https://doi.org/10.1186/s12864-024-10971-2

BMC Genomics

RESEARCH

Open Access

Leveraging large-scale multi-omics evidences to identify therapeutic targets from genome-wide association studies

Samuel Lessard[1], Michael Chao[1], Kadri Reis[2], FinnGen, Estonian Biobank Research Team, Mathieu Beauvais[3], Deepak K. Rajpal[4,5], Jennifer Sloane[6], Priit Palta[2], Katherine Klinger[7], Emanuele de Rinaldis[1], Khader Shameer[1] and Clément Chatelain[1*]

Abstract

Background Therapeutic targets supported by genetic evidence from genome-wide association studies (GWAS) show higher probability of success in clinical trials. GWAS is a powerful approach to identify links between genetic variants and phenotypic variation; however, identifying the genes driving associations identified in GWAS remains challenging. Integration of molecular quantitative trait loci (molQTL) such as expression QTL (eQTL) using mendelian randomization (MR) and colocalization analyses can help with the identification of causal genes. Careful interpretation remains warranted because eQTL can affect the expression of multiple genes within the same locus.

Methods We used a combination of genomic features that include variant annotation, activity-by-contact maps, MR, and colocalization with molQTL to prioritize causal genes across 4,611 disease GWAS and meta-analyses from biobank studies, namely FinnGen, Estonian Biobank and UK Biobank.

Results Genes identified using this approach are enriched for gold standard causal genes and capture known biological links between disease genetics and biology. In addition, we find that eQTL colocalizing with GWAS are statistically enriched for corresponding disease-relevant tissues. We show that predicted directionality from MR is generally consistent with matched drug mechanism of actions (> 85% for approved drugs). Compared to the nearest gene mapping method, genes supported by multi-omics evidences displayed higher enrichment in approved therapeutic targets (risk ratio 1.75 vs. 2.58 for genes with the highest level of support). Finally, using this approach, we detected anassociation between the IL6 receptor signal transduction gene *IL6ST* and polymyalgia rheumatica, an indication for which sarilumab, a monoclonal antibody against IL-6, has been recently approved.

Conclusions Combining variant annotation, activity-by-contact maps, and molQTL increases performance to identify causal genes, while informing on directionality which can be translated to successful target identification and drug development.

Deepak K. Rajpal was an employee of Sanofi US at the time of study.

*Correspondence:
Clément Chatelain
Clement.chatelain@sanofi.com
Full list of author information is available at the end of the article

Keywords Genome-wide association study, Molecular quantitative trait loci, Causal inference, Therapeutic targets, Interleukin 6, Polymyalgia rheumatica, Mendelian randomization

Background

Genome-wide associations studies (GWAS) have been successful in identifying genes associated with traits, diseases, and molecular phenotypes [1, 2]. Discoveries from GWAS have increased substantially over the years due to low cost of genomic profiling technologies, an increased number of studies, larger cohorts, and meta-analyses, as well as the formation of deeply phenotyped datasets [3]. The latter include large-scale biobank projects such as UK Biobank (UKB) [4, 5], Estonian Biobank [6], and FinnGen [7]. As an example, the UK Biobank alone has contributed to over 3,200 publications (https://www.ukbiobank.ac.uk/enable-your-research/publications), and the FinnGen project is set to increase the number of discoveries emerging from rare variants enriched in the Finnish population [7]. Similarly, the Estonian Biobank, with its extensive dataset, has enhanced rare and low-frequency genetic variation discoveries [8–10].

Discoveries from genetic studies provide a highly valuable resource for drug discoveries. For example, therapeutic targets with genetic support are >2 times more likely to succeed in clinical trials [11, 12]. A notable example is the association between a loss-of-function missense variant in *IL23R* gene and Crohn's disease, suggesting that IL-23 blockage could be beneficial [13–16]. Drugs targeting the IL-23 receptor including Ustekinumab and Risankizumab have recently been approved by the FDA for the treatment of Crohn's disease following successful clinical trials [17–19]. Other notable examples of targets supported by GWAS include *IL6R* for rheumatoid arthritis (Sarilumab, Tocilizumab) and *HMGCR* for high levels of low-density lipoprotein (statins) [20, 21].

While these examples clearly show that genetic disease associations provide important information for drug development, it remains a challenge to accurately assign causal genes driving disease risk from GWAS as most variants identified in GWAS fall in non-coding regions of the genome [22–24]. While it's been observed that the nearest gene often is the causal gene, this is not a guarantee as genetic variants can influence traits over large genomic distances [25]. In addition, this observation may be biased towards genes that have been well-characterized because they fall at the center of genetic association signals [26].

Several approaches have been used to predict causal genes, including selecting the nearest gene, variant pathogenicity predictions, epigenetic interactions, and integration of molecular quantitative trait loci (molQTL) such as expression QTL (eQTL). Mendelian randomization (MR) integrating GWAS and molQTL can help identify causal relationships while informing on directionality but may be confounded due to linkage disequilibrium (LD) [27–29]. On the other hand, colocalization approaches can be used to detect whether molQTL and GWAS signals share a common causal variant in a specific locus [30, 31]. While colocalization approaches can link genetic variation to changes in gene expression in specific tissue or cell-type contexts, they also tend to be pleiotropic and often impact the expression of multiple genes within the same locus [26, 32, 33]. They can also impact expression across multiple tissues and cell types, decreasing their utility to identify pathogenic cell types [32, 34, 35]. In addition, a large fraction of GWAS loci don't show eQTL signals, potentially due to the unavailability of data for relevant cell types or specific biological contexts or variants affecting disease risk due to different mechanisms such as splicing [32, 36, 37]. Despite these challenges, eQTL has successfully been used to identify causal genes [38, 39]. In addition, recent prioritization approaches such as the Locus to Gene (L2G) scores from Open Targets have shown that incorporating molecular trait information does increase performance to identify relevant genes [26].

Here, we sought to use currently available eQTL information to identify disease relevant genes in the context of drug discovery. We first derived a simple approach to prioritize causal genes based on MR [40], eQTL colocalization [31], activity-by-contact (ABC) enhancer-promoter interactions [41], and variant annotations [42]. We used this combinatorial approach as a way to mitigate the pleiotropic effect of eQTL while retaining important information about directionality. We show that this approach enriches for gold standard genes [26] and captures known target biology. In addition, genes prioritized by this approach are enriched for drug targets with successful clinical trials, and directionality inferred by MR or coding variants recapitulate drug mechanisms of action (MoA). Finally, we show that this approach can be used to identify drug indication expansion opportunities using genes related to the IL6 receptor as a case study and identify an association between *IL6ST* and polymyalgia rheumatica.

Methods

Estonian Biobank GWAS

The Estonian Biobank (EstBB) is a population-based biobank with 200k participants. The 198k data freeze was used for the analyses described here. All biobank participants have signed a broad informed consent form.

All EstBB participants have been genotyped at the Core Genotyping Lab of the Institute of Genomics, University of Tartu, using Illumina Global Screening Array v1.0 and v2.0. Samples were genotyped and PLINK format files were created using Illumina GenomeStudio v2.0.4. Individuals were excluded from the analysis if their call-rate was <95% or if sex defined based on heterozygosity of X chromosome did not match sex in phenotype data. Before phasing and imputation, variants were filtered by call-rate<95%, HWE p value<1e-4 (autosomal variants only), and minor allele frequency<1%. Variant positions were updated to b37 and all variants were changed to be from TOP strand using GSAMD-24v1-0_20011747_A1-b37.strand.RefAlt.zip files from https://www.well.ox.ac.uk/~wrayner/strand/ webpage. Chip data pre-phasing was done using Eagle v2.3 software [43] (number of conditioning haplotypes Eagle uses when phasing each sample was set to:–Kpbwt=20000) and imputation was done using Beagle v.28 Sep18.7932 [44] with effective population size ne=20,000. Population specific imputation reference panel of 2297 WGS samples was used [44].

FinnGen

The FinnGen study (https://www.finngen.fi/en) was described previously [7]. The study is a public-private research project that combines genetic and healthcare data of over 500,000 Finns. The objective of the FinnGen study is to identify novel medical and therapeutical insight into human diseases. It is a pre-competitive partnership of Finnish biobanks, universities and university hospitals, international pharmaceutical industry partners, and Finnish biobank cooperative (FINBB). A full list of FinnGen partners is published here: https://www.finngen.fi/en/partners.

Disease GWAS processing

We retrieved GWAS results from FinnGen release 10 (R10), UK Biobank pan-ancestry analysis [45], and a meta-analyses between FinnGen, UK Biobank, and Estonian biobank. For simplicity, we use the term GWAS to refer to both single study GWAS and meta-analyses throughout the manuscript. In total, we included 4,611 GWAS with at least one variant with $P<1\times10^{-6}$. When appropriate, we lifted over variants from hg38 to hg19 using the liftOver R package [46]. Variant with a minor allele frequency (MAF)<0.0001 were excluded from the analysis. For each GWAS, we considered genes located within 250 kb of a variant with $P<1\times10^{-6}$ for further analysis. For gold standard and clinical trial enrichment analyses (described below), only genome-wide significant loci were included ($P<5\times10^{-8}$). We excluded the human leukocyte antigen (HLA) region in all analyses.

Disease EFO mapping

In order to perform semantic integration of genetic data and clinical trial data, the ontological system Experimental Factor Ontology (EFO) was used. We used the EFO to map traits to their corresponding EFO categories and when multiple EFO terms could be mapped to the same trait, we assigned the trait to each possible term. We used the EFO version 3.52.0 (https://www.ebi.ac.uk/efo/).

