nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0084-1

Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes

Anubha Mahajan*

We aggregated coding variant data for 81,412 type 2 diabetes cases and 370,832 controls of diverse ancestry, identifying 40 coding variant association signals ($P < 2.2 \times 10^{-7}$); of these, 16 map outside known risk-associated loci. We make two important observations. First, only five of these signals are driven by low-frequency variants: even for these, effect sizes are modest (odds ratio ≤1.29). Second, when we used large-scale genome-wide association data to fine-map the associated variants in their regional context, accounting for the global enrichment of complex trait associations in coding sequence, compelling evidence for coding variant causality was obtained for only 16 signals. At 13 others, the associated coding variants clearly represent 'false leads' with potential to generate erroneous mechanistic inference. Coding variant associations offer a direct route to biological insight for complex diseases and identification of validated therapeutic targets; however, appropriate mechanistic inference requires careful specification of their causal contribution to disease predisposition.

Genome-wide association studies (GWAS) have identified thousands of association signals influencing multifactorial traits such as type 2 diabetes (T2D) and obesity[1–7]. Most of these associations involve common variants that map to noncoding sequence, and identification of their cognate effector transcripts has proved challenging. Identification of coding variants causally implicated in trait predisposition offers a more direct route from association signal to biological inference.

The exome occupies 1.5% of the overall genome sequence, but, for many common diseases, coding variants make a disproportionate contribution to trait heritability[8,9]. This enrichment indicates that coding variant association signals have an enhanced probability of being causal when compared to signals involving an otherwise equivalent noncoding variant. This does not, however, guarantee that all coding variant associations are causal. Alleles driving common variant (minor allele frequency (MAF) ≥5%) GWAS signals typically reside on extended risk haplotypes that, owing to linkage disequilibrium (LD), incorporate many common variants[10,11]. Consequently, the presence of a coding allele on the risk haplotype does not constitute sufficient evidence that it represents the causal variant at the locus or that the gene within which it lies is mediating the association signal. Because much coding variant discovery has proceeded through exome-specific analyses (either exome array genotyping or exome sequencing), researchers have often been poorly placed to position coding variant associations in the context of regional genetic variation. It is unclear how often this may have led to incorrect assumptions regarding the causal role of coding variants.

In our recent study of T2D predisposition[12], we surveyed the exomes of 34,809 cases and 57,985 controls, of predominantly European descent, and identified 13 distinct coding variant associations reaching genome-wide significance. Twelve of these associations involved common variants, but the data hinted at a substantial pool of lower-frequency coding variants of moderate impact, potentially amenable to detection in larger samples. We also reported that, while many of these signals fell within common variant loci previously identified by GWAS, it was far from trivial to determine, using available data, whether those coding variants were causal or 'hitchhiking' on risk haplotypes.

Here we report analyses that address these two issues. First, we extended the scope of our exome array genotyping to include data from 81,412 T2D cases and 370,832 controls of diverse ancestry, substantially increasing power to detect coding variant associations across the allele frequency spectrum. Second, to understand the extent to which identification of coding variant associations provides a reliable guide to causal mechanisms, we undertook high-resolution fine-mapping of identified coding variant association signals in 50,160 T2D cases and 465,272 controls of European ancestry with genome-wide genotyping data.

Results

Discovery study overview. First, we set out to discover coding variant association signals by aggregating T2D association summary statistics in up to 452,244 individuals (effective sample size of 228,825) across five ancestry groups, performing both European-specific (EUR) and trans-ethnic (TE) meta-analyses (Supplementary Tables 1 and 2). Analysis was restricted to the 247,470 variants represented on the exome array. Genotypes were assembled from (i) 58,425 cases and 188,032 controls genotyped with the exome array; (ii) 14,608 cases and 174,322 controls from UK Biobank and GERA (Resource for Genetic Epidemiology on Adult Health and Aging) genotyped with GWAS arrays enriched for exome content and/or coverage of low-frequency variation across ancestry groups[13,14]; and (iii) 8,379 cases and 8,478 controls with whole-exome sequence from the GoT2D/T2D-GENES[12] and SIGMA[15] studies. Overall, this represented a threefold increase in effective sample size over our previous study of T2D predisposition within coding sequence[12]. To deconvolute the impact of obesity on T2D-associated variants, association analyses were conducted with and without adjustment for body mass index (BMI).

We considered $P < 2.2 \times 10^{-7}$ as significant for protein-truncating variants (PTVs) and moderate-impact coding variants (including missense, in-frame indel, and splice-region variants) on the basis of a weighted Bonferroni correction that accounts for the observed enrichment in complex trait association signals across sequence annotation[16]. This threshold matches those obtained through other approaches such as simple Bonferroni correction for the number of coding variants on the exome array (Methods). In comparison to our

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

previous study[12], the expanded sample size substantially increased power to detect association for common variants of modest effect (for example, power was increased from 14.4% to 97.9% for a variant with 20% MAF and odds ratio (OR) = 1.05) and lower-frequency variants with larger effects (for example, power was increased from 11.8% to 97.5% for a variant with 1% MAF and OR = 1.20) assuming homogenous allelic effects across ancestry groups (Methods).

Insights into coding variant association signals underlying T2D susceptibility. We detected significant associations at 69 coding variants under an additive genetic model (either in BMI-unadjusted or BMI-adjusted analysis), mapping to 38 loci (Supplementary Fig. 1 and Supplementary Table 3). We observed minimal evidence of heterogeneity in allelic OR between ancestry groups (Supplementary Table 3) and no compelling evidence for non-additive allelic effects (Supplementary Fig. 2 and Supplementary Table 4). Reciprocal conditional analyses (Methods) indicated that the 69 coding variants represented 40 distinct association signals (conditional $P < 2.2 \times 10^{-7}$) across the 38 loci, with 2 distinct signals each at *HNF1A* and *RREB1* (Supplementary Table 5). These 40 signals included the 13 associations reported in our earlier publication[12], each featuring more significant associations in this expanded meta-analysis (Supplementary Table 6). Twenty-five of the 40 signals were significant in both EUR and TE analyses. Of the other 15, 3 (*PLCB3*, *C17orf58*, and *ZHX3*) were significant in EUR analysis and all reached $P_{TE} < 6.8 \times 10^{-6}$ in the TE analysis: for *PLCB3* and *ZHX3*, risk allele frequencies were substantially lower outside European-descent populations. Twelve loci (Supplementary Table 3) were significant in TE analysis alone, but for these (except *PAX4*, which is specific to East Asians) the evidence for association was proportionate in the smaller EUR component ($P_{EUR} < 8.4 \times 10^{-5}$).

Sixteen of the 40 distinct association signals mapped outside regions previously implicated in T2D susceptibility (Table 1 and Methods). These included missense variant signals in *POC5* (p.His36Arg, rs2307111, $P_{TE} = 1.6 \times 10^{-15}$), *PNPLA3* (p.Ile148Met, rs738409, P_{TE} BMI adjusted $= 2.8 \times 10^{-11}$), and *ZZEF1* (p.Ile2014Val, rs781831, $P_{TE} = 8.3 \times 10^{-11}$).

In addition to the 69 coding variant signals, we detected significant ($P < 5 \times 10^{-8}$) and new T2D associations for 20 noncoding variants (at 15 loci) that were also assayed on the exome array (Supplementary Table 7). Three of these (*POC5*, *LPL*, and *BPTF*) overlapped with new coding signals reported here.

Contribution of low-frequency and rare coding variation to T2D susceptibility. Despite increased power and good coverage of low-frequency variants on the exome array[12], 35 of the 40 distinct coding variant association signals were common, with modest effects (allelic ORs of 1.02–1.36) (Supplementary Fig. 3 and Supplementary Table 3). The five signals attributable to lower-frequency variants were also of modest effect (allelic ORs of 1.09–1.29) (Supplementary Fig. 3). Two of the lower-frequency variant signals were new, and, for both, the minor allele was protective against T2D: *FAM63A* p.Tyr95Asn (rs140386498, MAF = 1.2%, OR = 0.82 (0.77–0.88, 95% confidence interval), $P_{EUR} = 5.8 \times 10^{-8}$) and *ANKH* p.Arg187Gln (rs146886108, MAF = 0.4%, OR = 0.78 (0.69–0.87), $P_{EUR} = 2.0 \times 10^{-7}$). Both variants were very rare or monomorphic in individuals not of European descent.

In our previous study[12], we highlighted a set of 100 low-frequency coding variants with allelic ORs between 1.10 and 2.66 that, despite relatively large estimates for liability-scale variance explained, had not reached significance. In this expanded analysis, only five of these variants, including two new associations (at *FAM63A* p.Tyr95Asn and *ANKH* p.Arg187Gln), attained significance. More precise effect size estimation in the larger sample size indicates that OR estimates in the earlier study were subject to a substantial upward bias (Supplementary Fig. 3).

To detect additional rare variant association signals, we performed gene-based analyses (burden and SKAT[17]) using previously defined 'strict' and 'broad' masks, filtered for annotation and MAF[12,18] (Methods). We identified gene-based associations with T2D susceptibility ($P < 2.5 \times 10^{-6}$, Bonferroni correction for 20,000 genes) for *FAM63A* (10 variants, combined MAF = 1.9%, $P_{EUR} = 3.1 \times 10^{-9}$) and *PAM* (17 variants, combined MAF = 4.7%, $P_{TE} = 8.2 \times 10^{-9}$). On conditional analysis (Supplementary Table 8), the gene-based signal at *FAM63A* was entirely attributable to the low-frequency allele encoding p.Tyr95Asn described earlier (conditional $P_{EUR} = 0.26$). The gene-based signal for *PAM* was also driven by a single low-frequency variant (p.Asp563Gly; conditional $P_{TE} = 0.15$). A second previously described, low-frequency variant (*PAM* p.Ser539Trp[19]) is not represented on the exome array and did not contribute to these analyses.

Fine-mapping of coding variant association signals with T2D susceptibility. These analyses identified 40 distinct coding variant associations with T2D, but this information is not sufficient to determine that these variants are causal for disease. To assess the role of these coding variants given regional genetic variation, we fine-mapped these association signals using a meta-analysis of 50,160 T2D cases and 465,272 controls (European descent only; partially overlapping with the discovery samples), which we aggregated from 24 GWAS. Each component GWAS was imputed using appropriate high-density reference panels (for most, the Haplotype Reference Consortium[20]; Methods and Supplementary Table 9). Before fine-mapping, distinct association signals were delineated using approximate conditional analyses (Methods and Supplementary Table 5). We included 37 of the 40 identified coding variants in this fine-mapping analysis, excluding 3 (those in the major histocompatibility complex (MHC) region, *PAX4*, and *ZHX3*) that were, for various reasons (see the Methods), not amenable to fine-mapping in the GWAS data.

For each of these 37 signals, we first constructed 'functionally unweighted' credible variant sets, which collectively account for 99% of the posterior probability of association (PPA), based exclusively on the meta-analysis summary statistics[21] (Methods and Supplementary Table 10). For each signal, we calculated the proportion of PPA attributable to coding variants (missense, in-frame indel, and splice-region variants; Fig. 1 and Supplementary Figs. 4 and 5). There were only two signals at which coding variants accounted for ≥80% of PPA: *HNF4A* p.Thr139Ile (rs1800961, PPA > 0.999) and *RREB1* p.Asp1171Asn (rs9379084, PPA = 0.920). However, at other signals, including those for *GCKR* p.Pro446Leu and *SLC30A8* p.Arg276Trp, for which robust empirical evidence has established a causal role[22,23], genetic support for coding variant causation was weak. This is because coding variants were typically in high LD ($r^2 > 0.9$) with large numbers of noncoding variants, such that the PPA was distributed across many sites with broadly equivalent evidence for association.

