ANALYSIS

nature genetics

Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps

Valentina Iotchkova[1,2], Jie Huang[2,3], John A Morris[4,5], Deepti Jain[6], Caterina Barbieri[2,7], Klaudia Walter[2], Josine L Min[8], Lu Chen[2,9], William Astle[10], Massimilian Cocca[11,12], Patrick Deelen[13,14], Heather Elding[2], Aliki-Eleni Farmaki[15], Christopher S Franklin[2], Mattias Franberg[16], Tom R Gaunt[8], Albert Hofman[17,18], Tao Jiang[10], Marcus E Kleber[19], Genevieve Lachance[20], Jian'an Luan[21], Giovanni Malerba[22], Angela Matchan[2], Daniel Mead[2], Yasin Memari[2], Ioanna Ntalla[15,23], Kalliope Panoutsopoulou[2], Raha Pazoki[17], John R B Perry[20,21], Fernando Rivadeneira[17,24], Maria Sabater-Lleal[16], Bengt Sennblad[16], So-Youn Shin[2,8], Lorraine Southam[2,25], Michela Traglia[7], Freerk van Dijk[13,14], Elisabeth M van Leeuwen[17], Gianluigi Zaza[26], Weihua Zhang[27], UK10K Consortium[28], Najaf Amin[17], Adam Butterworth[10,29], John C Chambers[27], George Dedoussis[15], Abbas Dehghan[17], Oscar H Franco[17], Lude Franke[14], Mattia Frontini[9], Giovanni Gambaro[30], Paolo Gasparini[11,12,31], Anders Hamsten[16], Aaron Issacs[17], Jaspal S Kooner[32], Charles Kooperberg[33], Claudia Langenberg[21], Winfried Marz[19,34,35], Robert A Scott[21], Morris A Swertz[13,14,36], Daniela Toniolo[7], Andre G Uitterlinden[24], Cornelia M van Duijn[17], Hugh Watkins[25,37], Eleftheria Zeggini[2], Mathew T Maurano[38], Nicholas J Timpson[8], Alexander P Reiner[33,39,41], Paul L Auer[40,41] & Nicole Soranzo[2,9,29,41]

Large-scale whole-genome sequence data sets offer novel opportunities to identify genetic variation underlying human traits. Here we apply genotype imputation based on whole-genome sequence data from the UK10K and 1000 Genomes Project into 35,981 study participants of European ancestry, followed by association analysis with 20 quantitative cardiometabolic and hematological traits. We describe 17 new associations, including 6 rare (minor allele frequency (MAF) < 1%) or low-frequency (1% < MAF < 5%) variants with platelet count (PLT), red blood cell indices (MCH and MCV) and HDL cholesterol. Applying fine-mapping analysis to 233 known and new loci associated with the 20 traits, we resolve the associations of 59 loci to credible sets of 20 or fewer variants and describe trait enrichments within regions of predicted regulatory function. These findings improve understanding of the allelic architecture of risk factors for cardiometabolic and hematological diseases and provide additional functional insights with the identification of potentially novel biological targets.

Heritable influences on cardiometabolic and hematological traits have been identified across the allele frequency spectrum. Rare (MAF < 1%) and highly penetrant variants with large phenotypic effects have been identified, but these account for a small proportion of phenotypic variance[1,2]. At the other end of the allelic frequency spectrum, genome- and exome-wide association analyses based on sparse arrays have identified thousands of common (MAF ≥ 5%) and low-frequency (MAF = 1–5%) single-nucleotide variants (SNVs) with modest effects[3–11]. To investigate the influence of rare, less frequent and common variation on complex traits, we applied whole-genome sequencing in individuals from two UK cohorts, the St Thomas' Twin Registry (TwinsUK)[12] and the Avon Longitudinal Study of Parents and Children (ALSPAC)[13], as part of the UK10K project. Sequencing was performed at an average depth of 7× across 3,781 individuals. The final data set is described in ref. 14 and consists of 42 million SNVs, 3.5 million indel polymorphisms and nearly 18,000 large deletions.

The initial phase of the UK10K project applied different statistical tests to identify rare alleles associated with a broad range of complex phenotypes. Besides yielding the first examples of new trait associations identified through population-based whole-genome sequencing[15,16], the project provided large-scale empirical evaluation of strategies for testing association of variants in the low-frequency and rare ranges. First, the study demonstrated a lack of low-frequency alleles with high penetrance and consequent power for detection (defined by an effect for each variant of >1.2 s.d. and MAF ~0.5%). This confirms expectations that, in this frequency range, new discoveries require larger samples with greater statistical power. Further, the study defined, through simulations and empirical evidence, the allelic space where genotype imputation is expected to be most beneficial for association studies. Finally, it developed a new genotype imputation panel based on whole-genome sequence that enhances imputation accuracy for low-frequency and rare variants in populations of European descent[17], substantially improving resolution and power in this frequency range.

Capitalizing on these discoveries, we sought to increase the representation of rare variation in association studies of cardiometabolic

A full list of affiliations appears at the end of the paper.

Received 18 January; accepted 15 August; published online 26 September 2016; doi:10.1038/ng.3668

and hematological traits through imputation using the UK10K and 1000 Genomes Project haplotype reference panels, studying up to 35,981 individuals of European descent from 18 studies. After testing for association between 17 million sequence variants and 20 quantitative traits, we report 17 new variants associated with 7 different traits. We applied fine-mapping approaches that exploit these more comprehensive imputation reference panels to identify sets of variants with high (>95%) joint probability of being causal at 59 different loci. By expanding the number of discovered loci for seven cardiometabolic traits and narrowing known association signals to small sets of variants, our results demonstrate the utility of large imputation reference panels for the discovery and refinement of associations with complex quantitative traits.

RESULTS

Common, low-frequency and rare variant associations

We considered 20 quantitative traits representing five biomedical trait groups: lipids (HDL, LDL, TC and TG), inflammatory biomarkers (CRP and IL6), renal function (uric acid and creatinine), fasting glycemic traits (glucose, insulin, HOMA-B and HOMA-IR) and hematological indices (HGB, RBC, MCH, MCHC, MCV, PCV, PLT and WBC) (defined in **Fig. 1**). In the discovery stage, we tested associations of up to 15,188,514 autosomal and 468,312 X-linked SNVs and 1,311,244 biallelic indels (MAF ≥ 0.1%) in up to 3,210 participants with low-coverage whole-genome sequencing data available (depending on the trait) and combined these individuals with up to 32,904 participants from independent population-based samples with SNPs imputed to the UK10K panel or a combination of whole-genome sequence reference panels[17,18] (**Supplementary Table 1** and **Supplementary Note**). We tested associations within each study using linear regression (Online Methods, **Supplementary Fig. 1** and **Supplementary Tables 1** and **2**) and combined summary statistics from different studies with inverse-variance-weighted meta-analyses.

This approach yielded 171 independent associations ($P \leq 5 \times 10^{-8}$) in the discovery meta-analysis, of which 110 represented previously reported GWAS signals, 48 mapped to conditionally independent variants at known GWAS signals (secondary signals) and 13 corresponded to putative new associations. We obtained replication for 58 of 61 variants in up to 102,505 independent samples from five studies. We detected 17 new associations that were robustly replicated (replication $P < 0.05/58$; meta-analysis $P < 8.31 \times 10^{-9}$) in independent samples (**Table 1** and **Supplementary Table 3**). Of these, ten were new loci (primary signals), defined as genomic regions not previously associated with the trait of interest. We identified seven variants defined as secondary signals, where the genetic variant mapped to within 1 Mb of a locus already associated with the trait but was statistically independent of any previously reported association (Online Methods). Of the 17 variants reported, 3 were coding and the rest were located in noncoding putative regulatory regions (**Box 1**).

The 10 new associations involved hematological traits, including 7 variants associated with PLT, 2 associated with WBC and 1 associated with PCV. Two loci were previously associated with other traits. The rs1801689 missense variant (p.Cys325Gly) in *APOH* associated with higher PLT was previously associated with higher LDL cholesterol[19]. *SHROOM3* rs10008637 associated with higher PCV is a linkage disequilibrium (LD) proxy ($r^2 = 0.98$) for rs13146355, a common intronic variant associated with lower serum creatinine levels in East Asians[20] and higher serum magnesium levels in Europeans[21]. One of the PLT-associated loci, synonymous variant rs150813342 of *GFI1B*, was reported in an independent exome sequencing data set[22].

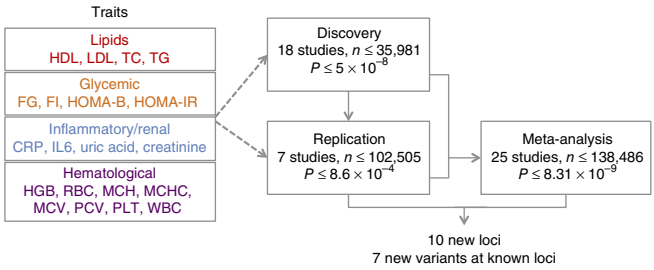

Figure 1 Study design. Summary of traits and studies investigated in this study. Study-specific information is given in **Supplementary Table 1**. HDL, high-density lipoprotein cholesterol; LDL, low-density lipoprotein cholesterol; TC, total cholesterol; TG, triglycerides; FG, fasting glucose; FI, fasting insulin; HOMA-B, homeostatic model assessment of beta cell function; HOMA-IR, homeostatic model assessment of insulin resistance; CRP, C-reactive protein; IL6, interleukin-6; HGB, hemoglobin; RBC, red blood cell count; MCH, mean cell hemoglobin; MCHC, mean cell hemoglobin concentration; MCV, mean cell volume; PCV, packed cell volume, or hematocrit; PLT, platelet count; WBC, white blood cell count.

