Variant calling

SNV and INDEL calls were made using samtools/bcftools (version 0.1.18-r579) [6] by pooling the alignments from 3,910 individual low read-depth BAM files. All-samples and all-sites genotype likelihood files (bcf) were created with the samtools mpileup command

samtools mpileup -EDVSp -C50 -m3 -F0.2 -d 8000 -P ILLUMINA -g –f <reference_fasta>

with the flags

C=Coefficient for downgrading mapping quality for reads containing excessive mismatches

d=At a position, read maximally d reads per input BAM.

Variants were then called using the following bcftools command to produce a VCF file [6]

bcftools view -m 0.9 -vcgN.

For calling on chromosome X and Y, the following settings were applied. The pseudo-autosomal region (PAR) was masked on chromosome Y in the reference fasta file. Male samples were called as diploid in the PAR on chromosome X, and haploid otherwise. Diploid/haploid calls were made using the -s option in bcftools view. The PAR regions were: X-PAR1 (60,001-2,699,520); X-PAR2 (154,931,044-155,260,560); Y-PAR1 (10,001-2,649,520); Y-PAR2 (59,034,050-59,363,566). The pipeline (run-mpileup) used to create the calls is available from https://github.com/VertebrateResequencing/vr-codebase/tree/develop.

Filtering

INDEL pre-filtering

The observation of spikes in the insertion/deletion ratio in sequencing cycles of a subset of the sequencing runs were linked to the appearance of bubbles in the flow cell during sequencing. To counteract this, the following post-calling filtering was applied. The bamcheck utility from the samtools package was used to create a distribution of INDELs per sequencing cycle. Lanes with INDELs predominantly clustered at certain read cycles were marked as problematic, specifically where the highest peak was 5x bigger than the median of the distribution. The list of problematic lanes included 159 samples. In the next step we checked mapped positions of the affected reads to see if they overlapped with called INDELs, which they did for 1,694,630 called sites. The genotypes and genotype likelihoods of affected samples were then set to the reference genotype unless there was a support for the INDEL also in a different, unaffected lane from the same sample. In total, 140,163 genotypes were set back to reference and 135,647 sites were excluded by this procedure. Note that this step was carried out on raw, unfiltered calls prior to VQSR filtering.

Site filtering

Variant Quality Score Recalibration (VQSR) [7] was used to filter sites. For SNVs, the GATK (version 1.3-21) UnifiedGenotyper was used to recall the sites/alleles discovered by samtools in order to generate annotations to be used for recalibration. Recalibration for the INDELs used annotations derived from the built-in samtools annotations. The GATK VariantRecalibrator was then used to model the variants, followed by GATK ApplyRecalibration, which assigns VQSLOD (variant quality score log odds ratio) values to the variants. SNVs and INDELs were modeled separately, with parameters given below:

	SNVs	INDELs
Annotations	QD, DP, FS, MQ, HaplotypeScore, MQRankSum, ReadPosRankSum, InbreedingCoeff	MSD, MDV, MSQ, ICF, DP, SB, VDB
Training set	HapMap 3.3: hapmap_3.3.b37.sites.vcf, Omni 2.5M chip: 1000G_omni2.5.b37.sites.vcf	Mills-Devine [8], 1000 Genomes Phase I
Truth set	HapMap 3.3: hapmap_3.3.b37.sites.vcf	Mills-Devine
Known set	dbSNP build 132: dbsnp_132.b37.vcf	Mills-Devine

The truth set included sites defined as truly showing variation from the reference (GRCh37). VQSLOD scores were calibrated by how many of the truth sites were retained when sites with a VQSLOD score below a given threshold were filtered out. For SNV sites a truth sensitivity of 99.5%, which corresponded to a minimum VQSLOD score of -0.6804, was selected (i.e. for this threshold 99.5% of truth sites were retained). For INDEL sites a truth sensitivity of 97%, which corresponded to a minimum VQSLOD score of 0.5939, was chosen. Post VQSLOD filtering, we also introduced the filter $p<10^{-6}$ to remove sites that failed Hardy-Weinberg equilibrium (HWE, 302,388 sites removed). Finally, logistic regression models were tested for the whole genotype set (N=3,621) where sequencing centre (BGI vs WTSI) and cohort (TwinsUK vs ALSPAC) were fitted as fixed effects. We removed 277,563 sites with evidence for differential frequency between samples sequenced at BGI and WTSI (logistic regression $p<10^{-2}$). After removal of batch effects, we re-computed the pairwise IBS metrics using an LD-pruned genotype set and performed a multidimensional scaling analysis (MDS) on 10 dimensions (PLINK,v1.07), which confirmed the removal of the original structure. We note that the batch-effect correction applied in this case is over-conservative. However, as shown in the QQ plots inflation of summary statistics is well controlled in all tests applied.

The final data set includes the VQSLOD score and other annotations from GATK (BaseQRankSum, Dels, FS, HRun, HaplotypeScore, InbreedingCoeff, MQ0, MQRankSum, QD, ReadPosRankSum, culprit), but it excludes annotations that already existed in or did not apply to the samtools VCFs (DP and MQ, AC, AN). Each VCF further contained the filters LowQual (a low quality variant according to GATK) and MinVQSLOD (variant's VQSLOD score is less than the cut-off). All sites that did not fail these filters were marked as PASS and brought forward to the genotype refinement stage.

Post-genotyping sample QC

Of the 4,030 samples (1,990 TwinsUK and 2,040 ALSPAC) that were submitted for sequencing, 3,910 samples (1,934 TwinsUK and 1,976 ALSPAC) were sequenced and went through the variant calling procedure. Low quality samples were identified before the genotype refinement by comparing the samples to their GWAS genotypes [9] using about 20,000 sites on chromosome 20. Comparing the raw genotype calls to these existing GWAS genotypes data, we removed a total of 112 samples (64 TwinsUK and 48 ALSPAC) because of one or more of the following causes: (i) high overall discordance to GWAS genotype data (>3%) (55 TwinsUK and 36 ALSPAC), (ii) heterozygosity rate > 3SD from population mean (1 TwinsUK and 1 ALSPAC), suggesting possible contamination (iii) no GWAS genotype data available for that sample (7 TwinsUK and 0 ALSPAC) and (iv) sample below 4x mean read-depth (1 TwinsUK and 11 ALSPAC). Overall, 3,798 samples (1,870 TwinsUK and 1,928 ALSPAC) were brought forward to the genotype refinement step.

Genotype refinement

The missing and low confidence genotypes in the filtered VCFs were refined through an imputation procedure with BEAGLE 4, rev909 [10]. The program was run with default parameters. VCFs were split into chunks each containing a maximum of 3,000 sites plus 1,000 sites in buffer regions, that is 500 on either side. Multiallelic sites were included in the imputation. It took 882 CPU weeks to complete. After imputation, chunks were recombined using the vcf-phased-join script from the vcftools [6] package.

Post-refinement sample QC

Additional sample-level QC steps were carried out on refined genotypes, leading to the exclusion of additional 17 samples (16 TwinsUK and 1 ALSPAC) due to one or more of the following causes: (i) post-refinement non-reference discordance (NRD) with GWAS data > 5% (12 TwinsUK and 1 ALSPAC), (ii) multiple relations to other samples, i.e. more than 25 relations with IBS>0.125 were deemed indicative of contamination (13 TwinsUK and 1 ALSPAC), (iii) discordance with manifest gender (3 TwinsUK and 0 ALSPAC). To identify these samples we pruned the WGS data to a set of independent SNVs and calculated genome-wide average identity by state between each pair of samples across the two cohorts. The resulting set of contaminated samples corresponded almost completely to the set of samples with NRD>5%. This left a final set of 3,781 samples (1,854 TwinsUK and 1,927 ALSPAC).

Re-phasing

SHAPEIT2 [11] was then used to rephase the genotype data. The VCF files were converted to binary ped format. Multiallelic and MAF<0.02% (singleton and monomorphic) sites were removed. Files were then split into 3Mbp chunks with +/-250kbp flanking regions. SHAPEIT (v2.r727) was used to rephase the haplotypes with the following command line option in phase mode:

```
--thread 4 --window 0.5 --states 200 --effective-size 11418 -B chr20.$chunk --input-map
genetic_map_chr20_combined_b37.txt --output-log $log --output-max chr20.$chunk.hap.gz
chr20.$chunk.sample
```

vcf-gensample [vcftools] was used to combine the original VCF with new phase information. Sites not rephased with SHAPEIT had any existing phase information removed. vcf-phased-join was used to stitch the chunked VCFs back together with phase determined by matching overlapping heterozygous sites.

These are the final VCF files released for the project and submitted to the EGA (**Table S13**). An imputation reference panel in the IMPUTE2 format created from these VCF files are also made available.

Removal of sequencing centre batch effects

To investigate the presence of batch effects between sequencing centres in the cohorts dataset (WTSI and BGI), we computed pairwise IBS metrics for a joint dataset of 3,621 individuals using an LD-pruned genotype set of 2,203,581 markers and performed a multidimensional scaling analysis (MDS) on 10 dimensions (PLINK,v1.07, options: --indep-pairwise, window size: 5000 SNVs, step: 1000 SNVs, r^2: 0.2; --mds-plot 10). Each sample was labelled by cohort and sequencing centre. Case/control status was assigned to each individual based on their sequencing centre ("BGI" vs "SANGER") and logistic regression models were applied to test for differences in allele frequency between the two centres, with cohort of origin ("ALSPAC" and "TwinsUK") treated as covariate. Here we used the whole genotype set (MAF≥1%, 46,857,518 SNPs). A total of 335,982 SNVs displayed significant association with sequencing centre (p-value ≤ 0.01) and were thus removed from analysis.

Removal of related and non-European ancestry samples for association analyses

The final release set that passed all QC contains non-European and related samples, both of which we sought to exclude to simplify the association testing. To identify participants of non-European ancestry we merged a pruned dataset to the 11 HapMap3 populations [12] and performed a principal components analysis (PCA) using EIGENSTRAT [13]. A total of 44 participants (12 TwinsUK and 32 ALSPAC) did not cluster to the European (CEU) cluster of samples and were removed from association analyses. Individuals with > 20,000 singleton variants (**Extended Figure 1a**) represent ethnic outliers that were excluded from association analyses. We further sought to flag related individuals for exclusion in association tests. Overall, 69 samples (36 TwinsUK and 33 ALSPAC) were flagged because of relatedness greater than third degree (IBS>0.125). Finally 63 co-twin samples (42 dizygotic and 21 monozygotic) and three duplicate samples were removed from TwinsUK. The final sequence data set that was used for the association analyses comprises 3,621 samples (1,754 TwinsUK and 1,867 ALSPAC).

Data quality evaluation

Determination of sequencing accuracy against high read-depth exomes

We retrieved sequence data from 61 individuals from the TwinsUK data set who were both low-coverage whole-genome and high read-depth exome[14] sequenced. Comparisons were carried out restricting the whole genome data to the bait regions that were used for the exome variant calling. The bait regions in these high read-depth sequence data samples covered 35,066,769 bp of sequence, and included a total of 82,998 sites called in the UK10K data. In total 74,621 exome sites (out of the 86,322 sites, or 86.4%) were shared with the UK10K low-genome samples. For the 74,621 shared sites we calculated the percent concordance by allele frequency (AF) bins, as well as the percentage of non-reference discordance (NRD). We also estimated the false discovery rate (FDR = FP / (FP + TP)), where we consider the exomes as the truth set. Additionally we calculated the number of false positives (FP) and the FDR against the exomes split by sites that are found in the 1000 Genomes Project, Phase I (1000GP) and sites that were not found in 1000GP. To estimate the false negative rate (FNR = FN / (FN + TP)) we used AF bins estimated from the 61 exomes (**Extended Data Figure 1**).

Determination of sequencing accuracy using MZ twin pairs

We used 22 monozygotic (MZ) twin pairs from TwinsUK who were all whole-genome sequenced to calculate the mean percentage of concordant genotypes and the mean NRD on chromosome 20 for 910,745 bi-allelic SNV sites by AF (**Extended Data Figure 1** and **Table S2**). On average ~67,000 sites were found to be variable in each twin pair. Out of the 22 MZ twins, 5 pairs were both sequenced at BGI, 8 pairs were both sequenced at the Sanger Institute (SC), and 9 pairs were split between BGI and SC. The overall discordance for bi-allelic SNVs of BGI/BGI pairs was 0.11% with an NRD of 1.52%, for SC/SC pairs the overall discordance was 0.07% with an NRD of 0.97%, and for BGI/SC pairs the discordance was 0.08% with an NRD of 1.14%.

