Citation for published version (APA):
UK10K Consortium, Metrustry, S., Hubbard, T., & Small, K. S. (2015). The UK10K project identifies rare variants in health and disease. DOI: 10.1038/nature14962

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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) studies1 and custom genotyping arrays2,3, and impute poorly with current reference panels. Rare and low-frequency variants also tend to be population- or sample-specific, requiring direct ascertainment through resequencing4,5. Recent exome-wide resequencing studies have begun to explore the contribution of rare coding variants to complex traits4, but comparatively little is known of the non-coding part of the genome where most complex trait-associated loci lie6. 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 imputation4, association discovery for bone mineral density5, thyroid function6 and circulating lipid levels7 and provide access to the study results through novel web tools8,9.

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)10 and TwinsUK11. 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 quantity15 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 orders 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
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, \(\sim 50 \times\) 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\(^3\).

When compared to two large-scale European sequencing repositories, 1000GP and the Genome of the Netherlands (GoNL, 12 \(\times\) read depth\(^19\)), UK10K-cohorts discovered over 24M novel SNVs. Overall, 96.5% of variants with MAF \(\geq 1\%\) 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\(^6\) (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\(^4\).

**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 1, 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, 1,122,562 bi-allelic INDELs (MAF \(\geq 0.1\%\)) 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\(^7\). The combined WGS+GWA sample had 80% power to detect associations of SNVs of low-frequency and rare down to \(\sim\)MAF 0.5%, for a per-alleles trait change (the regression beta coefficient or Beta) of \(\sim 1.2\) 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 \(\approx 1\)). 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

<table>
<thead>
<tr>
<th>Study name and design</th>
<th>(n)</th>
<th>Sequencing strategy, mean read depth and Ts/Tv ratio</th>
<th>SNVs/INDELs</th>
<th>SNVs/INDELs by allele frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cohorts.</strong> Unselected samples from two population-based cohorts</td>
<td>3,781</td>
<td>WGS, 7 (\times) Ts/Tv = 2.15</td>
<td>42,001,210/3,490,825</td>
<td>(&lt;1%: 34,247,969/2,296,962)</td>
</tr>
<tr>
<td><strong>Rare.</strong> Eight rare diseases with expected different allelic architectures (ciilopathy, coloboma, congenital heart disease, familial hypercholesterolaemia, intellectual disability, neuromuscular, severe insulin resistance and thyroid disease)</td>
<td>961 (397)</td>
<td>WES, 77 (\times) Ts/Tv = 3.02</td>
<td>252,809/1,621</td>
<td>(&lt;1%: 171,564/1,384)</td>
</tr>
<tr>
<td><strong>Obesity.</strong> Severely obese children (BMI &gt; 3 s.d. from population mean) and adults with extreme obesity</td>
<td>1,468 (1,359)</td>
<td>WES, 82 (\times) Ts/Tv = 3.02</td>
<td>484,931/3,370</td>
<td>(&lt;1%: 403,684/3,133)</td>
</tr>
<tr>
<td><strong>Neurodevelopmental.</strong> 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)</td>
<td>2,753 (1,707)</td>
<td>WES, 77 (\times) Ts/Tv = 3.02</td>
<td>538,526/3,826</td>
<td>(&lt;1%: 457,278/3,589)</td>
</tr>
</tbody>
</table>

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 as 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.

