

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

**Authors:** Samuel Lessard<sup>1</sup>, Michael Chao<sup>1</sup>, Kadri Reis<sup>2</sup>, FinnGen, Estonian Biobank Research Team, Mathieu Beauvais<sup>3</sup>, Deepak K. Rajpal<sup>4a,5</sup>, Srinivas Shankara<sup>1</sup>, Jennifer Sloane<sup>6</sup>, Priit Palta<sup>2</sup>, Katherine Klinger<sup>7</sup>, Emanuele de Rinaldis<sup>1</sup>, Shameer Khader<sup>1\*</sup>, Clément Chatelain<sup>1\*</sup>

- 1. Precision Medicine & Computational Biology, Sanofi US, Cambridge, MA, USA**
  - 2. Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia**
  - 3. Digital R&D Data & Computational Sciences, Sanofi, Chilly-Mazarin, France**
  - 4. Translational Sciences, Sanofi US, Framingham, MA, USA**
  - 5. Pre-Clinical and Translational Sciences, Takeda, MA, USA**
  - 6. Immunology & Inflammation Development, Sanofi US, Cambridge, MA, USA**
  - 7. Genetics Research, Sanofi US, Cambridge, MA, USA**
- a. Employee of Sanofi US at the time of study**

**\*These authors contributed equally**

**Corresponding author**

<b>Name</b>	Samuel Lessard
<b>Address</b>	Sanofi
	Cambridge, MA, USA
<b>E-mail</b>	<a href="mailto:Samuel.lessard@sanofi.com">Samuel.lessard@sanofi.com</a>

## ABSTRACT

1 **BACKGROUND:** Therapeutic targets supported by genetic evidence from genome-wide association  
2 studies (GWAS) show higher probability of success in clinical trials. GWAS is a powerful approach to  
3 identify links between genetic variants and phenotypic variation; however, identifying the genes driving  
4 associations identified in GWAS remains challenging. Integration of molecular quantitative trait loci  
5 (molQTL) such as expression QTL (eQTL) using mendelian randomization (MR) and colocalization  
6 analyses can help with the identification of causal genes. Careful interpretation remains warranted  
7 because eQTL can affect the expression of multiple genes within the same locus. **METHODS:** We used a  
8 combination of genomic features that include variant annotation, activity-by-contact maps, MR, and  
9 colocalization with molQTL to prioritize causal genes across 4,611 disease GWAS and meta-analyses  
10 from biobank studies, namely FinnGen, Estonian Biobank and UK Biobank. **RESULTS:** Genes identified  
11 using this approach are enriched for gold standard causal genes and capture known biological links  
12 between disease genetics and biology. In addition, we find that eQTLs colocalizing with GWAS are  
13 statistically enriched for corresponding disease-relevant tissues. We show that predicted directionality  
14 from MR is generally consistent with matched drug mechanism of actions (>78% for approved drugs).  
15 Compared to the nearest gene mapping method our approach also shows a higher enrichment in  
16 approved therapeutic targets (risk ratio 1.38 vs 2.06). Finally, using this approach, we detected a novel  
17 association between the IL6 receptor signal transduction gene *IL6ST* and polymyalgia rheumatica, an  
18 indication for which sarilumab, a monoclonal antibody against IL-6, has been recently approved.

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

22 **KEYWORDS:** Genome-wide association study; molecular quantitative trait loci; causal inference;  
23 therapeutic targets; interleukin 6; polymyalgia rheumatica; mendelian randomization

## 24 **BACKGROUND**

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

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

43 While these examples clearly show that disease-associated genetic information is important for drug  
44 development, it remains a challenge to accurately assign causal genes driving disease risk from GWAS as

45 most variants identified in GWAS fall in non-coding regions of the genome.[22-24] While it's been  
46 observed that the nearest gene often is the causal gene, this is not a guarantee as genetic variants can  
47 influence traits over large genomic distances.[25] In addition, this observation may be biased towards  
48 genes that have been well-characterized because they fall at the center of genetic association  
49 signals.[26]

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

66 Here, we sought to use currently available eQTL information to identify disease relevant genes in the  
67 context of drug discovery. We first derived a simple approach to prioritize causal genes based on  
68 MR[40], eQTL colocalization[31], activity-by-contact (ABC) enhancer-promoter interactions[41], and

69 variant annotations[42]. We used this combinatorial approach as a way to mitigate the pleiotropic  
70 effect of eQTL while retaining important information about directionality. We show that this approach  
71 enriches for gold standard genes[26] and captures known target biology. In addition, genes prioritized  
72 by this approach are enriched for drug targets with successful clinical trials, and directionality inferred  
73 by MR or coding variants recapitulate drug mechanisms of action (MoA). Finally, we show that this  
74 approach can be used to identify drug indication expansion opportunities using genes related to the IL6-  
75 R as a case study and identify a novel association between *IL6ST* and polymyalgia rheumatica.

76

## 77 **METHODS**

### 78 Estonian Biobank GWAS

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

82 All EstBB participants have been genotyped at the Core Genotyping Lab of the Institute of Genomics,  
83 University of Tartu, using Illumina Global Screening Array v1.0 and v2.0. Samples were genotyped and  
84 PLINK format files were created using Illumina GenomeStudio v2.0.4. Individuals were excluded from the  
85 analysis if their call-rate was <95% or if sex defined based on heterozygosity of X chromosome did not  
86 match sex in phenotype data. Before phasing and imputation, variants were filtered by call-rate <95%,  
87 HWE p value  $< 1e-4$  (autosomal variants only), and minor allele frequency <1%. Variant positions were  
88 updated to b37 and all variants were changed to be from TOP strand using GSAMD-24v1-  
89 0\_20011747\_A1-b37.strand.RefAlt.zip files from <https://www.well.ox.ac.uk/~wrayner/strand/> webpage.  
90 Chip data pre-phasing was done using Eagle v2.3 software [43] (number of conditioning haplotypes  
91 Eagle uses when phasing each sample was set to:  $-Kpbwt=20000$ ) and imputation was done using Beagle

92 v.28 Sep18.7932 [44] with effective population size  $n_e \approx 20,000$ . Population specific imputation  
93 reference panel of 2297 WGS samples was used.[44]

#### 94 FinnGen

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

#### 101 Disease GWAS processing

102 We retrieved GWAS results from FinnGen release 10 (R10), UK Biobank pan meta-analysis[45], and a  
103 meta-analyses between FinnGen, UK Biobank, and Estonian biobank. For simplicity, we use the term  
104 GWAS to refer to both single study GWAS and meta-analyses throughout the manuscript. In total, we  
105 included 4,611 GWAS with at least one variant with  $P < 1 \times 10^{-6}$ . When appropriate, we lifted over variants  
106 from hg38 to hg19 using the liftOver R package[46]. Variant with a minor allele frequency (MAF) <  
107 0.0001 were excluded from the analysis. For each GWAS, we considered genes located within 250kb of a  
108 variant with  $P < 1 \times 10^{-6}$  for further analysis. For gold standard and clinical trial enrichment analyses  
109 (described below), only genome-wide significant loci were included ( $P < 5 \times 10^{-8}$ ). We excluded the human  
110 leukocyte antigen (HLA) region in all analyses.

#### 111 Disease EFO mapping

112 In order to perform semantic integration of genetic data and clinical trial data, the ontological system  
113 Experimental Factor Ontology (EFO) was used. We used the EFO to map traits to their corresponding

114 EFO categories and when multiple EFO terms could be mapped to the same trait, we assigned the trait  
115 to each possible term. We used the EFO version 3.52.0 (<https://www.ebi.ac.uk/efo/>).

#### 116 Variant annotation

117 We used variant effect predictor (VEP v102) [42] to annotate the impact of variants with the following  
118 options: --everything --offline --check\_existing. Coding variants were defined as those impacting protein  
119 coding transcript annotated as missense variant or predicted to have “high” impact. We also retrieved  
120 predicted gain or loss of function (GoLoF) variants from LoGoFunc[47], and linked non-coding variants to  
121 genes using activity-by-contact (ABC) maps[41]. ABC scores represent the contribution of an enhancer to  
122 the regulation of gene, measured by multiplying the estimates of enhancer activity and three-  
123 dimensional contact frequencies between enhancers and promoters. ABCmax refers to variant-gene  
124 pairs with the highest ABC score. We also retrieved disease mutations from the Human Gene Mutation  
125 Database (HGMD) (license acquired via Qiagen, Maryland)[48]

#### 126 Mendelian randomization & colocalization

127 We performed transcriptome wide MR using the R package TwoSampleMR [40]. When more than one  
128 instrument was present, we used the inverse variant weighted approach, otherwise we used the Wald  
129 Ratio approach. We considered the following exposures: protein quantitative trait loci (pQTL) from Sun  
130 et al [49], and expression quantitative trait loci from Blueprint[50], eQTLGen [51] and other datasets  
131 from the EBI eQTL catalogue[51-75]. In total, 110 molQTL from 26 studies were included. For each of  
132 those studies, we excluded variants with a MAF < 1%. We clumped variants using PLINK[76] using the  
133 options --clump-p1 1 --clump-p2 1 --clump-r2 0.01 --clump-kb 10000 and using the European ancestry  
134 subset of the 1000 Genomes Project phase 3 data as reference[77]. We only considered genes 250kb  
135 around significant loci in this analysis. For each QTL, independent variants with  $P < 1 \times 10^{-4}$  were used as  
136 instruments. For genes with significant MR results (false discovery rate < 0.05), we also performed

137 colocalization analysis using COLOC[31], using a region of 250kb around the local lead GWAS variant.

138 Harmonization between the QTL and GWAS datasets was performed using the harmonise\_data function  
 139 in the TwoSampleMR package[40]. Only autosomes were included in this analysis.

140 Causal gene prioritization

141 We prioritized genes as putatively causal using a combination of evidence including MR, colocalization  
 142 H4 posterior probabilities (PP) with molQTL, presence of an associated GoLoF variant[47] or other  
 143 coding variants, distance to lead variant, and enhancer-promoter ABC scores[41]. Specifically, we ranked  
 144 genes as follow:

Rank	Criteria
<b>Very High</b>	Lead GoLoF 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
<b>High</b>	Lead coding variant; Or Associated ( $P < 1 \times 10^{-6}$ ) GoLoF variant; Or Colocalization (H4 PP> 80%) with molQTL of the target gene in >2 dataset or significant MR with protein QTL (q-value < 0.05); and maximum ABC score for an associated variant overlapping a regulatory element ( $P < 1 \times 10^{-6}$ )
<b>Moderate</b>	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
<b>Weak</b>	Colocalization with molQTL of the target gene (H4 PP>30%) Or 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
<b>Very weak</b>	Significant MR with eQTL Or ABC link (any score) between an element overlapping the lead variant and target



	gene
--	------

145

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

#### 149 Enrichment of gold standard genes

150 We retrieved GWAS causal gene gold standards supported by functional experiments or observations or  
151 expert curation from Open Targets (version 191108).[26, 78] We linked the current analysis with the  
152 gold standard gene list using Ensembl gene identifiers and EFO codes. That is, for a given gene-disease  
153 pair in the current analysis, we consider it a gold standard association if the gene and GWAS EFO code  
154 are present in the Open Targets gold standard gene-disease set. For each indication, we filtered out  
155 genes not represented in loci where a gold standard gene is located. We calculated the enrichment of  
156 gold standard genes in prioritized genes by different features or rankings as described above using  
157 Fisher exact tests. In addition, we calculated the precision (number of prioritized genes that are gold  
158 standards over all prioritized genes), recall (number of prioritized genes that are gold standards over the  
159 total number of gold standard genes), and F1 scores for each feature.