Determination of site overlap with 1000 Genomes Project and GoNL

To search for variant sites that are shared between UK10K Cohorts and the 1000GP or GoNL datasets, only bi-allelic SNVs were taken into consideration. Variant overlap was assessed for allele frequencies (AF) bins calculated separately for the UK10K Cohorts and for the 1000GP/GoNL set, in order to allow comparison between variant discovery in the two datasets pairs. We anticipated at least 95% of the variants with AF > 1% would be found in the 1000GP data set [5].

Copy-number variation

Deletions of size 100bp to 1Mb were identified using the Genome STRiP (version 1.04.1068) software [15]. Genome STRiP default parameters were used for the preprocessing, discovery and genotyping steps. The preprocessing step to generate the metadata for the subsequent discovery and genotyping steps was performed in batches of 100 samples. The metadata was then merged, and the deletion discovery step for each chromosome was carried out in batches of 200 samples to increase sensitivity. The genotyping step was performed on all discovered deletion sites. The deletion discovery and genotyping pipeline was carried separately for deletions size 100bp to 100kb, and 100,001 bp to 1Mb. The GATK SVAnnotator and VariantFiltration tools were used for post-genotyping annotation and filtering. For chromosome X, female and males were genotyped separately. After merging the male and female genotype calls into a single file, the SVAnnotator and VariantFiltration tools were used to annotate and filter the sites. Individual genotypes were annotated as low quality if the genotype quality score (GQ) was <13. The false discovery rate was estimated using the IntensityRankSum (IRS) Annotator tool, which is part of the SVAnnotator framework. SNP array intensities from the Illumina human 610-quad beadchip was used to calculate the IRS p-values for 927 samples, and sites with values ≥ 0.5 were considered false deletions. Only sites that passed all VariantFiltration filters and duplicate removal were used to calculate the FDR. Twenty-eight samples were determined as outliers due to unusually high number of variants per sample, and were excluded from the IRS analysis.

Loss of function annotations

Loss-of-function (LoF) annotation was performed using LOFTEE (Loss-of-function Transcript Effect Estimator, available at https://github.com/konradjk/loftee), a plugin to the Variant Effect Predictor (VEP) [16]. LOFTEE considers all stop-gained, splice-disrupting and frameshift variants, and filters out many known false-positive modes, such as variants near the end of transcripts and in non-canonical splice sites, as described in the code documentation. VEP version 77 (LOFTEE version 0.2) was used with Gencode v19 (basic annotation set) on GRCh37.

Only high-confidence (HC) LoF variants were selected for further analysis, i.e. a LoF variant was predicted as high confidence (HC) if there was one transcript that passes all filters, otherwise it was predicted as low confidence (LC). LC variants were filtered out as well as variants in coding exons with very weak support for coding status from conservation patterns (PhyloCSF [17]). Additionally, putative HC LoF variants were filtered out if their transcripts were not found in GRCh38 or were not annotated as protein coding anymore (see **Table S3** for the number of LoF alleles per gene, the number of homozygous LoF alleles and compound heterozygous alleles per gene).

High read-depth whole exome sequencing (rare, neurodevelopmental and obesity disease sets)

All whole-exome sequencing was performed at the WTSI. DNA (1-3µg) was sheared to 100-400 bp using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA). Sheared DNA underwent Illumina paired-end DNA library preparation and was enriched for target sequences (Agilent Technologies; Human All Exon 50 Mb -- ELID S02972011) according to manufacturer's recommendations (Agilent Technologies; SureSelectXT Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing). The bait regions comprised 204,609 autosomal intervals covering 49.4 Mb, and 8,338 intervals on the sex chromosomes covering 2.1

Mb. Enriched libraries were sequenced using the HiSeq platform (Illumina) as paired-end 75 base reads according to manufacturer's protocol.

Alignment and BAM processing

This was identical to the alignment and BAM processing procedure for UK10K WGS samples sequenced at the WTSI. See above section for details.

Variant calling

Calls were made using samtools/bcftools version 0.1.19-3-g4b70907 from 5,263 UK10K sample BAMs, split by chromosome. A BCF file was created with the following samtools mpileup command, calculating genotype likelihoods for every site in the bait (+/-100bp) regions file:

```
samtools mpileup -EDVSp -C50 -L500 -m3 -F0.2 -d 2000 -P ILLUMINA -l
SureSelect_All_Exon_50mb.hg19.w100.nr.bed –f <reference_fasta>
```

then variants (SNVs and Indels) were called by bcftools with the command

```
bcftools view -Ngvm0.99
```

Detected SNV sites were recalled with the GATK (v1.6-13-g91f02df) UnifiedGenotyper, providing additional quality metrics.

Detection of cross-sample contamination

Two methods were used in combination to assess whether the sample data showed evidence of contamination with another sample. VerifyBamID (version 1.0) [18] was applied to estimate the FREEMIX value. The estimation was made from sequence genotypes at biallelic SNVs in the bait regions that had an alternative allele frequency in European populations >5% in the 1000GP, and these European 1000GP allele frequencies. The second approach was developed in the UK10K project and provides a metric termed “fraction skewed hets”, which is the fraction of heterozygous sites that show a skewed percentage of reads with the alternative allele. All autosomal biallelic SNV sites called as heterozygous with a sample depth of at least 50 were included. Samples were considered contaminated if they had FREEMIX ≥0.03, as recommended for VerifyBamID (http://genome.sph.umich.edu/wiki/VerifyBamID#Reference), and an empirically determined threshold of fraction skewed hets>0.0027. In total 31 samples were removed post-calling because of possible contamination.

SNV quality control

Site filtering

The SnpGap option in vcf-annotate from VCFtools was applied with a window of 10 bp. This treated the start point of an INDEL as the bp after the first base reported in the reference allele and the end point as the bp of the last base reported in the reference allele, and annotated SNVs ≤10 bp from these positions with SnpGap.

The GATK (v1.6-13-g91f02df) VariantRecalibrator was used for filtering. UnifiedGenotyper was run for the sites discovered by SAMtools/BCFtools. Sites below the UnifiedGenotyper minimum calling threshold (4.0) were marked as LowQual. Sites that were not annotated with LowQual or SnpGap were input into VariantRecalibrator and then ApplyRecalibration was used to generate VQSLOD scores. The metrics supplied to VariantRecalibrator were QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, MQ, and InbreedingCoeff. HapMap 3 was used as the truth set and part of the training set, with a prior likelihood of 15.0. The polymorphic sites from the Omni2.5 array were also included in the training set, with a prior likelihood of 12.0. A truth sensitivity of 99.48% corresponding to a minimum VQSLOD score of -1.8768 was chosen. Sites with a lower VQSLOD score, as well as those marked with SnpGap or LowQual, were filtered out.

Genotype call filtering

The impact of filtering genotype calls using sample-level metrics was investigated in a preliminary data release. Called genotypes in 257 samples were compared with data from the Illumina 660W1 BeadChip. 12,862 sites inside the bait regions and 8,052 sites within 100 bp of the bait regions were examined. Calls that matched between the sequencing and array results were assumed to be true calls and we evaluated the drop in the number of false and true calls with increasingly stringent filtering. We determined that genotype quality (GQ), which is the phred-scaled probability of the call being wrong, conditional on the site being variant, was the most effective filter and selected a minimum threshold of GQ=20. Excluding calls with GQ<20, improved the non-reference concordance from 99.2 to 99.8% in the bait regions, whilst only removing 1.7% of true calls. Outside the baits, the starting concordance was lower and a higher proportion of true calls were lost relative to false calls. Calls with GQ<20 were changed to missing. This stage of the SNV filtering corresponds to the release file. Further filtering was applied for downstream analysis.

Further site filtering

Sites outside the bait regions were excluded. After GQ filtering, some sites had missing calls in multiple samples. Sites with missing calls in >10% of samples were filtered out. Sites that were fixed for an alternative allele were removed. Sites on the sex chromosomes were excluded.

Filtering around INDELs

We developed a more comprehensive method for filtering around INDELs. Excluding SNV sites close to INDELs was considered important because clusters of multiple SNVs can be called near INDELs, which are likely to be false calls. This method evaluates each INDEL allele separately and takes into account the surrounding bases from the reference sequence to determine the left-most and right-most possible positions of the allele. The potential position of an INDEL can span tens of base pairs if the region is repetitive. All INDEL alleles reported after removing the results from the contaminated samples were included, before any filtering. SNV sites that were less than 11 bp from the left-most or right-most position of an INDEL allele were excluded (a deletion of a base 10 bp away or closer; an insertion between the bases 10 and 11 bp away or closer).

INDEL quality control

Site filtering

INDEL sites that met any of the following criteria were filtered out: (i) Strand bias p-value<0.0001; (ii) Base quality bias p-value<1e-100; (iii) End distance bias p-value<0.0001; (iv) Depth (for all samples) <16000 or >8000000; (v) Number of alternate bases<2; (vi) Average RMS mapping quality<10; (vii) QUAL<10; (viii) N in the reference.

Genotype call filtering

Calls at INDEL sites with GQ<60, or sample-level depth <4 or >2000 were set to missing. This stage of the INDEL filtering corresponds to the release file. Further filtering was applied for downstream analysis.

Further site filtering

Sites with missing calls in >10% of samples were excluded. Sites that were multiallelic or fixed for an alternative allele were also excluded. A small number of INDEL alleles were reported more than once as different sites with different alternative allele counts. These sites were excluded. Sites on the sex chromosomes and sites outside the bait regions were also removed.

Filtering around INDELs

We filtered out INDELs that were close to other INDELs because this could be an indication of a region that is difficult to align. INDELs with a left-most or right-most position less than 6 bp from the left-most or right-most position of another INDEL allele were filtered out (5 bp or closer when both were deletions or both were insertions; an insertion between the bases 5 and 6 bp away from a deleted base or closer). When INDELs were in close proximity, both were removed. INDELs were compared against all reported INDEL alleles (excluding results from contaminated samples), before any filtering.

Sample selection

We restricted our sample group to a smaller set of unrelated, specific disease cases of UK or Irish origin. To assess ethnicity and relatedness, a subset of SNVs was selected. SNVs were filtered as described above except that sites with missing calls in >1% of samples were excluded and the more comprehensive filtering around INDELS was not performed. Additionally, only biallelic (before GQ filtering) sites, with MAF >5% and $p<10^{-6}$ for HWE (calculated in PLINK v1.07 [19]) were used.

Ethnicity was evaluated by projecting the UK10K samples onto the first two principal components (PC) calculated using the 1000GP samples. Identified relatives were removed from the 1000GP samples, and SNVs were restricted to region P of the strict accessibility mask, genotype imputation quality RSQ >0.9, MAF≥5% and $p<10^{-6}$ for HWE (calculated in PLINK). The overlap of SNVs with the high quality UK10K set was then taken. The remaining SNVs were LD-pruned in the unrelated 1000GP samples with the command --indep 50 5 2 in PLINK, leaving 17,850 SNVs. The PCA and projection was carried out with EIGENSTRAT. A threshold for European origin was defined by a circle in PC1-PC2 space, centred on the mean values of PC1 and PC2 for the CEU, GBR, IBS and TSI 1000GP samples, with radius 1.5 times the maximum distance of these samples to the centre. There were 709 UK10K samples that fell outside this radius and were removed as non-European. An additional 441 samples with records indicating an ethnic origin outside the UK and Ireland were also excluded. Another 409 individuals were removed because they were unaffected or did not have the specified disease phenotype.

Relatedness was estimated in PLINK. SNVs within two known extended regions of high LD (chr 6 25,000,000-35,000,000 bp and chr 8 7,000,000-13,000,000 bp) were removed and the remaining SNVs were LD-pruned with command --indep 200 2 1.5, leaving 18,470 SNVs. Pairwise proportions of alleles identical by descent (IBD) were calculated from these SNVs for all non-contaminated samples. For the samples remaining after the previous filtering, all pairs with IBD>0.125 (corresponding to third-degree relatives) were extracted. When a sample was in more than one pair, all the pairs were clustered into a family group. Each family group or relative pair was then reduced as follows, until no pairs of relatives remained. Firstly, if there were duplicate pairs, defined as a pair with IBD>0.99, the member of the pair with the highest number of missing calls in the LD-pruned SNVs was removed. Then, successively, the sample in the most pairs was removed; if several qualified, the one with the highest number of missing calls in the LD-pruned SNVs was removed. A further 210 samples were removed due to relatedness.

The patient set of unrelated UK and Irish cases comprised 3,463 samples and was released to EGA. There was a final set of 842,646 SNVs in these patients, of which 1.6% were multiallelic, and 6,067 INDELs.