Among the seven secondary signals within 1 Mb of a known locus, one was associated with HDL cholesterol levels (an intronic variant of *ABCA1*), one was associated with uric acid levels (an intronic variant of *SLC2A9*) and five were associated with hematological indices (PLT, WBC, MCV and MCH). Four loci harbored both common and independent, lower-frequency variants (*CCDN3* for MCV; *THPO* for PLT; *GCSAML* for PLT; and *ABCA1* for HDL cholesterol). The low-frequency *ABCA1* intronic variant rs3824477 (MAF = 0.02) was in strong LD ($r^2 = 0.94$) with an *ABCA1* missense variant (rs2066718, encoding p.Val771Leu) nominally associated with HDL cholesterol ($P = 1 \times 10^{-4}$) in a targeted lipid gene resequencing study[23].

Three of the ten new loci, and three of the seven secondary signal associations, were observed for low-frequency or rare variants, extending understanding of the genetic architecture of cardiometabolic traits. To illustrate, we considered the effect sizes and allele frequencies of both known and new variants for HDL cholesterol and PLT (**Fig. 2a**). Although we identified one rare variant with a large effect size (rs150813342 in *GFI1B*), the effect sizes of the other new low-frequency variants were similar to those that have been previously reported in genome-wide association studies of common variants. Indeed, for variants with MAF ≥ 0.5%, we had 80% power to detect associations with effect sizes of 0.25, 0.25, 0.35 and 0.55 trait s.d. for HGB, LDL, HOMA-B and IL6, respectively (**Fig. 2b**). Although there may be rare variants of large effect that we were unable to identify, we likely did not miss large-effect variants with MAF ≥ 0.5% and sufficient sequencing quality in European populations.

Functional enrichment analysis of trait-associated variants

The majority of the associations we identified were found in noncoding regions, where the underlying molecular mechanisms are poorly defined. To evaluate the functional properties of these variants, we estimated the extent to which associations for each of the 20 traits were non-randomly distributed across various coding, noncoding regulatory and cell-type-specific elements across the genome. We retrieved experimentally derived annotations from 1,005 genome-wide data sets from the GENCODE, Encyclopedia of DNA Elements (ENCODE) and Roadmap projects (**Supplementary Table 4**). We then used a novel nonparametric approach (GARFIELD) (**Supplementary Note**) to derive fold enrichment statistics for trait-associated SNPs within each annotation, where SNPs were selected from genome-wide data sets on the basis of their strength of association with each trait

Table 1 List of new variants and loci identified in this study

Associated trait	Marker	Chr.	Position (bp) (hg19)	Locus/nearest gene	Coded allele	Non-coded allele	MAF (WGS)	β (joint)	SE (joint)	*P* value (joint)	*n* (joint)	Primary/ secondary signal	Variant annotation
PCV	rs10008637	4	77,414,144	*SHROOM3*	C	T	0.463	0.032	0.004	1.08×10^{-14}	124,890	Primary	Intronic
PLT	rs2546979	5	159,595,612	*FABP6*	C	G	0.291	–0.049	0.004	1.81×10^{-31}	134,858	Primary	Intergenic
WBC	rs3130725	6	29,118,747	*ZNF311*	G	T	0.131	–0.008	0.001	2.70×10^{-26}	121,238	Primary	Intergenic
WBC	rs113164910	6	32,427,005	*HLA-DRA*	AAC	A	0.327	0.008	0.000	4.19×10^{-54}	122,412	Primary	Intergenic
PLT	rs61750929	9	91,495,135	*S1PR3*	T	C	0.059	–0.081	0.008	2.20×10^{-21}	134,858	Primary	Intergenic
PLT	rs150813342	9	135,864,513	*GFI1B*	T	C	0.004	–0.408	0.026	4.73×10^{-57}	111,278	Primary	Synonymous
PLT	rs113373353	12	65,007,682	*RASSF3*	T	C	0.111	0.055	0.006	1.76×10^{-17}	134,858	Primary	Intronic
PLT	rs575505283	15	43,703,277	*TP53BP1*	AT	A	0.014	–0.160	0.019	6.89×10^{-17}	121,073	Primary	Intronic
PLT	rs1801689	17	64,210,580	*APOH*	C	A	0.033	0.106	0.012	3.92×10^{-19}	134,858	Primary	Nonsynonymous
PLT	rs75570992	22	50,570,755	*TRABD–MOV10L1*	C	G	0.072	0.096	0.008	7.75×10^{-32}	134,377	Primary	Intronic
PLT	rs41315846	1	247,712,303	*GCSAML*	C	T	0.479	0.048	0.004	3.03×10^{-34}	134,858	Secondary	Intronic
PLT	rs78565404	3	184,090,242	*THPO*	T	C	0.057	0.136	0.009	1.65×10^{-50}	134,858	Secondary	3′ UTR
Uric acid	rs56223908	4	9,918,492	*SLC2A9*	C	A	0.080	0.137	0.018	9.21×10^{-15}	26,727	Secondary	Intronic
WBC	rs2442735	6	31,346,653	*HLA-B*	G	A	0.140	–0.010	0.001	1.93×10^{-46}	121,528	Secondary	Intergenic
MCV	rs112233623	6	41,924,998	*CCND3*	T	C	0.011	0.723	0.049	5.65×10^{-49}	107,036	Secondary	Intronic
HDL	rs3824477	9	107,588,328	*ABCA1*	A	G	0.026	0.122	0.016	1.43×10^{-13}	56,306	Secondary	Intronic
MCH	rs117747069	16	170,076	*NPRL3*	C	G	0.037	–0.172	0.024	4.20×10^{-13}	119,687	Secondary	Intronic

Chr., chromosome; WGS, whole-genome sequencing; SE, standard error.

(Online Methods). For an example of results for one trait (PLT) and one annotation type (DNase I hypersensitivity site (DHS) hotspots), see **Figure 3**, with all results summarized in **Supplementary Figure 2** and **Supplementary Table 5**.

Lipid and hematological traits displayed ubiquitous and marked enrichment patterns, with 151 ($P < 1 \times 10^{-8}$) and 906 ($P < 1 \times 10^{-5}$) overall significant fold enrichment statistics for serum lipids and 237 ($P < 1 \times 10^{-8}$) and 749 ($P < 1 \times 10^{-5}$) significant statistics for hematological traits. We found that associations with RBC were enriched in enhancers of the erythroid cell line K562 (fold enrichment = 39.63, empirical $P = 2 \times 10^{-5}$), while associations with WBC were enriched in footprints of $CD20^+$ cells (fold enrichment = 22.16, empirical $P < 1 \times 10^{-5}$). The most significant association for LDL cholesterol was within the transcription start site (TSS) chromatin states measured in the liver HepG2 cell line (fold enrichment = 19.53, empirical $P < 1 \times 10^{-5}$). Conversely, inflammatory and renal traits displayed weak patterns of enrichment. There was significant enrichment of associations (fold enrichment = 4.44, empirical $P = 1 \times 10^{-5}$) with creatinine levels within DHS hotspots of fetal kidney. Uric acid associations were weakly enriched in a small number of liver and fetal intestine annotations. Unexpectedly, we observed enrichment of TG in HMVEC-LLy (lymphatic microvascular endothelial cell) footprints for SNPs with $P < 1 \times 10^{-5}$ (fold enrichment = 9.75, empirical $P < 1 \times 10^{-5}$), which was much larger than that observed for the broader DHS hotspots (fold enrichment = 4.30, empirical $P < 1 \times 10^{-5}$). By contrast, there was no significant enrichment for footprints of the expected most relevant HepG2 cell type (well-established hepatocyte cellular model for cholesterol metabolism).

Fine-mapping of loci using dense imputation from whole-genome sequencing

LD and incomplete ascertainment of variants in a region of interest present challenges for pinpointing the causal variant(s) driving an association. To fine-map the causal variant(s) at associated loci, we exploited the high density of our whole-genome sequencing reference panels to define the posterior probability of each variant being causal given all other variants in the region. We selected 417 regions with informative associations ($P \leq 1 \times 10^{-5}$; Online Methods) in the initial discovery meta-analysis and applied three distinct Bayesian approaches (Maller[24], FINEMAP[25] and CAVIARBF[26]) (Online Methods). For each method, we created 95% credible sets by ranking variants on the basis of their decreasing posterior probability (PP) of association. These credible sets contain the minimum list of variants that jointly have at least 95% probability of including the causal variant. We focused on 59 known or new loci where the three methods identified a credible set of fewer than 20 variants and where all variants were either directly genotyped or well imputed (**Fig. 4**, **Supplementary Fig. 3** and **Supplementary Table 6**).

Overall, the 95% credible sets contained an average of 6.9 (s.d. = 5.9) variants per locus when considering the union of all methods, or 5.5 (s.d. = 4.7) variants when considering the intersection. In 45 cases, the three methods yielded identical 95% credible sets, including 13 known and 5 new loci where a single variant was predicted to be causative with PP ~1 by all methods. Of these 18 loci, 5 involved well-characterized missense variants (rs11591147 at *PCSK9*, rs1260326 at *GCKR*, rs855791 at *TMPRSS6*, rs7412 at *APOE* and rs429358 as a secondary signal at *APOE*). Missense variants were included in the 95% credible sets at several other loci (*ABCG2*, *APOB*, *CD300LG*, *CILP2*, *HFE*, *PSORCS1*, *SH2B3*, *SLC30A8* and *APOH*). At four loci, the credible interval included a variant predicted to alter an essential splicing donor or acceptor motif (*GCSAML*, *MLXIPL*, *BET1L* and

