nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0385-z

In the format provided by the authors and unedited.

Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg[1], Nasa Sinnott-Armstrong[2], Nicholas Mancuso[3], Alvaro N. Barbeira[4], David A. Knowles[5,6], David Golan[2], Raili Ermel[7], Arno Ruusalepp[7,8], Thomas Quertermous[9], Ke Hao[10], Johan L. M. Björkegren[8,10,11,12]*, Hae Kyung Im[4]*, Bogdan Pasaniuc[3,13,14]*, Manuel A. Rivas[15]* and Anshul Kundaje[1,2]*

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Genetics, Stanford University, Stanford, CA, USA. [3]Department of Pathology & Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [4]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA. [5]New York Genome Center, New York, NY, USA. [6]Department of Computer Science, Columbia University, New York, NY, USA. [7]Department of Cardiac Surgery, Tartu University Hospital, Tartu, Estonia. [8]Clinical Gene Networks AB, Stockholm, Sweden. [9]Division of Cardiovascular Medicine, Stanford University, Stanford, CA, USA. [10]Department of Genetics & Genomic Sciences, Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [11]Department of Pathophysiology, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia. [12]Integrated Cardio Metabolic Centre, Department of Medicine, Karolinska Institutet, Karolinska Universitetssjukhuset, Huddinge, Sweden. [13]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [14]Department of Computational Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [15]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. *e-mail: johan.bjorkegren@mssm.edu; haky@uchicago.edu; pasaniuc@ucla.edu; mrivas@stanford.edu; akundaje@stanford.edu

Methods

TWAS with Fusion

TWAS were performed with the Fusion software using default settings and also including polygenic risk score as a possible model during cross-validation in addition to BLUP, LASSO, and ElasticNet. TWAS p values from Fusion were Bonferroni-corrected according to the number of genes tested in the TWAS when assessing statistical significance. Variants in the STARNET reference panel were filtered for quality control using PLINK[1] with the options "--maf 1e-10 --hwe 1e-6 midp --geno". STARNET expression was processed as described in the STARNET paper[2], including probabilistic estimation of expression residuals[3] (PEER) covariate correction. Because Fusion only supports training on PLINK version 1 hard-call genotype files and not genotype dosages, we trained expression models on only the variants both genotyped in STARNET and either genotyped or imputed in the GWAS, filtering out variants without matching strands between the GWAS and STARNET. Expression models were trained on all remaining variants within 500 kb of a gene's TSS, using Ensembl v87 TSS annotations for hg19[4]. LD and total and predicted expression correlations were calculated across individuals in STARNET.

TWAS with S-PrediXcan

To run S-PrediXcan, ElasticNet prediction models and LD reference were generated using the same PEER-corrected STARNET data from the previous section, filtered to match each GWAS. Variants within 1MB of the TSS or TES were used to predict the expression of genes annotated as either protein-coding, lincRNA or pseudogene in Ensembl v87.

Statistics

The two-tailed Wilcoxon signed-rank p-values in the discussion to Supplementary Table 4 were computed with the *scipy.stats.wilcoxon* method in Python from the data shown in Supplementary Table 4. For details of how variant- and gene-level TWAS p-values were calculated, please refer to the Fusion and PrediXcan/S-PrediXcan manuscripts.

Simulations

For the simulations, we sampled independent genomic regions as defined by LDetect[5]. We then annotated each region with overlapping gene transcription start sites using all available genes in RefSeq v65. To simulate GWAS (N = 300,000) and expression panel (N = 500) genotypes, we sampled standardized genotypes using the multivariate normal approximation with mean 0 and covariance defined by LD among the 489 individuals from 1000 Genomes with European ancestry.

Next, we simulated heritable gene expression for all genes at a region with 80% of genes having a single causal eQTL and the remaining 20% having 2 causal eQTLs. Causal eQTLs were preferentially sampled within 50 kb of transcription start sites to exhibit a 50x enrichment on average compared with non-overlapping SNPs. Effect sizes for causal eQTLs were drawn from a normal distribution such that genetic variation explained 20% of variance in total expression.

Finally, given expression at genes causal for the complex trait, we sampled gene-level effect sizes from a normal distribution so that 20% of the variance in trait is explained by gene expression. To simulate dichotomous case/control disease traits, we assumed an underlying liability model, where liability for cases exceeds a given threshold. We set the liability threshold such that population-level prevalence of cases is 1%. We randomly assigned one gene at the locus to be causal and looked at what percent of the time other genes at the locus had a larger TWAS z-score than this causal gene, as a function of the predicted expression correlation magnitude with the causal gene.

To compute risk scores for quantitative and dichotomous traits using predicted expression in simulations, we fit a penalized linear model across all predicted expression levels and measured goodness of fit. We used both L1 (i.e. LASSO) and L2 (i.e. Ridge) models for estimation. To set the penalty parameters we adopted heuristics from SNP-based methods. Specifically, for LASSO we set the penalty term to the heritability explained by predicted expression[6]. For Ridge we set the penalty term to the ratio between the residual variance and effect-size variance, which is equivalent to the ridge-regression BLUP[7]. We measured goodness-of-fit in-sample, which provides an upper bound on risk prediction performance. For simulated continuous traits, goodness-of-fit is measured by the adjusted R^2 estimate. For simulated dichotomous traits we measured area under the precision-recall curve (AUPRC). We generated predicted expression for N=50,000 individuals across 10 genomic regions.

We also investigated performance of penalized linear models for association between predicted expression and simulated trait. We fit a LASSO (and Ridge) model using predicted expression of all genes to estimate jointly gene-level effects. To obtain confidence intervals, we computed 100 bootstrap estimates of gene-level effects and computed the empirical lower 2.5%- and upper 97.5%-quantiles. In this setting, we determined a gene to be associated if its empirical 95%-confidence interval did not overlap 0. Here, we generated predicted expression at a single risk region using N=300,000 and N=500 for the reference expression panel.

Code availability

Code to replicate the post-TWAS analysis is available at https://github.com/Wainberg/TWAS_challenges_and_opportunities. The version of Fusion used for this analysis is available at https://github.com/gusevlab/fusion_twas/tree/9142723485b38610695cea4e7ebb508945ec006c.