Data analyses in the UK10K Cohorts data set

Association analyses

Imputation from the UK10K reference panel

Genome-wide SNP array data

For association tests, we also considered additional GWA data for each cohort. For ALSPAC, there were another 6,557 samples available, which were measured on Illumina HumanHap550 arrays [20]. For TwinsUK, there were another 2,575 samples that were unrelated to the sequence dataset (IBS>0.125) with genotypes on Illumina HumanHap300 or Illumina Human610 arrays [21]. Both datasets passed QC criteria (gender check, heterozygosity, European ancestry, relatedness (ALSPAC) and zygosity (TwinsUK). Variants discovered through WGS of the TwinsUK and ALSPAC cohorts were imputed into the full GWAS genotyped cohorts increasing the sample size for single point association analysis up to 9,132 subjects.

The UK10K haplotype reference panel

The UK10K final release WGS data of 3,781 samples and 49,826,943 sites was used for creation of a new haplotype reference panel. For each chromosome, a summary file was first generated and merged with that of the 1000GP WGS data to identify multi-allelic sites, sites with inconsistent alleles with that of the 1000GP data, and singletons not existing in 1000GP. These sites were excluded to create a new set of VCF files, leaving 28,615,640 sites. The VCF-QUERY tool was used to convert the new VCF files into phased haplotypes and legend files for IMPUTE2. VCF files were converted to binary ped (bed) format and multi-allelic sites excluded, and files were then split into 3MB chunks with +/-250kb flanking regions. SHAPEIT v2 was used to rephrase the haplotypes. Phasing information from the SHAPEIT output was copied back to the original VCF files, with the phase removed for sites missing due to the MAF cut-off. The phased chunks were then recombined with vcf-phased-join from the vcftools package[6].

Imputation of UK10K data into SNP arrays

Prior to imputation, the two GWAS datasets were pre-phased using SHAPEIT v2 [22] to increase phasing accuracy. SHAPEIT v2 was also used for re-phasing the reference haplotypes provided by the UK10K project. Per the recommendation of the software, the mean size of the windows in which conditioning haplotypes are defined was set to 0.5MB, instead of 2MB used for pre-phasing GWAS. Due to the significantly higher number of variants in the WGS data, the re-phasing was conducted by 3MB chunk with 250kb buffering regions, rather than by whole chromosomes. Imputation was carried out on the same chunks with the same flanking regions as those of the reference panel using standard parameters with IMPUTE2.

Trait preparation protocols

All traits were available from previous studies. Information on trait measurements is summarised in **Table S18**, while **Table S4** contains trait specific details of phenotype preparation steps for analysis. For ALSPAC, we combined the sequenced and imputed samples before phenotype preparation. Traits were residualised on associated covariates using the following steps. First checks were undertaken to examine differences in trait distribution by males and females. Where different, data was handled in a sex specific manner from this point. Outliers greater than 5 SD were manually checked for data entry errors. Outliers greater than 3 or 5 SD (depending on trait) from the mean were removed and data then transformed to obtain a normal distribution using either an inverse normal, log or square-root transformation. Potential covariates were tested for association with phenotypes in the full ALSPAC data set and in sequenced and imputed TwinsUK datasets separately. Adjustment for covariates was only undertaken given evidence of association between covariate and phenotype. Technical covariates (instrument, assay, date of visit) were fitted as random effect (TwinsUK only) whereas other covariates (age, gender) were fitted as fixed effect. Traits were residualised on associated covariates to generate standardised residuals with a mean of zero and a SD of 1. Where a sex-specific transformation was used, females and males were standardised separately before being combined.

Single-marker tests

WGS data

We performed a single-variant analysis for each trait using the UK10K sequence data as a 'discovery' panel and the imputed data as a 'replication' panel. For 31 overlapping traits (core traits) with phenotype data in both ALSPAC and TwinsUK, we used the meta-analysis of the two cohorts for discovery ('Total WGS'). Of note, because of phenotype missingness the total number of samples available for each trait is smaller than 3,621, with the correct values given in **Table S1** for both the WGS and GWA samples.

The analysis software SNPTEST v 2.4.0 was used for analysing the discovery set employing an additive model within a frequentist test [23]. For each trait residual y_i and genotypes x_i a linear model $y_i = \beta_0 + \beta_1 x_i$ was fitted, for $i = 1,2,\dots,n$, where *n* is the number of samples.

To account for the genotype uncertainty that might arise from sequencing, we used genotype dosages, where each genotype was expressed on a quantitative scale between [0:2] (using in SNPTEST the function -method expected). We also used the option -use_raw_phenotypes to disable the default quantile normalization since the phenotype residuals were already standardised. Given the weak statistical power for rare variants we chose to exclude the variants that did not pass a low allele frequency threshold (MAF<0.1%). Meta-analyses of ALSPAC and TwinsUK sequence data were performed using GWAMA v 2.1 [24] where we assumed a fixed effect model and where we used genomic control to adjust the summary statistics for both input and output data. GWAMA calculates the combined allelic effect B_j across all studies at the *j*-th SNV as

$$B_j = \frac{\sum_{i=1}^{N} \beta_{ij} w_{ij}}{\sum_{i=1}^{N} w_{ij}}$$

where β_{ij} represents the strand-aligned effect of the reference allele at the *j*-th SNV in the *i*-th study and w_{ij} represents the inverse of the variance of the estimated allelic effect. The combined variance is given by $V_j = \left(\sum_{i=1}^{N} w_{ij}\right)^{-1}$.

GWA data

We attempted to replicate the genome-wide and suggestive signals using the GWA panel (variants imputed to the UK10K reference panel described earlier). For ALSPAC replication analysis comprising unrelated individuals only, we used the same SNPTEST approach and parameters. However the imputed TwinsUK data, although unrelated to the WGS set, contained related individuals (mainly co-twins), which amount to family structure in the replication panel. Therefore for TwinsUK replication, we employed GEMMA v0.92 [25], which is a standard linear mixed model for single-variant association while controlling for such structure. The algorithm requires pre-computation of a kinship matrix, which we estimated using its centered genotypes model. Additionally we carried out a 4-way meta-analysis across the four panels (TwinsUK/ALSPAC, WGS+GWA sample). For this, we used GWAMA with the options described above.

Large deletions

Similar to the single-variant analysis described above, we tested associations of large deletions and the 31 core traits using the software SNPTEST v2.4.0. Additionally we performed a 2-way meta-analysis with GWAMA v2.1 across the TwinsUK and ALSPAC WGS panels. None of the deletions reached the experiment-wide threshold of 4.6×10^{-10}; the results for the 31 core traits are given in **Table S19**.

Genome-wide significance threshold in single-marker and genome-wide rare variant tests

The necessary significance thresholds for family-wise control of type 1 error, for low frequency and rare variants and single-marker analysis, were estimated using the method of Xu et al. [26], at several different minor allele frequency thresholds. Furthermore, by examining the correlations between the 31 core phenotypes, we estimated that there were approximately 18 independent phenotypes [27] and used this correction factor when estimating the required thresholds in **Figure 3**. The resulting genome-wide significance cut-offs were p-value ≤ 4.6×10^{-10} (=$8.31 \times 10^{-9}/18$) for both single-marker and genome-wide rare variant tests and p-value ≤ 1.97×10^{-7} (=$3.55 \times 10^{-6}/18$) for exome-wide tests.

Follow-up strategies for single-marker associations

Selection of associated variants. For single-marker associations, we first identified all variants meeting the genome-wide significance cut-off (p-value ≤ 4.6×10^{-10}) from either the WGS or WGS+GWA meta-analysis. For suggestive associations, an arbitrary threshold of 1×10^{-5} was applied to select significant results for the analysis of single marker association statistics across the 31 core phenotypes. This threshold was chosen to select a reasonable number of SNPs for replication and validation, and also since several of the phenotype-specific QQ-plots showed some evidence of a change-point at approximately this threshold. Suggestive evidence for validation was based on a meta-analysis p-value of sequenced and imputed samples (WGS+GWA) that was at least 1×10^{-7}, i.e. two orders of magnitude smaller than the original threshold. False discovery rates at these thresholds were calculated using the method of Benjamini and Hochberg [28] for each phenotype. Finally, for all variants meeting these two significance cut-offs, we assessed independence between variants, and independence from known variants reported in the literature, as follows.

Annotation of index variants for previously reported loci. For each trait we compiled a list of known loci by selecting all SNPs associated with a trait of interest from the NHGRI GWAS catalog (p-value ≤ 5×10^{-8}, last updated in May 2014), supplemented by manual curation of all associations reported in the literature reaching the same genome-wide significance cut-off. Only index variants with a marginal significance in the UK10K single-marker association statistics (p-value≤0.05) were considered for conditional tests. Where a region contained multiple correlated index variants associated with a given trait in the GWAS catalog, we clumped the set of index variants to remove highly correlated ones (using a LD metric r^2>0.8 applied to within a 2Mb sliding window from each known index SNP (+/-1Mb)). This avoids collinearity errors when a variant is conditioned against multiple correlated index variants.

***Clumping of UK10K summary statistics*.** We next applied a clumping procedure to thin the list of variants associated with each traits, assigning sets of variants to discrete LD bins if their pairwise metrics r^2 was ≥ 0.2. For each LD bin, the variant most associated with the trait in question was retained for assessment in conditional analyses. Index variants for previously reported loci that mapped to within +/- 1Mb of an index variant for a known locus were also annotated.

***Conditional analyses*.** Finally, sequential conditional single-variant association analyses were carried out to confirm statistical independence between associations. In the initial round of conditional analysis, associations of SNVs with the respective quantitative trait were conditioned on the index variants for known loci clumped (r^2>0.8) as described before (this step was carried out only for SNVs within +/-1Mb of a known locus); in further rounds, associations were conditioned against all nearby known loci plus the best novel UK10K variant identified in the previous round of conditional analysis. The conditional analysis was tested independently for each WGS cohort (TwinsUK and ALSPAC), and a meta-analysis was conducted at the end of each round until the conditional association p-value was no longer significant (p-value>10^{-5}). A variant was considered independent if it has a conditional p-value ≤ 10^{-5} (corresponding to r^2<0.2 in our data) or has a p-value difference between conditional and unconditional analysis of magnitude less than 2.

Finally, variants were classified as **known** (denoting either a known variant, or a variant for which the association signal disappears after conditioning on the known locus) or **novel** (denoted as variant which still is conditionally independent on known loci, and on eventual other novel independent signals in that region). For novel signals, the variant with the lowest conditional p-value between multiple associated variants was reported.

Rare variant tests

Sequence Kernel Association Tests (SKAT[29] and SKAT-O[30]) were used to sequence data to investigate the aggregated effect of multiple rare variants on each trait. SKAT is a variance-component multiple regression test which retains power in settings where neutral variants or variants with opposite direction of effects could result in loss of power. SKAT-O represents the best linear combination of SKAT and burden tests, which is supposed to maximize power. For the analyses we used SKAT v0.93 with default parameters including Beta(1,25) weights and ‘linear.weighted’ kernel. P-values for SKAT and SKAT-O tests were computed using the methods 'davies' and 'optimal.adj' respectively. For the meta-analyses of summary statistics we used MetaSKAT v0.33 with default options [30].

Exome-wide analyses

We included all rare variants with MAF<1% which fell into coding exons, splice sites or UTR regions of known genes (35,796 genes on autosomes and chromosome X) based on GENCODEv15[31], only some types of pseudogenes (IG_C, IG_J, IG_V, TR_J, TR_V) were removed. Overlapping exons (470,312 exons) were merged within each gene, resulting in 254,530 exonic regions.

Three different exome-wide SKAT analyses were carried out: (i) naïve tests of all exonic variants, splice sites and variants residing in UTR; (ii) functional tests of LoF and missense variants and (iii) LoF variant tests (**Table S9**).

(i) In the naïve approach windows were generated by keeping the exon structure intact as far as possible and allowing between ~5 and ~50 variants per window. If there were less than 5 variants within an exon or more than 50 variants per gene, then windows were created by combining neighbouring exons so that the number of variants was similar between windows. Additionally we added a layer of windows on top by tiling across the concatenated exons with maximal ~50 variants per window but starting halfway into the first window that was generated by the initial approach. In the naïve approach we tested 26,226 genes in 50,717 windows with 35 variants per window on average, and a minimum of 2 and a maximum of 135 variants.

(ii) In the *functional* approach we tested all LoF and missense variants together within each gene. In total we analysed 14,909 genes with 17 variants per gene on average, and at least 5 variants and maximal 1,058 variants per gene.