Box 1 Biological and functional annotation of new genetic variants and loci

Locus–trait	Description of most likely functional SNP
GFI1B–PLT	Index SNP rs150813342 is a synonymous variant altering a predicted *GFI1B* exon 5 splice site. *GFI1B* encodes a transcription factor involved in the regulation of red blood cell and platelet production[34]. Rare, heterozygous loss-of-function mutations in *GFI1B* have been reported in hereditary thrombocytopenia (MIM 187900). rs150813342 has no LD proxies; it is predicted to be causative by CAVIARBF (PP = 1). Furthermore, it lies within a region enriched for H3K4me1 and H3K36me3 in megakaryocytes[52].
NPRL3–MCH	Index SNP rs117747069 is a low-frequency intronic variant of *NPRL3* with no LD proxies. It is predicted to be the most likely causal variant (CAVIARBF PP = 0.84) and is conditionally independent of the common *NPRL3* variant rs11248850 previously associated with MCH[8]. *NPRL3* is known to contain nucleosome-depleted regions involved in regulation of the α-globin genes on chromosome 16 (ref. 8). rs117747069 is located in an erythroid-specific super-enhancer[52–55], which is hypersensitive, is enriched for H3K27ac marks in erythroblasts and overlaps ChIP–seq signal for the erythroid transcription factors GATA1, GATA2 and TAL1 in K562 cells[56]. Although the nearest gene, *NPRL3*, is a potential target of the enhancer element, chromatin interactions in K562 cells[56] suggest that the super-enhancer element interacts with several downstream genes, including *HBA1* and *HBA2*.
CCND3–MCV	Index SNP rs112233623 is a low-frequency intronic variant of *CCND3*, conditionally independent of the previously reported common association of rs9349204 with red blood cell traits[8]. Cyclin D3 has a critical role in cell cycle regulation. The index SNP is located within an erythroid-specific enhancer[37,52,55] enriched for the H3K27ac mark in erythroblasts and is bound by GATA2 and TAL1 in K562 cells[56]. The association of rs112233623 with hemoglobin A2 levels[36] also supports the role of this variant in the regulation of α-globin.
HLA-DRA–WBC	Index variant rs113164910 is a 2-bp indel lying in the class II major histocompatibility complex (MHC) region, 14 kb 3′ of *HLA-DRA*. The most likely functional SNP, rs9268781 (8 kb 3′ of *HLA-DR*), is a strong expression quantitative trait locus (eQTL) for various *HLA-DR* and *HLA-DQ* genes in blood[57] and overlaps a DHS in blood monocytes[58]. Another LD proxy, rs7763262, has previously been associated with IgA nephropathy[59].
HLA-B–WBC	Index SNP rs2442735 is located ~20 kb 5′ of the *HLA-B* locus and is conditionally independent of another *HLA-B* intronic SNP in the class I MHC region, rs2853946, associated with WBC[60]. The most likely functional SNP, rs2853999, located 1 kb 5′ of *HLA-B*, is a blood eQTL for *HLA-C*, *C4A* and *C4B* and overlaps a blood cell promoter and enhancer, DHSs and histone marks. A proxy SNP has been associated with marginal zone lymphoma[61].
THPO–PLT	Index SNP rs78565404 is a second *THPO* signal, conditionally independent of the previously reported platelet GWAS variant rs6141 (ref. 62). Both SNPs fall in the 3′ UTR and have no LD proxies. THPO is a key regulator of platelet production. *THPO* gain-of-function mutations have been identified in hereditary thrombocythemia (MIM 187950). rs78565404 binds the transcription factor MAFK (ChIP–seq in HepG2 cells), a component of the NF-E2 complex involved in erythropoiesis and megakaryopoiesis[38,39].
GCSAML–PLT	Index SNP rs41315846 is located in a hematopoietic cell lineage–specific promoter of *GCSAML* (*C1orf150*)[58]. It is conditionally independent of previously reported *GCSAML* intronic index SNP rs7550918 and has no LD proxies. *GCSAML* encodes a protein thought to be a signaling molecule associated with germinal centers, the sites of proliferation and differentiation of mature B lymphocytes. rs41315846 lies within a putative enhancer overlapping a DHS, RUNX1, GATA1 and FLI1 ChIP–seq peaks and an H3K27ac-enriched region in megakaryocytes[52].
FABP6–PLT	Index SNP rs2546979 is a common intronic variant of *FABP6*, which encodes a fatty-acid-binding protein not known to have a role in platelet biology. It lies in a region of high LD spanning the region 5′ to the first intron of *FABP6*. The most likely functional SNP (r^2 = 0.7), rs2546372 (located ~22 kb upstream of *FABP6*), overlaps regions enriched for H3K4me1 and H3K27ac signal in megakaryocytes, DHSs, and RUNX1 and FLI1 ChIP–seq peaks[52]. Another gene in this region, the transcription factor gene *PTTG1*, is highly expressed in bone marrow stem cells[63] and in megakaryocytes and erythroid precursors. Platelet promoter capture data from BLUEPRINT show that rs2546979 physically interacts with neighboring gene *CCNJL*, which belongs to the family of cyclin genes involved in cell cycle regulation. The presence of H3K27ac (active promoter/enhancer) in the *CCNJL* promoter region and H3K36me3 (elongation) in the body of this gene indicates that *CCNJL* is actively expressed in megakaryocytes[52].
TRABD and *MOV10L1*–PLT	Index SNP rs75570992 is intronic to *MOV10L1*, a predicted RNA helicase of unknown function. It is predicted to be causal (CAVIARBF PP = 1) and is associated with expression of the neighboring gene *TRABD* in transformed fibroblasts and colon and lymphoblastoid cells[33,57]. However, another likely functional SNP is proxy SNP rs75107793 (r^2 = 0.5), which overlaps promoter and enhancer histone marks in many cell types[58] but, more importantly, is located in a putative enhancer overlapping RUNX1 ChIP–seq peaks and a DHS- and H3K4me1-enriched region in megakaryocytes[52]. Functional SNP rs75107793 is also located within a DHS peak in erythroblasts and lies upstream of the *TRABD* promoter (GENCODE and FANTOM5). On the basis of RNA–seq and epigenetic marks (H3K27ac, H3K4me3 and H3K36me3), *TRABD* is expressed in megakaryocytes[52].
ZNF311–WBC	Index SNP rs3130725 is located in an intergenic region on chromosome 6 containing extensive LD (>50 proxy SNPs, r^2 > 0.8), all of which (including rs3130725) are eQTLs in whole blood for several genes in the class I HLA region, including *ZFP57*, *HLA-F* and *HLA-H*[57]. The most likely functional SNP is rs3129794, which is located in the promoter region of *ZNF311* and overlaps an active promoter in K562 cells[58].

Box 1 Biological and functional annotation of new genetic variants and loci

Locus–trait	Description of most likely functional SNP
APOH–PLT	Index SNP rs1801689 (p.Cys325Arg) is located in *APOH*, which encodes β_2-GPI, a platelet phospholipid-binding protein. It is the most likely functional SNP (CAVIARBF PP = 0.37), although another proxy SNP, rs8178824 (r^2 = 1; PP = 0.22), is located in a liver-specific promoter (Roadmap Epigenomics). Platelet promoter capture data (BLUEPRINT) show that rs1801689 physically interacts with neighboring gene *PRKCA* (protein kinase Cα), which also has a role in platelet function and platelet production in mouse models of megakaryopoiesis[64,65].
S1PR3–PLT	Index SNP rs61750929 is located ~100 kb upstream of *S1PR3*, which encodes a receptor for sphingosine-1-phosphate (S1P) and likely contributes to the regulation of angiogenesis and vascular endothelial cell function[66,67]. *S1PR3* overlaps *C9orf47*, a gene of unknown function. The index SNP has 33 strong LD proxies in an intergenic region between *MIR4289* and *S1PR3* (*C9orf47*), several of which are *cis*-eQTLs for *S1PR3* in whole blood[68], positioned within megakaryocytic DHSs (rs62549698 and rs9410336) or H3K4me1-enriched enhancer regions (rs9410196, rs142550358 and rs9410336)[52]. Two proxies in weaker LD (r^2 = 0.5) are synonymous (rs11795137) or 3′ UTR (rs62551536) variants of *C9orf47*.
RASSF3–PLT	Index SNP rs113373353 and all 33 of its proxies are intronic to *RASSF3*, a tumor suppressor that also promotes apoptosis. The most likely functional SNP, rs77164989 (r^2 = 0.8), lies within a putative enhancer that overlaps DHSs, H3K4me1 marks and RUNX1 ChIP–seq peaks in megakaryocytes[52].
SHROOM3–HCT	Index SNP rs10008637 is intronic to *SHROOM3*, which encodes a protein that binds and regulates the subcellular distribution of F actin[69]. An intronic LD proxy, rs13146355, located in *SHROOM3* is associated with lower serum creatinine[20] and higher serum magnesium[21] levels. Another LD proxy, rs17319721 (r^2 = 0.8), overlaps DHSs in endothelial cells and is located in a TCF7L2-dependent enhancer, increasing *SHROOM3* transcription and influencing transforming growth factor (TGF)-β1 signaling and renal function[70].
ABCA1–HDL	The *ABCA1* intronic variant rs3824477 (MAF = 0.02) is in strong LD (r^2 = 0.94) with an *ABCA1* missense variant (rs2066718, p.Val771Leu) previously nominally associated with HDL ($P = 1 \times 10^{-4}$)[23]. Both SNPs are independent of the common *ABCA1* index SNP rs1883025 for HDL[19] and the secondary *ABCA1* signal rs11789603 (ref. 71). *ABCA1* regulates cholesterol and phospholipid homeostasis. Rare loss-of-function variants of *ABCA1* are associated with Tangier's disease (MIM 205400).
TP53BP1–PLT	Index variant chr15:43703277 is a 1-bp intronic indel of *TP53BP1* located at a DHS and binding sites for several hematopoietic transcription factors, including MAFK, GATA1, GATA2 and TAL1. A chromosomal aberration involving *TP53BP1* is found in a form of myeloproliferative disorder with eosinophilia[72]. The translocation t(5;15)(q33;q22) with *PDGFRB* creates a TP53BP1–PDGFRβ fusion protein.
SLC2A9–uric acid	Index SNP rs56223908 (MAF = 0.08) is intronic to the urate transporter gene *SLC2A9* (ref. 73). It has no LD proxies and is conditionally independent of the more common, known *SLC2A9* uric acid GWAS variant rs12498742 (ref. 74). Rare mutations in *SLC2A9* are a cause of autosomal recessive renal hypouricemia-2 (MIM 612076). The index SNP overlaps H3K4me1 enhancer histone marks in several Roadmap Epigenomics cell lines and tissues (blood, adrenal, muscle, heart and lung) and is predicted to be an active promoter in pancreas.

