














Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome

Eleonora Porcu ^{1,2,3✉}, Marie C. Sadler ^{2,3}, Kaido Lepik^{4,5}, Chiara Auwerx ^{1,2,3}, Andrew R. Wood ⁶, Antoine Weihs ⁷, Maroun S. Bou Sleiman⁸, Diogo M. Ribeiro ^{2,9}, Stefania Bandinelli¹⁰, Toshiko Tanaka¹¹, Matthias Nauck ^{12,13}, Uwe Völker ^{13,14}, Olivier Delaneau ^{2,9}, Andres Metspalu ¹⁵, Alexander Teumer ^{13,16}, Timothy Frayling ¹⁷, Federico A. Santoni¹⁸, Alexandre Reymond¹ & Zoltán Kutalik ^{2,3,6,9}

Comparing transcript levels between healthy and diseased individuals allows the identification of differentially expressed genes, which may be causes, consequences or mere correlates of the disease under scrutiny. We propose a method to decompose the observational correlation between gene expression and phenotypes driven by confounders, forward- and reverse causal effects. The bi-directional causal effects between gene expression and complex traits are obtained by Mendelian Randomization integrating summary-level data from GWAS and whole-blood eQTLs. Applying this approach to complex traits reveals that forward effects have negligible contribution. For example, BMI- and triglycerides-gene expression correlation coefficients robustly correlate with trait-to-expression causal effects ($r_{BMI} = 0.11$, $P_{BMI} = 2.0 \times 10^{-51}$ and $r_{TG} = 0.13$, $P_{TG} = 1.1 \times 10^{-68}$), but not detectably with expression-to-trait effects. Our results demonstrate that studies comparing the transcriptome of diseased and healthy subjects are more prone to reveal disease-induced gene expression changes rather than disease causing ones.

¹Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. ²Swiss Institute of Bioinformatics, Lausanne, Switzerland. ³University Center for Primary Care and Public Health, Lausanne, Switzerland. ⁴Institute of Computer Science, University of Tartu, Tartu, Estonia. ⁵Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia. ⁶Genetics of Complex Traits, College of Medicine and Health, University of Exeter, Exeter, Devon, UK. ⁷Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany. ⁸Laboratory of Integrative Systems Physiology, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland. ⁹Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. ¹⁰Local Health Unit Toscana Centro, Florence, Italy. ¹¹Clinical Res Branch, National Institute of Aging, Baltimore, MD, USA. ¹²Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany. ¹³DZHK (German Centre for Cardiovascular Research), partner site Greifswald, Greifswald, Germany. ¹⁴Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. ¹⁵Estonian Biobank, University of Tartu, Tartu, Estonia. ¹⁶Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany. ¹⁷University of Exeter Medical School, University of Exeter, Exeter Devon, UK. ¹⁸Endocrine, Diabetes, and Metabolism Service, Lausanne University Hospital, Lausanne, Switzerland. ✉email: eleonora.porcu@unil.ch

To unravel the genetics of complex diseases and traits causes, multiple approaches have concentrated on contrasting the expression of mRNA transcripts in two different groups of samples to understand how genes are expressed in health and disease^{1–4}. This allows identifying differentially expressed genes (DEGs) that can be used to obtain mechanistic insights from diseases or serve as clinical biomarkers for early diagnostics. However, DEG analyses are unable to distinguish between causes, consequences, or mere correlations between gene expression and phenotypes. To understand the contributions to observed trait-expression correlations, both the assessment of bidirectional causal effects and the impact of (unmeasured) confounders are needed. We argue that if the observed correlations and bidirectional causal effects are estimated, the contribution of such confounders can be evaluated.

