The missing link between genetic association and regulatory function

Noah Connally[1,2,3], Sumaiya Nazeen[1,2,4], Daniel Lee[1,2,3], Huwenbo Shi[3,5], John Stamatoyannopoulos[6,7], Sung Chun[8], Chris Cotsapas[3,9,10]*, Christopher Cassa[1,3]*, Shamil Sunyaev[1,2,3]*

[1] Brigham and Women's Hospital, Division of Genetics, Harvard Medical School, Boston, MA, USA
[2] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[3] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[4] Brigham and Women's Hospital, Department of Neurology, Harvard Medical School, Boston, MA, USA
[5] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[6] Altius Institute for Biomedical Sciences, Seattle, WA, USA
[7] Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA
[8] Division of Pulmonary Medicine, Boston Children's Hospital, Boston, MA, USA
[9] Department of Neurology, Yale Medical School, New Haven, CT, USA
[10] Department of Genetics, Yale Medical School, New Haven, CT, USA
* Co-corresponding authors

The genetic basis of most complex traits is highly polygenic and dominated by non-coding alleles. It is widely assumed that such alleles exert small regulatory effects on the expression of *cis*-linked genes. However, despite the availability of expansive gene expression and epigenomic data sets, few variant-to-gene links have emerged. We identified 134 gene-trait pairs in which protein-coding variants cause severe or familial forms of nine human traits. For most of these genes, we find that adjacent non-coding variation is associated with common complex forms of the same traits. However, we found limited evidence of colocalization—the same variant influencing both the physiological trait and gene expression—for only 7% of genes, and transcriptome-wide association evidence with correct direction of effect for only 6% of genes, despite the presence of eQTLs in most loci. Fine-mapping variants to regulatory elements and assigning these to genes by linear distance similarly failed to implicate most genes in complex traits. These results contradict the hypothesis that most complex trait-associated variants coincide with currently ascertained expression quantitative trait loci. The field must confront this deficit, and pursue this "missing regulation."

Modern complex trait genetics has uncovered surprises at every turn, including the paucity of associations between traits and coding variants of large effect, and the "mystery of missing heritability," in which no combination of common and rare variants can explain a large fraction of trait heritability[1]. Further work has revealed unexpectedly high polygenicity for most human traits and very small effect sizes for individual variants. Bulk enrichment analyses have demonstrated that a large fraction of heritability resides in regions with gene regulatory potential, predominantly tissue-specific accessible chromatin and enhancer elements, suggesting that trait-associated variants influence gene regulation[2–4]. Furthermore, genes in trait-associated loci are more likely to have genetic effects on their expression levels (expression QTLs, or eQTLs), and the variants with the strongest trait associations are more likely also to be associated with transcript abundance of at least one proximal gene[5]. Combined, these observations have led to the inference that most trait-associated variants are eQTLs, exerting their effect on phenotype by altering transcript abundance, rather than protein sequence. The mechanism may involve a knock-on effect on gene regulation, with the variant altering transcript abundances for genes elsewhere in the genome (a *trans*-eQTL), but the consensus view is that this must be mediated by the variant influencing a gene in the region (a *cis*-eQTL)[6]. As most eQTL studies profile cell populations or tissues from healthy donors at homeostatic equilibrium, the further assumption has been tacitly made that these trait-associated variants affect genes in *cis* under resting conditions. Equivalent QTL analyses of exon usage data have revealed a more modest overlap with trait-associated alleles, suggesting that a fraction of trait-associated variants influence splicing, and hence the relative abundance of different transcript isoforms, rather than overall expression levels. Thus, a model has emerged in which most trait-associated variants influence proximal gene regulation.

Several observations have challenged this basic model. One challenge comes from the difference between spatial distributions of eQTLs, which are dramatically enriched in close proximity of genes, and GWAS peaks, which are usually distal[7]. Another comes from colocalization analyses, attempting to map shared genetic associations between human traits and gene expression. If the model is correct, most trait associations should also be eQTLs; trait and expression phenotype should thus share an association in that locus (rather than two overlapping association peaks). However, only 5-40% of trait associations co-localize with eQTLs in relevant tissues or cell types[6,8–10], and only 15% of genes colocalize with any of 74 different complex traits[11]. Finally, expression levels mediated a minority of complex trait heritability[12]. This has led to the suggestion that most trait-associated alleles influence gene regulation in a context-specific manner[13]—either altering expression during development or in response to specific physiological stimuli—or that they act indirectly in *trans* to affect the regulation of a small number of genes involved in trait biology (the omnigenic model[14,15]). Without a set of

true positive cases, in which the gene driving trait variation is known, it remains difficult to assess either the basic model or the proposed variations.

One source of true positives is to identify genes that are both in loci associated with a complex trait and are also known to harbor coding mutations causing severe or early onset forms of related traits (e.g. related Mendelian disorders). The strong expectation is that a variant of small effect influences the gene identified in the severe form of the trait. This expectation is supported by several lines of evidence. Comorbidity between Mendelian and complex traits has been used to identify common variants associated with the complex traits[16]. A handful of genes have been conclusively identified in both Mendelian and complex forms of the same trait, including *APOE*, which is involved in cholesterol metabolism[17,18], and *SNCA*, which contributes to Parkinson's disease risk. Early genome-wide association studies (GWAS) found associations near genes identified through familial studies of severe disease[19,20], and more recent analyses have found that GWAS associations are enriched in regions near causative genes for cognate Mendelian traits in both blood traits[8] and a diverse collection of 62 traits[21].

To test the model that trait-associated variants influence baseline gene expression, therefore, we assembled a list of such "putatively causative" genes. We selected nine polygenic common traits with available large-scale GWAS data, each of which also has an extreme form in which coding mutations of large effect size affect one or more genes with well-characterized biology (Table 1). Our selection included four common diseases: type II diabetes[22], where early onset familial forms are caused by rare coding mutations (insulin-independent MODY; neonatal diabetes; maternally inherited diabetes and deafness; familial partial lipodystrophy); ulcerative colitis and Crohn disease[23,24], which have Mendelian pediatric forms characterized by severity of presentation; and breast cancer[25], where coding mutations in the germline (e.g. *BRCA1*) or somatic tissue (e.g. *PIK3CA*) are sufficient for disease. We also chose five quantitative traits: low and high density lipoprotein levels (LDL and HDL); systolic and diastolic blood pressure; and height. By manual literature search, we selected 134 genes harboring large-effect-size coding variants for one of the nine phenotypes (Table 1; specifically, we selected 134 gene-trait pairs, which represent 127 unique genes). These genes were identified in familial studies, rare disease exome sequencing, and, for breast cancer, using the MutPanning method[26] (citations for each gene are included in table 1).

We first examined whether these genes are more likely than chance to be in close proximity harboring variants associated with the polygenic form of each trait. In agreement with existing literature[21], we observe a significant enrichment for all traits except for height and breast cancer. However, in well-powered GWAS, even relatively rare large-effect coding alleles (mutations in *BRCA1* which cause breast cancer, for

instance) may be detectable as an association to common variants. To account for this possibility, we computed association statistics in each GWAS locus conditional on coding variants. We applied a direct conditional test to datasets with available individual-level genotype data (height, LDL, HDL); for those studies without available genotype data, we computed conditional associations from summary statistics using COJO[27,28]. After controlling for coding variation, we still detected a significant enrichment of our genes under GWAS peaks for all traits but height and breast cancer (Supp. Fig. 1). Of our 134 genes, 84 (63%) fell within 1 Mb of a GWAS locus for the cognate complex trait. Our window of 1 Mb represents roughly the upper bound for distances identified between enhancer-promoter pairs, but most pairs are closer[29], so we would expect enrichment to increase as the window around genes decreases; this proves to be the case (Supp. Fig. 1). After fine-mapping the GWAS associations in each locus using the SuSiE algorithm[30], we found that 23/134 (17%) putative causal genes are closer to the GWAS fine-mapped SNPs (posterior inclusion probability > 0.7) than any other gene in the locus, as measured from the transcription start site. Given their known causal roles in the severe forms of each phenotype, we thus suggest that the 84 genes near GWAS signals are likely to be the targets of trait-associated non-coding variants. For example, we see a significant GWAS association between breast cancer risk and variants in the estrogen receptor (*ESR1*) locus even after controlling for coding variation; the baseline expression model would thus predict that non-coding risk alleles alter *ESR1* expression to drive breast cancer risk.

We next looked for evidence that the trait-associated variants were also altering the expression of our 84 genes in relevant tissues. If these variants act through changes in gene expression, phenotypic associations should be driven by the same variants as eQTLs in relevant tissue types. We therefore looked for co-localization between our GWAS signals and eQTLs in relevant tissues (Supp. Tab. 1) drawn from the GTEx Project, using three well-documented methods: coloc[10], JLIM[9], and eCAVIAR[31]. We found support for the colocalization of trait and eQTL association for only four (coloc), seven (JLIM), and three (eCAVIAR) genes. Accounting for overlap, this represents 10 of our 84 putatively causative genes, even before correcting for multiple-hypothesis testing, which is not obviously better than random chance. We note that our estimates of the number of putatively causative *genes* with colocalization of eQTL and GWAS signal is conceptually distinct from and not directly comparable to the existing estimates of the fraction of *GWAS associations* colocalizing with eQTLs. This distinction matters because it illuminates the role of eQTLs in known trait biology rather than examining the locus for the presence of a colocalizing eQTL which may or may not be relevant to the complex trait.

A different way to identify potential causative genes under GWAS peaks using gene expression is the transcriptome-wide association study design (TWAS)[32–34]. This approach measures local genetic correlation between a complex trait and gene expression. Though not designed to avoid correlation signals caused by LD[35], the approach has higher power than colocalization methods in cases of allelic heterogeneity or poorly typed causative variants[32]. We used the FUSION implementation of TWAS, which accounts for the possibility of multiple cis-eQTLs linked to the trait-associated variant by jointly calling sets of genes predicted to include the causative gene, to interrogate our 84 loci[34].

FUSION included our putatively causative genes in the set of genes identified as likely relevant to the GWAS peak in 42 loci. TWAS does not require a genome-wide significant GWAS hit, and three of the genes returned (MAP2K4, BC; KCNJ11, T2D; ORC4, height) were not within 1 Mb of a GWAS peak. This leaves 39/84 (46%) of genes near GWAS peaks as positive TWAS results, in addition to the three not near a peak. Genes were often identified as hits in multiple tissues, but with an inconsistent direction of effect—that is, increased gene expression correlated with an increase in the quantitative trait or disease risk in some tissues, but a decrease in others. This may indicate that different tissues have relevant genes that are different, but still called within the same joint set. Because of this possibility, and the known biological role of many of our genes, we restricted our results to tissues with established relevance to our traits. Only 9/84 (11%) genes were identified by FUSION when we restricted the analysis to relevant tissues, and of these, only 5/84 (6%) had a direction of effect on the complex trait consistent with what is known from hypomorphic and amorphic Mendelian mutations. This fact, combined with the inconsistent direction of effect across tissues, may indicate that even when putatively causative genes fall within a set of genes jointly called by TWAS, their baseline expression may not be mediating the association.

Our results so far are consistent with trait-associated variants altering the regulation of causative genes in ways that are not well-represented by steady-state gene expression measurements. We thus tried to find whether fine-mapped GWAS variants are enriched in regulatory sites located within +/- 1 Mb of the transcription start sites (TSS) of our genes of interest. We found that 73 fine-mapped variants with a high posterior probability of association (PIP > 0.7) to a trait fall within some active histone modification feature (marked by narrow peaks of H3K27ac, H3K4me1, or H3K4me3 measurements) near putatively causative genes across all the tissue types examined. Despite our 1 Mb window, all identified features are located within a 100 kb window around the transcription starts sites of 27/84 (32%) putatively causative genes (*ATG16L1* is putatively causative for both CD and UC and is counted twice). Extending our search to include not only fine-mapped variants within chromatin modification

features, but also those within 500 bp of features, identifies only two additional putatively causative genes. Restricting our analysis to chromatin features in relevant tissues, 45 fine-mapped variants fall within chromatin features, corresponding to 25/84 (30%) putatively causative genes.