CETP), and at the other three (*DNAH11*, *IKZF1* and *GFI1B*) the 95% credible set included synonymous sites. For all other loci, the causative set included UTR, intergenic and intronic sites.

For each known locus, we compared the variants in the fine-mapped set with published evidence from functional validation studies (**Supplementary Table 6**). Of the 59 discrete regions, 40 were associated with one trait and 19 were associated with multiple traits. Further, 25 (42%) were known to have at least one causative variant previously functionally validated. At 20 of the 25 loci, the previously validated functional variant was within the 95% credible interval identified using one or more fine-mapping methods. In 11 regions, the known causal variant was ranked with the highest posterior probability by at least one fine-mapping method. We also identified several examples where the credible sets defined high-priority variants for downstream follow-up. Among these are *CRP* rs1205, a 3′ UTR variant associated with CRP levels that is located in a predicted liver enhancer region that alters a glucocorticoid receptor (NR3C1) transcription factor binding site; rs1822534, a regulatory region variant upstream of *PPARG*, associated with PLT; *ARHGEF3* rs1354034, an intronic variant associated with PLT located in a predicted enhancer region in hematopoietic and primary T cells (Roadmap Epigenomics chromatin state) and predicted to alter a GATA motif; the TC-associated variant rs2169387 located in a predicted liver and muscle enhancer region several hundred kilobases upstream of *PPP1R3B*; the TC-associated *ABCA1* rs2740488 variant located in a liver-specific promoter region;

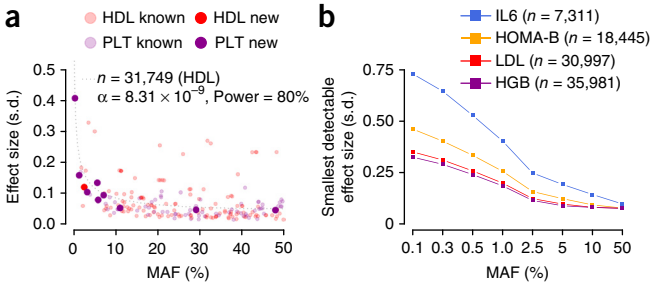

Figure 2 Allelic spectrum of cardiometabolic trait variants. (**a**) For each variant surpassing the genome-wide threshold in this study, the effect size (measured in s.d.) is plotted as a function of the MAF proportion. Associations with HDL (red) and PLT (purple) are shown, where loci discovered in this study are plotted with larger symbols. The dotted line represents the curve for 80% power with a sample size of 31,749 (for HDL) and α of 8.31×10^{-9}. The power line for PLT (sample size of 31,555) was similar and is therefore not shown here. (**b**) Plot of the smallest detectable effect size for a range of MAF proportions. Power calculations were performed for four traits from different trait groups with different sample sizes: IL6, HOMA-B, LDL and HGB.

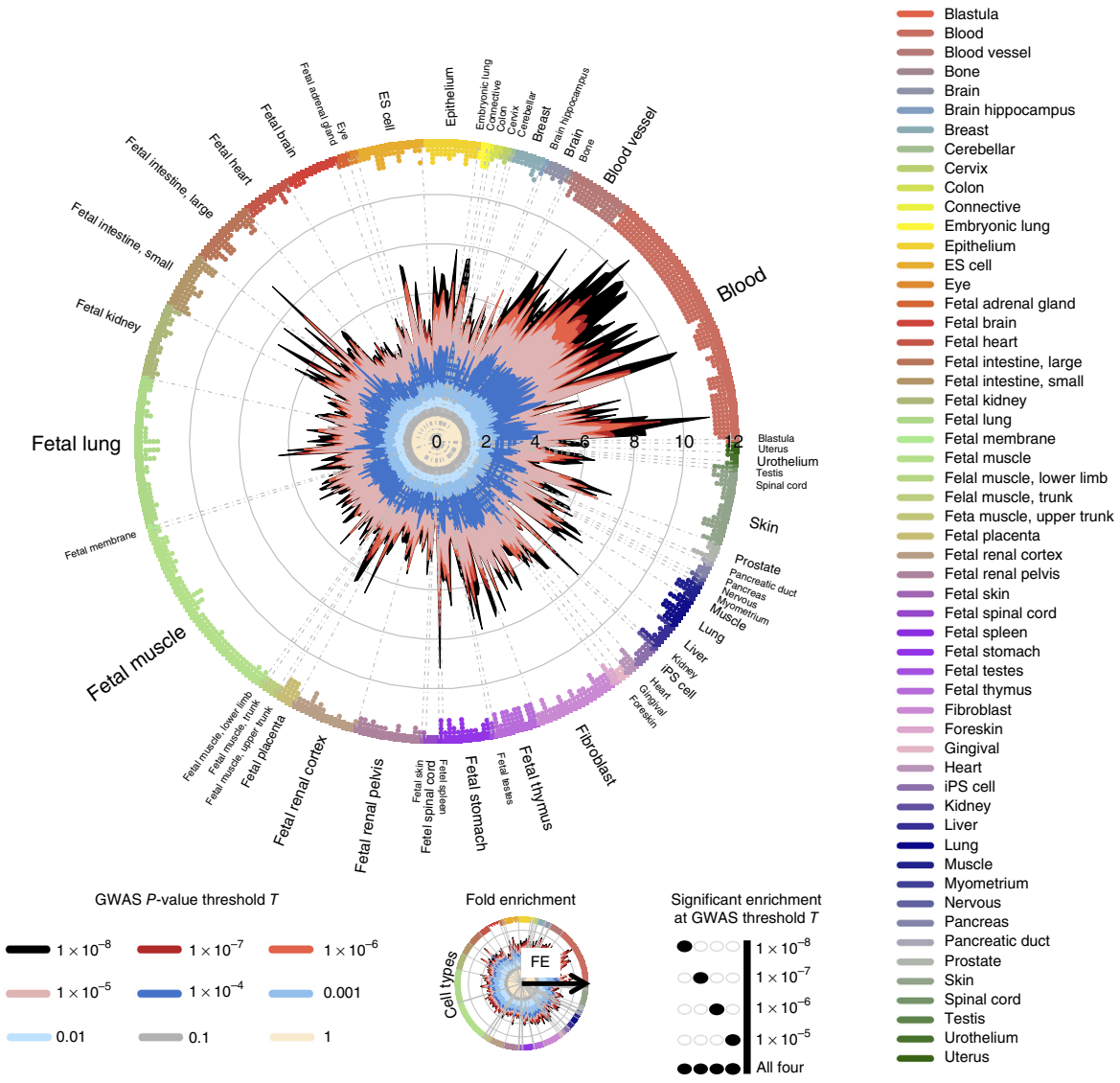

Figure 3 GARFIELD functional enrichment analyses. The wheel plot displays functional enrichment for associations with PLT within DHS hotspot regions in ENCODE and Roadmap Epigenomics studies. The radial axis shows fold enrichment (FE) calculated at each of eight GWAS *P*-value thresholds ($T < 1 \times 10^{-1}$ to $T < 1 \times 10^{-8}$) for each of 424 cell types. Cell types are sorted by tissue, represented along the outside edge of the plot with font size proportional to the number of cell types from that tissue. Boxes and dots next to the tissue labels are colored with respect to tissue. Fold enrichment values at the different thresholds are plotted with different colors inside the plot (for example, values at $T < 1 \times 10^{-8}$ are in black). Dots along the inside edge of the plot denote significant enrichment (if present) for a given cell type at $T < 1 \times 10^{-5}$ (outermost dot) to $T < 1 \times 10^{-8}$ (innermost dot). Results show overall well-spread enrichment, with the largest fold enrichment values obtained in blood, fetal spleen and fetal intestine tissues.

the PLT-associated variant rs12005199 located in a putative enhancer region upstream of *AK3* bound by GATA1, GATA2 and TAL1; the PCV-associated *HK1* intronic variant rs17476364 located in a hematopoietic cell enhancer region; and the TG-associated variant rs964184 located in a liver and fat enhancer within the *ZNF259* 3′ UTR.

Regulatory annotation of locus-specific findings

To inform our statistical fine-mapping approach, for every variant in a credible set we applied two scores for regulatory function based on cell-type-specific DHSs: the deltaSVM score and the Contextual Analysis of Transcription Factor Occupancy (CATO) score (Online Methods)[27,28] (**Supplementary Table 6**). The functional activity of a variant's effect allele is predicted by the magnitude of the deltaSVM score, with the sign indicating increase or decrease in DNase I hypersensitivity and, therefore, transcription factor binding potential at the site. Similarly, the functional activity of a variant's effect allele is predicted by the CATO score, where scores of 0.1 have a 51% true positive rate for perturbing known transcription factor motifs, with the true positive rate increasing as the score increases to 1 (ref. 28). To identify putatively causal variants, we considered deltaSVM scores greater than 10 in absolute value, CATO scores >0.1 and high posterior probability from the statistical fine-mapping methods.