Data availability

GWAS summary statistics are publicly available from the CARDIoGRAMplusC4D consortium and Global Lipids Genetics Consortium. STARNET genotypes are available from Johan LM Björkegren on reasonable request. STARNET expression data is available from dbGAP (accession phs001203.v1.p1).

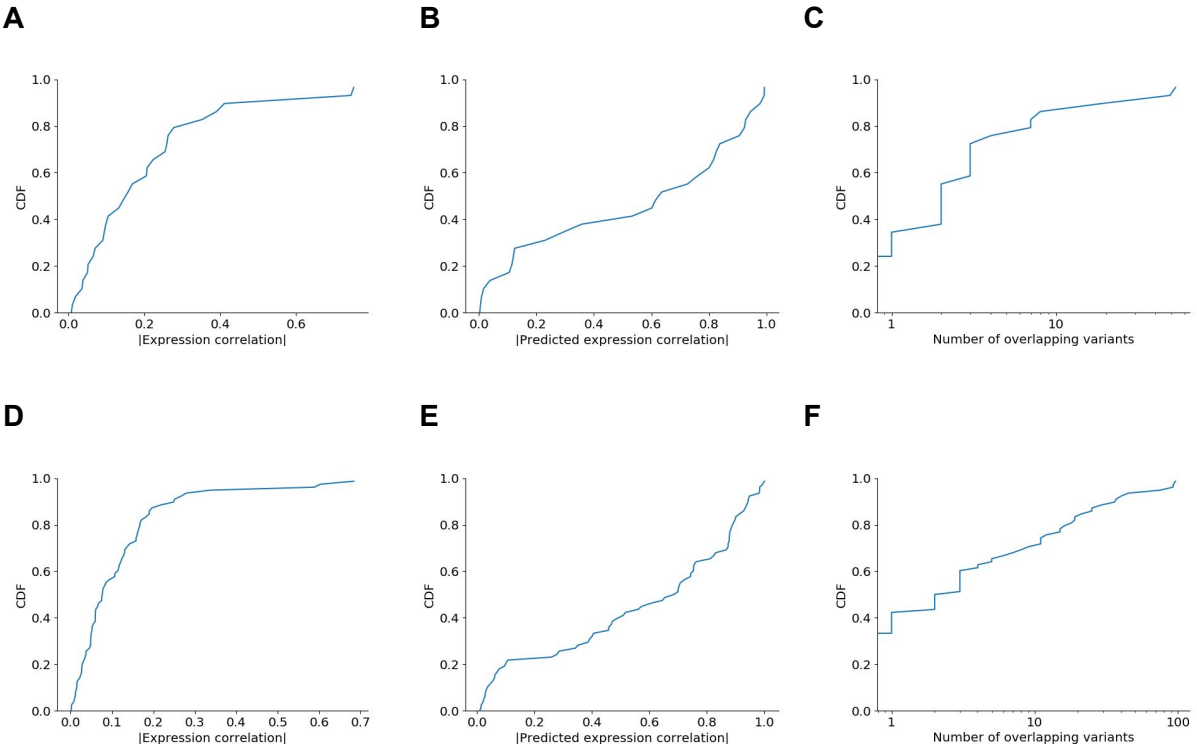
Life Sciences Reporting Summary

For more details on the study design and methods, please refer to the Life Sciences Reporting Summary, published alongside this paper.

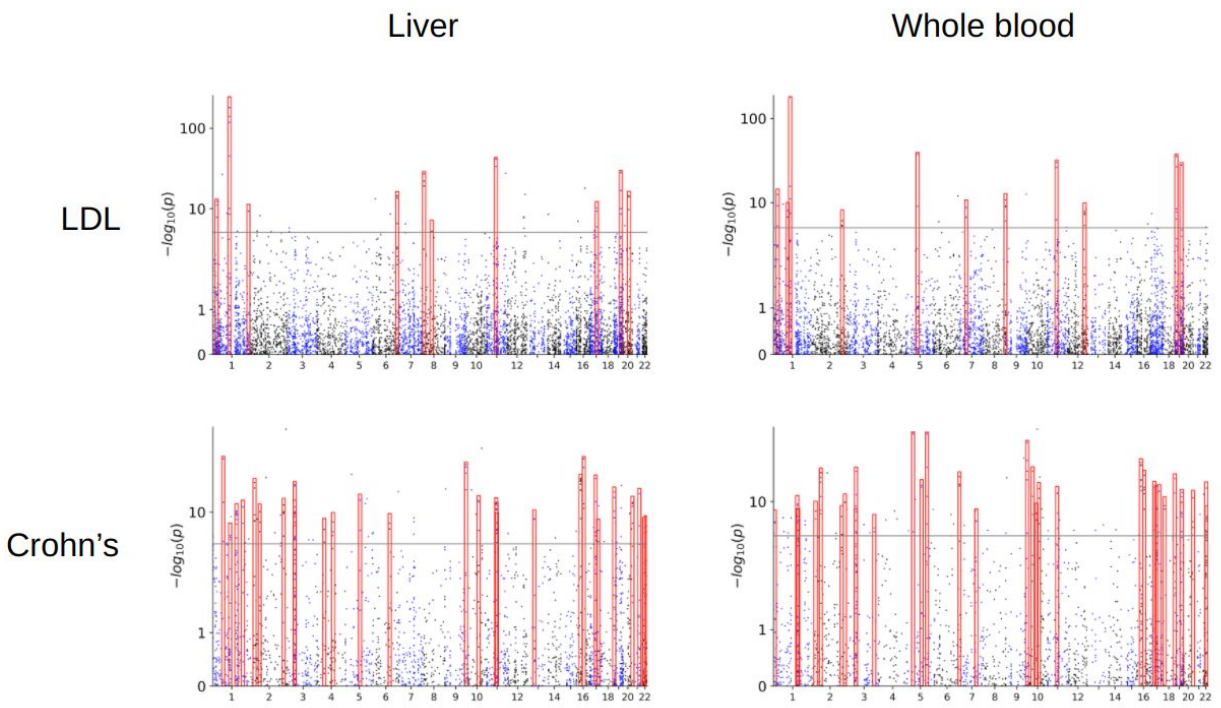
References

1. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
2. Franzén, O. *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* **353**, 827–830 (2016).
3. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
4. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database* (2016). doi:10.1093/database/baw093
5. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
6. Vattikuti, S., Lee, J. J., Chang, C. C., Hsu, S. D. H. & Chow, C. C. Applying compressed sensing to genome-wide association studies. *Gigascience* **3**, 10 (2014).
7. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* **4**, 250 (2011).