While the appearance of a fine-mapped GWAS variant inside or near a chromatin modification feature is a useful predictor of its importance, we also need to consider the specificity of the feature in relation to the genes surrounding it. To quantify the specificity of chromatin features for our putatively causative genes, we evaluated an "activity-by-distance" measure, a simplified version of the "activity-by-contact" method[36] for all chromatin features within +/- 1 Mb of the causative genes across different tissue types. Activity-by-distance (ABD) combines the activity of the above histone modification marks (specifically the strength of the acetylation or methylation peak) and linear distance (in basepairs) along the genome instead of the chromatin contact frequency between feature and TSS (Fig. 2). The two measures are multiplied together and normalized to generate a score that can be used as a proxy for measuring the specificity of a particular feature for a target gene in a tissue of relevance. When we projected our fine-mapped variants onto the chromatin modification features across different tissues, we found that 17 of them appear in features with the highest ABD scores in the loci, corresponding to 11/84 (13%) putatively causative genes. Restricting ourselves to only the relevant tissue types per trait, only nine fine-mapped variants fall inside features with the highest ABD scores, corresponding to 5/84 (6%) causative genes.

Next, we relaxed the requirement of proximity to a specific feature and selected all enhancer regions annotated by the ChromHMM[37] method in any measured cell or tissue type. Overall, within +/- 1 Mb windows of our putatively causative genes 120/335 fine-mapped variants fall in an enhancer region (i.e. enhancer, bivalent enhancer, genetic enhancer) highlighted by ChromHMM's core 15-state model. These enhancers correspond to 41/84 (49%) putatively causative genes. Restricting our analysis to relevant tissues, 51/335 fine-mapped variants fall in enhancers, corresponding to 27/84 (32%) putatively causative genes.

In sum, we observe that a sizable minority of our fine-mapped variants appear near sites of regulatory activity—suggested by the presence of activating chromatin marks or ChromHMM annotation. However, 47/84 (56%) putatively causative genes, no fine-mapped variants are associated with regulatory regions in relevant tissues according to either chromatin marks or ChromHMM. Furthermore, because we connect regulatory features to genes based solely on proximity, it is possible that our finding of

fine-mapped variants falling inside features specific to 30% of the putatively causative genes in relevant tissue types represents an overestimate.

Overall, our results do not support the assertion that most common non-coding variants associated with human traits alter baseline gene expression in trait-relevant tissues. Several explanations may account for this: incorrect assumptions, lack of statistical power, biological context, and alternative regulatory mechanisms. We discuss each below.

Incorrect assumptions: it is possible that our putatively causative genes may simply not be causative in complex trait forms. This would invalidate our underlying premise that they should be targets of trait-associated variants in the common, complex forms of phenotypes. This implies that in the vast majority of cases, a common variant associated with the polygenic form of a trait near a gene known to cause a severe form actually targets a different gene. For instance, the risk alleles driving the breast cancer GWAS signal near *BRCA2*, do not alter *BRCA2* expression in breast tissue, but instead influence another gene. This would also explain why 50 putatively causal genes do not fall near a GWAS peak. The implication is that the underlying biological causes of an extreme phenotypic presentation are different from the causes of the polygenic form across all nine of the traits we have studied. This, to our minds, stretches credulity given the highly significant enrichment of our genes near significant GWAS loci for cognate phenotypes. We suggest it is more likely that our putatively causative genes are relevant but influenced in some other way by polygenic risk alleles. For the 50 genes not near GWAS peaks, more parsimonious explanations are that currently available GWAS are incompletely powered, and thus have not detected association with alleles in those loci; or that strong purifying selection acting on noncoding regions of these genes is preventing noncoding variants from reaching population frequencies detectable by GWAS.

Lack of statistical power: it is possible that complex trait GWAS are insufficiently powered to allow accurate fine-mapping and hence accurate colocalization; that eQTL studies do not detect all eQTLs; that epigenetic studies do not identify all elements; or that colocalization and regulatory element mapping methods lack power to detect overlaps. We have ascertained GWAS associations at genome-wide significance, and fine-mapped the majority of these signals using a Bayesian approach; we believe it is unlikely that genetic associations to complex traits are the limiting factor for our analysis. Though the GTEx Consortium eQTL studies have identified eQTLs for 95% of protein-coding genes[6], tissues with fewer samples are likely to remain underpowered.

The upper bound on the power of colocalization methods, under near-ideal circumstances, is 66% at P < 0.01 (Barbeira et al. 2020). Under more typical conditions, the portion of GWAS peaks which colocalize with an eQTL is 25% or higher[9,10,31]. As not all GWAS peaks will share a causative SNP with a *cis*-eQTL, these estimates represent a lower bound on power, with empirical power likely to be much higher. Given our assumption that putatively causative genes are mediating association signals, we would expect that 25% of these associations would colocalize, and that in each case, the gene they colocalize with is our putatively causative gene. We would thus expect *at least* 21/84 (25%) of putatively causative genes near a polygenic trait association signal to have a colocalizing eQTL in relevant tissue. Here, we report all associations without correcting for multiple testing, so we would expect to find most of these colocalizations. We thus cannot attribute the absence of such events to lack of power. This conclusion is supported directly by our analyses: coloc explicitly tests the hypothesis that GWAS and eQTL signals are distinct, and finds strong statistical support for this hypothesis in three times as many loci as it finds evidence for colocalization. This suggests that, in many cases, genetically induced changes to baseline expression of putatively causative genes do not translate into downstream phenotypic effects. At the same time, most GWAS peaks over these genes are not eQTLs in available tissues.

The power of TWAS is comparable to colocalization methods in cases of a single typed causative SNP. Its relative power increases in cases of poorly-typed SNPs, allelic heterogeneity, or apparent heterogeneity (when multiple SNPs tag a single untyped causative SNP)[32]. Thus, the paucity of TWAS signals in the correct tissue and with the correct direction of effect cannot be explained by low power.

Biological context: causative eQTLs may only manifest in certain developmental windows, under specific conditions, or in a crucial cell subpopulation. We used data from the GTEx project, which profiled bulk post-mortem adult tissue samples. If causative eQTLs are only present in early development, or under specific exposures or conditions not applicable to the GTEx donors, they would not be captured in these contexts, even though *cis*-eQTLs have been detected for essentially every gene in the genome in the GTEx data[6].

Single-cell RNA sequencing (scRNA-seq) studies have identified eQTLs that are present in only a subset of the cell types captured in bulk-tissue analysis, but these appear to be limited—van der Wjist et al. found that 60% of cell type-specific eQTLs replicate in bulk-tissue analysis, and their use of scRNA-seq found only 13% more eQTLs than bulk-tissue analysis[38]. However, it has been posited that cell type-specific eQTLs may be enriched in disease association[39]. Additionally, genes causative for disease tend to have more enhancers, which may lead to more complex spatiotemporal

expression[40]. Nonetheless, using this tendency to explain the many putatively causative genes whose expression was not linked to GWAS requires us to believe most genes both have *cis*-eQTLs that do not show up in bulk-tissue analysis, and do not influence traits via those *cis*-eQTLs which do show up in bulk-tissue analysis. There is not, to our knowledge, evidence that this phenomenon is common, but given the nascence of research in cell type-specific eQTLs, we are withholding our judgement on this possibility.

A new cell type-specific TWAS method leverages large sample sizes for human bulk tissues and high-resolution mouse scRNA-seq data. It infers cell type-specific gene expression for each GTEx sample with respect to each Tabula Muris cell type under an empirical Bayes framework and produces gene expression prediction models at cell-type resolution. This method found no additional disease-associated gene in type II diabetes, and only one, targeting *FGFR2*, in breast cancer (albeit not in breast mammary tissue; Huwenbo Shi and Alkes Price, unpublished correspondence). This argues against cell type-specific eQTLs being the most prevalent effect of trait-associated variants.

It is possible for eQTLs to change or disappear over the course of development[41]. Because colocalization and TWAS methods rely on eQTL-mapping, such dynamic eQTLs present a potential blind spot. Chromatin marks provide an orthogonal source of information generally. Chromatin marks within a tissue, especially H3K4me3, can remain stable across developmental time[42]—though this is by no means universal—providing a useful, if imperfect, check on this blind spot.

The spatiotemporal patterns of gene expression may depend on tissue and cell types, stages of development, and environmental context. All such factors undoubtedly matter, complicating the question considerability. However, for this complexity to explain the majority of our negative result, it must be the case that: 1) context-dependent eQTLs exist for most of our gold-standard genes; 2) these eQTLs are not captured by steady state bulk-tissue RNA-seq; 3) most gold-standard genes do have steady state eQTLs captured by bulk-tissue analysis; 4) these measured eQTLs do not exist in the contexts in which the gene's expression matters.

Alternative regulatory mechanisms: finally, it is conceivable that most non-coding trait-associated variants act not on expression levels, but on other aspects of gene regulation. For example, splicing QTLs (sQTLs) are enriched in GWAS peaks to the same extent as eQTLs[43,44]. However, only 29% of our trait-associated variants that are highly likely to be causal (fine-mapping posterior probability > 0.7) fall in introns, despite

introns composing 45% of the genome[45]. Thus sQTLs do not immediately appear as a viable hypothesis to explain the majority of trait-associated variation.

We thus have to explain the observation that putatively causative genes are often near GWAS signals driven by non-coding variants, and that these genes are influenced by baseline eQTLs in relevant tissues, but that trait-associated variants are not driving those eQTLs. This result questions the basic assumption that trait variants act by perturbing baseline gene expression, so that eQTLs in GWAS peaks are necessarily relevant to the mapped trait. That these genes are more likely than chance to be near such non-coding trait-associated variants suggests that both the structure and regulation of these genes is relevant to complex traits. However, our results demonstrate that the mechanism by which our genes influence complex traits is generally not their baseline expression.

Regardless of the root cause, our results have consequences for efforts to uncover the biology underlying human traits by linking variants to molecular function through baseline expression measurements. These variant-to-function methods are currently the most common computational strategies for identifying the biological significance and therapeutic potential of non-coding genetic associations. Though they have successfully identified many genes of biological consequence and clinical promise, most causative genes likely go undiscovered. Given the difficulties many tissues present in obtaining expression data across diverse developmental and environmental contexts, the limitations of examining baseline expression may present a difficult obstacle to overcome.

There are limited mechanistic models to explain the function of non-coding variants besides their action as *cis*-eQTLs. Besides sQTLs, another possibility is *trans*-eQTLs that are not mediated by a *cis* effect on a gene, such as variants affecting CTCF binding sites[39], but this fails to explain the enrichment in GWAS signal near putatively causative genes. Though it is likely that power and context play a role in the lack of overlap we observe, for the reasons above it seems improbable that they explain it entirely. Cumulatively, our analysis shows that whilst gold standard genes are often the closest to a genetic association, more sophisticated analyses incorporating functional genomic data fail to identify them as relevant to the trait in meaningful numbers. There are currently no prominent models to fill this gap, but we must remember that complex trait genetics has overturned our assumptions time and time again.