Variant annotation

We used variant effect predictor (VEP v102) [42] to annotate the impact of variants with the following options: --everything --offline --check_existing --distance 250,000. Coding variants were defined as those impacting protein coding transcript annotated as missense variant or predicted to have "high" impact, including stop gain, splice-site, and frameshift variants. We also retrieved pathogenicity predictions for missense variants from ProtVar [47], considering conservation, structure stability predictions, and EVE [48] and ESM1b scores [49]. We defined pathogenic variants as those with "high" impact, predicted to be pathogenic, destabilizing, or in a conserved region. In addition, we linked non-coding variants to genes using activity-by-contact (ABC) maps [41]. ABC scores represent the contribution of an enhancer to the regulation of genes, measured by multiplying the estimates of enhancer activity and three-dimensional contact frequencies between enhancers and promoters. ABCmax refers to variant-gene pairs with the highest ABC score. We also retrieved disease mutations from the Human Gene Mutation Database (HGMD) (licensed from Qiagen, Maryland) [50]. We annotated all variants with $P<1\times10^{-6}$ and within 5 orders of magnitude of the lead variant at the locus.

Mendelian randomization & colocalization

We performed transcriptome wide MR using the R package TwoSampleMR [40]. When more than one instrument was present, we used the inverse variant weighted approach, otherwise we used the Wald Ratio approach. We considered the following exposures: protein quantitative trait loci (pQTL) from Sun et al. [51], and eQTL from Blueprint [52], eQTLGen [53] and other datasets from the EBI eQTL catalogue [53–77]. In total, 110 molQTL from 26 studies were included. For each of those studies, we excluded variants with a MAF<1%. We clumped variants using PLINK [78] using the options –clump-p1 1 –clump-p2 1 –clump-r2 0.01 – clump-kb 10,000 and using the European ancestry subset of the 1000 Genomes Project phase 3 data as reference [79]. We included all genes within 250 kb of a GWAS variant with $P < 1 \times 10^{-6}$. For each QTL, independent variants with $P < 1 \times 10^{-4}$ were used as instruments. For genes with significant MR results (false discovery rate < 0.05), we also performed colocalization analysis using COLOC [31] in order to account for pleiotropy due to linkage, using a region of 250 kb each side of the local lead GWAS variant. Harmonization between molQTL and GWAS datasets was performed using the harmonise_data function in the TwoSampleMR package [40]. Only autosomes were included in this analysis.

Causal gene prioritization

We prioritized genes as putatively causal using a combination of evidence including MR, colocalization H4 posterior probabilities (PP) with molQTL, presence of an associated pathogenic variant or other coding variants, distance to lead variant, and enhancer-promoter ABC scores [41]. Specifically, we ranked genes as follow:

Rank	Criteria
Very High	Lead pathogenic variant; Or Colocalization (H4 PP > 80%) with molQTL of the target gene in > 2 dataset; and maximum ABC score for a regulatory element overlapping the lead variant
High	Lead coding variant; Or Associated ($P < 1 \times 10^{-6}$) pathogenic variant; Or Colocalization (H4 PP > 80%) with molQTL of the target gene in > 2 dataset and maximum ABC score for an associated variant overlapping a regulatory element ($P < 1 \times 10^{-6}$) Or Colocalization (H4 PP > 80%) with molQTL of the target gene in one dataset; and maximum ABC score for a regulatory element overlapping the lead variant
Moderate	Colocalization with molQTL of the target gene (H4 PP > 80%) Or Significant MR with genome-wide protein QTL (q-value < 0.05) Or Maximum ABC score for an element overlapping the lead variant Or Associated ($P < 1 \times 10^{-6}$) coding variant
Weak	Nearest gene to the lead variant Or Maximum ABC score for an element overlapping an associated variant ($P < 1 \times 10^{-6}$) Or ABC link (any score) between an element overlapping the lead variant and target gene
Very weak	Significant MR with eQTL Or ABC link (any score) between an element overlapping the lead variant and target gene

For a given locus, we then prioritized the best gene(s) as the one with the highest rank. In case of ties, we prioritized the nearest gene to lead variant if it is within the set of genes with highest scores, otherwise all highest ranked genes were prioritized equally.

Enrichment of gold standard genes

We retrieved GWAS causal gene gold standards supported by functional experiments or observations or expert curation from Open Targets (version 191108) [26, 80]. We linked the current analysis with the gold standard gene list using Ensembl gene identifiers and EFO codes. That is, for a given gene-disease pair in the current analysis, we consider it a gold standard association if the gene and GWAS EFO code are present in the Open Targets gold standard gene-disease set. For each indication, we filtered out genes not represented in loci where a gold standard gene is located. We calculated the enrichment of gold standard genes in prioritized genes by different features or rankings as described above using Fisher exact tests. In addition, we calculated the precision (number of prioritized genes that are gold standards over all prioritized genes), recall (number of prioritized genes that are gold standards over the total number of gold standard genes), and F1 scores for each feature.

Single gene colocalizing cell-type molQTL enrichment

To identify enriched cell types with colocalizing molQTL at single genes, we calculated the ratio of indications for which this gene is prioritized to be causal by a given molQTL dataset (H4 PP > 80%) over the total number of prioritized indications (as defined by unique EFO) for that gene. We collapsed GWAS by corresponding EFO code so that a gene was only counted once per indication (and not multiple times for GWAS of the same disease). We then compared this ratio to the fraction of prioritized indications via colocalization of the same eQTL dataset over all prioritized indications genome wide. In other words, we are looking for genes that show an overrepresentation of colocalizing eQTL cell types across all associated indications compared to the genome-wide distribution. This corresponds to the following contingency table:

$$\begin{matrix} \sum_i C_{iJK} & \sum_i \sum_{k \neq K} C_{iJK} \\ \sum_i \sum_{j \neq J} C_{ijK} & \sum_i \sum_{k \neq K} \sum_{j \neq J} C_{ijk} \end{matrix}$$

Where C_{ijk}=1 if disease i colocalize with prioritized gene j in tissue k and 0 if not. P-values and odds ratios were calculated using Fisher exact tests. False discovery rate (FDR) adjusted P-values < 0.05 were considered significant.

Enrichment of disease categories for single genes

To identify enrichment of disease categories for single genes, we calculated the ratio of the number of GWAS where the genes is prioritized for a given EFO category over the total number of prioritized GWAS for that gene. We then compared this ratio to the genome-wide ratio of GWAS for this EFO category over the total number of tested GWAS. This corresponds to the following contingency table:

$$\begin{array}{cc} \sum_{i} D_{iJC} & \sum_{i} \sum_{c\neq C} D_{iJc} \\ \sum_{i} \sum_{j\neq J} D_{ijC} & \sum_{i} \sum_{c\neq C} \sum_{j\neq J} D_{ijc} \end{array}$$

Where D_{ijk}=1 if disease i is prioritized for gene j and belongs to category c and 0 if not. P-values and odds ratios were calculated using Fisher exact tests. FDR adjusted P-values < 0.05 were considered significant.

Disease colocalizing molQTL cell-type enrichment

We identify enriched cell types in GWAS disease EFO categories supported by colocalization as in King et al. 2021 [81]. Briefly, we extracted all GWAS colocalizing molQTL (H4 probability > 0.8). Then, for a given cell type K and disease category I, we generated the following contingency table:

$$\begin{array}{cc} \sum_{j} C_{IjK} & \sum_{j} \sum_{k\neq K} C_{Ijk} \\ \sum_{j} \sum_{i\neq I} C_{ijK} & \sum_{j} \sum_{k\neq K} \sum_{i\neq I} C_{ijk} \end{array}$$

Where C_{ijk}=1 if at least one disease GWAS of category i colocalize with gene j in tissue k and 0 if not. P-values and odds ratios were calculated using Fisher exact tests. We performed the analysis considering all molQTL separately, as well as by grouping similar cell types and tissues together prior to testing for enrichment. FDR adjusted P-values < 0.05 were considered significant.

Drug target- indication pairs in clinical trials

Information about drugs approved or in clinical trials was obtained from the Citeline data from Informa Pharma Intelligence, which is a superset of the most used data sources. In addition to multiple data streams, including nightly feeds from official sources such as ClinicalTrials.gov, Citeline also contains data from primary sources such as institutional press releases, financial reports, study reports, and drug marketing label applications, and secondary sources such as analyst reports by consulting companies. Secondary sources are particularly important to reduce potential biases to the organizations' tenancy to report only successful trials, especially those before the FDA Amendments Act of 2007, which requires all clinical trials to be registered and tracked by ClinicalTrials.gov. Citeline database contains information from both US and non-US sources. Any cancer or cancer related indications were excluded from this analysis.

In order to map gene-disease pairs in the genetic data to target-indication pairs in the drug data, we used EFO, which provided a systematic description of many data elements available in EBI databases. A target-indication pair is said to have genetic evidence if there is genetic evidence of association between the gene and disease sufficiently similar to the indication, based on semantic similarity. Two methods were used to calculate semantic similarity matrix [82, 83]. Semantic similarities between each pair of EFO headings were computed in the ontologySimilarity R package [84]. The average of the two methods was calculated and standardized similarities had a maximum value of 1 for each disease or indication. Two diseases are considered similar if the similarity is greater than or equal to a previously published value of 0.7 [11].

Prediction of drug mechanism of action directionality

We retrieved information about drug mechanism of action (MoA) from the Informa Pharma Intelligence dataset described above. Drug MoA were linked to GWAS using a semantic similarity threshold 0.7. When multiple GWAS could be connected to the same drug target and indication, we kept only the GWAS with the most significant p-value at the locus. For targets for which *decreased* expression or loss of function (LoF) is beneficial, we considered datasets with the following keywords: "antagonist", "inhibitor", and "degrader". For targets for which *increased* expression or function is beneficial, we considered the following keyworks: "agonist", and "activator". We considered drugs and targets in phase II clinical trial or above. We performed two analyses to infer directionality from GWAS. First, we assess directionality using the effect size of low-frequency lead coding variant (MAF < 5%). We assumed that these variants are disruptive or LoF. Therefore, a LoF coding variant associated with increased risk suggests that a drug MoA of agonist or activator would be beneficial, whereas for a protective LoF coding variant, an inhibitor or antagonist would be beneficial. Next, we assessed directionality based on the direction of effect of gene expression on disease risk predicted by MR using molQTL as exposure (q-value < 0.05). We included only molQTL colocalizing with local GWAS signal (H4 PP > 80%). For gene-disease pairs supported by multiple colocalizing molQTL, a consensus direction was inferred if the MR direction of effect was consistent across > 75% of the molQTL. Here, a negative consensus MR direction suggests that increased gene expression leads to decreased disease risk. Therefore, an activator or agonist drug targeting this gene would be beneficial. Conversely, a positive consensus MR direction suggests that increased gene expression increases disease risk, and an inhibitor or antagonist drug would be beneficial. We calculated enrichment of concordant direction of effect between GWAS and drug MoA using Fisher exact tests.

