ARTICLE

OPEN

doi:10.1038/nature14962

The UK10K project identifies rare variants in health and disease

The UK10K Consortium*

The contribution of rare and low-frequency variants to human traits is largely unexplored. Here we describe insights from sequencing whole genomes (low read depth, 7×) or exomes (high read depth, 80×) of nearly 10,000 individuals from population-based and disease collections. In extensively phenotyped cohorts we characterize over 24 million novel sequence variants, generate a highly accurate imputation reference panel and identify novel alleles associated with levels of triglycerides (*APOB*), adiponectin (*ADIPOQ*) and low-density lipoprotein cholesterol (*LDLR* and *RGAG1*) from single-marker and rare variant aggregation tests. We describe population structure and functional annotation of rare and low-frequency variants, use the data to estimate the benefits of sequencing for association studies, and summarize lessons from disease-specific collections. Finally, we make available an extensive resource, including individual-level genetic and phenotypic data and web-based tools to facilitate the exploration of association results.

Assessment of the contribution of rare genetic variation to many human traits is still largely incomplete. In common and complex diseases, a lack of empirical data has to date hampered the systematic assessment of the contribution of rare and low-frequency genetic variants (defined throughout this paper as minor allele frequency (MAF) <1% and 1–5%, respectively). Rare variants are incompletely represented in genome-wide association (GWA) studies[1] and custom genotyping arrays[2,3], and impute poorly with current reference panels. Rare and low-frequency variants also tend to be population- or sample-specific, requiring direct ascertainment through resequencing[4,5]. Recent exome-wide resequencing studies have begun to explore the contribution of rare coding variants to complex traits[6], but comparatively little is known of the non-coding part of the genome where most complex trait-associated loci lie[7]. At the other end of the human disease spectrum, the widespread application of exome-wide sequencing is accelerating the rate at which genes and variants causal for rare diseases are being identified. Despite this, many Mendelian diseases still lack a genetic diagnosis and the penetrance of apparently disease-causing loci remains inadequately assessed.

The UK10K project was designed to characterize rare and low-frequency variation in the UK population, and study its contribution to a broad spectrum of biomedically relevant quantitative traits and diseases with different predicted genetic architectures. Here we describe the data and initial findings generated by the different arms of the UK10K project. In addition to this paper, UK10K companion papers describe the utility of this resource for imputation[8], association discovery for bone mineral density[9], thyroid function[10] and circulating lipid levels[11] and provide access to the study results through novel web tools[12].

Study designs in the UK10K project

The UK10K project includes two main project arms (Table 1). The UK10K-cohorts arm aimed to assess the contribution of genome-wide genetic variation to a range of quantitative traits in 3,781 healthy individuals from two intensively studied British cohorts of European ancestry, namely the Avon Longitudinal Study of Parents and Children (ALSPAC)[13] and TwinsUK[14]. A low read depth (average 7×) whole-genome sequencing (WGS) strategy was employed in order to maximize total variation detected for a given total sequence quantity[15] while allowing interrogation of noncoding variation. Sixty-four different phenotypes were analysed, including traits of primary clinical relevance in 11 major phenotypic groups (obesity, diabetes, cardiovascular and blood biochemistry, blood pressure, dynamic measurements of ageing, birth, heart, lung, liver and renal function; Supplementary Table 1). Of these, 31 phenotypes were available in both studies (referred to as 'core' and reported in association analyses), 18 were unique to TwinsUK and 15 were unique to ALSPAC.

The UK10K-exomes arm aimed to identify causal mutations through high read depth (mean ~80× across studies) whole-exome sequencing of approximately 6,000 individuals from three different collections: rare disease, severe obesity and neurodevelopmental disorders. The disorders studied in the UK10K-exomes arm have been shown to have a substantial genetic component at least partially driven by very rare, highly penetrant coding mutations. The rare disease collection includes 125 patients and family members in each of eight rare disease areas (Table 1). Disease types were selected with different degrees of locus heterogeneity, prior evidence for monogenic causation and likely modes of inheritance (for example, dominant or recessive). The obesity collection comprises of samples with severe obesity phenotypes, including approximately 1,000 subjects from the Severe Childhood Onset Obesity Project (SCOOP)[16], plus severely obese adults from several population cohorts. The neurodevelopmental collection comprises of ~3,000 individuals selected to study two related neuropsychiatric disorders (autism spectrum disorder and schizophrenia).

Discovery of 24 million novel genetic variants

In total, 3,781 individuals were successfully whole-genome sequenced in the UK10K-cohorts arm. After conservative quality control filtering (Extended Data Figs 1 and 2 and Supplementary Table 2), the final call set contained over 42M single nucleotide variants (SNVs, 34.2M rare and 2.2M low-frequency), ~3.5M insertion/deletion polymorphisms (INDELs; 2,291,553 rare and 415,735 low-frequency) and 18,739 large deletions (median size 3.7 kilobase). Each individual on average contained 3,222,597 SNVs (5,073 private), 705,684 INDELs (295 private) and 215 large deletions (less than 1 private). Of 18,903 analysed

*A list of authors and affiliations appears at the end of the manuscript.

protein-coding genes, 576 genes contained at least one homozygous or compound heterozygous variant predicted to result in the loss of function of a protein (LoF, Supplementary Information, 14,516 variants in total). As previously shown[5,17], variants predicted to have the greatest phenotypic impact (LoF and missense variants, and variants mapping to conserved regions), were depleted at the common end of the derived allele spectrum (Extended Data Fig. 3). There were 495 homozygous LoF variants, a subset of which associated with phenotypic outliers (Supplementary Table 3).

We assessed sequence data quality by comparison with an exome sequencing data set (WES, ~50 × coverage)[18] and in 22 pairs of monozygotic twins (Extended Data Fig. 1). The non-reference discordance (NRD, or the fraction of discordant genotypes for non-reference homozygous or heterozygous alleles) was 0.6% for common variants and 3.2% (range 0.1–3.3%; Extended Data Fig. 1) for low-frequency and rare variants. False discovery rates (FDR) were comparable between newly discovered sites and sites previously reported in the 1000 Genomes Project phase 1 (1000GP) data set[5].

When compared to two large-scale European sequencing repositories, 1000GP and the Genome of the Netherlands (GoNL, 12 × read depth[19]), UK10K-cohorts discovered over 24M novel SNVs. Overall, 96.5% of variants with MAF > 1% were shared, reflecting a common reservoir within Europe (Fig. 1 and Extended Data Fig. 2). Conversely, 94.7% of singleton (allele count (AC) = 1) and 55.0% of rare (AC > 1 and MAF < 1%) SNVs were study-specific. In a similar comparison, 64.4% (AC = 1) and 15.8% of variants (AC > 1 and MAF < 1%) found in GoNL were found to be study-specific compared to 1.2% of variants above 1% MAF.

This deeper characterization of European genetic and haplotype diversity will benefit future studies by creating a novel genotype imputation panel with substantially increased coverage and accuracy compared to the 1000GP reference panel[8] (see ref. 9 and the next section for its application). It further informs a detailed empirical assessment of the geographical structure of rare variation in the UK where we detected geographical structure for very rare alleles (AC = 2–7) in Northern and Western UK regions, although this did not show evidence of substantial correlation with variation in phenotype (Box 1).

Findings from single-marker association tests

A main aim of the UK10K-cohorts project was to assess associations of low-frequency and rare variants under different analytical strategies (Fig. 2). We used a unified analysis strategy for the parallel evaluation of all quantitative traits (Supplementary Information, Supplementary Table 4). Here we describe results for the 31 core traits shared in ALSPAC and TwinsUK, with other results reported elsewhere[12].

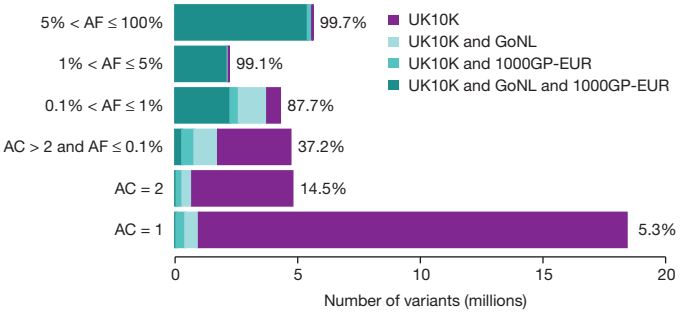


Figure 1 | The UK10K-cohorts resource for variation discovery. Number of SNVs identified in the UK10K-cohorts data set in all autosomal regions in different allele frequency (AF) bins, and percentages that were shared with samples of European ancestry from the 1000 Genomes Project (phase I, EUR $n = 379$) and/or the Genomes of the Netherlands (GoNL, $n = 499$) study, or unique to the UK10K-cohorts data set. AF bins were calculated using the UK10K data set, for allele count (AC) = 1, AC = 2, and non-overlapping AF bins for higher AC. All numerical values are in Extended Data Fig. 2.

We first carried out single-marker association tests, as in standard genome-wide association studies of common variants[20]. Assuming an additive genetic model, we used standard approaches to model relationships between standardised traits, residualized for relevant covariates, and allele dosages of 13,074,236 SNVs, 1,122,542 biallelic INDELs (MAF ≥ 0.1%) and 18,739 large deletions in whole-genome sequenced samples ('WGS sample'). We further assessed associations in an independent study sample of genome-wide genotyped individuals ('GWA' sample) including up to 6,557 ALSPAC and 2,575 TwinsUK participants who were not part of UK10K (actual numbers per trait are given in Supplementary Table 1). In the GWA sample, genotypes were imputed from genome-wide single nucleotide polymorphism (SNP) data using the UK10K haplotype reference panel, described in a companion manuscript[8]. The combined WGS+GWA sample had 80% power to detect associations of SNVs of low-frequency and rare down to ~MAF 0.5%, for a per-alleles trait change (the regression beta coefficient or Beta) of ~1.2 standard deviations or greater (Fig. 3). To combine WGS and GWA data we carried out a fixed effect meta-analysis using the inverse variance method, which showed no evidence of inflation of summary statistics at the traits investigated (GC lambda ≈ 1). We used a conservative stepwise procedure for reporting loci from single-variant analysis (Supplementary Table 5), and we discuss elsewhere replication and technical validation of associations of rare variants not supported in the combined WGS+GWA sample (Supplementary Information, Supplementary Table 6).

Table 1 | Summary of sample collections and sequencing metrics for the four main studies of the UK10K project

Study name and design	*n*	Sequencing strategy, mean read depth and Ts/Tv ratio	SNVs/INDELs	SNVs/INDELs by allele frequency
Cohorts. Unselected samples from two population-based cohorts	3,781	WGS, 7 × Ts/Tv = 2.15	42,001,210/3,490,825	<1%: 34,247,969/2,296,962 1–5%: 2,298,220/412,168 >5%: 5,869,317/1,496,955
Rare. Eight rare diseases with expected different allelic architectures (ciliopathy, coloboma, congenital heart disease, familial hypercholesterolaemia, intellectual disability, neuromuscular, severe insulin resistance and thyroid disease)	961 (397)	WES, 77 × Ts/Tv = 3.02	252,809/ 1,621	<1%: 171,564/1,384 ≥1%: 81,245/237
Obesity. Severely obese children (BMI > 3 s.d. from population mean) and adults with extreme obesity	1,468 (1,359)	WES, 82 × Ts/Tv = 3.02	484,931/ 3,370	<1%: 403,684/3,133 ≥1%: 81,247/237
Neurodevelopmental. Autism and schizophrenia (individual probands, families with one affected and other healthy individuals sampled, families with data from multiple affected individuals and individuals with comorbid intellectual disability and psychosis)	2,753 (1,707)	WES, 77 × Ts/Tv = 3.02	538,526/ 3,826	<1%: 457,278/3,589 ≥1%: 81,248/237

For the cohorts arm, numbers are for the set of 3,781 samples passing quality control, while a subset of 3,621 was used for association testing. For the exome arm, numbers of sites are based on the joint call set, and are calculated for a subset of all individuals that represent the patient subset (in brackets). The total number of individuals sequenced in each study is also given (see Supplementary Methods). The transition to transversion ratio (Ts/Tv) was calculated for the final set of SNVs excluding multiallelic sites. WGS, whole-genome sequencing; WES, whole-exome sequencing.

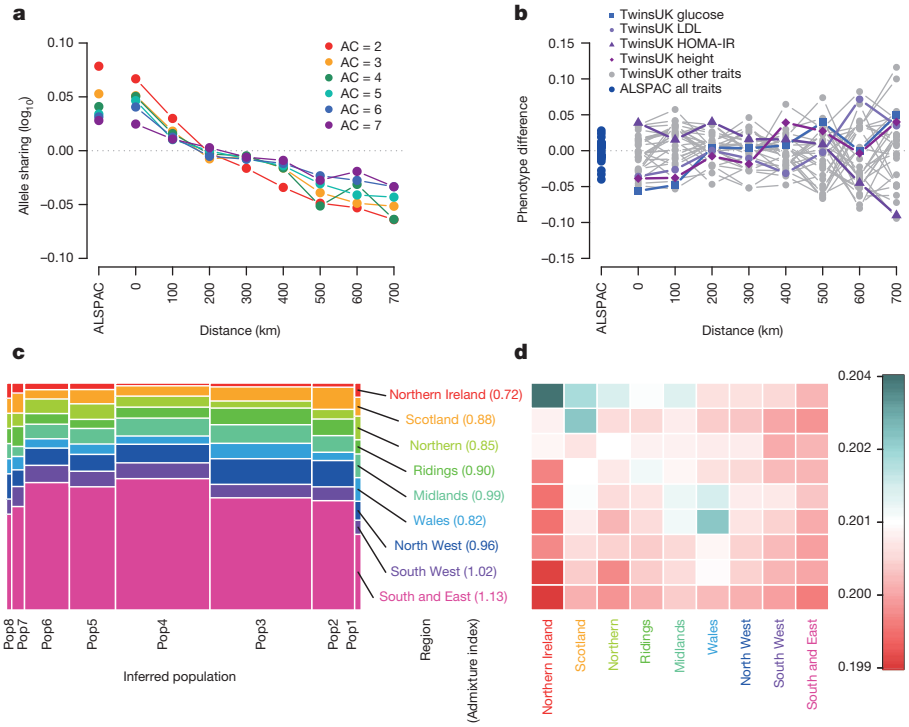
BOX 1

Genetic structure of rare variation within the UK

We used the ALSPAC cohort (from the Bristol region) and a subset of TwinsUK individuals (UK-wide origin) to investigate the spatial structure of rare genetic variants (Supplementary Table 16). We first sought to define the extent to which variants of different MAF were geographically structured. We estimated the excess of allele sharing between pairs of individuals as a function of their physical distance, as compared to expectations under a neutral model (Supplementary Information)[46]. Rare genetic variants showed excess allele sharing at distances smaller than about 200 km, and reduced sharing for more than about 300 km. There was a steeper geographical cline for doubletons (AC = 2), which decreased with increasing allele counts (3 up to 7, equivalent to a MAF of ~0.1–0.3%; **a**). No corresponding geographical structure was observed for phenotypic variation (**b**).

We next assessed the extent to which the non-random distribution of rare SNVs could be accounted for by regional differences at the level of 13 main regions within the UK[47]. Overall, patterns of allele sharing were indicative of a larger degree of genetic homogeneity in Southern and Eastern England compared to individuals of Welsh, Northern, Scottish or Northern Irish origin. Doubletons were the most structured both within and between regions (Wilcoxon rank sum P value <0.05, Extended Data Fig. 8).

Finally, we used "chromosome painting"[48] to gain insights into possible demographic events underlying the observed genetic structure. We first estimated the average length of DNA tracts shared between individuals, and used the number of such tracts to identify fine population structure in our data set. The tract length distribution showed weak geographic structure reflecting the rare variant analysis. A fine structure analysis suggested that the identified populations were not strongly geographically defined, indicative of a large degree of movement between regions compared to the samples in the Peoples of the British Isles study[45], which were chosen to have all four grandparents born in the same location (Extended Data Fig. 9).