These functionally unweighted sets are based on genetic fine-mapping data alone and do not account for the disproportionate representation of coding variants among GWAS associations for complex traits[8,9]. To accommodate this information, we extended the fine-mapping analyses by incorporating an 'annotation-informed prior' model of causality. We derived priors from estimates of the enrichment of association signals by sequence annotation from analyses conducted by deCODE across 96 quantitative and 123 binary phenotypes[16] (Methods). This model 'boosts' the prior and, hence, the posterior probabilities (we use '$_{ai}$PPAs' to denote annotation-informed PPAs) of coding variants. It also takes into account (in a tissue-non-specific manner) the GWAS enrichment of variants within enhancer elements (as assayed through DNase I hypersensitivity) when compared to noncoding variants mapping elsewhere. The annotation-informed model generated smaller 99% credible sets across most signals, corresponding to fine-mapping at higher resolution (Supplementary Table 10). As expected, the contribution of coding variants was increased under

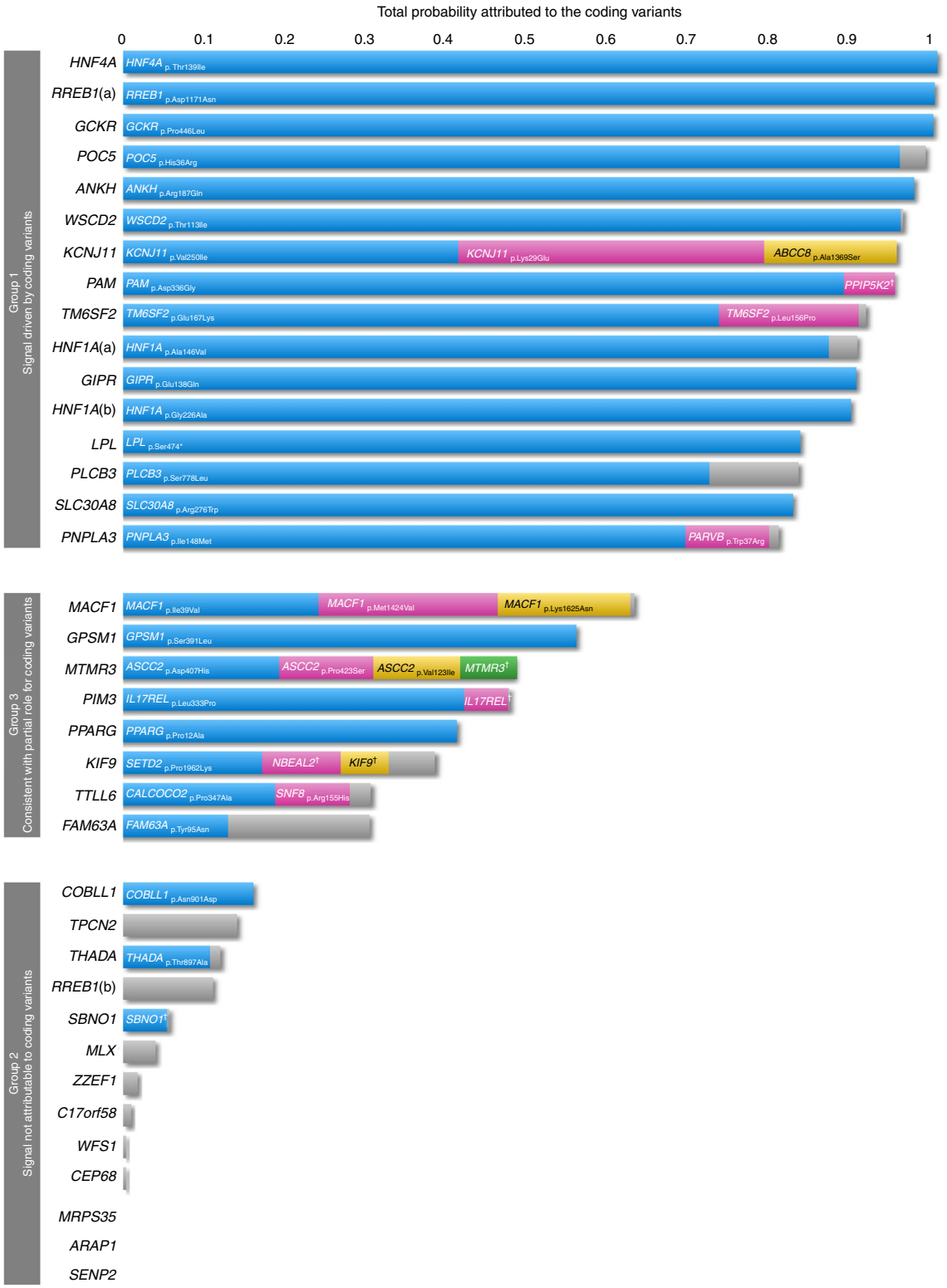

Fig. 1 | Posterior probabilities for coding variants across loci with annotation-informed priors. Fine-mapping of 37 distinct association signals was performed using European-ancestry GWAS meta-analysis including 50,160 T2D cases and 465,272 controls. For each signal, we constructed a credible set of variants accounting for 99% of the posterior probability of driving the association, incorporating an annotation-informed prior model of causality, which boosts the posterior probability of driving the association signal that is attributed to coding variants. Each bar represents a signal, with the total probability attributed to the coding variants within the 99% credible set plotted on the *y* axis. When the probability (bar) is split across multiple coding variants (at least 0.05 probability attributed to a variant) at a particular locus, these variants are indicated by blue, pink, yellow, and green. The combined probability of the remaining coding variants is highlighted in gray. *RREB1*(a), *RREB1* p.Asp1171Asn; *RREB1*(b), *RREB1* p.Ser1499Tyr; *HNF1A*(a), *HNF1A* p.Ala146Val; *HNF1A*(b), *HNF1A* p.Ile75Leu; *PPIP5K2*†, *PPIP5K2* p.Ser1207Gly; *MTMR3*†, *MTMR3* p.Asn960Ser; *IL17REL*†, *IL17REL* p.Gly70Arg; *NBEAL2*†, *NBEAL2* p.Arg511Gly, *KIF9*†, *KIF9* p.Arg638Trp

Table 1 | Summary of discovery and fine-mapping analyses of the 40 index coding variants associated with T2D ($P < 2.2 \times 10^{-7}$)

Locus	Index variant	rsID	Chr.	Pos. (bp)	Alleles (R/O)	RAF	Discovery meta-analysis using exome array component: 81,412 T2D cases and 370,832 controls from diverse ancestries								Fine-mapping meta-analysis using GWAS: 50,160 T2D cases and 465,272 controls of European ancestry					
							BMI unadjusted				BMI adjusted				RAF	OR	L95	U95	*P*	Group
							OR	L95	U95	*P*	OR	L95	U95	*P*						
Previously reported T2D-associated loci																				
MACF1	*MACF1* p.Met1424Val	rs2296172	1	39,835,817	G/A	0.193	1.06	1.05	1.08	6.7×10^{-16}	1.04	1.03	1.06	5.9×10^{-8}	0.22	1.08	1.06	1.10	1.6×10^{-15}	3
GCKR	*GCKR* p.Pro446Leu	rs1260326	2	27,730,940	C/T	0.630	1.06	1.05	1.08	5.3×10^{-25}	1.06	1.04	1.07	3.2×10^{-18}	0.607	1.05	1.04	1.07	9.1×10^{-10}	1
THADA	*THADA* p.Cys845Tyr	rs35720761	2	43≠,519,977	C/T	0.895	1.08	1.05	1.10	4.6×10^{-15}	1.07	1.05	1.10	8.3×10^{-16}	0.881	1.1	1.07	1.12	3.4×10^{-12}	2
GRB14	*COBLL1* p.Asn901Asp	rs7607980	2	165,551,201	T/C	0.879	1.08	1.06	1.11	8.6×10^{-20}	1.09	1.07	1.12	5.0×10^{-23}	0.871	1.08	1.06	1.11	3.6×10^{-10}	2
PPARG	*PPARG* p.Pro12Ala	rs1801282	3	12,393,125	C/G	0.887	1.09	1.07	1.11	1.4×10^{-17}	1.10	1.07	1.12	2.7×10^{-19}	0.876	1.12	1.09	1.14	3.7×10^{-17}	3
IGF2BP2	*SENP2* p.Thr291Lys	rs6762208	3	185,331,165	A/C	0.367	1.03	1.01	1.04	1.6×10^{-6}	1.03	1.02	1.05	3.0×10^{-8}	0.339	1.02	1.01	1.04	0.01	2
WFS1	*WFS1* p.Val333Ile	rs1801212	4	6,302,519	A/G	0.748	1.07	1.06	1.09	1.1×10^{-24}	1.07	1.05	1.08	7.1×10^{-21}	0.703	1.07	1.05	1.09	4.1×10^{-13}	2
PAM-PPIP5K2	*PAM* p.Asp336Gly	rs35658696	5	102,338,811	G/A	0.045	1.13	1.10	1.17	1.2×10^{-16}	1.13	1.09	1.17	7.4×10^{-15}	0.051	1.17	1.13	1.22	2.5×10^{-17}	1
RREB1	*RREB1* p.Asp1171Asn	rs9379084	6	7,231,843	G/A	0.884	1.08	1.06	1.11	1.1×10^{-13}	1.10	1.07	1.13	1.5×10^{-17}	0.888	1.09	1.06	1.12	1.1×10^{-9}	1
	RREB1 p.Ser1499Tyr	rs35742417	6	7,247,344	C/A	0.836	1.04	1.03	1.06	5.5×10^{-8}	1.04	1.02	1.06	2.2×10^{-7}	0.817	1.04	1.02	1.07	0.00012	2
MHC	*TCF19* p.Met131Val	rs2073721	6	31,129,616	G/A	0.749	1.04	1.02	1.05	1.6×10^{-10}	1.04	1.02	1.05	2.3×10^{-9}	NA	NA	NA	NA	NA	NA
PAX4	*PAX4* p.Arg190His	rs2233580	7	127,253,550	T/C	0.029	1.36	1.25	1.48	1.8×10^{-12}	1.38	1.26	1.51	4.2×10^{-13}	0	NA	NA	NA	NA	NA
SLC30A8	*SLC30A8* p.Arg276Trp	rs13266634	8	118,184,783	C/T	0.691	1.09	1.08	1.11	1.9×10^{-47}	1.09	1.08	1.11	1.3×10^{-47}	0.683	1.12	1.10	1.14	8.2×10^{-36}	1
GPSM1	*GPSM1* p.Ser391Leu	rs60980157	9	139,235,415	C/T	0.771	1.06	1.05	1.08	3.2×10^{-16}	1.06	1.05	1.08	6.6×10^{-16}	0.756	1.06	1.04	1.09	8.3×10^{-8}	3
KCNJ11-ABCC8	*KCNJ11* p.Lys29Glu	rs5219	11	17,409,572	T/C	0.364	1.06	1.05	1.07	5.7×10^{-22}	1.07	1.05	1.08	1.5×10^{-22}	0.381	1.07	1.05	1.09	8.1×10^{-16}	1
CENTD2	*ARAP1* p.Gln802Glu	rs56200889	11	72,408,055	G/C	0.733	1.04	1.02	1.05	4.8×10^{-8}	1.05	1.03	1.06	5.2×10^{-10}	0.727	1.05	1.03	1.07	2.3×10^{-8}	2
KLHDC5	*MRPS35* p.Gly43Arg	rs1127787	12	27,867,727	G/A	0.850	1.06	1.04	1.08	1.4×10^{-11}	1.05	1.03	1.07	1.5×10^{-8}	0.842	1.06	1.04	1.09	2.2×10^{-7}	2
HNF1A	*HNF1A* p.Ile75Leu	rs1169288	12	121,416,650	C/A	0.323	1.04	1.03	1.06	1.1×10^{-11}	1.04	1.02	1.06	1.9×10^{-10}	0.33	1.05	1.04	1.07	4.6×10^{-9}	1
	HNF1A p.Ala146Val	rs1800574	12	121,416,864	T/C	0.029	1.11	1.06	1.15	6.1×10^{-8}	1.10	1.06	1.15	1.3×10^{-7}	0.03	1.16	1.10	1.21	5.0×10^{-9}	1
MPHOSPH9	*SBNO1* p.Ser729Asn	rs1060105	12	123,806,219	C/T	0.815	1.04	1.02	1.06	5.7×10^{-7}	1.04	1.02	1.06	1.1×10^{-7}	0.787	1.04	1.02	1.06	3.6×10^{-5}	2
CILP2	*TM6SF2* p.Glu167Lys	rs58542926	19	19,379,549	T/C	0.076	1.07	1.05	1.10	4.8×10^{-12}	1.09	1.06	1.11	3.4×10^{-15}	0.076	1.09	1.05	1.12	2.0×10^{-7}	1
GIPR	*GIPR* p.Glu318Gln	rs1800437	19	46,181,392	C/G	0.200	1.03	1.02	1.05	7.1×10^{-5}	1.06	1.04	1.07	6.8×10^{-12}	0.213	1.09	1.06	1.12	4.6×10^{-9}	1
HNF4A	*HNF4A* p.Thr139Ile	rs1800961	20	43,042,364	T/C	0.032	1.09	1.05	1.13	2.6×10^{-8}	1.10	1.06	1.14	5.0×10^{-8}	0.037	1.17	1.12	1.22	1.4×10^{-12}	1
MTMR3ASCC2	*ASCC2* p.Asp407His	rs28265	22	30,200,761	C/G	0.925	1.09	1.06	1.11	2.1×10^{-12}	1.09	1.07	1.12	4.4×10^{-14}	0.916	1.1	1.07	1.14	9.6×10^{-11}	3
New T2D-associated loci																				
FAM63A	*FAM63A* p.Tyr95Asn	rs140386498	1	150,972,959	A/T	0.988	1.21	1.14	1.28	7.5×10^{-8}	1.19	1.12	1.26	6.7×10^{-7}	0.986	1.15	1.06	1.25	0.00047	3
CEP68	*CEP68* p.Gly74Ser	rs7572857	2	65,296,798	G/A	0.846	1.05	1.04	1.07	8.3×10^{-9}	1.05	1.03	1.07	6.6×10^{-7}	0.830	1.06	1.03	1.08	6.6×10^{-7}	2
KIF9	*KIF9* p.Arg638Trp	rs2276853	3	47,282,303	A/G	0.588	1.02	1.01	1.04	8.0×10^{-5}	1.03	1.02	1.05	5.3×10^{-8}	0.602	1.04	1.02	1.05	2.6×10^{-5}	3
ANKH	*ANKH* p.Arg187Gln	rs146886108	5	14,751,305	C/T	0.996	1.29	1.16	1.45	1.4×10^{-7}	1.27	1.13	1.41	3.5×10^{-7}	0.995	1.51	1.29	1.77	3.5×10^{-7}	1
POC5	*POC5* p.His36Arg	rs2307111	5	75,003,678	T/C	0.562	1.05	1.04	1.07	1.6×10^{-15}	1.03	1.01	1.04	2.1×10^{-5}	0.606	1.06	1.05	1.08	1.1×10^{-12}	1
LPL	*LPL* p.Ser474*	rs328	8	19,819,724	C/G	0.903	1.05	1.03	1.08	6.8×10^{-9}	1.05	1.03	1.07	2.3×10^{-7}	0.901	1.08	1.05	1.11	7.1×10^{-8}	1
PLCB3[†]	*PLCB3* p.Ser778Leu	rs35169799	11	64,031,241	T/C	0.071	1.05	1.02	1.08	1.3×10^{-5}	1.06	1.03	1.09	1.8×10^{-7}	0.065	1.07	1.04	1.11	3.8×10^{-5}	1
TPCN2	*TPCN2* p.Val219Ile	rs72928978	11	68,831,364	G/A	0.890	1.05	1.02	1.07	5.2×10^{-7}	1.05	1.03	1.07	1.8×10^{-8}	0.847	1.03	1.00	1.05	0.042	2
WSCD2	*WSCD2* p.Thr113Ile	rs3764002	12	108,618,630	C/T	0.719	1.03	1.02	1.05	3.3×10^{-8}	1.03	1.02	1.05	1.2×10^{-7}	0.736	1.05	1.03	1.07	8.1×10^{-7}	1