(iii) In the LoF approach we tested only LoF variants (stop gained, splice-disrupting and frameshift variants). In total we analysed 3,208 genes with 2 variants per gene on average, and minimal 2 and maximal 17 variants per gene.

Genome-wide analyses

We partitioned the genome into 3kb half-overlapping tiling windows with an average of 37 variants per window. The same SKAT parameters and allele frequency cut-off were applied as above. In total, for each trait, we generated SKAT and SKAT-O p-values and their corresponding MetaSKAT p-values for more than 1.8 million windows across the genome.

Technical validation of rare variant associations through bespoke genotyping

We evaluated a random sample of 9 rare SNVs selected to be strongly associated in the WGS sample (MAF<1%, *p-value* < 10^{-5}), but with no evidence of replication in the GWA sample (*p-value* < 10^{-7}) through bespoke validation genotyping independently of the UK10K data production pipeline. For ALSPAC, the entire cohort (10,145 young participants with DNA samples immediately available) was genotyped using KASP™ at KBioscience (www.lgcgenomics.com/). Assays are based on competitive allele-specific PCR and enable bi-allelic scoring of SNVs. The SNP-specific KASP Assay mix and the universal KASP Master mix are added to DNA samples, a thermal cycling reaction is then performed, followed by an end-point fluorescent read. Bi-allelic discrimination is achieved through the competitive binding of two allele-specific forward primers, each with a unique tail sequence that corresponds with two universal FRET (fluorescence resonant energy transfer) cassettes; one labelled with FAMTM dye and the other with HEXTM dye. This effort demonstrated good technical validation of WGS genotypes (**Table S6**), and validation of GWA genotypes (albeit to a separate degree). This suggests that false positives of initial discoveries, rather than poor imputation, may account for the lack in replication.

Variance explained of WGS statistics

We used the Restricted Maximum Likelihood (REML) [32] method implemented in GCTA (http://www.complextraitgenomics.com/software/gcta/reml.html) to estimate phenotypic variance explained by SNV sets in our discovery sequence data. Four genome-wide sets of autosomal QC'd SNVs in 3,621 individuals were considered from different reference panels: HapMap2 (Variant N=2,331,713), Hapmap3 (N=1,168,695), 1000 Genomes (N=7,475,230) and the entire UK10K reference panel (N=8,317,582), which was each used to generate a single genetic relationship matrix (GRM). Only SNVs with minor allele frequencies > 1% were considered. Each GRM was individually tested against the 31 traits with phenotypic values present in both cohort studies, producing a beta, se and p-value for total trait variance explained by the given SNV set.

Power calculations

Relative power for single marker tests

Let N be the sample size, p be the minor allele frequency, β represent the standardised effect of a SNV on a continuous phenotype (standardised so that β is the effect per standard deviation of the phenotype), and let r^2 represent the square of the correlation between a true genotype and a genotype measured with error. The non-centrality parameter of the chi-squared distribution for a single SNV has been shown to be $NCP = 2(N-1)p(1-p)\beta^2 r^2$ [33,34]. Using a significance threshold of 4.62x10^{-10} (a genome-wide threshold of 6.7x10^{-8} that takes into account the large number of variants identified by WGS [35] divided by an effective number of 18 independent traits, calculated by using correlations between all phenotypes [27] we calculated the smallest β detectable at 80% power for a range of values of p and r^2, and for sample sizes corresponding to UK10K (N = 3,621) as well as larger sample sizes of N = 10,000, 20,000 and 50,000. Using chromosome 20 data from UK10K, we obtained the observed r^2 values for the 300,489 variants on chromosome 20 that could be imputed using both the 1000GP and 1000GP+UK10K reference databases [36].

We then averaged the corresponding smallest detectable β values across the 300,489 variants for the two imputation schemes, for minor allele frequencies binned into (0 – 0.0025], (0.0025-0.0075], (0.0075-0.02], (0.02 – 0.10]. In **Figure 4a**, the data for these four bins are displayed at MAF=(0.001, 0.005, 0.01, and 0.05).

Estimation of genotype error rates

Genotypes obtained for the same individuals from two sources were compared and agreement was measured by r^2 **(Table S11).** Three different comparisons were performed: (1) WES versus WGS, for exons only in N=92 individuals, (2) WGS versus GWAS plus imputation (using IMPUTE2) of the GWAS data against 1000Genomes data N=1672, and (3) WGS versus GWAS plus imputation against a combined reference set from the 1000Genomes and UK10K (N=59 individuals). Note that the 59 individuals for comparison (3) were included in the reference set, and therefore the agreement will be overly optimistic in this case.

Relative power for region based tests

Power for the SKAT rare variant tests [29,30] was calculated by assuming a causal model for the relationship between the SNVs and the phenotype. Power is calculated by averaging the results of the region-based tests across several randomly selected windows of a desired size. We modified the authors' original approach to accept both true haplotype data and haplotype data where errors had been introduced. Errors were randomly introduced into the haplotypes following the observed r^2 distributions between genotypes assayed on the same individuals via two methods. Specifically:

1. 10 regions were randomly selected on each autosome, for a total of 220 regions each containing 30 SNVs. Haplotypes for these regions were extracted from the estimated haplotypes in the UK10K WGS data, with sample size of N = 3,621 Haplotypes had been previously estimated with Beagle[37].

2. Within each of the 220 regions, genotype errors were randomly introduced via simulation, following the distributions in **Table S11.** For a marker with MAF in a particular bin, we randomly sampled an r^2 bin from the MAF-specific conditional distribution, and generated a genotype with error using this value of r^2. Bin midpoints were used for generation of genotypes.

3. The SKAT power program was modified to accept both the true genotype data and the data with errors at each of the 220 regions, and power was calculated as follows: The true model is $y = \alpha + G_0\beta + \epsilon$ where G_0 (size nxp) is the true data without error. However the model that is fit in the SKAT test is $y = \alpha + G_e\beta + \epsilon$ where G_e are the genotype data containing errors. Following through the development in the supplement of [30], we can write $Q = (y - \bar{y}1)^T K_e\,(y - \bar{y}1)$, where the phenotype y depends on the true causal variant genotypes, but the kernel K_e used in the test statistic is measured with genotypes containing errors. Then, we can write $(y - \bar{y}1) = (E + \mu_0)$ where E is an independent Gaussian random variable and $\mu_0 = G_0\beta$ depends on the true genotypes. We finally obtain $A = E(\frac{G_e^T G_e W}{n})$ and $B = E(\frac{G_e^T \mu_0 \mu_0^T G_e W}{n^2})$, where W is a matrix of weights.

4. The following parameters were used in the simulations:

   a. Causal variants could only be variants with MAF less than a chosen threshold, and this threshold was varied from 0.001 to 0.10.

   b. The maximum effect of a causal variant ranged from 2.0 to 4.0 (the standardised effect size), and this is referred to as MaxBeta in the Results and Figures. This maximum would apply to variants with MAF of 0.001 or even more rare; for variants with higher MAF values, the effect of a causal variant was smaller and decreased as MAF increased.

   c. The percentage of causal variants in the window was set to either 20% or 5%.

   d. 20% of variants decrease phenotype, 80% increase phenotype.

   e. The significance threshold was set to $\alpha = 6.7e - 08$ [26], an estimated genome-wide threshold for region-based tests of rare genetic variation with SKAT.

5. Power was then averaged across the 220 regions and across 500 different simulations. This was performed for the data without error and then three additional times, using each of the sub-tables in **Table S11**, one looking at the benefit of high-depth WES sequencing compared to WGS (only for regions and individuals sequenced by both methods), once comparing WGS to GWAS plus imputation against 1000GP data, and once comparing WGS to GWAS plus imputation against the combined data from the 1000GP and the UK10K WGS data.

Analyses of genomic architecture

Functional annotation of WGS statistics

The genome-wide single-variant analysis *p*-value distributions of 5 UK10K lipid traits were used in order to assess whether we observe enrichment in various functional and regulatory features - using genic annotations, chromatin states, DNaseI hypersensitive sites, transcription factor (TF) binding sites, conservation scores and histone modifications. To do this, a set of independent SNVs for each phenotype was selected and annotated such that each variant was said to have a certain feature if it itself or if any of its LD proxies fell into an appropriate region. A "n-fold" enrichment score was computed to quantify the observed enrichment at various GWAS p-value cut-offs and finally permutations were performed matching on MAF, distance to nearest TSS and number of LD proxies in order to assess the significance of the observed enrichment. The matching is introduced to control for genomic features of the data, which could otherwise lead to biased results.

Genome functional annotation maps

a) Genic annotations were obtained from GENCODEv13 and variants were split into categories defined as follows: Intron, Exon (coding), 3'UTR, 5'UTR, Downstream genetic variant (within 5KB of the end of a gene), Upstream genetic variant (within 5KB of the start of a gene). Insufficient datapoints were available for 5'UTRs, which were excluded from further analysis.

b) DNaseI data was obtained from ENCODE [38] (ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011) on the liver HepG2 cell line for DHS hotspots. DHS data was processed following DHS data processing protocol described in an ENCODE study [39].

c) Six chromatin states (Enhancer, Transcribed, CTCF, Promoter, TSS and Repressed) were obtained from ENCODE in the HepG2 cell line, informative for lipids measurements.

GWAS data

Data Processing

To remove possible biases due to linkage disequilibrium (LD) or dependence between variants we compute the r^2 between all SNVs within 1Mb windows and consider r^2 of less than 0.1 between two variants to mean (approximate) independence. Next, from the full set of genetic variants for each phenotype, we create an independent set of SNVs, where in order to keep all possible GWAS signals we sequentially find and retain the next most significant (lowest P) variant independent of all other variants in our independence set. Then we annotate each independent SNV and consider it as overlapping a functional element if (1) the SNV itself resides in such a genomic region or (2) at least one of its proxies in LD ($r^2 \geq 0.8$) and within 500Kb with it does. We include the latter as the association of a SNV in GWAS potentially tags the effect of other variants, which could underlie the observed association signal.

Quantifying enrichment

To find the enrichment of GWAS signals within a given annotation, we calculate fold enrichment as the fraction of variants that fall in that annotation and have p less than threshold t, divided by the fraction of total number of variants in that annotation.

Specifically,

$$Fold\ Enrichment(t) = \frac{N_a^t}{N^t} \Big/ \frac{N_a}{N}$$

where N denotes the total number of variants, N_a - the total number of variants that fall in the annotation of interest, N^t - the total number of variants with p less than threshold t, $N_a^{\square}$ - the number of variants with p less than t that fall in the annotation of interest.

Statistical testing

We consider that perhaps variants with specific annotations may be more likely to be at specific positions of the genome. To account for this we use permutation testing, where we shuffle the p associated to each variant in our independence set in such a way as to match SNVs according to MAF, distance to nearest TSS and number of LD proxies (r^2>=0.8) they have. Specifically, we bin variants according to 10 quantiles of MAF and number of LD proxies and 7 quantiles of distance to nearest TSS, resulting in 700 bins overall. We then permute variants within each bin separately.

Multiple testing

We test at 95% significance level and correct for multiple testing by applying a Bonferroni correction for the effective number of distinct annotation used, which we determine by adapting an approach proposed in Galwey, 2009 [40].

Enrichment differences between MAF-classes

Additionally to the general enrichment analyses carried out, we split our variants into sets of low MAF (1-5%) and common (>5%) and performed the same enrichment analyses in both classes separately to look for differences in the levels of enrichment.

Population genetic analyses

Allele sharing by distance and WTCCC1-defined regions

Place of birth was available for 1,139 individuals from TwinsUK from which it was possible to calculate mean latitude longitude for further analyses. We searched for excess allele sharing by distance as suggested by Mathieson and McVean [41]. However we adapted their method, which divides the number of shared alleles at a given distance first by the total number of individuals at that distance and then by the allele frequency to get the excess allele-sharing probability. We found that their method was biased if the allele frequency was very low, since it only counts events where at least one rare allele was observed. In contrast our method simply divides the number of observed shared alleles by the number of expected alleles which is the total number of observed alleles multiplied by the proportion of samples in each distance bin. This analysis was carried out for two types of distances (i) pairwise Euclidean distance between the geographic location of all sample pairs (**Box 1**) and (ii) distance to Bristol but excluding individuals from London.

We also compared the mean shared alleles (iii) within and between counties of origin (not shown) and (iv) within and between regions of origin (based on WTCCC regions) for allele counts (AC) from two to seven with AC values derived on the whole data set of 3,781 samples (**Extended Data Figure 8**). The mean shared alleles were calculated as the number of shared alleles between two samples divided by all possible sample pairs, which is $N \times (N-1)/2$ for N individuals within a region and $N \times M$ between two regions for N individuals in one region and M individuals in the other region.