This union-of-methods approach identified several strong cases for causal variants. At the *TRIB1* locus associated with the TG, TC and LDL cholesterol traits, rs112875651 had the strongest evidence for causality from all three fine-mapping methods (0.517, 0.532 and 0.526) and from extreme CATO and deltaSVM scores (0.315 and −12.31, respectively). Other functional variants have been suggested for the

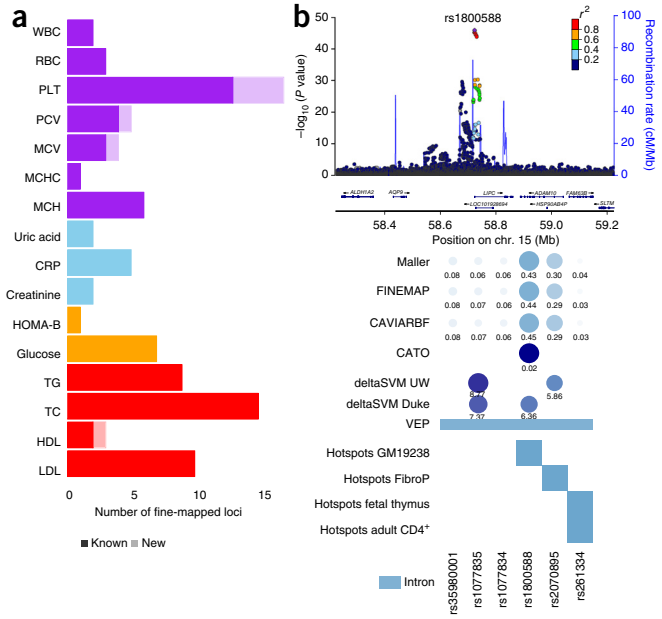

Figure 4 Fine-mapping experiments. Regional association plots for loci show fine-mapped variants. (**a**) Numerical summary of 59 loci that were fine-mapped. (**b**) Example of fine-mapping and annotation at the *LIPC* locus for association with HDL. The panels (from top to bottom) show the LocusZoom regional association plot; posterior probability (PP) statistics from the fine-mapping methods and CATO and deltaSVM scores; Variant Effect Predictor (VEP) genic annotation; and overlap with regulatory annotations found to be significant (1×10^{-4}; blue) in GARFIELD enrichment analysis. Circle sizes and colors for all scores were scaled with respect to score type (PP, CATO or deltaSVM), and numbers are plotted below each circle.

TRIB1 region, namely rs2001844 (ref. 29) ($r^2 = 0.8$) and rs6982502 (ref. 30) ($r^2 = 0.7$), but these SNPs were four orders of magnitude less significant than rs112875651 in our TG analysis, suggesting that rs112875651 may be a causal variant at *TRIB1*. At the *CELSR2* locus associated with LDL and TC, all three fine-mapping methods provided evidence for causality (0.205, 0.202 and 0.200) of rs12740374, although rs646776 ($r^2 > 0.8$) was a stronger predicted causal variant from the posterior probability estimates. However, additional evidence for rs12740374 as the causal variant came from a high CATO score (0.199) and an extreme deltaSVM score (14.37) for cell types with significant enrichment predicted by GARFIELD (liver and epithelial cells). The CATO and deltaSVM scores are also helpful when there are no obvious causal candidates from statistical fine-mapping. At the *CXCL2* locus associated with WBC, the posterior probabilities did not provide sufficiently strong evidence for a single causal variant. However, index variant rs13128896 had strong functional evidence from its high CATO score (0.146) and its extreme deltaSVM score (−10.71) for blood and skin cell types, with the former cell type being enriched for WBC associations in the GARFIELD analysis.

Integration of methods to prioritize variants for follow-up

We further combined information from fine-mapping analysis, genome-wide functional enrichment results and regulatory scores to assess the overall evidence supporting functional and causal interpretation at 66 independent regions (in 59 loci). There were 17 regions with at least one coding variant, 33 regions with support from both functional enrichment and regulatory scores, 9 regtions with functional scores only, and 6 regions with enrichment only (**Fig. 5a**). Variants with functional enrichment overlap and those with regulatory scores had larger posterior probabilities of causality (average PP increase of 0.3 and 0.1, respectively) (**Fig. 5b**), in contrast to variants with no such regulatory support, highlighting them as statistically more likely to be causal. For 24 of the 66 regions, we found functional or regulatory support for only a fraction of the variants within credible sets (**Fig. 5c**), ranging from 29–94% of variants with annotation from at least one type of evidence (mean = 74%, s.d. = 18%), resulting in up to a 71% reduction of the credible set. There was one fine-mapped region (*G6PC2* locus associated with glucose levels) with only statistical and no regulatory support; however, the credible set contained a single causal variant with PP >0.999 from all three fine-mapping approaches, and the variant has previously been shown to enhance *G6PC2* pre-mRNA splicing[31].

DISCUSSION

Our analysis demonstrates the utility of deep imputation from whole-genome sequence reference panels for informing studies of quantitative cardiometabolic and hematological traits. By combining the UK10K and 1000 Genomes Project sequence data, we constructed a dense imputation reference panel that substantially improves upon the HapMap 2 and 1000 Genomes Project panels. With this reference panel, we investigated associations with variants with a frequency as low as 0.5%.

Consistent with previous reports[17,32], our imputation accuracy declined with decreasing allele frequencies. Therefore, we did not consider very rare variants (MAF ≤ 0.001) or variants with poor imputation quality (INFO ≤ 0.4). This resulted in a substantial culling of the total number of variants that were identified in the UK10K project. Thus, our study may have missed rare variant associations that would be identifiable in a larger study. Because genotype imputation provides model-based estimates of allelic probabilities

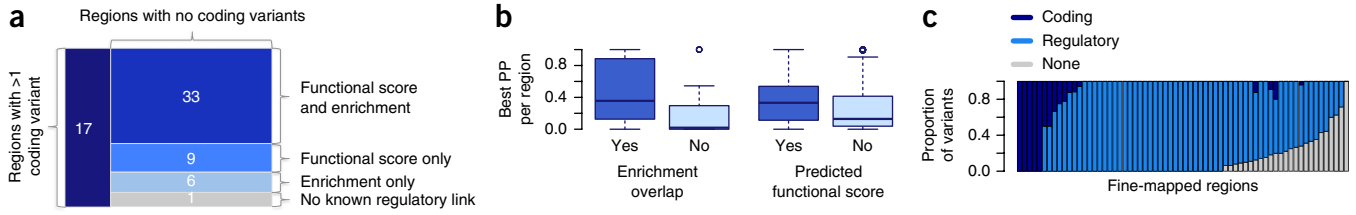

Figure 5 Summary of variant consequences for fine-mapped variants. (**a**) Number of fine-mapped trait–region pairs containing at least one variant in the 95% credible set with the following consequence: (i) being a coding variant; (ii) functional score and overlapping annotation significantly enriched for the given trait; (iii) functional score only; (iv) significant enrichment overlap only; or (v) none of the above. (**b**) Posterior probability distribution of the statistically most likely causal variant per region stratified by significant enrichment overlaps (left) and predicted functional scores (right) (after removing regions containing a coding variant). Box plots depict the median (thick horizontal line), first and third quartiles (colored box), maximum and minimum values (whiskers) and outlying values (circles). (**c**) Proportions of variants in credible sets with coding, regulatory or no annotation.

in the study subjects, rather than hard-called, empirically based genotypes, we could not reference cluster plots or intensity files to validate our findings. In this context, independent replication serves a critical function for validating associations from an imputation-based discovery effort.

Our dense imputation reference panels expanded the set of variants amenable to association analysis. Only one of the 17 new loci we report was well tagged ($r^2 > 0.8$) in HapMap 2 or 1000 Genomes Project Phase 1. Markers assessed in previous GWAS of PLT, hemoglobin and WBC poorly tagged nine of the new loci associated with hematological traits. However, for PLT (the trait for which we observed the most and strongest associations), the new loci identified here increased the percentage of phenotypic variance explained from 7.71% to 8.23%. Although increasingly large imputation panels are useful for investigating low-frequency and rare variants, considerably larger sample sizes are needed to identify rare variants of modest-to-large effect.

For each new locus identified, we undertook epigenomic, tissue expression and fine-mapping analyses to describe the potential mechanism of the association (**Box 1**). Our results implicate several genes or loci not previously known to be involved in regulation of blood cell counts. For example, the chromosome 22 PLT index variant rs75570992 is located upstream of *TRABD*, a gene of unknown function. On the basis of RNA–seq and epigenomic data from BLUEPRINT, *TRABD* is expressed in megakaryocytes. The index variant rs75570992 is associated with differential expression of *TRABD* in blood cells[33]. The index variant is in partial LD with rs75107793 ($r^2 = 0.5$), which lies upstream of the *TRABD* promoter in a putative megakaryocyte enhancer enriched for monomethylation of histone H3 at lysine 4 (H3K4me1) overlapping a ChIP–seq site for the hematopoietic transcription factor RUNX1.

Another newly discovered locus leading to new mechanistic insights is *GFI1B* rs150813342, a synonymous variant predicted to alter an exonic splicing enhancer. *GFI1B* is a hematopoietic transcription factor required for normal red blood cell and platelet production[34]. The rs150813342 variant influences the relative amounts of two *GFI1B* transcript isoforms, a full-length (long) isoform and a short isoform lacking the alternatively spliced exon 5 (ref. 22). We further demonstrate the lineage-specific role of the long *GFI1B* isoform in megakaryocyte development. Previous studies have suggested that the short *GFI1B* isoform is required for red blood cell production[35].

We identified several secondary, independent signals in genes previously implicated in regulation of blood cell counts (*CCDN3*, *NLPR3* and *THPO*). The new MCV-associated *CCDN3* low-frequency variant rs112233623 was also associated with hemoglobin A2 levels[36]. rs112233623 is located within an erythroid-specific enhancer[37] and is bound by the hematopoietic transcription factors GATA2 and TAL1. Similarly, *NLPR3* rs117747069 is located in an erythroid enhancer element involved in α-globin gene regulation and overlaps GATA2 and TAL1 ChIP–seq sites. A 3′ UTR variant of the thrombopoietin gene (*THPO* rs6141) was previously associated with higher PLT. We identify a second, independent 3′ UTR *THPO* signal, rs78565404. By ChIP–seq, rs78565404 is bound in liver HepG2 cells by musculoaponeurotic fibrosarcoma oncogene homolog K (MAFK), a component of the hematopoietic NF-E2 transcription factor complex involved in megakaryopoiesis[38,39].