Figures

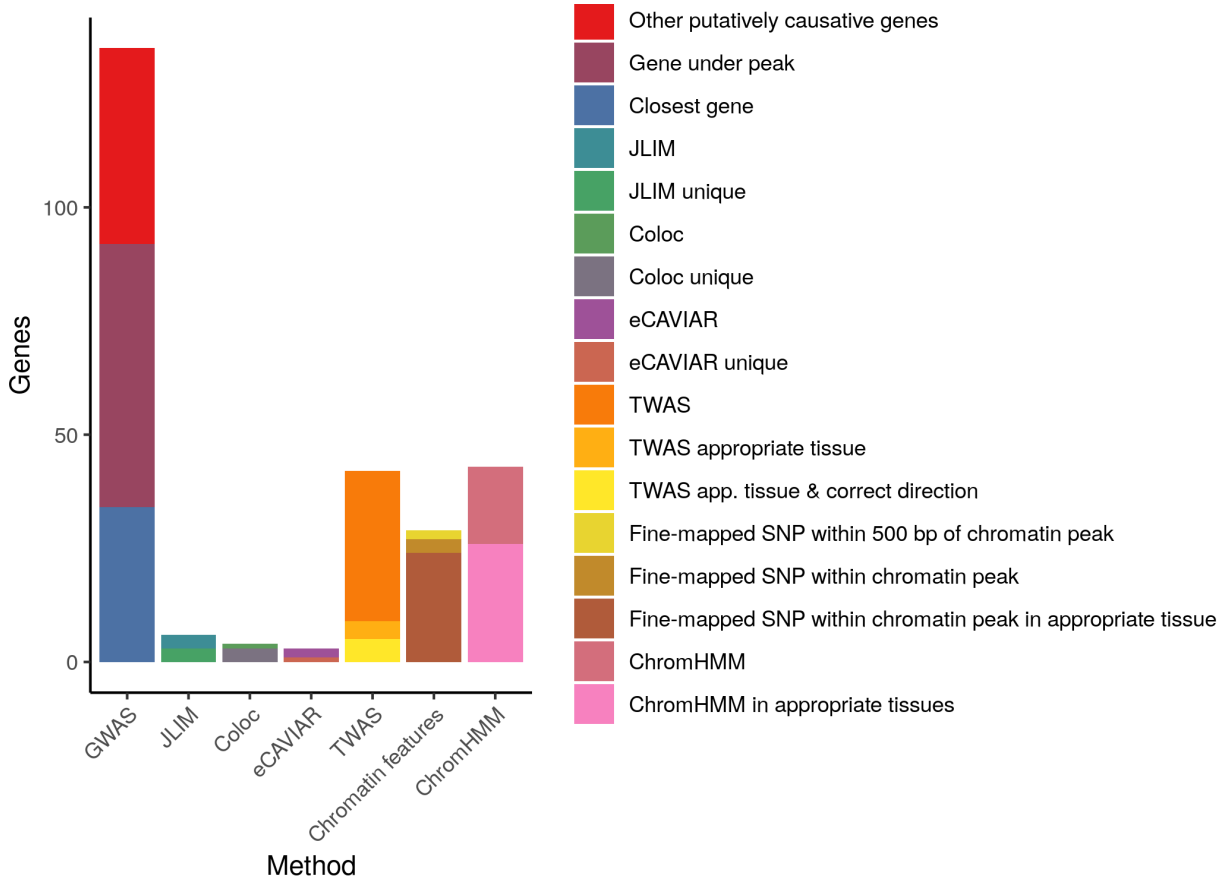


Figure 1. Putatively causative genes identified by each method.
The leftmost column displays the entire set of putatively causative genes, along with the subset near a linkage peak, and its subset of genes closest to the peak. For JLIM, Coloc, and eCAVIAR, the portion of genes that were the only gene to colocalize in their locus is noted. The numbers for these methods represent nominal significance thresholds. For TWAS results, the subsets of genes which are in an appropriate tissue and in an appropriate tissue in the right direction are indicated.

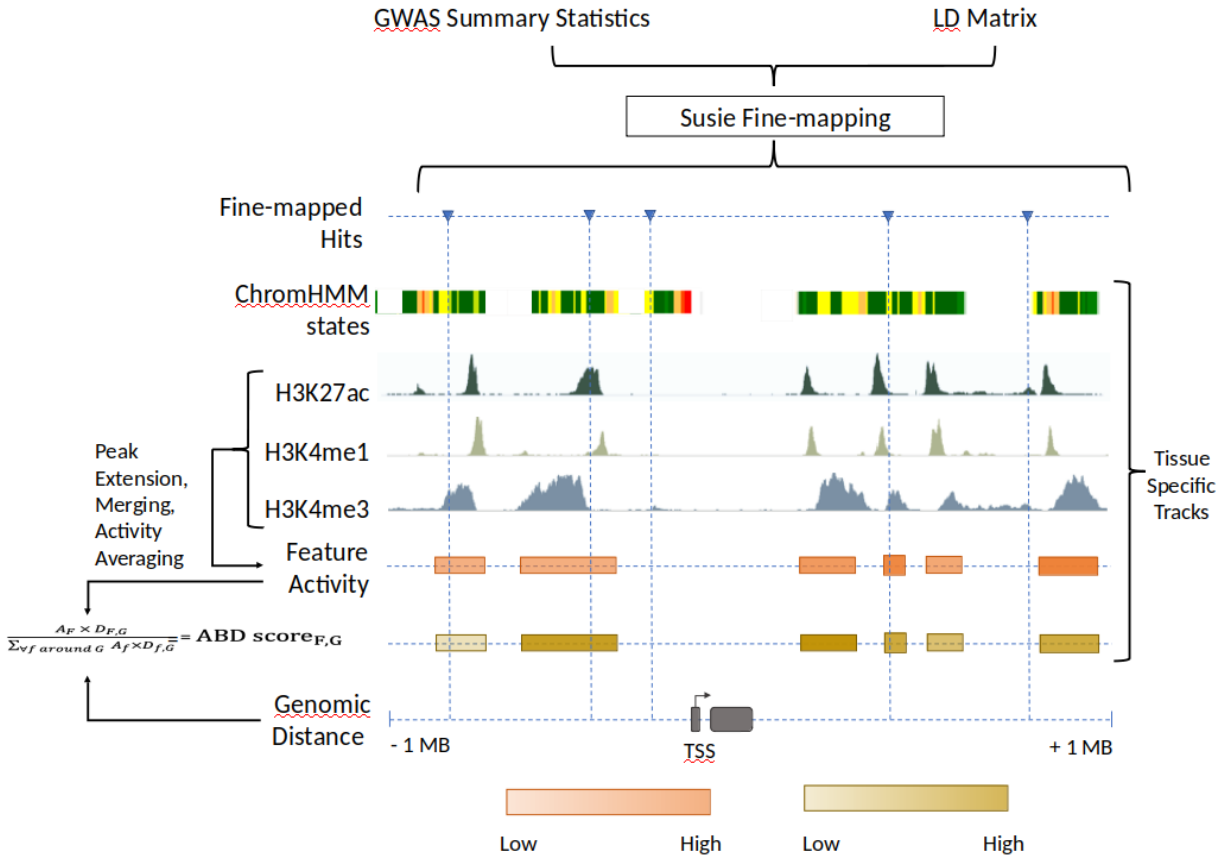
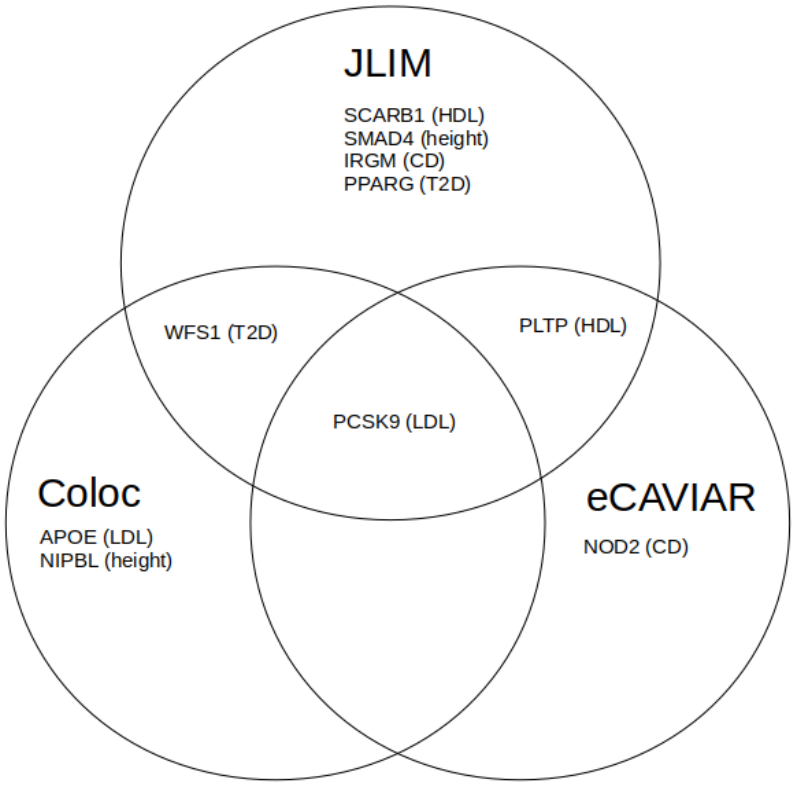
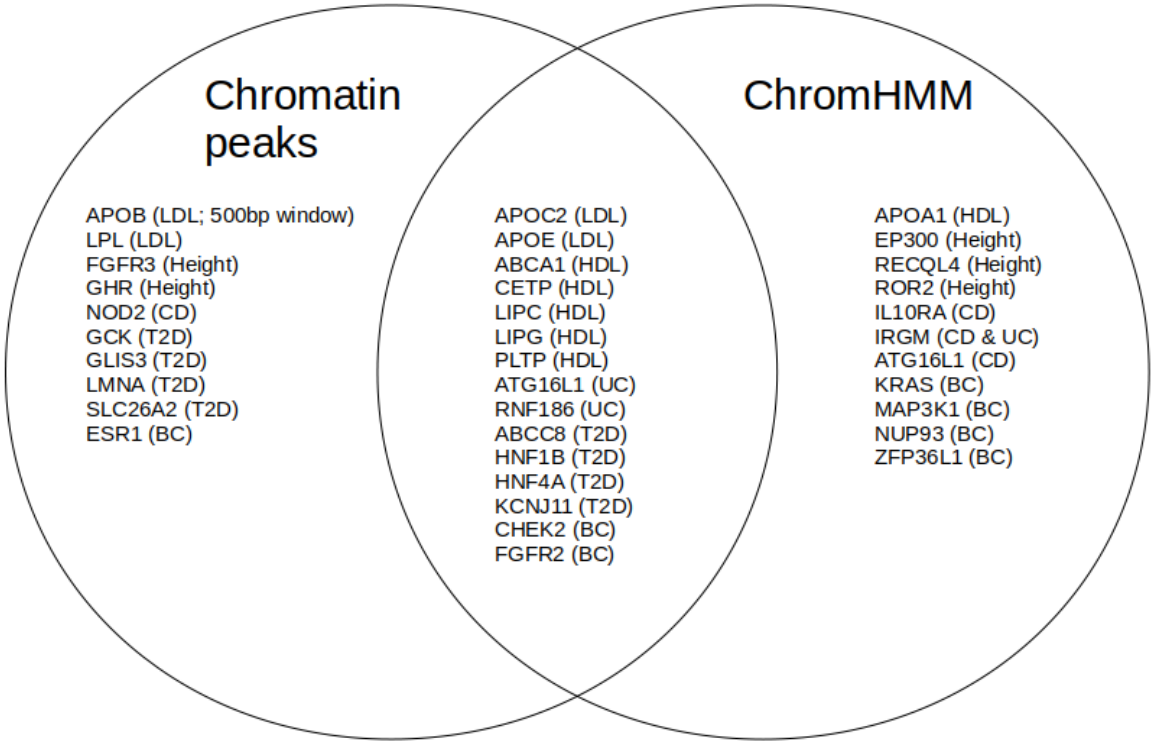


Figure 2. Chromatin-based causative gene identification.
Following the fine-mapping of GWAS variants, two parallel methods were used. One identified variants falling within regions annotated as enhancers by ChromHMM. The other identified variants within histone modification features, and evaluated their relevance using an ABD score that combined the strength of the feature (i.e. the strength of the acetylation or methylation peak) with its genomic distance to the gene of interest (see methods).