Genotype-phenotype structure within the UK

In order to assess potential confounding effects of phenotypes and genotypes by geographic location, we used the mean latitude and mean longitude of the county of origin for 1,139 samples from TwinsUK. We computed (i) the pairwise Euclidean distance of the geographic location (latitude and longitude) of the samples (distance matrix), (ii) the pairwise absolute difference between the normalised residuals of the phenotypes (phenotype matrix) (**Extended Data Figure 8**) and (iii) the pairwise number of variants shared for allele counts from 2 to 7 (genotype matrices) where for each allele count a separate matrix has been generated. Then we carried out Mantel tests [42] to test the correlation between the phenotype and the genotype matrices, but also between the distance and the phenotype, and between the distance and the genotype matrices (not shown).

Fine structure of genetic variation

1,139 unrelated individuals within the UK10K project have both sequence data and geographical information available. We then excluded singletons and doubletons and thinned the dataset using PLINK [19] (with the option --indep-pairwise 50 5 0.002) to obtain a working and computationally manageable sample of 6,076,814 SNVs. To explore fine structure within these samples, the data were processed with ChromoPainter [43], first estimating the parameter values via Expectation Maximisation and then "painting" all individuals against all other individuals for all autosomal chromosomes. This painting is combined into the coancestry matrix, which counts most-recent shared recombination events. We display the average length of DNA tracts shared between each population (called chunks), obtained by dividing the total recombination distance shared between populations (the sum of chunk lengths) by the number of chunks.

FineSTRUCTURE [43] was run for 4 Million iterations with the final 1M (thinned to 1000) used to form a posterior over population assignments, with convergence checked by visual comparison of the pairwise coincidence matrix between two independent runs. Stochastic optimization was used to obtain the maximum aposteriori solution, displayed. Geographic regions were taken from the UK10K dataset and the data first analysed with the complete set of regional labels. **Extended Data Figure 9** shows the FineSTRUCTURE MCMC results without simplifying the population. The coancestry matrix and chunk length matrix for all geographical regions and all inferred populations are shown, as is the detailed population breakdown and relationship between populations in the MCMC sample. Because there is very little variation within 'Eastern', 'Southern', 'North Midland', 'Southeast' and 'London' labels we merged these into 'South & East' for **Box 1**. Additionally, small populations inferred by FineSTRUCTURE were merged into larger similar populations, chosen if they were deemed `similar' by the FineSTRUCTURE dendrogram. **Extended Data Figure 9c** provides a mapping from FineSTRUCTURE inferred populations to populations in **Box 1**.

Admixture index: We treated the 8 populations as an admixture cline in the order shown in **Figure Box 1c**. We first computed the proportion of individuals in each inferred population from each region (**Figure Box 1c**). We then computed the variance of these proportions when the admixture cline is seen as 8 evenly spaced populations on the [0,1] interval, and reported the "admixture index" as (variance expected under a uniform distribution)/(observed variance). This takes the value 1 when the observed distribution is uniform; infinity when all individuals are found in a theoretical 'central' population, and 1/[12p(1-p)] when individuals are observed with frequency p in population 1 and frequency (1-p) in population 8. Empirically, p is close to 0.5 for all populations so the minimum value is 1/3.

Haplotype ages

For estimating the distribution of times to the most recent common ancestors of recent haplotype pairs, we applied a novel method [44] that focuses on haplotypes surrounding rare variants of allele count 2 in all of the TwinsUK samples. Any such rare variant defines a pair of haplotypes that necessarily coalesce with each other before any of them coalesces with any other haplotype in the population. This assumes that each variant is caused by a single mutation and neglects double-mutations. The method explicitly models the mutation and recombination process in the variant-surrounding haplotypes to estimate their age. The result is a distribution of f2 haplotype ages within and between subgroups, which we defined by the WTCCC regions previously defined and the median age estimates as a summary statistics for each age distribution within and between each region are shown in **Extended Data Figure 9**.

Meta-analysis of single marker association studies in lipid levels

Samples

In silico GWAS samples from non-UK10K cohorts

1958 Birth Cohort. Participants to the cohort have been followed-up regularly since birth with prospective information collected on a wide range of indicators related to health, health behaviour, lifestyle, growth and development. There have been 9 contacts with the participants since their birth (ages 7, 11, 16, 23, 33, 41, 45, 47, and 50 years). The biomedical survey at age 45 years included collection of blood samples and DNA from about 8000 participants. The survey was approved by the South East multicentre research ethics committee (MREC). There was an informed consent process conducted by the National Centre for Social Research [45]. **Lipid measurements:** Venous blood samples were obtained without prior fasting; participants could choose whether to sit or lie down when blood was taken. Serum triglycerides, and total and HDL-cholesterol were measured in serum by Olympus model AU640 autoanalyser in a central lab in Newcastle. Enzymatic colorimetric determination GPO-PAP method was used to determine triglycerides, CHOD-PAP method for total cholesterol and for HDL-cholesterol. Blood samples were stored in a -80 C freezer.

INGI-Val Borbera. The INGI-Val Borbera population is a collection of 1,785 genotyped samples collected in the Val Borbera Valley, a geographically isolated valley located within the Appennine Mountains in Northwest Italy [46]. The valley is inhabited by about 3,000 descendants from the original population, living in 7 villages along the valley and in the mountains. Participants were healthy people 18-102 years of age that had at least one grandfather living in the valley. A standard battery of tests were performed by the laboratory of ASL 22 - Novi Ligure (AL), on sera from fasting blood collected in the morning. The project was approved by the Ethical committee of the San Raffaele Hospital and of the Piemonte Region. All participants signed an informed consent. **Lipid measurements:** Lipids were measured using HITACHI 917 ROCHE and Unicel Dx-C 800 BECKMAN devices.

INGI FVG. The INGI Friuli Venezia Giulia (FVG) cohort comprised of about 1700 samples from six isolated villages covering a total area of 7858 km^2 in a hilly part of Friuli-Venezia Giulia (FVG) county located in north-eastern Italy. A recent study [47] characterized this population as a genetic isolate with high level of genomic homozygosity and elevated linkage disequilibrium. The cohort accounts for 1590 genotyped samples. Participants were randomly selected people 3-92 years of age. Genotyping and phenotypic data for 1590 samples are available. People with age < 18 were excluded from analyses. A written informed consent for participation was obtained from all subjects. The project was approved by the Ethical committee of the IRCCS Burlo-Garofolo. **Lipid measurements:** Lipids were measured using BIOTECNICA BT-3000 TARGA chemistry analyser.

INGI Carlantino. Carlantino is a small village in the Province of Foggia in southern Italy. Genetic analyses of chromosome Y haplotypes as well as mitochondrial DNA show that Carlantino is a genetically homogeneous population and not only a geographically isolated village [47]. Participants were randomly selected in a range of 15 – 90 years of age. Genotyping and phenotypic data are available for 630 individuals. People with age < 18 were excluded from analyses. The local administration of Carlantino, the Health Service of Foggia Province, Italy, and ethical committee of the IRCCS Burlo-Garofolo of Trieste approved the project. Written informed consent was obtained from every participant to the study. **Lipid measurements:** Lipids were measured using a BIOTECNICA BT-3000 TARGA chemistry analyser.

INCIPE. For the INCIPE study, 6200 randomly chosen individuals, all Caucasians and at least 40 years of age as of 1 January 2006, received a letter inviting them to participate in the study. A total of 3870 subjects (62%) accepted and were enrolled. Two studies were included in the analysis:

1. INCIPE1: Individuals genotyped on Affymetrix 500k
2. INCIPE2: Individuals genotyped on HumanCoreExome-12v1

The ethics committees of the involved institutions approved the study protocol. **Lipid measurements:** Enzymatic determination of cholesterol and triglycerides was performed on Dimension RxL apparatus (Siemens Diagnostics). HDL cholesterol was determined by the homogeneous method; LDL cholesterol by the Friedewald equation [48].

The **Ludwigshafen Risk and Cardiovascular Health (LURIC) study.** The LURIC study is a prospective study of more than 3,300 individuals of German ancestry in whom cardiovascular and metabolic phenotypes (CAD, MI, dyslipidaemia, hypertension, metabolic syndrome and diabetes mellitus) have been defined or ruled out using standardised methodologies in all study completed participants. A 10-year clinical follow-up for total and cause specific mortality has been completed. [49] From 1997 to 2002 about 3,800 patients were recruited at the Heart Center of Ludwigshafen (Rhein). Inclusion criteria were: German ancestry, clinical stability (except for acute coronary syndromes) and existence of a coronary angiogram. Exclusion criteria were: any acute illness other than acute coronary syndromes, any chronic disease where non-cardiac disease predominated and a history of malignancy within the last five years. The study was approved by the ethics review committee at the Landesärztekammer Rheinland-Pfalz in Mainz, Germany, and written informed consent was obtained from the participants. **Lipid measurements:** Total cholesterol and triglycerides were obtained by ß-quantification from serum and measured enzymatically using WAKO reagents on a WAKO 30R analyser (Neuss, Germany). LDL-C, HDL-C and apolipoprotein B were measured after separating lipoproteins with a combined ultracentrifugation-precipitation method.

HELIC MANOLIS. The HELIC (Hellenic Isolated Cohorts; www.helic.org) MANOLIS (Minoan Isolates) collection focuses on Anogia and surrounding Mylopotamos villages. Recruitment of this population-based sample was primarily carried out at the village medical centres. All individuals were older than 17 years and had to have at least one parent from the Mylopotamos area. 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. The study was approved by the Harokopio University Bioethics Committee and informed consent was obtained from every participant. **Lipid measurements.** Lipid traits were assessed using enzymatic colorimetric assays and included; total cholesterol (cholesterol oxidase - phenol aminophenazone method), high density lipoprotein (HDL)-cholesterol and triglycerides (glycerol-3-phosphate oxidase -phenol aminophenazone). Low Density Lipoprotein (LDL)-cholesterol levels were calculated according to the Friedewald equation [48].

HELIC Pomak. The HELIC (Hellenic Isolated Cohorts; www.helic.org) Pomak collection focuses on the Pomak villages, a set of isolated mountainous villages in the North of Greece. Recruitment of this population-based sample 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. The study was approved by the Harokopio University Bioethics Committee and informed consent was obtained from every participant. **Lipid measurements.** Lipid traits were assessed using enzymatic colorimetric assays and included; total cholesterol (cholesterol oxidase - phenol aminophenazone method), high density lipoprotein (HDL)-cholesterol and triglycerides (glycerol-3-phosphate oxidase -phenol aminophenazone). Low Density Lipoprotein (LDL)-cholesterol levels were calculated according to the Friedewald equation [48].

TEENAGE (TEENs of Attica: Genes and Environment). Participants were drawn from the TEENAGE (TEENs of Attica: Genes and Environment) study. A random sample of 857 adolescent students attending public secondary schools located in the wider Athens area of Attica in Greece were recruited in the study from

2008 to 2010. Our sample comprised 707 (55.9% females) adolescents of Greek origin aged 13.42 ± 0.88 years. Details of recruitment and data collection have been described elsewhere [50]. Prior to recruitment all study participants gave their verbal assent along with their parents'/guardians' written consent forms. The Harokopio University Bioethics Committee and the Greek Ministry of Education, Lifelong Learning and Religious Affairs approved the study. **Lipid measurements.** Lipid traits were assessed using enzymatic colorimetric assays and included; total cholesterol (cholesterol oxidase - phenol aminophenazone method), high density lipoprotein (HDL)-cholesterol and triglycerides (glycerol-3-phosphate oxidase -phenol aminophenazone). Low Density Lipoprotein (LDL)-cholesterol levels were calculated according to the Friedewald equation [48].

London Life Sciences Prospective Population Study (LOLIPOP). LOLIPOP is an ongoing community cohort of approximately 30,000 individuals aged 35-75 years, recruited in West London, UK to study the environmental and genetic factors that contribute to cardiovascular disease among UK Indian Asians. The study includes both European and Indian Asian subjects. Indian Asian participants reported having all four grandparents born on the Indian subcontinent, while European participants are self-classified whites born in Europe. For the current study, only white individuals were included in the primary meta-analysis. All participants provided written consent including for genetic studies. The LOLIPOP study is approved by the Ealing and St Mary's Hospitals Research Ethics Committees.