Several of our newly identified variants are located within genes for congenital (*GFI1B* and *THPO*) or acquired (*APOH*) platelet disorders, underscoring the idea that more subtle genetic variation within genes known to contain loss-of-function variants may reflect between-individual differences in these complex traits. Rare loss-of-function *GFI1B* mutations have been identified in patients with congenital thrombocytopenia[40,41], while *THPO* mutations have more often been found in pedigrees with hereditary thrombocytosis. Most of the *THPO* mutations described in patients with familial thrombocytosis have involved noncoding sequence (splice-site, 5′ UTR and intronic) gain-of-function mutations that lead to enhanced *THPO* mRNA translation efficiency[42–45]. It remains to be determined whether the two common 3′ UTR variants of *THPO* associated with higher PLT similarly enhance mRNA translation and thrombopoietin synthesis. Recently, the first 'loss-of-function' *THPO* missense mutation (p.Arg38Cys) was associated with aplastic anemia in the homozygous state and mild thrombocytopenia in the heterozygous state[46].

Apolipoprotein H (APOH) is also known as β_2 glycoprotein I (β_2-GPI), a major autoantigen for antiphospholipid antibody syndrome (APS), a clinical disorder characterized by arterial and venous thrombosis[47,48]. Thrombocytopenia is also sometimes a feature of APS. The p.Cys325Gly variant encoded by *APOH* rs1801689 disrupts the β_2-GPI phospholipid-binding site[49] APOH/β_2-GPI is also a component of LDL and binds to members of the LDL receptor family. The same *APOH* rs1801689 missense variant associated with higher platelet count was recently associated with higher LDL[19]. β2-GPI and antiphospholipid antibody complexes bind to LRP8, an LDL receptor present on platelets and endothelial cells; this interaction has been postulated to have a role in β_2-GPI–mediated thrombosis[50,51]. However, even when we controlled for LDL levels, the rs1801689 association with platelet count remained intact, suggesting independent mechanisms driving the associations.

We undertook extensive fine-mapping of previously reported loci, identifying 59 loci where we could reduce associated signals to credible sets of 20 variants or fewer. We observed that the number of variants in the credible set was negatively correlated with the allele frequency of the index SNP, as expected because rare variants have fewer proxies on average. The newly identified loci had lower average MAFs and a lower number of proxies, making the identification of causative variants more straightforward. Rare variants were also more likely to have severe consequences or lead to changes in the protein, facilitating the identification of likely causative genes.

Our enrichment analyses showed that SNPs significantly associated with a phenotype of interest are over-represented within 'functional' regions that were derived in a broad range of cell types and tissues. We evaluated the extent to which genetic associations for each of the 20 traits were enriched in different functional domains and found that lipids and platelet counts were enriched in a large number of tissues and cell types in comparison to other traits displaying more localized (red blood cell traits) or null (renal and inflammatory traits) enrichment patterns. In combination with the fine-mapping experiments, we observed a positive correlation between the posterior probability of causality and overlap with significantly enriched annotations. Overall, this suggests that the process of sifting through putative causal variants can benefit from multiple-pronged approaches incorporating fine-mapping analysis to additional regulatory information obtained from epigenomes and deltaSVM and CATO scores. This information in turn empowers downstream functional experiments by guiding explorations of the functional consequences for sets of associated variants.

By performing detailed epigenomic and functional annotation, we were able to suggest several novel mechanisms for variants at known loci (for example, differential splicing for *GFI1B*, experimentally demonstrated in ref. 22) or posit strong biological candidates for further functional and cellular study on platelet production (for example, *TRABD*) and highlight potential genetic connections between

platelet count and traditional cardiovascular disease risk factors such as cholesterol levels (*APOH*). Imputation using dense genotype maps affords a greater understanding of the relative contribution of rare and low-frequency variants to complex traits and allows the fine-mapping of common variant association signals to manageable credible sets. In parallel, the development of robust functional enrichment methods and the overlap of fine-mapped associations with genome functional maps allowed us to pinpoint variants with high probability of being causal.

URLs. GARFIELD software is available in a standalone version at http://www.ebi.ac.uk/birney-srv/GARFIELD/ and as a Bioconductor package at http://bioconductor.org/packages/release/bioc/html/garfield.html. DeltaSVM scores were downloaded from http://www.beerlab.org/deltasvm/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This study makes use of data generated by the UK10K Consortium, derived from samples from the ALSPAC and TwinsUK data sets. A full list of the investigators who contributed to the generation of the data is available from http://www.UK10K.org/. Funding for UK10K was provided by the Wellcome Trust under award WT091310. The research of N.S. is supported by the Wellcome Trust (grants WT098051 and WT091310), the European Union Framework Programme 7 (EPIGENESYS grant 257082 and BLUEPRINT grant HEALTH-F5-2011-282510) and the National Institute for Health Research Blood and Transplant Research Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge in partnership with NHS Blood and Transplant (NHSBT). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or NHSBT. P.L.A. was supported by NHLBI R21 HL121422-02. A full list of grant support and acknowledgements can be found in the **Supplementary Note** and ref. 14.

AUTHOR CONTRIBUTIONS

Designed and/or managed individual studies and contributed data: A.B., A.D., A.G.U., A. Hamsten, A. Hofman, A.P.R., C.L., C.K., C.M.v.D., D.M., D.T., E.Z., G.G., H.W., J.C.C., J.S.K., L.F., M.A.S., M. Franberg, M. Frontini, N.J.T., N.S., P.G., P.L.A., R.A.S., R.P., W.M. Generated and/or performed quality control of data: A.-E.F., A. Hamsten, A. Hofman, A.I., A.M., B.S., C.S.F., E.M.v.L., F.R., G.L., G.M., G.Z., H.E., I.N., J.H., J.L., J.L.M., J.R.B.P., K.P., K.W., L.C., L.S., M.C., M.E.K., M.S.-L., M.T., N.A., O.H.F., S.-Y.-S., T.J., T.R.G., W.A., Y.M. Analyzed the data and provided critical interpretation of results: A.-E.F., A. Hamsten, A. Hofman, C.B., C.S.F., D.J., F.v.D., H.E., J.A.M., J.H., J.L.M., J.R.B.P., K.P., K.W., L.C., M.C., M.T.M., P.D., P.L.A., S.-Y.S., T.J., T.R.G., V.I., W.A., W.Z., Y.M. Provided tools or materials: A.P.R., E.Z., F.v.D., G.D., M.T.M., N.J.T., N.S., P.D. Wrote the manuscript: A.P.R., C.B., D.J., J.A.M., J.H., J.L.M., K.W., L.C., L.F., M.A.S., N.J.T., N.S., P.L.A., V.I. Evaluated the manuscript: A.B., A.D., A.-E.F., A.G.U., A. Hamsten, A. Hofman, A.I., A.M., A.P.R., B.S., C.B., C.L., C.K., C.S.F., C.M.v.D., D.J., D.M., D.T., E.M.v.L., E.Z., F.R., F.v.D., G.D., G.G., G.L., G.M., G.Z., H.E., H.W., I.N., J.A.M., J.C.C., J.H., J.L., J.L.M., J.R.B.P., J.S.K., K.P., K.W., L.C., L.F., L.S., M.A.S., M.C., M.E.K., M. Franberg, M. Frontini, M.S.-L., M.T., M.T.M., N.A., N.J.T., N.S., O.H.F., P.D., P.G., P.L.A., R.A.S., R.P., S.-Y.S., T.J., T.R.G., V.I., W.A., W.M., W.Z., Y.M. Designed and/or managed the project: A.P.R., N.J.T., N.S., P.L.A.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
2. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42**, 684–687 (2010).
3. Auer, P.L. *et al.* Rare and low-frequency coding variants in *CXCR2* and other genes are associated with hematological traits. *Nat. Genet.* **46**, 629–634 (2014).
4. Willer, C.J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
5. Huyghe, J.R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* **45**, 197–201 (2013).
6. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
7. Peloso, G.M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94**, 223–232 (2014).
8. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).
9. Auer, P.L. *et al.* Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am. J. Hum. Genet.* **91**, 794–808 (2012).
10. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
11. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
12. Moayyeri, A., Hammond, C.J., Hart, D.J. & Spector, T.D. Effects of age on genetic influence on bone loss over 17 years in women: the Healthy Ageing Twin Study (HATS). *J. Bone Miner. Res.* **27**, 2170–2178 (2012).
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. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
15. Timpson, N.J. *et al.* A rare variant in *APOC3* is associated with plasma triglyceride and VLDL levels in Europeans. *Nat. Commun.* **5**, 4871 (2014).
16. Taylor, P.N. *et al.* Whole-genome sequence-based analysis of thyroid function. *Nat. Commun.* **6**, 5681 (2015).
17. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
18. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
19. Do, R. *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45**, 1345–1352 (2013).
20. Okada, Y. *et al.* Meta-analysis identifies multiple loci associated with kidney function–related traits in east Asian populations. *Nat. Genet.* **44**, 904–909 (2012).
21. Meyer, T.E. *et al.* Genome-wide association studies of serum magnesium, potassium, and sodium concentrations identify six loci influencing serum magnesium levels. *PLoS Genet.* **6**, e1001045 (2010).
22. Polfus, L.M. *et al.* Whole-exome sequencing identifies loci associated with blood cell traits and reveals a role for alternative *GFI1B* splice variants in human hematopoiesis. *Am. J. Hum. Genet.* **99**, 481–488 (2016).
23. Service, S.K. *et al.* Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet.* **10**, e1004147 (2014).
24. Maller, J.B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
25. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
26. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
27. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
28. Maurano, M.T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy *in vivo*. *Nat. Genet.* **47**, 1393–1401 (2015).
29. Douvris, A. *et al.* Functional analysis of the *TRIB1* associated locus linked to plasma triglycerides and coronary artery disease. *J. Am. Heart Assoc.* **3**, e000884 (2014).
30. Iwamoto, S. *et al.* The role of TRIB1 in lipid metabolism; from genetics to pathways. *Biochem. Soc. Trans.* **43**, 1063–1068 (2015).
31. Baerenwald, D.A. *et al.* Multiple functional polymorphisms in the *G6PC2* gene contribute to the association with higher fasting plasma glucose levels. *Diabetologia* **56**, 1306–1316 (2013).
32. Duan, Q., Liu, E.Y., Croteau-Chonka, D.C., Mohlke, K.L. & Li, Y. A comprehensive SNP and indel imputability database. *Bioinformatics* **29**, 528–531 (2013).
33. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
34. Möröy, T., Vassen, L., Wilkes, B. & Khandanpour, C. From cytopenia to leukemia: the role of Gfi1 and Gfi1b in blood formation. *Blood* **126**, 2561–2569 (2015).
35. Laurent, B. *et al.* A short Gfi-1B isoform controls erythroid differentiation by recruiting the LSD1-CoREST complex through the dimethylation of its SNAG domain. *J. Cell Sci.* **125**, 993–1002 (2012).
36. Danjou, F. *et al.* Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nat. Genet.* **47**, 1264–1271 (2015).
37. Sankaran, V.G. *et al.* Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev.* **26**, 2075–2087 (2012).