A)



B)



C)

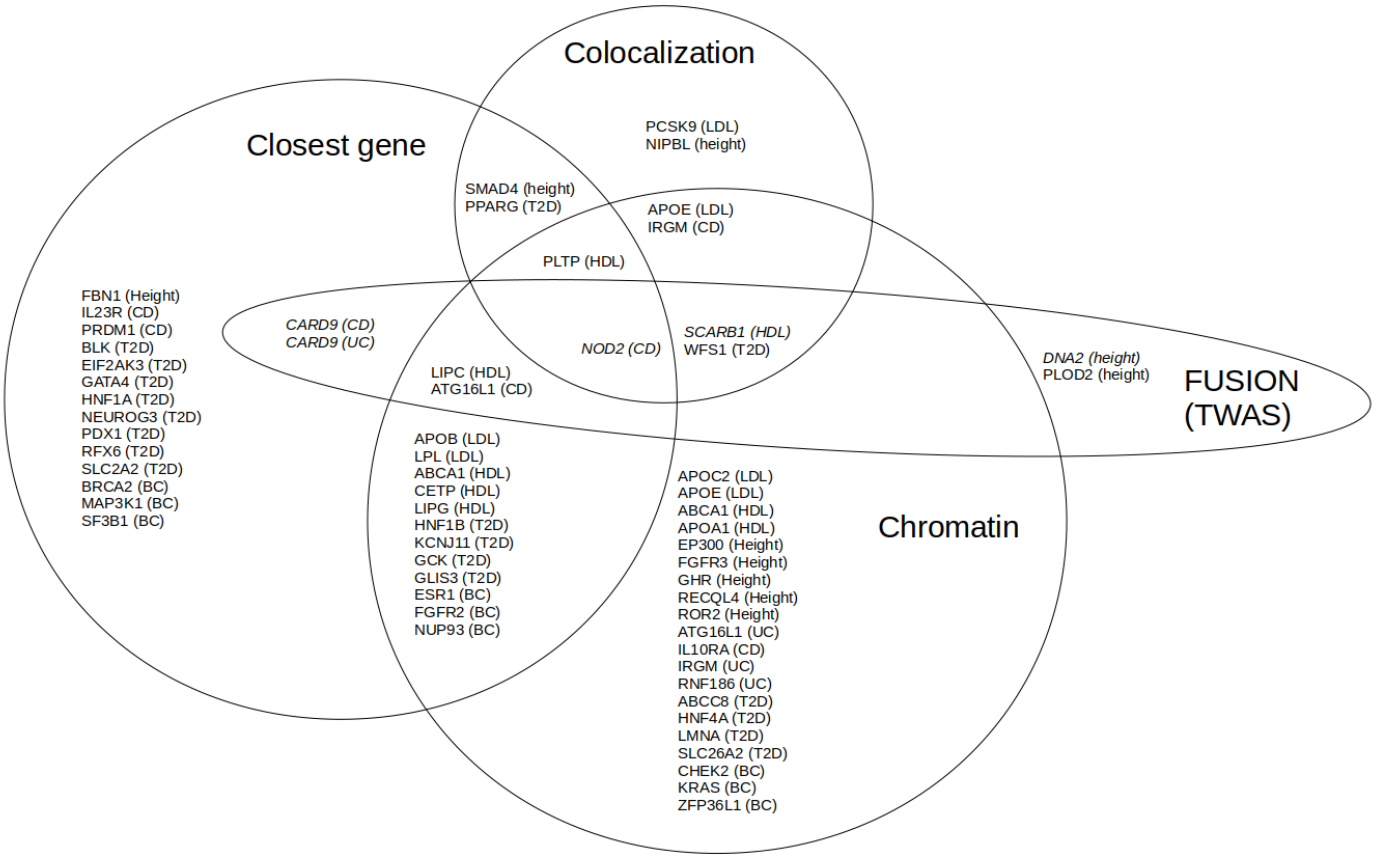


Figure 3. Genes identified as associated with a complex trait by each method. A) Positive results for each of the three colocalization methods. B) Positive results for each of the two chromatin methods. C) Positive results for all methods, collapsing A) to “colocalization” and B) to “chromatin.” The FUSION area reports all genes identified in the correct tissues, and genes for which the direction of effect is consistent with known biology are italicized (italication has no significance for other methods).

Phenotype	Genes
LDL	APOB[46,47] APOC2[48] APOE[49] LDLR[50] LPL[51,52] PCSK9[53]
HDL	ABCA1[54–56] APOA1[57] CETP[58]

	LIPC[59–61] LIPG[62] PLTP SCARB1[63,64]
Height	ANTXR1[65,66] ATR[67,68] BLM[69,70] CDC6[71] CDT1[71,72] CENPJ[73] COL1A1[74] COL1A2[75,76] COMP[77,78] CREBBP[79–81] DNA2[82] DTDST[83] EP300[84,85] EVC[86,87] EVC2[87,88] FBN1[89–92] FGFR3[93–95] FKBP10[96–98] GHR[99–102] KRAS[103–105] NBN[106,107] NIPBL[108,109] ORC1[71,72,110] ORC4[71,72] ORC6L[71,111] PCNT[112–114] PLOD2[115–117] PTPN11[118–120] RAD21[121–123] RAF1[124,125] RECQL4[126–128] RIT1[129–131] RNU4ATAC *should remove snRNA* ROR2[132–134] SLC26A2[135–137] SMAD4[138–140] SMC3 *milder form of trait, remove* SOS1 *same as above* SRCAP[141,142] WRN[143–145]

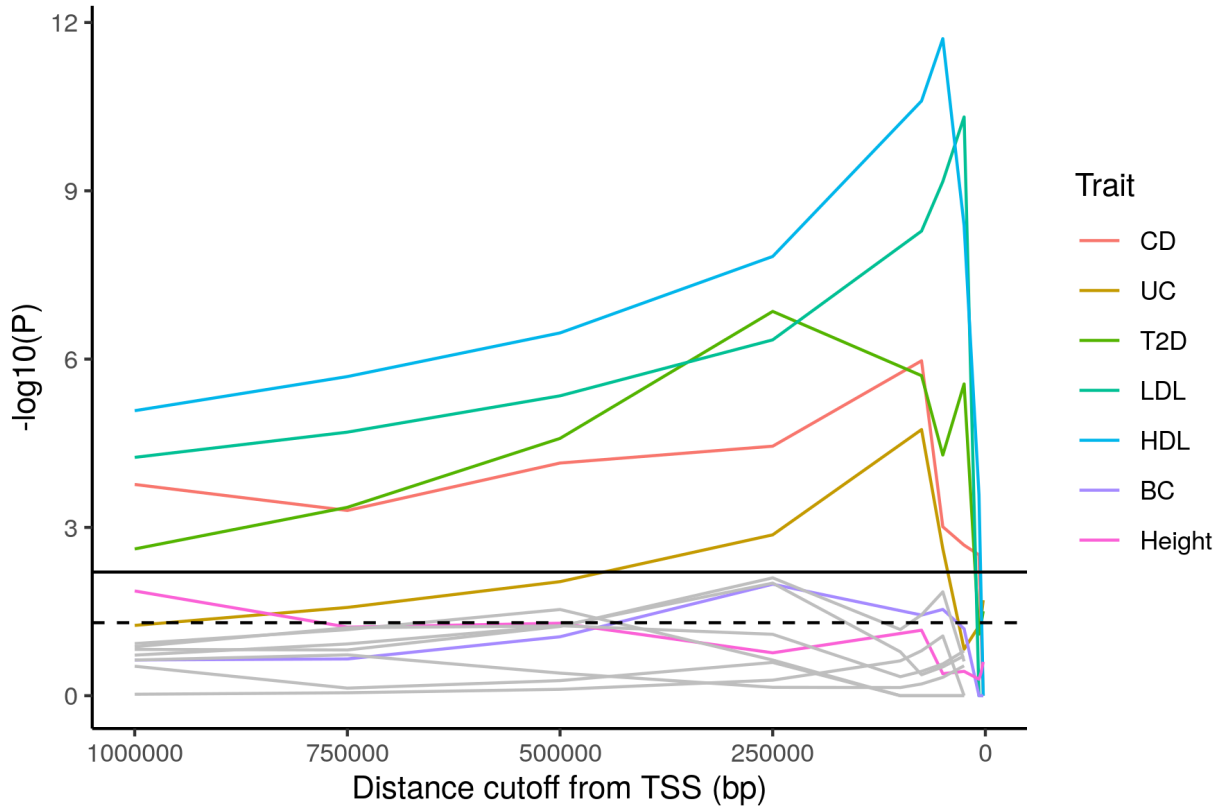
Blood pressure (systolic and diastolic)	KCNJ1[146,147] SLC12A1[148,149] SLC12A3[150,151] WNK1[152] WNK4[153]
Crohn disease	ATG16L1[154] CARD9[155] IL10[156] IL10RA[157,158] IL10RB[159,160] IL23R[161–163] IRGM[164–166] NOD2[167,168] PRDM1[169] PTPN22[170] RNF186
Ulcerative colitis	ATG16L1[171] CARD9[155] IL23R[163,172] IRGM[164] PRDM1[169] PTPN22[170] RNF186[173,174]
Type II diabetes	ABCC8[175] BLK[176] CEL[177,178] EIF2AK3[179–181] GATA4[182] GATA6[183,184] GCK[185] GLIS3[186] HNF1A[187,188] HNF1B[189,190] HNF4A[191,192] IER3IP1[193–195] INS[196] KCNJ11[197,198] KLF11[199] LMNA[200] NEUROD1[201] NEUROG3[202–204] PAX4[205–207] PDX1[208–210]

	PPARG[211,212] PTF1A[213] RFX6[214,215] SLC19A2[216–218] SLC2A2[219,220] WFS1[221–223] ZFP57[224,225]
Breast cancer (selected using MutPanning[26])	AKT1 ARID1A ATM BRCA1 BRCA2 CBFB CDH1 CDKN1B CHEK2 CTCF ERBB2 ESR1 FGFR2 FOXA1 GATA3 GPS2 HS6ST1 KMT2C KRAS LRRC37A3 MAP2K4 MAP3K1 NCOR1 NF1 NUP93 PALB2 PIK3CA PTEN RB1 RUNX1 SF3B1 STK11 TBX3 TP53 ZFP36L1

Table 1. Putatively causative genes

Mendelian trait	GWAS trait	Tissues examined
Breast cancer	Breast cancer	Breast mammary tissue
Crohn disease	Crohn disease	Small intestine terminal ileum Colon Sigmoid Colon Transverse
Ulcerative colitis	Ulcerative colitis	Small intestine terminal ileum Colon Sigmoid Colon Transverse
Dyslipidemia Hyperlipidemia Tangier's disease	HDL	Liver Adipose (subcutaneous) Whole blood
Dyslipidemia Hyperlipidemia	LDL	Liver Adipose (subcutaneous) Whole blood
Mendelian short stature	Height	Skeletal muscle
Blood pressure	Blood pressure	Heart atrial appendage Heart left ventricle
Monogenic diabetes	Type II diabetes	Pancreas Skeletal muscle Adipose (subcutaneous) Small intestine terminal ileum

Table 2. Tissue-trait pairs
Tissues were selected for each trait based on *a priori* knowledge of disease biology.