Three studies were included in the analysis:

1. LOLIPOP_EW_A: European whites from the general population, genotyped on Affymetrix 500K arrays.
2. LOLIPOP_EW_P: European whites from the general population, genotyped on Perlegen custom array.
3. LOLIPOP_EW610: European whites from the general population, genotyped on Illumina Human610 array.

Lipid measurements: Blood was taken from all participants for a standard biochemistry profile, including fasting glucose and lipids.[51].

The UCL-London-School-Edinburgh-Bristol (UCLEB) consortium of population-based prospective studies. The UCL-LSHTM-Edinburgh-Bristol (UCLEB) Consortium has been established to allow interrogation of genetic associations. The consortium consists of 12 well-established prospective observational studies comprising over 30,000 participants: Northwick Park Heart Study II (NPHS II), Whitehall-II Study (WHII), British Regional Heart Study (BRHS), English Longitudinal Study of Ageing (ELSA), MRC National Survey of Health and Development (MRC NSHD), 1958 Birth Cohort (1958BC), Edinburgh Artery Study (EAS), Edinburgh Type 2 Diabetes Study (ET2DS), Edinburgh Heart Disease Prevention Study (EHDPS), Aspirin for Asymptomatic Atherosclerosis Trial (AAAT), Caerphilly Prospective Study (CaPS) and the British Women's Heart and Health Study (BWHHS). All 12 studies in the UCLEB consortium are UK-based with wide geographic representation. The age at recruitment ranges from birth (MRC NSHD and 1958BC) to >90 years (ELSA), with most cohorts recruiting subjects in mid-life[52]. Individual study characteristics for studies used for replication are described below:

UCLEB-British Regional Heart Study (BRHS). The British Regional Heart study is a prospective study of CVD involving 7,735 British men drawn from general practices in 24 British towns followed from 1978 to 1980. Blood samples were taken along with a range of physiological measures and questionnaire data. The men were re-surveyed 20 years later in 1998-2000 when aged 60-79, and a whole blood sample taken at that time was used for extracting DNA, for 3,945 of the participants [53]. The BRHS has local (from each of the districts in which the study was based) and multi-centre ethical committee approvals. All men provided informed written consent to the investigation, which was performed in accordance with the Declaration of Helsinki. **Lipid measurements:** Blood samples (nonfasting at baseline, fasting at 20 years) were analyzed for serum total cholesterol by a modified Liebermann-Burchard method on a Technicon SMA 12/60 analyzer (Technicon Instruments, Tarrytown, NY) at baseline and with a Hitachi 747 automated analyzer (Roche Diagnostics, Indianapolis, Indiana) at 20 years. HDL cholesterol was measured by the Liebermann-Burchard method or enzymatic procedures after precipitation with magnesium phosphotungstate. The assays were cross-calibrated by remeasuring a small number of residual baseline samples for total and HDL cholesterol levels with the assay techniques applied at the 20-year examination [54].

UCLEB-British Women's Heart and Health Study (BWHHS). The British Women's Heart and Health Study (BWHHS) is a prospective cohort study of heart disease in over 4000 British women between the ages of 60 and 79. Study participants aged 60-79 years were randomly selected from the age sex register of a single general practice in each of 23 towns in England, Wales and Scotland. Women were recruited into the study between 1999 and 2001. Ethical committee approval was obtained for the study. **Lipid measurements:** During the medical examinations, fasting and resting blood Samples were taken. TC, HDL-C, and TG were measured on frozen serum samples using an Hitachi 747 analyser (Roche Diagnostics) and standard reagents LDL-C was estimated using the Friedewald equation [48].

Rotterdam Study cohort I (RS-I). The Rotterdam Study is an ongoing prospective population-based cohort study, focused on chronic disabling conditions of the elderly. The study comprises an outbred ethnically homogenous population of Dutch Caucasian origin. The rationale of the study has been described in detail elsewhere [55]. In summary, 7,983 men and women aged 55 years or older, living in Ommoord, a suburb of Rotterdam, the Netherlands, were invited to participate in the first phase. Fasting blood samples were taken during the participant's third visit to the research center.

Rotterdam Study cohort II (RS-II). The Rotterdam Study cohort II prospective population-based cohort study comprises 3,011 residents aged 55 years and older from the same district of Rotterdam. The rationale and study design of this cohort is similar to that of the RS-I [55]. The baseline measurements, including the fasting HDL measurements, took place during the first visit. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the “Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study)”. All participants provided written informed consent to participate in the study and to obtain information from their treating physicians [55]. **Lipid measurements:** Serum lipid measurements of total cholesterol and HDL were determined enzymatically, using an automated procedure [56].

FENLAND. The Fenland Study is an ongoing, population-based cohort study (started in 2005) designed to investigate the association between genetic and lifestyle environmental factors and the risk of obesity, insulin sensitivity, hyperglycemia and related metabolic traits in men and women aged 30 to 55 years. Potential volunteers were recruited from General Practice sampling frames in the Fenland, Ely and Cambridge areas of the Cambridgeshire Primary Care Trust in the UK. Exclusion criteria for the study were: prevalent diabetes, pregnant and lactating women, inability to participate due to terminal illness, psychotic illness, or inability to walk unaided. All participants had measurements done at the MRC Epidemiology Unit Clinical Research Facilities in Ely, Wisbech and Cambridge. Participants attended after an overnight fast for a detailed clinical examination, and blood samples were collected. The Local Research Ethics Committee granted ethical approval for the study and all participants gave written informed consent. **Lipid Measurements:** HDL-cholesterol, total cholesterol and triglycerides were measured by enzymatic assays (Siemens Healthcare), and LDL-cholesterol was derived by the Friedewald formula [48].

Copenhagen General Population Study (CGPS). All participants were white and of Danish descent; this information is available through the national Danish Central Person Registry. No participants appeared in more than one of the three studies. The studies were approved by Danish ethical committees and Herlev Hospital. This general population study was initiated in 2003 with ongoing enrolment. IHD endpoints have been collected from 1976 to May 2009. Individuals were selected on the basis of the national Danish Civil Registration System to reflect the adult Danish population aged 20–100y. Data were obtained from a questionnaire, a physical examination, blood samples, and from DNA. At the time of genotyping 59,883 participants had been included; of these, 5,270 were used as controls in the CIHDS (see below), leaving 54,613 for analyses in the CGPS [57-60]. **Genotyping**. KASP genotyping of rs7265886 and rs5985471 was undertaken using KASP at KBioscience (www.lgcgenomics.com/). Assays are based on competitive allele-specific PCR and enable bi-allelic scoring of single nucleotide polymorphisms (SNPs). The SNP-specific KASP Assay mix and the universal KASP Master mix are added to DNA samples, a thermal cycling reaction is then performed, followed by an end-point fluorescent read. Bi-allelic discrimination is achieved through the competitive binding of two allele-specific forward primers, each with a unique tail sequence that corresponds with two universal FRET (fluorescence resonant energy transfer) cassettes; one labelled with FAMTM dye and the other with HEXTM dye. DNA for this genotyping was prepared using whole genome amplification using a primer extension pre-amplification (PEP-PCR) protocol (http://goo.gl/pXYCPW). For the autosomal variant rs7265886, quality control of bespoke genotyping data involved visual inspection of genotyping intensity plots along with assessment of Hardy Weinberg equilibrium (p>0.05). For the X-chromosome variant rs5985471, visual inspection of genotyping intensity plots was stratified by sex assuming that all males would present only as homozygotes. Following visual inspection, any mis-called heterozygote males were dropped from further analysis (0.6% of the total sample size). Database records of sex were assessed prior to these quality control analyses.

Case-control status. Information on diagnosis of IHD (World Health Organization International Classification of Diseases: ICD8 410–414; ICD10 I20–I25) was collected and verified from existing data from 1976 until May 2009 by reviewing all hospital admissions and diagnoses entered in the national Danish Patient Registry and all causes of death entered in the national Danish Causes of Death Registry. Even though some individuals entered into our studies after 1976, we have complete information on all participants on any hospitalisation or death from IHD from 1976 through 2009 through these registries. IHD was angina pectoris and/or myocardial infarction (ICD8 410; ICD10 I21–I22), based on characteristic chest pain, electrocardiographic changes, and/or elevated cardiac enzymes. Follow-up was 100% complete, that is, no individual was lost to follow-up in any of the studies. Information on diagnosis of MI (World Health Organization; code 410 from the *International Classification of Diseases, Eighth Revision* [*ICD-8*] and codes I21-I22 from the *ICD-10*) was collected and verified by reviewing hospital admissions and diagnoses entered in the national Danish Patient Registry, causes of death entered in the national Danish Causes of Death Registry, and medical records from hospitals and general practitioners.

Association testing

Discovery and replication of novel lipid loci using single-marker association tests

We assessed single point analysis to test associations between genetic variants (including SNPs and biallelic indels) and lipid residuals (LDL, HDL, TG, TC, VLDL) using linear regression models assuming additive genetic models. Single point analysis was carried out in each cohort separately. We combined 14 cohorts with the UK10K+1000GP reference panel increasing our sample size to a maximum of 22,814 study participants (including the four UK10K cohorts with 1958BC, VBI, FVG, CARLANTINO, INCIPE1, INCIPE2, TEENAGE, HELIC-HA, HELIC-HP, LURIC; **Table S7**). We used SNPTESTv4.2 for cohorts with unrelated samples and GEMMA v0.92 for cohorts with related samples. We accounted for relatedness by computing a kinship matrix using its centered genotypes model. We used fixed-effect inverse variance methods implemented in GWAMA v2.1 to calculate meta-analysis statistics. The p-values in the 14-way meta-analysis were adjusted for the genomic inflation factor. For the 14-way meta-analysis we filtered out variants not found in the UK10K populations, variants with an overall MAF<0.1% among the 14 populations and a mean info score <0.4 among the imputed cohorts.

Locus selection from single-variant tests

We compiled a list of variants previously associated with lipid measurements from online resources, including the NHGRI GWAS catalog [61], complemented by manual curation in PubMed and of large-scale genetic association studies for lipids [62]. To identify whether potentially novel variants corresponded to known associated lipid regions, a LD clumping procedure was applied to the 14-way meta-analysis results using PLINK. The procedure clumps all variants in LD (r^2>0.2) within 500kb from an index variant. These index variants were clumped against the genotypes from the WGS analysis (3,621 individuals) and against the known variant list. To examine which genes were spanning the clumped region, we annotated the clumped region to the list of GENCODEv15 genes.

We next selected from the 14-way meta-analysis all loci reaching a suggestive threshold (p-value≤10^{-5}), and annotated their physical distance and linkage disequilibrium (LD) metrics r^2 to the nearest known variant. Variants with high linkage di r^2>0.2 were declared as known associations. We used conditional analyses to test the independence of putative novel associations from known variants, if the two variants were within a physical distance of 1Mb and had moderate to low LD (r^2>0.2). The analysis was assessed in each UK10K cohort separately (ALSPAC WGS, TwinsUK WGS, ALSPAC GWA, TwinsUK GWA) by re-running single-point analysis conditioning upon positive controls in the same region. For the WGS and the ALSPAC GWA datasets, SNPTEST was used whereas for the TwinsUK GWA, GEMMA was used using allele counts of the conditioned SNPs as a covariate. We then performed a 2-way meta-analysis on the conditioned single-point statistics for variants from the WGS analysis and a 4-way meta-analysis for variants from the WGS and GWA combined analysis.

Each independent novel variant selected from the 14-way meta-analysis as described above was carried forward for replication in up to 20,000 additional individuals from 8 cohorts using a combination of imputation using the UK10K+1000GP reference panel (LOLI-EW610, LOLI-EWA, LOLI-EWP, RS-1, RS-2) and direct genotyping (Fenland, UCLEB-BRHS and UCLEB-BWHHS). Two loci reaching the experiment-wide threshold for association (p-value≤6.42×10^{-10}) were additionally genotyped in the Copenhagen General Population Study. For replication and bespoke genotyping data linear regression of the transformed and standardised variable on coded genetic variant assuming an additive genetic model. Covariables were included in the formation of transformed lipid variables through residualisation. Analyses of both ischaemic heart disease (IHD) and myocardial infarction (MI) (binary outcomes including all available cases) were undertaken using a logistic regression including age, age^2 and sex in the model. Effects are reported as odds ratios for IHD or MI according to each additional minor effect allele. Sensitivity analyses are reported where the use of lipid lowering drugs are included in the model. As for LDL analyses, all tests assume an additive genetic model.