Box 1 | Population structure in UK10K-cohorts. All ALSPAC (from Bristol), and 1,139 TwinsUK (UK-wide) participants with a complete set of genotype, phenotype and place of birth data.**a**, Excess of allele sharing as a function of geographical distance, expressed as the proportion of shared alleles between sample pairs for AC from 2 to 7 against their geographical distance. **b**, Phenotypic sharing, estimated for the 31 core phenotypes as the absolute difference between pairs of individuals, averaged within distance bins, rescaled and plotted against their geographical distance. The four traits with the most extreme structure are highlighted. HOMA-IR, homeostatic model assessment for insulin. **c**, Geographical decomposition of each population. Populations are shown proportional to size; historically 'Celtic' and 'Briton' regions are closer to the edges, whereas 'Anglo-Saxon' England is more homogeneous and at the centre (see ref. 45). Ridings refers to East and West Ridings, Yorkshire. **d**, Average length of DNA tracts shared between individuals when clustered by sampling location. The 'admixture' index is given in brackets, with one-third corresponding to regions containing completely unadmixed populations and infinity to completely admixed populations. See also Extended Data Fig. 9.

Overall, across the 31 traits 27 independent loci reached our experiment-wide significance threshold[21] P value $\leq 4.62 \times 10^{-10}$ in the combined WGS+GWA sample (Fig. 3 and Supplementary Table 5). Two associations have been newly discovered by this project, and were conditionally independent of other variants previously reported at the same loci. The first was a low-frequency intronic variant in *ADIPOQ* associated with decreased adiponectin levels (rs74577862-A, effect allele frequency (EAF) = 2.6%, P value = 3.04×10^{-64}). The second was a rare splice variant (rs138326449) in *APOC3* described in advance of this manuscript[11,22,23]. The remaining 25 loci reaching experiment-wide significance in the combined WGS+GWA sample included common, low-frequency and rare variants tagging known associations with adiponectin levels (*CDH13* and *ADIPOQ*), lipid traits (*APOB*, *APOC3-APOA1*, *APOE*, *CETP*, *LIPC*, *LPL*, *PCSK9*, *SORT1-PSRC1-CELSR2*), C-reactive protein (*LEPR*), haemoglobin levels (*HFE*) and fasting glycaemic traits (*G6PC2-ABCB11*, Supplementary Table 5). In contrast to previous projections[24], from this analysis of a wide range of biomedical traits there was no evidence of low-frequency alleles with large effects upon traits (Fig. 3)[25], with classical lipid alleles identifying extremes of single-variant genetic contributions for these traits. This suggests that few, if any, low-frequency variants with stronger effects than those we see

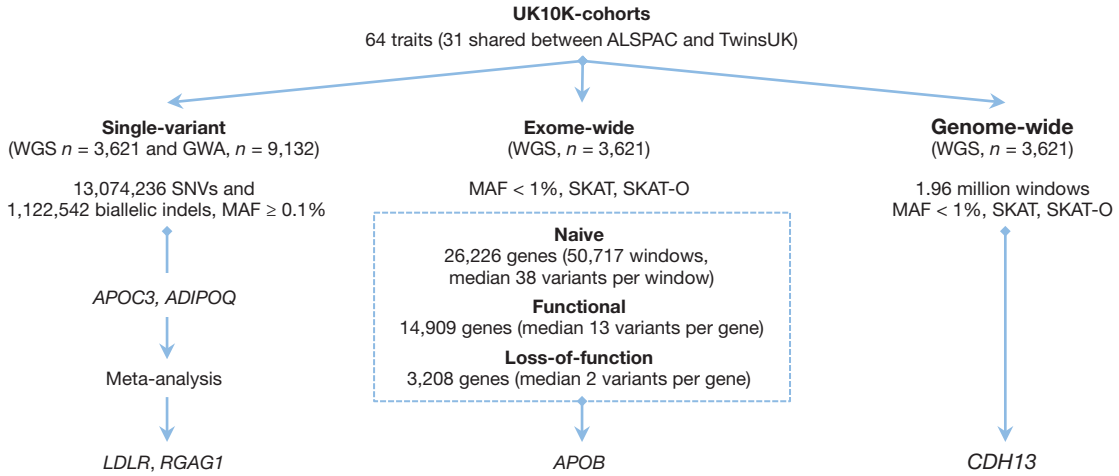


Figure 2 | Study design for associations tested in the UK10K-cohorts study. Summary of phenotype–genotype association testing strategies employed in the UK10K-cohorts study.

are likely to be detected in the general European population for the wide range of traits that we considered.

Increasing sample size may identify additional moderate effect variants, or variants with rarer frequency. We therefore sought to assess the extent to which the more accurate imputation offered by the UK10K reference panel, applied to larger study samples, could discover additional associations. A restricted maximum likelihood (REML)[26] analysis suggested that using the UK10K data could increase the estimated variance explained, compared to the sparser HapMap2, HapMap3 and 1000GP data sets (Extended Data Table 1). We tested four lipid traits (high-density and low-density lipoprotein cholesterol, total cholesterol and triglycerides) in up to 22,082 additional samples from 14 cohorts imputed to the combined UK10K+1000GP phase I panel (Supplementary Table 7).

This effort identified two novel associations with low-density lipoprotein cholesterol (Fig. 3, Supplementary Table 8), which we further replicated in an independent imputation data set of 15,586 samples from 8 cohorts and through genotyping in 95,067 samples from the Copenhagen General Population Study (CGPS[27]). The first was a rare intronic variant in *LDLR* (rs72658867-A, c.2140 + 5G > A; EAF = 0.01, combined sample *P* value = 1.27×10^{-46}); per allele effect Beta (s.e.m.) = −0.23 mmol l^{-1} (0.02), *P* value = 7.63×10^{-30} (CGPS, *n* = 95,079). The second was a common, X-linked variant near *RGAG1* (rs5985471-T, EAF = 0.403, *P* value = 1.53×10^{-12}); per allele effect Beta (s.e.m.) = −0.02 mmol l^{-1} (0.004), *P* value = 1.8×10^{-5} (CGPS, *n* = 93,639). The *LDLR* variant was previously classified to be of uncertain impact in ClinVar, and reported to have no effect on plasma cholesterol levels in a small sample of familial hypercholesterolaemia patients[28]. The *LDLR*-A allele is almost perfectly imputed in our sample (info = 0.96), but absent in previous imputation panels[29]; the *RGAG1*-T allele is common but was missed in previous studies, which focused predominantly on autosomal variation[29]. Within CGPS, these variants were weakly associated with ischaemic heart disease (odds ratio (OR) = 0.77(0.66, 0.92), *P* = 0.003 for rs72658867; 0.96(0.94, 0.99), *P* = 0.005 for rs5985471) and rs72658867 with myocardial infarction (OR = 0.65(0.49, 0.87), *P* = 0.003; Supplementary Table 8). These results demonstrate the value of our expanded haplotype reference panel for discovery of trait associations driven by low-frequency and rare variants, as also shown in refs 9, 10.

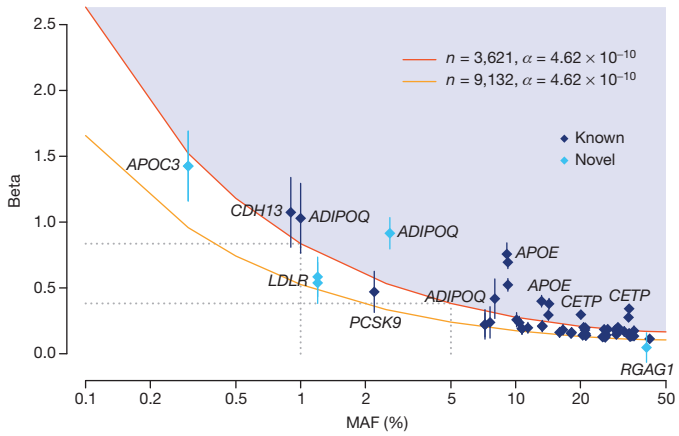


Figure 3 | Summary of association results across the UK10K-cohorts study. Allelic spectrum for single-marker association results for independent variants identified in the single-variant analysis (Supplementary Table 5). A variant's effect (absolute value of Beta, expressed in standard deviation units) is given as a function of minor allele frequency (MAF, *x* axis). Error bars are proportional to the standard error of the beta, variants identifying known loci are dark blue and variants identifying novel signals replicated in independent studies are coloured in light blue. The red and orange lines indicate 80% power at experiment-wide significance level (*t*-test; *P* value ≤4.62×10^{-10}) for the maximum theoretical sample size for the WGS sample and WGS+GWA, respectively.

Findings from rare variant association tests

Single-marker association tests are typically underpowered for rare variants[30]. Many questions remain regarding the optimal choice of test, owing to the unknown allelic architecture of rare variant contribution to traits, in particular outside protein-coding regions. We first evaluated associations by considering genes (GENCODE v15) as functional units of analysis using three separate variant selection strategies. Naive tests considered all variants in exons, untranslated regions (UTRs) and essential splice sites, weighted equally. Functional tests considered missense and LoF variants, the latter defined as being predicted to cause essential splice site changes, stop codon gains or frameshifts. For each scenario we applied two separate statistical models with different properties, sequence kernel association tests (SKAT) and burden tests implemented in SKAT and SKAT-O[31,32], to rare variants (MAF < 1%).

Overall, there was an excess of test statistics with *P* values ≤10^{-4} for functional and loss-of-function tests (Extended Data Figs 4 and 5), with a total of 9, 70 and 196 genes associated with the 31 core traits with the LoF, functional and naive tests, respectively (Supplementary Table 9). A signal driven by loss-of-function variants in the *APOB* gene (encoding apolipoprotein B) achieved our threshold for experiment-wide significance (*P* value ≤1.97×10^{-7}), in a burden-type test (min *P* value for TG = 7.02×10^{-9}). Overall, 3 singleton LoF variants

were responsible for this signal, of which two were not previously reported (rs141422999 and Chr2:21260958). Examples of novel rare variants in complex trait-associated loci (for example, *G6PC2* associated with fasting glucose) were also seen for genes reaching suggestive levels of association (P value $\leq 10^{-4}$). Lastly, we tested the value of a genome-wide naive approach to explore associations outside protein-coding genes by combining variants across ~1.8 million genome-wide tiled windows of 3 kb in size (median 37 SNVs per window, MAF < 1%, assigning an equal weight to all variants in the window). Overall association statistics appeared underpowered to detect true signals, apart from an association signal for adiponectin driven by a known rare intronic variant at the *CDH13* locus (rs12051272, EAF = 0.09%, P value = 6.52×10^{-12}; Supplementary Table 10)[33,34]. As previously shown for single-variant tests, in this study adiponectin and lipid traits yielded the greatest evidence for associations for region-based tests.

Informing studies of low-frequency and rare variants

The UK10K-cohorts data allow an empirical evaluation of the relative importance of increasing sample size, genotyping accuracy or variant coverage for increasing power of genetic discoveries across the allele frequency spectrum. In a companion paper[8] we show that common variants are exhaustively and accurately imputed using current haplotype reference panels, so increasing sample size is likely to be the single most beneficial approach for discovering novel loci driven by common variants. We further show that the UK10K haplotype reference panel, with tenfold more European samples compared to 1000GP, yields substantial improvements in imputation accuracy and coverage for low-frequency and rare variants. To obtain realistic estimates of the power benefit due to imputation with 1000GP+UK10K compared to 1000GP alone, we averaged the smallest value of Beta (the magnitude of a per-allele effect measured in standard deviations) detectable at 80% power, across variants imputable from both reference panels on chromosome 20. Fig. 4a shows sizable reductions in the magnitude of the effect sizes that can be identified at any sample size through use of the UK10K reference panel, compared to the 1000GP panel alone. For instance, for a variant of MAF = 0.3% we have equivalent power when imputing from UK10K+1000GP into a 3,621 sample as we have when using the 1000GP imputation panel alone with 10,000 samples.

Similar, although weaker, increases in power were seen for region-based tests of rare variants. Using the WGS autosome data from UK10K, we used simulation to introduce genotype errors into 220 randomly selected regions of 30 variants each. For each variant, errors were simulated to match the MAF and the observed r^2 values between imputation and sequencing, and between whole-exome and whole-genome sequencing (Supplementary Table 11). We modified the SKAT power calculator[35] to estimate power both for the true genotypes in a region and the data containing error, and averaged results across the 220 regions (see Supplementary Information). Although absolute power in Fig. 4b is generally poor, we can also see demonstrable power improvements when data are better imputed or are directly sequenced (Fig. 4c).

Tests involving non-coding rare variants may further benefit from aggregation strategies driven by biological annotation that takes into consideration the context- and trait-specific impact of non-coding variation[36–38]. Exploiting the denser sequence ascertainment of the UK10K-cohorts, we developed a robust approach to quantify fold-enrichment statistics for different categories of non-coding variants compared to null sets matched for minor allele frequency, local linkage disequilibrium and gene density (Supplementary Information). We used this approach to assess the relative contribution of low-frequency and common variants to associations with five exemplar lipid measures (the study did not have sufficient signal for rarer variants). We considered twelve different functional annotation domains, five in or near protein-coding regions and seven main chromatin segmentation states, defined using data from a cell line informative for lipid traits (HepG2; Supplementary Table 12). Low-frequency variants

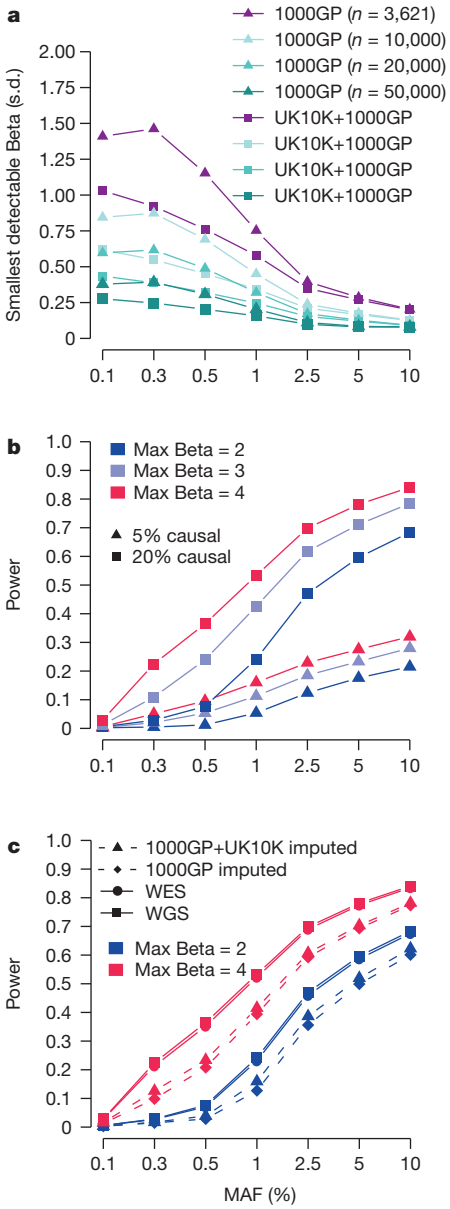


Figure 4 | Power for single-variant and region-based tests. **a**, Strength of single-variant associations detectable at 80% power as a function of MAF and sample size. Using data from chromosome 20[8], we calculated the smallest value of the strength of association Beta (measured in standard deviations), that would be detectable under a linear dosage model, given the MAF and r^2 of each variant imputable from both the 1000GP and the UK10K+1000GP reference panels, for various sample sizes, n. The averages of these minimum detectable beta values by MAF and sample size are shown. **b**, Power of region-based tests in the UK10K-cohorts sample. Evaluations assume $n = 3{,}621$, $\alpha = 6.7 \times 10^{-8}$ and that the proportion of causal variants in the regions is either 5% or 20%, for maximum association (Max Beta) in a region = 2, 3, 4 s.d. **c**, Power of region-based tests and the impact of genotype imputation. Ten regions of 30 variants were randomly sampled from each autosome, and then genotype errors were randomly added to the data following observed r^2 values between genotypes from data imputed from different sources (WGS, high depth WES, GWAS imputed against 1000GP, GWAS imputed against the combined reference panel of 1000GP and UK10K; Supplementary Table 11), and matching the MAF of each variant using the same parameters as in **b**, with the proportion of causal variants in the regions set to 20%.