#### 160 Single gene colocalizing cell-type eQTL enrichment

161 To identify enriched colocalizing cell types for single genes, we calculated the ratio of indications for  
162 which this gene is prioritized to be causal by a given molQTL dataset (H4 PP > 80%) over the total  
163 number of prioritized indications (as defined by unique EFO) for that gene. We collapsed GWAS by  
164 corresponding EFO code so that a gene was only counted once per indication (and not multiple times for  
165 GWAS of the same disease). We then compared this ratio to the fraction of prioritized indications via  
166 colocalization of the same eQTL dataset over all prioritized indications genome wide. In other words, we

167 are looking for genes that show an overrepresentation of colocalizing eQTL cell types across all  
168 associated indications compared to the genome-wide distribution. This corresponds to the following  
169 contingency table:

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

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

#### 175 Enrichment of disease categories for single genes

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

$$\begin{array}{cc} \sum_i D_{ijc} & \sum_i \sum_{c \neq C} D_{ijc} \\ \sum_i \sum_{j \neq J} D_{ijc} & \sum_i \sum_{c \neq C} \sum_{j \neq J} D_{ijc} \end{array}$$

182 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  
183 ratios were calculated using Fisher exact tests. FDR adjusted  $P$ -values  $< 0.05$  were considered  
184 significant.

#### 185 Disease colocalizing molQTL cell-type enrichment

186 We identify enriched cell types in GWAS disease EFO categories supported by colocalization as in King et  
187 al. 2021.[79] Briefly, we extracted all GWAS colocalizing molQTL (H4 probability > 0.8). Then, for a given  
188 cell type  $K$  and disease category  $l$ , we generated the following contingency table:

$$\begin{array}{cc} \sum_j C_{ljk} & \sum_j \sum_{k \neq K} C_{ljk} \\ \sum_j \sum_{i \neq l} C_{ijK} & \sum_j \sum_{k \neq K} \sum_{i \neq l} C_{ijk} \end{array}$$

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

#### 195 Drug target- indication pairs in clinical trials

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

206 In order to map gene-disease pairs in the genetic data to target-indication pairs in the drug data, we  
207 used experimental factor ontology (EFO), which provided a systematic description of many data

208 elements available in EBI databases. A target-indication pair is said to have genetic evidence if there is  
209 genetic evidence of association between the gene and disease sufficiently similar to the indication,  
210 based on semantic similarity. Two methods were used to calculate semantic similarity matrix.[80, 81]  
211 Semantic similarities between each pair of EFO headings were computed in the ontologySimilarity R  
212 package.[82] The average of the two methods was calculated and standardized similarities had a  
213 maximum value of 1 for each disease or indication. Two diseases are considered similar if the similarity  
214 is greater than or equal to a previously published value of 0.7.[11]

#### 215 Prediction of drug mechanism of action directionality

216 We retrieved information about drug mechanism of action from the Informa Pharma Intelligence  
217 dataset described above. For targets for which *decreased* expression or loss of function (LoF) is  
218 beneficial, we considered datasets with the following keywords: “antagonist”, “inhibitor”, and  
219 “degrader”. For targets for which *increased* expression or function is beneficial, we considered the  
220 following keywords: “agonist”, and “activator”. We considered drugs and targets in phase II clinical trial  
221 or above. We performed two analyses to infer directionality from GWAS. First, we assess directionality  
222 using the effect size of low-frequency lead coding variant (MAF < 5%). We assumed that these variants  
223 are disruptive or LoF. Therefore, a LoF coding variant associated with increased risk suggests that a drug  
224 MoA of agonist or activator would be beneficial, whereas for a protective LoF coding variant, an  
225 inhibitor or antagonist would be beneficial. Next, we assessed directionality based on the direction of  
226 effect of gene expression on disease risk predicted by MR using molQTL as exposure (q-value < 0.05).  
227 We included only molQTL colocalizing with local GWAS signal (H4 PP > 80%). For gene-disease pairs  
228 supported by multiple colocalizing molQTL, a consensus direction was inferred if the MR direction of  
229 effect was consistent across > 75% of the molQTL. Here, a negative consensus MR direction suggests  
230 that increased gene expression leads to decreased disease risk. Therefore, an activator or agonist drug  
231 targeting this gene would be beneficial. Conversely, a positive consensus MR direction suggests that

232 increased gene expression increases disease risk, and an inhibitor or antagonist drug would be  
233 beneficial. We calculated enrichment of concordant direction of effect between GWAS and drug MoA  
234 using Fisher exact tests.

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

236 We aimed to apply our proposed approach to a specific case example. Using the causal gene  
237 prioritization and GWAS datasets described above, we extracted all disease GWAS for which *IL6*, *IL6R*, or  
238 *IL6ST* were predicted to be causal. We predicted directionality of effect of gene expression on disease  
239 risk by MR as above using a threshold of q-value < 0.05. We generated local association of plots molQTL  
240 and GWAS using LocusZoom[83]. We performed fine-mapping of *IL6ST* genetic variants associated with  
241 polymyalgia rheumatica using SuSIE[84] as previously described for FinnGen[7].

242

## 243 **RESULTS**

### 244 *Prioritization of putative causal genes in thousands of GWAS*

245 We aimed to prioritize causal genes across 4,611 GWAS from 3 different sources (**Table 1**): UK Biobank  
246 (UKB)[45], FinnGen release 10 (R10), and meta-analyses of UK Biobank, FinnGen R10, and Estonian  
247 biobank.[6] For simplicity, we refer to both single studies and meta-analyses as GWAS throughout the  
248 manuscript. While molecular QTLs (molQTL) such as expression quantitative trait loci (eQTL) have been  
249 used previously to prioritize causal genes, they are often pleiotropic with the same variant associated  
250 with multiple genes within the same locus.[26, 32, 33] Additional genomic information such as the ABC  
251 model have been shown to increase performance to identify causal genes, in particular when selecting  
252 genes with the highest ABC score (ABCmax).[41] Therefore, we derived a ranking scheme to prioritize  
253 genes using different features including ABC, molQTL, presence of an associated coding or gain or loss of

254 function (GoLoF) variants, and distance to lead variant (**Figure 1A, methods**). We integrated 110 molQTL  
255 datasets from 26 studies using MR to infer causality and directionality of gene expression on disease  
256 risk. We also performed colocalization analysis to confirm that both GWAS or meta-analyses and molQTL  
257 signals shared at least one causal variant. Top ranking genes were selected as those that either  
258 contained an associated lead coding variant or were supported by both ABCmax and colocalization  
259 across >2 cell types or tissues. We did not include distance to lead variant for higher ranks because we  
260 wanted to first prioritize genes for which we could identify potential biological mechanisms. However,  
261 for loci without such evidence, or in cases where multiple genes showed identical ranks, the nearest  
262 gene to the lead variant was selected as the putative causal gene if it was among the best candidates.  
263 Overall, between 1.1 and 1.4 genes were prioritized per locus (before breaking ties with the nearest  
264 gene), with 17-49% of loci supported by molQTL colocalization or coding variants (**Table 1**).

#### 265 *Enrichment of genomic features for gold standard genes*

266 Comparing the enrichment of different genomic features alone for curated gold standard genes[26], we  
267 found a strong enrichment for genes supported by ABCmax with lead variant (Odds ratio (OR)=8.0-18.7,  
268  $P=0.0002-4 \times 10^{-6}$ ) (**Additional file 1: Figure S1; Additional file 2: Table S1**). molQTL colocalization also  
269 enriched for gold standard genes (colocalization H4 posterior probability (PP) > 95%, OR=3.4-17.7,  
270  $P=0.001-2 \times 10^{-12}$ ). However, the strongest enrichment was generally observed for genes with associated  
271 lead coding variants[47] (OR>36.2,  $P=0.0002-2 \times 10^{-10}$ ) and the nearest gene (OR=17.7-38.7,  $P=3 \times 10^{-9}-$   
272  $1 \times 10^{-25}$ ). The strong enrichment for nearest genes is expected given that the gene closest to the lead  
273 variant is often the causal gene. In addition, several of the gold standard genes have been selected  
274 because they are supported by coding variants or tend to fall in the center of GWAS peaks and have  
275 been investigated more closely[26]. However, when using these features in combination, we found that  
276 our ranking approach performed well and generally better than selecting the nearest gene alone, with a

277 mean increase in F1 score of 0.08 (-0.03 – 0.23) (**Additional file 1: Figure S2-S3; Additional file 2: Table**  
278 **S1**).

### 279 *Gain and Loss of function variants identify genes linked to monogenic disorders*

280 Integrating information about GoLoF variants retrieved variants linked to monogenic disorders including  
281 *PSEN1* with Alzheimer’s disease (AD)[85] (rs764971634, p.Ile437Val,  $P=2 \times 10^{-12}$ ), *SQSTM1* and Paget’s  
282 disease[86] (rs104893941, p.Pro392Leu,  $P=6 \times 10^{-11}$ ), and *HFE* and disorders of iron metabolism[87]  
283 (rs1800562, p.Cys282Tyr,  $P=1 \times 10^{-178}$ ) (**Figure 1B; Additional file 2: Table S3**). We also identified  
284 protective GoLoF variants such as *APP* p.Ala673Thr (rs63750847,  $P=7 \times 10^{-11}$ ) reducing odds of developing  
285 AD[88], and *ALOX15* p.Thr560Met protecting against nasal polyps (rs34210653,  $P=2 \times 10^{-15}$ )[89]. Of 208  
286 genes prioritized with at least one predicted GoLoF variant, 179 had at least one disease mutation  
287 reported in the Human Gene Mutation Database (HGMD)[48] (OR = 2.3 [1.5-3.6],  $P=5 \times 10^{-6}$ ). Potential  
288 novel associations included *COLGALT2* and arthrosis (rs35937944, p.Tyr212Cys,  $P=2 \times 10^{-14}$ ), *LRG5* and  
289 carcinoid syndrome (rs200138614, p.Cys712Phe,  $P=4 \times 10^{-9}$ ), and *GREB1* and female infertility  
290 (rs755857714, p.Arg1339His,  $P=4 \times 10^{-9}$ ).