Continued

Table 1 | Summary of discovery and fine-mapping analyses of the 40 index coding variants associated with T2D ($P < 2.2 \times 10^{-7}$) (Continued)

Discovery meta-analysis using exome array component: 81,412 T2D cases and 370,832 controls from diverse ancestries														Fine-mapping meta-analysis using GWAS: 50,160 T2D cases and 465,272 controls of European ancestry					
Locus	Index variant	rsID	Chr.	Pos. (bp)	Alleles (R/O)	RAF	BMI unadjusted				BMI adjusted				RAF	OR	L95	U95	P	Group
							OR	L95	U95	P	OR	L95	U95	P						
ZZEF1	*ZZEF1* p.Ile402Val	rs781831	17	3,947,644	C/T	0.422	1.04	1.03	1.05	8.3×10^{-11}	1.03	1.02	1.05	1.8×10^{-7}	0.407	1.04	1.02	1.05	2.1×10^{-5}	2
MLX	*MLX* p.Gln139Arg	rs665268	17	40,722,029	G/A	0.294	1.04	1.02	1.05	2.0×10^{-8}	1.03	1.02	1.04	1.1×10^{-5}	0.280	1.04	1.02	1.06	5.2×10^{-6}	2
TTLL6	*TTLL6* p.Glu712Asp	rs2032844	17	46,847,364	C/A	0.754	1.04	1.02	1.06	1.2×10^{-7}	1.03	1.01	1.04	0.00098	0.750	1.04	1.02	1.06	9.5×10^{-5}	3
C17orf58[†]	*C17orf58* p.Ile92Val	rs9891146	17	65,988,049	T/C	0.277	1.04	1.02	1.06	1.3×10^{-7}	1.02	1.00	1.04	0.00058	0.269	1.05	1.03	1.07	1.7×10^{-7}	2
ZHX3[†]	*ZHX3* p.Asn310Ser	rs17265513	20	39,832,628	C/T	0.211	1.05	1.03	1.07	9.2×10^{-8}	1.04	1.02	1.05	2.9×10^{-6}	0.208	1.02	1.00	1.04	0.068	NA
PNPLA3	*PNPLA3* p.Ile148Met	rs738409	22	44,324,727	G/C	0.239	1.04	1.03	1.05	2.1×10^{-10}	1.05	1.03	1.06	2.8×10^{-11}	0.230	1.05	1.03	1.07	5.8×10^{-6}	1
PIM3	*PIM3* p.Val300Ala	rs4077129	22	50,356,693	T/C	0.276	1.04	1.02	1.05	1.9×10^{-7}	1.04	1.02	1.06	3.5×10^{-8}	0.280	1.04	1.02	1.06	8.7×10^{-5}	3

Chr., chromosome; Pos., position build 37; RAF, risk allele frequency; R, risk allele; O, other allele; BMI, body mass index; OR, odds ratio; L95, lower 95% confidence interval; U95, upper 95% confidence interval; GWAS, genome-wide association studies. Fine-mapping group 1, signal is driven by coding variants; group 2, signal attributable to noncoding variants; group 3, consistent with a partial role for coding variants. *P* values are based on the meta-analyses of discovery-stage and fine-mapping studies as appropriate.[a]Summary statistics from European-ancestry-specific meta-analyses of 48,286 cases and 250,671 controls.

the annotation-informed model. At these 37 association signals, we distinguished three broad patterns of causal relationship between coding variants and T2D risk.

Group 1: T2D association signal driven by coding variants. At 16 of the 37 distinct signals, coding variation accounted for >80% of the $_{ai}$PPA (Fig. 1, Table 2, and Supplementary Table 10). This was attributable to a single coding variant at 11 signals and multiple coding variants at 5 signals. Reassuringly, group 1 signals confirmed coding variant causation for several loci (*GCKR*, *PAM*, *SLC30A8*, and *KCNJ11–ABCC8*) at which functional studies have established strong mechanistic links to T2D pathogenesis (Table 2). T2D association signals at the 12 remaining signals (Fig. 1 and Supplementary Table 10) had not previously been shown to be driven by coding variation, but our fine-mapping analyses pointed to causal coding variants with high $_{ai}$PPA values: these included *HNF4A*, *RREB1* (p.Asp1171Asn), *ANKH*, *WSCD2*, *POC5*, *TM6SF2*, *HNF1A* (p.Ala146Val; p.Ile75Leu), *GIPR*, *LPL*, *PLCB3*, and *PNPLA3* (Table 2). At several of these, independent evidence corroborates the causal role of the genes harboring the associated coding variants. For example, rare coding mutations at *HNF1A* and *HNF4A* are causal for monogenic, early-onset forms of diabetes[24], and at *TM6SF2* and *PNPLA3* the associated coding variants are implicated in the development of non-alcoholic fatty liver disease (NAFLD)[25,26].

The use of priors to capture the enrichment of coding variants seems a reasonable model, across the genome. However, at any given locus, strong priors (especially for PTVs) might elevate to apparent causality variants that would have been excluded from a causal role on the basis of genetic fine-mapping alone. Comparison of the annotation-informed and functionally unweighted credible sets for group 1 signals indicated that this scenario was unlikely. For 11 of the 16 (*GCKR*, *PAM*, *KCNJ11–ABCC8*, *HNF4A*, *RREB1* (p.Asp1171Asn), *ANKH*, *POC5*, *TM6SF2*, *HNF1A* (p.Ala146Val), *PLCB3*, and *PNPLA3*), the coding variant had the highest PPA in the fine-mapping analysis (Table 2), even under the functionally unweighted model. At *SLC30A8*, *WSCD2*, and *GIPR*, the coding variants had similar PPAs to the lead noncoding SNPs under the functionally unweighted prior (Table 2). At these 14 signals, therefore, coding variants have either greater or equivalent PPA to the best flanking noncoding SNPs under the functionally unweighted model, but receive a boost in PPA after incorporating the annotation weights.

The situation is less clear at *LPL*. Here fine-mapping resolution is poor under the functionally unweighted prior and the coding variant sits on an extended haplotype in strong LD with noncoding variants, some with higher PPA, such as rs74855321 (PPA = 0.048) (compared to *LPL* p.Ser474* (rs328, PPA = 0.023)). However, *LPL* p.Ser474* is annotated as a PTV and benefits from a substantially increased prior that boosts its annotation-informed ranking (Table 2). Ultimately, decisions regarding the causal role of any such variant must rest on the amalgamation of evidence from diverse sources, including detailed functional evaluation of the coding variants and of other variants with which they are in LD.

Group 2: T2D association signals not attributable to coding variants. At 13 of the 37 distinct signals, coding variation accounted for <20% of the PPA, even after applying the annotation-informed prior model. These signals are likely to be driven by local noncoding variation and mediated through regulatory mechanisms. Five of these signals (*TPCN2*, *MLX*, *ZZEF1*, *C17orf58*, and *CEP68*) represent new T2D association signals identified in the exome-focused analysis. Given the exome array discoveries, it would have been natural to consider the named genes at these and other loci in this group as candidates for mediation of their respective association signals. However, the fine-mapping analyses indicate that these coding variants do not provide useful mechanistic inference given low $_{ai}$PPA (Fig. 1 and Table 2).

The coding variant association at the *CENTD2* (*ARAP1*) locus is a case in point. The association with the p.Gln802Glu variant in *ARAP1* (rs56200889, $P_{TE} = 4.8 \times 10^{-8}$ but $_{ai}$PPA < 0.001) is seen in the fine-mapping analysis to be secondary to a substantially stronger noncoding association signal involving a cluster of variants including rs11603334 ($P_{TE} = 9.5 \times 10^{-18}$, $_{ai}$PPA = 0.0692) and rs1552224 ($P_{TE} = 2.5 \times 10^{-17}$, $_{ai}$PPA = 0.0941). The identity of the effector transcript at this locus has been the subject of detailed investigation, and some early studies used islet expression data to promote *ARAP1*[27]. However, a more recent study integrating human islet genomics and mouse gene knockout data has established *STARD10* as the gene mediating the GWAS signal, consistent with the reassignment of the *ARAP1* coding variant association as irrelevant to causal inference[28].

While at these loci the coding variant associations represent false leads, this does not necessarily exclude the genes concerned from a causal role. At *WFS1*, for example, coding variants too rare to be visible to the array-based analyses we performed, and statistically

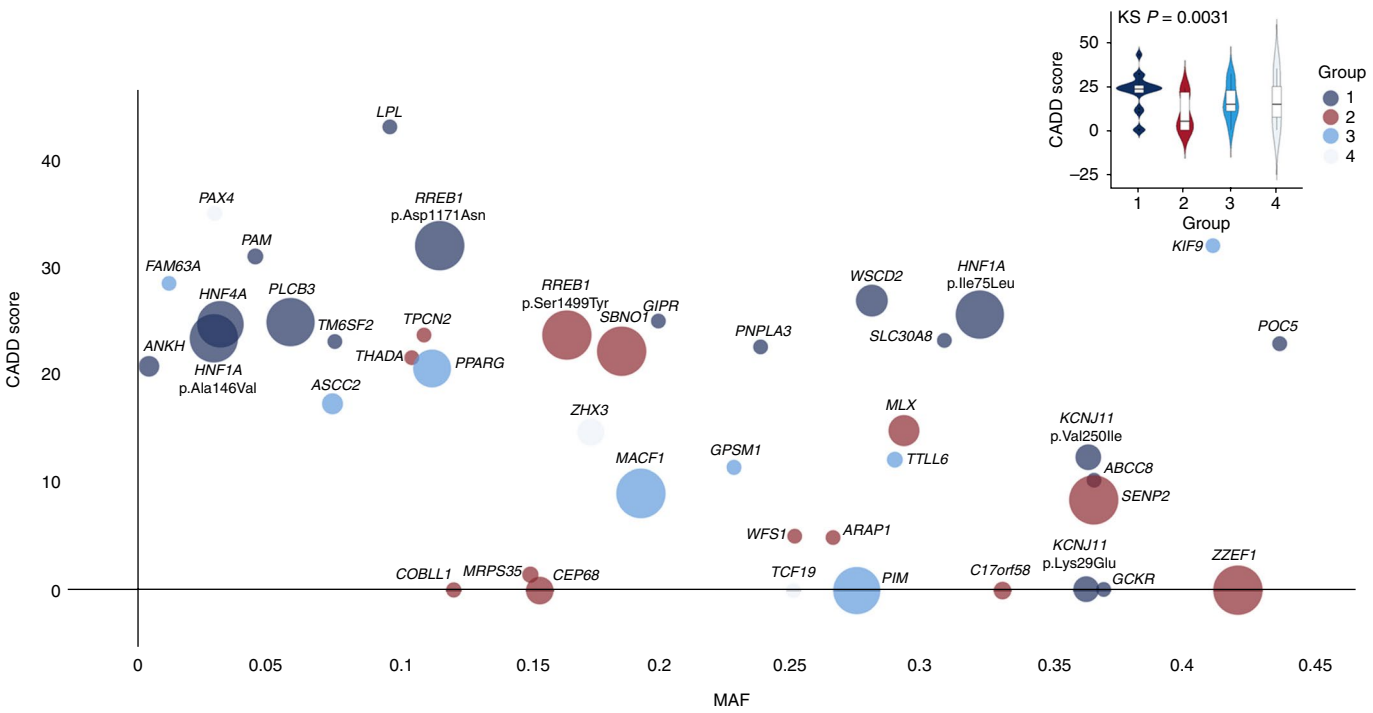

Fig. 2 | Plot of measures of variant-specific and gene-specific features of distinct coding signals to access the functional impact of coding alleles. Each point represents a coding variant with the MAF plotted on the *x* axis and the CADD score plotted on the *y* axis. The size of each point varies with the measure of intolerance of the gene to LoF variants (pLI), and the color represents the fine-mapping group to which each variant is assigned. In group 1, signal is driven by the coding variant; in group 2, signal is attributable to noncoding variants; in group 3, signal is consistent with a partial role for coding variants. Group 4 represents an unclassified category and includes *PAX4*, *ZHX3*, and signal at *TCF19* within the MHC region where we did not perform fine-mapping. Inset, plot showing the distribution of CADD scores between different groups. The plot is a combination of violin plots and box plots; the width of each violin corresponds to the frequency at the corresponding CADD score, and box plots show the median and the 25% and 75% quantiles. The *P* value indicates significance from a two-sample Kolmogorov–Smirnov test.

independent of the common p.Val333Ile variant we detected, cause an early-onset form of diabetes that renders *WFS1* the strongest local candidate for T2D predisposition.