in exonic regions displayed the strongest degree of enrichment (25-fold, compared to fivefold for common variants, Fig. 5), compatible with the effect of purifying selection[39]. Importantly, however, we showed nearly as strong levels of functional enrichment at both sets of variants for

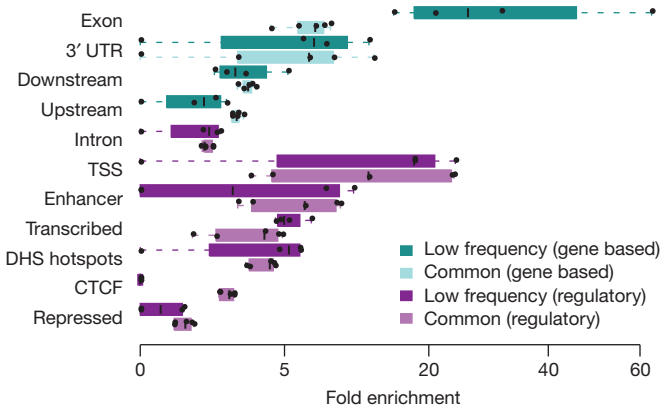


Figure 5 | Enrichment of single-marker associations by functional annotation in the UK10K-cohorts study. Distribution of fold enrichment statistics for single-variant associations of low-frequency (MAF 1–5%) and common (MAF ≥ 5%) SNVs in near-genic elements or selected chromatin states and DNase I hotspots (DHS). Boxplots represent distributions of fold enrichment statistics estimated across the five (out of 31 core) traits where at least 10 independent SNVs were associated with the trait at 10^{-7} *P* value (permutation test) threshold (HDL, LDL, TC, APOA1 and APOB). Chromatin state and DHS regions were inferred from ENCODE data in a liver cell line, HepG2, which is informative for lipids. Promoter and 5′ UTR are not shown, but corresponding statistics are given in Supplementary Table 12.

several non-coding domains (~10- to 20-fold for transcription start sites, DNase I hotspots and 3′ UTRs of genes), confirming the important contribution of non-coding low-frequency alleles to phenotypic trait variance.

Findings from the exome arm of UK10K

In the UK10K-exomes arm studies (see Supplementary Table 13), 5,182 individuals passed sequencing quality control with an average read depth of 80× in the bait regions. We analysed variation discovered in 3,463 disease-affected, unrelated, European-ancestry samples (Supplementary Information). We discovered 842,646 SNVs (of which 1.6% were multiallelic) and 6,067 INDELs. Both variant types were dominated by very rare variants, with more than 60% observed in only one individual. (Extended Data Fig. 6). When compared to European-American samples from the NHLBI Exome Sequencing Project (ESP)[39], we found near-complete overlap at sites with MAF ≥ 1%: 99% of SNVs that are well covered by both projects and pass quality control are present in both data sets. By contrast, 72% of well-covered SNVs seen only once or twice in UK10K are present in ESP. To inform the functional annotation of these variants, we used the Illumina Body Map to determine if the frequency of LoF and functional variants changed when transcripts are selected based on their expression level (Extended Data Fig. 7). When only consequences from highly expressed transcripts and especially those highly expressed in all the Body Map tissues were considered, LoF and functional changes declined. This demonstrates that the choice of transcript can affect the consequence and this should be taken into account when annotating patient exomes.

The rare disease collection studied 1,000 exomes, or ~125 from each of eight rare diseases. Thus far, 25 novel genetic causes have been identified for five of the eight diseases: ciliopathies ($n = 14$), neuromuscular disorders ($n = 7$), eye malformations ($n = 2$), congenital heart defects ($n = 1$) and intellectual disability ($n = 1$; Supplementary Table 14). Notably, there was marked variation in our ability to identify causal variants based on familial recurrence risk, with the primary factors appearing to be: (1) the proportion of patients with a monogenic cause, (2) the strength of prior information about the mode of inheritance (for example, dominant, recessive), and (3) the extent of prior knowledge of the relevant functional pathways. In contrast with our success identifying single-diagnostic variants in these rare diseases, our analysis of three complex diseases (obesity, autism spectrum disorder and schizophrenia) on their own did not yield replicating disease-associated loci. This is perhaps unsurprising given expected locus and allelic heterogeneity, and modest sample size[40]. We therefore engaged in a collaborative meta-analysis as part of the Autism Sequencing Consortium[41] which identified 13 associated genes (FDR < 0.01), many of which have been previously shown to cause intellectual disability or developmental disorders. This suggests that rare variation in single genes can have a large role causing a subset of autism spectrum disorder, but these effects only become apparent when large numbers of individuals are studied.

We also used the UK10K-exomes sequence data to explore the occurrence of incidental findings. We focused on disease-specific genes identified in current guidelines for the analysis of exome/whole-genome data by the American College of Medical Genetics and Genomics (ACMG)[42], and used objective criteria described in the Supplementary Information. We identified a total of 29 distinct reportable variants affecting a total of 2.3% of the UK10K cases considered in this analysis (42 out of 1,805 individuals), a number similar to previous estimates (2% estimate in adults of European ancestry[43]). The incidental findings were predominantly associated with cardiovascular disorders (Supplementary Table 15).

Two main challenges of reporting incidental findings from whole-exome surveys emerge. The need for clinical expertise, the difficulty of interpreting a fraction of variants, and the lack of completeness of the ClinVar database[44] all highlighted the need to further consolidate knowledge from the community into freely accessible and more exhaustive databases. Furthermore, for some disorders, the frequency of carriers is likely to be too high compared to the disease frequency, despite our strict assessment criteria. This suggests that reported estimates of the penetrance of recognized variants for specific disorders are too high. Given these challenges, we suggest that, in the absence of additional evidence, scientific publications describing proposed penetrant associations for rare variants need to be complemented by accurate estimates of population frequencies.

Conclusions

In summary we have generated a high-quality whole-genome sequence data repository including 24 million novel variants from nearly 4,000 European-ancestry individuals. We showed that the UK10K haplotype reference panel greatly increases accuracy and coverage of low-frequency and rare variants compared to existing panels such as the 1000GP phase 1 panel. We carried out a large-scale empirical exploration of association testing of common, low-frequency and rare genetic variants with a large variety of biomedically important quantitative traits. For each of the different association scenarios tested, we report first examples of novel alleles associated with lipid and adiponectin traits. This provides proof-of-principle evidence on the value of the large-scale sequencing data for complex traits, while also indicating that there are few low-frequency large effect 'quick wins' that make substantial contributions to population trait variation and that can be discovered from sequencing studies of few thousands individuals. Our power calculations, informed by the sequence data, provide realistic estimates of the benefit of sequencing versus imputation in future association studies. Finally, rare variation tests showed limited evidence for confounding owing to population stratification at the traits investigated, likely to be due to a weakening of historical patterns of population structure in the current general UK population[45].

Overall, this effort has given us both new genomic tools[12] and insights into the role of low-frequency and rare variation on human complex traits, and will inform strategies for future association studies. Our exploration of non-coding variants supports the need for incorporating functional genome information in association tests of rare

variants outside protein-coding regions. Improved study power through larger numbers, and a better understanding of the observed heterogeneity in allelic architecture between different loci, are likely to provide the best route forward to describe the contribution of rare variants to phenotypic variance in health and disease, and for assessing their utility in healthcare.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 March 2015; accepted 17 July 2015.

Published online 14 September; corrected online 30 September 2015 (see full-text HTML version for details).

1. Manolio, T. A. Bringing genome-wide association findings into clinical use. *Nature Rev. Genet.* **14,** 549–558 (2013).
2. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8,** e1002793 (2012).
3. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13,** 101 (2011).
4. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genet.* **46,** 220–224 (2014).
5. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
6. Lange, L. A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94,** 233–245 (2014).
7. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
8. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Commun.* **6,** 8111 (2015).
9. Zheng, H. *et al.* Whole-genome sequencing identifies *EN1* as a determinant of bone density and fracture. *Nature* http://dx.doi.org/10.1038/nature14878 (2015).
10. Taylor, P. N. *et al.* Whole-genome sequence-based analysis of thyroid function. *Nature Commun.* **6,** 5681 (2015).
11. Timpson, N. J. *et al.* A rare variant in *APOC3* is associated with plasma triglyceride and VLDL levels in Europeans. *Nature Commun.* **5,** 4871 (2014).
12. Geihs, M. *et al.* An interactive genome browser of association results from the UK10K cohorts project. *Bioinformatics.* http://dx.doi.org/10.1093/bioinformatics/btv491 (2015).
13. 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,** 111–127 (2013).
14. Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **16,** 144–149 (2013).
15. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21,** 940–951 (2011).
16. Wheeler, E. *et al.* Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nature Genet.* **45,** 513–517 (2013).
17. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genet.* **47,** 435–444 (2015).
18. Williams, F. M. *et al.* Genes contributing to pain sensitivity in the normal population: an exome sequencing study. *PLoS Genet.* **8,** e1003095 (2012).
19. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genet.* **46,** 818–825 (2014).
20. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).
21. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* **38,** 281–290 (2014).
22. Jørgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjærg-Hansen, A. Loss-of-function mutations in *APOC3* and risk of ischemic vascular disease. *N. Engl. J. Med.* **371,** 32–41 (2014).
23. The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute. Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. *N. Engl. J. Med.* **371,** 22–31 (2014).
24. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9,** 356–369 (2008).
25. Park, J. H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl Acad. Sci. USA* **108,** 18026–18031 (2011).
26. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42,** 565–569 (2010).
27. 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. *J. Am. Med. Assoc.* **298,** 299–308 (2007).
28. Whittall, R. A., Matheus, S., Cranston, T., Miller, G. J. & Humphries, S. E. The intron 14 2140+5G>A variant in the low density lipoprotein receptor gene has no effect on plasma cholesterol levels. *J. Med. Genet.* **39,** e57 (2002).
29. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
30. Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* **44,** 293–308 (2010).
31. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).
32. Liu, D. J. & Leal, S. M. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *Am. J. Hum. Genet.* **91,** 585–596 (2012).
33. Morisaki, H. *et al. CDH13* gene coding T-cadherin influences variations in plasma adiponectin levels in the Japanese population. *Hum. Mutat.* **33,** 402–410 (2012).
34. Dastani, Z. *et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8,** e1002607 (2012).
35. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13,** 762–775 (2012).
36. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **488,** 57–74 (2012).
37. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnol.* **30,** 224–226 (2012).
38. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
39. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
40. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111,** E455–E464 (2014).
41. Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15,** 565–574 (2013).
42. Kaye, J. *et al.* Managing clinically significant findings in research: the UK10K example. *Eur. J. Hum. Genet.* **22,** 1100–1104 (2014).
43. Amendola, L. M. *et al.* Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res.* **25,** 305–315 (2015).
44. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2014).
45. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519,** 309–314 (2015).
46. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genet.* **44,** 243–246 (2012).
47. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
48. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8,** e1002453 (2012).
49. Benjamini, Y. & Hochberg, Y. controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).

Supplementary Information is available in the online version of the paper.

Acknowledgements This study makes use of data generated by the UK10K Consortium. The Wellcome Trust provided funding for UK10K (WT091310). Additional grant support and acknowledgements can be found in the Supplementary Information.

Author Contributions Project management: D.M., K.R.; designed individual studies and contributed data: A.A., A.Do., A.G.M., A.I., A.Ma., A.McI., A.McQ., A.Mor., A.O., A.R.F., A.T.H., A.Val., A.Var., B.H.S., B.N., C.B., C.C., C.M.v., C.W., Cl.L., D.A., D.B., D.B.S., D.Co., D.Cu., D.Ge., D.Gr., D.H., D.J.P., D.R.F., D.S.-C., D.S., D.T., E.M.v., E.St., E.Z., F.M., F.Z., G.B., G.Cl., G.D., G.G., G.L., G.Mal., G.S., G.Z., H.Gu., H.M.M., H.W., I.L., I.N.M.D., I.S.F., J.B., J.C., J.C.C., J.H., J.J., J.Keo., J.L.M., J.Lö., J.Lu., J.Mo., J.R.P., J.S.K., J.Suv., J.Wal., K.A.W., K.Ch., K.J.W., K.N., L.G., F.L.R., L.S., M.A., M.Be., M.C.O., M.Ca., M.Co., M.D.T., M.E.K., M.J.O., M.M., M.S., M.T., N.C., N.J.T., N.R., N.Sc., N.So., O.S., P.Be., P.Bo., P.G., P.Ho., P.M., P.Sc., P.W., R.A., R.B., R.K.S., R.M., Ri.S., S.Bh., S.Ci., S.Cu., S.E.H., S.G.W., S.I.S., S.O., S.R., T.D.S., T.G., T.P., T.W., V.I., V.Pa., W.M., UCLEB Consortium†; generated and/or quality controlled sequence data: A.Da., A.K.-K., C.J., Co.L., D.K.J., D.M., F.Z., G.Co., G.W., H.L., J.H., J.Li., J.Mas., J.St., J.Sun., J.T., K.Wo., M.A.Q., P.C., P.D., P.E., P.F., R.D., Ru.L., Ry.L., S.Ba., S.E., S.McC., T.C., T.K., Xi.G., Y.D.; designed new statistical or bioinformatics tools: A.C., A.H., B.H., C.M.T.G., C.X., E.Bi., E.Z., G.R.S.R., H.S., I.D., I.T., J.Mar., K.O., N.So., Ru.L., S.Me., T.D., T.H., V.I.; analysed the data and provided critical interpretation of results: A.D.-W., A.H., A.M.V., A.Moa., A.P., A.S., J.B.R., B.S., C.A., C.M.T.G., C.K.R., C.S.F., C.W., D.E., D.G.M., D.L., E.Bo., E.Se., E.W., E.Z., F.K., F.P., F.Z., G.Mar., G.R.S.R., H.Z., I.B., I.M., I.T., J.C.B., J.F., J.H., J.Kem., J.L.M., J.Mo., J.R.B.P., J.Y., K.Ca., K.P., K.S., K.Wa., K.Wo., L.Ch., L.Cr., L.P., L.Q., L.R.L., L.S., L.V.W., M.Co., M.E.H., M.F., M.G., M.Le., M.S.-A., M.v., N.J.T., N.M., N.So., O.P., P.D., P.Hy., P.M.V., P.Sy., P.V., R.C., R.C.P., R.D., R.E., R.L.R., R.T., Ri.S., S.-Y.S., S.A., S.E.H., S.G.W., S.McC., S.Me., S.P., S.S., T.G., V.I., V.Pl., Y.J., Y.M.; ethics: A.K., C.S., D.M., D.R.F., F.M., H.Gr., J.Ka., K.K., F.L.R., M.Bo., M.E.H., N.J.T., P.Bo., R.D., R.K.S., T.D.S.; designed and/or managed the study: A.P., J.B.R., Co.L., D.M., D.R.F., E.Z., G.D.-S., I.S.F., J.C.B., J.Ka., J.St., K.K., M.E.H., M.J.O., N.J.T., N.So., R.D., S.McC., T.D.S.; wrote the manuscript: A.H., J.B.R., C.M.T.G., C.X., D.L., E.Z., I.B., J.C.B., J.F., J.H., J.L.M., J.R.B.P., K.Wa., L.Cr., M.E.H., M.F., N.J.T., N.So., P.D., R.D., Ru.L., S.E.H., S.McC., S.S., V.I., V.Pl., Y.M.