Genome-wide association studies (GWAS) identified thousands of common genetic variants associated with complex human traits⁵ and studies on expression quantitative trait loci (eQTLs) showed how genetic variants contribute to the regulation of gene expression levels⁶. The overlay of the two methodologies showed that trait-associated SNPs are three times more likely to be eQTLs^{7–10}, suggesting that gene expression is a reliable intermediary between DNA variation and higher-order complex phenotypes. Starting from this hypothesis, many statistical approaches integrating GWAS and eQTLs summary statistics have been proposed to detect these overlapping associations^{9,11,12}. However, while these studies aim to identify genes whose (genetically determined) expression is significantly associated with complex traits, they do not aim to estimate the strength of the causal effect and are unable to distinguish causation from pleiotropy (i.e., when a genetic variant independently affects gene expression and phenotype). This challenge can be addressed by combining summary-level data from eQTL and GWAS studies in a two-sample Mendelian Randomization framework¹³ to evaluate whether gene expression has a causal influence on a complex trait. Such methods successfully identified thousands of genes associated with complex traits.

Yet, these transcriptome-wide approaches only use *cis*-eQTLs as instruments to tease out the causal effect of gene expression on a complex trait even though the variation in gene expression may be secondary to, rather than causal for, the disease process (“reverse causation”). Disease-associated genetic variants affect expression levels more often in *trans* than in *cis*¹⁴. Hence, polygenic risk scores (PRS) have been used to evaluate the association between genetically predicted complex traits and gene expression levels¹⁴. However, PRS-based approaches are prone to detect associations merely due to pleiotropic SNPs.

In this work, to circumvent this issue and elucidate the impact of diseases on the transcriptome program at a large scale and in a principled way, we propose a reverse transcriptome-wide Mendelian randomization approach (revTWMR), which integrates summary-level data from GWAS and *trans*-eQTLs studies in an MR framework to estimate the causal effect of phenotypes on gene expression. By combining revTWMR results with the causal effects

of gene expression on phenotypes—estimated by transcriptome-wide Mendelian randomization (TWMR)¹⁵—we obtain a clear picture of the bidirectional causal effects between gene expression and complex traits (Fig. 1) and evaluate their contribution to their observational correlation.

Results

Overview of the approach. We recently developed a transcriptome-wide summary statistics-based Mendelian randomization approach (TWMR¹⁵) integrating summary-level data from GWAS and *cis*-eQTL studies. Applying TWMR to summary data from whole blood *cis*-eQTL meta-analyses from >32,000 individuals (eQTLGen Consortium¹⁴) and publicly available GWAS summary statistics revealed an atlas of putative functionally relevant genes for several complex human traits¹⁵. This approach can be reversed to design a multi-instrument MR approach to estimate the causal effect of a phenotype (exposure) on gene expression (outcome) (revTWMR, Fig. 1). For each gene, using the inverse-variance weighted meta-analysis of ratio estimates from summary statistics¹⁶, we estimate the causal effect of a phenotype on the expression of the probed gene as

$$\hat{\alpha} = \frac{\sum_{j=1}^N \beta_j \gamma_j}{\sum_{j=1}^N \beta_j^2} \quad (1)$$

where β_j and γ_j are the standardized effect sizes of SNP_{*j*} on the phenotype and on the expression level of the probed gene, respectively, and *N* is the number of independent SNPs used as instrumental variables.

Applying revTWMR to GWAS and eQTL summary statistics.

We applied revTWMR to assess causal associations between 12 complex traits—body mass index (BMI), Crohn’s disease (CD), educational attainment (EDU), fasting glucose (FG), high-density lipoprotein (HDL), height, low-density lipoprotein (LDL), rheumatoid arthritis (RA), schizophrenia (SCZ), total cholesterol (TC), triglycerides (TG), and waist-to-hip ratio adjusted for BMI (WHRadjBMI)—and the expression of 19,942 genes. We combined summary whole blood *trans*-eQTLs data from the eQTLGen Consortium¹⁴, with large publicly available GWAS for the traits of interest^{17–22} (see Methods). Together, we identified 46 genes significantly affected by at least one phenotype ($P_{\text{revTWMR}} < 2.5 \times 10^{-6} = 0.05/19,942$), often corroborating known biological associations (Supplementary Data 1). In parallel, we performed TWMR analyses on the same set of traits, allowing testing for the presence of bidirectional effects (see Methods) (Supplementary Data 2).