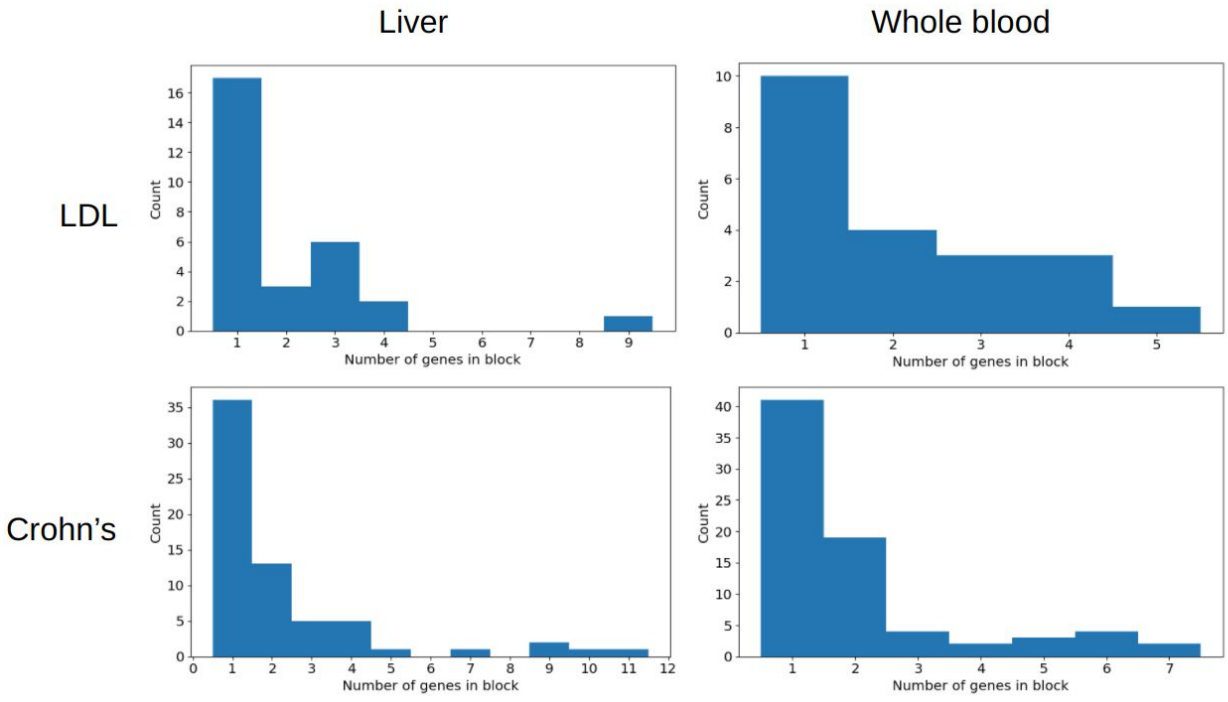
Supplementary Figures



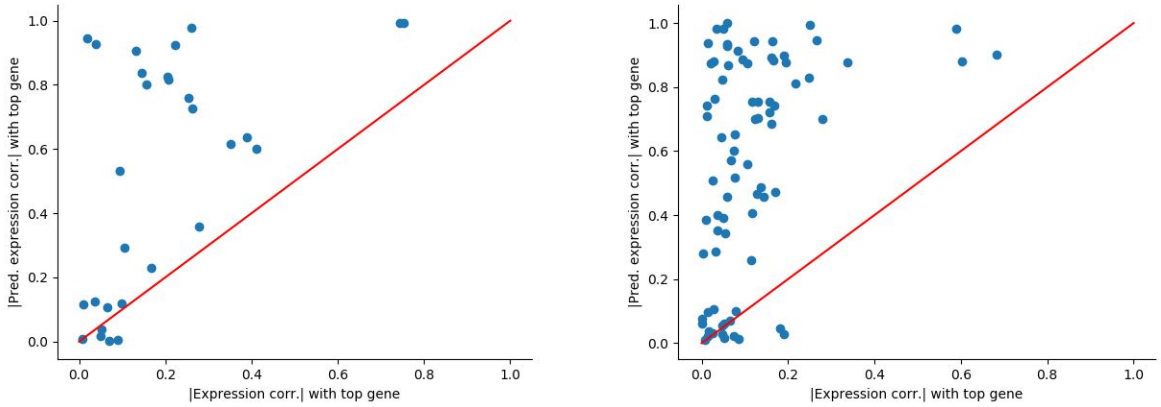
Supplementary Figure 1: Distributions of co-regulation across putative non-causal genes in multi-hit Fusion TWAS loci. Since many multi-hit loci do not have a clear causal gene or have multiple plausible candidates, we make the approximation that only the most significant gene at each locus is causal. We then plot the cumulative distribution functions (CDFs) of (a, d) expression correlations, (b, e) predicted expression correlations and (c, f) number of shared variants between these most significant genes and all the other genes at their loci, separately for LDL/liver (a-c) and Crohn's/whole blood (d-f). To collapse these CDFs into a single estimate of the percent of affected non-causal genes (Fig. 2c), we combine genes across the two studies and threshold to correlation $r^2 \geq 0.2$, a threshold commonly used for weak LD in GWAS, or ≥ 1 shared variant. Note that counting only exact sharing of variants does not account for LD, for simplicity.