Author Information Data access form is available at http://www.uk10k.org/data_access.html, raw and processed data files at https://www.ebi.ac.uk/ega/, imputation panel at https://www.ebi.ac.uk/ega/, UK10K Genome Browser at http://www.uk10k.org/dalliance.html, single-marker loci navigator at http://fathmm.biocompute.org.uk/UK10K_Browser/ and dynamic power calculator at http://fathmm.biocompute.org.uk/UK10K_Browser/Power.htm. All sequence and

phenotype data were deposited to the European Genome-Phenome archive (EGA, https://www.ebi.ac.uk/ega/), with accession numbers EGAD00001000740, EGAD00001000789, EGAD00001000741, EGAD00001000790, EGAD00001000776, EGAD00001000433, EGAD00001000434, EGAD00001000435, EGAD00001000436, EGAD00001000613, EGAD00001000614, EGAD00001000437, EGAD00001000438, EGAD00001000615, EGAD00001000439, EGAD00001000440, EGAD00001000441, EGAD00001000442, EGAD00001000443, EGAD00001000430, EGAD00001000431, EGAD00001000432, EGAD00001000429, EGAD00001000413, EGAD00001000414, EGAD00001000415, EGAD00001000416, EGAD00001000417, EGAD00001000418, EGAD00001000419 and EGAD00001000420. A breakdown of studies is given in Supplementary Table 13. All study participants provided informed consent. Details of REC approvals are given in Supplementary Table 17.

Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.S. (ns6@sanger.ac.uk) or R.D. (rd@sanger.ac.uk).

The UK10K Project Consortium

Writing group

Klaudia Walter[1]*, Josine L. Min[3]*, Jie Huang[1]*, Lucy Crooks[1,4]*, Yasin Memari[1], Shane McCarthy[1], John R. B. Perry[5,6], ChangJiang Xu[7,8], Marta Futema[9], Daniel Lawson[10], Valentina Iotchkova[1,11], Stephan Schiffels[1], Audrey E. Hendricks[1,12], Petr Danecek[1], Rui Li[7,13,14], James Floyd[1,15], Louise V. Wain[16], Inês Barroso[1,17], Steve E. Humphries[9], Matthew E. Hurles[1], Eleftheria Zeggini[1], Jeffrey C. Barrett[1], Vincent Plagnol[18], J. Brent Richards[5,7,8,13,14], Celia M. T. Greenwood[7,8,14,19], Nicholas J. Timpson[3], Richard Durbin[1], Nicole Soranzo[1,2]

Production group

Senduran Bala[1], Peter Clapham[1], Guy Coates[1], Tony Cox[1], Allan Daly[1], Petr Danecek[1], Yuanping Du[20], Richard Durbin[1], Sarah Edkins[1], Peter Ellis[1], Paul Flicek[1,11], Xiaosen Guo[20,21], Xueqin Guo[20], Liren Huang[20], David K. Jackson[1], Chris Joyce[1], Thomas Keane[1], Anja Kolb-Kokocinski[1], Cordelia Langford[1], Yingrui Li[20], Jieqin Liang[20], Hong Lin[20], Ryan Liu[22], John Maslen[1], Shane McCarthy (co-chair)[1], Dawn Muddyman[1], Michael A. Quail[1], Jim Stalker (co-chair)[1], Jianping Sun[7,8], Jing Tian[20], Guangbiao Wang[20], Jun Wang[20,21,23,24,25], Yu Wang[20], Kim Wong[1], Pingbo Zhang[20]

Cohorts group

Inês Barroso[1,17], Ewan Birney[11], Chris Boustred[26], Lu Chen[1,2], Gail Clement[5], Massimiliano Cocca[27,28], Petr Danecek[1], George Davey Smith[3], Ian N. M. Day[29], Aaron Day-Williams[1,30], Thomas Down[1,31], Ian Dunham[11], Richard Durbin[1], David M. Evans[3,32], Tom R. Gaunt[3], Matthias Geihs[1], Celia M. T. Greenwood[7,8,14,19], Deborah Hart[5], Audrey E. Hendricks[1,12], Bryan Howie[33], Jie Huang[1], Tim Hubbard[1,31], Pirro Hysi[5], Valentina Iotchkova[1,11], Yalda Jamshidi[34], Konrad J. Karczewski[35,36], John P. Kemp[3,32], Genevieve Lachance[5], Daniel Lawson[10], Monkol Lek[35], Margarida Lopes[1,37,38], Daniel G. MacArthur[35,36], Jonathan Marchini[37,39], Massimo Mangino[5,40], Iain Mathieson[41], Shane McCarthy[1], Yasin Memari[1], Sarah Metrustry[5], Josine L. Min[3], Alireza Moayyeri[5,42], Dawn Muddyman[1], Kate Northstone[3], Kalliope Panoutsopoulou[1], Lavinia Paternoster[3], John R. B. Perry[5,6], Lydia Quaye[5], J. Brent Richards (co-chair)[5,7,8,13,14], Susan Ring[43], Graham R. S. Ritchie[1,11], Stephan Schiffels[1], Hashem A. Shihab[3], So-Youn Shin[1,3], Kerrin S. Small[5], María Soler Artigas[16], Nicole Soranzo (co-chair)[1,2], Lorraine Southam[1,37], Timothy D. Spector[5], Beate St Pourcain[3,44,45], Gabriela Surdulescu[5], Ioanna Tachmazidou[1], Nicholas J. Timpson (co-chair)[3], Martin D. Tobin[16,46], Ana M. Valdes[5], Peter M. Visscher[32,47], Louise V. Wain[16], Klaudia Walter[1], Kirsten Ward[5], Scott G. Wilson[5,48,49], Kim Wong[1], Jian Yang[32,47], Eleftheria Zeggini[1], Feng Zhang[5], Hou-Feng Zheng[7,13,14]

Neurodevelopmental disorders group

Richard Anney[50], Muhammad Ayub[51], Jeffrey C. Barrett[1], Douglas Blackwood[52], Patrick F. Bolton[53,54,55], Gerome Breen[54,55], David A. Collier[55,56], Nick Craddock[57], Lucy Crooks[1,4], Sarah Curran[53,58,59], David Curtis[60], Richard Durbin[1], Louise Gallagher[50], Daniel Geschwind[61], Hugh Gurling[62]‡, Peter Holmans[57], Irene Lee[63], Jouko Lönnqvist[64], Shane McCarthy[1], Peter McGuffin[55], Andrew M. McIntosh[52], Andrew G. McKechanie[52,65], Andrew McQuillin[62], James Morris[1], Dawn Muddyman[1], Michael C. O'Donovan[57], Michael J. Owen (co-chair)[57], Aarno Palotie (co-chair)[1,66,67], Jeremy R. Parr[68], Tiina Paunio[64,69], Olli Pietilainen[1,64,66], Karola Rehnström[1], Sally I. Sharp[62], David Skuse[63], David St Clair[70], Jaana Suvisaari[64], James T. R. Walters[57], Hywel J. Williams[57,71]

Obesity group

Inês Barroso (co-chair)[1,17], Elena Bochukova[17], Rebecca Bounds[17], Anna Dominiczak[72], Richard Durbin[1], I. Sadaf Farooqi (co-chair)[17], Audrey E. Hendricks[1,12], Julia Keogh[17], Gaëlle Marenne[1], Shane McCarthy[1], Andrew Morris[73], Dawn Muddyman[1], Stephen O'Rahilly[17], David J. Porteous[74], Blair H. Smith[75], Ioanna Tachmazidou[1], Eleanor Wheeler[1], Eleftheria Zeggini[1]

Rare disease group

Saeed Al Turki[1,76], Carl A. Anderson[1], Dinu Antony[77], Inês Barroso[1,17], Phil Beales[77], Jamie Bentham[78], Shoumo Bhattacharya[78], Mattia Calissano[79], Keren Carss[1], Krishna Chatterjee[17], Sebahattin Cirak[79,80], Catherine Cosgrove[78], Richard Durbin[1], David R. Fitzpatrick (co-chair)[81], James Floyd[1,15], A. Reghan Foley[79], Christopher S. Franklin[1], Marta Futema[9], Detelina Grozeva[82], Steve E. Humphries[9], Matthew E. Hurles (co-chair)[1], Shane McCarthy[1], Hannah M. Mitchison[77], Dawn Muddyman[1], Francesco Muntoni[79], Stephen O'Rahilly[17], Alexandros Onoufriadis[31], Victoria Parker[17], Felicity Payne[1], Vincent Plagnol[18], F. Lucy Raymond[82], Nicola Roberts[82], David B. Savage[17], Peter Scambler[77], Miriam Schmidts[77,83], Nadia Schoenmakers[17], Robert K. Semple[17], Eva Serra[1], Olivera Spasic-Boskovic[82], Elizabeth Stevens[79], Margriet van Kogelenberg[1], Parthiban Vijayarangakannan[1], Klaudia Walter[1], Kathleen A. Williamson[81], Crispian Wilson[82], Tamieka Whyte[79]

Statistics group

Antonio Ciampi[8], Celia M. T. Greenwood (co-chair)[7,8,14,19], Audrey E. Hendricks[1,12], Rui Li[7,13,14], Sarah Metrustry[5], Karim Oualkacha[84], Ioanna Tachmazidou[1], ChangJiang Xu[7,8], Eleftheria Zeggini (co-chair)[1]

Ethics group

Martin Bobrow[82], Patrick F. Bolton[53,54,55], Richard Durbin[1], David R. Fitzpatrick[81], Heather Griffin[85], Matthew E. Hurles (co-chair)[1], Jane Kaye (co-chair)[85], Karen Kennedy[1,86], Alastair Kent[87], Dawn Muddyman[1], Francesco Muntoni[79], F. Lucy Raymond[82], Robert K. Semple[17], Carol Smee[1], Timothy D. Spector[5], Nicholas J. Timpson[3]

Incidental findings group

Ruth Charlton[88], Rosemary Ekong[89], Marta Futema[9], Steve E. Humphries[9], Farrah Khawaja[90], Luis R. Lopes[91,92], Nicola Migone[93], Stewart J. Payne[94], Vincent Plagnol (chair)[18], Rebecca C. Pollitt[95], Sue Povey[89], Cheryl K. Ridout[96], Rachel L. Robinson[88], Richard H. Scott[77,97], Adam Shaw[98], Petros Syrris[91], Rohan Taylor[90], Anthony M. Vandersteen[99]

Management committee

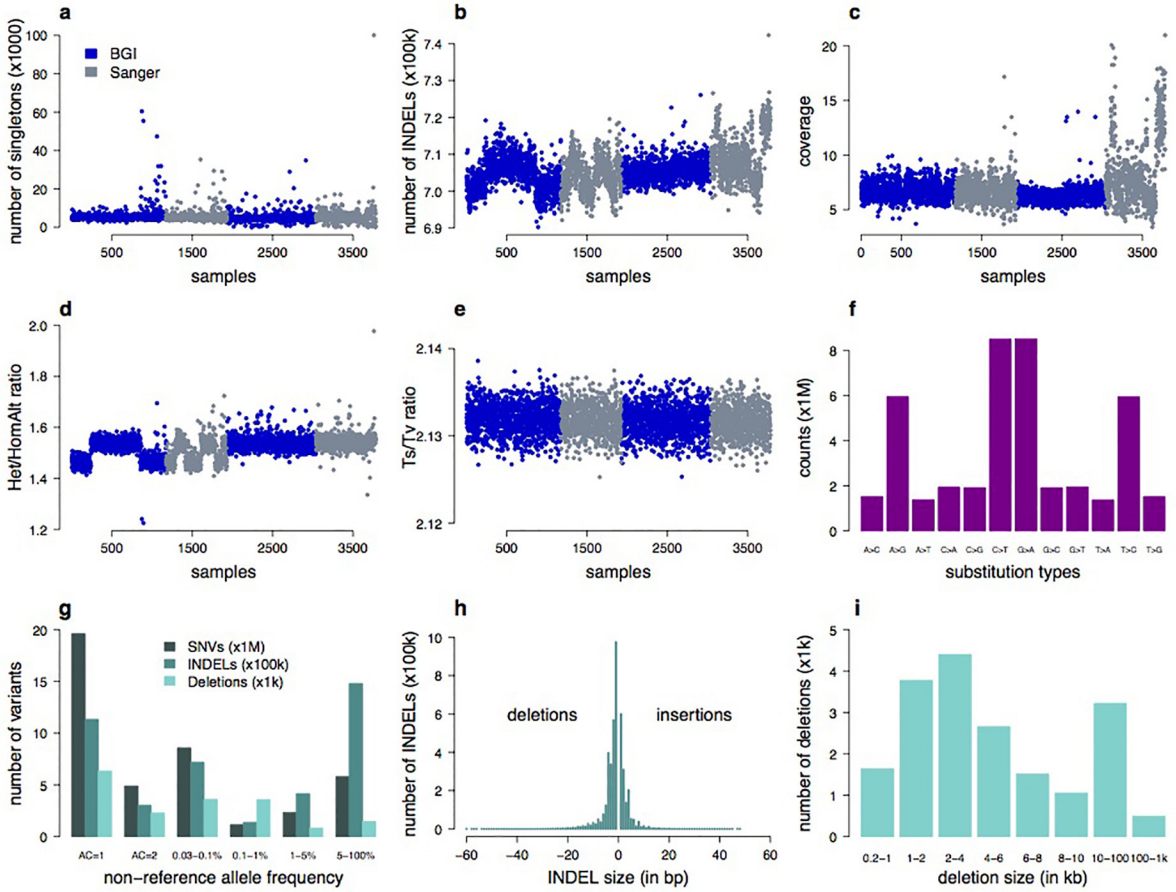
Jeffrey C. Barrett[1], Inês Barroso[1,17], George Davey Smith[3], Richard Durbin (chair)[1], I. Sadaf Farooqi[17], David R. Fitzpatrick[81], Matthew E. Hurles[1], Jane Kaye[85], Karen Kennedy[1,86], Cordelia Langford[1], Shane McCarthy[1], Dawn Muddyman[1], Michael J. Owen[57], Aarno Palotie[1,66,67], J. Brent Richards[5,7,8,13,14], Nicole Soranzo[1,2], Timothy D. Spector[5], Jim Stalker[1], Nicholas J. Timpson[3], Eleftheria Zeggini[1]

Lipid meta-analysis group

Antoinette Amuzu[100], Juan Pablo Casas[91,100], John C. Chambers[101], Massimiliano Cocca[27,28], George Dedoussis[102], Giovanni Gambaro[103], Paolo Gasparini[27,28,104], Tom R. Gaunt[3], Jie Huang[1], Valentina Iotchkova[1,11], Aaron Isaacs[105], Jon Johnson[106], Marcus E. Kleber[107], Jaspal S. Kooner[108], Claudia Langenberg[109], Jian'an Luan[109], Giovanni Malerba[110], Winfried März[111,112,113], Angela Matchan[1], Josine L. Min[3], Richard Morris[114], Børge G. Nordestgaard[115,116], Marianne Benn[115,116], Susan Ring[43], Robert A. Scott[109], Nicole Soranzo[1,2], Lorraine Southam[1,37], Nicholas J. Timpson[3], The UCLEB Consortium†, Daniela Toniolo[117], Michela Traglia[117], Anne Tybjaerg-Hansen[116,118], Cornelia M. van Duijn[105], Elisabeth M. van Leeuwen[105], Anette Varbo[115,116], Peter Whincup[119], Gianluigi Zaza[120], Eleftheria Zeggini[1], Weihua Zhang[101]