Supplementary figure 1. Enrichment of Mendelian genes near GWAS peaks.
As the window around GWAS peaks shrinks, the enrichment of Mendelian genes within the window becomes increasingly significant, while the enrichment of non-matching trait pairs used as controls (gray lines; see methods) is not consistently increased. Some controls achieve nominal significance (dotted horizontal line), but none reach significance once multiple-testing is corrected for (solid horizontal line). We suspect that the non-significant results for breast cancer and height are due at least in part to the large number of GWAS peaks these traits have. At a distance cutoff of 25 kb, most traits no longer have significant enrichment, as the absolute number of genes captured within the window becomes too small.

Alias	Descriptive Name
E027	Breast myoepithelial primary Cells
E028	Breast luminal epithelial Cells / Human mammary epithelial cells
E062	Primary mononuclear cells from peripheral blood
E063	Adipose nuclei
E066	Liver
E075	Colonic mucosa
E076	Colon smooth muscle
E087	Pancreatic islets
E095	Heart left ventricle
E096	Lung
E104	Heart right atrium
E107	Skeletal Muscle Male
E108	Skeletal Muscle Female
E109	Small Intestine

Supplementary Table 1. Roadmap epigenomics aliases of tissue types used for functional genomic analysis.

Supplementary methods

Gene selection

Our gold-standard genes were selected by manual literature search. Review papers, as well as the OMIM database[226], were generally used as starting points, but an examination of the original literature was needed to confirm genes' suitability. For example, though SMC3 is known to cause Cornelia de Lange syndrome, which is characterized in part by short stature, SMC3 mutations lead to a milder form of the syndrome, usually without a marked reduction in stature[227].

Identifying coding variants

Because GWAS sample sizes are large enough to detect the low-frequency coding variants used to select some of our genes, it is possible that a coding SNP would distort

the association signal of nearby eQTLs. To minimize this concern, we removed the effects of coding variants on GWAS. Many variants can fall within coding sequences in rare splice variants, so it is important to remove only those variants that appear commonly as coding. These coding SNPs were selected based on the pext (proportion of expression across transcripts) data[228]. Two filters were used. First, we removed genes whose expression in a trait-relevant tissue was below 50% of their maximum expression across tissues. Second, we removed variants that fell within the coding sequence of less than 25% of splice isoforms in that tissue. The remaining variants were used to correct GWAS signal, as explained below.

GWAS

For height, LDL cholesterol, and HDL cholesterol, GWAS were performed using genotypic and phenotypic data from the UKBB. In order to avoid confounding, we restricted our sample to the 337K unrelated individuals with genetically determined British ancestry identified by Bycroft et al.[229] The GWAS were run using Plink 2.0[230], with the covariates age, sex, BMI (for LDL and HDL only), 10 principal components, and coding SNPs.

Conditional analysis

Because UKBB has limited power for breast cancer, Crohn disease, ulcerative colitis, and type II diabetes, we used publicly available summary statistics. The Conditional and Joint Analysis (COJO)[27,28] program can condition summary statistics on selected variants—in our case, coding variants—by using an LD reference panel. For this reference, we used TOPMed subjects of European ancestry[231]. The ancestry of these subjects was confirmed with FastPCA[232,233] and the relevant data were extracted using bcftools[234].

Enrichment analysis

At each distance, the number of Mendelian and non-Mendelian genes within that window around GWAS peaks are counted. *P*-values are calculated using Fisher's exact test (Supp. Fig. 1). Because Mendelian genes may be unusually important beyond our chosen traits, we conduct a set of controls by measuring the enrichment of non-matching Mendelian and complex traits (CD genes & BC GWAS; BC genes & LDL GWAS; LDL genes & UC GWAS; UC genes & height GWAS; height genes & T2D GWAS; T2D genes & HDL GWAS; HDL genes & CD GWAS).

Colocalization

JLIM[9] was run using GWAS summary statistics and GTEx v7 genotypes and phenotypes. Coloc[10] was run using GWAS and GTEx v7 summary statistics. eCAVIAR[31]

was run using GWAS and GTEx v7 summary statistics, and a reference dataset of LD from UKBB[235].

Fine-mapping GWAS hits

We fine-mapped the GWAS variants located within +/- 1 Mb of our putatively causative genes by applying the SuSiE algorithm[30] on the unconditional summary statistics from the GWAS of breast cancer, Crohn disease, ulcerative colitis, type II diabetes, height, LDL cholesterol, and HDL cholesterol. An LD reference panel from UKBB subjects of European ancestry was used for this analysis. Fine-mapped variants were annotated using snpEff (v4.3t). Only non-coding variants were kept for further analysis.

Functional genomic annotation of fine-mapped hits

We downloaded imputed narrowPeak sets for acetylation of histone H3 lysine 27 residues (H3K27ac), mono-methylation of histone H3 lysine 4 residues (H3K4me1), and tri-methylation of lysine 4 residues (H3K4me3) from the Roadmap Epigenomics Project[37] ftp site (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidatedImputed/narrowPeak/) for 14 different tissue types (Supp. Tab. 1). For each tissue type, we extracted the narrow peaks that are within +/- 5 Mb of our putatively causative genes. Then following the approach described in Fulco et al.[36], we extended the 150 bp narrow peaks by 175 bp on both sides to arrive at candidate features of 500 bp in length. All features mapping to blacklisted regions (https://sites.google.com/site/anshulkundaje/projects/blacklists) were removed. Remaining features were re-centered around the peak and overlapping features were merged to give the final set of features per histone modification track. Next, we calculated the mean activity/strength of a feature (A_F) by taking the geometric mean of the corresponding peak strengths from H3K27ac, H3K4me1, and H3K4me3 marks. We then combined these activity measurements with the linear distances between the features and the transcription start sites of causative genes to compute “activity-by-distance” scores (a simplified version of ABC scores[36]) for gene-feature pairs using the following formula.

$$ABD\ score_{F,G} = \frac{A_F \times D_{F,G}}{\sum_{all\ f\ within \pm 5\ Mb\ of\ G} A_f \times D_{f,G}}$$

The ABD score can be thought of as a measure of the contribution of a feature, F to the combined regulatory signals acting on gene, G. A high ABD score may serve as a proxy for an increased specificity between a chromatin feature and the gene of interest. We projected the fine-mapped variants onto the chromatin features in different tissue types to assess whether there is an enrichment of likely causal GWAS hits in regulatory features near our putatively causative genes. Both proximity (genomic distance) and

specificity (ABD scores) were considered to determine the regulatory contribution of the fine-mapped hits.

Chromatin state predictions (chromHMM core 15-state model[37]) for the same tissue types (Supplementary Table 1) were downloaded from the Roadmap Epigenomics Project[37] ftp site (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/). We considered a fine-mapped variant to fall in an enhancer region if it mapped to a chromHMM segment described as enhancer, bivalent enhancer, or genic enhancer.

References

1. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
2. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
4. 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).
5. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLOS Genet.* **6**, e1000888 (2010).
6. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
7. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
8. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231.e11 (2020).
9. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and

autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).

10. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
11. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *bioRxiv* 814350 (2020) doi:10.1101/814350.
12. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
13. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
14. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
15. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).
16. Blair, D. R. *et al.* A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell* **155**, 70–80 (2013).
17. Schneider, W. J. *et al.* Familial dysbetalipoproteinemia. Abnormal binding of mutant apoprotein E to low density lipoprotein receptors of human fibroblasts and membranes from liver and adrenal of rats, rabbits, and cows. *J. Clin. Invest.* **68**, 1075–1085 (1981).
18. Boerwinkle, E. & Utermann, G. Simultaneous effects of the apolipoprotein E polymorphism on apolipoprotein E, apolipoprotein B, and cholesterol metabolism. *Am. J. Hum. Genet.* **42**, 104–112 (1988).
19. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
20. Chan, Y. *et al.* Genome-wide Analysis of Body Proportion Classifies Height-Associated Variants by Mechanism of Action and Implicates Genes Important for Skeletal Development.

Am. J. Hum. Genet. **96**, 695–708 (2015).

21. Freund, M. K. *et al.* Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. *Am. J. Hum. Genet.* **103**, 535–552 (2018).

22. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

23. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

24. Goyette, P. *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat. Genet.* **47**, 172–179 (2015).

25. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581 (2020).

26. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).

27. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

28. 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).

29. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

30. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

31. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

32. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

33. GTEx Consortium *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

34. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).

35. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).

36. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).

37. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

38. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

39. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* (2020) doi:10.1016/j.tig.2020.08.009.

40. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).

41. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).

42. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).

43. Walker, R. L. *et al.* Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. *Cell* **179**, 750-771.e22 (2019).

44. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).

45. Francis, W. R. & Wörheide, G. Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. *Genome Biol. Evol.* **9**, 1582–1598 (2017).

46. Soria, L. F. *et al.* Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. *Proc. Natl. Acad. Sci.* **86**, 587–591 (1989).

47. Pullinger, C. R. *et al.* Familial ligand-defective apolipoprotein B. Identification of a new mutation that decreases LDL receptor binding affinity. *J. Clin. Invest.* **95**, 1225–1234 (1995).

48. Hegele, R. A. *et al.* An apolipoprotein CII mutation, CIILys19----Thr' identified in patients with hyperlipidemia. *Dis. Markers* **9**, 73–80 (1991).

49. de Knijff, P., van den Maagdenberg, A. M. J. M., Frants, R. R. & Havekes, L. M. Genetic heterogeneity of apolipoprotein E and its influence on plasma lipid and lipoprotein levels. *Hum. Mutat.* **4**, 178–194 (1994).

50. Brown, M. S. & Goldstein, J. L. Analysis of a mutant strain of human fibroblasts with a defect in the internalization of receptor-bound low density lipoprotein. *Cell* **9**, 663–674 (1976).

51. Heizmann, C. *et al.* DNA polymorphism haplotypes of the human lipoprotein lipase gene: possible association with high density lipoprotein levels. *Hum. Genet.* **86**, 578–584 (1991).

52. Clee, S., Loubser, O., Collins, J., Kastelein, J. & Hayden, M. The LPL S447X cSNP is associated with decreased blood pressure and plasma triglycerides, and reduced risk of coronary artery disease. *Clin. Genet.* **60**, 293–300 (2001).

53. Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat. Genet.* **34**, 154–156 (2003).

54. Brooks-Wilson, A. *et al.* Mutations in ABC1 in Tangier disease and familial high-density lipoprotein deficiency. *Nat. Genet.* **22**, 336–345 (1999).

55. Bodzioch, M. *et al.* The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. *Nat. Genet.* **22**, 347–351 (1999).

56. Rust, S. *et al.* Tangier disease is caused by mutations in the gene encoding ATP-binding cassette transporter 1. *Nat. Genet.* **22**, 352–355 (1999).

57. Ordovas, J. M. *et al.* Apolipoprotein A-I Gene Polymorphism Associated with Premature Coronary Artery Disease and Familial Hypoalphalipoproteinemia. *http://dx.doi.org/10.1056/NEJM198603133141102* https://www.nejm.org/doi/10.1056/NEJM198603133141102 (1986) doi:10.1056/NEJM198603133141102.

58. Glueck, C. J., Fallat, R. W., Millett, F. & Steiner, P. M. Familial Hyperalphalipoproteinemia. *Arch. Intern. Med.* **135**, 1025–1028 (1975).

59. Isaacs, A., Sayed-Tabatabaei, F. A., Njajou, O. T., Witteman, J. C. M. & van Duijn, C. M. The −514 C→T Hepatic Lipase Promoter Region Polymorphism and Plasma Lipids: A Meta-Analysis. *J. Clin. Endocrinol. Metab.* **89**, 3858–3863 (2004).

60. Grarup, N. *et al.* The −250G>A Promoter Variant in Hepatic Lipase Associates with Elevated Fasting Serum High-Density Lipoprotein Cholesterol Modulated by Interaction with Physical Activity in a Study of 16,156 Danish Subjects. *J. Clin. Endocrinol. Metab.* **93**, 2294–2299 (2008).