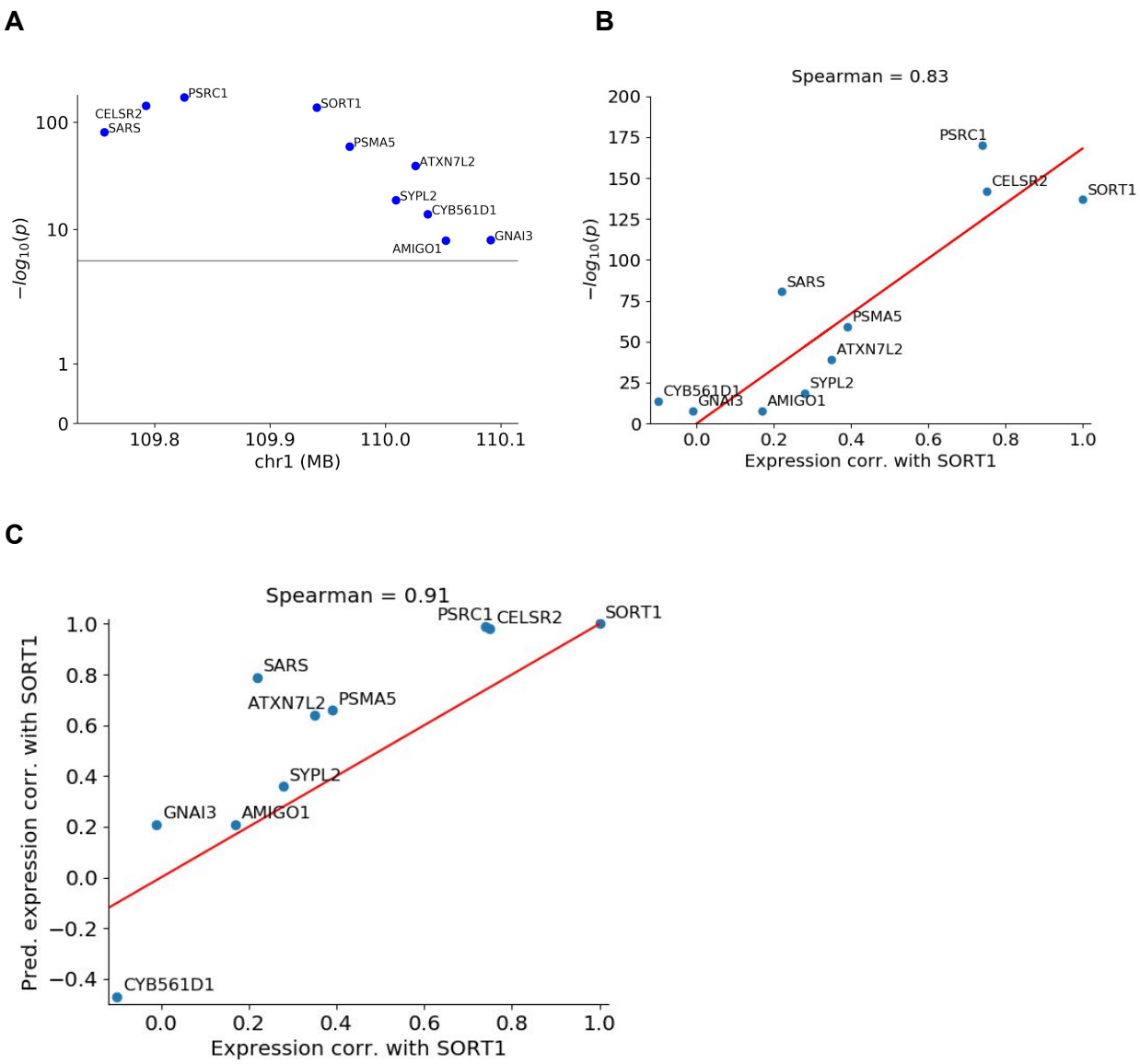
Supplementary Figure 2: Manhattan plots of the 4 Fusion TWAS conducted in this study. As in Fig. 2, clusters of multiple adjacent TWAS hit genes are highlighted in red.



Supplementary Figure 3: Number of Fusion TWAS hit genes per locus after 2.5-MB clumping.

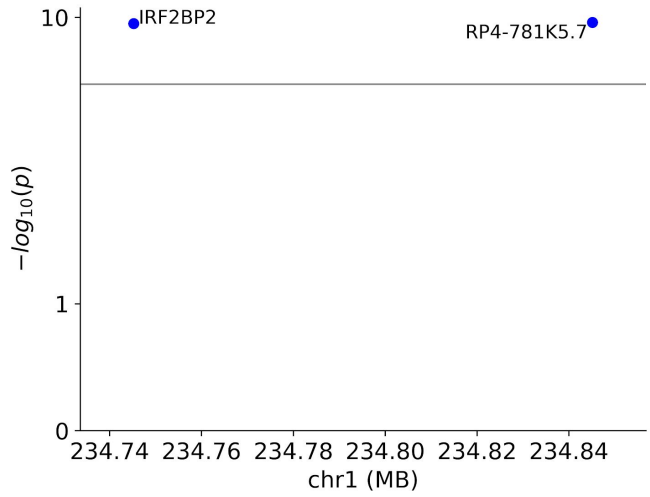


Supplementary Figure 4: Total versus predicted expression correlation versus the top hit, for all genes in Fusion TWAS multi-hit blocks that are not the top hits. a) Liver, LDL. b) Crohn's, whole blood. Note that predicted expression correlation is generally higher than total expression correlation, as discussed in the Results section.

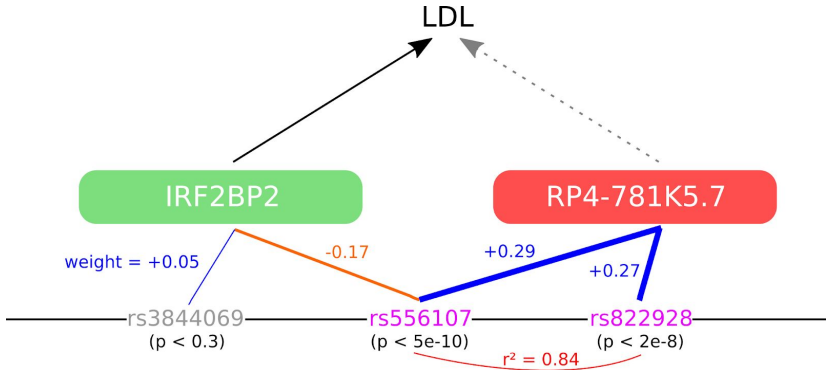


Supplementary Figure 5: The *SORT1* locus with S-Predixcan. a) S-PrediXcan Manhattan plot of the *SORT1* locus. b) Expression correlation with *SORT1* versus TWAS *p*-value, for each gene in the *SORT1* locus. c) For nearby genes, S-PrediXcan predicted expression correlations tend to be higher than total expression correlations, e.g. at the *SORT1* locus.