Identification of causal links between diseases and genes related to the IL6 receptor

We aimed to apply our proposed approach to a specific case example. Using the causal gene prioritization and GWAS datasets described above, we extracted all disease GWAS for which *IL6*, *IL6R*, or *IL6ST* were predicted to be causal. We predicted directionality of effect of gene expression on disease risk by MR as above using a threshold of q-value < 0.05. We generated local association of plots molQTL and GWAS using LocusZoom [85]. We performed fine-mapping of *IL6ST* genetic variants associated with polymyalgia rheumatica using SuSIE [86] as previously described for FinnGen [7].

Results

Prioritization of putative causal genes in thousands of GWAS

We aimed to prioritize causal genes across 4,611 GWAS from 3 different sources (Table 1): UKB [45], FinnGen release 10 (R10), and meta-analyses of UKB, FinnGen R10, and Estonian biobank [6]. For simplicity, we refer to both single studies and meta-analyses as GWAS throughout the manuscript. While molQTL such as eQTL have been used previously to prioritize causal genes, they are often pleiotropic with the same variant associated with multiple genes within the same locus [26, 32, 33]. Additional genomic information such as the ABC model have been shown to increase performance to identify causal genes, in particular when selecting genes with the highest ABC score (ABCmax) [41]. Therefore, we derived a ranking scheme to prioritize genes using different features including ABC, molQTL, coding variant annotations and pathogenicity predictions, and distance to lead variant (Fig. 1A, methods). We integrated 110 molQTL datasets from 26 studies using MR to infer causality and directionality of gene expression on disease risk. We also performed colocalization analyses to confirm that both GWAS or meta-analyses and molQTL signals shared at least one causal variant. Top ranking genes were selected as those that either contained an associated lead coding variant or were supported by both ABCmax and colocalization across >2 cell types or tissues. We did not include distance to lead variant for higher ranks because we wanted to first prioritize genes for which we could identify potential biological mechanisms. However, for loci without such evidence, or in cases where multiple genes showed identical ranks, the nearest gene to the lead variant was selected as the putative causal gene if it was among the best candidates. Overall, between 1.1 and 1.4 genes were prioritized per locus on average (before breaking ties with the nearest gene), with 17–45% of loci supported by molQTL colocalization or coding variants (Table 1).

Enrichment of genomic features for gold standard genes

Comparing the enrichment of different genomic features alone for curated gold standard genes [26], we found a strong enrichment for genes supported by ABCmax with lead variant (Odds ratio (OR) = 16.3, $P = 5\times10^{-19}$i (Additional file 1: Figure S1; Additional file 2: Table S1). molQTL colocalization also enriched for gold standard genes (colocalization H4 posterior probability (PP) > 95%, OR = 13.3, $P = 3\times10^{-31}$). However, the strongest enrichment was observed for genes with associated coding variants (OR = 50.5, $P = 7\times10^{-60}$) and the nearest genes (OR = 28.5,

Table 1 GWAS included in this study.The table reports the maximum GWAS sample size for each study, the total number of GWAS with at least one associated gene. The number of loci with at least one variant with GWAS $P < 1\times10^{-6}$. To calculate the number of loci, we defined 250 kb regions on each side of the lead variant. Overlapping regions were then merged. The table reports the total number of non-overlapping regions. The mean number of prioritized genes corresponds to the average number of genes prioritized across each GWAS. The mean number of prioritized gene per locus corresponds to the average number of genes with the highest scores in a locus. For the analyses reported throughout this manuscript, ties are broken using the shortest distance to the lead variant. Finally, the last column reports the average number of prioritized gene supported by coding variants or molQTL colocalization

Study ID	Max sample size	Number of GWAS	Mean *N* loci ($P < 1\times10^{-6}$)	Mean *N* prioritized genes	Mean *N* prioritized genes per locus	Mean *N* prioritized genes supported by molQTL or coding variants
FinnGen R10	412,181	2,297	16.36	19.9	1.17	0.21
FinnGen, UK biobank, Estonian biobank meta-analysis (R10)	1,073,998	95	123.44	164.76	1.32	0.44
UKBB pan ICD-10 (European)	420,531	898	9.01	10.23	1.08	0.17
UKBB pan phecodes (European)	42,0531	1,321	10.52	12.21	1.09	0.19

molQTL molecular QTL, *N* Number

$P = 3 \times 10^{-81}$). These strong enrichments are expected given that the gene closest to the lead variant is often the causal gene. In addition, several of the gold standard genes have been selected because they are supported by coding variants or tend to fall in the center of GWAS peaks and have been investigated more closely [26]. However, when using these features in combination, we found that our ranking approach performed well and generally better than selecting the nearest gene alone, with a mean increase in F1 score of 0.13 across studies (range 0.08–0.23) (Additional file 1: Figure S2-S3; Additional file 2: Table S1).

Pathogenicity annotations identify genes linked to monogenic disorders

Integrating information about variant pathogenicity retrieved variants linked to monogenic disorders including *PSEN1* with Alzheimer's disease (AD) [87] (rs764971634, p.Ile437Val, $P = 2 \times 10^{-12}$), *SQSTM1* and Paget's disease [88] (rs104893941, p.Pro392Leu, $P = 6 \times 10^{-11}$), and *HFE* and disorders of iron metabolism [89] (rs1800562, p.Cys282Tyr, $P = 1 \times 10^{-178}$) (Fig. 1B; Additional file 2: Table S3). We also identified protective variants such as *APP* p.Ala673Thr (rs63750847, $P = 7 \times 10^{-11}$) reducing odds of developing AD [90], and *ALOX15* p.Thr560Met protecting against nasal polyps (rs34210653, $P = 2 \times 10^{-15}$) [91]. Of 504 genes prioritized with at least one predicted pathogenic lead variant, 287 had at least one disease mutation reported in the Human Gene Mutation Database (HGMD) [50] (OR = 2.4 [2.0-2.9], $P = 4 \times 10^{-21}$). Potential novel associations included *COLGALT2* and arthrosis (rs35937944, p.Tyr212Cys, $P = 2 \times 10^{-14}$), *LGR5* and carcinoid syndrome (rs200138614, p.Cys712Phe, $P = 4 \times 10^{-9}$), and *GREB1* and female infertility (rs755857714, p.Arg1339His, $P = 4 \times 10^{-9}$).

Colocalizing molQTL link genes to diseases and pathogenic tissues

Prioritized candidate causal genes showed enrichment in disease-colocalizing molQTL related to their known function. For instance, colocalizing molQTL for prioritized genes supported associations with disease categories such as *EDNRA*, *LPA* and *FGF5* with cardiovascular diseases ($P < 5 \times 10^{-10}$), *TSLP*, *IL33*, *CHRNA3*, and *CHRNA5* and respiratory system diseases ($P < 2 \times 10^{-16}$), and *IL23R*, *TYK2*, *IL10* and immune system disease ($P < 2 \times 10^{-9}$) (Fig. 1C-D; Additional file 2: Table S4). In addition, disease-colocalizing molQTL tended to be enriched in specific tissues and cell types. For instance, we found an enrichment of disease-colocalizing eQTL in kidney cortex for *FGF5*, a gene expressed during kidney development and associated with kidney function ($P = 2 \times 10^{-18}$) [92] (Fig. 1E; Additional file 2: Table S5). Other examples include artery eQTL for the cardiovascular diseases associated gene *PHACTR1* [93] ($P = 8 \times 10^{-10}$); the lysosomal acid lipase (*LIPA*) gene and microglia eQTL ($P = 2 \times 10^{-11}$); and the *ABO* blood group gene with plasma pQTL ($P = 1 \times 10^{-21}$). Finally, we confirmed that enriched colocalizing eQTL matched the expected pathogenic tissues and cell-types of different disease categories (Fig. 1F; Additional file 2: Table S6). For instance, after grouping eQTL of similar tissues and cell types together, we found a strong enrichment of genes with artery and heart eQTL colocalizing with cardiovascular disease GWAS ($P < 6 < \text{x}10^{-17}$). We found similar enrichment for T cell and thyroid eQTL in endocrine system diseases ($P < 3 \times 10^{-7}$); blood, lymphoblastoid cell line, monocytes, neutrophil, and T cells with immune system diseases ($P < 1 \times 10^{-6}$); and fibroblasts and musculoskeletal diseases ($P = 7 \times 10^{-6}$). Treating each eQTL dataset separately revealed additional associations with tissues or cell subsets including brain cortex and diseases of the visual system ($P = 7 \times 10^{-6}$); cerebellum

(See figure on next page.)