### 291 *Colocalizing molQTL link genes to diseases and pathogenic tissues*

292 Prioritized candidate causal genes showed enrichment in disease colocalizing molQTLs related to their  
293 known function. For instance, colocalizing molQTL for prioritized genes supported associations with  
294 disease categories such as *EDNRA*, *LPA* and *FGF5* with cardiovascular diseases ( $P < 2 \times 10^{-16}$ ), *TSLP*, *IL33* and  
295 *CHRNA3* and respiratory system diseases ( $P < 7 \times 10^{-21}$ ), and *IL23R*, *TYK2*, *IL10* and immune system disease  
296 ( $P < 5 \times 10^{-11}$ ) (**Figure 1C-D; Additional file 2: Table S4**). In addition, we found an enrichment of disease  
297 colocalizing eQTLs in kidney cortex for *FGF5*, a gene expressed during kidney development and  
298 associated with kidney function ( $P=4 \times 10^{-15}$ )[90] (**Figure 1E; Additional file 2: Table S5**). Other examples  
299 include artery eQTLs for the cardiovascular diseases associated gene *PHACTR1*[91] ( $P=1 \times 10^{-9}$ ); the

300 lysosomal acid lipase (*LIPA*) gene and microglia eQTLs ( $P=1\times 10^{-10}$ ); and the *ABO* with plasma pQTL  
301 ( $P=1\times 10^{-20}$ ). Finally, we confirmed that enriched colocalizing eQTLs matched the expected pathogenic  
302 tissues and cell-types of different disease categories (**Figure 1F; Additional file 2: Table S6**). For instance,  
303 after grouping eQTL of similar tissues and cell types together, we found a strong enrichment of genes  
304 with artery and heart eQTL colocalizing with cardiovascular disease GWAS ( $P<9\times 10^{-17}$ ). We found  
305 similar enrichment for T cell and thyroid eQTLs in endocrine system diseases ( $P<3\times 10^{-8}$ ); blood,  
306 lymphoblastoid cell line, monocytes, neutrophil, and T cells with immune system diseases ( $P<4\times 10^{-6}$ );  
307 and fibroblasts and musculoskeletal diseases ( $P<4\times 10^{-6}$ ). Treating each eQTL data separately revealed  
308 additional associations with tissues or cell subsets including brain cortex and diseases of the visual  
309 system ( $P<6\times 10^{-6}$ ); cerebellum and nervous system diseases ( $P<4\times 10^{-6}$ ); regulatory T cells and endocrine  
310 system diseases ( $P<9\times 10^{-9}$ ); and T helper 17 cells and digestive system diseases ( $P<5\times 10^{-7}$ ) (**Additional**  
311 **file 1: Figure S4; Additional file 2: Table S7**). Overall, the analyses illustrate that in contrast to the  
312 nearest gene approach, inclusion of eQTL can help identify potential pathogenic cell types and tissues.

### 313 *Prioritized genes increase clinical trial probability of success*

314 Building on these results, we tested whether we could use molQTL information of putative causal gene  
315 to drive drug repurposing opportunities or identify potential safety concerns. First, we evaluated  
316 whether the prioritized genes enriched for therapeutic targets with clinical trial success. Clinical trial  
317 information was retrieved from the Citeline Pharma Intelligence project. Consistent with previous  
318 observations, we found that targets with clinical trial success were enriched for features such as  
319 presence of coding variation (**Figure 2A, Additional file 2: Table S8**). For example, gain or loss of  
320 function lead variants demonstrated some of the best predictive performances, in particular using  
321 genetic evidence from the UKB EUR ICD10 (Phase I: Risk ratio (RR)=1.23,  $P=0.104$ ; Phase II: RR=1.33,  
322  $P=0.0688$ ; Phase III: RR=2.08,  $P=0.0023$ ; Approved: RR=2.67,  $P=0.00378$ ). Similar results were observed  
323 across all studies. Use of epigenetic evidence also improved predictions, for example, lead SNPs linked



324 by the ABC model in UKB EUR ICD10 (Phase I: RR=1.33,  $P=0.00484$ ; Phase II: RR=1.4,  $P=0.0162$ ; Phase III:  
325 RR=2.15,  $P=0.000304$ ; Approved: RR=2.82,  $P=0.000622$ ). However, molQTL information alone did not  
326 enrich as much for clinical trial success, for example, colocalizing molQTL with posterior probability >  
327 80% in UKB EUR ICD10 (Phase I: RR=1.22,  $P=0.013$ ; Phase II: RR=1.18,  $P=0.154$ ; Phase III: RR=1.43,  
328  $P=0.0581$ ; Approved: RR=1.71,  $P=0.044$ ). While the overall prioritized genes did not show the strongest  
329 enrichment (UKB ICD10 Phase I: RR=1.24,  $P=0.0006$ ; Phase II: RR=1.17,  $P=0.008$ ; Phase III: RR=1.51,  
330  $P=0.003$ ; Approved: RR=1.60,  $P=0.03$ ), this was likely due to the inclusion of genes with no supportive  
331 evidence other than distance (**Figure 2A**). Indeed, we found that “High” and “Very High” prioritization  
332 ranks were more predictive of successful clinical trial progression (higher risk ratios) than lower-ranking  
333 genes, especially at later clinical trial phases or approval (High + Very high ranks in UKB ICD10 Phase I:  
334 RR=1.16,  $P=0.103$ ; Phase II: RR=1.18,  $P=0.174$ ; Phase III: RR=1.78,  $P=0.00149$ ; Approved: RR=2.06,  
335  $P=0.00637$ ) (**Figure 2B; Additional file 2: Table S9**). In our analysis, distance itself was seldom predictive  
336 or clinical trial success (UKB ICD10 Phase I: RR=1.18,  $P=0.03$ ; Phase II: RR=1.06,  $P=0.061$ ; Phase III:  
337 RR=1.24,  $P=0.61$ ; Approved: RR=1.38,  $P=0.19$ ) especially after excluding loci potentially driven by coding  
338 variants (**Figure 2B**).

### 339 *Inferred directionality from GWAS recapitulate drug mechanisms of action*

340 To understand whether inferred directionality could inform on clinical trial success, we first investigated  
341 the consistency between the direction of effect of coding variants and drug mechanism of action (MoA)  
342 (methods). When considering prioritized genes with lead low-frequency coding variants (minor allele  
343 frequency < 0.05) and clinical trials phase II and above, between 83% and 96% of showed consistent  
344 effect between the minor allele and drug MoA (Fisher  $P=0.08-6 \times 10^{-8}$ , **Figure 2C**). We then asked whether  
345 molQTL could similarly inform on directionality. Using prioritized gene-disease pairs supported by MR (q-  
346 value < 0.05) and colocalization (PP > 80%), we inferred the direction of effect when the predicted MR  
347 effect was consistent across >75% of molQTL datasets for a given gene. This was the case for most gene-

348 disease pairs (**Additional file 1: Figure S5**). Again, direction of effect was generally in agreement with  
349 drug MoA (64-81% agreement, Fisher  $P=4\times 10^{-8}$ - $5\times 10^{-41}$ , **Figure 2D**). Consistency increased when  
350 considering only approved drugs (78-93% agreement, Fisher  $P=3\times 10^{-5}$ - $1\times 10^{-23}$ , **Additional file 1: Figure**  
351 **S6**). Overall, these data suggest that molQTL can be used to inform on drug MoA.

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

353 Finally, we applied our causal gene prioritization approach to a specific use case, that is identifying  
354 potential new indications for drugs targeting the IL6 receptor such as Sarilumab and Tocilizumab, both  
355 drugs approved for rheumatoid arthritis. We extracted diseases prioritized by our approach for genes  
356 related to the receptor, namely *IL6*, *IL6ST*, and *IL6R*. We identified putative causal links between  
357 increased *IL6* expression in CD16 monocytes and increased risk of varicose veins, ischemic heart disease,  
358 coronary atherosclerosis, and atrial fibrillation (MR beta > 0), but decreased risk of asthma and allergy  
359 (MR beta < 0) (**Additional file 1: Figure S7; Additional file 2: Table S10**). eQTL of *IL6* in whole blood also  
360 supported these disease associations, albeit with an opposite predicted direction of effect. Similarly,  
361 *IL6R* expression in multiple tissues including artery, colon, and esophagus was associated with increased  
362 risk of coronary revascularization, coronary atherosclerosis, and abdominal aortic aneurysm (AAA), but  
363 lower risk of lower respiratory diseases and atopic dermatitis. Again, we observed opposite direction of  
364 effect predicted by MR using monocyte or macrophage eQTL as exposure. The associations with  
365 coronary atherosclerosis and AAA were further driven by a lead coding variant in *IL6R*, rs2228145  
366 (Asp358Ala, **Additional file 2: Table S10**). Finally, we found that increased *IL6ST* expression in T cells and  
367 whole blood is predicted to increase the risk of rheumatoid arthritis, systemic connective tissue  
368 disorders, polyarthropathies, other arthritis, autoimmune diseases, and polymyalgia rheumatica (**Figure**  
369 **3A**). The later association has not been reported previously to our knowledge. These associations were  
370 driven by rs7731626 (SuSIE fine-mapping probability >0.99). This variant is located within an intron of  
371 *ANKRD55* and colocalizes with eQTLs for both *ANKRD55* and *IL6ST* (PP > 80%). However, this variant also

372 overlaps an enhancer that shows highest ABC score for *IL6ST* for genes in the region, suggesting the  
373 latter is the causal gene, in line with previous studies[92, 93] (**Figure 3B**). Overall, our approach was able  
374 to capture known associations with IL6-R related genes and identified a new association between *IL6ST*  
375 and polymyalgia rheumatica.

376

## 377 **DISCUSSION**

378 We prioritized disease-associated genes across 4,611 GWAS and meta-analyses from biobank studies  
379 using a combination of MR with molQTL, colocalization analysis, variant effect prediction, and epigenetic  
380 annotations (ABC model). This approach allows the use of molQTL to infer directionality of gene  
381 expression on disease risk, while improving the causal gene prediction compared to using molQTL alone.  
382 Based on combination of these features, we used a ranking approach to prioritize genes within loci and  
383 showed that this approach enriched for gold standard genes. We recover known coding variant  
384 associations, including rare variants in genes linked to monogenic disorders such as *PSEN1* and *APP1* and  
385 Alzheimer's disease, and *SQSTM1* and Paget's disease (**Figure 1B**). Genes prioritized by molQTL also  
386 show enrichment in disease categories related to their function with pathogenic tissue contexts (**Figure**  
387 **1C-F**). Of note, when multiple genes show evidence of colocalization within the same locus, the addition  
388 of epigenetic (ABCmax) information can help prioritize one gene over the others. We note as an  
389 example the association of variants with polymyalgia rheumatica at the *ANRKD55* locus where this gene  
390 would be prioritized using the nearest gene approach. Whereas colocalization alone did not identify a  
391 single causal gene, combination of colocalization and ABCmax identified *IL6ST* as the putative causal  
392 gene. To our knowledge, this is the first report of a GWAS association between *IL6ST* and polymyalgia  
393 rheumatica. *IL6ST* encodes a protein involved in signal transduction for the IL6 receptor pathway.