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1HH, Cambridge, UK. [2]Department of Haematology, University of Cambridge, Long Road, Cambridge CB2 0PT, UK. [3]MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. [4]Sheffield Diagnostic Genetics Service, Sheffield Childrens' NHS Foundation Trust, Western Bank, Sheffield S10 2TH, UK. [5]The Department of Twin Research & Genetic Epidemiology, King's College London, St Thomas' Campus, Lambeth Palace Road, London SE1 7EH, UK. [6]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Box 285, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. [7]Lady Davis Institute, Jewish General Hospital, Montreal, Quebec H3T 1E2, Canada. [8]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec H3A 1A2, Canada. [9]Cardiovascular Genetics, BHF Laboratories, Rayne Building, Institute of Cardiovascular Sciences, University College London, London WC1E 6JJ, UK. [10]Schools of Mathematics and Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. [11]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [12]Department of Mathematical and Statistical Sciences, University of Colorado, Denver, Colorado 80204, USA. [13]Department of Medicine, Jewish General Hospital, McGill University, Montreal, Quebec H3A 1B1, Canada. [14]Department of Human Genetics, McGill University, Montreal, Quebec H3A 1B1, Canada. [15]The Genome Centre, John Vane Science Centre, Queen Mary, University of London, Charterhouse Square, London EC1M 6BQ, UK. [16]Departments of Health Sciences and Genetics, University of Leicester, Leicester LE1 7RH, UK. [17]University of Cambridge Metabolic Research Laboratories, and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [18]University College London (UCL) Genetics Institute (UGI) Gower Street, London WC1E 6BT, UK. [19]Department of Oncology, McGill University, Montreal, Quebec H2W 1S6, Canada. [20]BGI-Shenzhen, Shenzhen 518083, China. [21]Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark. [22]BGI-Europe, London

EC2M 4YE, UK. [23]Princess Al Jawhara Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, P.O. Box 80200, Jeddah 21589, Saudi Arabia. [24]Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. [25]Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong. [26]North East Thames Regional Genetics Service, Great Ormond Street Hospital NHS Foundation Trust, London WC1N 3JH, UK. [27]Medical Genetics, Institute for Maternal and Child Health IRCCS "Burlo Garofolo", 34100 Trieste, Italy. [28]Department of Medical, Surgical and Health Sciences, University of Trieste, 34100 Trieste, Italy. [29]Bristol Genetic Epidemiology Laboratories, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. [30]Computational Biology & Genomics, Biogen Idec, 14 Cambridge Center, Cambridge, Massachusetts 02142, USA. [31]Department of Medical and Molecular Genetics, Division of Genetics and Molecular Medicine, King's College London School of Medicine, Guy's Hospital, London SE1 9RT, UK. [32]University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland 4102, Australia. [33]Adaptive Biotechnologies Corporation, Seattle, Washington 98102, USA. [34]Human Genetics Research Centre, St George's University of London, London SW17 0RE, UK. [35]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [36]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [37]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [38]Illumina Cambridge Ltd, Chesterford Research Park, Cambridge CB10 1XL, UK. [39]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [40]National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' Foundation Trust, London SE1 9RT, UK. [41]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [42]Institute of Health Informatics, Farr Institute of Health Informatics Research, University College London (UCL), 222 Euston Road, London NW1 2DA, UK. [43]ALSPAC & School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. [44]School of Oral and Dental Sciences, University of Bristol, Lower Maudlin Street, Bristol BS1 2LY, UK. [45]School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK. [46]National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester LE3 9QP, UK. [47]Queensland Brain Institute, University of Queensland, Brisbane, Queensland 4072, Australia. [48]School of Medicine and Pharmacology, University of Western Australia, Perth, Western Australia 6009, Australia. [49]Department of Endocrinology and Diabetes, Sir Charles Gairdner Hospital, Nedlands, Western Australia 6009, Australia. [50]Department of Psychiatry, Trinity Centre for Health Sciences, St James Hospital, James's Street, Dublin 8, Ireland. [51]Division of Developmental Disabilities, Department of Psychiatry, Queen's University, Kingston, Ontario N6C 0A7, Canada. [52]Division of Psychiatry, The University of Edinburgh, Royal Edinburgh Hospital, Edinburgh EH10 5HF, UK. [53]Department of Child Psychiatry, Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, London SE5 8AF, UK. [54]NIHR BRC for Mental Health, Institute of Psychiatry, Psychology and Neuroscience and SLaM NHS Trust, King's College London, 16 De Crespigny Park, London SE5 8AF, UK. [55]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, Denmark Hill, London SE5 8AF, UK. [56]Lilly Research Laboratories, Eli Lilly & Co. Ltd., Erl Wood Manor, Sunninghill Road, Windlesham GU20 6PH, UK. [57]MRC Centre for Neuropsychiatric Genetics & Genomics, Institute of Psychological Medicine & Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff CF24 4HQ, UK. [58]University of Sussex, Brighton BN1 9RH, UK. [59]Sussex Partnership NHS Foundation Trust, Swandean, Arundel Road, Worthing BN13 3EP, UK. [60]University College London (UCL), UCL Genetics Institute, Darwin Building, Gower Street, London WC1E 6BT, UK. [61]UCLA David Geffen School of Medicine, Los Angeles, California 90095, USA. [62]University College London (UCL), Molecular Psychiatry Laboratory, Division of Psychiatry, Gower Street, London WC1E 6BT, UK. [63]Behavioural and Brain Sciences Unit, UCL Institute of Child Health, London WC1N 1EH, UK. [64]National Institute for Health and Welfare (THL), Helsinki FI-00271, Finland. [65]The Patrick Wild Centre, The University of Edinburgh, Edinburgh EH10 5HF, UK. [66]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki FI-00014, Finland. [67]Program in Medical and Population Genetics and Genetic Analysis Platform, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02132, USA. [68]Institute of Neuroscience, Henry Wellcome Building for Neuroecology, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [69]University of Helsinki, Department of Psychiatry, Helsinki FI-00014, Finland. [70]Institute of Medical Sciences, University of Aberdeen, Aberdeen AB25 2ZD, UK. [71]The Centre for Translational Omics – GOSgene, UCL Institute of Child Health, London WC1N 1EH, UK. [72]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Wolfson Medical School Building, University Avenue, Glasgow, G12 8QQ, UK. [73]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, 9 Little France Road, Edinburgh EH16 4UX, UK. [74]Centre for Genomic and Experimental Medicine, Institute of Genetics and Experimental Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. [75]Mackenzie Building, Kirsty Semple Way, Ninewells Hospital and Medical School, Dundee DD2 4RB, UK. [76]Department of Pathology, King Abdulaziz Medical City, P.O. Box 22490, Riyadh 11426, Saudi Arabia. [77]Genetics and Genomic Medicine and Birth Defects Research Centre, UCL Institute of Child Health, London WC1N 1EH, UK. [78]Department of Cardiovascular Medicine and Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [79]Dubowitz Neuromuscular Centre, UCL Institute of Child Health & Great Ormond Street Hospital, London WC1N 1EH, UK. [80]Institut für Humangenetik, Uniklinik Köln, Kerpener Strasse 34, 50931 Köln, Germany. [81]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, at the University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [82]Academic Laboratory of Medical Genetics, Box 238, Lv 6 Addenbrooke's Treatment Centre, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [83]Human Genetics Department, Radboudumc and Radboud Institute for Molecular Life Sciences (RIMLS), Geert Grooteplein 25, Nijmegen 6525 HP, The Netherlands. [84]Department of Mathematics, Université de Québec À Montréal, Montréal, Québec H3C 3P8, Canada. [85]HeLEX – Centre for Health, Law and Emerging Technologies, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK. [86]National Cancer Research Institute, Angel Building, 407 St John Street, London EC1V 4AD, UK. [87]Genetic Alliance UK, 4D Leroy House, 436 Essex Road, London N1 3QP, UK. [88]Leeds Genetics Laboratory, St James University Hospital, Beckett Street, Leeds LS9 7TF, UK. [89]University College London (UCL) Department of Genetics, Evolution & Environment (GEE), Gower Street, London WC1E 6BT, UK. [90]SW Thames Regional Genetics Lab, St George's University, Cranmer Terrace, London SW17 0RE, UK. [91]Institute of Cardiovascular Science, University College London, Gower Street, London WC1E 6BT, UK. [92]Cardiovascular Centre of the University of Lisbon, Faculty of Medicine, University of Lisbon, Avenida Professor Egas Moniz,1649-028 Lisbon, Portugal. [93]Department of Medical Sciences, University of Torino, 10124 Torino, Italy. [94]North West Thames Regional Genetics Service, Kennedy-Galton Centre, Northwick Park Hospital, Watford Road, Harrow HA1 3UJ, UK. [95]Connective Tissue Disorders Service, Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Foundation Trust, Western Bank, Sheffield S10 2TH, UK. [96]Molecular Genetics, Viapath at Guy's Hospital, London SE1 9RT, UK. [97]Department of Clinical Genetics, Great Ormond Street Hospital, London, WC1N 3JH, UK. [98]Clinical Genetics, Guy's & St Thomas' NHS Foundation Trust, London SE1 9RT, UK. [99]Maritime Medical Genetics Service, 5850/5980 University Avenue, PO Box 9700, Halifax, Nova Scotia B3K 6R8, Canada. [100]London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [101]The Department of Epidemiology and Biostatistics, Imperial College London, St Mary's campus, Norfolk Place, Paddington, London W2 1PG, UK. [102]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens 17671, Greece. [103]Division of Nephrology and Dialysis, Institute of Internal Medicine, Renal Program, Columbus-Gemelli University Hospital, Catholic University, 00168 Rome, Italy. [104]Experimental Genetics Division, Sidra, P.O. Box 26999 Doha, Qatar. [105]Genetic Epidemiology Unit, Department of Epidemiology, Erasmus MC, Rotterdam 3000 CA, Netherlands. [106]Department of Quantitative Social Science, UCL Institute of Education, University College London, 20 Bedford Way, London WC1H 0AL, UK. [107]Vth Department of Medicine, Medical Faculty, Mannheim 68167, Germany. [108]National Heart and Lung Institute, Imperial College London, London W12 0NN, UK. [109]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. [110]Biology and Genetics, Department of Life and Reproduction Sciences, University of Verona, 37134 Verona, Italy. [111]Clinical Institute of Medical and Chemical Laboratory Diagnostics, Medical University of Graz, Graz 8036, Austria. [112]Synlab Academy, Synlab Services GmbH, D-68161 Mannheim, Germany. [113]Medical Clinic V (Nephrology, Hypertensiology, Rheumatology, Endocrinolgy, Diabetology), Mannheim Medical Faculty, Heidelberg University, Mannheim 68167, Germany. [114]School of Social and Community Medicine, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK. [115]Department of Clinical Biochemistry and The Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark. [116]The Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark. [117]Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Milan 20132, Italy. [118]Department of Clinical Biochemistry KB3011, Rigshospitalet, Copenhagen University Hospital, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. [119]Population Health Research Institute, St George's University of London, London SW17 0RE, UK. [120]Renal Unit, Department of Medicine, University of Verona, 37126 Verona, Italy.
†A list of authors and affiliations appears in the Supplementary Information.
‡Deceased.
*These authors contributed equally to this work.



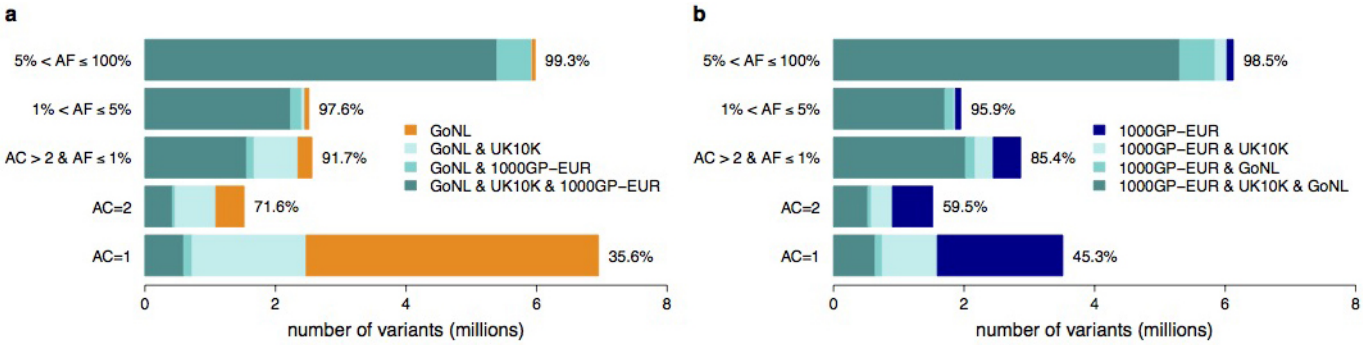
j

	Number of variants	
AF	SNVs	INDELs
AC = 1	19,596,845	1,132,608
AC = 2	4,890,329	301,105
0.03 − 1%	9,719,435	857,840
1 − 5%	2,332,114	415,735
> 5%	5,787,895	1,479,767

k

	WGS versus Exomes						MZ Twins	
AF	Total sites (concordant, %)	Non-Ref genotypes (NRD, %)	FP (FDR, %)	FP in 1000GP (FDR, %)	FP not in 1000GP (FDR, %)	FNR (%)	Total sites (concordant, %)	Non-ref genotypes (NRD, %)
AC = 1	2,963 (99.999)	2,965 (0.1)	125 (4.0)	11 (3.8)	114 (4.1)	n.a.	411,583 (99.995)	3,534 (12.7)
AC = 2	1,566 (99.998)	1,577 (0.1)	147 (8.6)	25 (7.9)	122 (8.7)	n.a.	101,116 (99.989)	1,594 (15.1)
0.03−1%	16,303 (99.928)	21,114 (3.3)	1,160 (6.6)	766 (5.5)	394 (11.3)	27.2	193,531 (99.954)	19,034 (10.2)
1 − 5%	16,356 (99.829)	53,165 (3.2)	1,038 (6.0)	980 (5.7)	58 (68.2)	6.4	50,360 (99.776)	56,554 (4.4)
> 5%	37,433 (99.688)	1,151,178 (0.6)	2,668 (6.7)	2,653 (6.6)	15 (46.9)	7.3	123,690 (99.574)	1,382,934 (0.8)

Extended Data Figure 1 | UK10K-cohorts, sequence and sample quality and variation metrics. **a**–**e**, Sample quality metrics for UK10K-cohorts ($n = 3{,}781$) where $n = 1$–$1{,}927$ corresponds to ALSPAC and 1,928 to 3,781 to TwinsUK. This sample includes all individuals passing sample quality control, including related pairs and non-European individuals that were later removed from association tests. A subset of 3,621 individuals was included in association analyses. Samples sequenced at BGI are coloured in blue and samples sequenced at Sanger are coloured in grey. **a**, Number of singletons (AC = 1) by sample ($\times 10^3$). **b**, Number of INDELs by sample ($\times 10^5$). **c**, Read depth (sequence coverage) by sample. **d**, Ratio of heterozygous and homozygous non-reference (=homozygous alternative) SNV genotypes (mean for females = 1.54, mean for males = 1.47). **e**, Transition to transversion ratio (Ts/Tv) by sample. **f**–**i**, Sequence variation metrics for UK10K-cohorts. **f**, Types of substitution ($\times 10^6$). **g**, Number of SNVs ($\times 10^6$), INDELs ($\times 10^5$) and large deletions ($\times 10^3$) by non-overlapping non-reference allele frequency (AF) bins. **h**, Size distribution of INDELs. Negative INDEL lengths represent deletions and positive INDEL lengths represent insertions. **i**, Large deletion size distribution in unequal bin sizes where the smallest deletions were 200 bp to 1 kb long and the largest deletions 100 kb to 1 Mb. In total 18,739 deletions were called with GenomeSTRiP[14]. The average deletion size was ~13 kb and the median size was ~3.7 kb. **j**, Total number of SNVs and INDELs by AF bin (based on 3,781 samples), multi-allelic variants are treated as separate variants. **k**, Sequence quality and variation metrics for UK10K-cohorts. For 61 overlapping TwinsUK individuals we compared the variant sites and genotypes of the low-coverage sequences with high-coverage exome data by non-overlapping AF bins (WGS versus Exomes). We considered 74,621 shared sites in non-overlapping AF bins. We calculated the fraction of concordant over total sites, the number of non-reference genotypes and non-reference genotype discordance (NRD, in %) between WGS and Exomes; false discovery rate (FDR = FP/(FP + TP); TP, true positive; FP, false positive), where we consider the exomes as the truth set; number of false positives (FP) and FDR for sites that are or not shared with the 1000 Genomes Project, phase I (1000GP); false negative rate (FNR = FN/(FN + TP); FN, false negative; TP, true positive), where AF bins were defined based on the 61 exomes. Furthermore, we compared 22 monozygotic twin pairs at 880,280 bi-allelic SNV sites on chromosome 20, reporting the percentage of concordant genotypes, non-reference genotypes and NRD. AFs are from the set of 3,621 samples, which contains at most one of the two monozygotic twins from each pair. We note that discrepancies can be caused by errors in either twin, so the expected NRD to the truth would be half the NRD value given.