Fig. 1 Characteristics of prioritized genes via gain or loss of function variants and molQTL. **A** Features used to prioritize genes in GWAS loci. Genes are ranked based on a combination of features including molQTL, activity-by-contact (ABC) maps, and variant annotations, including variant effect predictions (VEP) and pathogenicity predictions. **B** Disease-associated predicted pathogenic variants capture disease associations with high effect sizes. Lead pathogenic variants with GWAS *P*-value $< 5 \times 10^{-8}$ are reported in the figure. Effect of the risk allele (odds ratio) is reported on the y-axis. The x-axis corresponds to the frequency of the risk allele. **C** Disease category overrepresentation for single genes predicted to be causal. Each dot represents a different associated disease category. Top 30 enrichments are shown. **D** Same as B, but filtered for genes predicted to be causal and enriched in "Immune system diseases". Each dot represents a different associated disease category. Top 30 genes are shown. **E** Overrepresentation of eQTL colocalization for single genes predicted to be causal. Gene-tissue pairs are included only if the gene has the highest rank in a locus for a given associated disease. Top 30 enriched eQTL are shown. Each dot represents a different enriched tissue or cell-type. **F** Enriched colocalizing cell types and tissues by disease categories. Only disease categories and tissues or cell types with at least one significant enrichment are reported in the heatmap. Enrichment *P*-values are calculated using Fisher exact test, testing for the enrichment of genes with eQTL colocalizing with GWAS belonging to specific disease categories as in [81]. Tissues and cell-types were collapsed into broader categories before testing for enrichment. For example, tibial, coronary, and aorta arteries were grouped into "artery" molQTL: Molecular QTL; ABC: Activity-By-Contact; LCL: Lymphoblastoid cell lines; iPSC: induced Pluripotent Stem Cells .: Adjusted *P* < 0.1; *: Adjusted *P* < 0.05; **: Adjusted *P* < 0.01; ***: Adjusted *P* < 0.001

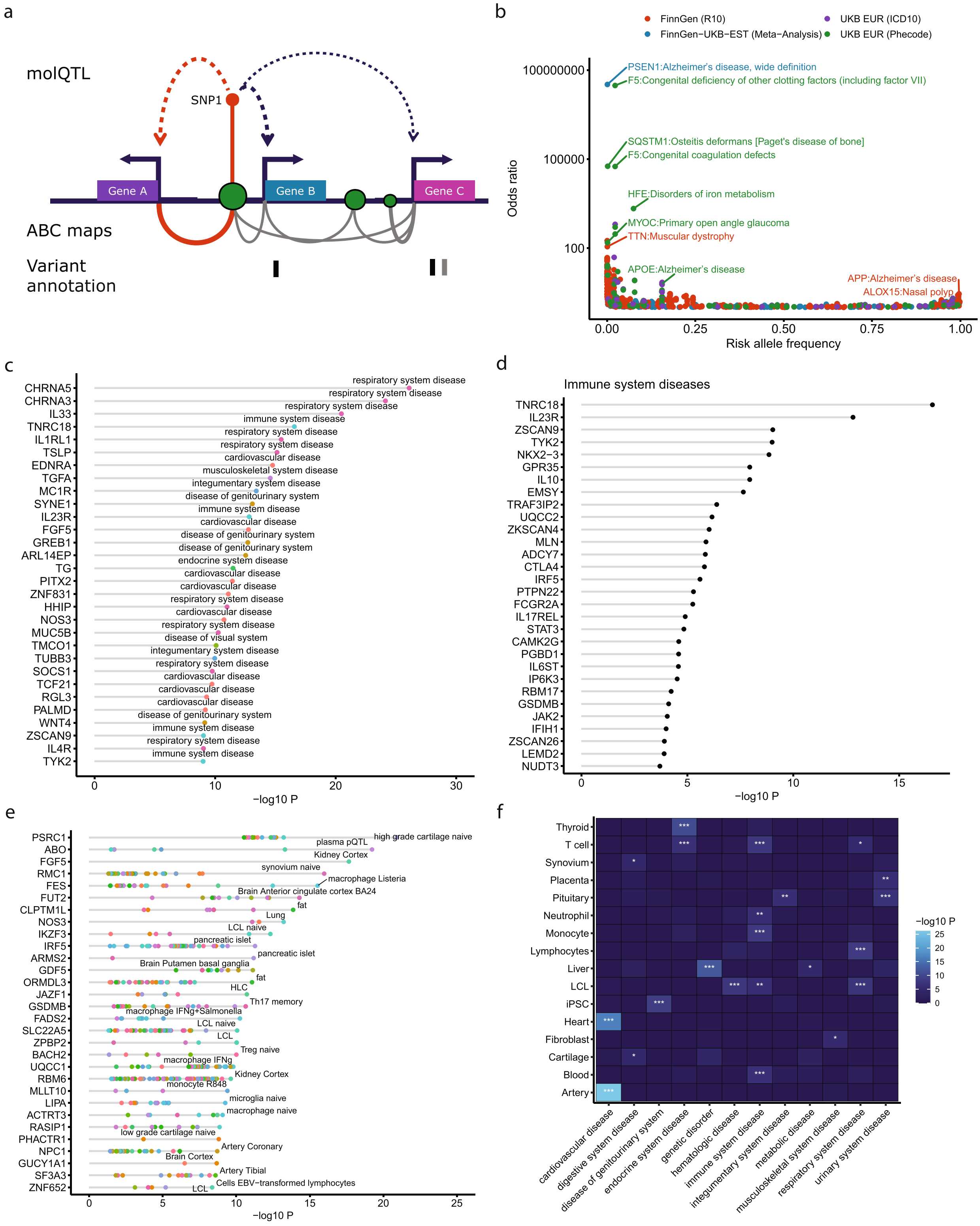

Fig. 1 (See legend on previous page.)

and nervous system diseases ($P=4\times10^{-6}$); regulatory T cells and endocrine system diseases ($P=9\times10^{-9}$); and T helper 17 cells and digestive system diseases ($P=5\times10^{-7}$) (Additional file 1: Figure S4; Additional file 2: Table S7). Overall, the analyses illustrate that in contrast to the nearest gene approach, inclusion of molQTL can help contextualize genetic associations to potential pathogenic cell types and tissues.

Prioritized genes increase clinical trial probability of success

Building on these results, we tested whether we could use molQTL information of putative causal gene to drive drug repurposing opportunities. First, we evaluated whether the prioritized genes enriched for therapeutic targets with clinical trial success. Clinical trial information was retrieved from the Citeline Pharma Intelligence project. Consistent with previous observations, we found that targets with clinical trial success were enriched for features such as presence of coding variants (Fig. 2A, Additional file 2: Table S8). For example, likely pathogenic coding variants demonstrated some of the best predictive performances (Phase I: Risk ratio (RR) = 1.20, $P=0.007$; Phase II: RR = 1.26, $P=0.008$; Phase III: RR = 2.07, $P=3\times10^{-8}$; Approved: RR = 2.84, $P=1\times10^{-9}$). Similar results were observed analyzing each study separately (Additional file 1: Figure S5). Use of epigenetic evidence also improved predictions, in particular lead SNPs linked by the ABC model (Phase I: RR = 1.23, $P=0.004$; Phase II: RR = 1.36, $P=6\times10^{-4}$; Phase III: RR = 1.67, $P=5\times10^{-4}$; Approved: RR = 2.06, $P=4\times10^{-4}$). However, molQTL information alone did not enrich as much for clinical trial success, for example, colocalizing molQTL with posterior probability > 80% (Phase I: RR = 1.15, $P=0.003$; Phase II: RR = 1.21, $P=0.001$; Phase III: RR = 1.29, $P=0.01$; Approved: RR = 1.57, $P=0.002$). While the overall prioritized genes did not show the strongest enrichment (Phase I: RR = 1.18, $P=1\times10^{-5}$; Phase II: RR = 1.24, $P=3\times10^{-6}$; Phase III: RR = 1.49, $P=7\times10^{-7}$; Approved: RR = 1.83, $P=4\times10^{-8}$), this was likely due to the inclusion of genes with no supportive evidence other than distance (Fig. 2A). Indeed, we found that "High" and "Very High" prioritization ranks were more predictive of successful clinical trial progression (higher risk ratios) than lower-ranking genes, especially at later clinical trial phases or after approval (High + Very high ranks Phase I: RR = 1.24, $P=4\times10^{-6}$; Phase II: RR = 1.33, $P=8\times10^{-7}$; Phase III: RR = 1.71, $P=3\times10^{-8}$; Approved: RR = 2.24, $P=1\times10^{-10}$) (Fig. 2B, Additional file 1: Figure S5, Additional file 2: Table S9). In our analysis, distance itself was not as predictive of clinical trial success especially after excluding loci likely driven by coding variants (Phase I: RR = 1.13, $P=0.01$; Phase II: RR = 1.21, $P=0.001$; Phase III: RR = 1.32, $P=0.009$; Approved: RR = 1.54, $P=0.003$) (Fig. 2B).

Inferred directionality from GWAS recapitulate drug MoA

To understand whether inferred directionality could be informative of clinical trial success, we first investigated the consistency between the direction of effect of coding variants and drug MoA (methods). When considering prioritized genes with lead low-frequency coding variants (minor allele frequency < 0.05) and clinical trials phase II and above, between 92% showed consistent effect between the minor allele and drug MoA (Fisher $P=2\times10^{-16}$, Fig. 2C). Results were similar when stratifying GWAS by data source (Additional file 1: Figure S6). We then asked whether molQTL could similarly inform on directionality. Using prioritized gene-disease pairs supported by MR (q-value < 0.05) and colocalization (PP > 80%), we inferred the direction of effect when the predicted MR effect was consistent across >75% of molQTL datasets for a given gene. This was the case for most gene-disease pairs (Additional file 1: Figure S7). Again, direction of effect was generally in agreement with drug MoA (73% agreement, Fisher $P=4\times10^{-8}$-5×10^{-41}, Fig. 2D, Additional file 1: Figure S6). Consistency across all studies increased when considering only approved drugs (85–94% agreement, Fisher $P=4\times10^{-7}$-5×10^{-26}, Additional file 1: Figure S8). Overall, these data suggest that molQTL can be used to inform on drug MoA.