©2015 Macmillan Publishers Limited. All rights reserved

1 OCTOBER 2015 | VOL 526 | NATURE | 83
**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)\(^4\). 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\(^4\). 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”\(^4\) 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\(^4\), 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\(^2\) \(P\) value \(\leq 4.62 \times 10^{-10}\) in the combined WGS+GWA sample included common, low-frequency and rare variants tagging known associations with adiponectin levels (CDH13 and ADIPQ)\(^2\), 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\(^2\), 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)\(^3\), 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
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 data could discover additional associations. A restricted maximum likelihood (REML) 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). The first was a rare intronic variant in LDLR (rs72658867-A, c.2140 + 5G > A; EAF = 0.01, combined sample $P$ value $= 1.27 \times 10^{-46}$); per allele effect Beta (s.e.m.) $= -0.23$ mmol $L^{-1}$ (0.02), $P$ value $= 7.63 \times 10^{-30}$ (CGPS, $n = 95,079$). The second was a common, X-linked variant near RGAG1 (rs5985471-T, EAF = 0.403, $P$ value $= 1.53 \times 10^{-12}$); per allele effect Beta (s.e.m.) $= -0.02$ mmol $L^{-1}$ (0.004), $P$ value $= 1.8 \times 10^{-5}$ (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 patients29. The LDLR-A allele is almost perfectly imputed in our sample (info = 0.96), but absent in previous imputation panels29; the RGAG1-T allele is common but was missed in previous studies, which focused predominately on autosomal variation29. 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.

Findings from rare variant association tests

Single-marker association tests are typically underpowered for rare variants30. 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-O31,32, to rare variants (MAF < 1%). Overall, there was an excess of test statistics with $P$ values $\leq 10^{-4}$ 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 $\leq 1.97 \times 10^{-7}$), in a burden-type test (min $P$ value for TG $= 7.02 \times 10^{-9}$). 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 (\( P \leq 10^{-4} \)). Lastly, we tested the value of a genome-wide naive approach to explore associations outside protein-coding genes by combining variants across \( \sim 1.8 \) 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 = \( 6.52 \times 10^{-12} \); 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\(^4 \) 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 autosomal 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\(^{28–30} \). 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 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\(^{36} \). Importantly, however, we showed nearly as strong levels of functional enrichment at both sets of variants for

![Figure 4](https://example.com/figure4.png)
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 size40. We therefore engaged in a collaborative meta-analysis as part of the Autism Sequencing Consortium41 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 ancestry43). 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 database44 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 population45.

Overall, this effort has given us both new genomic tools 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 genetic diseases.
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 to these sections appear only in the online paper.

and Source Data, are available in the online version of the paper; references unique

The UK10K Project Consortium

Klaudia Walter1, Josie M. Lin2, Jie Huang3, Lucy Crooks4, Yasmin Memari, Shane McCarthy1, John R. Perry4, Changjiang Xu5, Martha Futema6, Daniel Lawson1, Valentina Litchkov1,14, Stephen Schifflers7, Audrey E. Hendricks1,12, Peter Clapham1, Richard Durbin1, Nicola Soranzo1,2,20,21,22,77,83, Nadia Schoenmakers17, Robert K. Semple17, Carol Smee7, Timothy D. Spector7, Nicholas J. Timkinson3

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article’s Creative Commons licence, unless otherwise stated in a third party notice. If you copy, alter, transform, or build upon this article, you may not distribute the modified material, in whole or in part, without the express written consent of the licence holder. To reproduce the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0

ARTICLE
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 (×10^3). b, Number of INDELs by sample (×10^3). 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 (×10^6). g, Number of SNVs (×10^6), INDELs (×10^5) and large deletions (×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.
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 ≥ 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).
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 $\leq 1\%$ 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.
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 \( > 1 \) 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.