Group 3: Fine-mapping data consistent with a partial role for coding variants. At 8 of the 37 distinct signals, the $_{ai}$PPA attributable to coding variation lay between 20% and 80%. At these signals, the evidence is consistent with 'partial' contributions from coding variants, although the precise inference is likely to be locus specific, dependent on subtle variations in LD, imputation accuracy, and the extent to which global priors accurately represent the functional impact of the specific variants concerned.

This group includes *PPARG*, for which independent evidence corroborates the causal role of this specific effector transcript with respect to T2D risk. *PPARG* encodes the target of antidiabetic thiazolidinedione drugs and harbors very rare coding variants causal for lipodystrophy and insulin resistance, conditions highly relevant to T2D. The common variant association signal at this locus has generally been attributed to the p.Pro12Ala coding variant (rs1801282), although empirical evidence that this variant influences PPARγ function is scant[29–31]. In the functionally unweighted analysis, p.Pro12Ala had an unimpressive PPA (0.0238); after including annotation-informed priors, the same variant emerged with the highest $_{ai}$PPA (0.410), although the 99% credible set included 19 noncoding variants, spanning 67 kb (Supplementary Table 10). These credible set variants included rs4684847 ($_{ai}$PPA = 0.0089), at which the T2D-associated allele has been reported to impact *PPARG* expression and insulin sensitivity by altering binding of the homeobox transcription factor PRRX1[32]. These data are consistent with a model whereby regulatory variants contribute to altered PPARγ activity in combination with, or potentially to the exclusion of, p.Pro12Ala. Future improvements in functional annotation for regulatory variants (gathered from relevant tissues and cell types) should provide increasingly granular priors that allow fine-tuned assignment of causality at loci such as this.

Functional impact of coding alleles. In other contexts, the functional impact of coding alleles is correlated with (i) variant-specific features, including measures of conservation and predicted impact on protein structure, and (ii) gene-specific features, such as extreme selective constraints as quantified by the intolerance to functional variation[33]. To determine whether similar measures could capture information pertinent to T2D causation, we compared coding variants falling into the different fine-mapping groups for a variety of measures, including MAF, Combined Annotation-Dependent Depletion (CADD) score[34], and the loss-of-function (LoF) intolerance metric pLI[33] (Fig. 2 and Methods). Variants from group 1 had significantly higher CADD scores than those in group 2 (Kolmogorov–Smirnov $P = 0.0031$). Except for the variants at *KCNJ11–ABCC8* and *GCKR*, all group 1 coding variants considered likely to be driving T2D association signals had a CADD score ≥20. On this basis, we predict that the East Asian–specific coding variant at *PAX4*, for which the fine-mapping data were not informative, is also likely causal for T2D.

T2D loci and physiological classification. The development of T2D involves dysfunction of multiple mechanisms. Systematic analysis of the physiological effects of known T2D risk alleles has improved understanding of the mechanisms through which these alleles exert their primary impact on disease risk[35]. We obtained association summary statistics for diverse metabolic traits

Table 2 | Posterior probabilities for coding variants within 99% credible sets across loci with annotation-informed and functionally unweighted priors based on fine-mapping analysis performed using 50,160 T2D cases and 465,272 controls of European ancestry

Locus	Variant	rsID	Chr.	Pos. (bp)	Posterior probability		Cumulative posterior probability attributed to coding variants	
					PPA	$_{ai}$PPA	PPA	$_{ai}$PPA
MACF1	*MACF1* p.Ile39Val	rs16826069	1	39,797,055	0.012	0.240	0.032	0.628
	***MACF1* p.Met1424Val**	**rs2296172**	**1**	**39,835,817**	0.011	0.224		
	MACF1 p.Lys1625Asn	rs41270807	1	39,801,815	0.008	0.163		
FAM63A	*FAM63A* p.Tyr95Asn	rs140386498	1	150,972,959	0.005	0.129	0.012	0.303
GCKR	*GCKR* p. Pro 446Leu	rs1260326	2	27,730,940	0.773	0.995	0.773	0.995
THADA	***THADA* p.Cys845Tyr**	**rs35720761**	**2**	**43,519,977**	**<0.001**	**0.011**	0.003	0.120
	THADA p.Thr897Ala	rs7578597	2	43,732,823	0.003	0.107		
CEP68	*CEP68* p.Gly74Ser	rs7572857	2	65,296,798	<0.001	0.004	<0.001	0.004
GRB14	*COBLL1* p.Asn901Asp	rs7607980	2	165,551,201	0.006	0.160	0.006	0.160
PPARG	*PPARG* p.Pro12Ala	rs1801282	3	12,393,125	0.023	0.410	0.024	0.410
KIF9	*SETD2* p.Pro1962Lys	rs4082155	3	47,125,385	0.008	0.171	0.018	0.384
	NBEAL2 p.Arg511Gly	rs11720139	3	47,036,756	0.005	0.097		
	***KIF9* p.Arg638Trp**	**rs2276853**	**3**	**47,282,303**	0.003	0.059		
IGF2BP2	*SENP2* p.Thr291Lys	rs6762208	3	185,331,165	<0.001	<0.001	<0.001	<0.001
WFS1	*WFS1* p.Val333Ile	rs1801212	4	6,302,519	<0.001	0.001	<0.001	0.004
ANKH	*ANKH* p.Arg187Gln	rs146886108	5	14,751,305	0.459	0.972	0.447	0.972
POC5	*POC5* p.His36Arg	rs2307111	5	75,003,678	0.697	0.954	0.702	0.986
PAM-PPIP5K2	*PAM* p.Asp336Gly	rs35658696	5	102,338,811	0.288	0.885	0.309	0.947
	PPIP5K2 p.Ser1207Gly	rs36046591	5	102,537,285	0.020	0.063		
RREB1 p.Asp1171Asn	*RREB1* p.Asp1171Asn	rs9379084	6	7,231,843	0.920	0.997	0.920	0.997
RREB1 p.Ser1499Tyr	*RREB1* p.Ser1499Tyr	rs35742417	6	7,247,344	<0.001	0.013	0.005	0.111
LPL	*LPL* p.Ser474*	rs328	8	19,819,724	0.023	0.832	0.023	0.832
SLC30A8	*SLC30A8* p.Arg276Trp	rs13266634	8	118,184,783	0.295	0.823	0.295	0.823
GPSM1	*GPSM1* p.Ser391Leu	rs60980157	9	139,235,415	0.031	0.557	0.031	0.557
KCNJ11–ABCC8	*KCNJ11* p.Val250Ile	rs5215	11	17,408,630	0.208	0.412	0.481	0.951
	***KCNJ11* p.Lys29Glu**	**rs5219**	**11**	**17,409,572**	0.190	0.376		
	ABCC8 p.Ala1369Ser	rs757110	11	17,418,477	0.083	0.163		
PLCB3	*PLCB3* p.Ser778Leu	rs35169799	11	64,031,241	0.113	0.720	0.130	0.830
TPCN2	*TPCN2* p.Val219Ile	rs72928978	11	68,831,364	<0.001	0.004	0.006	0.140
CENTD2	*ARAP1* p.Gln802Glu	rs56200889	11	72,408,055	<0.001	<0.001	<0.001	<0.001
KLHDC5	*MRPS35* p.Gly43Arg	rs1127787	12	27,867,727	<0.001	<0.001	<0.001	<0.001
WSCD2	*WSCD2* p.Thr113Ile	rs3764002	12	108,618,630	0.281	0.955	0.282	0.958
HNF1A p.Ile75Leu	*HNF1A* p.Gly226Ala	rs56348580	12	121,432,117	0.358	0.894	0.358	0.894
	***HNF1A* p.Ile75Leu**	**rs1169288**	**12**	**121,416,650**	<0.001	<0.001		
HNF1A p.Ala146Val	*HNF1A* p.Ala146Val	rs1800574	12	121,416,864	0.269	0.867	0.280	0.902
MPHOSPH9	*SBNO1* p.Ser729Asn	rs1060105	12	123,806,219	0.002	0.054	0.002	0.057
ZZEF1	*ZZEF1* p.Ile402Val	rs781831	17	3,947,644	<0.001	0.001	<0.001	0.018
MLX	*MLX* p.Gln139Arg	rs665268	17	40,722,029	0.002	0.038	0.002	0.039
TTLL6	***TTLL6* p.Glu712Asp**	**rs2032844**	**17**	**46,847,364**	<0.001	<0.001	0.016	0.305
	CALCOCO2 p.Pro347Ala	rs10278	17	46,939,658	0.0100	0.187		
	SNF8 p.Arg155His	rs57901004	17	47,011,897	0.005	0.092		
C17orf58	*C17orf58* p.Ile92Val	rs9891146	17	65,988,049	<0.001	0.009	<0.001	0.009
CILP2	***TM6SF2* p.Glu167Lys**	**rs58542926**	**19**	**19,379,549**	0.211	0.732	0.263	0.913
	***TM6SF2* p.Leu156Pro**	rs187429064	19	19,380,513	0.049	0.172		
GIPR	*GIPR* p.Glu318Gln	rs1800437	19	46,181,392	0.169	0.901	0.169	0.901

Continued

Table 2 | Posterior probabilities for coding variants within 99% credible sets across loci with annotation-informed and functionally unweighted priors based on fine-mapping analysis performed using 50,160 T2D cases and 465,272 controls of European ancestry (Continued)

Locus	Variant	rsID	Chr.	Pos. (bp)	Posterior probability		Cumulative posterior probability attributed to coding variants	
					PPA	$_{ai}$PPA	PPA	$_{ai}$PPA
ZHX3	*ZHX3* p.Asn310Ser	rs17265513	20	39,832,628	<0.001	0.003	0.003	0.110
HNF4A	*HNF4A* p.Thr139Ile	rs1800961	20	43,042,364	1.000	1.000	1.00	1.000
MTMR3–ASCC2	***ASCC2* p.Asp407His**	**rs28265**	**22**	**30,200,761**	0.011	0.192	0.028	0.481
	ASCC2 p.Pro423Ser	rs36571	22	30,200,713	0.007	0.116		
	ASCC2 p.Val123Ile	rs11549795	22	30,221,120	0.006	0.107		
	MTMR3 p.Asn960Ser	rs41278853	22	30,416,527	0.004	0.065		
PNPLA3	*PNPLA3* p.Ile148Met	rs738409	22	44,324,727	0.112	0.691	0.130	0.806
	PARVB p.Trp37Arg	rs1007863	22	44,395,451	0.017	0.103		
PIM3	*IL17REL* p.Leu333Pro	rs5771069	22	50,435,480	0.041	0.419	0.047	0.475
	IL17REL p.Gly70Arg	rs9617090	22	50,439,194	0.005	0.054		
	***PIM3* p.Val300Ala**	**rs4077129**	22	**50,356,693**	<0.001	0.002		

Chr., chromosome; Pos., position build 37; PPA, functionally unweighted prior; $_{ai}$PPA, annotation-informed prior. Index coding variants are highlighted in bold.

(and other outcomes) for 94 T2D-associated index variants. These 94 variants were restricted to sites represented on the exome array and included the 40 coding signals plus 54 distinct noncoding signals (12 new and 42 previously reported GWAS lead SNPs). We applied clustering techniques (Methods) to generate multi-trait association patterns, allocating 71 of the 94 loci to one of three main physiological categories (Supplementary Fig. 6 and Supplementary Table 11). The first category, comprising nine T2D risk loci with strong BMI and dyslipidemia associations, included three of the new coding signals: *PNPLA3*, *POC5*, and *BPTF*. The T2D associations at both *POC5* and *BPTF* were substantially attenuated (>2-fold decrease in $-\log_{10} P$) after adjusting for BMI (Table 1, Supplementary Fig. 7, and Supplementary Table 3), indicating that their impact on T2D risk is likely mediated by a primary effect on adiposity. *PNPLA3* and *POC5* are established loci for NAFLD[25] and BMI[6], respectively. The second category featured 39 loci at which multi-trait profiles indicated a primary effect on insulin secretion. This set included four of the new coding variant signals (*ANKH*, *ZZEF1*, *TTLL6*, and *ZHX3*). The third category encompassed 23 loci with primary effects on insulin action, including signals at the *KIF9*, *PLCB3*, *CEP68*, *TPCN2*, *FAM63A*, and *PIM3* loci. For most variants in this category, the T2D risk allele was associated with lower BMI, and T2D association signals were more pronounced after adjustment for BMI. At a subset of these loci, including *KIF9* and *PLCB3*, the T2D risk alleles were associated with higher waist–hip ratio and lower body fat percentage, indicating that the mechanism of action likely reflects limitations in storage capacity of peripheral adipose tissue[36].

Discussion

The present study adds to mounting evidence constraining the contribution of lower-frequency variants to T2D risk. Although the exome array interrogates only a subset of the universe of coding variants, it captures the majority of low-frequency coding variants in European populations. The substantial increase in sample size in the present study over our previous effort[12] (effective sample sizes of 228,825 and 82,758, respectively) provides more robust evaluation of the effect size distribution in this low-frequency variant range and indicates that previous analyses are likely, if anything, to have overestimated the contribution of low-frequency variants to T2D risk.