Causal gene predication from GWAS identifies a link between IL6ST and polymyalgia rheumatica

Finally, we applied our causal gene prioritization approach to a specific use case, that is to identify potential new indications for drugs targeting the IL6 receptor such as Sarilumab and Tocilizumab, both drugs approved for rheumatoid arthritis. We extracted diseases prioritized by our approach for genes related to the receptor, namely *IL6*, *IL6ST*, and *IL6R*. We identified putative causal links between increased *IL6* expression in CD16 monocytes and increased risk of varicose veins, ischemic heart disease, coronary atherosclerosis, and atrial fibrillation (MR beta > 0), but decreased risk of asthma and allergy (MR beta < 0) (Additional file 1: Figure S9; Additional file 2: Table S10). eQTL of *IL6* in whole blood also supported these disease associations, albeit with an opposite predicted direction of effect. Similarly, *IL6R* expression in multiple tissues including artery, colon, and esophagus was associated with increased risk of coronary revascularization, coronary atherosclerosis, and abdominal aortic aneurysm (AAA), but lower risk of lower respiratory diseases and atopic dermatitis (Additional file 1: Figure S10). Again, we observed opposite direction of effect predicted by MR for these diseases when using monocyte

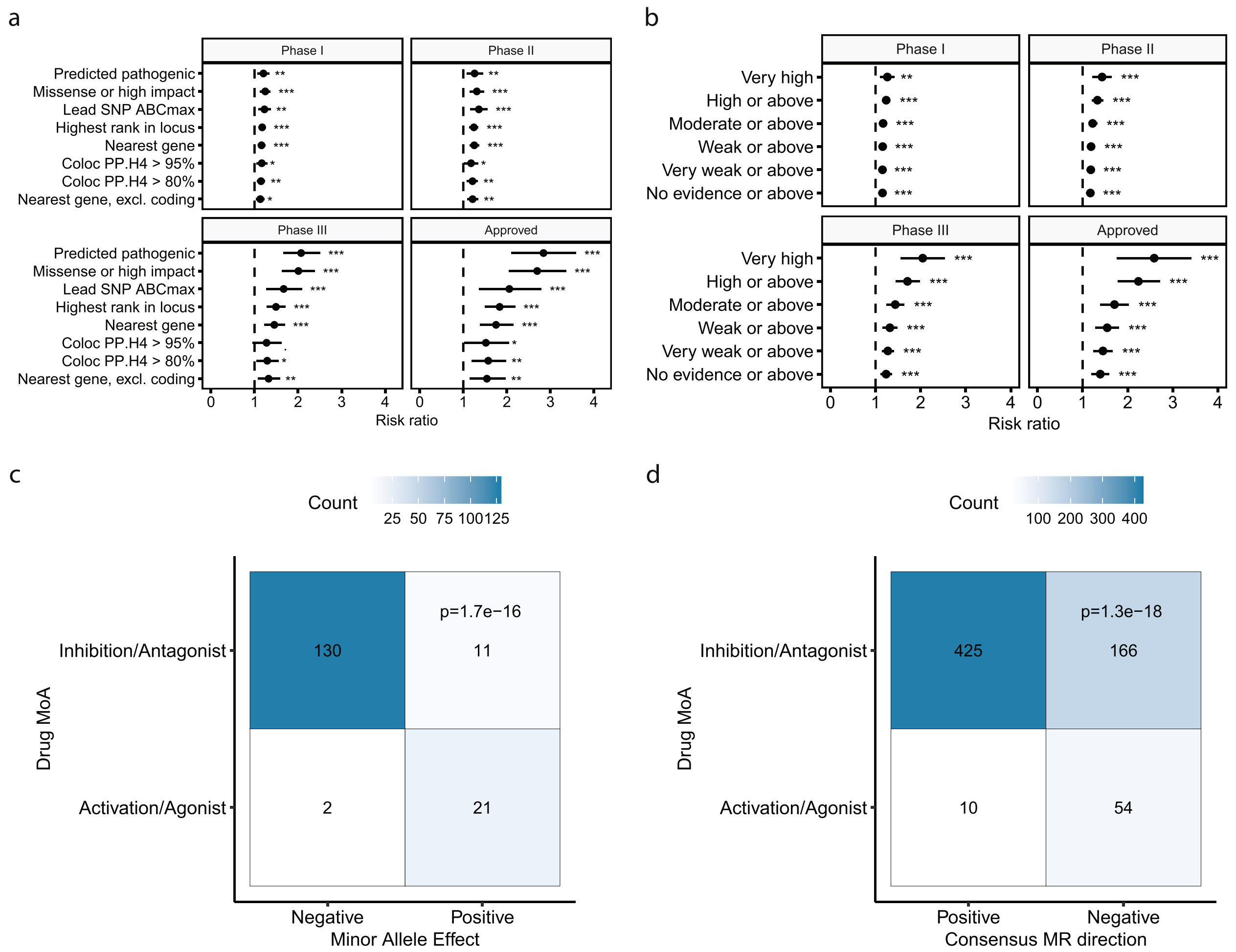

Fig. 2 Prioritized genes predict clinical trial success. **A** Enrichment of targets of approved drugs or drugs in clinical trials (phase I-III) using genetic evidence aggregated from FinnGen, UKB, and biobank meta-analyses prioritizing genes using colocalization (posterior probability of colocalization [H4] > 80% or > 95%), predicted pathogenic variants, genes with highest prioritization rank, ABC score for lead variant, or nearest gene excluding loci with associated coding variants. **B** Enrichment of targets of approved drugs or drugs in clinical trials (phase I-III) using causal gene prioritization ranks across all studies. **C** Concordance between direction of effect of lead low-frequency coding variants on disease risk, and drug MoA for targets in phase II clinical trials or above. We retrieved information about targets, clinical trials, and drug MoA from the Citeline Pharmacogenomics dataset. We connected this dataset to GWAS phenotypes using EFO codes and a semantic similarity score > 0.7. We assume that low-frequency coding variants (minor allele frequency < 5%) are disruptive (LoF). Therefore, a negative (protective) direction of effect would translate into inhibition or antagonism being beneficial (and vice-versa). **D** Concordance between the predicted impact of gene expression on disease risk predicted by MR, and drug MoA for targets in phase II clinical trials or above. Information about targets, clinical trials, and drug MoA were collected from the Citeline Pharmacogenomics dataset and connected to GWAS phenotypes using EFO codes and a semantic similarity score > 0.7. The direction of effect of gene expression on disease risk was assessed by MR using molQTL as exposure (q-value < 0.05). Only molQTL colocalizing with local GWAS signal (H4 posterior probability > 80%) were included. A consensus direction was inferred if the MR direction of effect was consistent across > 75% of molQTL for a given gene and disease GWAS. A negative consensus MR direction suggests that increased gene expression leads to decreased disease risk. Therefore, an activator or agonist drug targeting this gene would be beneficial. Conversely, a positive consensus MR direction suggests that increased gene expression increases disease risk, and an inhibitor or antagonist drug would be beneficial. Reported *P*-values were calculated by Fisher exact test .: $P < 0.1$; *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$

or macrophage eQTL as exposure. The associations with coronary atherosclerosis and AAA were further driven by a lead coding variant in *IL6R*, rs2228145 (Asp358Ala, Additional file 2: Table S10). Finally, we found that increased *IL6ST* expression in T cells and whole blood is predicted to increase the risk of rheumatoid arthritis, systemic connective tissue disorders, polyarthropathies, other arthritis, autoimmune diseases, and polymyalgia rheumatica (Fig. 3A, Additional file 2: Table S10). These associations were driven by rs7731626 (SuSIE fine-mapping probability > 0.99). This variant is located within an intron of *ANKRD55* and colocalizes with eQTL for both

ANKRD55 and *IL6ST* (PP > 80%). However, this variant also overlaps with an enhancer that shows the highest ABC score for *IL6ST* for genes in the region, suggesting the latter is the causal gene, in line with previous studies [94, 95] (Fig. 3B). Overall, our approach was able to capture known associations with IL6-R related genes and identified an association between *IL6ST* and polymyalgia rheumatica.

Discussion

We prioritized disease-associated genes across 4,611 GWAS and meta-analyses from biobank studies using a combination of MR with molQTL, colocalization analysis, variant effect prediction, and epigenetic annotations (ABC model). This approach allows the use of molQTL to infer directionality of gene expression on disease risk, while improving the causal gene prediction compared to using molQTL alone. Based on combination of these features, we used a ranking approach to prioritize genes within loci and showed that this approach is enriched for gold standard genes. We recover known coding variant associations, including rare variants in genes linked to monogenic disorders such as *PSEN1* and *APP1* and Alzheimer's disease, and *SQSTIM1* and Paget's disease (Fig. 1B). Genes prioritized by molQTL also show enrichment in disease categories related to their function with pathogenic tissue contexts (Fig. 1C-F). Of note, when multiple genes show evidence of colocalization within the same locus, the addition of epigenetic (ABCmax) information can help prioritize one gene over the others. We note as an example the association of variants with polymyalgia rheumatica at the *ANRKD55* locus where this gene would be prioritized using the nearest gene approach. Whereas colocalization alone did not identify a single causal gene, combination of colocalization and ABCmax identified *IL6ST* as the putative causal gene, consistent with recent reports [96, 97]. *IL6ST* encodes a protein involved in signal transduction for the IL6 receptor pathway. Inhibitors of the IL6 receptor have recently shown success in clinical trials for this indication leading to a recent approval by the FDA [98].