38. Ono, Y. *et al.* Induction of functional platelets from mouse and human fibroblasts by p45NF-E2/Maf. *Blood* **120**, 3812–3821 (2012).
39. Shavit, J.A. *et al.* Impaired megakaryopoiesis and behavioral defects in mafG-null mutant mice. *Genes Dev.* **12**, 2164–2174 (1998).
40. Stevenson, W.S. *et al. GFI1B* mutation causes a bleeding disorder with abnormal platelet function. *J. Thromb. Haemost.* **11**, 2039–2047 (2013).
41. Monteferrario, D. *et al.* A dominant-negative *GFI1B* mutation in the gray platelet syndrome. *N. Engl. J. Med.* **370**, 245–253 (2014).
42. Wiestner, A., Schlemper, R.J., van der Maas, A.P. & Skoda, R.C. An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. *Nat. Genet.* **18**, 49–52 (1998).
43. Ghilardi, N., Wiestner, A., Kikuchi, M., Ohsaka, A. & Skoda, R.C. Hereditary thrombocythaemia in a Japanese family is caused by a novel point mutation in the thrombopoietin gene. *Br. J. Haematol.* **107**, 310–316 (1999).
44. Kondo, T. *et al.* Familial essential thrombocythemia associated with one-base deletion in the 5′-untranslated region of the thrombopoietin gene. *Blood* **92**, 1091–1096 (1998).
45. Liu, K. *et al.* A *de novo* splice donor mutation in the thrombopoietin gene causes hereditary thrombocythemia in a Polish family. *Haematologica* **93**, 706–714 (2008).
46. Dasouki, M.J. *et al.* Exome sequencing reveals a thrombopoietin ligand mutation in a Micronesian family with autosomal recessive aplastic anemia. *Blood* **122**, 3440–3449 (2013).
47. Giannakopoulos, B. & Krilis, S.A. The pathogenesis of the antiphospholipid syndrome. *N. Engl. J. Med.* **368**, 1033–1044 (2013).
48. De Groot, P.G., Meijers, J.C. & Urbanus, R.T. Recent developments in our understanding of the antiphospholipid syndrome. *Int. J. Lab. Hematol.* **34**, 223–231 (2012).
49. Sanghera, D.K., Wagenknecht, D.R., McIntyre, J.A. & Kamboh, M.I. Identification of structural mutations in the fifth domain of apolipoprotein H (β2-glycoprotein I) which affect phospholipid binding. *Hum. Mol. Genet.* **6**, 311–316 (1997).
50. Korporaal, S.J. *et al.* Binding of low density lipoprotein to platelet apolipoprotein E receptor 2′ results in phosphorylation of p38MAPK. *J. Biol. Chem.* **279**, 52526–52534 (2004).
51. Lutters, B.C. *et al.* Dimers of β2-glycoprotein I increase platelet deposition to collagen via interaction with phospholipids and the apolipoprotein E receptor 2′. *J. Biol. Chem.* **278**, 33831–33838 (2003).
52. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30**, 224–226 (2012).
53. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
54. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D1, D164–D171 (2016).
55. Xu, J. *et al.* Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev. Cell* **23**, 796–811 (2012).
56. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
57. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
58. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
59. Kiryluk, K. *et al.* Discovery of new risk loci for IgA nephropathy implicates genes involved in immunity against intestinal pathogens. *Nat. Genet.* **46**, 1187–1196 (2014).
60. Keller, M.F. *et al.* Trans-ethnic meta-analysis of white blood cell phenotypes. *Hum. Mol. Genet.* **23**, 6944–6960 (2014).
61. Vijai, J. *et al.* A genome-wide association study of marginal zone lymphoma shows association to the HLA region. *Nat. Commun.* **6**, 5751 (2015).
62. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–208 (2011).
63. Menicanin, D., Bartold, P.M., Zannettino, A.C. & Gronthos, S. Identification of a common gene expression signature associated with immature clonal mesenchymal cell populations derived from bone marrow and dental tissues. *Stem Cells Dev.* **19**, 1501–1510 (2010).
64. Konopatskaya, O. *et al.* PKCα regulates platelet granule secretion and thrombus formation in mice. *J. Clin. Invest.* **119**, 399–407 (2009).
65. Williams, C.M., Harper, M.T. & Poole, A.W. PKCα negatively regulates *in vitro* proplatelet formation and *in vivo* platelet production in mice. *Platelets* **25**, 62–68 (2014).
66. Kong, Y., Wang, H., Lin, T. & Wang, S. Sphingosine-1-phosphate/S1P receptors signaling modulates cell migration in human bone marrow–derived mesenchymal stem cells. *Mediators Inflamm.* **2014**, 565369 (2014).
67. Yang, L. *et al.* Sphingosine 1-phosphate receptor 2 and 3 mediate bone marrow–derived monocyte/macrophage motility in cholestatic liver injury in mice. *Sci. Rep.* **5**, 13423 (2015).
68. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
69. Hildebrand, J.D. Shroom regulates epithelial cell shape via the apical positioning of an actomyosin network. *J. Cell Sci.* **118**, 5191–5203 (2005).
70. Menon, M.C. *et al.* Intronic locus determines *SHROOM3* expression and potentiates renal allograft fibrosis. *J. Clin. Invest.* **125**, 208–221 (2015).
71. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
72. Grand, F.H. *et al.* p53-binding protein 1 is fused to the platelet-derived growth factor receptor β in a patient with a t(5;15)(q33;q22) and an imatinib-responsive eosinophilic myeloproliferative disorder. *Cancer Res.* **64**, 7216–7219 (2004).
73. Caulfield, M.J. *et al.* SLC2A9 is a high-capacity urate transporter in humans. *PLoS Med.* **5**, e197 (2008).
74. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* **45**, 145–154 (2013).

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [2]Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK. [3]Boston VA Research Institute, Boston, Massachusetts, USA. [4]Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, Quebec, Canada. [5]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [6]Department of Biostatistics, University of Washington, Seattle, Washington, USA. [7]Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Milan, Italy. [8]MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, UK. [9]Department of Hematology, University of Cambridge, Cambridge, UK. [10]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [11]Medical Genetics, Institute for Maternal and Child Health IRCCS 'Burlo Garofolo', Trieste, Italy. [12]Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste, Italy. [13]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands. [14]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands. [15]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece. [16]Cardiovascular Medicine Unit, Department of Medicine, Karolinska Institute, Stockholm, Sweden. [17]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands. [18]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [19]Medical Clinic V (Nephrology, Hypertensiology, Rheumatology, Endocrinology, Diabetology), Mannheim Medical Faculty, Heidelberg University, Mannheim, Germany. [20]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [21]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. [22]Department of Neurosciences, Biomedicine and Movement Sciences, Section of Biology and Genetics, University of Verona, Verona, Italy. [23]William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [24]Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands. [25]Wellcome Trust Centre for Human Genetics, Oxford, UK. [26]Renal Unit, Department of Medicine, University of Verona, Verona, Italy. [27]Department of Epidemiology and Biostatistics, Imperial College London, St Mary's Campus, London, UK. [28]A list of consortium members and affiliations can be found at http://www.uk10k.org/. [29]National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge, University of Cambridge, Cambridge, UK. [30]Division of Nephrology and Dialysis, Institute of Internal Medicine, Renal Program, Columbus-Gemelli University Hospital, Catholic University, Rome, Italy. [31]Experimental Genetics Division, Sidra, Doha, Qatar. [32]National Heart and Lung Institute, Imperial College London, London, UK. [33]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [34]Clinical Institute of Medical and Chemical Laboratory Diagnostics, Medical University of Graz, Graz, Austria. [35]Synlab Academy, Synlab Holding Deutschland, Mannheim, Germany. [36]LifeLines Cohort Study, University Medical Center Groningen, Groningen, the Netherlands. [37]Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. [38]Institute for Systems Genetics, New York University Langone Medical Center, New York, New York, USA. [39]Department of Epidemiology, University of Washington, Seattle, Washington, USA. [40]Zilber School of Public Health, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA. [41]These authors contributed equally to this work. Correspondence should be addressed to A.P.R. (apreiner@uw.edu), N.S. (ns6@sanger.ac.uk) or P.L.A. (pauer@uwm.edu).

ONLINE METHODS

Imputation. *Whole-genome-sequence-based haplotype reference panel.* A joint reference panel was created as described in ref. 17 by combining two large-scale, low-read-depth whole-genome sequencing data sets, TwinsUK and ALSPAC. The UK10K final-release whole-genome sequencing data for 3,781 samples and 49,826,943 sites were used. From this data set, multiallelic sites, sites containing alleles inconsistent with those from 1000 Genomes Project data and singletons not existing in the 1000 Genomes Project were removed, leaving 28,615,640 sites. SHAPEIT v2 (ref. 75) was used to rephase the haplotypes in 3-Mb chunks with flanking regions of ±250 kb. The phased chunks were then recombined with vcf-phased-join from the vcftools package[76]. The 1000 Genomes Project Phase I integrated variant set release (v3) for low-coverage whole genomes in NCBI Build 37 (hg19) coordinates was downloaded from the 1000 Genomes Project FTP site (23 November 2010 data freeze). This call set includes phased haplotypes for 1,092 individuals and 39,293,751 variants (22 autosomes and the X chromosome). For each chromosome, a summary file was generated and merged with that of the UK10K whole-genome sequencing data to identify multiallelic sites and singletons not polymorphic in UK10K. These sites were excluded to create a new set of VCF files. The final reference panel included all 1,092 samples and 32,506,604 sites. The VCFQUERY tool was used to convert the new VCF files into phased haplotypes and legend files for IMPUTE v2 (ref. 77).