61. Iijima, H. *et al.* Association of an intronic haplotype of the LIPC gene with hyperalphalipoproteinemia in two independent populations. *J. Hum. Genet.* **53**, 193–200 (2008).

62. Yamakawa-Kobayashi, K., Yanagi, H., Endo, K., Arinami, T. & Hamaguchi, H. Relationship between serum HDL-C levels and common genetic variants of the endothelial lipase gene in Japanese school-aged children. *Hum. Genet.* **113**, 311–315 (2003).

63. Tai, E. *et al.* Polymorphisms at the SRBI locus are associated with lipoprotein levels in subjects with heterozygous familial hypercholesterolemia. *Clin. Genet.* **63**, 53–58 (2003).

64. McCarthy, J. J. *et al.* Association of genetic variants in the HDL receptor, SR-B1, with abnormal lipids in women with coronary artery disease. *J. Med. Genet.* **40**, 453–458 (2003).

65. Stránecký, V. *et al.* Mutations in ANTXR1 Cause GAPO Syndrome. *Am. J. Hum. Genet.* **92**, 792–799 (2013).

66. Bayram, Y. *et al.* Whole exome sequencing identifies three novel mutations in ANTXR1 in families with GAPO syndrome. *Am. J. Med. Genet. A.* **164**, 2328–2334 (2014).

67. O'Driscoll, M., Ruiz-Perez, V. L., Woods, C. G., Jeggo, P. A. & Goodship, J. A. A splicing mutation affecting expression of ataxia–telangiectasia and Rad3–related protein (ATR) results in Seckel syndrome. *Nat. Genet.* **33**, 497–501 (2003).

68. Ogi, T. *et al.* Identification of the First ATRIP–Deficient Patient and Novel Mutations in ATR Define a Clinical Spectrum for ATR–ATRIP Seckel Syndrome. *PLOS Genet.* **8**, e1002945 (2012).

69. Ellis, N. A. *et al.* The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* **83**, 655–666 (1995).

70. Foucault, F. *et al.* Characterization of a New BLM Mutation Associated with a Topoisomerase IIα Defect in a Patient with Bloom's Syndrome. *Hum. Mol. Genet.* **6**, 1427–1434 (1997).

71. Bicknell, L. S. *et al.* Mutations in the pre-replication complex cause Meier-Gorlin syndrome. *Nat. Genet.* **43**, 356–359 (2011).

72. Guernsey, D. L. *et al.* Mutations in origin recognition complex gene ORC4 cause Meier-Gorlin syndrome. *Nat. Genet.* **43**, 360–364 (2011).

73. Al-Dosari, M. S., Shaheen, R., Colak, D. & Alkuraya, F. S. Novel CENPJ mutation causes Seckel syndrome. *J. Med. Genet.* **47**, 411–414 (2010).

74. Wallis, G. A., Starman, B. J., Zinn, A. B. & Byers, P. H. Variable expression of osteogenesis imperfecta in a nuclear family is explained by somatic mosaicism for a lethal point mutation in the alpha 1(I) gene (COL1A1) of type I collagen in a parent. *Am. J. Hum. Genet.* **46**, 1034–1040 (1990).

75. Spotila, L. D., Sereda, L. & Prockop, D. J. Partial isodisomy for maternal chromosome 7

and short stature in an individual with a mutation at the COL1A2 locus. *Am. J. Hum. Genet.* **51**, 1396–1405 (1992).

76. Paepe, A. D., Nuytinck, L., Raes, M. & Fryns, J.-P. Homozygosity by descent for a COL1A2 mutation in two sibs with severe osteogenesis imperfecta and mild clinical expression in the heterozygotes. *Hum. Genet.* **99**, 478–483 (1997).

77. Briggs, M. D. *et al.* Pseudoachondroplasia and multiple epiphyseal dysplasia due to mutations in the cartilage oligomeric matrix protein gene. *Nat. Genet.* **10**, 330–336 (1995).

78. Mabuchi, A. *et al.* Novel types of COMP mutations and genotype-phenotype association in pseudoachondroplasia and multiple epiphyseal dysplasia. *Hum. Genet.* **112**, 84–90 (2003).

79. Menke, L. A. *et al.* CREBBP mutations in individuals without Rubinstein–Taybi syndrome phenotype. *Am. J. Med. Genet. A.* **170**, 2681–2693 (2016).

80. Menke, L. A. *et al.* Further delineation of an entity caused by CREBBP and EP300 mutations but not resembling Rubinstein–Taybi syndrome. *Am. J. Med. Genet. A.* **176**, 862–876 (2018).

81. Angius, A. *et al.* Confirmation of a new phenotype in an individual with a variant in the last part of exon 30 of CREBBP. *Am. J. Med. Genet. A.* **179**, 634–638 (2019).

82. Shaheen, R. *et al.* Genomic analysis of primordial dwarfism reveals novel disease genes. *Genome Res.* **24**, 291–299 (2014).

83. Hästbacka, J. *et al.* Identification of the Finnish founder mutation for diastrophic dysplasia (DTD). *Eur. J. Hum. Genet.* **7**, 664–670 (1999).

84. Woods, S. A. *et al.* Exome sequencing identifies a novel EP300 frame shift mutation in a patient with features that overlap cornelia de lange syndrome. *Am. J. Med. Genet. A.* **164**, 251–258 (2014).

85. Tsai, A. C.-H. *et al.* Exon deletions of the EP300 and CREBBP genes in two children with Rubinstein–Taybi syndrome detected by aCGH. *Eur. J. Hum. Genet.* **19**, 43–49 (2011).

86. Polymeropoulos, M. H. *et al.* The Gene for the Ellis–van Creveld Syndrome Is Located

on Chromosome 4p16. *Genomics* **35**, 1–5 (1996).

87. Ruiz-Perez, V. L. *et al.* Mutations in Two Nonhomologous Genes in a Head-to-Head Configuration Cause Ellis-van Creveld Syndrome. *Am. J. Hum. Genet.* **72**, 728–732 (2003).

88. Galdzicka, M. *et al.* A new gene, EVC2, is mutated in Ellis–van Creveld syndrome. *Mol. Genet. Metab.* **77**, 291–295 (2002).

89. Faivre, L. *et al.* In frame fibrillin-1 gene deletion in autosomal dominant Weill-Marchesani syndrome. *J. Med. Genet.* **40**, 34–36 (2003).

90. Le Goff, C. *et al.* Mutations in the TGFβ Binding-Protein-Like Domain 5 of FBN1 Are Responsible for Acromicric and Geleophysic Dysplasias. *Am. J. Hum. Genet.* **89**, 7–14 (2011).

91. Horn, D. & Robinson, P. N. Progeroid facial features and lipodystrophy associated with a novel splice site mutation in the final intron of the FBN1 gene. *Am. J. Med. Genet. A.* **155**, 721–724 (2011).

92. Takenouchi, T. *et al.* Severe congenital lipodystrophy and a progeroid appearance: Mutation in the penultimate exon of FBN1 causing a recognizable phenotype. *Am. J. Med. Genet. A.* **161**, 3057–3062 (2013).

93. Hyland, V. J. *et al.* Somatic and germline mosaicism for a R248C missense mutation in FGFR3, resulting in a skeletal dysplasia distinct from thanatophoric dysplasia. *Am. J. Med. Genet. A.* **120A**, 157–168 (2003).

94. Toydemir, R. M. *et al.* A Novel Mutation in FGFR3 Causes Camptodactyly, Tall Stature, and Hearing Loss (CATSHL) Syndrome. *Am. J. Hum. Genet.* **79**, 935–941 (2006).

95. Makrythanasis, P. *et al.* A Novel Homozygous Mutation in FGFR3 Causes Tall Stature, Severe Lateral Tibial Deviation, Scoliosis, Hearing Impairment, Camptodactyly, and Arachnodactyly. *Hum. Mutat.* **35**, 959–963 (2014).

96. Alanay, Y. *et al.* Mutations in the Gene Encoding the RER Protein FKBP65 Cause Autosomal-Recessive Osteogenesis Imperfecta. *Am. J. Hum. Genet.* **86**, 551–559 (2010).

97. Kelley, B. P. *et al.* Mutations in FKBP10 cause recessive osteogenesis imperfecta and bruck syndrome. *J. Bone Miner. Res.* **26**, 666–672 (2011).

98. Barnes, A. M. *et al.* Kuskokwim Syndrome, a Recessive Congenital Contracture Disorder, Extends the Phenotype of FKBP10 Mutations. *Hum. Mutat.* **34**, 1279–1288 (2013).

99. Berg, M. A. *et al.* Diverse growth hormone receptor gene mutations in Laron syndrome. *Am. J. Hum. Genet.* **52**, 998–1005 (1993).

100. Woods, K. A., Fraser, N. C., Postel-Vinay, M. C., Savage, M. O. & Clark, A. J. A homozygous splice site mutation affecting the intracellular domain of the growth hormone (GH) receptor resulting in Laron syndrome with elevated GH-binding protein. *J. Clin. Endocrinol. Metab.* **81**, 1686–1690 (1996).

101. Goddard, A. D. *et al.* Mutations of the Growth Hormone Receptor in Children with Idiopathic Short Stature. *http://dx.doi.org/10.1056/NEJM199510263331701* https://www.nejm.org/doi/10.1056/NEJM199510263331701 (1995) doi:10.1056/NEJM199510263331701.

102. Ayling, R. M. *et al.* A dominant-negative mutation of the growth hormone receptor causes familial short stature. *Nat. Genet.* **16**, 13–14 (1997).

103. Aoki, Y. *et al.* Germline mutations in HRAS proto-oncogene cause Costello syndrome. *Nat. Genet.* **37**, 1038–1040 (2005).

104. Schubbert, S. *et al.* Germline KRAS mutations cause Noonan syndrome. *Nat. Genet.* **38**, 331–336 (2006).

105. Carta, C. *et al.* Germline Missense Mutations Affecting KRAS Isoform B Are Associated with a Severe Noonan Syndrome Phenotype. *Am. J. Hum. Genet.* **79**, 129–135 (2006).

106. Varon, R. *et al.* Nibrin, a Novel DNA Double-Strand Break Repair Protein, Is Mutated in Nijmegen Breakage Syndrome. *Cell* **93**, 467–476 (1998).

107. Tanzarella, C. *et al.* Chromosome instability and nibrin protein variants in NBS heterozygotes. *Eur. J. Hum. Genet.* **11**, 297–303 (2003).

108. Tonkin, E. T., Wang, T.-J., Lisgo, S., Bamshad, M. J. & Strachan, T. NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. *Nat. Genet.* **36**, 636–641 (2004).

109. Krantz, I. D. *et al.* Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B. *Nat. Genet.* **36**, 631–635 (2004).

110. Bicknell, L. S. *et al.* Mutations in ORC1, encoding the largest subunit of the origin recognition complex, cause microcephalic primordial dwarfism resembling Meier-Gorlin syndrome. *Nat. Genet.* **43**, 350–355 (2011).

111. de Munnik, S. A. *et al.* Meier–Gorlin syndrome genotype–phenotype studies: 35 individuals with pre-replication complex gene mutations and 10 without molecular diagnosis. *Eur. J. Hum. Genet.* **20**, 598–606 (2012).

112. Rauch, A. *et al.* Mutations in the Pericentrin (PCNT) Gene Cause Primordial Dwarfism. *Science* **319**, 816–819 (2008).

113. Griffith, E. *et al.* Mutations in pericentrin cause Seckel syndrome with defective ATR-dependent DNA damage signaling. *Nat. Genet.* **40**, 232–236 (2008).

114. Piane, M. *et al.* Majewski osteodysplastic primordial dwarfism type II (MOPD II) syndrome previously diagnosed as Seckel syndrome: Report of a novel mutation of the PCNT gene. *Am. J. Med. Genet. A.* **149A**, 2452–2456 (2009).