c

AF	Total UK10K	UK10K & GoNL & 1000GP-EUR (% total)	UK10K & GoNL (% total)	UK10K & 1000GP-EUR (% total)	UK10K only (% total)
AC = 1	18,414,734	49,267 (0.3)	605,284 (3.3)	412,700 (2.2)	17,446,017 (94.7)
AC = 2	4,787,082	57,915 (1.2)	456,853 (9.5)	294,748 (6.2)	4,093,396 (85.5)
0.03 − 0.1%	4,706,597	287,250 (6.1)	1,251,513 (26.6)	784,729 (16.7)	2,957,605 (62.8)
0.1 − 1%	4,277,695	2,252,555 (52.7)	3,404,131 (79.6)	2,598,915 (60.8)	527,204 (12.3)
1 − 5%	2,198,285	2,136,712 (97.2)	2,156,903 (98.1)	2,158,931 (98.2)	19,163 (0.9)
> 5%	5,611,847	5,406,635 (96.3)	5,413,902 (96.5)	5,587,743 (99.6)	16,837 (0.3)

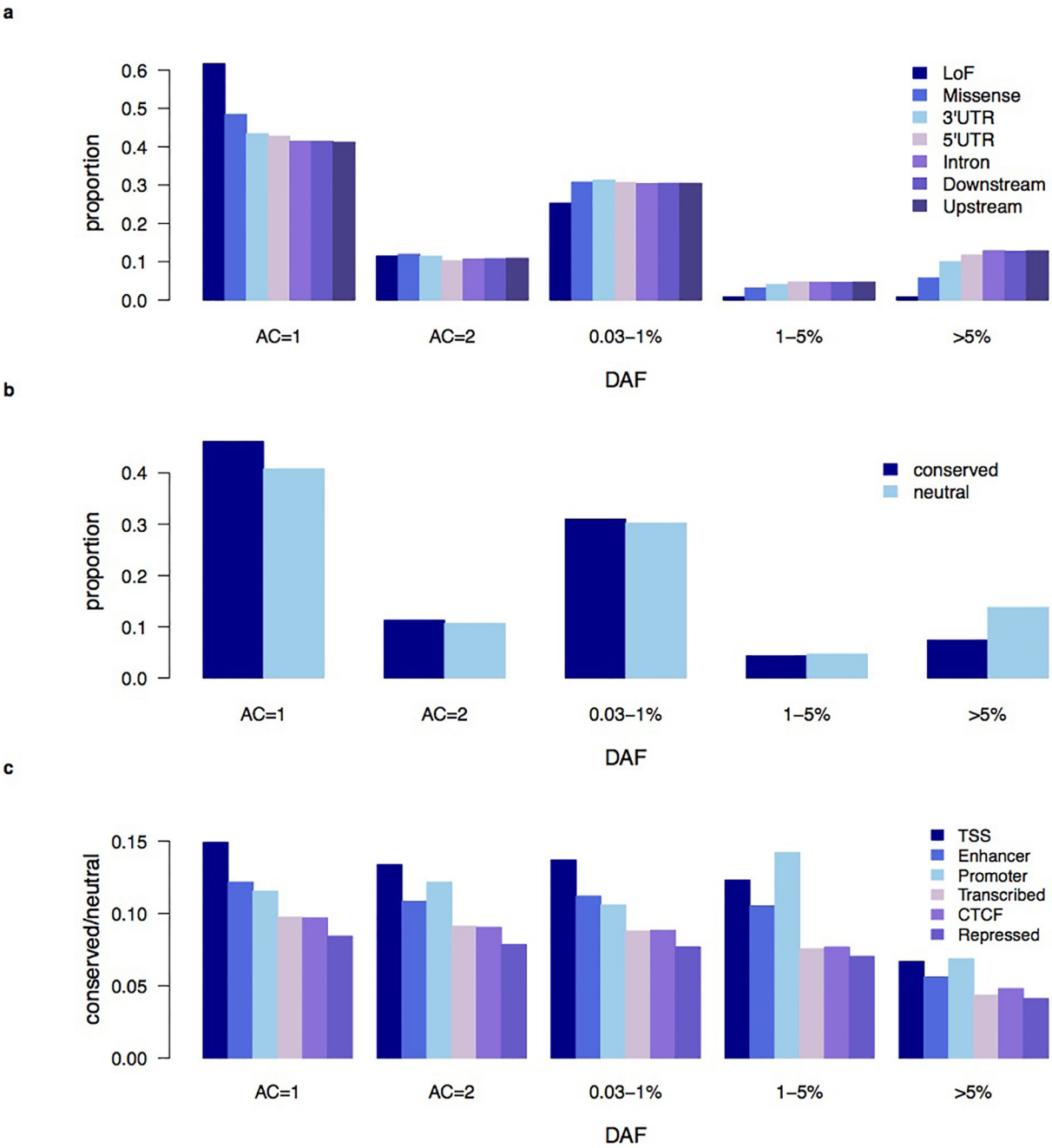
d

AF	Total GoNL	GoNL & UK10K & 1000GP-EUR (% total)	GoNL & UK10K (% total)	GoNL & 1000GP-EUR (% total)	GoNL only (% total)
AC = 1	6,947,211	593,530 (8.5)	2,340,962 (33.7)	723,843 (10.4)	4,475,936 (64.4)
AC = 2	1,518,021	420,990 (27.7)	1,041,369 (68.6)	466,613 (30.7)	431,029 (28.4)
0.03 − 1%	2,557,845	1,556,624 (60.9)	2,230,239 (87.2)	1,671,746 (65.4)	212,484 (8.3)
1 − 5%	2,510,847	2,228,873 (88.8)	2,278,668 (90.8)	2,401,884 (95.7)	59,168 (2.4)
> 5%	5,976,721	5,390,317 (90.2)	5,397,348 (90.3)	5,925,087 (99.1)	44,603 (0.7)

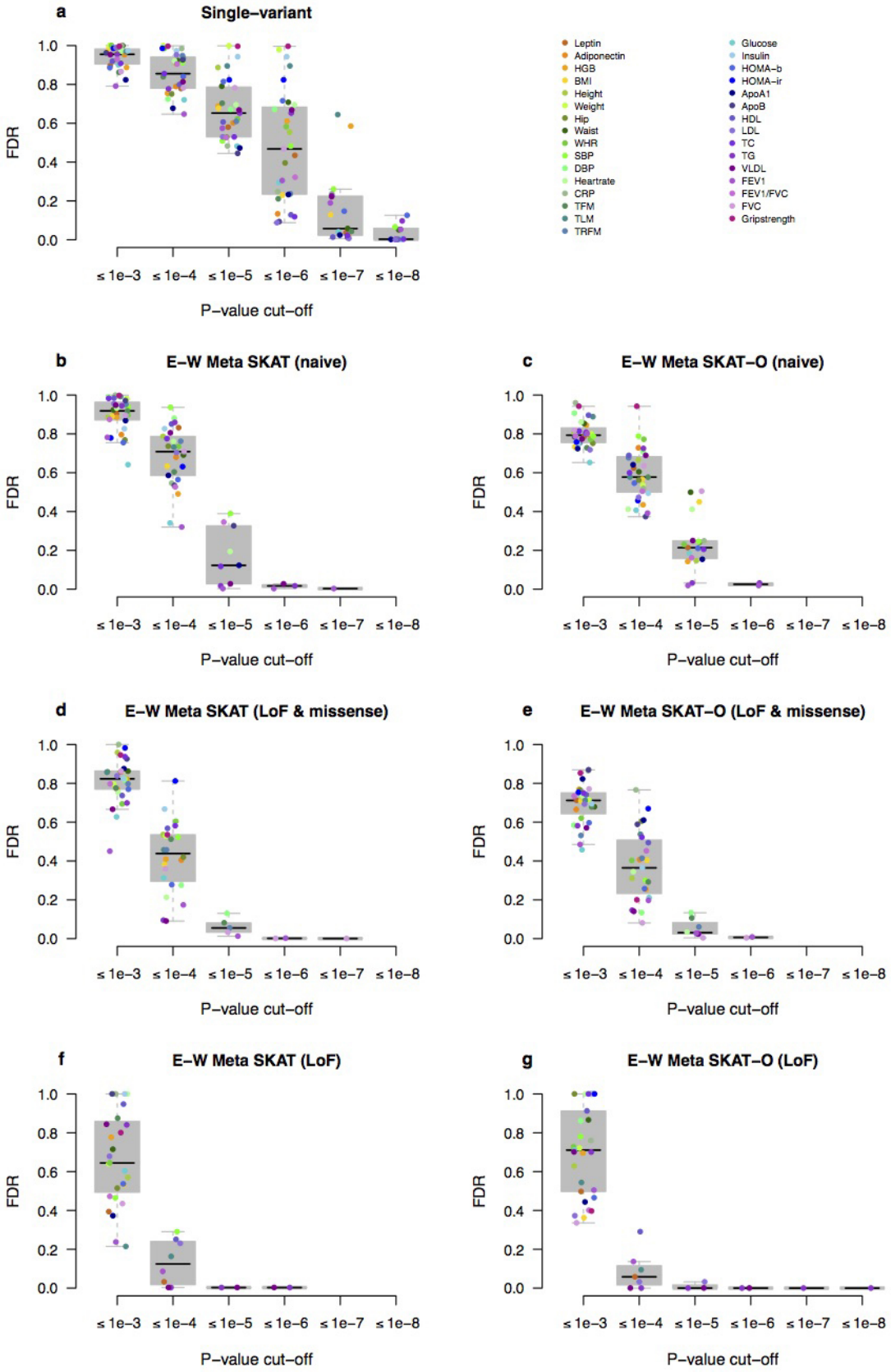
e

AF	Total 1000GP-EUR	1000GP-EUR & UK10K & GoNL (% total)	1000GP-EUR & UK10K (% total)	1000GP-EUR & GoNL (% total)	1000GP-EUR only (% total)
AC = 1	3,511,032	638,386 (18.2)	1,481,516 (42.2)	747,118 (21.3)	1,920,784 (54.7)
AC = 2	1,517,131	525,808 (34.7)	846,398 (55.8)	581,707 (38.3)	614,834 (40.5)
0.03 − 1%	2,866,389	2,019,676 (70.5)	2,297,344 (80.1)	2,170,476 (75.7)	418,245 (14.6)
1 − 5%	1,947,958	1,704,958 (87.5)	1,729,262 (88.8)	1,843,712 (94.6)	79,942 (4.1)
> 5%	6,122,720	5,301,506 (86.6)	5,483,246 (89.6)	5,846,160 (95.5)	94,820 (1.5)

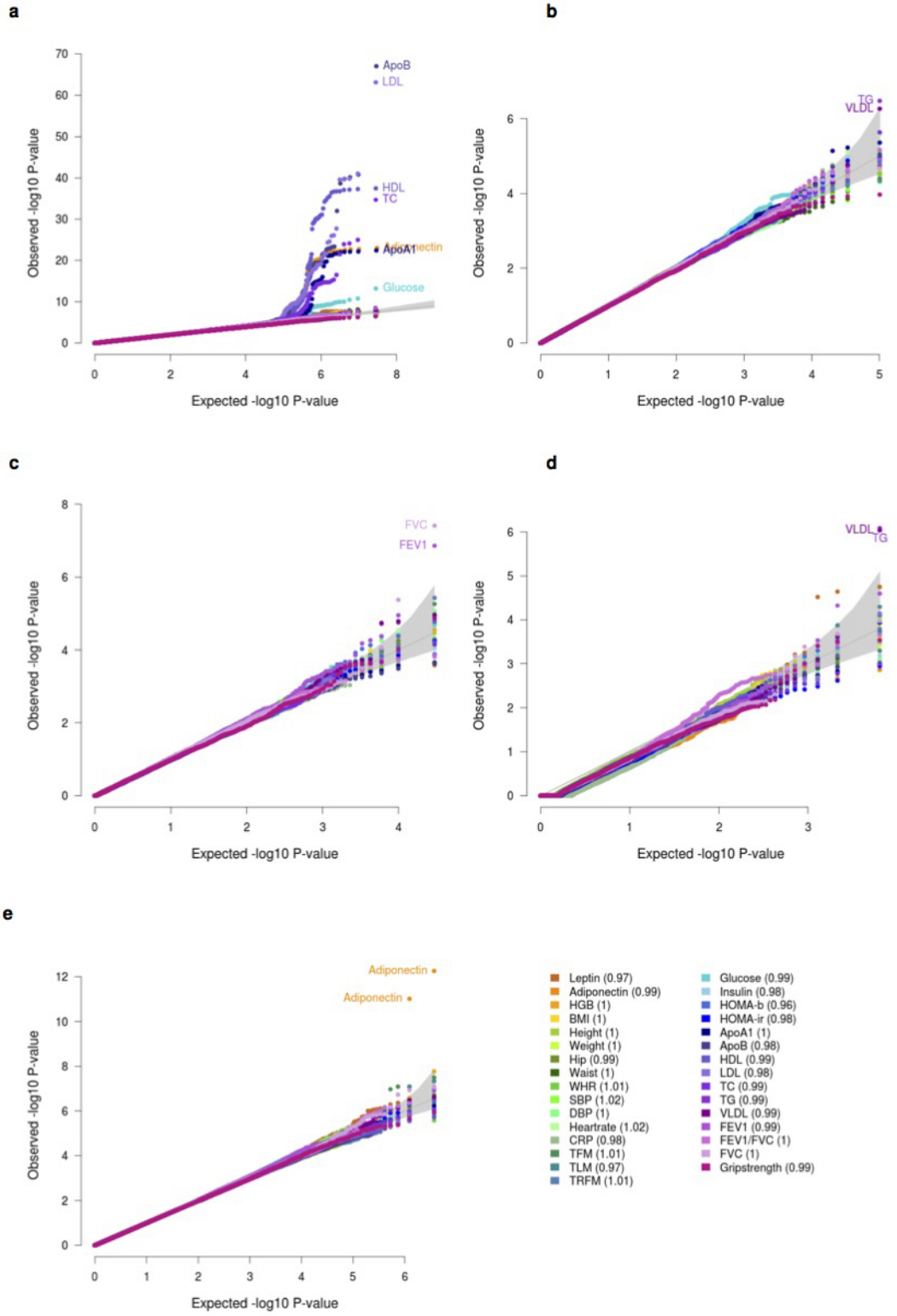
Extended Data Figure 2 | UK10K-cohorts, comparison with GoNL and 1000GP-EUR. Percentage of autosomal SNVs that are either shared between UK10K ($n = 3{,}781$), GoNL ($n = 499$) and 1000GP-EUR ($n = 379$), or unique to each set, for allele counts (AC) AC = 1, AC = 2, and non-overlapping allele frequency (AF) bins for higher AC. **a**, Shared and unique variants for GoNL with AF based on GoNL, and **b**, for 1000GP-EUR. AF bins are not directly comparable owing to the different sample sizes in each call set. The *x*-axis shows the number of variants in millions. The percentages next to the bars represent the percentage of variants from GoNL (**a**) and 1000GP-EUR (**b**) that are shared with at least one of the other data sets. All numerical values used in **a** can be found in **d** and for **b** in **e**. **c**, Numerical values for Fig. 1.