In line with previous studies [11, 12], we show that therapeutic targets with genetic evidence are enriched at later clinical trial phases and as targets of approved drugs. In our analysis, using the nearest gene information alone was not strongly predictive of clinical trial success. The most predictive features were coding variant annotations and ABC maps. While the latter performs well to link causal genes to diseases, it does not provide information about directionality. We used coding variants and MR with molQTL to infer directionality of a target on disease risk. Both approaches were generally consistent with drug MoA matched for the target and disease. These data support that molQTL can be used to predict drug MoA. However, while we found that in general eQTL were consistent across cell type and tissues for a given gene and disease (Additional file 1: Figure S7), we note that this isn't always the case. This is exemplified by the IL6-R case study, where all three queried genes displayed inconsistent direction of effect predicted by MR depending on the molQTL dataset. Future improvement of this approach should consider prior knowledge on pathogenic cell types

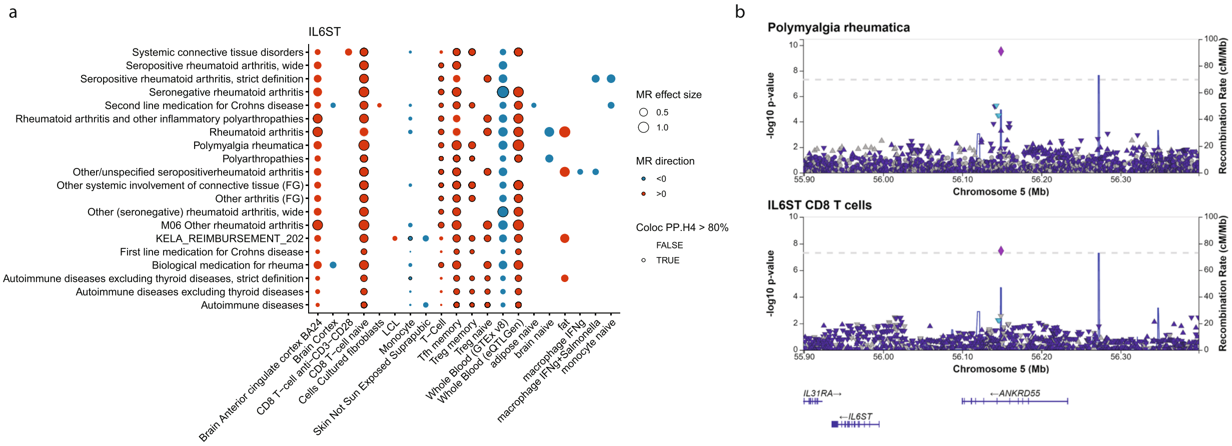

Fig. 3 *IL6ST* is predicted to be causal for rheumatoid arthritis and polymyalgia rheumatica. **A** Diseases associations supported by MR, colocalization and ABC. The figure shows tissues and cell-types with significant MR (q-value < 0.05) using *IL6ST* eQTL as exposure and diseases as outcome (red: positive effect size estimate [MR beta]; blue: negative effect size estimate). The size of the dots represents absolute effect size. Disease-eQTL pairs with a colocalization posterior probability > 80% are highlighted with a dark border. **B** LocusZoom [85] plot showing the top association for polymyalgia rheumatica at the *ANKRD55-IL6ST* locus. Both *IL6ST* and *ANKRD55* eQTL colocalize with the polymyalgia rheumatica signal, but *IL6ST* has the highest ABC score

or tissues to infer directionality in relevant contexts. Overall, our analysis suggests that using features such as ABCmax in combination with molQTL can increase the performance of causal gene inference approaches while informing on directionality which is crucial for translating GWAS hits to therapies.

We note that this study has some limitations. First, we did not perform fine-mapping analyses nor colocalization approaches that use LD references. Indeed, we opted to avoid methods that do not rely on LD references as we used GWAS from various sources, including meta-analyses where these methods may not be well calibrated [99]. Nevertheless, using fine-mapping information likely would improve performance, especially in cases where there are multiple causal variants underlying molQTL or GWAS signals, and would reduce LD contamination [30, 100]. In addition, using MR approaches like SMR and HEIDI or MRLocus, are likely to perform better in case of pleiotropy or allelic heterogeneity [101, 102]. This is evident in the case of *IL6ST*, where MR using eQTL from whole blood from different sources (GTEx, eQTLGen) lead to inversed estimate of directionality (Fig. 3A). This difference was due to different instruments used as only one genetic instrument was included in GTEx whereas 5 independent instruments were included for eQTLGen. We also assume that there is one causal gene per locus, although it is possible that multiple genes contribute to disease risk. Finally, integrating other sources of molQTL such as metabolite or splice QTL could help further identify putative causal genes as coding variants and eQTL only cover a fraction of loci (17–47% in this study) [103]. Similarly, considering additional cell types in both the molQTL and ABC annotations would further help identify functional links between variants and genes. While these approaches can be useful to nominate candidate causal genes and their relationship to diseases, proper functional validation remains of high importance.

Conclusions

We nominated putative causal genes across 4,611 GWAS from biobank studies and public resources by integrating variant annotations as well as molQTL. We show that these prioritized genes recover known biological relationships in terms of disease and tissue enrichment and are enriched for therapeutic targets that succeeded in clinical trials. We show that directionality predicted by molQTL and coding variants generally recapitulate drug MoA. Finally, we applied this approach to genes related to the IL6 receptor and identified an association between *IL6ST* and polymyalgia rheumatica supporting the recent approval of Sarilumab for this indication.

Abbreviations

AAA	Abdominal aortic aneurysm
ABC	Activity-by-contact
CI	Confidence interval
EFO	Experimental factor ontology
eQTL	Expression quantitative trait loci
EstBB	Estonian Biobank
GWAS	Genome-wide association study
GoF	Gain of function
HLA	Human leukocyte antigen
iPSC	Induced Pluripotent Stem Cells
LCL	Lymphoblastoid cell lines
LD	Linkage disequilibrium
LoF	Loss of function
MAF	Minor allele frequency
MoA	Mechanism of action
MR	Mendelian randomization
molQTL	Molecular quantitative trait loci
OR	Odds ratio
pQTL	Protein quantitative trait loci
PP	Posterior probability
QTL	Quantitative trait loci
RR	Risk ratio
UKB	UK Biobank
VEP	Variant effect predictor

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10971-2.

Additional file 1: Figure S1. Gold standard gene enrichment by genomic features. Figure S2. Precision and recall of gold standard genes for different genomic features as well as causal candidate prioritization approach. Figure S3. F1 scores for each considered features and prioritization scheme. Figure S4. Enriched colocalizing cell types and tissues by disease categories. Figure S5. Enrichment of clinical success stratified by GWAS source. Figure S6. Predicted directionality and drug mechanism of action stratified by GWAS source. Figure S7. Predicted direction of effect of gene expression on disease risk. Figure S8. Concordance between the predicted effect of gene expression on disease risk by MR and mMoA of approved drugs. Figure S9. Association between *IL6* and diseases, supported by MR, colocalization and ABC. Figure S10. Association between *IL6R* and diseases, supported by MR, colocalization and ABC.

Additional file 2: Table S1. Enrichment of gold standard genes by feature and GWAS study source. Table S2. Precision and recall of different features to recover gold standard genes. Table S3. Genes with predicted gain or loss of function variants (P<1x10^{-6}). Table S4. Genes with overrepresented disease categories of GWAS in which they are prioritized as causal. Table S5. Genes with overrepresented cell-type colocalizing QTL with GWAS in which they are prioritized as causal. Table S6. Significantly enriched colocalizing QTL cell types and tissues in disease GWAS categories, after grouping similar tissues and cell-types together. Table S7. Significantly enriched colocalizing QTL cell types and tissues in disease GWAS categories, treating each eQTL dataset separately. Table S8. Enrichment of prioritized genes by feature across clinical trial phases and approved drugs. Table S9. Enrichment of prioritized genes by rank across clinical trial phases and approved drugs. Table S10. Putative causal association between diseases and IL6, IL6ST, or IL6R.

Additional file 3. List of FinnGen authors and their affiliations.

Additional file 4. Funding statements and references for all eQTL and pQTL datasets used for this manuscript.

Acknowledgements

We thank all participants and contributors to the datasets used in this study. We thank Hao He (Sanofi US) for his valuable feedback on the manuscript, and Omar Stradella (Sanofi US) for his support with ontology mapping.

UK biobank

This research has been conducted using the UK Biobank Resource (https://www.ukbiobank.ac.uk.), a large-scale biomedical database and research resource containing genetic, lifestyle and health information from 500,000 UK participants. UK Biobank is supported by its founding funders the Wellcome Trust and UK Medical Research Council, as well as the Department of Health, Scottish Government, the Northwest Regional Development Agency, British Heart Foundation and Cancer Research UK. The UK biobank pan-ancestry analysis was conducted under project ID 31063 (https://pan.ukbb.broadinstitute.org).

FinnGen

The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Celgene Corporation, Celgene International II Sàrl, Genentech Inc., Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, and Novartis AG. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish Biobanks are members of BBMRI. fi infrastructure (www.bbmri.fi). Finnish Biobank Cooperative -FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland.

Estonian biobank

Estonian Biobank research was supported by the European Union through Horizon 2020 research and innovation programme under grant no 810645 and through the European Regional Development Fund project no. MOBEC008, by the Estonian Research Council grant PUT (PRG1291, PRG687 and PRG184) and by the European Union through the European Regional Development Fund project no. MOBERA21 (ERA-CVD project DETECT ARRHYTHMIAS, GA no JTC2018-009), Project No. 2014-2020.4.01.15-0012 and Project No. 2014-2020.4.01.16–0125. Computations were performed in the High Performance Computing Center, University of Tartu.

molQTL datasets

This manuscript makes use of previously published molecular QTL data. Except for pQTL from Sun 2018 and eQTL Gen, formatted summary statistics were obtained from the EBI eQTL catalogue. We wish to thank all participants and contributors to these datasets. We list funding sources of each of the dataset in a supplementary note.

Estonian biobank and FinnGen banners

The Estonian Biobank team is composed of: Andres Metspalu, Mari Nelis, Lili Milani, Reedik Mägi, Georgi Hudjashov, and Tõnu Esko (University of Tartu, Tartu, Estonia). FinnGen authors and their institution are listed in a supplementary file.

Authors' contributions

SL, MC, and CC performed data analysis, interpreted the results, designed analyses, and are major contributors in writing the manuscript. FinnGen authors defined endpoints, performed GWAS, generated summary statistics, and performed meta-analyses. Estonian Biobank authors defined matching disease endpoints, performed GWAS and generated EstBB summary statistics. MB acquired and formatted data and performed data analysis. KK, EdR, JS, SK, DR designed the study, interpreted results, and contributed to writing the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by Sanofi (Cambridge, MA, United States). The funder had the following involvement with the study: Sanofi reviewed the manuscript.

The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Celgene Corporation, Celgene International II Sàrl, Genentech Inc., Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, and Novartis AG.

UK Biobank is supported by its founding funders the Wellcome Trust and UK Medical Research Council, as well as the Department of Health, Scottish Government, the Northwest Regional Development Agency, British Heart Foundation and Cancer Research UK.