A

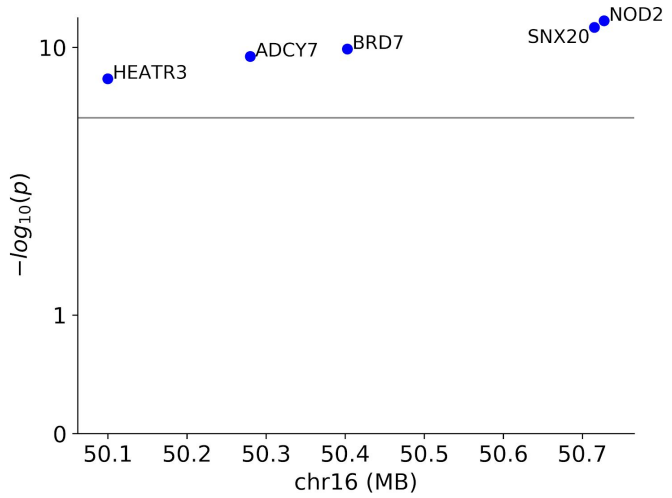


B

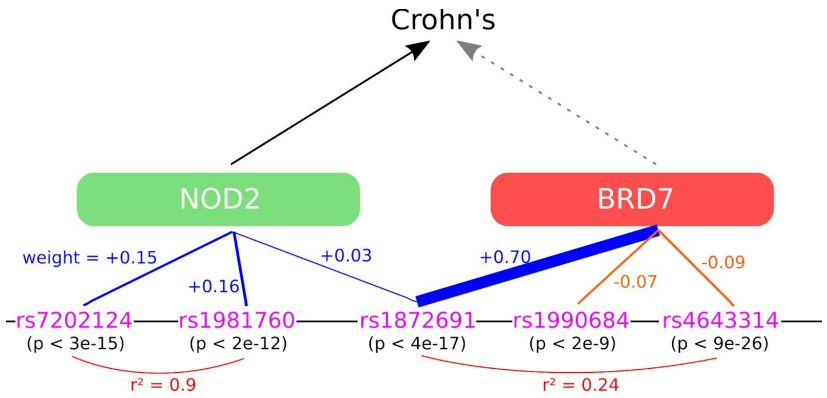


Supplementary Figure 6: The *IRF2BP2* locus with S-PrediXcan. a) S-PrediXcan Manhattan plot of the *IRF2BP2* locus, where *RP4-781K5.7* is a likely non-causal hit due to predicted expression correlation with *IRF2BP2*. b) Details of the two genes' S-PrediXcan expression models: a line between a variant's rs number and a gene indicates the variant is included in the gene's expression model with either a positive weight (blue) or negative weight (orange), with the thickness of the line increasing with the magnitude of the weight; red arcs indicate LD.