394 Inhibitors of the IL6 receptor have recently shown success in clinical trials for this indication leading to a  
395 recent approval by the FDA.[94]

396 In line with previous studies[11, 12], we show that therapeutic targets with genetic evidence are  
397 enriched at later clinical trial phases and as targets of approved drugs. In our analysis, using the nearest  
398 gene information alone was not strongly predictive of clinical trial success. The most predictive features  
399 were coding variant annotations and ABC maps. While the later performs well to link causal genes to  
400 diseases, it does not provide information about directionality. We used coding variants and MR with  
401 molQTL to infer directionality of a target on disease risk. Both approaches were generally consistent  
402 with drug MoA matched for the target and disease. These data support that molQTL can be used to  
403 predict drug MoA. However, while we found that in general eQTL were consistent across cell type and  
404 tissues for a given gene and disease (**Additional file 1: Figure S5**), we note that this isn't always the case.  
405 This is exemplified by the IL6-R case study, where all three queried gene displayed inconsistent direction  
406 of effect predicted by MR depending on the molQTL dataset. Future improvement of this approach  
407 should consider prior knowledge on pathogenic cell types or tissues to infer directionality in relevant  
408 contexts. Overall, our analysis suggests that using features such as ABCmax in combination to molQTL  
409 can increase the performance of causal gene inference approaches while informing on directionality  
410 which is crucial for translating GWAS hits to therapies.

411 We note that this study has some limitations. First, we did not perform fine-mapping analyses nor  
412 colocalization approaches that use linkage disequilibrium references. Indeed, we opted to avoid  
413 methods that do not rely on LD references as we used GWAS from various sources, including meta-  
414 analyses where these methods may not be well calibrated.[95] Nevertheless, using fine-mapping  
415 information likely would improve performance, especially in cases where there are multiple causal  
416 variants underlying molQTL or GWAS signals, and would reduce LD contamination[30, 96]. In addition,  
417 we performed MR and colocalization analyses as separate steps. Tools that use a combination of these

418 approaches have been recently developed, which are likely to perform better in case of allelic  
419 heterogeneity[97]. This is evident in the case of *IL6ST*, where MR using eQTL from whole blood from  
420 different sources (GTEx, eQTLGen) lead to inversed estimate of directionality (**Figure 3A**). This difference  
421 was due to different instrument used as only one genetic instrument was included in GTEx whereas 5  
422 independent instruments were included for eQTLGen. We also assume that there is one causal gene per  
423 locus, although it is possible that multiple genes contribute to disease risk. Finally, integrating other  
424 sources of molQTL such as metabolite or splice QTL could help identify putative causal genes as coding  
425 variants and eQTL only cover a fraction of loci (18-45% in this study).[98] While these approaches can be  
426 useful to nominate candidate causal genes and their relationship to diseases, proper functional  
427 validation remains of high importance.

428

## 429 **CONCLUSIONS**

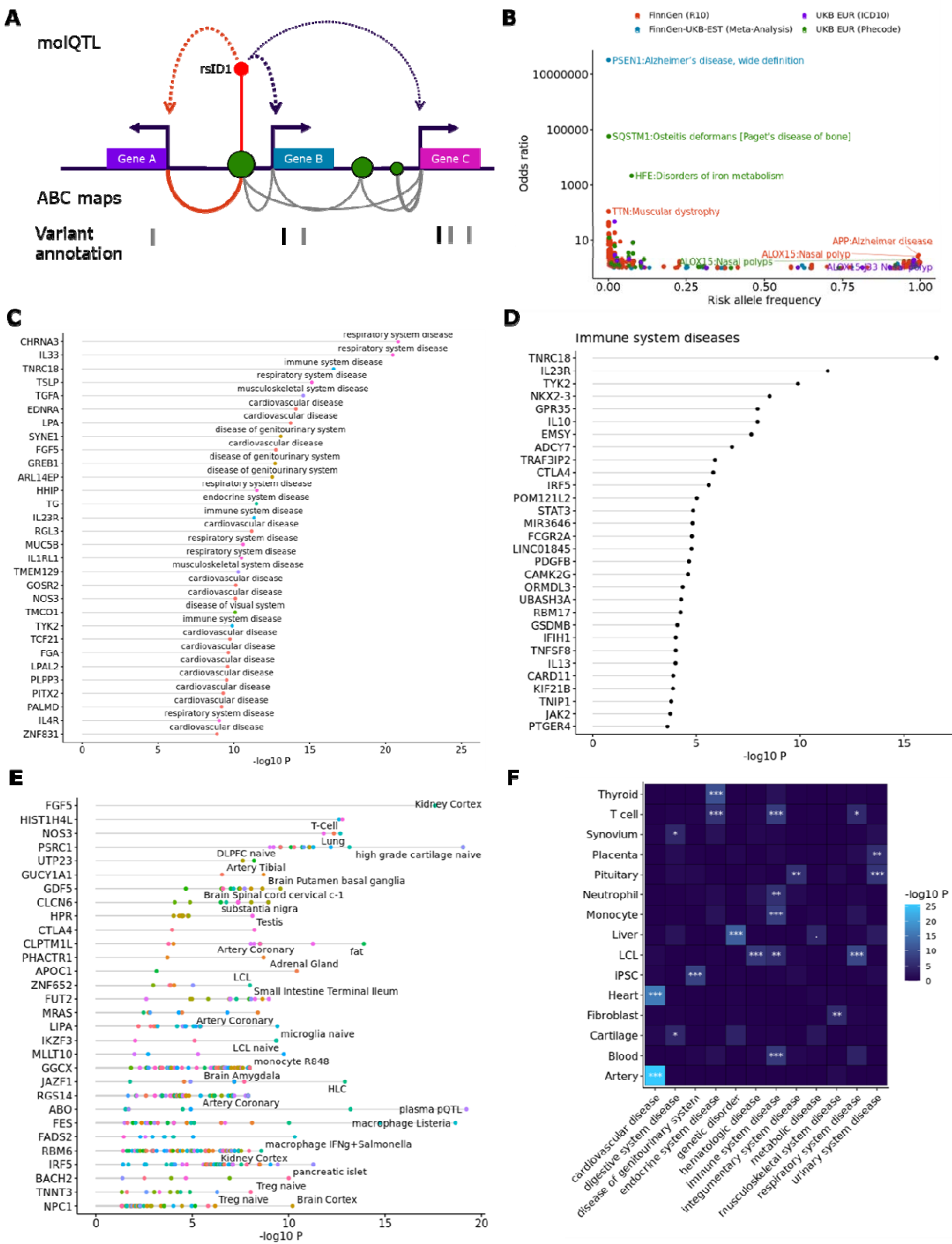
430 We nominated putative causal genes across 4,611 GWAS from biobank studies and public resources by  
431 integrating variant annotations as well as molecular QTL. We show that these prioritized genes recover  
432 known biological relationships in terms of disease and tissue enrichment and are enriched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

437

## 438 **ABBREVIATIONS**

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

447 **FIGURE LEGENDS**



448

449 **Figure 1. Characteristics of prioritized genes via gain or loss of function variants and molQTLs. A)**

450 Features used to prioritize genes in GWAS loci. Genes are ranked based on a combination of features

451 including molQTLs, activity-by-contact (ABC) maps, and variant annotations, including variant effect

452 predictions (VEP) and loss-of-function (LoF) and gain-of-function (GoF) predictions. **B)** Disease-

453 associated predicted GoF and LoF variants captures disease associations with high effect sizes. Lead GoF

454 and LoF variant with GWAS  $P$ -value  $< 5 \times 10^{-8}$  are reported in the figure. Effect of the risk allele (odds

455 ratio) is reported on the y-axis. The x-axis corresponds to the frequency of the risk allele. **C)** Disease

456 category overrepresentation for single genes predicted to be causal. Each dot represents a different

457 associated disease category. Top 30 enrichments are shown. **D)** Same as B, but filtered for genes

458 predicted to be causal and enriched in “Immune system diseases”. Each dot represents a different

459 associated disease category. Top 30 genes are shown. **E)** Overrepresentation of eQTL colocalization for

460 single genes predicted to be causal. Gene-tissue pairs are included only if the gene has the highest rank

461 in a locus for a given associated disease. Top 30 colocalized eQTLs are shown. Each dot represents a

462 different enriched tissue or cell-type. **F)** Enriched colocalizing cell types and tissues by disease

463 categories. Only disease categories and tissues or cell types with at least one significant enrichment are

464 reported in the heatmap. Enrichment  $P$ -values are calculated using Fisher exact test, testing for the

465 enrichment of genes with eQTL colocalizing with GWAS belonging to specific disease categories as in

466 [79]. Tissues and cell-types were collapsed into broader categories before testing for enrichment. For

467 example, tibial, coronary, and aorta arteries were grouped into “artery”.