Data analyses in the UK10K Exome data set

Comparison with results from the NHLBI Exome Sequencing Project (ESP)

The final set of SNVs in the patient set was compared with variants in 4,300 European-American individuals (EA) from the ESP (ESP6500) [63]. SNVs were compared by position and alternative allele. Multiallelic sites were included and different alternative alleles at the same site were treated as separate variants. UK10K and the ESP used different baits regions and this was considered in the comparison. Four bait regions were used in the ESP and we defined the ESP bait region as the overlap of these four sets. From the coverage files provided by ESP, 92-93% of UK10K SNVs in this bait region were in sequence blocks with an average coverage of at least 30X. As well as comparing the EA ESP variants with our final SNVs, we also determined how many EA ESP SNVs had been called in our patient set but filtered out during quality control.

Examination of SNV functional consequences using the Illumina Body Map dataset

While exome sequencing is occasionally used to study complex traits, the major focus of early exome sequencing has been on identifying single diagnostic mutations in individuals with rare, severe disease (as for the UK10K *rare disease* collection described here). The design of the UK10K study allowed insight into the clinical interpretation of exome data. For instance, a critical step in the use of sequence data in a clinical setting is to assign a specific consequence (loss of function [LoF], missense and other functional, silent or possibly functional, and other) to a given SNV. Variants often affect more than one transcript and the predicted consequences can differ between transcripts. Usually each SNV is annotated with the most severe consequence of any overlapping transcript. This approach may incorrectly classify some variants as deleterious if the relevant transcript is either an artefact of computational gene models, or expression is only at low levels or in tissues where it is unlikely to contribute to the disease of interest. Recently, it has been shown that most protein-coding genes have a dominant transcript, which is expressed at a considerably higher level than other transcripts, and that often the same transcript has the highest expression across tissues [64]. Here, we evaluate how inferred consequences change when the set of transcripts is restricted based on their expression patterns. Our aim was to increasingly focus on effects that are most likely to be important at the protein level.

We used expression values from the Illumina Body Map dataset. A threshold of 1 fragment per kilobase per million fragments mapped (FPKM) was applied to split transcripts into high and low expression groups because this has been suggested as the minimum level corresponding to detectable protein. To address transcript expression in appropriate tissues, given that our exome samples are from diverse diseases groups, we also defined a subset of widely expressed transcripts, which had at least 1 FPKM in all tissues in the Body Map dataset. Consistent with other studies, we see a strong depletion of common LoF and functional variants (**Extended Data Figure 7**). We show that when effects are restricted to transcripts with high expression there are relatively fewer LoF and functional variants than when all protein-coding transcripts are considered. LoF and functional variants are further depleted when effects are restricted to widely expressed transcripts. In SNVs where all the protein-coding transcripts had low expression, these transcripts had a higher incidence of LoF and functional variants than found for transcripts with high expression. These findings are consistent with selection against LoF and functional changes that have an impact at the protein level. In transcripts with high expression, the incidence of LoF SNVs was 1.8% for singletons and 0.2% for variants with frequencies over 1%. For widely expressed transcripts these numbers dropped to 1.3% and 0.04%, respectively. Our observations could have important implications for future annotation of patient exomes, as the consequence of a particular SNV can vary dramatically across transcripts and therefore needs to be assessed in a context dependent manner.

Methods. The final set of SNVs was annotated with functional effects using VEP v73 [16]. Different alternative alleles at the same site were treated as separate variants. The option to flag canonical transcripts was selected and Sequence Ontology (http://www.ensembl.org/info/genome/variation/predicted_data.html#consequences) terms were used. We followed the approach of Gonzàlez-Porta *et al.* [64] to obtain expression data for human tissues. Transcript expression levels were downloaded from Expression Atlas for the Illumina Body Map dataset [65]. The Illumina Body Map consists of RNA-seq reads from several individuals, covering 16 tissues, with one sample per tissue. The Expression Atlas results were generated by the iRAP pipeline v0.3.3 (unpublished - see http://dx.doi.org/10.1101/005991), which included mapping against Ensembl v73 with tophat1 v1.4.1 and expression quantification with cufflinks1 v1.3.0. The experimental protocol selected for mRNA and transcripts at least 300 bp long. We therefore only included protein-coding transcripts that were at least 300 bp long in the expression analysis. Transcript and gene biotypes and exon genomic start and end points were downloaded from Ensembl 73 using BioMart. We defined protein-coding transcripts as those with a biotype of protein _coding from a gene with a biotype of protein_coding. Transcript length was calculated from the exon genomic start and end points. Transcripts were classed as having high expression if they had a fragment per kilobase per million fragments mapped (FPKM) ≥1 in any tissue and low expression if they had an FPKM <1 in every tissue. Widely expressed transcripts were those with a FPKM ≥1 in all 16 tissues. Some transcripts had negative expression values; these are measurements that cufflinks determined were low quality. There were also some transcripts with no expression data. SNVs were excluded from the expression analysis if any of the protein-coding transcripts listed was <300 bp long or had missing or low quality expression data. Of 842,646 SNVs, 820,845 affected a protein-coding transcript and for 645,870 (79%) all protein-coding transcripts were ≥300 bp and had good quality expression data.

The most severe consequence for each SNV was found for the following subsets of transcripts:

1. protein-coding
2. high expression
3. widely expressed
4. low expression where there are no high expression transcripts for that variant

Consequences were grouped into categories according to the SO annotations as follows:

1. Loss of function (LoF) – splice_donor_variant, splice_acceptor_variant, stop_gained
2. Functional – stop_lost, initiator_codon_variant, missense_variant
3. Possibly functional – splice_region_variant, synonymous_variant, stop_retained_variant, 5_prime_UTR_variant, 3_prime_UTR_variant
4. Other – all other annotations.

For each transcript subtype, the relative numbers of SNVs in each functional category were calculated as a percentage of the SNVs meeting the criteria for that subset.

Assessment of incidental findings in exome data

In asymptomatic individuals, the frequency of exonic disease causing variants, also known as incidental findings (IF), has important practical implications for population-wide genetic screens. However, this frequency depends the exact filtering criteria used. The combination of this lack of objectivity with the limited availability of large-scale sequencing population datasets contributes to the uncertainty surrounding the IF frequency. The rich catalogue of variants identified in the UK10K exome study (rare diseases, obesity and neuro-developmental diseases) provides an opportunity to answer this question in the UK population. To this end, we defined a set of objective criteria based on the implicated genes as well as the available annotations for these variants.

First, to be considered as incidental findings, genetic variants had to be located in the list of genes recently released by the American College of Medical Genetics and Genomics (ACMG) in their guidelines for the analysis of exome/whole genome data [66].

Second, using the 1,805 participants whose consent allowed for such a broad search [67], we consulted disease-specific clinicians with expertise in the disease area and provided them with the list of variant calls in the genes cited by the ACMG. All rare genetic variants (MAF < 1%) for each ACMG gene/disorder combination were extracted and assessed using objective criteria by a clinical geneticist with the relevant disease-specific expertise for its potential as an IF. The clinicians then were asked to provide for each variant four quantitative values (each either 0, 1 or 2) based on: (i) evidence based on disease specific databases (when available for this disease), (ii) scientific literature search, (iii) in silico prediction and (iv) ClinVar database. The scores were defined as follows:

ClinVar scoring

0: no review, or review with no evidence of pathogenicity

1: reviewed variant, classified as pathogenic but submitted by a single lab

2: variant reviewed by at least two independent sources, classified as pathogenic by all submitters without a conflict

Locus Database scoring

0- variant not previously ascribed as pathogenic in a LSDB

1- variant previously ascribed as pathogenic in a LSDB by 1 lab

2- variant previously ascribed as pathogenic in a LSDB by 2+ labs

Literature search scoring

0- variant not previously ascribed as pathogenic in a peer-reviewed publication

1- variant previously ascribed as pathogenic in 1 peer-reviewed publication

2- variant previously ascribed as pathogenic in 2+ independent peer-reviewed publication (different lab/authors).

In silico prediction scoring

0- in silico prediction does not suggest disease causing.

1- moderate confidence in silico prediction of pathogenicity (high values for nsSNP prediction for example)

2- high confidence in silico prediction of pathogenicity (that would be a LOF variant in a known haplo-sufficient gene)

Variants were defined as reportable IF if they scored a value of 2 in at least one of these categories. We further excluded from this list the variants that were reviewed by ClinVar and were flagged to have either a non-pathogenic interpretation or some level of discordance, on the basis that only variants without contentious interpretation should be reported as IF. In order to more appropriately mirror a real-world clinical setting, only one disease-specific expert assessed each set of variants. However, we note that this approach did not allow us to quantify concordance between experts. While participants in the exome arm were recruited for particular diseases, we only considered as reportable mutations unrelated to the disease participants were recruited for (hence excluding FH cases from the FH IF computation).

Supplementary References

1. Golding, J., Pembrey, M., Jones, R. & Team, A.S. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr Perinat Epidemiol* **15**, 74-87 (2001).
2. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **42**, 111-27 (2013).
3. Moayyeri, A., Hammond, C.J., Hart, D.J. & Spector, T.D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res Hum Genet*, 1-6 (2012).
4. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078-2079 (2009).
5. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
6. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).
7. Depristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498 (2011).
8. Mills, R.E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* **21**, 830-9 (2011).
9. Shin, S.Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* (2014).
10. 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. *American journal of human genetics* **81**, 1084-1097 (2007).
11. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2012).
12. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010).
13. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909 (2006).
14. Williams, F.M. *et al.* Genes contributing to pain sensitivity in the normal population: an exome sequencing study. *PLoS Genet* **8**, e1003095 (2012).
15. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269-76 (2011).
16. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).
17. Lin, M.F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-82 (2011).
18. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-48 (2012).
19. Purcell, S. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American journal of human genetics* **81**, 559-575 (2007).
20. Bonnelykke, K. *et al.* Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. *Nat Genet* **45**, 902-6 (2013).
21. Soranzo, N. *et al.* Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. *PLoS Genetics* **5**, e1000445 (2009).
22. Delaneau, O., Howie, B., Cox, A.J., Zagury, J.-F. & Marchini, J. Haplotype Estimation Using Sequencing Reads. *The American Journal of Human Genetics* **93**, 687-696 (2013).
23. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906-913 (2007).
24. Magi, R. & Morris, A.P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
25. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821-824 (2012).
26. Xu, C. *et al.* Estimating Genome-Wide Significance for Whole-Genome Sequencing Studies. *Genetic Epidemiology*, n/a-n/a (2014).
27. Li, M.X., Yeung, J.M., Cherny, S.S. & Sham, P.C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet* **131**, 747-56 (2012).
28. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300 (1995).

29. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* **89**, 82-93 (2011).
30. Lee, S., Wu, M.C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762-75 (2012).
31. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-74 (2012).
32. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* (2010).
33. Spencer, C.C., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* **5**, e1000477 (2009).
34. Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 18-31 (2003).
35. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet Epidemiol* **38**, 281-90 (2014).
36. Huang, J. *et al.* A reference panel of 3,781 genomes from the UK10K Project increases imputation performance of low frequency variants. *Nature Communications (Under peer review)*.
37. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics* **84**, 210-223 (2009).
38. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **488**, 57-74 (2012).
39. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **488**, 75-82 (2012).
40. Galwey, N.W. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genet Epidemiol* **33**, 559-68 (2009).
41. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics* **44**, 243-246 (2012).
42. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res* **27**, 209-20 (1967).
43. Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453 (2012).
44. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet* **10**, e1004528 (2014).
45. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* **35**, 34-41 (2006).
46. Traglia, M. *et al.* Heritability and Demographic Analyses in the Large Isolated Population of Val Borbera Suggest Advantages in Mapping Complex Traits Genes. *PLoS ONE* **4**, e7554 (2009).
47. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet* **21**, 659-65 (2013).
48. Friedewald, W.T., Levy, R.I. & Fredrickson, D.S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* **18**, 499-502 (1972).
49. Winkelmann, B.R. *et al.* Rationale and design of the LURIC study--a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. *Pharmacogenomics* **2**, S1-73 (2001).
50. Ntalla, I. *et al.* Body composition and eating behaviours in relation to dieting involvement in a sample of urban Greek adolescents from the TEENAGE (TEENs of Attica: Genes & Environment) study. *Public Health Nutr* **17**, 561-8 (2014).
51. Chambers, J.C. *et al.* Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat Genet* **40**, 716-8 (2008).
52. Shah, T. *et al.* Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PLoS One* **8**, e71345 (2013).
53. Walker, M., Whincup, P.H. & Shaper, A.G. The British Regional Heart Study 1975-2004. *Int J Epidemiol* **33**, 1185-92 (2004).
54. Hardoon, S.L. *et al.* How much of the recent decline in the incidence of myocardial infarction in British men can be explained by changes in cardiovascular risk factors? Evidence from a prospective population-based study. *Circulation* **117**, 598-604 (2008).
55. Hofman, A. *et al.* The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol* **28**, 889-926 (2013).
56. van Gent, C.M., van der Voort, H.A., de Bruyn, A.M. & Klein, F. Cholesterol determinations. A comparative study of methods with special reference to enzymatic procedures. *Clin Chim Acta* **75**, 243-51 (1977).
57. Nordestgaard, B.G., Benn, M., Schnohr, P. & Tybjaerg-Hansen, A. Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women. *JAMA* **298**, 299-308 (2007).