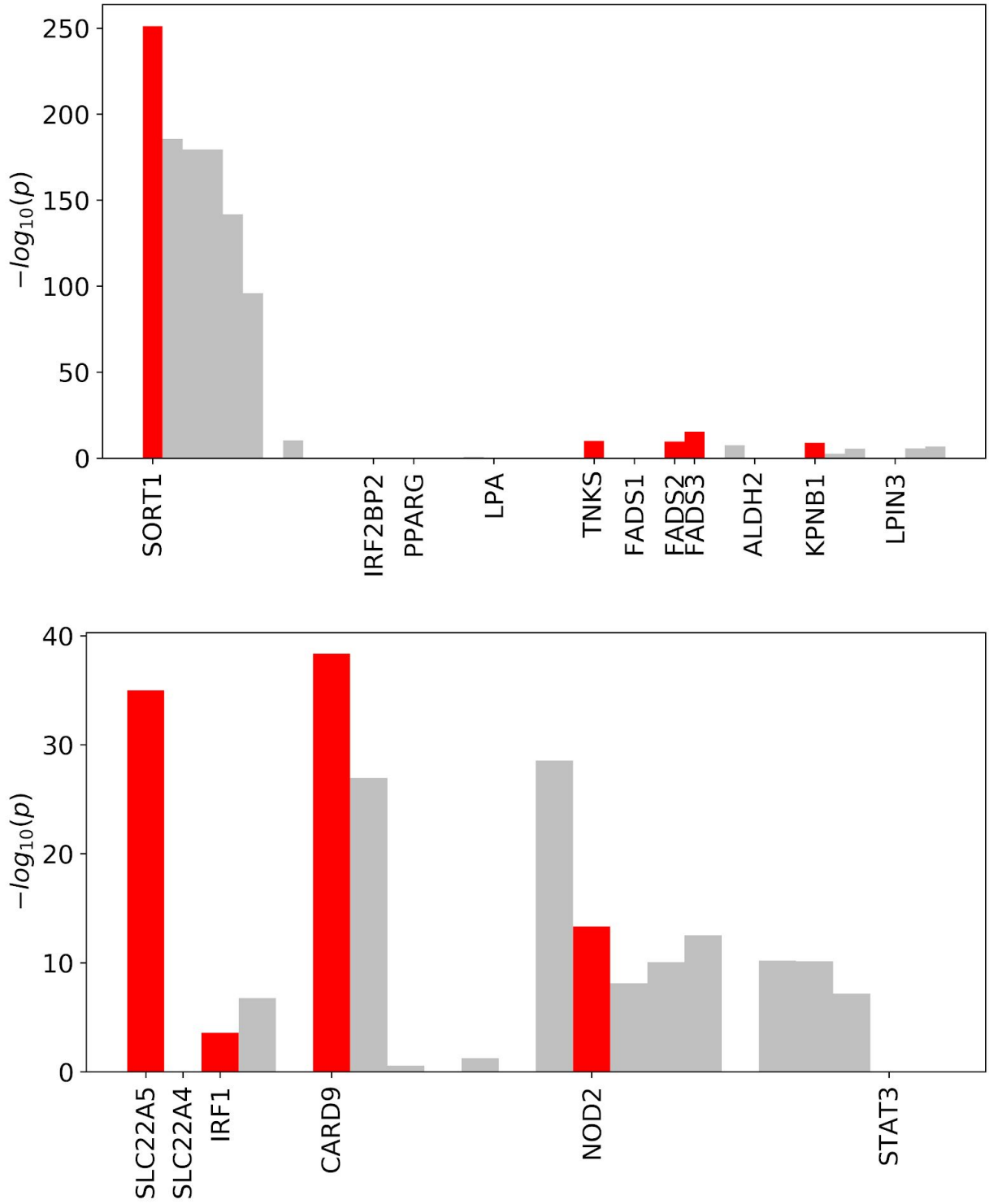
A



B



Supplementary Figure 7: The *NOD2* locus with S-Predixcan. a) S-PrediXcan Manhattan plot of the NOD2 locus. b) Details of the expression models of NOD2 and BRD7. For clarity, 5 variants for BRD7 (rs12925755, p = 6e-34, weight = 0.002; rs2066852, p = 3e-10, weight = -0.02; rs17227589, p = 2e-7, weight = -0.02; rs11642187, p = 0.04, weight = +0.007; rs2241258, p = 0.3, weight = -0.05) are not shown. A line between a variant's rs number and a gene indicates the variant is included in the gene's expression model with either a positive weight (blue) or negative weight (orange), with the thickness of the line increasing with the magnitude of the weight; red arcs indicate LD.



Supplementary Figure 8: Performance of combined whole-blood and liver reference panel on the loci in Figure 6.

Gene	Lowest TWAS p-value in any brain tissue	Lowest TWAS p-value in any tissue
C4A	4e-18 (Hypothalamus)	2e-20 (Pancreas)
ATF6B	3e-9 (Anterior cingulate cortex)	3e-9 (Anterior cingulate cortex)
CYP21A2	5e-7 (Cortex)	9e-19 (Aorta)
NELFE	7e-7 (Cerebellum)	7e-7 (Cerebellum)
STK19	1e-5 (Frontal Cortex, BA9)	4e-12 (Adrenal gland)
SKIV2L	5e-5 (Cerebellum)	5e-5 (Cerebellum)
C4B	6e-5 (Nucleus accumbens, basal ganglia)	1e-21 (Testis)
C2	0.008 (Cortex)	1e-18 (Whole blood)
DXO	0.03 (Putamen, basal ganglia)	0.02 (Thyroid)
CFB	Not significant	2e-13 (Whole blood)
EHMT2	Not significant	3e-10 (Skin, sun-exposed lower leg)
TNXB	Not significant	3e-6 (Adrenal gland)
ZBTB12	Not significant	7e-5 (Ovary)

Supplementary Table 1: The *C4A* locus, a success story where TWAS p-values accurately prioritize the causal gene. Lowest schizophrenia p-value in any GTEx brain tissue, and in any GTEx tissue, for each gene within 100 kb of *C4A* with available S-PrediXcan TWAS results (http://metabeta.gene2pheno.org). The TWAS used schizophrenia summary GWAS data from the Psychiatric Genomics Consortium[8] and expression data from GTEx[9].

	Predicted expression correlation magnitude with causal gene			
	0-0.05	0.05-0.1	0.1-0.15	0.15-0.2
Number of genes	3143	502	284	70
% genes with $\|z\| > \|z_{causal}\|$	9.4%	12.0%	31.7%	45.7%
Power (% of causal genes with $p < 0.05$)	76%			
False positive rate (% of non-causal genes with $p < 0.05$)	14.9%	50.8%	77.1%	71.4%

Supplementary Table 2: Simulation of percent of genes with larger TWAS z-score than the causal gene, binned by predicted expression correlation. The number of genes in each bin (among all genes at the 1000 random loci being simulated) is shown in brackets for each bin. Predicted expression correlations were computed as

the vector-matrix-vector product of the causal gene's model weights, the LD matrix among the variants included in the models, and the other gene's model weights.

Gene	Trait	Evidence	Details
SORT1	LDL	Strong	In mouse models, overexpression of *SORT1* in liver reduced plasma LDL levels and siRNA knockdown increased plasma LDL levels[10,11], though in other studies deletion of *SORT1* counter-intuitively reduced, rather than increased, atherosclerosis in mice without affecting plasma LDL levels[12,13,14].
IRF2BP2	LDL	Moderate	A loss-of-function variant in *IRF2BP2* has been associated with increased susceptibility to CAD[15] *IRF2BP2* knockout has been shown in mouse models to increase atherosclerosis, albeit via an inflammatory mechanism[15].
PPARG	LDL	Strong	*PPARG* activation increases LDL metabolism via induction of *LDLR* and *CYP7A1*[16]; PPAR agonists decrease glycated LDL uptake into macrophages via regulation of lipoprotein lipase[17].
LPA	LDL	Strong	*LPA* encodes is a primary constituent of lipoprotein(a), a class of lipoproteins related to LDL. The LDL GWAS used in this study is a meta-analysis of 60 studies, most of which do not measure LDL levels directly but instead calculate them indirectly using the Friedewald formula, which does not distinguish between LDL and lipoprotein(a) and instead reports the sum of LDL and lipoprotein(a) levels[18]. Thus, although *LPA* abundance may not causally influence true LDL levels, it does causally determine LDL levels as calculated by the Friedewald formula.
TNKS	LDL	Moderate	Inhibition of *TNKS* inhibits Wnt signalling[19] and upregulates genes involved in cholesterol biosynthesis[20]. Wnt signalling has independently been implicated in lipid homeostasis[21,22].
FADS1-3	LDL	Strong	*FADS1*-knockout mice had lower triglyceride and total cholesterol levels[23]. *FADS2*-knockout mice had roughly doubled cholesterol synthesis rate in macrophages[24] and altered levels of multiple cholesterol esters in liver[25]. *FADS3* is least well-characterized but has 52% and 62% sequence homology with FADS1 and FADS2, respectively[26].
ALDH2	LDL	Moderate	*ALDH2* is required for alcohol metabolism, and alcohol consumption has long been known to have wide-ranging influences on lipid levels[27]. Both *ALDH2* and another alcohol metabolic enzyme, *ADH1B*, have been associated with alcohol consumption[28], variants at both loci have been associated with LDL among alcoholic men[29], and Mendelian randomization using variants near *ADH1B* and other alcohol metabolic enzymes recapitulated the causal role of alcohol consumption on LDL levels[30], suggesting that *ALDH2* causally influences LDL levels via its effect on alcohol consumption.
KPNB1	LDL	Strong	*KPNB1* knockdown reduced cellular internalization of fluorescence-labeled LDL[31].
LPIN3	LDL	Moderate	*LPIN3* is one of three lipin genes; lipins genes catalyze the synthesis of diacylglycerols[32], constituents of LDL and other lipoproteins[33] and intermediates in the synthesis of multiple classes of lipids[34].
SLC22A4/5	Crohn's	Weak	Also known as *OCTN1/2*, these genes encode proteins that transport substrates such as ergothioneine and acetylcholine, and a Crohn's-associated variant, L503F, increases SLC22A4's transport efficiency[35,36]. However, the link between altered transport efficiency and