115. van der Slot, A. J. *et al.* Identification of PLOD2 as Telopeptide Lysyl Hydroxylase, an Important Enzyme in Fibrosis*. *J. Biol. Chem.* **278**, 40967–40972 (2003).

116. Ha-Vinh, R. *et al.* Phenotypic and molecular characterization of Bruck syndrome (osteogenesis imperfecta with contractures of the large joints) caused by a recessive mutation in PLOD2. *Am. J. Med. Genet. A.* **131A**, 115–120 (2004).

117. Puig-Hervás, M. T. *et al.* Mutations in PLOD2 cause autosomal-recessive connective tissue disorders within the Bruck syndrome—Osteogenesis imperfecta phenotypic spectrum. *Hum. Mutat.* **33**, 1444–1449 (2012).

118. Tartaglia, M. *et al.* Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nat. Genet.* **29**, 465–468 (2001).

119. Maheshwari, M. *et al.* PTPN11 Mutations in Noonan syndrome type I: detection of recurrent mutations in exons 3 and 13. *Hum. Mutat.* **20**, 298–304 (2002).

120. Kosaki, K. *et al.* PTPN11 (Protein-Tyrosine Phosphatase, Nonreceptor-Type 11) Mutations in Seven Japanese Patients with Noonan Syndrome. *J. Clin. Endocrinol. Metab.* **87**, 3529–3533 (2002).

121. Deardorff, M. A. *et al.* RAD21 Mutations Cause a Human Cohesinopathy. *Am. J. Hum. Genet.* **90**, 1014–1027 (2012).

122. Kruszka, P. *et al.* Cohesin complex-associated holoprosencephaly. *Brain* **142**, 2631–2643 (2019).

123. Goel, H. & Parasivam, G. Another case of holoprosencephaly associated with RAD21 loss-of-function variant. *Brain* **143**, e64 (2020).

124. Pandit, B. *et al.* Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nat. Genet.* **39**, 1007–1012 (2007).

125. Razzaque, M. A. *et al.* Germline gain-of-function mutations in RAF1 cause Noonan syndrome. *Nat. Genet.* **39**, 1013–1017 (2007).

126. Lindor, N. M. *et al.* Rothmund-Thomson syndrome due to RECQ4 helicase mutations: Report and clinical and molecular comparisons with Bloom syndrome and Werner syndrome. *Am. J. Med. Genet.* **90**, 223–228 (2000).

127. Beghini, A., Castorina, P., Roversi, G., Modiano, P. & Larizza, L. RNA processing defects of the helicase gene RECQL4 in a compound heterozygous Rothmund–Thomson patient. *Am. J. Med. Genet. A.* **120A**, 395–399 (2003).

128. Wang, L. L. *et al.* Association Between Osteosarcoma and Deleterious Mutations in the RECQL4 Gene in Rothmund–Thomson Syndrome. *JNCI J. Natl. Cancer Inst.* **95**, 669–674 (2003).

129. Aoki, Y. *et al.* Gain-of-Function Mutations in RIT1 Cause Noonan Syndrome, a RAS/MAPK Pathway Syndrome. *Am. J. Hum. Genet.* **93**, 173–180 (2013).

130. Bertola, D. R. *et al.* Further evidence of the importance of RIT1 in Noonan syndrome. *Am. J. Med. Genet. A.* **164**, 2952–2957 (2014).

131. Gos, M. *et al.* Contribution of RIT1 mutations to the pathogenesis of Noonan syndrome: Four new cases and further evidence of heterogeneity. *Am. J. Med. Genet. A.* **164**, 2310–2316 (2014).

132. Afzal, A. R. *et al.* Recessive Robinow syndrome, allelic to dominant brachydactyly type B, is caused by mutation of ROR2. *Nat. Genet.* **25**, 419–422 (2000).

133. van Bokhoven, H. *et al.* Mutation of the gene encoding the ROR2 tyrosine kinase causes autosomal recessive Robinow syndrome. *Nat. Genet.* **25**, 423–426 (2000).

134. Tufan, F. *et al.* Clinical and molecular characterization of two adults with autosomal recessive Robinow syndrome. *Am. J. Med. Genet. A.* **136A**, 185–189 (2005).

135. Hästbacka, J., Salonen, R., Laurila, P., Chapelle, A. de la & Kaitila, I. Prenatal diagnosis of diastrophic dysplasia with polymorphic DNA markers. *J. Med. Genet.* **30**, 265–268 (1993).

136. Rossi, A. & Superti-Furga, A. Mutations in the diastrophic dysplasia sulfate transporter (DTDST) gene (SLC26A2): 22 novel mutations, mutation review, associated skeletal phenotypes, and diagnostic relevance. *Hum. Mutat.* **17**, 159–171 (2001).

137. Barreda-Bonis, A. C. *et al.* Multiple SLC26A2 mutations occurring in a three-generational family. *Eur. J. Med. Genet.* **61**, 24–28 (2018).

138. Le Goff, C. *et al.* Mutations at a single codon in Mad homology 2 domain of SMAD4 cause Myhre syndrome. *Nat. Genet.* **44**, 85–88 (2012).

139. Caputo, V. *et al.* A Restricted Spectrum of Mutations in the SMAD4 Tumor-Suppressor Gene Underlies Myhre Syndrome. *Am. J. Hum. Genet.* **90**, 161–169 (2012).

140. Lindor, N. M., Gunawardena, S. R. & Thibodeau, S. N. Mutations of SMAD4 account for both LAPS and Myhre syndromes. *Am. J. Med. Genet. A.* **158A**, 1520–1521 (2012).

141. Hood, R. L. *et al.* Mutations in SRCAP, Encoding SNF2-Related CREBBP Activator Protein, Cause Floating-Harbor Syndrome. *Am. J. Hum. Genet.* **90**, 308–313 (2012).

142. Goff, C. L. *et al.* Not All Floating-Harbor Syndrome Cases are Due to Mutations in Exon 34 of SRCAP. *Hum. Mutat.* **34**, 88–92 (2013).

143. Yu, C.-E. *et al.* Positional Cloning of the Werner's Syndrome Gene. *Science* **272**, 258–262 (1996).

144. Goto, M. *et al.* Analysis of helicase gene mutations in Japanese Werner's syndrome patients. *Hum. Genet.* **99**, 191–193 (1997).

145. Yu, C.-E. *et al.* Mutations in the Consensus Helicase Domains of the Werner Syndrome Gene. *Am J Hum Genet* 12 (1997).

146. Simon, D. B. *et al.* Genetic heterogeneity of Barter's syndrome revealed by mutations in the K+ channel, ROMK. *Nat. Genet.* **14**, 152–156 (1996).

147. Károlyi, L. *et al.* Mutations in the Gene Encoding the Inwardly-Rectifying Renal Potassium Channel, ROMK, Cause the Antenatal Variant of Bartter Syndrome: Evidence for Genetic Heterogeneity. *Hum. Mol. Genet.* **6**, 17–26 (1997).

148. Quaggin, S. E., Payne, J. A., Forbush, B. & Igarashi, P. Localization of the renal Na−K−Cl cotransporter gene (Slc 12a1) on mouse Chromosome 2. *Mamm. Genome* **6**, 557–558 (1995).

149. Simon, D. B. *et al.* Bartter's syndrome, hypokalaemic alkalosis with hypercalciuria, is caused by mutations in the Na–K–2Cl cotransporter NKCC2. *Nat. Genet.* **13**, 183–188 (1996).

150. Simon, D. B. *et al.* Gitelman's variant of Barter's syndrome, inherited hypokalaemic alkalosis, is caused by mutations in the thiazide-sensitive Na–Cl cotransporter. *Nat. Genet.* **12**, 24–30 (1996).

151. Glaudemans, B. *et al.* Novel NCC mutants and functional analysis in a new cohort of patients with Gitelman syndrome. *Eur. J. Hum. Genet.* **20**, 263–270 (2012).

152. Wilson, F. H. *et al.* Human Hypertension Caused by Mutations in WNK Kinases. *Science* **293**, 1107–1112 (2001).

153. Lalioti, M. D. *et al.* Wnk4 controls blood pressure and potassium homeostasis via regulation of mass and activity of the distal convoluted tubule. *Nat. Genet.* **38**, 1124–1132 (2006).

154. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat. Genet.* **39**, 207–211 (2007).

155. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43**, 1066–1073 (2011).

156. Fowler, E. V. *et al.* TNFα and IL10 SNPs act together to predict disease behaviour in Crohn's disease. *J. Med. Genet.* **42**, 523–528 (2005).

157. Gasche, C. *et al.* Novel Variants of the IL-10 Receptor 1 Affect Inhibition of Monocyte TNF-α Production. *J. Immunol.* **170**, 5578–5582 (2003).

158. Mao, H. *et al.* Exome sequencing identifies novel compound heterozygous mutations of IL-10 receptor 1 in neonatal-onset Crohn's disease. *Genes Immun.* **13**, 437–442 (2012).

159. Glocker, E.-O. *et al.* Inflammatory Bowel Disease and Mutations Affecting the Interleukin-10 Receptor. *http://dx.doi.org/10.1056/NEJMoa0907206* https://www.nejm.org/doi/10.1056/NEJMoa0907206 (2009) doi:10.1056/NEJMoa0907206.

160. Begue, B. *et al.* Defective IL10 Signaling Defining a Subgroup of Patients With Inflammatory Bowel Disease. *Off. J. Am. Coll. Gastroenterol. ACG* **106**, 1544–1555 (2011).

161. Duerr, R. H. *et al.* A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science* **314**, 1461–1463 (2006).

162. Libioulle, C. *et al.* Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4. *PLOS Genet.* **3**, e58 (2007).

163. Glas, J. *et al.* rs1004819 Is the Main Disease-Associated IL23R Variant in German

Crohn's Disease Patients: Combined Analysis of IL23R, CARD15, and OCTN1/2 Variants. *PLOS ONE* **2**, e819 (2007).

164. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).

165. Craddock, N. *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).

166. Prescott, N. J. *et al.* Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. *Hum. Mol. Genet.* **19**, 1828–1839 (2010).

167. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).

168. Hugot, J.-P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).

169. Ellinghaus, D. *et al.* Association Between Variants of PRDM1 and NDP52 and Crohn's Disease, Based on Exome Sequencing and Functional Studies. *Gastroenterology* **145**, 339–347 (2013).

170. Diaz-Gallo, L.-M. *et al.* Differential association of two PTPN22 coding variants with Crohn's disease and ulcerative colitis. *Inflamm. Bowel Dis.* **17**, 2287–2294 (2011).

171. Fowler, E. V. *et al.* ATG16L1 T300A Shows Strong Associations With Disease Subgroups in a Large Australian IBD Population: Further Support for Significant Disease Heterogeneity. *Off. J. Am. Coll. Gastroenterol. ACG* **103**, 2519–2526 (2008).

172. Fisher, S. A. *et al.* Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. *Nat. Genet.* **40**, 710–712 (2008).

173. Beaudoin, M. *et al.* Deep Resequencing of GWAS Loci Identifies Rare Variants in CARD9, IL23R and RNF186 That Are Associated with Ulcerative Colitis. *PLOS Genet.* **9**, e1003723 (2013).

174. Rivas, M. A. *et al.* A protein-truncating R179X variant in RNF186 confers protection

against ulcerative colitis. *Nat. Commun.* **7**, 12342 (2016).

175. Reis, A. F. *et al.* Association of a variant in exon 31 of the sulfonylurea receptor 1 (SUR1) gene with type 2 diabetes mellitus in French Caucasians. *Hum. Genet.* **107**, 138–144 (2000).

176. Borowiec, M. *et al.* Mutations at the BLK locus linked to maturity onset diabetes of the young and β-cell dysfunction. *Proc. Natl. Acad. Sci.* **106**, 14460–14465 (2009).