The present study is less informative regarding rare variants. These are sparsely captured on the exome array. In addition, the combination of greater regional diversity in rare allele distribution and the enormous sample sizes necessary to detect rare variant associations (likely to require meta-analysis of data from diverse populations) acts against their identification. Our complementary genome and exome sequence analyses have thus far failed to register strong evidence for a substantial rare variant component to T2D risk[12]. It is therefore highly unlikely that rare variants missed in our analyses are causal for any of the common or low-frequency variant associations we have detected and fine-mapped. On the other hand, it is probable that rare coding alleles, with associations that are distinct from the common variant signals we have examined and detected only through sequence-based analyses, will provide additional clues to the most likely effector transcripts at some of these signals (*WFS1* provides one such example).

Once a coding variant association is detected, it is natural to assume a causal connection between that variant, the gene in which it sits, and the phenotype of interest. While such assignments may be robust for many rare protein-truncating alleles, we demonstrate that this implicit assumption is often inaccurate, particularly for associations attributable to common, missense variants. One-third of the coding variant associations we detected were, when assessed in the context of regional LD, highly unlikely to be causal. At these loci, the genes within which they reside are consequently deprived of their implied connection to disease risk and attention is redirected toward nearby noncoding variants and their impact on regional gene expression. As a group, coding variants we assign as causal are predicted to have a more deleterious impact on gene function than those that we exonerate, but, as in other settings, coding annotation methods lack both sensitivity and specificity. It is worth emphasizing that empirical evidence that the associated coding allele is 'functional' (that is, can be shown to influence cognate gene function in some experimental assay) provides limited reassurance that the coding variant is responsible for the T2D association, unless that specific perturbation of gene function can itself be plausibly linked to the disease phenotype.

Our fine-mapping analyses make use of the observation that coding variants are globally enriched across GWAS signals[8,9,16], with greater prior probability of causality assigned to those with more severe impact on biological function. We assigned diminished priors to noncoding variants, with the lowest support for those mapping outside of DNase I–hypersensitive sites. The extent to which

our findings corroborate previous assignments of causality (often substantiated by detailed, disease-appropriate functional assessment and other orthogonal evidence) suggests that even these sparse annotations provide valuable information to guide target validation. Nevertheless, there are inevitable limits to the extrapolation of these 'broad-brush' genome-wide enrichments to individual loci: improvements in functional annotation for both coding and regulatory variants, particularly when gathered from trait-relevant tissues and cell types, should provide more granular, trait-specific priors to fine-tune assignment of causality within associated regions. These will motivate target validation efforts that benefit from the synthesis of both coding and regulatory mechanisms of gene perturbation. It also needs to be acknowledged that, without whole-genome sequencing data on sample sizes comparable to those we have examined here, imperfections arising from the imputation may confound fine-mapping precision at some loci and that robust inference will inevitably depend on integration of diverse sources of genetic, genomic, and functional data.

The term 'smoking gun' has often been used to describe the potential of functional coding variants to provide causal inference with respect to pathogenetic mechanisms[37]. This study provides a timely reminder that, even when a suspect with a smoking gun is found at the scene of a crime, it should not be assumed that they fired the fatal bullet.

URLs. Type 2 Diabetes Knowledge Portal, http://www.type2diabetesgenetics.org/.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0084-1.

Received: 19 May 2017; Accepted: 30 January 2018;
Published online: 09 April 2018

References

1. Kooner, J. S. et al. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43**, 984–989 (2011).
2. Cho, Y. S. et al. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.* **44**, 67–72 (2011).
3. 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).
4. Mahajan, A. et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
5. Ng, M. C. et al. Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS. Genet.* **10**, e1004517 (2014).
6. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
7. Shungin, D. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
8. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
9. Walter, K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
10. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
11. Horikoshi, M. et al. Transancestral fine-mapping of four type 2 diabetes susceptibility loci highlights potential causal regulatory mechanisms. *Hum. Mol. Genet.* **25**, 2070–2081 (2016).
12. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
13. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS. Med.* **12**, e1001779 (2015).
14. Cook, J. P. & Morris, A. P. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur. J. Hum. Genet.* **24**, 1175–1180 (2016).
15. Estrada, K. et al. Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a Latino population. *J. Am. Med. Assoc.* **311**, 2305–2314 (2014).
16. Sveinbjornsson, G. et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).
17. Liu, D. J. et al. Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
18. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
19. 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).
20. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
21. Maller, J. B. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
22. Flannick, J. et al. Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).
23. Beer, N. L. et al. The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum. Mol. Genet.* **18**, 4081–4088 (2009).
24. Murphy, R., Ellard, S. & Hattersley, A. T. Clinical implications of a molecular genetic classification of monogenic beta-cell diabetes. *Nat. Clin. Pract. Endocrinol. Metab.* **4**, 200–213 (2008).
25. Romeo, S. et al. Genetic variation in *PNPLA3* confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* **40**, 1461–1465 (2008).
26. Kozlitina, J. et al. Exome-wide association study identifies a *TM6SF2* variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* **46**, 352–356 (2014).
27. Kulzer, J. R. et al. A common functional regulatory variant at a type 2 diabetes locus upregulates *ARAP1* expression in the pancreatic beta cell. *Am. J. Hum. Genet.* **94**, 186–197 (2014).
28. Carrat, G. R. et al. Decreased *STARD10* expression is associated with defective insulin secretion in humans and mice. *Am. J. Hum. Genet.* **100**, 238–256 (2017).
29. Deeb, S. S. et al. A Pro12Ala substitution in PPARγ2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nat. Genet.* **20**, 284–287 (1998).
30. Majithia, A. R. et al. Rare variants in *PPARG* with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc Natl Acad Sci USA* **111**, 13127–13132 (2014).
31. Majithia, A. R. et al. Prospective functional classification of all possible missense variants in *PPARG*. *Nat. Genet.* **48**, 1570–1575 (2016).
32. Claussnitzer, M. et al. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell* **156**, 343–358 (2014).
33. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
34. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
35. Dimas, A. S. et al. Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes.* **63**, 2158–2171 (2014).
36. Lotta, L. A. et al. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat. Genet.* **49**, 17–26 (2017).
37. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nat. Genet.* **39**, 813–815 (2007).

Acknowledgements

A full list of acknowledgments appears in the Supplementary Note. Part of this work was conducted using the UK Biobank Resource under application number 9161.

Author contributions

Project coordination: A. Mahajan, A.P.M., J.I.R., M.I.M. **Core analyses and writing:** A. Mahajan, J.W., S.M.W., W. Zhao, N.R.R., A.Y.C., W.G., H.K., R.A.S., I. Barroso, T.M.F., M.O.G., J.B.M., M. Boehnke, D.S., A.P.M., J.I.R., M.I.M. **Statistical analysis in individual studies:** A. Mahajan, J.W., S.M.W., W. Zhao, N.R.R., A.Y.C., W.G., H.K., D.T., N.W.R., X.G., Y. Lu, M. Li, R.A.J., Y. Hu, S. Huo, K.K.L., W. Zhang, J.P.C., B.P.P., J. Flannick, N.G., V.V.T., J. Kravic, Y.J.K., D.V.R., H.Y., M.M.-N., K.M., R.L.-G., T.V.V., J. Marten, J. Li, A.V.S., P. An, S.L., S.G., G.M., A. Demirkan, J.F.T., V. Steinthorsdottir, M.W., C. Lecoeur, M. Preuss, L.F.B., P. Almgren, J.B.-J., J.A.B., M.C., K.-U.E., K.F.,H.G.d.H., Y. Hai, S. Han,

S.J., F. Kronenberg, K.L., L.A.L., J.-J.L., H.L., C.-T.L., J. Liu, R.M., K.R., S.S., P.S., T.M.T., G.T., A. Tin, A.R.W., P.Y., J.Y., L.Y., R.Y., J.C.C., D.I.C., C.v.D., J. Dupuis, P.W.F., A. Köttgen, D.M.-K., N. Soranzo, R.A.S., A.P.M. **Genotyping:** A. Mahajan, N.R.R., A.Y.C., Y. Lu, Y. Hu, S. Huo, B.P.P., N.G., R.L.-G., P. An, G.M., E.A., N.A., C.B., N.P.B., Y.-D.I.C., Y.S.C., M.L.G., H.G.d.H., S. Hackinger, S.J., B.-J.K., P.K., J. Kriebel, F. Kronenberg, H.L., S.S.R., K.D.T., E.B., E.P.B., P.D., J.C.F., S.R.H., C. Langenberg, M.A.P., F.R., A.G.U., J.C.C., D.I.C., P.W.F., B.-G.H., C.H., E.I., S.L.R.K., J.S.K., Y. Liu, R.J.F.L., N. Soranzo, N.J.W., R.A.S., T.M.F., A.P.M., J.I.R., M.I.M. **Cross-trait lookups in unpublished data:** S.M.W., A.Y.C., Y. Lu, M. Li, M.G., H.M.H., A.E.J., D.J.L., E.M., G.M.P., H.R.W., S.K., C.J.W. **Phenotyping:** Y. Lu, Y. Hu, S. Huo, P. An, S.L., A. Demirkan, S. Afaq, S. Afzal, L.B.B., A.G.B., I. Brandslund, C.C., S.V.E., G.G., V. Giedraitis, A.T.-H., M.-F.H., B.I., M.E.J., T.J., A. Käräjämäki, S.S.K., H.A.K., P.K., F. Kronenberg, B.L., H.L., K.-H.L., A.L., J. Liu, M. Loh, V.M., R.M.-C., G.N., M.N., S.F.N., I.N., P.A.P., W.R., L.R., O.R., S.S., E.S., K.S.S., A.S., B.T., A. Tönjes, A.V., D.R.W., H.B., E.P.B., A. Dehghan, J.C.F., S.R.H., C. Langenberg, A.D. Morris, R.d.M., M.A.P., A.R., P.M.R., F.R.R., V. Salomaa, W.H.-H.S., R.V., J.C.C., J. Dupuis, O.H.F., H.G., B.-G.H., T.H., A.T.H., C.H., S.L.R.K., J.S.K., A. Köttgen, L.L., Y. Liu, R.J.F.L., C.N.A.P., J.S.P., O.P., B.M.P., M.B.S., N.J.W., T.M.F., M.O.G. **Individual study design and principal investigators:** N.G., P. An, B.-J.K., P. Amouyel, H.B., E.B., E.P.B., R.C., F.S.C., G.D., A. Dehghan, P.D., M.M.F., J. Ferrières, J.C.F., P. Frossard, V. Gudnason, T.B.H., S.R.H., J.M.M.H., M.I., F. Kee, J. Kuusisto, C. Langenberg, L.J.L., C.M.L., S.M., T.M., O.M., K.L.M., M.M., A.D. Morris, A.D. Murray, R.d.M., M.O.-M., K.R.O., M. Perola, A.P., M.A.P., P.M.R., F.R., F.R.R., A.H.R., V. Salomaa, W.H.-H.S., R.S., B.H.S., K. Strauch, A.G.U., R.V., M. Blüher, A.S.B., J.C.C., D.I.C., J. Danesh, C.v.D., O.H.F., P.W.F., P. Froguel, H.G., L.G., T.H., A.T.H., C.H., E.I., S.L.R.K., F. Karpe, J.S.K., A. Köttgen, K.K., M. Laakso, X.L., L.L., Y. Liu, R.J.F.L., J. Marchini, A. Metspalu, D.M.-K., B.G.N., C.N.A.P., J.S.P., O.P., B.M.P., R.R., N. Sattar, M.B.S., N. Soranzo, T.D.S., K. Stefansson, M.S., U.T., T.T., J.T., N.J.W., J.G.W., E.Z., I. Barroso, T.M.F., J.B.M., M. Boehnke, D.S., A.P.M., J.I.R., M.I.M.