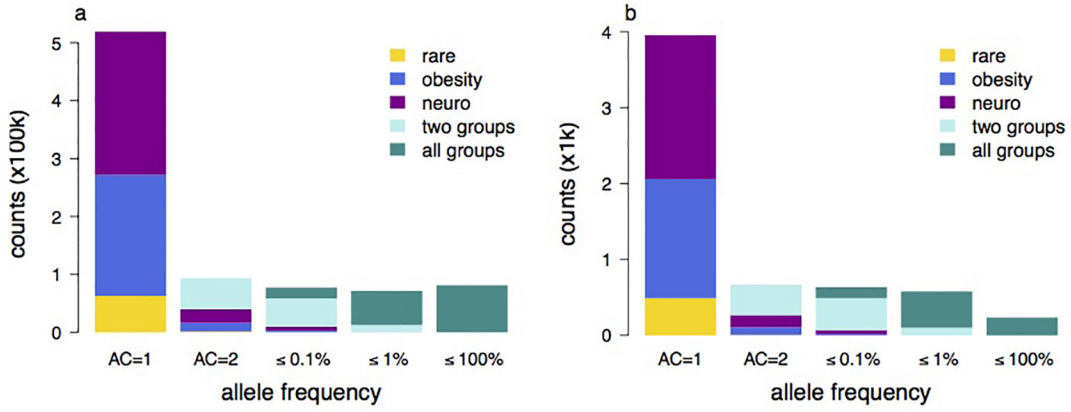
Extended Data Figure 3 | UK10K-cohorts, derived allele frequency spectrum by functional annotation. Derived allele frequency (DAF) spectrum for UK10K-cohorts chromosome 20 variants divided by functional class. **a**, Proportion of total variants (standardized across DAF bins) as a function of DAF for different genic elements. **b**, Standardized proportion of all variants by DAF bin, and divided into conserved (GERP >2) versus neutral (GERP ≤ 2) sites. **c**, Ratio of conserved versus neutral variants by DAF bin, and classified by chromatin segmentation domains defined by ENCODE as detailed in the methods.



Extended Data Figure 4 | UK10K-cohorts, false discovery rate (FDR). **a**–**g**, FDR values for reporting associations at different *P* value cut-offs for all analyses reported in this study and the 31 core traits for single-variant analysis (**a**); naive exome-wide Meta SKAT (**b**); naive exome-wide Meta SKAT-O (**c**); functional exome-wide Meta SKAT (LoF and missense) (**d**); functional exome-wide Meta SKAT-O (LoF and missense) (**e**); functional exome-wide Meta SKAT (LoF) (**f**); functional exome-wide Meta SKAT-O (LoF) (**g**).



Extended Data Figure 5 | UK10K-cohorts, QQ plots. QQ plots for the association tests of the 31 core traits in the WGS data set ($n = 3{,}621$ individuals). **a**, Single-variant analysis (~14 million variants with MAF $\geq$ 0.1%); **b**, naive exome-wide Meta SKAT (1,783,548 variants with MAF $<$ 1% in 50,717 windows); **c**, functional exome-wide Meta SKAT (LoF and missense; 256,733 variants with MAF $<$ 1% in 14,909 windows); **d**, loss-of-function functional exome-wide Meta SKAT (LoF; 9,113 variants with MAF $<$ 1% in 3,208 windows); **e**, genome-wide Meta SKAT (35,858,684 variants with MAF $<$ 1% in 1,845,982 windows).



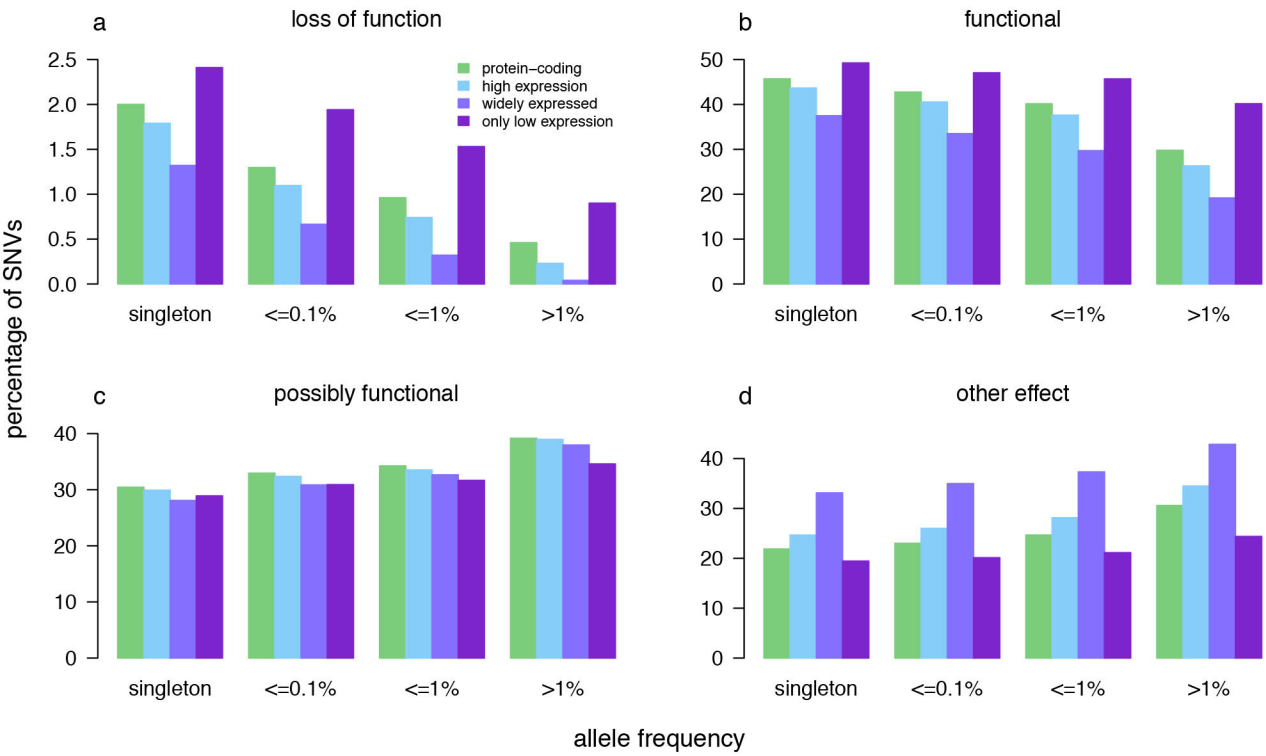
c

		Disease Collection					
	AF	Total	All groups	Two groups	Rare	Obesity	Neurodevelop-mental
SNVs	$AC = 1$	518,966	0	0	62,989	208,700	247,277
	$AC = 2$	93,601	0	54,213	1,392	15,103	22,893
	$AC = 2 < AF \leq 0.1\%$	77,199	18,644	49,140	86	3,008	6,321
	$0.1 < AF \leq 1\%$	71,634	58,958	12,575	0	19	82
	$AF > 1\%$	81,246	81,242	4	0	0	0
	AF	Total	All groups	Two groups	Rare	Obesity	Neurodevelop-mental
INDELs	$AC = 1$	3,954	0	0	491	1,568	1,895
	$AC = 2$	666	0	400	6	100	160
	$AC = 2 < AF \leq 0.1\%$	635	144	422	0	16	53
	$0.1 < AF \leq 1\%$	578	477	101	0	0	0
	$AF > 1\%$	234	234	0	0	0	0

d

	UK10K exomes		EA ESP	
AF	Total (% shared with EA ESP)	Inside baits (% shared with EA ESP)	Total (% shared with UK10K)	Inside baits (% shared with UK10K)
$AC = 1$	518,966 (21)	273,856 (23)	681,351 (16)	587,453 (15)
$AC = 2$	93,601 (48)	48,863 (53)	135,426 (41)	112,556 (39)
$AC = 2 < AF \leq 0.1\%$	77,199 (74)	38,935 (81)	132,937 (72)	107,041 (70)
$AF > 0.1\%$	152,880 (89)	69,989 (99)	201,062 (96)	141,215 (99)
Total	842,646 (41)	431,643 (44)	1,150,776 (40)	948,265 (37)

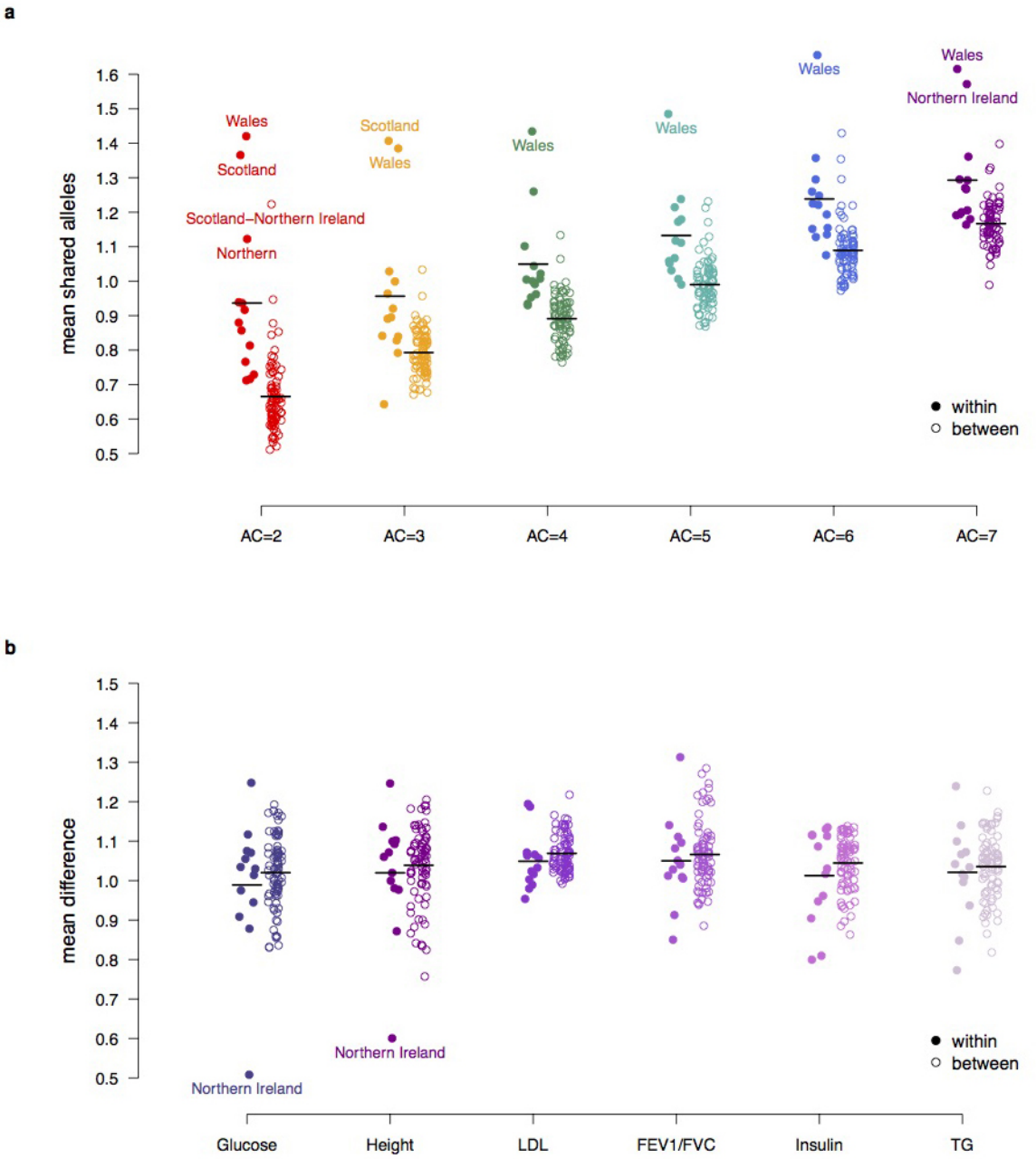
Extended Data Figure 6 | UK10K-exomes, sequence variant statistics. Number of variants ($\times 10^3$) that are found in one or more of the three UK10K-exomes disease data sets, as a function of allele frequency (AF) of the non-reference allele. Variants are split into allele counts (AC) $AC = 1$, $AC = 2$ and non-overlapping AF bins for $AC > 2$. Allele frequency is the frequency of the alternative allele. The distributions of SNVs and INDELs across frequencies and disease collections are similar, except that there is a lower proportion of INDELs with $AF > 1\%$ compared to SNVs. **a**, SNVs. Multiallelic sites are included (1.6%), and non-reference alleles at the same site are treated as separate variants. **b**, INDELs. Counts are given in **c**. **c**, Variants are classed by whether they were found in more than one disease collection or unique to a specific group. **d**, Comparison of UK10K patient set with European-Americans individuals from the NHLBI Exome Sequencing project (EA ESP). The left panel shows the variants identified in UK10K and the percentage shared with EA ESP. Both the total number of variants and the number within the EA ESP bait regions (intersection of bait sets) are given. The right panel shows the variants identified in EA ESP and the percentage shared with UK10K. Both the total number of variants, and the number within the UK10K baits after removing any that failed UK10K quality control, are given. There is some overlap in the ranges of AC and AF for EA ESP variants because different numbers of individuals were included.



e

Transcript set	Frequency	LoF	Functional	Possibly functional	Other	No qualifying transcripts
All	AC = 1	10,647	234,169	178,067	95,667	416
	AC = 2	1,338	40,013	34,430	17,747	73
	AF ≤ 0.1%	988	32,020	28,905	15,226	60
	AF ≤ 1%	751	28,264	27,651	14,900	68
	AF > 1%	446	23,721	36,084	20,860	135
Protein-coding	AC = 1	10,127	231,825	154,433	110,928	11,653
	AC = 2	1,254	39,538	29,836	20,634	2,339
	AC = 2 < AF ≤ 0.1%	905	31,613	25,007	17,661	2,013
	AF ≤ 1%	664	27,876	23,776	17,124	2,194
	AF > 1%	361	23,123	30,413	23,747	3,602
High expression	AC = 1	6,343	154,247	105,711	87,097	56,948
	AC = 2	760	26,062	20,365	16,097	10,846
	AC = 2 < AF ≤ 0.1%	504	20,663	16,946	13,866	9,167
	AF ≤ 1%	352	17,981	16,023	13,450	9,007
	AF > 1%	123	13,787	20,422	18,073	12,831
Widely expressed	AC = 1	1,304	36,957	27,694	32,654	311,737
	AC = 2	131	5,909	5,301	5,863	56,926
	AC = 2 < AF ≤ 0.1%	76	4,525	4,308	5,030	47,207
	AF ≤ 1%	40	3,684	4,050	4,629	44,410
	AF > 1%	5	2,440	4,834	5,456	52,501
Only low expression	AC = 1	1,093	22,305	13,092	8,805	365,051
	AC = 2	171	4,047	2,608	1,681	65,623
	AC = 2 < AF ≤ 0.1%	133	3,320	2,232	1,469	53,992
	AF ≤ 1%	104	3,113	2,157	1,439	50,000
	AF > 1%	83	3,705	3,192	2,249	56,007

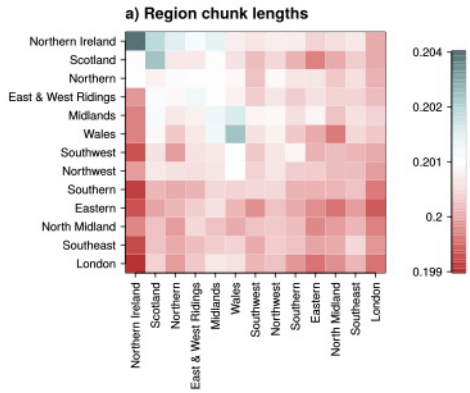
Extended Data Figure 7 | UK10K-exomes, functional consequences. **a**–**d**, Percentage of SNVs in each allele frequency bin that are loss of function (**a**), functional (**b**), possibly functional (**c**) and other (**d**), when consequences are restricted to given subsets of transcripts, and where the most severe consequence in qualifying transcripts is used. Values are percentages of SNVs that have transcripts of a given type. Protein-coding is transcripts with a biotype of protein coding. High expression is transcripts with FPKM (fragments per kilobase of transcript per million mapped reads) ≥1 in any tissue. Widely expressed is transcripts with FPKM ≥ 1 in 16 tissues. Only low expression is transcripts expressed at FPKM < 1 in all 16 tissues where there were no transcripts with high expression in that variant. Expression was determined from the Illumina Body Map data set. Variants mapping to protein-coding transcripts <300-bp long or with missing or low quality expression data were excluded. Frequency bins are singletons and non-overlapping allele frequency ranges for allele counts above 1. Allele frequency is the frequency of the alternative allele. Multi-allelic sites were included with alternative alleles at the same site treated as separate variants. **e**, Counts of single nucleotide polymorphisms in each consequence class by allele frequency and transcript subset.