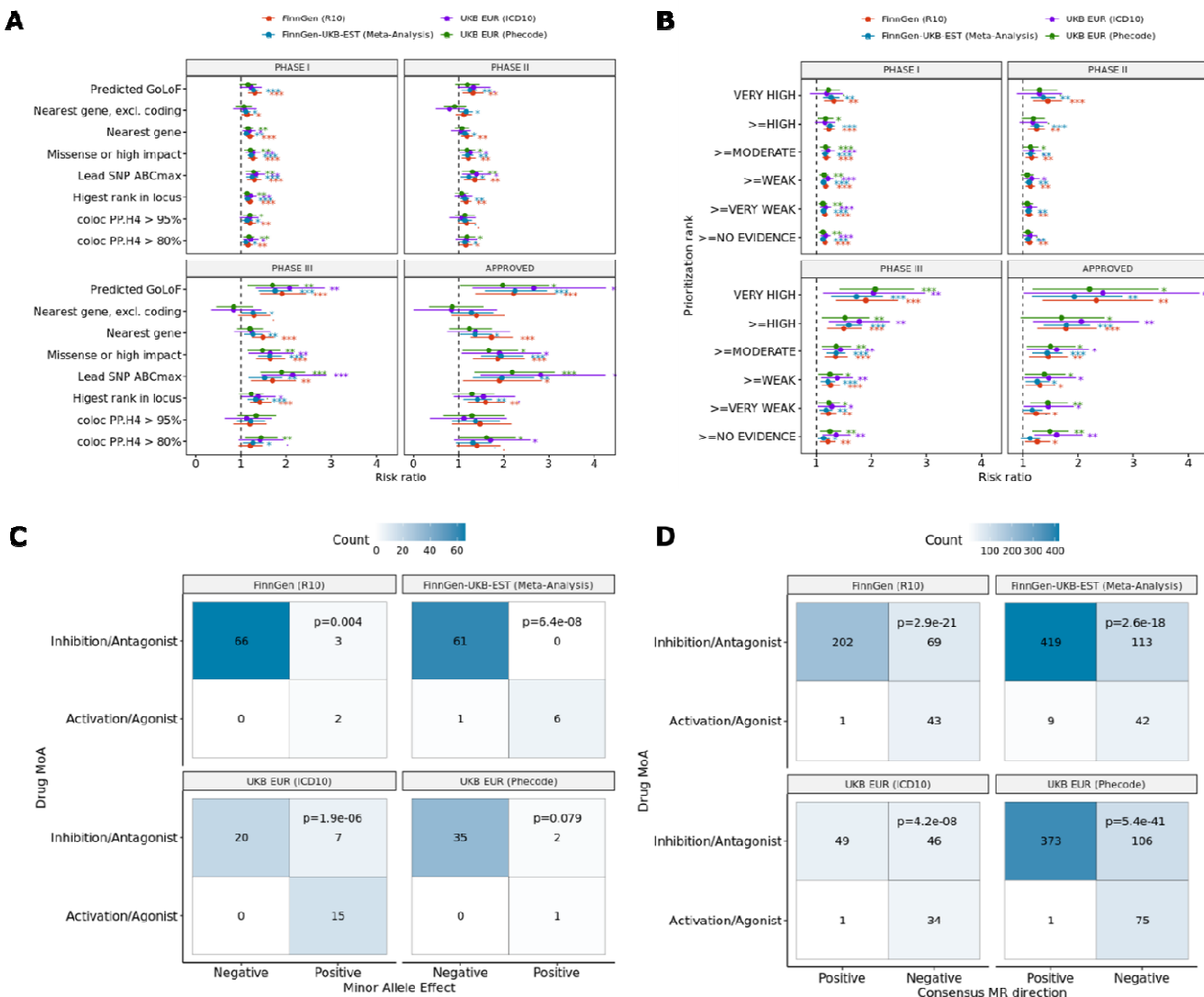
468 molQTL: Molecular QTL; ABC: Activity-By-Contact; LCL: Lymphoblastoid cell lines; iPSC: induced

469 Pluripotent Stem Cells

470 .: Adjusted  $P < 0.1$ ; \*: Adjusted  $P < 0.05$ ; \*\*: Adjusted  $P < 0.01$ ; \*\*\*: Adjusted  $P < 0.001$

471





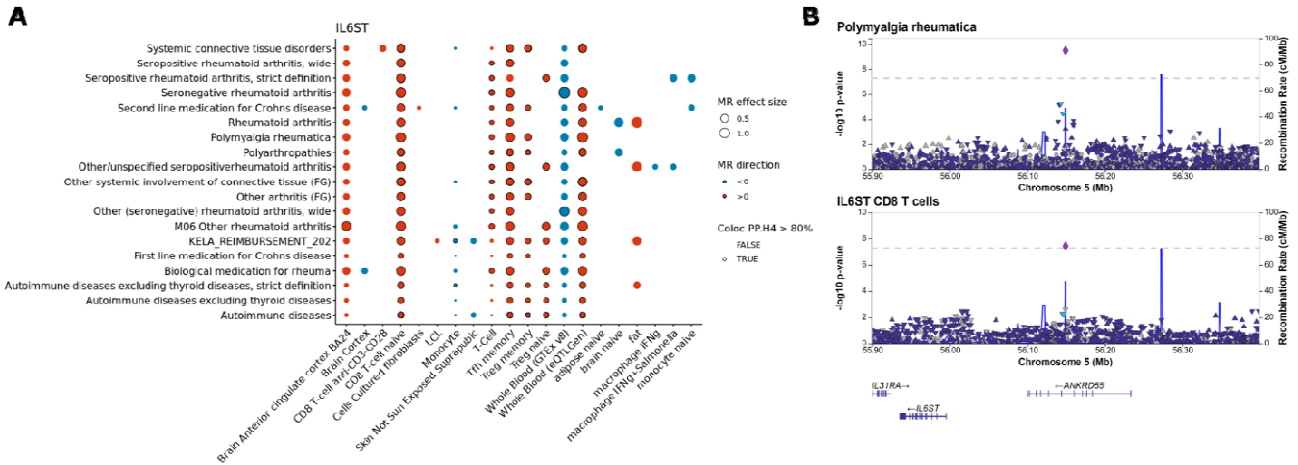
472

473 **Figure 2. Prioritized genes predict clinical trial success.** Enrichment of targets of approved drugs or  
 474 drugs in clinical trials (phase I-III) using genetic evidence from FinnGen, UK Biobank, and biobank meta-  
 475 analyses prioritizing genes using colocalization (posterior probability of colocalization [H4] > 80% or >  
 476 95%), predicted gain of function (GoF) or loss of function (LoF) variants[47], genes with highest  
 477 prioritization rank, ABC score for lead variant, or nearest gene excluding loci with associated coding  
 478 variants. **B)** Enrichment of targets of approved drugs or drugs in clinical trials (phase I-III) using causal  
 479 gene prioritization ranks in FinnGen, UK Biobank, and biobank meta-analyses. **C)** Concordance between  
 480 direction of effect of lead low-frequency coding variants on disease risk, and drug mechanism of action

481 (MoA) for targets in phase II clinical trials or above. We retrieved information about targets, clinical  
482 trials, and drug MoA from the Citeline Pharmacogenomics dataset. We connected this dataset to GWAS  
483 phenotypes using EFO codes and a semantic similarity score > 0.7. We assume that low-frequency  
484 coding variants (minor allele frequency < 5%) are disruptive (LoF). Therefore a negative (protective)  
485 direction of effect would translate into inhibition or antagonism being beneficial (and vice-versa). **D)**  
486 Concordance between the predicted impact of gene expression on disease risk predicted by mendelian  
487 randomization (MR), and drug MoA for targets in phase II clinical trials or above. Information about  
488 targets, clinical trials, and drug MoA were collected from the Citeline Pharmacogenomics dataset and  
489 connected to GWAS phenotypes using EFO codes and a semantic similarity score > 0.7. The direction of  
490 effect of gene expression on disease risk was assessed by MR using molQTL as exposure (q-value < 0.05).  
491 Only molQTL colocalizing with local GWAS signal (H4 posterior probability > 80%) were included. A  
492 consensus direction was inferred if the MR direction of effect was consistent across > 75% of molQTL for  
493 a given gene and disease GWAS. A negative consensus MR direction suggests that increased gene  
494 expression leads to decreased disease risk. Therefore, an activator or agonist drug targeting this gene  
495 would be beneficial. Conversely, a positive consensus MR direction suggests that increased gene  
496 expression increases disease risk, and an inhibitor or antagonist drug would be beneficial. Reported *P*-  
497 values were calculated by Fisher exact test.

498 ∴ *P*<0.1; \* : *P*<0.05; \*\* : *P*<0.01; \*\*\* : *P*<0.001

499



500

501 **Figure 3. *IL6ST* is predicted to be causal for rheumatoid arthritis and polymyalgia rheumatica. A)**

502 Diseases associations supported by MR, colocalization and ABC. thows tissues and cell-types with

503 significant MR (q-value < 0.05) using *IL6ST* eQTL as exposure and diseases as outcome (red: positive

504 effect size estimate [MR beta]; blue: negative effect size estimate). The size of the dots represents

505 absolute effect size. Disease-eQTL pairs with a colocalization posterior probability > 80% are highlighted

506 with a dark border. **B)** LocusZoom[83] plot showing the top association for polymyalgia rheumatica at

507 the *ANKRD55-IL6ST* locus. Both *IL6ST* and *ANKRD55* eQTL colocalize with the polymyalgia rheumatica

508 signal, but *IL6ST* has the highest ABC score.

509

510 **TABLES**

511 **Table 1. GWAS included in this study.** The table reports the maximum GWAS sample size for each study,  
 512 the total number of GWAS with at least one associated gene. The number of loci with at least one  
 513 variant with GWAS  $P < 1 \times 10^{-6}$ . To calculate the number of loci, we defined 250kb regions each side of the  
 514 lead variant. Overlapping regions were then merged. The table reports the total number of non-  
 515 overlapping regions. The mean number of prioritized genes corresponds to the average number of  
 516 genes prioritized across each GWAS. The mean number of prioritized gene per locus correspond to the  
 517 average number of genes with the highest scores in a locus. For the analyses reported throughout this  
 518 manuscript, ties are broken using the shortest distance to the lead variant. Finally, the last column  
 519 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 ( $P < 1 \times 10^{-6}$ )	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	22.86	1.18	0.22
FinnGen, UK biobank, Estonian biobank meta-analysis (R10)	1,073,998	95	123.44	183.59	1.38	0.45
UKBB pan ICD-10 (European)	420,531	898	9.01	10.83	1.14	0.17
UKBB pan	42,0531	1,321	10.52	13.11	1.15	0.19

phecodes (European)						
------------------------	--	--	--	--	--	--

520 molQTL: molecular QTL; N: Number

521 **DECLARATIONS**

522 Ethics approval and consent to participate

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

530 The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers:  
531 THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018,  
532 THL/283/6.02.00/2019, THL/1721/5.05.00/2019, THL/1524/5.05.00/2020, and THL/2364/14.02/2020),  
533 Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3,  
534 VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA  
535 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA  
536 16/522/2020 and Statistics Finland (permit numbers: TK-53-1041-17 and TK-53-90-20).

537 The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 6 include:  
538 THL Biobank BB2017\_55, BB2017\_111, BB2018\_19, BB\_2018\_34, BB\_2018\_67, BB2018\_71, BB2019\_7,  
539 BB2019\_8, BB2019\_26, BB2020\_1, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank  
540 HUS/359/2017, Auria Biobank AB17-5154, Biobank Borealis of Northern Finland\_2017\_1013, Biobank of  
541 Eastern Finland 1186/2018, Finnish Clinical Biobank Tampere MH0004, Central Finland Biobank 1-2017,  
542 and Terveystalo Biobank STB 2018001.

543 UK Biobank has received ethical approval from the NHS National Research Ethics Service North West  
544 (approval numbers 11/NW/0382 and 16/NW/0274). All participants provided written informed consent.

545 Estonian Biobank GWAS and consecutive meta-analyses were carried out under ethical approval permit  
546 number 1.1-12/1020 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry  
547 of Social Affairs).