177. Bengtsson-Ellmark, S. H. *et al.* Association between a polymorphism in the carboxyl ester lipase gene and serum cholesterol profile. *Eur. J. Hum. Genet.* **12**, 627–632 (2004).

178. Ræder, H. *et al.* Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nat. Genet.* **38**, 54–62 (2006).

179. Harding, H. P. *et al.* Diabetes Mellitus and Exocrine Pancreatic Dysfunction in Perk−/− Mice Reveals a Role for Translational Control in Secretory Cell Survival. *Mol. Cell* **7**, 1153–1163 (2001).

180. Brickwood, S. *et al.* Wolcott-Rallison syndrome: pathogenic insights into neonatal diabetes from new mutation and expression studies of EIF2AK3. *J. Med. Genet.* **40**, 685–689 (2003).

181. Durocher, F. *et al.* A novel mutation in the EIF2AK3 gene with variable expressivity in two patients with Wolcott–Rallison syndrome. *Clin. Genet.* **70**, 34–38 (2006).

182. Shaw-Smith, C. *et al.* GATA4 Mutations Are a Cause of Neonatal and Childhood-Onset Diabetes. *Diabetes* **63**, 2888–2894 (2014).

183. Yorifuji, T. *et al.* Dominantly inherited diabetes mellitus caused by GATA6 haploinsufficiency: variable intrafamilial presentation. *J. Med. Genet.* **49**, 642–643 (2012).

184. Franco, E. D. *et al.* GATA6 Mutations Cause a Broad Phenotypic Spectrum of Diabetes From Pancreatic Agenesis to Adult-Onset Diabetes Without Exocrine Insufficiency. *Diabetes* **62**, 993–997 (2013).

185. Froguel, P. *et al.* Familial Hyperglycemia Due to Mutations in Glucokinase -- Definition of

a Subtype of Diabetes Mellitus. *http://dx.doi.org/10.1056/NEJM199303113281005* https://www.nejm.org/doi/10.1056/NEJM199303113281005 (1993) doi:10.1056/NEJM199303113281005.

186. Senée, V. *et al.* Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. *Nat. Genet.* **38**, 682–687 (2006).

187. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young (MODY3). *Nature* **384**, 455–458 (1996).

188. Vaxillaire, M. *et al.* Identification of Nine Novel Mutations in the Hepatocyte Nuclear Factor 1 Alpha Gene Associated with Maturity-Onset Diabetes of the Young (MODY3). *Hum. Mol. Genet.* **6**, 583–586 (1997).

189. Horikawa, Y. *et al.* Mutation in hepatocyte nuclear factor–1β gene (TCF2) associated with MODY. *Nat. Genet.* **17**, 384–385 (1997).

190. Lindner, T. H. *et al.* A Novel Syndrome of Diabetes Mellitus, Renal Dysfunction and Genital Malformation Associated with a Partial Deletion of the Pseudo-POU Domain of Hepatocyte Nuclear Factor-1β. *Hum. Mol. Genet.* **8**, 2001–2008 (1999).

191. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-4α gene in maturity-onset diabetes of the young (MODY1). *Nature* **384**, 458–460 (1996).

192. Stoffel, M. & Duncan, S. A. The maturity-onset diabetes of the young (MODY1) transcription factor HNF4α regulates expression of genes required for glucose transport and metabolism. *Proc. Natl. Acad. Sci.* **94**, 13209–13214 (1997).

193. Poulton, C. J. *et al.* Microcephaly with Simplified Gyration, Epilepsy, and Infantile Diabetes Linked to Inappropriate Apoptosis of Neural Progenitors. *Am. J. Hum. Genet.* **89**, 265–276 (2011).

194. Abdel-Salam, G. M. H. *et al.* A homozygous IER3IP1 mutation causes microcephaly with simplified gyral pattern, epilepsy, and permanent neonatal diabetes syndrome (MEDS). *Am. J. Med. Genet. A.* **158A**, 2788–2796 (2012).

195. Shalev, S. A. *et al.* Microcephaly, epilepsy, and neonatal diabetes due to compound heterozygous mutations in IER3IP1: insights into the natural history of a rare disorder. *Pediatr. Diabetes* **15**, 252–256 (2014).

196. Støy, J. *et al.* Insulin gene mutations as a cause of permanent neonatal diabetes. *Proc. Natl. Acad. Sci.* **104**, 15040–15044 (2007).

197. Hani, E. H. *et al.* Missense mutations in the pancreatic islet beta cell inwardly rectifying K+ channel gene (KIR6.2/BIR): a meta-analysis suggests a role in the polygenic basis of Type II diabetes mellitus in Caucasians. *Diabetologia* **41**, 1511–1515 (1998).

198. Gloyn, A. L. *et al.* Activating Mutations in the Gene Encoding the ATP-Sensitive Potassium-Channel Subunit Kir6.2 and Permanent Neonatal Diabetes. *http://dx.doi.org/10.1056/NEJMoa032922* https://www.nejm.org/doi/10.1056/NEJMoa032922 (2004) doi:10.1056/NEJMoa032922.

199. Neve, B. *et al.* Role of transcription factor KLF11 and its diabetes-associated gene variants in pancreatic beta cell function. *Proc. Natl. Acad. Sci.* **102**, 4807–4812 (2005).

200. Cao, H. & Hegele, R. A. Nuclear lamin A/C R482Q mutation in Canadian kindreds with Dunnigan-type familial partial lipodystrophy. *Hum. Mol. Genet.* **9**, 109–112 (2000).

201. Malecki, M. T. *et al.* Mutations in NEUROD1 are associated with the development of type 2 diabetes mellitus. *Nat. Genet.* **23**, 323–328 (1999).

202. Gradwohl, G., Dierich, A., LeMeur, M. & Guillemot, F. neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc. Natl. Acad. Sci.* **97**, 1607–1611 (2000).

203. Rubio-Cabezas, O. *et al.* Permanent Neonatal Diabetes and Enteric Anendocrinosis Associated With Biallelic Mutations in NEUROG3. *Diabetes* **60**, 1349–1353 (2011).

204. Pinney, S. E. *et al.* Neonatal Diabetes and Congenital Malabsorptive Diarrhea Attributable to a Novel Mutation in the Human Neurogenin-3 Gene Coding Sequence. *J. Clin. Endocrinol. Metab.* **96**, 1960–1965 (2011).

205. Shimajiri, Y. *et al.* A Missense Mutation of Pax4 Gene (R121W) Is Associated With Type 2 Diabetes in Japanese. *Diabetes* **50**, 2864–2869 (2001).

206. Mauvais-Jarvis, F. *et al.* PAX4 gene variations predispose to ketosis-prone diabetes. *Hum. Mol. Genet.* **13**, 3151–3159 (2004).

207. Plengvidhya, N. *et al.* PAX4 Mutations in Thais with Maturity Onset Diabetes of the Young. *J. Clin. Endocrinol. Metab.* **92**, 2821–2826 (2007).

208. Staffers, D. A., Ferrer, J., Clarke, W. L. & Habener, J. F. Early-onset type-ll diabetes mellitus (MODY4) linked to IPF1. *Nat. Genet.* **17**, 138–139 (1997).

209. Macfarlane, W. M. *et al.* Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. *J. Clin. Invest.* **104**, R33–R39 (1999).

210. Hani, E. H. *et al.* Defective mutations in the insulin promoter factor-1 (*IPF-1*) gene in late-onset type 2 diabetes mellitus. *J. Clin. Invest.* **104**, R41–R48 (1999).

211. 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).

212. Savage, D. B. *et al.* Digenic inheritance of severe insulin resistance in a human pedigree. *Nat. Genet.* **31**, 379–384 (2002).

213. Sellick, G. S. *et al.* Mutations in PTF1A cause pancreatic and cerebellar agenesis. *Nat. Genet.* **36**, 1301–1305 (2004).

214. Smith, S. B. *et al.* Rfx6 directs islet formation and insulin production in mice and humans. *Nature* **463**, 775–780 (2010).

215. Sansbury, F. H. *et al.* Biallelic RFX6 mutations can cause childhood as well as neonatal onset diabetes mellitus. *Eur. J. Hum. Genet.* **23**, 1744–1748 (2015).

216. Labay, V. *et al.* Mutations in SLC19A2 cause thiamine-responsive megaloblastic anaemia associated with diabetes mellitus and deafness. *Nat. Genet.* **22**, 300–304 (1999).

217. Oishi, K. *et al.* Targeted disruption of Slc19a2, the gene encoding the high-affinity

thiamin transporter Thtr-1, causes diabetes mellitus, sensorineural deafness and megaloblastosis in mice. *Hum. Mol. Genet.* **11**, 2951–2960 (2002).

218. Shaw-Smith, C. *et al.* Recessive SLC19A2 mutations are a cause of neonatal diabetes mellitus in thiamine-responsive megaloblastic anaemia. *Pediatr. Diabetes* **13**, 314–321 (2012).

219. Laukkanen, O. *et al.* Polymorphisms in the SLC2A2 (GLUT2) Gene Are Associated With the Conversion From Impaired Glucose Tolerance to Type 2 Diabetes : The Finnish Diabetes Prevention Study. *Diabetes* **54**, 2256–2260 (2005).

220. Sansbury, F. H. *et al.* SLC2A2 mutations can cause neonatal diabetes, suggesting GLUT2 may have a role in human insulin secretion. *Diabetologia* **55**, 2381–2385 (2012).

221. Strom, T. M. *et al.* Diabetes Insipidus, Diabetes Mellitus, Optic Atrophy and Deafness (DIDMOAD) Caused by Mutations in a Novel Gene (Wolframin) Coding for a Predicted Transmembrane Protein. *Hum. Mol. Genet.* **7**, 2021–2028 (1998).

222. Hardy, C. *et al.* Clinical and Molecular Genetic Analysis of 19 Wolfram Syndrome Kindreds Demonstrating a Wide Spectrum of Mutations in WFS1. *Am. J. Hum. Genet.* **65**, 1279–1290 (1999).

223. Khanim, F., Kirk, J., Latif, F. & Barrett, T. G. WFS1/wolframin mutations, Wolfram syndrome, and associated diseases. *Hum. Mutat.* **17**, 357–367 (2001).

224. Mackay, D. J. G. *et al.* Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. *Nat. Genet.* **40**, 949–951 (2008).

225. Boonen, S. E. *et al.* Transient Neonatal Diabetes, ZFP57, and Hypomethylation of Multiple Imprinted Loci: A detailed follow-up. *Diabetes Care* **36**, 505–512 (2013).

226. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). *Online Mendelian Inheritance in Man, OMIM®*. (2021).

227. Deardorff, M. A. *et al.* Mutations in Cohesin Complex Members SMC3 and SMC1A

Cause a Mild Variant of Cornelia de Lange Syndrome with Predominant Mental Retardation. *Am. J. Hum. Genet.* **80**, 485–494 (2007).

228. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).

229. Bycroft, C. *et al. Genome-wide genetic data on ~500,000 UK Biobank participants*. 166298 https://www.biorxiv.org/content/10.1101/166298v1 (2017) doi:10.1101/166298.

230. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

231. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

232. Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).

233. Galinsky, K. J., Loh, P.-R., Mallick, S., Patterson, N. J. & Price, A. L. Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. *Am. J. Hum. Genet.* **99**, 1130–1139 (2016).

234. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).

235. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).

Acknowledgements

We thank Alkes Price, Alex Bloemendal, Benjamin Neale, Bogdan Pasanuic, Sasha (Alexander) Gusev, and Matt Warman for their helpful discussions. This research was supported by NIH grants HG010372, R35GM127131, R01HG010372, and R01MH101244. N.J.C was supported by NIH training grant T32GM74897. UK Biobank was accessed under projects 14048 and 10438. TOPMed data were used under dbGaP project 28674.