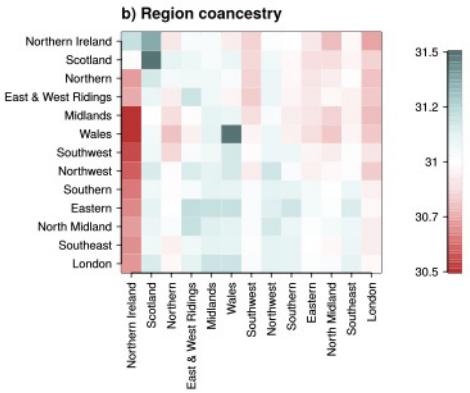


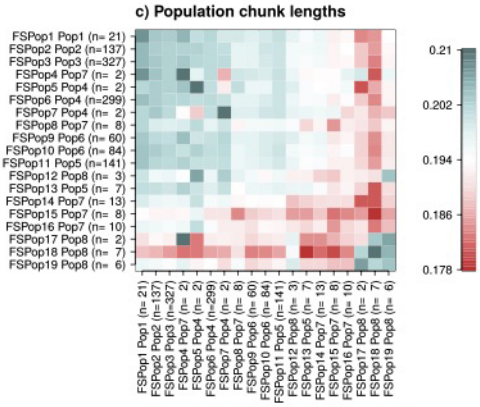
c

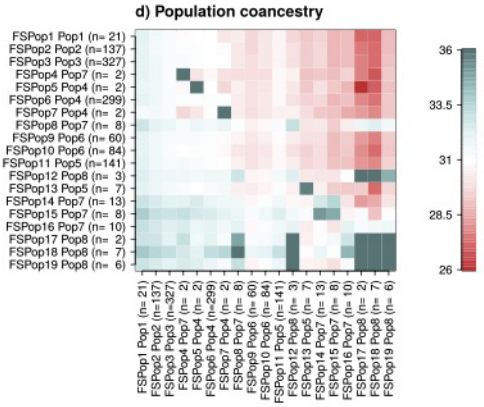
AC=2		AC=3		AC=4		AC=5		AC=6		AC=7	
Height	0.048	Height	0.052	Adiponectin	0.190	Gripstrength	0.137	FEV1/FVC	0.142	Insulin	0.100
LDL	0.063	Weight	0.055	TRFM	0.231	Adiponectin	0.206	Adiponectin	0.144	ApoA1	0.108
Adiponectin	0.071	Adiponectin	0.075	Insulin	0.297	ApoB	0.298	Height	0.144	Gripstrength	0.119
Weight	0.177	FEV1/FVC	0.117	Weight	0.317	Insulin	0.310	Glucose	0.144	TFM	0.175
Waist	0.192	Waist	0.183	Gripstrength	0.359	ApoA1	0.318	LDL	0.150	FEV1	0.206

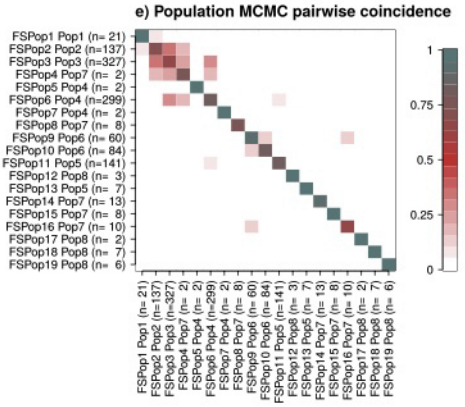
Extended Data Figure 8 | UK10K-cohorts, genotype and phenotype similarities within and between regions. **a**, **b**, Dot plots show the genetic (**a**) and phenotypic distribution (**b**) of the relationships of 1,139 unrelated TwinsUK individuals by their regional place of birth. To determine the genetic relationships we used the mean number of shared alleles between two individuals within and between regions for allele counts (AC) 2 to 7, where AC is calculated from the whole data set of 3,781 samples. To determine phenotypic similarities we calculated the mean difference between the residualized phenotypes. Genetically-related individuals are more closely related within a region than between regions, while the phenotypic distance measure has similar distributions within and between regions. The mean shared alleles increase with increasing allele count, and simultaneously the within and between distributions converge. **c**, The five lowest *P* values for AC 2 to 7 obtained from Mantel tests to determine similarities between genotypes and phenotypes by region. *P* values were not significant after correcting for multiple testing using the FDR method[49]. Full trait names are given in Supplementary Table 1.

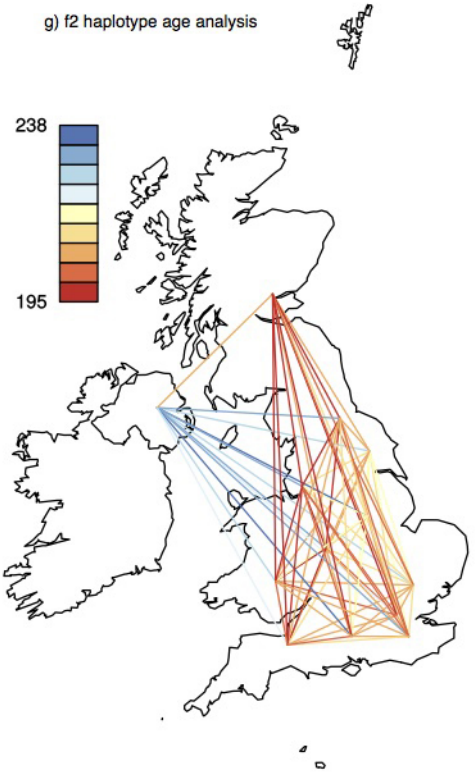


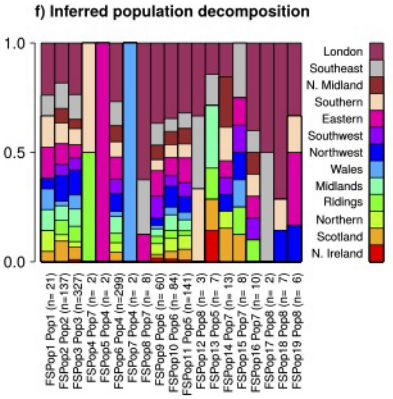












Extended Data Figure 9 | UK10K-cohorts, population fine structure in the TwinsUK sample. **a**, Chunk length matrix for all UK10K defined geographic regions, calculated as described in the methods. The bottom 5 regions are merged in Box 1 Figure. **b**, Coancestry matrix for all UK10K defined geographic regions, calculated as described in the methods. **c**, Chunk length matrix for all UK10K FineSTRUCTURE inferred populations, calculated as described in the methods. **d**, Coancestry matrix for all UK10K FineSTRUCTURE inferred populations. Details on calculation of these parameters are described in Methods. **e**, Pairwise coincidence matrix for the UK10K FineSTRUCTURE MCMC run, showing the fraction of the 1,000 retained iterations from the posterior in which each pair of individuals is in the same population, averaged for each pair of populations. The full posterior is extremely complex, which is indicative of a continuous admixture cline rather than discrete populations. **f**, Sources distribution for the FineSTRUCTURE inferred populations with the full set of inferred populations and geographic labels. Geographic labels of London, Southeast, North Midland, Southern and Eastern are merged into South and East for Box 1 Figure. FSPop labels are given to populations inferred by FineSTRUCTURE, which are merged into the Pop labels as shown in the main Box 1 Figure. **g**, The f2 haplotype age analysis estimates the time to the most recent common ancestor (tMRCA) between the two haplotypes underlying a given observed variant of allele count 2 in all of the TwinsUK samples. The observed IBD segment length around each f2 variant estimates the tMRCA, using an explicit model parameterized by the recombination and the mutation rates. Shown is the map of the UK with all regions used in this analysis depicted by their location, and lines colour-coding the observed median tMRCA of f2 haplotypes.

Extended Data Table 1 | UK10K-cohorts, estimated variance explained by SNVs across the 31 UK10K traits shared by both cohorts

		HapMap 2		HapMap3		1000GP		UK10K	
Trait	N	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value
Anthropometry/obesity									
Height	3,541	0.200 (0.087)	0.009	0.201 (0.093)	0.013	0.268 (0.102)	0.002	0.262 (0.101)	0.003
BMI	3,538	0.150 (0.088)	0.044	0.191 (0.095)	0.024	0.177 (0.104)	0.040	0.150 (0.103)	0.069
Hip	3,074	0.159 (0.100)	0.052	0.191 (0.108)	0.037	0.143 (0.116)	0.100	0.116 (0.116)	0.150
Waist	3,072	0.166 (0.098)	0.038	0.197 (0.106)	0.028	0.108 (0.115)	0.163	0.083 (0.114)	0.226
WHR	3,071	0.107 (0.097)	0.127	0.159 (0.105)	0.057	0.055 (0.114)	0.310	0.053 (0.113)	0.313
Weight	3,559	0.077 (0.087)	0.190	0.089 (0.094)	0.174	0.027 (0.103)	0.396	0.008 (0.102)	0.468
Adiponectin	2,325	0.312 (0.137)	0.012	0.359 (0.145)	0.007	0.407 (0.168)	0.011	0.401 (0.167)	0.011
Leptin	2,417	0.040 (0.125)	0.374	0.057 (0.135)	0.337	0.024 (0.148)	0.437	0.008 (0.146)	0.479
TFM	3,399	0.049 (0.089)	0.289	0.084 (0.097)	0.190	0.114 (0.104)	0.126	0.088 (0.103)	0.187
TLM	3,399	0.128 (0.090)	0.074	0.150 (0.097)	0.058	0.055 (0.108)	0.305	0.046 (0.107)	0.336
TRFM	3,197	0.000 (0.094)	0.500	0.026 (0.101)	0.398	0.055 (0.110)	0.302	0.033 (0.109)	0.378
Diabetes biochemistry									
Glucose	2,925	0.134 (0.106)	0.101	0.129 (0.112)	0.124	0.131 (0.125)	0.146	0.108 (0.125)	0.194
Insulin	2,896	0.210 (0.109)	0.026	0.225 (0.114)	0.022	0.170 (0.129)	0.096	0.164 (0.128)	0.103
HOMA-B	2,888	0.257 (0.109)	0.008	0.321 (0.115)	0.002	0.220 (0.129)	0.044	0.230 (0.128)	0.036
HOMA-IR	2,796	0.242 (0.113)	0.016	0.262 (0.119)	0.012	0.171 (0.134)	0.105	0.172 (0.133)	0.101
Cardiovascular and blood biochemistry									
CRP	2,046	0.056 (0.150)	0.356	0.012 (0.161)	0.472	0.000 (0.177)	0.500	0.000 (0.175)	0.500
LDL	3,191	0.117 (0.098)	0.117	0.133 (0.105)	0.106	0.100 (0.113)	0.183	0.123 (0.113)	0.134
HDL	3,210	0.217 (0.093)	0.007	0.299 (0.100)	7.0×10^{-4}	0.318 (0.115)	2.4×10^{-3}	0.341 (0.113)	7.9×10^{-4}
TC	3,206	0.097 (0.093)	0.139	0.119 (0.100)	0.104	0.077 (0.110)	0.235	0.116 (0.109)	0.132
TG	3,202	0.032 (0.094)	0.366	0.078 (0.101)	0.215	0.000 (0.111)	0.500	0.000 (0.111)	0.500
VLDL	3,197	0.033 (0.094)	0.361	0.081 (0.101)	0.206	0.000 (0.111)		0.000 (0.111)	0.500
ApoA1	2,914	0.174 (0.102)	0.033	0.256 (0.109)	0.005	0.176 (0.125)	0.076	0.238 (0.122)	0.019
ApoB	2,911	0.029 (0.107)	0.395	0.071 (0.117)	0.278	0.000 (0.124)	0.500	0.000 (0.123)	0.500
HGB	3,077	0.178 (0.099)	0.029	0.220 (0.104)	0.012	0.225 (0.121)	0.030	0.261 (0.118)	0.011
Heart, lung function, dynamic									
DBP	3,309	0.154 (0.093)	0.045	0.174 (0.099)	0.037	0.191 (0.109)	0.034	0.186 (0.109)	0.038
SBP	3,309	0.278 (0.094)	0.001	0.310 (0.101)	8.3×10^{-4}	0.345 (0.112)	7.0×10^{-4}	0.362 (0.112)	3.9×10^{-4}
Heart rate	2,975	0.150 (0.104)	0.071	0.180 (0.112)	0.051	0.129 (0.126)	0.156	0.134 (0.124)	0.140
FEV1	3,287	0.481 (0.094)	4.7×10^{-8}	0.534 (0.101)	2.6×10^{-8}	0.545 (0.114)	5.9×10^{-7}	0.562 (0.112)	1.3×10^{-7}
FVC	3,285	0.420 (0.094)	1.6×10^{-6}	0.467 (0.101)	9.9×10^{-7}	0.479 (0.114)	9.8×10^{-6}	0.487 (0.113)	5.0×10^{-6}
FEV1/FVC	3,280	0.294 (0.093)	5.1×10^{-4}	0.335 (0.101)	3.4×10^{-4}	0.361 (0.111)	3.2×10^{-4}	0.367 (0.110)	2.1×10^{-4}
Grip strength	3,196	0.270 (0.096)	0.002	0.248 (0.103)	0.007	0.323 (0.115)	0.002	0.334 (0.114)	0.001

We used the restricted maximum likelihood (REML) method implemented in GCTA to estimate phenotypic variance explained by SNV sets (MAF ≤ 1%) in our discovery sequence data (n = 3,621 individuals). SNVs were selected from the WGS data to correspond to the content of four different reference panels: HapMap2 (n = 2,331,713 SNVs), Hapmap3 (n = 1,168,695), 1000 Genomes (n = 7,475,230) and the entire UK10K reference panel (n = 8,317,582). Each GRM was individually tested against the 31 traits with phenotypic values present in both cohort studies, producing a beta, s.e. and P value for total trait variance explained by the given SNV set. Full trait names are given in Supplementary Table 1.