548

549 Consent for publication

550 Not applicable.

551

552 Availability of data and materials

553 The UK Biobank Pan ancestry GWAS[45] are available through <https://pan.ukbb.broadinstitute.org/>.

554 FinnGen GWAS[7] are available through [https://www.finngen.fi/en/access\\_results](https://www.finngen.fi/en/access_results). Processed and

555 formatted eQTL data used in this study are available through the eQTL catalogue [52-75]

556 <https://www.ebi.ac.uk/eqtl/>. pQTL from Sun et al. 2018[49] are available through

557 <http://www.phpc.cam.ac.uk/ceu/proteins/>. eQTLGen eQTL[51] are available through

558 <https://www.eqtlgen.org/phase1.html>. 1000 Genomes project phase 3 data[77] is available through

559 <https://www.internationalgenome.org/category/phase-3/>. Activity-by-contact maps[41] are available

560 through <https://www.engreitzlab.org/resources/>. LoGoFunc[47] gain and loss of function predictions are

561 available through <https://itanlab.shinyapps.io/goflof/>. Datasets supporting the conclusions of this article

562 are included within the article and its additional files. Additional datasets used and/or analysed during

563 the current study are available from the corresponding author on reasonable request.

564

565 Competing interests

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

569

570 Funding

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

573 The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH  
574 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc.,  
575 Celgene Corporation, Celgene International II Sàrl, Genentech Inc., Merck Sharp & Dohme Corp, Pfizer  
576 Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics  
577 Inc., Janssen Biotech Inc, and Novartis AG.

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

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



587

588 Authors' contributions

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

595

596 Acknowledgments

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

600 *UK Biobank*

601 This research has been conducted using the UK Biobank Resource ([www.ukbiobank.ac.uk](http://www.ukbiobank.ac.uk)), a large-scale  
602 biomedical database and research resource containing genetic, lifestyle and health information from  
603 500,000 UK participants. UK Biobank is supported by its founding funders the Wellcome Trust and UK  
604 Medical Research Council, as well as the Department of Health, Scottish Government, the Northwest  
605 Regional Development Agency, British Heart Foundation and Cancer Research UK. The UK biobank pan-  
606 ancestry analysis was conducted under project ID 31063 (<https://pan.ukbb.broadinstitute.org>).

607 *FinnGen*

608 The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH  
609 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc.,  
610 Celgene Corporation, Celgene International II Sàrl, Genentech Inc., Merck Sharp & Dohme Corp, Pfizer  
611 Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics  
612 Inc., Janssen Biotech Inc, and Novartis AG. Following biobanks are acknowledged for delivering biobank  
613 samples to FinnGen: Auria Biobank ([www.auria.fi/biopankki](http://www.auria.fi/biopankki)), THL Biobank ([www.thl.fi/biobank](http://www.thl.fi/biobank)), Helsinki  
614 Biobank ([www.helsinginbiopankki.fi](http://www.helsinginbiopankki.fi)), Biobank Borealis of Northern Finland  
615 (<https://www.ppsHP.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx>),  
616 Finnish Clinical Biobank Tampere ([www.tays.fi/en-](http://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere)  
617 [US/Research\\_and\\_development/Finnish\\_Clinical\\_Biobank\\_Tampere](http://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere)), Biobank of Eastern Finland  
618 ([www.ita-suomenbiopankki.fi/en](http://www.ita-suomenbiopankki.fi/en)), Central Finland Biobank ([www.ksshP.fi/fi-FI/Potilaalle/Biopankki](http://www.ksshP.fi/fi-FI/Potilaalle/Biopankki)),  
619 Finnish Red Cross Blood Service Biobank ([www.veripalvelu.fi/verenluovutus/biopankkitoiminta](http://www.veripalvelu.fi/verenluovutus/biopankkitoiminta)) and  
620 Terveystalo Biobank ([www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/](http://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/)). All  
621 Finnish Biobanks are members of BBMRI.fi infrastructure ([www.bbMRI.fi](http://www.bbMRI.fi)). Finnish Biobank Cooperative -  
622 FINBB (<https://finbb.fi/>) is the coordinator of BBMRI-ERIC operations in Finland.

### 623 *Estonian Biobank*

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

631 *molQTL datasets*

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

636 *Estonian Biobank and FinnGen banners*

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

640

#### 641 **ADDITIONAL INFORMATION**

642 **Additional file 1 (docx): Supplementary figures S1-S8.** Figure S1: Gold standard gene enrichment by  
643 genomic features. Figure S2: Precision and recall of gold standard genes for different genomic features  
644 as well as causal candidate prioritization approach. Figure S3: F1 scores for each considered features and  
645 prioritization scheme. Figure S4: Enriched colocalizing cell types and tissues by disease categories. Figure  
646 S5: Predicted direction of effect of gene expression on disease risk. Figure S6: Concordance between the  
647 predicted effect of gene expression on disease risk by mendelian randomization (MR) and mechanism of  
648 action (MoA) of approved drugs. Figure S7: Association between *IL6* and diseases, supported by MR,  
649 colocalization and ABC. Figure S8: Association between *IL6R* and diseases, supported by MR,  
650 colocalization and ABC.

651 **Additional file 2 (xlsx): Supplementary tables S1-S10.** Table S1: Enrichment of gold standard genes by  
652 feature and GWAS study source. Table S2: Precision and recall of different features to recover gold

653 standard genes. Table S3: Genes with predicted gain or loss of function variants ( $P < 1e-6$ ). Table S4:  
654 Genes with overrepresented disease categories of GWAS in which they are prioritized as causal. Table  
655 S5: Genes with overrepresented cell-type colocalizing QTL with GWAS in which they are prioritized as  
656 causal. Table S6: Significantly enriched colocalizing QTL cell types and tissues in disease GWAS  
657 categories, after grouping similar tissues and cell-types together. Table S7: Significantly enriched  
658 colocalizing QTL cell types and tissues in disease GWAS categories, treating each eQTL dataset  
659 separately. Table S8: Enrichment of prioritized genes by feature across clinical trial phases and approved  
660 drugs. Table S9: Enrichment of prioritized genes by rank across clinical trial phases and approved drugs.  
661 Table S10: Causal association between diseases and IL6, IL6ST, or IL6R.

662 **Additional file 3: FinnGen banner authors and affiliations.** List of FinnGen authors and their affiliations.

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

## 665 REFERENCES

- 666 1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J: **10 Years of GWAS**  
667 **Discovery: Biology, Function, and Translation.** *Am J Hum Genet* 2017, **101**:5-22.
- 668 2. Loos RJF: **15 years of genome-wide association studies and no signs of slowing down.** *Nat*  
669 *Commun* 2020, **11**:5900.
- 670 3. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales  
671 J, Mountjoy E, Sollis E, et al: **The NHGRI-EBI GWAS Catalog of published genome-wide**  
672 **association studies, targeted arrays and summary statistics 2019.** *Nucleic Acids Res* 2019,  
673 **47**:D1005-D1012.
- 674 4. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O,  
675 O'Connell J, et al: **The UK Biobank resource with deep phenotyping and genomic data.** *Nature*  
676 2018, **562**:203-209.
- 677 5. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray  
678 M, et al: **UK biobank: an open access resource for identifying the causes of a wide range of**  
679 **complex diseases of middle and old age.** *PLoS Med* 2015, **12**:e1001779.
- 680 6. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Magi R, Milani  
681 L, et al: **Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu.**  
682 *Int J Epidemiol* 2015, **44**:1137-1147.

- 683 7. Kurki MI, Karjalainen J, Palta P, Sipila TP, Kristiansson K, Donner KM, Reeve MP, Laivuori H,  
684 Aavikko M, Kaunisto MA, et al: **FinnGen provides genetic insights from a well-phenotyped**  
685 **isolated population.** *Nature* 2023, **613**:508-518.
- 686 8. Laisk T, Lepamets M, Koel M, Abner E, Estonian Biobank Research T, Magi R: **Genome-wide**  
687 **association study identifies five risk loci for pernicious anemia.** *Nat Commun* 2021, **12**:3761.
- 688 9. Tyrmi JS, Arffman RK, Pujol-Gualdo N, Kurra V, Morin-Papunen L, Sliz E, FinnGen Consortium  
689 EBRT, Piltonen TT, Laisk T, Kettunen J, Laivuori H: **Leveraging Northern European population**  
690 **history: novel low-frequency variants for polycystic ovary syndrome.** *Hum Reprod* 2022,  
691 **37**:352-365.
- 692 10. Alver M, Palover M, Saar A, Lall K, Zekavat SM, Tonisson N, Leitsalu L, Reigo A, Nikopensius T,  
693 Ainla T, et al: **Recall by genotype and cascade screening for familial hypercholesterolemia in a**  
694 **population-based biobank from Estonia.** *Genet Med* 2019, **21**:1173-1180.
- 695 11. King EA, Davis JW, Degner JF: **Are drug targets with genetic support twice as likely to be**  
696 **approved? Revised estimates of the impact of genetic support for drug mechanisms on the**  
697 **probability of drug approval.** *PLoS Genet* 2019, **15**:e1008489.
- 698 12. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J,  
699 et al: **The support of human genetic evidence for approved drug indications.** *Nat Genet* 2015,  
700 **47**:856-860.
- 701 13. Reay WR, Cairns MJ: **Advancing the use of genome-wide association studies for drug**  
702 **repurposing.** *Nat Rev Genet* 2021, **22**:658-671.
- 703 14. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah  
704 T, et al: **Association analyses identify 38 susceptibility loci for inflammatory bowel disease and**  
705 **highlight shared genetic risk across populations.** *Nat Genet* 2015, **47**:979-986.
- 706 15. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C,  
707 Regueiro M, Griffiths A, et al: **A genome-wide association study identifies IL23R as an**  
708 **inflammatory bowel disease gene.** *Science* 2006, **314**:1461-1463.
- 709 16. Pidasheva S, Trifari S, Phillips A, Hackney JA, Ma Y, Smith A, Sohn SJ, Spits H, Little RD, Behrens  
710 TW, et al: **Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor**  
711 **function in the protective genetic variant R381Q.** *PLoS One* 2011, **6**:e25038.
- 712 17. Peyrin-Biroulet L, Ghosh S, Lee SD, Lee WJ, Griffith J, Wallace K, Berg S, Liao X, Panes J, Loftus EV,  
713 Jr., Louis E: **Effect of risankizumab on health-related quality of life in patients with Crohn's**  
714 **disease: results from phase 3 MOTIVATE, ADVANCE and FORTIFY clinical trials.** *Aliment*  
715 *Pharmacol Ther* 2023, **57**:496-508.
- 716 18. Ferrante M, Panaccione R, Baert F, Bossuyt P, Colombel JF, Danese S, Dubinsky M, Feagan BG,  
717 Hisamatsu T, Lim A, et al: **Risankizumab as maintenance therapy for moderately to severely**  
718 **active Crohn's disease: results from the multicentre, randomised, double-blind, placebo-**  
719 **controlled, withdrawal phase 3 FORTIFY maintenance trial.** *Lancet* 2022, **399**:2031-2046.
- 720 19. Feagan BG, Sandborn WJ, Gasink C, Jacobstein D, Lang Y, Friedman JR, Blank MA, Johanns J, Gao  
721 LL, Miao Y, et al: **Ustekinumab as Induction and Maintenance Therapy for Crohn's Disease.** *N*  
722 *Engl J Med* 2016, **375**:1946-1960.
- 723 20. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, Zhernakova A, Stahl E, Viatte S, McAllister K,  
724 et al: **High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis.**  
725 *Nat Genet* 2012, **44**:1336-1340.
- 726 21. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP,  
727 Ripatti S, Chasman DI, Willer CJ, et al: **Biological, clinical and population relevance of 95 loci for**  
728 **blood lipids.** *Nature* 2010, **466**:707-713.
- 729 22. Ober C, Yao TC: **The genetics of asthma and allergic disease: a 21st century perspective.**  
730 *Immunol Rev* 2011, **242**:10-30.