Prephasing and imputation of target GWAS. Genome-wide SNP data were obtained from each individual study, having undergone study-specific quality control (**Supplementary Note**). These samples were prephased using SHAPEIT v2, with the mean size of the windows in which conditioning haplotypes were defined set to 0.5 Mb. Because of the significantly higher number of variants in the whole-genome sequencing data, the rephasing was conducted by 3-Mb chunks with 250-kb buffering regions. Phased genotypes were then imputed to one of the three whole-genome sequencing reference panels (UK10K alone, UK10K + 1000 Genomes Project, or 1000 Genomes Project + Genomes of the Netherland (GoNL)) as detailed in **Supplementary Table 1**. Imputation was carried out using IMPUTE v2 with standard settings[77].

Association testing. *Phenotype preparation.* All traits were available from previous studies. Information on trait measurements is summarized in **Supplementary Table 1**. Traits were transformed by inverse normalization (creatinine, glucose, HDL, HGB, HOMA-B, HOMA-IR, CRP, IL6, insulin, LDL, PCV, PLT, TC, TG and uric acid), square root transform (MCH) or log transform (WBC) or were left untransformed (MCHC, MCV and RBC) to meet the normality assumption for linear model association testing. Traits were further residualized on associated covariables for each trait and each population sample, following detailed information given in the UK10K project paper[14] (summarized in Supplementary Table 4 of ref. 14). Finally, ten principal components were additionally regressed out from all traits for cohorts with unrelated individuals to further control for potential confounding. Information on individual study characteristics, including trait values and potential additional cohort-specific covariates applied, is given in **Supplementary Table 2** and the **Supplementary Note**. Histograms of trait residuals for which inverse normalization was not applied are shown in **Supplementary Figure 4**.

Study design for association testing. The study design is shown in **Figure 1**. Briefly, a total of 12,267 to 35,981 participants from 18 different studies were included in the discovery sample. Each cohort carried out single-marker association testing using linear additive models. Genotype dosages were used to account for genotype uncertainty that might arise from sequencing, where each genotype was expressed on a quantitative scale [0:2]. Variants that did not pass a low-frequency threshold (MAF < 0.1%) or were imputed with low accuracy (defined by imputation info score <0.4) were excluded from the analysis. Meta-analyses of cohort summary statistics were performed using GWAMA v 2.1 (ref. 78) assuming a fixed-effect model. Genomic control was used to adjust the summary statistics for both input and output data. We prioritized for replication all variants that reached $P \leq 5 \times 10^{-8}$ from the meta-analysis of 23 studies. During the course of the study, we updated our meta-analyses several times; variants were prioritized for replication if they met our cutoff (5×10^{-8}) during any of these updates. These variants were taken forward into 2,141–102,505 additional independent samples from seven cohorts (**Supplementary Table 1**), depending on the trait. Evidence for validation was based on a Bonferroni-corrected stage 2 P value of 8.6×10^{-4} (0.05/58) and a joint meta-analysis P value of 8.31×10^{-9} (ref. 14).

Fine-mapping of associated loci (new and previously identified GWAS regions). *Annotation and selection of index variants for previously reported loci.* For each trait, we compiled a list of known loci by selecting all index SNPs associated with our traits of interest (lipids, fasting glucose, HOMA, uric acid, CRP, and blood cell counts and indices) from the National Human Genome Research Institute (NHGRI) GWAS catalog ($P \leq 5 \times 10^{-8}$; last updated in May 2014), supplemented by manual curation of all associations reported in the literature reaching the same genome-wide significance cutoff. Only index variants with marginal significance in the UK10K whole-genome sequencing cohort single-marker association statistics ($P \leq 0.05$) were considered for conditional tests. Using TwinsUK and ALSPAC sequence data, we selected variants with P values less than 1×10^{-3} in the two-way meta-analysis. For each such variant, we extracted regions for fine-mapping on the basis of HapMap estimates of recombination rates. When a region contained multiple correlated index variants associated with a given trait in the GWAS catalog, we clumped the set of index variants to remove ones that were highly correlated (using an LD metric of $r^2 > 0.8$ applied to a 2-Mb sliding window for each known index SNP (±1 Mb)). This approach avoids collinearity errors when a variant is conditioned against multiple correlated index variants.

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

Conditional analyses. Sequential conditional single-variant association analyses were carried out to confirm statistical independence of associations. In the initial round of conditional analysis, associations of SNVs with the respective quantitative trait were conditioned on the index variants for known loci clumped ($r^2 > 0.8$) as described before (this step was carried out only for SNVs within ±1 Mb of a known locus); in further rounds, associations were conditioned against all nearby known loci plus the best new variant identified in the previous round of conditional analysis. The conditional analysis was performed independently for each cohort and a meta-analysis was conducted at the end of each round until the conditional association P value was no longer significant ($P > 1 \times 10^{-5}$). A variant was considered independent if it had conditional $P \leq 1 \times 10^{-5}$ (corresponding to $r^2 < 0.2$ in our data).

Finally, variants were classified as 'known' (denoting either a previously reported GWAS index variant or a variant for which the association signal disappears after conditioning on the known locus) or 'new' (denoted as a variant that is still conditionally independent of known loci and eventual other new independent signals in that region). For new signals, the variant with the lowest conditional P value among multiple associated variants is reported.

Bayesian fine-mapping methods. For each previously reported (known) association and each new index variant, we extracted regions for fine-mapping on the basis of HapMap estimates of recombination rates according to Maller *et al.*[24]. Specifically, the boundaries were chosen to be at a distance of at least 0.1 cM on either side of the index or known SNP and, if necessary, were extended further to include all tagging variants ($r^2 > 0.1$ within 1-Mb windows). Of the previously reported loci, only informative associations ($P \leq 1 \times 10^{-5}$ in the discovery-stage analysis) were taken forward. Regions with multiple SNPs reported to be associated with the same trait were merged if overlapping. Analysis of each region was then performed separately using three different methods. We implemented the method of Maller *et al.*[24], by converting our discovery-stage meta-analysis P values to Bayes' factors of association using Wakefield's approximation[79]. Additionally, we employed the fine-mapping methods CAVIARBF[80] and FINEMAP[25], both Bayesian approaches that use association summary statistics (rather than original genotypic data) and SNP correlations to compute Bayes' factors. The Bayes' factors from each method were then used to calculate posterior probabilities, on the basis of the assumption that there is a single causal SNP in each region. Conditional association analysis on the top fine-mapped variant was additionally carried out and (conditional)

fine-mapping was performed to fine-map secondary associations. For all regions, 95% credible sets were constructed to assess the uncertainty of the fine-mapping analyses. To assess the suitability of our two-stage fine-mapping approach (conditional steps) in the presence of multiple causal variants, we further compared our results to those obtained from FINEMAP under a relaxed assumption of multiple causal variants (**Supplementary Table 7** and **Supplementary Note**).

Enrichment of GWAS SNPs in functional and regulatory elements. To systematically characterize the functional, cellular and regulatory contribution of genetic variation implicated for each quantitative trait, we used GARFIELD, a non-parametric enrichment analysis approach that takes genome-wide association summary statistics to calculate fold enrichment values at given significance thresholds and then tests them for significance via permutation testing while accounting for LD, MAF and local gene density. We used a range of functional annotations, including genic elements (GENCODE), DHSs, transcription factor binding sites, histone modifications and chromatin states (ENCODE and Roadmap Epigenomics) (**Supplementary Table 4**), and included different cell types and tissues to capture and characterize possible cell-type-specific patterns of enrichment. We calculated fold enrichment statistics at eight genome-wide significance thresholds T (in powers of 10) and tested their significance at the four most stringent ones (1×10^{-8} to 1×10^{-5}) to analyze both stringent association findings and nominal ones. Multiple-testing correction was further performed on the effective number of annotations used, resulting in an enrichment P-value threshold of 1×10^{-4}. Further information on the approach is provided in the **Supplementary Note**.

Scoring credible set variants for regulatory function. DeltaSVM scores were generated as previously published by training the gapped k-mer support vector machine (gkmSVM) on cell-type-specific DHSs, computing weights for all possible 10-mers of the genome on the basis of the SVM classifier and calculating the difference in weights of 10-mers encompassing the reference and effect alleles for the variant of interest[27]. Precomputed weights were available from a total of 222 ENCODE DHS samples—99 from the Duke University (Duke) set and 123 from the University of Washington (UW) set[81]. Genetic variants were scored for deltaSVM in all 222 cell lines and filtered for those with at least one deltaSVM score greater than an absolute value of 5, allowing putative inference of relevant cell types or tissues. CATO scores were generated as described in ref. 28. Briefly, logistic models were fit to imbalance in DNA accessibility in 443 DNase–seq data sets from the ENCODE and Roadmap Epigenomics projects. An independent model was fit for each of 44 transcription factor families and included terms for both the effect of the variant on the transcription factor position weight matrix and genomic context. Genetic variants were then scored by taking the maximum prediction for all overlapping transcription factor models. CATO scores greater than 0.1 were shown to have a 51% true positive rate on the initial training set and are therefore of interest[28].

75. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
76. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
77. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
78. Mägi, R. & Morris, A.P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
79. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
80. Chen, W. *et al.* Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200**, 719–736 (2015).
81. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

doi:10.1038/ng.3668