58. Frikke-Schmidt, R. *et al.* Association of loss-of-function mutations in the ABCA1 gene with high-density lipoprotein cholesterol levels and risk of ischemic heart disease. *JAMA* **299**, 2524-32 (2008).
59. Zacho, J. *et al.* Genetically elevated C-reactive protein and ischemic vascular disease. *N Engl J Med* **359**, 1897-908 (2008).
60. Kamstrup, P.R., Tybjaerg-Hansen, A., Steffensen, R. & Nordestgaard, B.G. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *JAMA* **301**, 2331-9 (2009).
61. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).
62. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713 (2010).
63. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
64. Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol* **14**, R70 (2013).
65. Petryszak, R. *et al.* Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* **42**, D926-32 (2014).
66. Green, R.C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* **15**, 565-74 (2013).
67. Kaye, J. *et al.* Managing clinically significant findings in research: the UK10K example. *Eur J Hum Genet* (2014).

Acknowledgements and grant support

This study makes use of data generated by the UK10K Consortium. The Wellcome Trust provided funding for UK10K (WT091310).

Additional grant support. Action Medical Research (RTF-1411 to M. Schmidts), Australian National Health and Medical Research Council (1010494 to S.G. Wilson, 1048216 and 1078037 to S.G. Wilson, J. B. Richards, T.D. Spector and P.M. Visscher) and the Sylvia & Charles Viertel Charitable Foundation (J. Yang), FRSQ, CIHR, CFI, McGill University, Lady Davis Institute and the Jewish General Hospital (J.B. Richards), British Heart Foundation (PG/07/045, CH/09/003/26631 to S. Bhattacharya, PG08/008 to S. E. Humphries, M. Futema, RG/15/13/28570 to P. Scambler, RG/10/12/28456 and SP/13/6/30554 to UCLEB consortium, RG/08/013/25942 to The British Regional Heart Study, BRHS, PG/12/66/30442 to British Women Heart and Health Study, BWHHS, MC_U106179471 to Fenland Study, SP/04/002 to LOLIPOP), Dutch Kidney Foundation (CP11.18 to M. Schmidts), EMBL (E. Birney, I. Dunham, P. Flicek), EMBL-WTSI (G.R.S. Ritchie, V. Iotchkova), Emma and Ernesto Rulfo Association for Medical Genetics (N. Migone), EU FP7 (257082, HEALTH-F5-2011-282510 to N. Soranzo and L. Chen, TwinsUK, EpiMigrant, 279143 to LOLIPOP, 201668 and 305739 to LURIC), Great Ormond Street Hospital Charity and the GOSH Biomedical Research Centre (F. Muntoni), Medical Research Council (S. Ring, MC_UU_12015/2 to J. R.B. Perry, G0801418, G0800509 to M. C. O'Donovan, M. J. Owen, MR/J012165/1 to L. Paternoster, MC_UU_12013/1-9 to N. J. Timpson, G. D. Smith, D. Evans, T. R. Gaunt, H. Shihab, G0601966, G0700931 to LOLIPOP, to ALSPAC), Muscular Dystrophy Campaign (A. R. Foley), National Alliance for Autism Research, Nancy Lurie Marks Family Foundation (D. Skuse), NHS (F. Khawaja, R. Taylor, S. J. Payne), NIHR (M. D. Tobin, P. Bolton, P. Syrris, R. K. Semple, S. E. Humphries, M. Futema, N. Soranzo, TwinsUK, RP-PG-0407-10371 to LOLIPOP), National Institutes of Health (GM104371 to D.G. MacArthur and K.J. Karczewski), Oak Foundation (S. Shin), RadboundMC Hypatia fellowship, RadboundMC, Nijmegen (M. Schmidts), Sigrid Juselius Foundation, Academy of Finland (J. Lönnqvist), The Gulbenkian Doctoral Programme for Advanced Medical Education (L. R. Lopes), Tuberous Sclerosis Association (UK) and TS Alliance (USA) (R. Ekong, S. Povey), Wellcome Trust (WT083228 to S. Bhattacharya, WT091551 to D. B. Savage, WT092731 to S. Ring, WT095908 to P. Flicek, WT098051, WT091310 to A. Day-Williams, A. Daly, A. Kolb-Kokocinski, A. Hendricks, C. A. Anderson, C. Smee, C. Joyce, C. S. Franklin, C. Langford, D. K. Jackson, D. Muddyman, E. Wheeler, E. Zeggini, E, Serra, F. Payne, G. Marenne, G. R.S. Ritchie, G. Coates, I. Barroso, I. Tachmazidou, J. Floyd, J. Morris, J. C. Barrett, J. Huang, J. Stalker, J. Maslen, K. Panoutsopoulou, K. Kennedy, K. Rehnström, K. Carss, K. Wong, K. Walter, L. Southam, L. Crooks, M. Lopes, M. van Kogelenberg, M. E. Hurles, M. A. Quail, N. Soranzo, O. Pietilainen, P. Vijayarangakannan, P. Clapham, P. Ellis, P. Danecek, R. Durbin, S. Farooqi, S. Edkins, S. Balasubramaniam, S. McCarthy, S. Shin, S. Schiffels, T. Keane, T. Cox, V. Iotchkova, Y. Memari, WT 095564/Z11/Z to K.Chatterjee, WT098498 to R. K. Semple, WT100585/Z/12/Z to N. Schoenmakers, WT092447/B/10/Z to J. R.B. Perry, Fenland Study, WT091310/C/10/Z to TwinsUK, 098051 to TEENAGE and HELIC, 084723/Z/08/Z to LOLIPOP, 092731 to ALSPAC), The Danish Heart Foundation, Copenhagen County Foundation, and Chief Physician Johan Boserup and Lise Boserup's Fund (CGPS) (The Danish Council for Independent Research, Medical Sciences (FSS)), CARIVR Foundation (INCIPE), Fundação Calouste Gulbenkian, Fundação Champalimaud, Ministério da Saúde and Fundação para a Ciência e Tecnologia (The Gulbenkian Doctoral Programme for Advanced Medical Education), TwinsUK receives support from the National Institute for Health Research (NIHR) BioResource Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London, Wellcome Trust Sanger Institute and National Eye Institute (TwinsUK), Heracleitus II (TEENAGE), European Social Fund (ESF) (TEENAGE), European Research Council (ERC-2011-StG 280559-SEPI to HELIC), National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre Imperial College Healthcare NHS Trust, Action on Hearing Loss (G51 to LOLIPOP), 7th Framework Program (Grant Agreement numbers 201668 and 305739 to LURIC), German Ministry for Education and Research (project e:AtheroSysMed, grant number 01ZX1313A-K to LURIC), European Regional Development Fund (LURIC, INGI FVG and Carlantino), Telethon Foundation (GGP09037 to INGI FVG and Carlantino, to INGI-Val Borbera), Fondo Trieste (2008 to INGI FVG and Carlantino), Regione FVG (L.26.2008 to INGI FVG and Carlantino), Italian Ministry of Health (RC16/06, ART. 13 D.LGS 297/99 to INGI FVG and Carlantino), Compagnia di San Paolo (INGI-Val Borbera), Fondazione Cariplo (INGI-Val Borbera), Ricerca Finalizzata 2008 and CCM 2010 (INGI-Val Borbera), University of Bristol (ALSPAC), The National Basic Research Program of China (973 program no. 2011CB809201, 2011CB809202 and 2011CB809203), the Shenzhen Municipal Government of China (ZYC201105170397A), The Pawsey Supercomputing Centre (with funding from the Australian Government and the Government of Western Australia; PG/13/511 to S.G. Wilson), The National Children's Research Centre, Crumlin, Ireland (Louise Gallagher), Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates CZD/16/6 and the Scottish Funding Council HR03006.

Other. We would especially like to thank Robert E Handsaker for useful discussions and technical support with Genome STRiP (Nature Genetics, 2011), which we used to analyse large deletions in the UK10K cohorts set. We would also like to thank Lorenz Wernisch for useful discussions regarding the analysis of population structure, Danielle Walker and Daniel Mead for manuscript editing, and Jennifer Harrow and Adam Frankish for LoF annotations. *UK10KMUIR cohort of the Neuro/schizophrenia samples*: "We are indebted to the late Walter Muir (1958-2009), Professor of Developmental Psychiatry at the University of Edinburgh whose commitment to patients with intellectual disability and psychosis led to the inclusion of this cohort in the study." *Copenhagen:* "We are indebted to staff and participants of the Copenhagen General Population Study, Copenhagen City Heart Study, and Copenhagen Ischemic Heart Disease Study for their important contributions". *BWHHS* The funders had no role in the study design; in the collection, analysis and interpretation of the data; in the writing of the report or in the decision to submit the paper for publication. *Rotterdam Study*. On behalf of the Rotterdam Study, we thank Pascal Arp, Mila Jhamai, Marijn Verkerk, Lizbeth Herrera and Marjolein Peters for their help in creating the GWAS database, and Karol Estrada and Maksim V. Struchalin for their support in creation and analysis of imputed data. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. *Fenland*. We are grateful to all the volunteers for their time and help, and to the General Practitioners and practice staff for assistance with recruitment. We thank the Fenland Study Investigators, Fenland Study Co-ordination team and the Epidemiology Field, Data and Laboratory teams. We acknowledge A. Monaco, A. Bailey, R. Holt, A. Le Couteur, G. Baird, J. Green, J. R. Parr. *TwinsUK*. We are extremely grateful to all the participants who took part in this study, those who helped recruitment and the whole TUK team. *ALSPAC*. We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. Work also carried out at the Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute, and LabCorp (Laboratory Corporation of America) using support from 23andMe. *INGI-Val Borbera***.** We thank the inhabitants of the VB that made this study possible, the local administrations, the Tortona and Genova archdiocese and the ASL-22, Novi Ligure (Al) for support. We also thank C. Camaschella for data collection supervision and organization of the clinical data collection, F. Viganò for technical help, C. Masciullo for building the analysis platform. *INGI FVG* *and Carlantino*. We would like to thank the people of the Friuli Venezia Giulia Region and of Carlantino for the everlasting support. *LURIC*. We extend our appreciation to the participants of the LURIC study; without their collaboration, this article would not have been written. We thank the LURIC study team who were either temporarily or permanently involved in patient recruitment as well as sample and data handling, in addition to the laboratory staff at the Ludwigshafen General Hospital and the Universities of Freiburg and Ulm, Germany. *HELIC (Hellenic Isolated Cohorts).* The MANOLIS cohort is named in honour of Manolis Giannakakis, 1978-2010. We thank the residents of the Mylopotamos villages, and of the Pomak villages, for taking part. The HELIC study has been supported by many individuals who have contributed to sample collection (including A. Athanasiadis, O. Balafouti, C. Batzaki, G. Daskalakis, E. Emmanouil, C. Giannakaki, M. Giannakopoulou, A. Kaparou, V. Kariakli, S. Koinaki, D. Kokori, M. Konidari, H. Koundouraki, D. Koutoukidis, V. Mamakou, E. Mamalaki, E. Mpamiaki, M. Tsoukana, D. Tzakou, K. Vosdogianni, N. Xenaki, E. Zengini), data entry (T. Antonos, D. Papagrigoriou, B. Spiliopoulou), sample logistics (S. Edkins, E. Gray), genotyping (R. Andrews, H. Blackburn, D. Simpkin, S. Whitehead), research administration (A. Kolb-Kokocinski, S. Smee, D. Walker) and informatics (M. Pollard, J. Randall). *TEENAGE*. We thank all study participants and their families as well as all volunteers for their contribution in this study. We thank the following staff from the Sample Management and Genotyping Facilities at the Wellcome Trust Sanger Institute for sample preparation, quality control and genotyping: D. Jones, D. Simpkin, E. Gray, H. Blackburn, S. Edkins. *Generation Scotland.* We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants and nurses. *LOLIPOP.* The work was carried out in part at the NIHR/Wellcome Trust Imperial Clinical Research Facility. We thank the participants and research staff who made the study possible. This report presents independent research funded partially by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.