<table>
<thead>
<tr>
<th>Transcript set</th>
<th>Frequency</th>
<th>LoF</th>
<th>Functional</th>
<th>Possibly functional</th>
<th>Other</th>
<th>No qualifying transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC = 1</td>
<td>10,647</td>
<td>234,169</td>
<td>175,067</td>
<td>95,667</td>
<td>416</td>
<td></td>
</tr>
<tr>
<td>AC = 2</td>
<td>1,338</td>
<td>40,013</td>
<td>34,430</td>
<td>17,747</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 0.1%</td>
<td>988</td>
<td>32,020</td>
<td>26,905</td>
<td>15,226</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 1%</td>
<td>751</td>
<td>28,264</td>
<td>27,651</td>
<td>14,900</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>AF &gt; 1%</td>
<td>446</td>
<td>23,721</td>
<td>36,084</td>
<td>20,840</td>
<td>135</td>
<td></td>
</tr>
<tr>
<td><strong>Protein-coding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC = 1</td>
<td>10,127</td>
<td>231,825</td>
<td>154,433</td>
<td>110,928</td>
<td>11,653</td>
<td></td>
</tr>
<tr>
<td>AC = 2</td>
<td>1,254</td>
<td>39,538</td>
<td>29,836</td>
<td>20,634</td>
<td>2,339</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 0.1%</td>
<td>905</td>
<td>31,613</td>
<td>25,007</td>
<td>17,661</td>
<td>2,013</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 1%</td>
<td>664</td>
<td>27,876</td>
<td>23,776</td>
<td>17,124</td>
<td>2,194</td>
<td></td>
</tr>
<tr>
<td>AF &gt; 1%</td>
<td>361</td>
<td>23,123</td>
<td>30,413</td>
<td>23,747</td>
<td>3,602</td>
<td></td>
</tr>
<tr>
<td><strong>High expression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC = 1</td>
<td>6,343</td>
<td>154,247</td>
<td>105,711</td>
<td>87,097</td>
<td>56,948</td>
<td></td>
</tr>
<tr>
<td>AC = 2</td>
<td>760</td>
<td>26,062</td>
<td>20,365</td>
<td>16,097</td>
<td>10,846</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 0.1%</td>
<td>504</td>
<td>20,663</td>
<td>16,946</td>
<td>13,886</td>
<td>9,167</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 1%</td>
<td>352</td>
<td>17,981</td>
<td>16,023</td>
<td>13,450</td>
<td>9,007</td>
<td></td>
</tr>
<tr>
<td>AF &gt; 1%</td>
<td>123</td>
<td>13,787</td>
<td>20,422</td>
<td>18,073</td>
<td>12,831</td>
<td></td>
</tr>
<tr>
<td><strong>Widely expressed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC = 1</td>
<td>1,304</td>
<td>36,957</td>
<td>27,694</td>
<td>32,654</td>
<td>311,737</td>
<td></td>
</tr>
<tr>
<td>AC = 2</td>
<td>131</td>
<td>5,909</td>
<td>5,301</td>
<td>5,863</td>
<td>56,926</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 0.1%</td>
<td>76</td>
<td>4,525</td>
<td>4,308</td>
<td>5,030</td>
<td>47,207</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 1%</td>
<td>40</td>
<td>3,684</td>
<td>4,050</td>
<td>4,629</td>
<td>44,410</td>
<td></td>
</tr>
<tr>
<td>AF &gt; 1%</td>
<td>5</td>
<td>2,440</td>
<td>4,834</td>
<td>5,456</td>
<td>52,501</td>
<td></td>
</tr>
<tr>
<td><strong>Only low expression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC = 1</td>
<td>1,093</td>
<td>22,305</td>
<td>13,092</td>
<td>8,865</td>
<td>355,051</td>
<td></td>
</tr>
<tr>
<td>AC = 2</td>
<td>171</td>
<td>4,047</td>
<td>2,608</td>
<td>1,681</td>
<td>65,623</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 0.1%</td>
<td>133</td>
<td>3,320</td>
<td>2,232</td>
<td>1,469</td>
<td>53,992</td>
<td></td>
</tr>
<tr>
<td>AF ≤ 1%</td>
<td>104</td>
<td>3,113</td>
<td>2,157</td>
<td>1,439</td>
<td>50,000</td>
<td></td>
</tr>
<tr>
<td>AF &gt; 1%</td>
<td>83</td>
<td>3,705</td>
<td>3,192</td>
<td>2,249</td>
<td>56,007</td>
<td></td>
</tr>
</tbody>
</table>
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. 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 UK10KFineSTRUCTURE 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 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.
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.