- 731 23. Valette K, Li Z, Bon-Baret V, Chignon A, Berube JC, Eslami A, Lamothe J, Gaudreault N, Joubert P,  
732 Obeidat M, et al: **Prioritization of candidate causal genes for asthma in susceptibility loci**  
733 **derived from UK Biobank.** *Commun Biol* 2021, **4**:700.
- 734 24. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R,  
735 Qu H, Brody J, et al: **Systematic localization of common disease-associated variation in**  
736 **regulatory DNA.** *Science* 2012, **337**:1190-1195.
- 737 25. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry  
738 JL, Puvion-Vandier V, et al: **FTO Obesity Variant Circuitry and Adipocyte Browning in Humans.** *N*  
739 *Engl J Med* 2015, **373**:895-907.
- 740 26. Mountjoy E, Schmidt EM, Carmona M, Schwartzenuber J, Peat G, Miranda A, Fumis L, Hayhurst  
741 J, Buniello A, Karim MA, et al: **An open approach to systematically prioritize causal variants and**  
742 **genes at all published human GWAS trait-associated loci.** *Nat Genet* 2021, **53**:1527-1533.
- 743 27. Smith GD, Ebrahim S: **'Mendelian randomization': can genetic epidemiology contribute to**  
744 **understanding environmental determinants of disease?** *Int J Epidemiol* 2003, **32**:1-22.
- 745 28. Davey Smith G, Hemani G: **Mendelian randomization: genetic anchors for causal inference in**  
746 **epidemiological studies.** *Hum Mol Genet* 2014, **23**:R89-98.
- 747 29. Richardson TG, Hemani G, Gaunt TR, Relton CL, Davey Smith G: **A transcriptome-wide**  
748 **Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across**  
749 **the human phenome.** *Nat Commun* 2020, **11**:185.
- 750 30. Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S,  
751 Pasaniuc B, Eskin E: **Colocalization of GWAS and eQTL Signals Detects Target Genes.** *Am J Hum*  
752 *Genet* 2016, **99**:1245-1260.
- 753 31. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V:  
754 **Bayesian test for colocalisation between pairs of genetic association studies using summary**  
755 **statistics.** *PLoS Genet* 2014, **10**:e1004383.
- 756 32. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods  
757 groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida,  
758 et al: **Genetic effects on gene expression across human tissues.** *Nature* 2017, **550**:204-213.
- 759 33. Ndungu A, Payne A, Torres JM, van de Bunt M, McCarthy MI: **A Multi-tissue Transcriptome**  
760 **Analysis of Human Metabolites Guides Interpretability of Associations Based on Multi-SNP**  
761 **Models for Gene Expression.** *Am J Hum Genet* 2020, **106**:188-201.
- 762 34. Liu X, Finucane HK, Gusev A, Bhatia G, Gazal S, O'Connor L, Bulik-Sullivan B, Wright FA, Sullivan  
763 PF, Neale BM, Price AL: **Functional Architectures of Local and Distal Regulation of Gene**  
764 **Expression in Multiple Human Tissues.** *Am J Hum Genet* 2017, **100**:605-616.
- 765 35. Consortium GT: **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis:**  
766 **multitissue gene regulation in humans.** *Science* 2015, **348**:648-660.
- 767 36. Connally NJ, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, Chun S, Cotsapas C, Cassa CA,  
768 Sunyaev SR: **The missing link between genetic association and regulatory function.** *Elife* 2022,  
769 **11**.
- 770 37. Mostafavi H, Spence JP, Naqvi S, Pritchard JK: **Limited overlap of eQTLs and GWAS hits due to**  
771 **systematic differences in discovery.** *bioRxiv* 2022.
- 772 38. Lessard S, Gatof ES, Beaudoin M, Schupp PG, Sher F, Ali A, Prehar S, Kurita R, Nakamura Y, Baena  
773 E, et al: **An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria**  
774 **susceptibility.** *J Clin Invest* 2017, **127**:3065-3074.
- 775 39. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, Aneas I, Credidio FL,  
776 Sobreira DR, Wasserman NF, et al: **Obesity-associated variants within FTO form long-range**  
777 **functional connections with IRX3.** *Nature* 2014, **507**:371-375.

- 778 40. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J,  
779 Langdon R, et al: **The MR-Base platform supports systematic causal inference across the**  
780 **human phenome.** *Elife* 2018, **7**.
- 781 41. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, Jones TR, Nguyen  
782 TH, Ulirsch JC, Lekschas F, et al: **Genome-wide enhancer maps link risk variants to disease**  
783 **genes.** *Nature* 2021, **593**:238-243.
- 784 42. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F: **The**  
785 **Ensembl Variant Effect Predictor.** *Genome Biol* 2016, **17**:122.
- 786 43. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, Schoenherr S, Forer L, McCarthy S,  
787 Abecasis GR, et al: **Reference-based phasing using the Haplotype Reference Consortium panel.**  
788 *Nat Genet* 2016, **48**:1443-1448.
- 789 44. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference**  
790 **for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum*  
791 *Genet* 2007, **81**:1084-1097.
- 792 45. team P-U: <https://pan.ukbb.broadinstitute.org>. 2020.
- 793 46. Maintainer BP: **liftOver: Changing genomic coordinate systems with rtracklayer::liftOver.** *R*  
794 *package version 1180* 2021.
- 795 47. Stein D, Bayrak ÇS, Wu Y, Stenson PD, Cooper DN, Schlessinger A, Itan Y: **Genome-wide**  
796 **prediction of pathogenic gain- and loss-of-function variants from ensemble learning of diverse**  
797 **feature set.** *bioRxiv* 2022.
- 798 48. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeyasinghe S, Krawczak M,  
799 Cooper DN: **Human Gene Mutation Database (HGMD): 2003 update.** *Hum Mutat* 2003, **21**:577-  
800 581.
- 801 49. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E,  
802 Surendran P, et al: **Genomic atlas of the human plasma proteome.** *Nature* 2018, **558**:73-79.
- 803 50. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, Watt S, Yan Y, Kundu K, Ecker S, et  
804 al: **Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells.** *Cell*  
805 2016, **167**:1398-1414 e1324.
- 806 51. Vosa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber  
807 R, Yazar S, et al: **Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and**  
808 **polygenic scores that regulate blood gene expression.** *Nat Genet* 2021, **53**:1300-1310.
- 809 52. Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, Samovica M, Sakthivel MP,  
810 Kuzmin I, Trevanion SJ, et al: **A compendium of uniformly processed human gene expression**  
811 **and splicing quantitative trait loci.** *Nat Genet* 2021, **53**:1290-1299.
- 812 53. Buil A, Brown AA, Lappalainen T, Vinuela A, Davies MN, Zheng HF, Richards JB, Glass D, Small KS,  
813 Durbin R, et al: **Gene-gene and gene-environment interactions detected by transcriptome**  
814 **sequence analysis in twins.** *Nat Genet* 2015, **47**:88-91.
- 815 54. Jaffe AE, Straub RE, Shin JH, Tao R, Gao Y, Collado-Torres L, Kam-Thong T, Xi HS, Quan J, Chen Q,  
816 et al: **Developmental and genetic regulation of the human cortex transcriptome illuminate**  
817 **schizophrenia pathogenesis.** *Nat Neurosci* 2018, **21**:1117-1125.
- 818 55. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, Ha B,  
819 Altay G, Greenbaum JA, McVicker G, et al: **Impact of Genetic Polymorphisms on Human**  
820 **Immune Cell Gene Expression.** *Cell* 2018, **175**:1701-1715 e1716.
- 821 56. Ng B, White CC, Klein HU, Sieberts SK, McCabe C, Patrick E, Xu J, Yu L, Gaiteri C, Bennett DA, et  
822 al: **An xQTL map integrates the genetic architecture of the human brain's transcriptome and**  
823 **epigenome.** *Nat Neurosci* 2017, **20**:1418-1426.