Estonian Biobank research was supported by the European Union through Horizon 2020 research and innovation programme under grant no 810645 and through the European Regional Development Fund project no. MOBEC008, by the Estonian Research Council grant PUT (PRG1291, PRG687 and PRG184) and by the European Union through the European Regional Development Fund project no. MOBERA21 (ERA-CVD project DETECT ARRHYTHMIAS, GA no JTC2018-009), Project No. 2014-2020.4.01.15-0012 and Project No. 2014-2020.4.01.16–0125.

Data availability

The UK Biobank Pan ancestry GWAS are available through https://pan.ukbb.broadinstitute.org/. FinnGen GWAS are available through https://www.finngen.fi/en/access_results. Processed and formatted eQTL data used in this study are available through the eQTL catalogue https://www.ebi.ac.uk/eqtl/. pQTL from Sun et al. 2018 are available through http://www.phpc.cam.ac.uk/ceu/proteins/. eQTLGen eQTL are available through https://www.eqtlgen.org/phase1.html. 1000 Genomes project phase 3 data [79] is available through https://www.internationalgenome.org/category/phase-3/. Activity-by-contact maps are available through https://www.engreitzlab.org/resources/. ProtVar annotations, including ESMb1 and EVE, are available through https://www.ebi.ac.uk/ProtVar/. Code for GWAS gene prioritization is available at https://github.com/Sanofi-Public/PMCB-GWAS_multi-omics_prioritization. Datasets supporting the conclusions of this article are included within the article and its additional files. Additional datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Patients and control subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea, the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) approved the FinnGen study protocol Nr HUS/990/2017.

The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019, THL/1524/5.05.00/2020, and THL/2364/14.02/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020 and Statistics Finland (permit numbers: TK-53-1041-17 and TK-53-90-20).

The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 6 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, Auria Biobank AB17-5154, Biobank Borealis of Northern Finland_2017_1013, Biobank of Eastern Finland 1186/2018, Finnish Clinical Biobank Tampere MH0004, Central Finland Biobank 1-2017, and Terveystalo Biobank STB 2018001.

UK Biobank has received ethical approval from the NHS National Research Ethics Service North West (approval numbers 11/NW/0382 and 16/NW/0274). All participants provided written informed consent. Estonian Biobank GWAS and consecutive meta-analyses were carried out under ethical approval permit number 1.1-12/1020 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs).

Consent for publication

Not applicable.

Competing interests

SL, MC, MB, CC, EdR, KK, JS, SK are employees of Sanofi US Services and hold shares and/or stock options in the company. DKR is currently an employee of Takeda and was an employee of Sanofi US Services at the time of study. All authors declare no other competing interests.

Author details

[1]Precision Medicine & Computational Biology, Sanofi, Cambridge, MA, USA. [2]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia. [3]Digital R&D Data & Computational Sciences, Sanofi, Gentilly, France. [4]Translational Sciences, Sanofi, Framingham, MA, USA. [5]Pre-Clinical and Translational Sciences, Takeda, MA, USA. [6]Immunology & Inflammation Development, Sanofi, Cambridge, MA, USA. [7]Genetics Research, Sanofi, Cambridge, MA, USA.

Received: 15 November 2023 Accepted: 28 October 2024
Published online: 19 November 2024

References

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 years of GWAS Discovery: Biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
2. Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. Nat Commun. 2020;11:5900.
3. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–1012.
4. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.
5. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, 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. 2015;12: e1001779.
6. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Magi R, Milani L, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015;44:1137–47.
7. Kurki MI, Karjalainen J, Palta P, Sipila TP, Kristiansson K, Donner KM, Reeve MP, Laivuori H, Aavikko M, Kaunisto MA, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613:508–18.
8. Laisk T, Lepamets M, Koel M, Abner E, Estonian Biobank Research T, Magi R. Genome-wide association study identifies five risk loci for pernicious anemia. Nat Commun. 2021;12:3761.
9. Tyrmi JS, Arffman RK, Pujol-Gualdo N, Kurra V, Morin-Papunen L, Sliz E, FinnGen Consortium EBRT, Piltonen TT, Laisk T, Kettunen J, Laivuori H. Leveraging northern European population history: novel low-frequency variants for polycystic ovary syndrome. Hum Reprod. 2022;37:352–65.
10. Alver M, Palover M, Saar A, Lall K, Zekavat SM, Tonisson N, Leitsalu L, Reigo A, Nikopensius T, Ainla T, et al. Recall by genotype and cascade screening for familial hypercholesterolemia in a population-based biobank from Estonia. Genet Med. 2019;21:1173–80.
11. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet. 2019;15:e1008489.
12. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47:856–60.
13. Reay WR, Cairns MJ. Advancing the use of genome-wide association studies for drug repurposing. Nat Rev Genet. 2021;22:658–71.
14. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47:979–86.
15. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006;314:1461–3.
16. Pidasheva S, Trifari S, Phillips A, Hackney JA, Ma Y, Smith A, Sohn SJ, Spits H, Little RD, Behrens TW, et al. Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q. PLoS ONE. 2011;6: e25038.
17. Peyrin-Biroulet L, Ghosh S, Lee SD, Lee WJ, Griffith J, Wallace K, Berg S, Liao X, Panes J, Loftus EV Jr, Louis E. Effect of risankizumab on health-related quality of life in patients with Crohn's disease: results from phase 3 MOTIVATE, ADVANCE and FORTIFY clinical trials. Aliment Pharmacol Ther. 2023;57:496–508.
18. Ferrante M, Panaccione R, Baert F, Bossuyt P, Colombel JF, Danese S, Dubinsky M, Feagan BG, Hisamatsu T, Lim A, et al. Risankizumab as maintenance therapy for moderately to severely active Crohn's disease: results from the multicentre, randomised, double-blind, placebo-controlled, withdrawal phase 3 FORTIFY maintenance trial. Lancet. 2022;399:2031–46.
19. Feagan BG, Sandborn WJ, Gasink C, Jacobstein D, Lang Y, Friedman JR, Blank MA, Johanns J, Gao LL, Miao Y, et al. Ustekinumab as induction and maintenance therapy for Crohn's Disease. N Engl J Med. 2016;375:1946–60.
20. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, Zhernakova A, Stahl E, Viatte S, McAllister K, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet. 2012;44:1336–40.
21. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466:707–13.
22. Ober C, Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. Immunol Rev. 2011;242:10–30.
23. Valette K, Li Z, Bon-Baret V, Chignon A, Berube JC, Eslami A, Lamothe J, Gaudreault N, Joubert P, Obeidat M, et al. Prioritization of candidate causal genes for asthma in susceptibility loci derived from UK Biobank. Commun Biol. 2021;4:700.
24. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.
25. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in humans. N Engl J Med. 2015;373:895–907.
26. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, Fumis L, Hayhurst J, Buniello A, Karim MA, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet. 2021;53:1527–33.
27. Smith GD, Ebrahim S. Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1–22.
28. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23:R89–98.
29. Richardson TG, Hemani G, Gaunt TR, Relton CL, Davey Smith G. A transcriptome-wide mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. Nat Commun. 2020;11:185.
30. Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. Colocalization of GWAS and eQTL signals detects Target genes. Am J Hum Genet. 2016;99:1245–60.
31. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs

of genetic association studies using summary statistics. PLoS Genet. 2014;10: e1004383.

32. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, Mohammadi P, Park Y, Parsana P, Segrè AV, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.
33. Ndungu A, Payne A, Torres JM, van de Bunt M, McCarthy MI. A multi-tissue transcriptome analysis of human metabolites guides interpretability of associations based on Multi-SNP models for gene expression. Am J Hum Genet. 2020;106:188–201.
34. Liu X, Finucane HK, Gusev A, Bhatia G, Gazal S, O'Connor L, Bulik-Sullivan B, Wright FA, Sullivan PF, Neale BM, Price AL. Functional architectures of local and distal regulation of Gene expression in multiple human tissues. Am J Hum Genet. 2017;100:605–16.
35. Consortium GT. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60.
36. Connally NJ, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, Chun S, Cotsapas C, Cassa CA, Sunyaev SR. The missing link between genetic association and regulatory function. Elife. 2022;11:e74970.
37. Mostafavi H, Spence JP, Naqvi S, Pritchard JK: Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. bioRxiv 2022.
38. Lessard S, Gatof ES, Beaudoin M, Schupp PG, Sher F, Ali A, Prehar S, Kurita R, Nakamura Y, Baena E, et al. An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria susceptibility. J Clin Invest. 2017;127:3065–74.
39. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014;507:371–5.
40. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7:7.
41. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, Jones TR, Nguyen TH, Ulirsch JC, Lekschas F, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;593:238–43.
42. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
43. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR HKF, Schoenherr S, Forer L, McCarthy S, Abecasis GR, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48:1443–8.
44. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.
45. Karczewski KJ, Gupta R, Kanai M, Lu W, Tsuo K, Wang Y, Walters RK, Turley P, Callier S, Baya N, et al. Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. medRxiv. 2024;03.13.2430386.
46. Maintainer BP. liftOver: Changing genomic coordinate systems with rtracklayer::liftOver. R package version 1180 2021.
47. Stephenson JD, Totoo P, Burke DF, Janes J, Beltrao P, Martin MJ. ProtVar: mapping and contextualizing human missense variation. Nucleic Acids Res. 2024;52:W140–7.
48. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y, Marks DS. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599:91–5.
49. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet. 2023;55:1512–22.
50. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. Human gene mutation database (HGMD): 2003 update. Hum Mutat. 2003;21:577–81.
51. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558:73–9.
52. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, Watt S, Yan Y, Kundu K, Ecker S, et al. Genetic drivers of epigenetic and transcriptional variation in Human Immune cells. Cell. 2016;167:1398–e14141324.
53. Vosa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Yazar S, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53:1300–10.
54. Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, Samovica M, Sakthivel MP, Kuzmin I, Trevanion SJ, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet. 2021;53:1290–9.
55. Buil A, Brown AA, Lappalainen T, Vinuela A, Davies MN, Zheng HF, Richards JB, Glass D, Small KS, Durbin R, et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat Genet. 2015;47:88–91.
56. Jaffe AE, Straub RE, Shin JH, Tao R, Gao Y, Collado-Torres L, Kam-Thong T, Xi HS, Quan J, Chen Q, et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. Nat Neurosci. 2018;21:1117–25.
57. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, Ha B, Altay G, Greenbaum JA, McVicker G, et al. Impact of genetic polymorphisms on human immune cell gene expression. Cell. 2018;175:1701–e17151716.
58. Ng B, White CC, Klein HU, Sieberts SK, McCabe C, Patrick E, Xu J, Yu L, Gaiteri C, Bennett DA, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci. 2017;20:1418–26.
59. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife. 2013;2: e00523.
60. van de Bunt M, Manning Fox JE, Dai X, Barrett A, Grey C, Li L, Bennett AJ, Johnson PR, Rajotte RV, Gaulton KJ, et al. Transcript expression data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for type 2 diabetes and glycemic traits to their downstream effectors. PLoS Genet. 2015;11: e1005694.
61. Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, Consortium H, Hale C, Dougan G, Gaffney DJ. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet. 2018;50:424–31.
62. Taylor DL, Jackson AU, Narisu N, Hemani G, Erdos MR, Chines PS, Swift A, Idol J, Didion JP, Welch RP, et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. Proc Natl Acad Sci U S A. 2019;116:10883–8.
63. Lepik K, Annilo T, Kukuskina V, e QC, Kisand K, Kutalik Z, Peterson P, Peterson H. C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. PLoS Comput Biol. 2017;13:e1005766.
64. Quach H, Rotival M, Pothlichet J, Loh YE, Dannemann M, Zidane N, Laval G, Patin E, Harmant C, Lopez M, et al. Genetic adaptation and neandertal admixture shaped the immune system of human populations. Cell. 2016;167:643–e656617.
65. Nedelec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier JC, Freiman A, Sams AJ, Hebert S, et al. Genetic ancestry and Natural Selection Drive Population differences in Immune responses to pathogens. Cell. 2016;167:657–e669621.
66. Panopoulos AD, D'Antonio M, Benaglio P, Williams R, Hashem SI, Schuldt BM, DeBoever C, Arias AD, Garcia M, Nelson BC, et al. iPSCORE: a resource of 222 iPSC lines enabling functional characterization of genetic variation across a Variety of Cell types. Stem Cell Rep. 2017;8:1086–100.
67. Pashos EE, Park Y, Wang X, Raghavan A, Yang W, Abbey D, Peters DT, Arbelaez J, Hernandez M, Kuperwasser N, et al. Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-Associated loci. Cell Stem Cell. 2017;20:558–e570510.
68. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, Bala S, Bensaddek D, Casale FP, Culley OJ, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature. 2017;546:370–5.
69. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.
70. Hoffman GE, Bendl J, Voloudakis G, Montgomery KS, Sloofman L, Wang YC, Shah HR, Hauberg ME, Johnson JS, Girdhar K, et al. CommonMind

Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. Sci Data. 2019;6:180.

71. Guelfi S, D'Sa K, Botia JA, Vandrovcova J, Reynolds RH, Zhang D, Trabzuni D, Collado-Torres L, Thomason A, Quijada Leyton P, et al. Regulatory sites for splicing in human basal ganglia are enriched for disease-relevant information. Nat Commun. 2020;11:1041.
72. Young AMH, Kumasaka N, Calvert F, Hammond TR, Knights A, Panousis N, Park JS, Schwartzentruber J, Liu J, Kundu K, et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. Nat Genet. 2021;53:861–8.
73. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
74. Theusch E, Chen YI, Rotter JI, Krauss RM, Medina MW. Genetic variants modulate gene expression statin response in human lymphoblastoid cell lines. BMC Genomics. 2020;21:555.
75. Peng S, Deyssenroth MA, Di Narzo AF, Cheng H, Zhang Z, Lambertini L, Ruusalepp A, Kovacic JC, Bjorkegren JLM, Marsit CJ, et al. Genetic regulation of the placental transcriptome underlies birth weight and risk of childhood obesity. PLoS Genet. 2018;14: e1007799.
76. Steinberg J, Southam L, Roumeliotis TI, Clark MJ, Jayasuriya RL, Swift D, Shah KM, Butterfield NC, Brooks RA, McCaskie AW, et al. A molecular quantitative trait locus map for osteoarthritis. Nat Commun. 2021;12:1309.
77. Schwartzentruber J, Foskolou S, Kilpinen H, Rodrigues J, Alasoo K, Knights AJ, Patel M, Goncalves A, Ferreira R, Benn CL, et al. Molecular and functional variation in iPSC-derived sensory neurons. Nat Genet. 2018;50:54–61.
78. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
79. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.
80. Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, Fumis L, Miranda A, Carvalho-Silva D, Buniello A, et al. Open targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res. 2021;49:D1311–20.
81. King EA, Dunbar F, Davis JW, Degner JF. Estimating colocalization probability from limited summary statistics. BMC Bioinformatics. 2021;22:254.
82. Lin D. An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1998;98:296–304.
83. Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J Artificial Intellig Res. 1999;11:95–130.
84. Greene D, Richardson S, Turro E. ontologyX: a suite of R packages for working with ontological data. Bioinformatics. 2017;33:1104–6.
85. Boughton AP, Welch RP, Flickinger M, VandeHaar P, Taliun D, Abecasis GR, Boehnke M. LocusZoom.js: interactive and embeddable visualization of genetic association study results. Bioinformatics. 2021;37:3017–8.
86. 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. 2020;82:1273–300.
87. Nicolas G, Wallon D, Charbonnier C, Quenez O, Rousseau S, Richard AC, Rovelet-Lecrux A, Coutant S, Le Guennec K, Bacq D, et al. Screening of dementia genes by whole-exome sequencing in early-onset Alzheimer disease: input and lessons. Eur J Hum Genet. 2016;24:710–6.
88. Laurin N, Brown JP, Morissette J, Raymond V. Recurrent mutation of the gene encoding sequestosome 1 (SQSTM1/p62) in Paget disease of bone. Am J Hum Genet. 2002;70:1582–8.
89. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R Jr, Ellis MC, Fullan A, et al. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet. 1996;13:399–408.
90. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, Bjornsson S, Stefansson H, Sulem P, Gudbjartsson D, Maloney J, et al. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature. 2012;488:96–9.
91. Kristjansson RP, Benonisdottir S, Davidsson OB, Oddsson A, Tragante V, Sigurdsson JK, Stefansdottir L, Jonsson S, Jensson BO, Arthur JG, et al. A loss-of-function variant in ALOX15 protects against nasal polyps and chronic rhinosinusitis. Nat Genet. 2019;51:267–76.
92. Morris AP, Le TH, Wu H, Akbarov A, van der Most PJ, Hemani G, Smith GD, Mahajan A, Gaulton KJ, Nadkarni GN, et al. Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. Nat Commun. 2019;10:29.
93. Beaudoin M, Gupta RM, Won HH, Lo KS, Do R, Henderson CA, Lavoie-St-Amour C, Langlois S, Rivas D, Lehoux S, et al. Myocardial infarction-Associated SNP at 6p24 interferes with MEF2 binding and associates with PHACTR1 expression levels in human coronary arteries. Arterioscler Thromb Vasc Biol. 2015;35:1472–9.
94. Lopez-Isac E, Smith SL, Marion MC, Wood A, Sudman M, Yarwood A, Shi C, Gaddi VP, Martin P, Prahalad S, et al. Combined genetic analysis of juvenile idiopathic arthritis clinical subtypes identifies novel risk loci, target genes and key regulatory mechanisms. Ann Rheum Dis. 2021;80:321–8.
95. Yang J, McGovern A, Martin P, Duffus K, Ge X, Zarrineh P, Morris AP, Adamson A, Fraser P, Rattray M, Eyre S. Analysis of chromatin organization and gene expression in T cells identifies functional genes for rheumatoid arthritis. Nat Commun. 2020;11:4402.
96. Zhao SS, Mackie SL, Larsson SC, Burgess S, Yuan S. Modifiable risk factors and inflammation-related proteins in polymyalgia rheumatica: genome-wide meta-analysis and Mendelian randomisation. *Rheumatology (Oxford)* 2024.
97. Zhao SS, Gill D. Genetically proxied IL-6 receptor inhibition and risk of polymyalgia rheumatica. Ann Rheum Dis. 2022;81(10):1480–2.
98. Dasgupta B, Unizony S, Warrington KJ, Sloane Lazar J, Giannelou A, Nivens C, Akinlade B, Wong W, Lin Y, Buttgereit F, et al. LB0006 sarilumab in patients with relapsing polymyalgia rheumatica: a phase 3, multicenter, randomized, double blind, placebo controlled trial (SAPHYR). Ann Rheum Dis. 2022;81:210–1.
99. Kanai M, Elzur R, Zhou W, Daly MJ, Finucane HK, Global Biobank Meta-analysis I. Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. Cell Genom. 2022;2:100210.
100. Hormozdiari F, Zhu A, Kichaev G, Ju CJ, Segre AV, Joo JWJ, Won H, Sankararaman S, Pasaniuc B, Shifman S, Eskin E. Widespread allelic heterogeneity in Complex traits. Am J Hum Genet. 2017;100:789–802.
101. Zhu A, Matoba N, Wilson EP, Tapia AL, Li Y, Ibrahim JG, Stein JL, Love MI. MRLocus: identifying causal genes mediating a trait through bayesian estimation of allelic heterogeneity. PLoS Genet. 2021;17: e1009455.
102. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481–7.
103. Yin X, Bose D, Kwon A, Hanks SC, Jackson AU, Stringham HM, Welch R, Oravilahti A, Fernandes Silva L, FinnGen, et al: Integrating transcriptomics, metabolomics, and GWAS helps reveal molecular mechanisms for metabolite levels and disease risk. Am J Hum Genet 2022, 109:1727–1741.

Publisher's Note