			disease is unclear, and *IRF1* is a stronger candidate at the locus (see below).
IRF1	Crohn's	Strong	Genome-wide, variants that increase binding of *IRF1* (a transcriptional activator of the innate immune response) tend to increase Crohn's risk, and vice versa[37]. High-density genotyping of the *IRF1*/*SLC22A4*/*5* locus indicates that *IRF1*, but not *SLC22A4*/*5*, associates with Crohn's disease risk, and *IRF1* expression, but not *SLC22A4* expression, in GI biopsies was increased among Crohn's cases[38].
CARD9	Crohn's	Strong	CARD9 plays critical roles in the innate immune response and has been implicated in a variety of autoimmune conditions[39]; a loss-of-function splice variant in *CARD9* is strongly protective against Crohn's disease[40].
NOD2	Crohn's	Strong	Multiple coding variants in *NOD2* are independently associated with Crohn's disease[40,41,42].
STAT3	Crohn's	Strong	*STAT3*-knockout mice develop Crohn's-like symptoms[43].

Supplementary Table 3: Candidate causal genes curated from the literature, with supporting evidence for causality. The strength of evidence for each gene is also stated: strong indicates clear experimental evidence (*SORT1*, *PPARG*, *FADS1-3*, *KPNB1*, *STAT3*), coding loss-of-function or fine-mapped GWAS association (*IRF1*, *CARD9*, *NOD2*) or functional inference (*LPA*) linking the gene to the trait; moderate indicates less direct experimental (*TNKS*) or functional (*ALDH2*, *LPIN3*) evidence, or clear experimental evidence linking the gene to a related trait (*IRF2BP2*); weak indicates disputed evidence of causality, where another gene at the locus is a stronger candidate (*SLC22A4/5*).

Candidate causal gene	Number of TWAS hit genes at locus	Rank - TWAS	Rank - proximity	Rank - expression
SORT1	9	1	4	4
IRF2BP2	2	2	2	2
PPARG	2	1	1	1
LPA	3	2	3	1
TNKS	3	3	3	3
FADS1	4	1	1	2
FADS2	4	3	2	4
FADS3	4	4	4	3
ALDH2	3	2	1	3
KPNB1	3	1	2	3
LPIN3	3	1	3	3
SLC22A4	4	2	3	3
SLC22A5	4	1	2	2

IRF1	4	3	1	4
CARD9	5	1	1	3
NOD2	5	2	1	4
STAT3	5	4	4	5
Mean	**3.9**	**2.0**	**2.2**	**2.9**

Supplementary Table 4: Performance comparison of Fusion TWAS, expression and proximity to lead variant at ranking candidate causal genes.

		Continuous Trait				*Dichotomous Trait*			
Inference Method	**N Sims**	**Mean PPV**	**SD**	**Mean Sensitivity**	**SD**	**Mean PPV**	**SD**	**Mean Sensitivity**	**SD**
LASSO	10	0.20	0.08	1.00	0.00	1.00	0.00	1.00	0.00
Ridge	10	0.17	0.09	1.00	0.00	0.20	0.09	1.00	0.00

Supplementary Table 5: Joint association testing with penalized linear models. Using our simulation pipeline, we compared the performance of Ridge and LASSO to identify causal genes genome-wide (N=300,000). We fit a model of all predicted expression levels jointly using either LASSO or Ridge and computed empirical 95%-confidence intervals with bootstrap (see Methods). PPV measures the positive predictive value, or the proportion of associated genes that are causal. Sensitivity measures the proportion of causal genes that are associated.

		Continuous Trait			*Dichotomous Trait*		
Inference Method	**N Sims**	**Mean Adjusted R2**	**SD**	**P-value for diff (t-test)**	**Mean AUPRC**	**SD**	**P-value for diff (t-test)**
LASSO	10	0.128	9.88E-03	1.80E-04	0.010	3.86E-04	1.27E-09
Ridge	10	0.107	1.07E-02		0.047	4.71E-03	

Supplementary Table 6: Risk prediction in simulations using predicted expression. Using our simulation pipeline, we generated a complex quantitative trait (or dichotomous) for N=50,000 GWAS individuals. We then predicted expression into the simulated group using fitted models in separate reference panels with varying sample size (Expr Ref Panel Size). We then measured how well a joint model (Inference Method) predicts downstream risk for GWAS individuals (see Methods).