Competing interests

J.C.F. has received consulting honoraria from Merck and from Boehringer-Ingelheim. D.I.C. received funding for exome chip genotyping in the WGHS from Amgen. O.H.F. works in ErasmusAGE, a center for aging research across the life course funded by Nestlé Nutrition (Nestec, Ltd.), Metagenics, Inc., and AXA. Nestlé Nutrition (Nestec, Ltd.), Metagenics, Inc., and AXA had no role in the design or conduct of the study; collection, management, analysis, or interpretation of the data; or preparation, review, or approval of the manuscript. E.I. is an advisor and consultant for Precision Wellness, Inc., and an advisor for Cellink for work unrelated to the present project. B.M.P. serves on the DSMB for a clinical trial funded by the manufacturer (Zoll LifeCor) and on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. I. Barroso and spouse own stock in GlaxoSmithKline and Incyte Corporation. T.F. has consulted for Boeringer–Ingelheim and Sanofi on the genetics of diabetes. D.S. has received support from Pfizer, Regeneron, Genentech, and Eli Lilly. M.I.M. has served on advisory panels for Novo Nordisk and Pfizer and received honoraria from Novo Nordisk, Pfizer, Sanofi-Aventis, and Eli Lilly.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0084-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.M. or J.I.R. or M.I.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Anubha Mahajan[1*], Jennifer Wessel[2], Sara M. Willems[3], Wei Zhao[4], Neil R. Robertson[1,5], Audrey Y. Chu[6,7], Wei Gan[1], Hidetoshi Kitajima[1], Daniel Taliun[8], N. William Rayner[1,5,9], Xiuqing Guo[10], Yingchang Lu[11], Man Li[12,13], Richard A. Jensen[14], Yao Hu[15], Shaofeng Huo[15], Kurt K. Lohman[16], Weihua Zhang[17,18], James P. Cook[19], Bram Peter Prins[9], Jason Flannick[20,21], Niels Grarup[22], Vassily Vladimirovich Trubetskoy[8], Jasmina Kravic[23], Young Jin Kim[24], Denis V. Rybin[25], Hanieh Yaghootkar[26], Martina Müller-Nurasyid[27,28,29], Karina Meidtner[30,31], Ruifang Li-Gao[32], Tibor V. Varga[33], Jonathan Marten[34], Jin Li[35], Albert Vernon Smith[36,37], Ping An[38], Symen Ligthart[39], Stefan Gustafsson[40], Giovanni Malerba[41], Ayse Demirkan[39,42], Juan Fernandez Tajes[1], Valgerdur Steinthorsdottir[43], Matthias Wuttke[44], Cécile Lecoeur[45], Michael Preuss[11], Lawrence F. Bielak[46], Marielisa Graff[47], Heather M. Highland[48], Anne E. Justice[47], Dajiang J. Liu[49], Eirini Marouli[50], Gina Marie Peloso[20,25], Helen R. Warren[50,51], ExomeBP Consortium[52], MAGIC Consortium[52], GIANT Consortium[52], Saima Afaq[17], Shoaib Afzal[53,54,55], Emma Ahlqvist[23], Peter Almgren[56], Najaf Amin[39], Lia B. Bang[57], Alain G. Bertoni[58], Cristina Bombieri[41], Jette Bork-Jensen[22], Ivan Brandslund[59,60], Jennifer A. Brody[14], Noël P. Burtt[20], Mickaël Canouil[45], Yii-Der Ida Chen[10], Yoon Shin Cho[61], Cramer Christensen[62], Sophie V. Eastwood[63], Kai-Uwe Eckardt[64], Krista Fischer[65], Giovanni Gambaro[66], Vilmantas Giedraitis[67], Megan L. Grove[68], Hugoline G. de Haan[32], Sophie Hackinger[9], Yang Hai[10], Sohee Han[24], Anne Tybjærg-Hansen[54,55,69], Marie-France Hivert[70,71,72], Bo Isomaa[73,74], Susanne Jäger[30,31], Marit E. Jørgensen[75,76], Torben Jørgensen[55,77,78], Annemari Käräjämäki[79,80], Bong-Jo Kim[24], Sung Soo Kim[24], Heikki A. Koistinen[81,82,83,84], Peter Kovacs[85], Jennifer Kriebel[31,86], Florian Kronenberg[87], Kristi Läll[65,88], Leslie A. Lange[89], Jung-Jin Lee[4], Benjamin Lehne[17], Huaixing Li[15], Keng-Hung Lin[90], Allan Linneberg[77,91,92], Ching-Ti Liu[25], Jun Liu[39], Marie Loh[17,93,94], Reedik Mägi[65], Vasiliki Mamakou[95], Roberta McKean-Cowdin[96], Girish Nadkarni[97], Matt Neville[5,98], Sune F. Nielsen[53,54,55], Ioanna Ntalla[50], Patricia A. Peyser[46], Wolfgang Rathmann[31,99], Kenneth Rice[100], Stephen S. Rich[101], Line Rode[53,54], Olov Rolandsson[102], Sebastian Schönherr[87], Elizabeth Selvin[12], Kerrin S. Small[103], Alena Stančáková[104], Praveen Surendran[105], Kent D. Taylor[10], Tanya M. Teslovich[8], Barbara Thorand[31,106], Gudmar Thorleifsson[43], Adrienne Tin[107], Anke Tönjes[108],

Anette Varbo[53,54,55,69], Daniel R. Witte[109,110], Andrew R. Wood[26], Pranav Yajnik[8], Jie Yao[10], Loïc Yengo[45], Robin Young[105,111], Philippe Amouyel[112], Heiner Boeing[113], Eric Boerwinkle[68,114], Erwin P. Bottinger[11], Rajiv Chowdhury[115], Francis S. Collins[116], George Dedoussis[117], Abbas Dehghan[39,118], Panos Deloukas[50,119], Marco M. Ferrario[120], Jean Ferrières[121,122], Jose C. Florez[70,123,124,125], Philippe Frossard[126], Vilmundur Gudnason[36,37], Tamara B. Harris[127], Susan R. Heckbert[14], Joanna M. M. Howson[115], Martin Ingelsson[67], Sekar Kathiresan[20,123,125,128], Frank Kee[129], Johanna Kuusisto[104], Claudia Langenberg[3], Lenore J. Launer[127], Cecilia M. Lindgren[1,20,130], Satu Männistö[131], Thomas Meitinger[132,133], Olle Melander[56], Karen L. Mohlke[134], Marie Moitry[135,136], Andrew D. Morris[137,138], Alison D. Murray[139], Renée de Mutsert[32], Marju Orho-Melander[140], Katharine R. Owen[5,98], Markus Perola[131,141], Annette Peters[29,31,106], Michael A. Province[38], Asif Rasheed[126], Paul M. Ridker[7,125], Fernando Rivadineira[39,142], Frits R. Rosendaal[32], Anders H. Rosengren[23], Veikko Salomaa[131], Wayne H.-H. Sheu[143,144,145], Rob Sladek[146,147,148], Blair H. Smith[149], Konstantin Strauch[27,150], André G. Uitterlinden[40,142], Rohit Varma[151], Cristen J. Willer[152,153,154], Matthias Blüher[85,108], Adam S. Butterworth[105,155], John Campbell Chambers[17,18,156], Daniel I. Chasman[7,125], John Danesh[105,155,157,158], Cornelia van Duijn[39], Josée Dupuis[6,25], Oscar H. Franco[39], Paul W. Franks[33,102,159], Philippe Froguel[45,160], Harald Grallert[31,86,161,162], Leif Groop[23,141], Bok-Ghee Han[24], Torben Hansen[22,163], Andrew T. Hattersley[164], Caroline Hayward[34], Erik Ingelsson[35,40], Sharon L. R. Kardia[46], Fredrik Karpe[5,98], Jaspal Singh Kooner[18,156,165], Anna Köttgen[44], Kari Kuulasmaa[132], Markku Laakso[104], Xu Lin[15], Lars Lind[166], Yongmei Liu[58], Ruth J. F. Loos[11,167], Jonathan Marchini[1,168], Andres Metspalu[65], Dennis Mook-Kanamori[32,169], Børge G. Nordestgaard[53,54,55], Colin N. A. Palmer[170], James S. Pankow[171], Oluf Pedersen[22], Bruce M. Psaty[14,172], Rainer Rauramaa[173], Naveed Sattar[174], Matthias B. Schulze[30,31], Nicole Soranzo[9,155,175], Timothy D. Spector[103], Kari Stefansson[37,43], Michael Stumvoll[176], Unnur Thorsteinsdottir[37,43], Tiinamaija Tuomi[74,82,141,177], Jaakko Tuomilehto[81,178,179,180], Nicholas J. Wareham[3], James G. Wilson[181], Eleftheria Zeggini[9], Robert A. Scott[3], Inês Barroso[9,182], Timothy M. Frayling[26], Mark O. Goodarzi[183], James B. Meigs[184], Michael Boehnke[8], Danish Saleheen[4,126,186], Andrew P. Morris[1,19,65,186], Jerome I. Rotter[10,185,186*] and Mark I. McCarthy[1,5,98,186*]

[1]Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [2]Departments of Epidemiology and Medicine, Diabetes Translational Research Center, Indiana University, Indianapolis, IN, USA. [3]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK. [4]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA. [5]Oxford Centre for Diabetes, Endocrinology, and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [6]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, USA. [7]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [8]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [9]Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK. [10]Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, LABioMed at Harbor-UCLA Medical Center, Torrance, CA, USA. [11]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [12]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [13]Division of Nephrology and Hypertension, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA. [14]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, USA. [15]Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, University of the Chinese Academy of Sciences, Shanghai, People's Republic of China. [16]Department of Biostatistical Sciences, Division of Public Health Sciences, Wake Forest University Health Sciences, Winston-Salem, NC, USA. [17]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [18]Department of Cardiology, Ealing Hospital, London North West Healthcare NHS Trust, Middlesex, UK. [19]Department of Biostatistics, University of Liverpool, Liverpool, UK. [20]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA. [21]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA. [22]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [23]Department of Clinical Sciences, Diabetes, and Endocrinology, Lund University Diabetes Centre, Malmö, Sweden. [24]Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, Republic of Korea. [25]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [26]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK. [27]Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [28]Department of Medicine I, University Hospital Großhadern, Ludwig-Maximilians-Universität, Munich, Germany. [29]DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany. [30]Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE),

Nuthetal, Germany. [31]German Center for Diabetes Research (DZD), Neuherberg, Germany. [32]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands. [33]Department of Clinical Sciences, Lund University Diabetes Centre, Genetic and Molecular Epidemiology Unit, Lund University, Malmö, Sweden. [34]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [35]Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [36]Icelandic Heart Association, Kopavogur, Iceland. [37]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [38]Department of Genetics, Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA. [39]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands. [40]Department of Medical Sciences, Molecular Epidemiology, and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [41]Section of Biology and Genetics, Department of Neurosciences, Biomedicine, and Movement Sciences, University of Verona, Verona, Italy. [42]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [43]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. [44]Institute of Genetic Epidemiology, Medical Center–University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. [45]CNRS, UMR 8199, Lille University, Lille Pasteur Institute, Lille, France. [46]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [47]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA. [48]Human Genetics Center, University of Texas Graduate School of Biomedical Sciences at Houston, University of Texas Health Science Center at Houston, Houston, TX, USA. [49]Department of Public Health Sciences, Institute of Personalized Medicine, Penn State College of Medicine, Hershey, PA, USA. [50]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [51]National Institute for Health Research, Barts Cardiovascular Biomedical Research Unit, Queen Mary University of London, London, UK. [52]A full list of members and affiliations appears in the Supplementary Note. [53]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [54]Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, Denmark. [55]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [56]Department of Clinical Sciences, Hypertension, and Cardiovascular Disease, Lund University, Malmö, Sweden. [57]Department of Cardiology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark. [58]Department of Epidemiology and Prevention, Public Health Sciences, Wake Forest University Health Sciences, Winston-Salem, NC, USA. [59]Institute of Regional Health Research, University of Southern Denmark, Odense, Denmark. [60]Department of Clinical Biochemistry, Vejle Hospital, Vejle, Denmark. [61]Department of Biomedical Science, Hallym University, Chuncheon, Republic of Korea. [62]Medical Department, Lillebælt Hospital Vejle, Vejle, Denmark. [63]Institute of Cardiovascular Science, University College London, London, UK. [64]Department of Nephrology and Medical Intensive Care, Charité, University Medicine Berlin, Berlin, Germany. [65]Estonian Genome Center, University of Tartu, Tartu, Estonia. [66]Institute of Internal and Geriatric Medicine, Università Cattolica del Sacro Cuore, Rome, Italy. [67]Department of Public Health and Caring Sciences, Geriatrics, Uppsala University, Uppsala, Sweden. [68]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA. [69]Department of Clinical Biochemistry, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark. [70]Diabetes Research Center (Diabetes Unit), Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [71]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA, USA. [72]Department of Medicine, Université de Sherbrooke, Sherbrooke, Quebec, Canada. [73]Malmska Municipal Health Care Center and Hospital, Jakobstad, Finland. [74]Folkhälsan Research Centre, Helsinki, Finland. [75]Steno Diabetes Center Copenhagen, Gentofte, Denmark. [76]National Institute of Public Health, Southern Denmark University, Copenhagen, Denmark. [77]Research Centre for Prevention and Health, Capital Region of Denmark, Glostrup, Denmark. [78]Faculty of Medicine, Aalborg University, Aalborg, Denmark. [79]Department of Primary Health Care, Vaasa Central Hospital, Vaasa, Finland. [80]Diabetes Center, Vaasa Health Care Center, Vaasa, Finland. [81]Department of Health, National Institute for Health and Welfare, Helsinki, Finland. [82]Endocrinology, Abdominal Center, Helsinki University Hospital, Helsinki, Finland. [83]Minerva Foundation Institute for Medical Research, Helsinki, Finland. [84]Department of Medicine, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland. [85]Integrated Research and Treatment (IFB) Center Adiposity Diseases, University of Leipzig, Leipzig, Germany. [86]Research Unit of Molecular Epidemiology, Institute of Epidemiology II, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [87]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria. [88]Institute of Mathematical Statistics, University of Tartu, Tartu, Estonia. [89]Department of Medicine, Division of Bioinformatics and Personalized Medicine, University of Colorado Denver, Aurora, CO, USA. [90]Department of Ophthalmology, Taichung Veterans General Hospital, Taichung, Taiwan. [91]Department of Clinical Experimental Research, Rigshospitalet, Glostrup, Denmark. [92]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [93]Institute of Health Sciences, University of Oulu, Oulu, Finland. [94]Translational Laboratory in Genetic Medicine (TLGM), Agency for Science, Technology, and Research (A*STAR), Singapore, Singapore. [95]Dromokaiteio Psychiatric Hospital, National and Kapodistrian University of Athens, Athens, Greece. [96]Department of Preventive Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA. [97]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [98]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. [99]Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [100]Department of Biostatistics, University of Washington, Seattle, WA, USA. [101]Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA, USA. [102]Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden. [103]Department of Twins Research and Genetic Epidemiology, King's College London, London, UK. [104]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. [105]MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [106]Institute of Epidemiology II, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [107]Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [108]Department of Medicine, University of Leipzig, Leipzig, Germany. [109]Department of Public Health, Aarhus University, Aarhus, Denmark. [110]Danish Diabetes Academy, Odense, Denmark. [111]Robertson Centre for Biostatistics, University of Glasgow, Glasgow, UK. [112]Institut Pasteur de Lille, INSERM U1167, Université Lille Nord de France, Lille, France. [113]Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany. [114]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [115]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [116]Genome Technology Branch, National Human Genome Research Institute, US National Institutes of Health, Bethesda, MD, USA. [117]Department of Nutrition and Dietetics, Harokopio University of Athens, Athens, Greece. [118]MRC–PHE Centre for Environment and Health, Imperial College London, London, UK. [119]Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdulaziz University, Jeddah, Saudi Arabia. [120]Research Centre on Epidemiology and Preventive Medicine (EPIMED), Department of Medicine and Surgery, University of Insubria, Varese, Italy. [121]INSERM, UMR 1027 Toulouse, France. [122]Department of Cardiology, Toulouse University School of Medicine, Rangueil Hospital, Toulouse, France. [123]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [124]Programs in Metabolism and Medical & Population Genetics, Broad Institute, Cambridge, MA, USA. [125]Department of Medicine, Harvard Medical School, Boston, MA, USA. [126]Center for Non-Communicable Diseases, Karachi, Pakistan. [127]Laboratory of Epidemiology and Population Sciences, National Institute on Aging, US National Institutes of Health, Bethesda, MD, USA. [128]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [129]UKCRC Centre of Excellence for Public Health (NI), Queens University of Belfast, Belfast, UK. [130]Big Data Institute, Li Ka Shing Centre For Health Information and Discovery, University of Oxford, Oxford, UK. [131]National Institute for Health and Welfare, Helsinki, Finland. [132]Institute of Human Genetics, Technische Universität München, Munich, Germany. [133]Institute of Human Genetics,

Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [134]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [135]Department of Epidemiology and Public Health, University of Strasbourg, Strasbourg, France. [136]Department of Public Health, University Hospital of Strasbourg, Strasbourg, France. [137]Clinical Research Centre, Centre for Molecular Medicine, Ninewells Hospital and Medical School, Dundee, UK. [138]Usher Institute to the Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK. [139]Aberdeen Biomedical Imaging Centre, School of Medicine, Medical Sciences, and Nutrition, University of Aberdeen, Aberdeen, UK. [140]Department of Clinical Sciences, Diabetes, and Cardiovascular Disease, Genetic Epidemiology, Lund University, Malmö, Sweden. [141]Finnish Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland. [142]Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands. [143]Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan. [144]National Yang-Ming University, School of Medicine, Taipei, Taiwan. [145]National Defense Medical Center, School of Medicine, Taipei, Taiwan. [146]McGill University and Génome Québec Innovation Centre, Montreal, QC, Canada. [147]Department of Human Genetics, McGill University, Montreal, QC, Canada. [148]Division of Endocrinology and Metabolism, Department of Medicine, McGill University, Montreal, QC, Canada. [149]Division of Population Health Sciences, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [150]Institute of Medical Informatics, Biometry, and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany. [151]USC Roski Eye Institute, Department of Ophthalmology, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA. [152]Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI, USA. [153]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [154]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [155]NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [156]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [157]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [158]British Heart Foundation, Cambridge Centre of Excellence, Department of Medicine, University of Cambridge, Cambridge, UK. [159]Department of Nutrition, Harvard School of Public Health, Boston, MA, USA. [160]Department of Genomics of Common Disease, School of Public Health, Imperial College London, London, UK. [161]Clinical Cooperation Group Type 2 Diabetes, Helmholtz Zentrum München, Ludwig-Maximillians University Munich, Munich, Germany. [162]Clinical Cooperation Group Nutrigenomics and Type 2 Diabetes, Helmholtz Zentrum München, Technical University of Munich, Munich, Germany. [163]Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark. [164]University of Exeter Medical School, University of Exeter, Exeter, UK. [165]National Heart and Lung Institute, Cardiovascular Sciences, Hammersmith Campus, Imperial College London, London, UK. [166]Department of Medical Sciences, Uppsala University, Uppsala, Sweden. [167]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [168]Department of Statistics, University of Oxford, Oxford, UK. [169]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands. [170]Pat Macpherson Centre for Pharmacogenetics and Pharmacogenomics, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [171]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA. [172]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA. [173]Foundation for Research in Health, Exercise, and Nutrition, Kuopio Research Institute of Exercise Medicine, Kuopio, Finland. [174]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK. [175]Department of Hematology, School of Clinical Medicine, University of Cambridge, Cambridge, UK. [176]Divisions of Endocrinology and Nephrology, University Hospital Leipzig, Leipzig, Germany. [177]Research Programs Unit, Diabetes and Obesity, University of Helsinki, Helsinki, Finland. [178]Dasman Diabetes Institute, Dasman, Kuwait. [179]Department of Neuroscience and Preventive Medicine, Danube University Krems, Krems, Austria. [180]Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia. [181]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [182]Metabolic Research Laboratories, Wellcome Trust–MRC Institute of Metabolic Science, University of Cambridge, Cambridge, UK. [183]Division of Endocrinology, Diabetes, and Metabolism, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [184]General Medicine Division, Massachusetts General Hospital and Department of Medicine, Harvard Medical School, Boston, MA, USA. [185]Department of Medicine, Institute for Translational Genomics and Population Sciences, LABioMed at Harbor-UCLA Medical Center, Torrance, CA, USA. [186]These authors jointly directed this work: Danish Saleheen, Andrew P. Morris, Jerome I. Rotter and Mark I. McCarthy *e-mail: anubha@well.ox.ac.uk; jrotter@labiomed.org; mark.mccarthy@drl.ox.ac.uk

Methods

Ethics statement. All human research was approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki. All participants provided written informed consent.

Derivation of significance thresholds. We considered five categories of annotation[16] of variants on the exome array in order of decreasing effect on biological function: (i) PTVs (stop-gain and stop-loss, frameshift indel, donor and acceptor splice-site, and initiator codon variants, $n_1 = 8{,}388$); (ii) moderate-impact variants (missense, in-frame indel, and splice-region variants, $n_2 = 216{,}114$); (iii) low-impact variants (synonymous, 3′ and 5′ UTR, and upstream and downstream variants, $n_3 = 8{,}829$); (iv) other variants mapping to DNase I–hypersensitive sites (DHSs) in any of 217 cell types[8] (DHSs, $n_4 = 3{,}561$); and (v) other variants not mapping to DHSs ($n_5 = 10{,}578$). To account for the greater prior probability of causality for variants with greater effect on biological function, we determined a weighted Bonferroni-corrected significance threshold on the basis of reported enrichment[16], denoted w_i, in each annotation category i: $w_1 = 165$; $w_2 = 33$; $w_3 = 3$; $w_4 = 1.5$; $w_5 = 0.5$. For coding variants (annotation categories 1 and 2)

$$\alpha = \frac{0.05 \sum_{i=1}^{2} n_i w_i}{\left(\sum_{i=1}^{2} n_i\right)\left(\sum_{i=1}^{5} n_i w_i\right)} = 2.21 \times 10^{-7}$$

We note that this threshold is similar to a simple Bonferroni correction for the total number of coding variants on the array, which would yield

$$\alpha = \frac{0.05}{224502} = 2.23 \times 10^{-7}$$

For noncoding variants (annotation categories 3–5), the weighted Bonferroni-corrected significance threshold is

$$\alpha = \frac{0.05 \sum_{i=3}^{5} n_i w_i}{\left(\sum_{i=3}^{5} n_i\right)\left(\sum_{i=1}^{5} n_i w_i\right)} = 9.45 \times 10^{-9}$$

Discovery exome array study-level analyses. Within each study, genotype calling and quality control were undertaken according to protocols developed by the UK Exome Chip Consortium or the CHARGE central calling effort[38] (Supplementary Table 1). Within each study, variants were then excluded for the following reasons: (i) not mapping to autosomes or the X chromosome; (ii) multiallelic and/or insertion–deletion; (iii) monomorphic; (iv) call rate <99%; or (v) exact $P < 10^{-4}$ for deviation from Hardy–Weinberg equilibrium (autosomes only).

We tested association of T2D with each variant in a linear mixed model, implemented in RareMetalWorker[17], using a genetic relationship matrix (GRM) to account for population structure and relatedness. For participants from family-based studies, known relationships were incorporated directly in the GRM. For founders and participants from population-based studies, the GRM was constructed from pairwise identity-by-descent (IBD) estimates based on LD-pruned ($r^2 < 0.05$) autosomal variants with MAF ≥1% (across cases and controls combined), after exclusion of those in high LD and complex regions[39,40] and those mapping to established T2D loci. We considered additive, dominant, and recessive models for the effect of the minor allele, adjusted for age and sex (where appropriate) and additional study-specific covariates (Supplementary Table 2). Analyses were also performed with and without adjustment for BMI (where available; Supplementary Table 2).

For single-variant association analyses, variants with minor allele count ≤10 in cases and controls combined were excluded. Association summary statistics for each analysis were corrected for residual inflation by means of genomic control[41], calculated after excluding variants mapping to established T2D susceptibility loci. For gene-based analyses, we made no variant exclusions on the basis of minor allele count.

Discovery exome sequence analyses. We used summary statistics of T2D association from analyses conducted on 8,321 T2D cases and 8,421 controls across different ancestries, all genotyped using exome sequencing. Details of samples included, sequencing, and quality control are described elsewhere[12,15] (http://www.type2diabetesgenetics.org/). Samples were divided into 15 subgroups according to ancestry and study of origin. Each subgroup was analyzed independently, with subgroup-specific principal components and GRMs. Association tests were performed with both a linear mixed model, as implemented in EMMAX[42], using covariates for sequencing batch, and the Firth test, using covariates for principal components and sequencing batch. Related samples were excluded from the Firth analysis but maintained in the linear mixed model analysis. Variants were then filtered from each subgroup analysis, according to call rate, differential case–control missingness, or deviation from Hardy–Weinberg equilibrium (as computed separately for each subgroup). Association statistics were then combined via a fixed-effects inverse-variance-weighted meta-analysis, both at the level of ancestry and across all samples. P values were taken from the linear mixed model analysis, while effect size estimates were taken from the Firth analysis. Analyses were performed with and without adjustment for BMI. From exome sequence summary statistics, we extracted variants passing quality control and present on the exome array.

Discovery GWAS analyses. The UK Biobank is a large detailed prospective study of more than 500,000 participants aged 40–69 years when recruited in 2006–2010[13]. Prevalent T2D status was defined using self-reported medical history and medication in UK Biobank participants[43]. Participants were genotyped with the UK Biobank Axiom Array or UK BiLEVE Axiom Array, and quality control and population structure analyses were performed centrally at UK Biobank. We defined a subset of samples of 'white European' ancestry ($n = 120{,}286$) as those who both self-identified as white British and were confirmed as ancestrally European descent from the first two axes of genetic variation from principal-components analysis. Imputation was also performed centrally at UK Biobank for the autosomes only, up to a merged reference panel from the 1000 Genomes Project (multi-ethnic, phase 3, October 2014 release)[44] and the UK10K Project[9]. We used SNPTESTv2.5[45] to test for association of T2D with each SNP in a logistic regression framework under an additive model and after adjustment for age, sex, six axes of genetic variation, and genotyping array as covariates. Analyses were performed with and without adjustment for BMI, after removing related individuals.

GERA is a large multi-ethnic population-based cohort, created for investigating the genetic and environmental basis of age-related diseases (dbGaP phs000674.p1). T2D status is based on ICD-9 codes in linked electronic medical health records, with all other participants defined as controls. Participants were previously genotyped using one of four custom arrays, which were designed to maximize coverage of common and low-frequency variants in non-Hispanic white, East Asian, African-American, and Latino ancestry groups[46,47]. Methods for quality control have been described previously[14]. Each of the four genotyping arrays was imputed separately, up to the 1000 Genomes Project reference panel (autosomes, phase 3, October 2014 release; X chromosome, phase 1, March 2012 release) using IMPUTEv2.3[48,49]. We used SNPTESTv2.5[45] to test for association of T2D with each SNP in a logistic regression framework under an additive model and after adjustment for sex and nine axes of genetic variation from principal-components analysis as covariates. BMI was not available for adjustment in GERA.

For UK Biobank and GERA, we extracted variants passing standard imputation quality-control thresholds (IMPUTE info ≥ 0.4)[50] and present on the exome array. Association summary statistics under an additive model were corrected for residual inflation by means of genomic control[41], calculated after excluding variants mapping to established T2D susceptibility loci: GERA ($\lambda = 1.097$ for BMI-unadjusted analysis) and UK Biobank ($\lambda = 1.043$ for BMI-unadjusted analysis, $\lambda = 1.056$ for BMI-adjusted analysis).

Discovery single-variant meta-analysis. We aggregated association summary statistics under an additive model across studies, with and without adjustment for BMI, using METAL[51]: (i) effective sample size weighting of z scores to obtain P values and (ii) inverse-variance weighting of log-odds ratios. For exome array studies, allelic effect sizes and standard errors obtained from the RareMetalWorker linear mixed model were converted to the log-odds scale before meta-analysis to correct for case–control imbalance[52].

The European-specific meta-analyses aggregated association summary statistics from a total of 48,286 cases and 250,671 controls from (i) 33 exome array studies of European ancestry; (ii) exome array sequence from individuals of European ancestry; and (iii) GWAS from UK Biobank. Note that noncoding variants represented on the exome array were not available in exome sequence. The European-specific meta-analyses were corrected for residual inflation by means of genomic control[41], calculated after excluding variants mapping to established T2D susceptibility loci: $\lambda = 1.091$ for BMI-unadjusted analysis and $\lambda = 1.080$ for BMI-adjusted analysis.