- 824 57. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger  
825 T, Romano L, Planchon A, et al: **Passive and active DNA methylation and the interplay with**  
826 **genetic variation in gene regulation.** *Elife* 2013, **2**:e00523.
- 827 58. van de Bunt M, Manning Fox JE, Dai X, Barrett A, Grey C, Li L, Bennett AJ, Johnson PR, Rajotte  
828 RV, Gaulton KJ, et al: **Transcript Expression Data from Human Islets Links Regulatory Signals**  
829 **from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their**  
830 **Downstream Effectors.** *PLoS Genet* 2015, **11**:e1005694.
- 831 59. Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, Consortium H, Hale C,  
832 Dougan G, Gaffney DJ: **Shared genetic effects on chromatin and gene expression indicate a role**  
833 **for enhancer priming in immune response.** *Nat Genet* 2018, **50**:424-431.
- 834 60. Taylor DL, Jackson AU, Narisu N, Hemani G, Erdos MR, Chines PS, Swift A, Idol J, Didion JP, Welch  
835 RP, et al: **Integrative analysis of gene expression, DNA methylation, physiological traits, and**  
836 **genetic variation in human skeletal muscle.** *Proc Natl Acad Sci U S A* 2019, **116**:10883-10888.
- 837 61. Lepik K, Annilo T, Kukuskina V, e QC, Kisand K, Kutalik Z, Peterson P, Peterson H: **C-reactive**  
838 **protein upregulates the whole blood expression of CD59 - an integrative analysis.** *PLoS*  
839 *Comput Biol* 2017, **13**:e1005766.
- 840 62. Quach H, Rotival M, Pothlichet J, Loh YE, Dannemann M, Zidane N, Laval G, Patin E, Harmant C,  
841 Lopez M, et al: **Genetic Adaptation and Neandertal Admixture Shaped the Immune System of**  
842 **Human Populations.** *Cell* 2016, **167**:643-656 e617.
- 843 63. Nedelec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier JC, Freiman A, Sams AJ,  
844 Hebert S, et al: **Genetic Ancestry and Natural Selection Drive Population Differences in**  
845 **Immune Responses to Pathogens.** *Cell* 2016, **167**:657-669 e621.
- 846 64. Panopoulos AD, D'Antonio M, Benaglio P, Williams R, Hashem SI, Schuldt BM, DeBoever C, Arias  
847 AD, Garcia M, Nelson BC, et al: **iPSCORE: A Resource of 222 iPSC Lines Enabling Functional**  
848 **Characterization of Genetic Variation across a Variety of Cell Types.** *Stem Cell Reports* 2017,  
849 **8**:1086-1100.
- 850 65. Pashos EE, Park Y, Wang X, Raghavan A, Yang W, Abbey D, Peters DT, Arbelaez J, Hernandez M,  
851 Kuperwasser N, et al: **Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like**  
852 **Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci.** *Cell Stem Cell* 2017,  
853 **20**:558-570 e510.
- 854 66. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, Bala S, Bensaddek D, Casale FP,  
855 Culley OJ, et al: **Common genetic variation drives molecular heterogeneity in human iPSCs.**  
856 *Nature* 2017, **546**:370-375.
- 857 67. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta  
858 M, Kurbatova N, Griebel T, Ferreira PG, et al: **Transcriptome and genome sequencing uncovers**  
859 **functional variation in humans.** *Nature* 2013, **501**:506-511.
- 860 68. Hoffman GE, Bendl J, Voloudakis G, Montgomery KS, Sloofman L, Wang YC, Shah HR, Hauberg  
861 ME, Johnson JS, Girdhar K, et al: **CommonMind Consortium provides transcriptomic and**  
862 **epigenomic data for Schizophrenia and Bipolar Disorder.** *Sci Data* 2019, **6**:180.
- 863 69. Guelfi S, D'Sa K, Botia JA, Vandrovцова J, Reynolds RH, Zhang D, Trabzuni D, Collado-Torres L,  
864 Thomason A, Quijada Leyton P, et al: **Regulatory sites for splicing in human basal ganglia are**  
865 **enriched for disease-relevant information.** *Nat Commun* 2020, **11**:1041.
- 866 70. Young AMH, Kumasaka N, Calvert F, Hammond TR, Knights A, Panousis N, Park JS,  
867 Schwartzenruber J, Liu J, Kundu K, et al: **A map of transcriptional heterogeneity and regulatory**  
868 **variation in human microglia.** *Nat Genet* 2021, **53**:861-868.
- 869 71. Consortium GT: **The GTEx Consortium atlas of genetic regulatory effects across human tissues.**  
870 *Science* 2020, **369**:1318-1330.



- 871 72. Theusch E, Chen YI, Rotter JI, Krauss RM, Medina MW: **Genetic variants modulate gene**  
872 **expression statin response in human lymphoblastoid cell lines.** *BMC Genomics* 2020, **21**:555.
- 873 73. Peng S, Deysenroth MA, Di Narzo AF, Cheng H, Zhang Z, Lambertini L, Ruusalepp A, Kovacic JC,  
874 Bjorkegren JLM, Marsit CJ, et al: **Genetic regulation of the placental transcriptome underlies**  
875 **birth weight and risk of childhood obesity.** *PLoS Genet* 2018, **14**:e1007799.
- 876 74. Steinberg J, Southam L, Roumeliotis TI, Clark MJ, Jayasuriya RL, Swift D, Shah KM, Butterfield NC,  
877 Brooks RA, McCaskie AW, et al: **A molecular quantitative trait locus map for osteoarthritis.** *Nat*  
878 *Commun* 2021, **12**:1309.
- 879 75. Schwartzenruber J, Foskolou S, Kilpinen H, Rodrigues J, Alasoo K, Knights AJ, Patel M, Goncalves  
880 A, Ferreira R, Benn CL, et al: **Molecular and functional variation in iPSC-derived sensory**  
881 **neurons.** *Nat Genet* 2018, **50**:54-61.
- 882 76. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to**  
883 **the challenge of larger and richer datasets.** *Gigascience* 2015, **4**:7.
- 884 77. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini  
885 JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.**  
886 *Nature* 2015, **526**:68-74.
- 887 78. Ghossaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, Fumis L, Miranda A,  
888 Carvalho-Silva D, Buniello A, et al: **Open Targets Genetics: systematic identification of trait-**  
889 **associated genes using large-scale genetics and functional genomics.** *Nucleic Acids Res* 2021,  
890 **49**:D1311-D1320.
- 891 79. King EA, Dunbar F, Davis JW, Degner JF: **Estimating colocalization probability from limited**  
892 **summary statistics.** *BMC Bioinformatics* 2021, **22**:254.
- 893 80. Lin D: **An Information-Theoretic Definition of Similarity.** In *Proceedings of the Fifteenth*  
894 *International Conference on Machine Learning.* pp. 296–304: Morgan Kaufmann Publishers Inc.;  
895 1998:296–304.
- 896 81. Resnik P: **Semantic similarity in a taxonomy: an information-based measure and its application**  
897 **to problems of ambiguity in natural language.** 1999, **11**:95–130.
- 898 82. Greene D, Richardson S, Turro E: **ontologyX: a suite of R packages for working with ontological**  
899 **data.** *Bioinformatics* 2017, **33**:1104-1106.
- 900 83. Boughton AP, Welch RP, Flickinger M, VandeHaar P, Taliun D, Abecasis GR, Boehnke M:  
901 **LocusZoom.js: interactive and embeddable visualization of genetic association study results.**  
902 *Bioinformatics* 2021, **37**:3017-3018.
- 903 84. Wang G, Sarkar A, Carbonetto P, Stephens M: **A simple new approach to variable selection in**  
904 **regression, with application to genetic fine mapping.** *J R Stat Soc Series B Stat Methodol* 2020,  
905 **82**:1273-1300.
- 906 85. Nicolas G, Wallon D, Charbonnier C, Quenez O, Rousseau S, Richard AC, Rovelet-Lecrux A,  
907 Coutant S, Le Guennec K, Bacq D, et al: **Screening of dementia genes by whole-exome**  
908 **sequencing in early-onset Alzheimer disease: input and lessons.** *Eur J Hum Genet* 2016, **24**:710-  
909 716.
- 910 86. Laurin N, Brown JP, Morissette J, Raymond V: **Recurrent mutation of the gene encoding**  
911 **sequestosome 1 (SQSTM1/p62) in Paget disease of bone.** *Am J Hum Genet* 2002, **70**:1582-1588.
- 912 87. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R, Jr.,  
913 Ellis MC, Fullan A, et al: **A novel MHC class I-like gene is mutated in patients with hereditary**  
914 **haemochromatosis.** *Nat Genet* 1996, **13**:399-408.
- 915 88. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, Bjornsson S, Stefansson H, Sulem P,  
916 Gudbjartsson D, Maloney J, et al: **A mutation in APP protects against Alzheimer's disease and**  
917 **age-related cognitive decline.** *Nature* 2012, **488**:96-99.

- 918 89. Kristjansson RP, Benonisdottir S, Davidsson OB, Oddsson A, Tragante V, Sigurdsson JK,  
919 Stefansdottir L, Jonsson S, Jensson BO, Arthur JG, et al: **A loss-of-function variant in ALOX15**  
920 **protects against nasal polyps and chronic rhinosinusitis.** *Nat Genet* 2019, **51**:267-276.
- 921 90. Morris AP, Le TH, Wu H, Akbarov A, van der Most PJ, Hemani G, Smith GD, Mahajan A, Gaulton  
922 KJ, Nadkarni GN, et al: **Trans-ethnic kidney function association study reveals putative causal**  
923 **genes and effects on kidney-specific disease aetiologies.** *Nat Commun* 2019, **10**:29.
- 924 91. Beaudoin M, Gupta RM, Won HH, Lo KS, Do R, Henderson CA, Lavoie-St-Amour C, Langlois S,  
925 Rivas D, Lehoux S, et al: **Myocardial Infarction-Associated SNP at 6p24 Interferes With MEF2**  
926 **Binding and Associates With PHACTR1 Expression Levels in Human Coronary Arteries.**  
927 *Arterioscler Thromb Vasc Biol* 2015, **35**:1472-1479.
- 928 92. Lopez-Isac E, Smith SL, Marion MC, Wood A, Sudman M, Yarwood A, Shi C, Gaddi VP, Martin P,  
929 Prahallad S, et al: **Combined genetic analysis of juvenile idiopathic arthritis clinical subtypes**  
930 **identifies novel risk loci, target genes and key regulatory mechanisms.** *Ann Rheum Dis* 2021,  
931 **80**:321-328.
- 932 93. Yang J, McGovern A, Martin P, Duffus K, Ge X, Zarrineh P, Morris AP, Adamson A, Fraser P,  
933 Rattray M, Eyre S: **Analysis of chromatin organization and gene expression in T cells identifies**  
934 **functional genes for rheumatoid arthritis.** *Nat Commun* 2020, **11**:4402.
- 935 94. Dasgupta B, Unizony S, Warrington KJ, Sloane Lazar J, Giannelou A, Nivens C, Akinlade B, Wong  
936 W, Lin Y, Buttgerit F, et al: **LB0006 SARILUMAB IN PATIENTS WITH RELAPSING POLYMYALGIA**  
937 **RHEUMATICA: A PHASE 3, MULTICENTER, RANDOMIZED, DOUBLE BLIND, PLACEBO**  
938 **CONTROLLED TRIAL (SAPHYR).** *Annals of the Rheumatic Diseases* 2022, **81**:210-211.
- 939 95. Kanai M, Elzur R, Zhou W, Global Biobank Meta-analysis I, Daly MJ, Finucane HK: **Meta-analysis**  
940 **fine-mapping is often miscalibrated at single-variant resolution.** *Cell Genom* 2022, **2**.
- 941 96. Hormozdiari F, Zhu A, Kichaev G, Ju CJ, Segre AV, Joo JWJ, Won H, Sankararaman S, Pasaniuc B,  
942 Shifman S, Eskin E: **Widespread Allelic Heterogeneity in Complex Traits.** *Am J Hum Genet* 2017,  
943 **100**:789-802.
- 944 97. Zhu A, Matoba N, Wilson EP, Tapia AL, Li Y, Ibrahim JG, Stein JL, Love MI: **MRLocus: Identifying**  
945 **causal genes mediating a trait through Bayesian estimation of allelic heterogeneity.** *PLoS*  
946 *Genet* 2021, **17**:e1009455.
- 947 98. Yin X, Bose D, Kwon A, Hanks SC, Jackson AU, Stringham HM, Welch R, Oravilahti A, Fernandes  
948 Silva L, FinnGen, et al: **Integrating transcriptomics, metabolomics, and GWAS helps reveal**  
949 **molecular mechanisms for metabolite levels and disease risk.** *Am J Hum Genet* 2022, **109**:1727-  
950 1741.