References

8. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
9. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
10. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
11. Strong, A. *et al.* Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism. *J. Clin. Invest.* **122**, 2807–2816 (2012).
12. Mortensen, M. B. *et al.* Targeting sortilin in immune cells reduces proinflammatory cytokines and atherosclerosis. *J. Clin. Invest.* **124**, 5317–5322 (2014).
13. Patel, K. M. *et al.* Macrophage sortilin promotes LDL uptake, foam cell formation, and atherosclerosis. *Circ. Res.* **116**, 789–796 (2015).
14. Westerterp, M. & Tall, A. R. SORTILIN: many headed hydra. *Circ. Res.* **116**, 764–766 (2015).
15. Chen, H.-H. *et al.* IRF2BP2 Reduces Macrophage Inflammation and Susceptibility to Atherosclerosis. *Circ. Res.* **117**, 671–683 (2015).
16. Duan, Y. *et al.* Peroxisome Proliferator-activated receptor γ activation by ligands and dephosphorylation induces proprotein convertase subtilisin kexin type 9 and low density lipoprotein receptor expression. *J. Biol. Chem.* **287**, 23667–23677 (2012).
17. Gbaguidi, F. G. *et al.* Peroxisome proliferator-activated receptor (PPAR) agonists decrease lipoprotein lipase secretion and glycated LDL uptake by human macrophages. *FEBS Lett.* **512**, 85–90 (2002).
18. Kronenberg, F. *et al.* Lipoprotein(a)- and low-density lipoprotein-derived cholesterol in nephrotic syndrome: Impact on lipid-lowering therapy? *Kidney Int.* **66**, 348–354 (2004).
19. Huang, S.-M. A. *et al.* Tankyrase inhibition stabilizes axin and antagonizes Wnt signalling. *Nature* **461**, 614–620 (2009).
20. Zedell, C. Identification of a novel function of tankyrase : inhibition of tankyrase catalytic activity leads to increased cellular cholesterol levels. (Norwegian University of Life Sciences, Ås, 2017).
21. Scott, C. C. *et al.* Wnt directs the endosomal flux of LDL-derived cholesterol and lipid droplet homeostasis. *EMBO Rep.* **16**, 741–752 (2015).
22. Behari, J. *et al.* Liver-Specific β-Catenin Knockout Mice Exhibit Defective Bile Acid and Cholesterol Homeostasis

and Increased Susceptibility to Diet-Induced Steatohepatitis. *Am. J. Pathol.* **176**, 744–753 (2010).

23. Powell, D. R. *et al.* Fatty acid desaturase 1 knockout mice are lean with improved glycemic control and decreased development of atheromatous plaque. *Diabetes Metab. Syndr. Obes.* **9**, 185–199 (2016).
24. Rosenblat, M., Volkova, N., Roqueta-Rivera, M., Nakamura, M. T. & Aviram, M. Increased macrophage cholesterol biosynthesis and decreased cellular paraoxonase 2 (PON2) expression in Delta6-desaturase knockout (6-DS KO) mice: beneficial effects of arachidonic acid. *Atherosclerosis* **210**, 414–421 (2010).
25. Stroud, C. K. *et al.* Disruption of FADS2 gene in mice impairs male reproduction and causes dermal and intestinal ulceration. *J. Lipid Res.* **50**, 1870–1880 (2009).
26. Buckley, M. T. *et al.* Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Mol. Biol. Evol.* **34**, 1307 (2017).
27. Baraona, E. & Lieber, C. S. Effects of ethanol on lipid metabolism. *J. Lipid Res.* **20**, 289–315 (1979).
28. Jorgenson, E. *et al.* Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol. Psychiatry* **22**, 1359–1367 (2017).
29. Yokoyama, A. *et al.* Alcohol Dehydrogenase-1B (rs1229984) and Aldehyde Dehydrogenase-2 (rs671) Genotypes Are Strong Determinants of the Serum Triglyceride and Cholesterol Levels of Japanese Alcoholic Men. *PLoS One* **10**, e0133460 (2015).
30. Vu, K. N. *et al.* Causal Role of Alcohol Consumption in an Improved Lipid Profile: The Atherosclerosis Risk in Communities (ARIC) Study. *PLoS One* **11**, e0148765 (2016).
31. Bartz, F. *et al.* Identification of cholesterol-regulating genes by targeted RNAi screening. *Cell Metab.* **10**, 63–75 (2009).
32. Eaton, J. M., Mullins, G. R., Brindley, D. N. & Harris, T. E. Phosphorylation of lipin 1 and charge on the phosphatidic acid head group control its phosphatidic acid phosphatase activity and membrane association. *J. Biol. Chem.* **288**, 9933–9945 (2013).
33. Lalanne, F., Pruneta, V., Bernard, S. & Ponsin, G. Distribution of diacylglycerols among plasma lipoproteins in control subjects and in patients with non-insulin-dependent diabetes. *Eur. J. Clin. Invest.* **29**, 139–144 (1999).
34. Reue, K. & Dwyer, J. R. Lipin proteins and metabolic homeostasis. *J. Lipid Res.* **50 Suppl**, S109–14 (2009).
35. Taubert, D., Grimberg, G., Jung, N., Rubbert, A. & Schömig, E. Functional role of the 503F variant of the organic cation transporter OCTN1 in Crohn's disease. *Gut* **54**, 1505–1506 (2005).
36. Pochini, L. *et al.* The human OCTN1 (SLC22A4) reconstituted in liposomes catalyzes acetylcholine transport

which is defective in the mutant L503F associated to the Crohn's disease. *Biochim. Biophys. Acta* **1818**, 559–565 (2012).

37. Reshef, Y. A. *et al.* Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. (2017). doi:10.1101/204685
38. Huff, C. D. *et al.* Crohn's disease and genetic hitchhiking at IBD5. *Mol. Biol. Evol.* **29**, 101–111 (2012).
39. Zhong, X., Chen, B., Yang, L. & Yang, Z. Molecular and physiological roles of the adaptor protein CARD9 in immunity. *Cell Death Dis.* **9**, 52 (2018).
40. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43**, 1066–1073 (2011).
41. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
42. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
43. Welte, T. *et al.* STAT3 deletion during hematopoiesis causes Crohn's disease-like pathogenesis and lethality: a critical role of STAT3 in innate immunity. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 1879–1884 (2003).