The trans-ethnic meta-analyses aggregated association summary statistics from a total of 81,412 cases and 370,832 controls across all studies (51 exome array studies, exome sequence, and GWAS from UK Biobank and GERA), irrespective of ancestry. Note that noncoding variants represented on the exome array were not available in exome sequence. The trans-ethnic meta-analyses were corrected for residual inflation by means of genomic control[41], calculated after excluding variants mapping to established T2D susceptibility loci: $\lambda = 1.073$ for BMI-unadjusted analysis and $\lambda = 1.068$ for BMI-adjusted analysis. Heterogeneity in allelic effect sizes between exome array studies contributing to the trans-ethnic meta-analysis was assessed by Cochran's Q statistic[53].

Discovery detection of distinct association signals. Conditional analyses were undertaken to detect association signals by inclusion of index variants and/or tags for previously reported noncoding GWAS lead SNPs as covariates in the regression model at the study level. Within each exome array study, approximate conditional analyses were undertaken under a linear mixed model using RareMetal[17], which uses score statistics and the variance–covariance matrix from the RareMetalWorker single-variant analysis to estimate the correlation in effect size estimates between variants due to LD. Study-level allelic effect sizes and standard errors obtained

from the approximate conditional analyses were converted to the log-odds scale to correct for case–control imbalance[52]. Within each GWAS, exact conditional analyses were performed under a logistic regression model using SNPTESTv2.5[45]. GWAS variants passing standard imputation quality-control thresholds (IMPUTE info ≥ 0.4)[50] and present on the exome array were extracted for meta-analysis.

Association summary statistics were aggregated across studies, with and without adjustment for BMI, using METAL[51]: (i) effective sample size weighting of z scores to obtain P values and (ii) inverse-variance weighting of log-odds ratios.

We defined new loci as mapping >500 kb from a previously reported lead GWAS SNP. We performed conditional analyses where a new signal mapped close to a known GWAS locus and the lead GWAS SNP at that locus was present (or tagged) on the exome array (Supplementary Table 5).

Discovery non-additive association models. For exome array studies only, we aggregated association summary statistics under recessive and dominant models across studies, with and without adjustment for BMI, using METAL[51]: (i) effective sample size weighting of z scores to obtain P values and (ii) inverse-variance weighting of log-odds ratios. Allelic effect sizes and standard errors obtained from the RareMetalWorker linear mixed model were converted to the log-odds scale before meta-analysis to correct for case–control imbalance[52]. The European-specific meta-analyses aggregated association summary statistics from a total of 41,066 cases and 136,024 controls from 33 exome array studies of European ancestry. The European-specific meta-analyses were corrected for residual inflation by means of genomic control[41], calculated after excluding variants mapping to established T2D susceptibility loci: $\lambda = 1.076$ and $\lambda = 1.083$ for BMI-unadjusted analysis, under the recessive and dominant models, respectively, and $\lambda = 1.081$ and $\lambda = 1.062$ for BMI-adjusted analysis, under the recessive and dominant models, respectively. The trans-ethnic meta-analyses aggregated association summary statistics from a total of 58,425 cases and 188,032 controls across all exome array studies, irrespective of ancestry. The trans-ethnic meta-analyses were corrected for residual inflation by means of genomic control[41], calculated after excluding variants mapping to established T2D susceptibility loci: $\lambda = 1.041$ and $\lambda = 1.071$ for BMI-unadjusted analysis, under the recessive and dominant models, respectively, and $\lambda = 1.031$ and $\lambda = 1.063$ for BMI-adjusted analysis, under the recessive and dominant models, respectively.

Discovery gene-based meta-analyses. For exome array studies only, we aggregated association summary statistics under an additive model across studies, with and without adjustment for BMI, using RareMetal[17]. This approach uses score statistics and the variance–covariance matrix from the RareMetalWorker single-variant analysis to estimate the correlation in effect size estimates between variants due to LD. We performed gene-based analyses using a burden test (assuming all variants had the same direction of effect on T2D susceptibility) and SKAT (allowing variants to have different directions of effect on T2D susceptibility). We used two previously defined filters for annotation and MAF[18] to define group files: (i) a strict filter, including 44,666 variants, and (ii) a broad filter, including all variants from the strict filter and 97,187 additional variants.

We assessed the contribution of each variant to gene-based signals by performing approximate conditional analyses. We repeated RareMetal analyses for the gene, excluding each variant in turn from the group file, and compared the strengths of the association signal.

Fine-mapping of coding variant association signals with T2D susceptibility. We defined a locus as mapping 500 kb upstream and downstream of each index coding variant (Supplementary Table 5), excluding the MHC. Our fine-mapping analyses aggregated association summary statistics from 24 GWAS incorporating 50,160 T2D cases and 465,272 controls of European ancestry from the DIAGRAM Consortium (Supplementary Table 9). Each GWAS was imputed using miniMAC[12] or IMPUTEv2[48,49] up to high-density reference panels: (i) 22 GWAS were imputed up to the Haplotype Reference Consortium[20]; (ii) the UK Biobank GWAS was imputed to a merged reference panel from the 1000 Genomes Project (multi-ethnic, phase 3, October 2014 release)[44] and the UK10K Project[9]; and (iii) the deCODE GWAS was imputed up to the deCODE Icelandic population-specific reference panel based on whole-genome sequence data[19]. Association with T2D susceptibility was tested for each remaining variant using logistic regression, adjusting for age, sex, and study-specific covariates, under an additive genetic model. Analyses were performed with and without adjustment for BMI. For each study, variants with minor allele count <5 (in cases and controls combined) or those with imputation quality r2-hat <0.3 (miniMAC) or proper-info <0.4 (IMPUTE2) were removed. Association summary statistics for each analysis were corrected for residual inflation by means of genomic control[41], calculated after excluding variants mapping to established T2D susceptibility loci.

We aggregated association summary statistics across studies, with and without adjustment for BMI, in a fixed-effects inverse-variance-weighted meta-analysis, using METAL[51]. The BMI-unadjusted meta-analysis was corrected for residual inflation by means of genomic control ($\lambda = 1.012$)[41], calculated after excluding variants mapping to established T2D susceptibility loci. No adjustment was required for BMI-adjusted meta-analysis ($\lambda = 0.994$). From the meta-analysis, variants were extracted that were present on the HRC panel and reported in at least 50% of the total effective sample size.

We included 37 of the 40 identified coding variants in fine-mapping analyses, excluding 3 that were not amenable to fine-mapping in the GWAS datasets: (i) the locus in the MHC region because of the extended and complex structure of LD across the region, which complicates fine-mapping efforts; (ii) the East Asian–specific *PAX4* p.Arg190His (rs2233580) signal, as the variant was not present in European-ancestry GWAS; and (iii) *ZHX3* p.Asn310Ser (rs4077129) because the variant was only weakly associated with T2D in the GWAS datasets used for fine-mapping.

To delineate distinct association signals in four regions, we undertook approximate conditional analyses, implemented in GCTA[54], to adjust for the index coding variants and noncoding lead GWAS SNPs: (i) *RREB1* p.Asp1171Asn (rs9379084), p.Ser1499Tyr (rs35742417), and rs9505118; (ii) *HNF1A* p.Ile75Leu (rs1169288) and p.Ala146Val (rs1800574); (iii) *GIPR* p.Glu318Gln (rs1800437) and rs8108269; and (iv) *HNF4A* p.Thr139Ile (rs1800961) and rs4812831. We made use of summary statistics from the fixed-effects meta-analyses (BMI unadjusted for *RREB1*, *HNF1A*, and *HNF4A*; BMI adjusted for *GIPR*, as this signal was only seen in BMI-adjusted analysis) and genotype data from 5,000 random individuals of European ancestry from the UK Biobank, as reference for LD between genetic variants across the region.

For each association signal, we first calculated an approximate Bayes' factor[55] in favor of association on the basis of allelic effect sizes and standard errors from the meta-analysis. Specifically, for the jth variant

$$\Lambda_j = \sqrt{\frac{V_j}{V_j+\omega}} \exp\left[\frac{\omega\beta_j^2}{2V_j(V_j+\omega)}\right],$$

where β_j and V_j denote the estimated allelic effect (log-OR) and corresponding variance from the meta-analysis. The parameter ω denotes the prior variance in allelic effects, taken here to be 0.04[55].

We then calculated the posterior probability that the jth variant drives the association signal, given by

$$\pi_j = \frac{\rho_j \Lambda_j}{\sum_k \rho_k \Lambda_k}.$$

In this expression, ρ_j denotes the prior probability that the jth variant drives the association signal and the summation in the denominator is over all variants across the locus. We considered two prior models: (i) functionally unweighted, for which $\rho_j = 1$ for all variants, and (ii) annotation informed, for which ρ_j is determined by the functional severity of the variant. For the annotation-informed prior, we considered five categories of variation[16], such that (i) $\rho_j = 165$ for PTVs; (ii) $\rho_j = 33$ for moderate-impact variants; (iii) $\rho_j = 3$ for low-impact variants; (iv) $\rho_j = 1.5$ for other variants mapping to DHSs; and (v) $\rho_j = 0.5$ for all other variants.

For each locus, the 99% credible set[21] under each prior was then constructed by (i) ranking all variants according to their posterior probability of driving the association signal and (ii) including ranked variants until their cumulative posterior probability of driving the association equaled or exceeded 0.99.

Functional impact of coding alleles. We used CADD[34] to obtain scaled CADD scores for each of the 40 significantly associated coding variants. The CADD method objectively integrates a range of different annotation metrics into a single measure (CADD score), providing an estimate of deleteriousness for all known variants and an overall rank for this metric across the genome. We obtained estimates of the intolerance of a gene to harboring LoF variants (pLI) from the ExAC dataset[33]. We used the Kolmogorov–Smirnov test to determine whether fine-mapping groups 1 and 2 had the same statistical distribution for each of these parameters.

T2D loci and physiological classification. To explore the different patterns of association between T2D and other anthropometric, metabolic, and/or endocrine traits and diseases, we performed hierarchical clustering analysis. We obtained association summary statistics for a range of metabolic traits and other outcomes for 94 coding and noncoding variants that were significantly associated with T2D through collaboration or by querying publically available GWAS meta-analysis datasets. The z score (allelic effect/standard error) was aligned to the T2D risk allele. We obtained the distance matrix among the z scores of the loci/traits using the Euclidean measure and performed clustering using the complete agglomeration method. Clustering was visualized by constructing a dendrogram and heat map.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. Summary-level data of the exome array component of this project can be downloaded from the DIAGRAM consortium website at http://diagram-consortium.org/ and the Accelerating Medicines Partnership T2D portal at http://www.type2diabetesgenetics.org/.

References

38. Grove, M. L. et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLoS. ONE.* **8**, e68095 (2013).
39. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 132–135 (2008). author reply 135–139.
40. Weale, M. E. Quality control for genome-wide association studies. *Methods. Mol. Biol.* **628**, 341–372 (2010).
41. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics.* **55**, 997–1004 (1999).
42. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
43. Eastwood, S. V. et al. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. *PLoS. ONE.* **11**, e0162388 (2016).
44. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
45. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
46. Hoffmann, T. J. et al. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics.* **98**, 79–89 (2011).
47. Hoffmann, T. J. et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics.* **98**, 422–430 (2011).
48. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS. Genet.* **5**, e1000529 (2009).
49. 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).
50. Winkler, T. W. et al. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
51. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
52. Cook, J. P., Mahajan, A. & Morris, A. P. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *Eur. J. Hum. Genet.* **25**, 240–245 (2017).
53. Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS. ONE.* **2**, e841 (2007).
54. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012). S1–S3.
55. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).

nature research

Corresponding author(s): Mark McCarthy

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

   Describe how sample size was determined.

   We aimed to bring together the largest possible sample size (N>80,000 T2D cases and >350,00 controls) to study the role of coding variants in T2D. Our sample size is adequate to recover known T2D associated regions, and identify 28 novel T2D associated regions. Also, analytical power calculation showed that our dataset has >97% power to identify variant with 20% allele frequency and 1.05 OR or variant with 1% allele frequency and OR 1.20.

2. Data exclusions

   Describe any data exclusions.

   We used established protocols to conduct rigorous data quality control for each exome-array study: variants were excluded for the following reasons: (i) not mapping to autosomes or X chromosome; (ii) multi-allelic and/or insertion-deletion; (iii) monomorphic; (iv) call rate <99%; or (v) exact p<10-4 for deviation from Hardy-Weinberg equilibrium (autosomes only) (details in Supplementary Tables 1 & 9 and Online methods pages 41-42). We made sure that the allele labels and strand were well aligned between studies. We also visually examined the allele frequencies from the sample and the reference dataset (1000 Genomes Project), and made sure that the allele frequencies are consistent.

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings.

   Experimental replication was not attempted.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   Not applicable.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Not applicable.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The software used has been described in Online Methods section. Softwares are: GenCall, zCall, optiCall, RAREMETALWORKER, RareMETALS, METAL, IMPUTE2, PLINK, SHAPEITv2. In addition, study-specific software, used by each study to perform analyses is listed in Supplementary Tables 1 and 9.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	Not applicable.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	Not applicable.
c. Report whether the cell lines were tested for mycoplasma contamination.	Not applicable.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Not applicable.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No animals were used in the study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

We provide a description of outcome and covariates used for each study in Supplementary Tables 1 and 9. In general, association analyses were conducted with adjustment for age, sex, kinship matrix, and any other study specific covariates. Where available, analysis was also conducting after adjustment for body mass index.