The most influential traits in our analysis were TG and RA, significantly influencing the expression of 26 and 15 genes, respectively. These were analyzed for functional enrichment with UniProtKB²³, KEGG pathway²⁴, Gene Ontology²⁵, and InterPro²⁶. For TG, cholesterol metabolism (UniProt KW-0153)

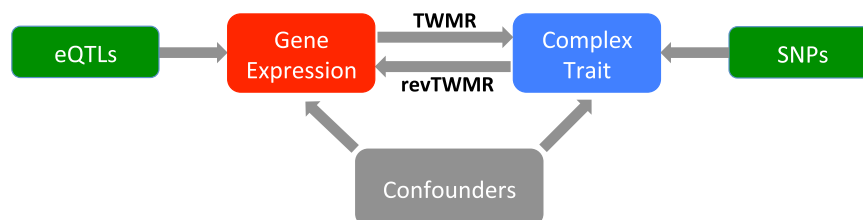


Fig. 1 TWMR and revTWMR. Schematic representation of how TWMR and revTWMR dissect bidirectional causal and confounder contributions to the observed correlation between gene expression and phenotype.

was the most significantly enriched class (Supplementary Data 3). For RA, immunoglobulin-related terms (InterPro IPR013106, IPR007110, and IPR013783) were the top significantly enriched classes (Supplementary Data 4). Closer investigation revealed that this enrichment is due to the presence of 7 T-cell receptor α and β variable genes (*TRAV* and *TRBV*). Interestingly, a bias in $V\beta$ gene utilization by T cells in patients suffering from RA was reported²⁷.

Focusing on serum lipid levels (Supplementary Data 5), revTWMR revealed that in addition to the 26 genes affected by TG, the expression of eight genes is altered by HDL-cholesterol levels. In line with the commonly observed negative correlation between HDL and TG²⁸, five genes were impacted by both traits with an opposite direction of the causal effect (Supplementary Data 1). Regarding the impact of high HDL levels, we found that it reduced the expression of squalene synthase (*FDFT1*; $\alpha_{\text{revTWMR}} = -0.14$, $P_{\text{revTWMR}} = 1.3 \times 10^{-10}$, Supplementary Fig. 1), a key enzyme of the cholesterol biosynthesis pathway²⁹. Interestingly, serum levels of squalene, the product of squalene synthase were found to negatively correlate with HDL-cholesterol³⁰. Genes involved in cholesterol transport were impacted too: high HDL negatively impacted the expression of the LDL receptor (*LDLR*; $\alpha_{\text{revTWMR}} = -0.11$, $P_{\text{revTWMR}} = 1.0 \times 10^{-06}$, Supplementary Fig. 1), while having a positive impact on *MYLIP* (also known as *IDOL*; $\alpha_{\text{revTWMR}} = 0.18$, $P_{\text{revTWMR}} = 2.2 \times 10^{-16}$, Supplementary Fig. 1), a ubiquitin ligase that induces degradation of the LDL receptor³¹. In parallel, HDL increased the expression of the *ABCA1* ($\alpha_{\text{revTWMR}} = 0.24$, $P_{\text{revTWMR}} = 9.5 \times 10^{-29}$, Supplementary Fig. 1) and *ABCG1* ($\alpha_{\text{revTWMR}} = 0.19$, $P_{\text{revTWMR}} = 4.1 \times 10^{-20}$, Supplementary Fig. 1), two transporters responsible for cholesterol efflux from macrophages³². While we did not observe a significant effect of *ABCA1* on HDL and TC levels through TWMR, an association between *ABCA1* and these two traits was previously reported by a GWAS³³, suggesting a complex regulatory mechanism. Together, these results are reminiscent of the well-described negative feedback mechanisms that tightly control cholesterol biosynthesis and uptake³⁴.

As the other traits influenced only a smaller number of genes, no further significant enrichments were found. Nevertheless, a gene-by-gene investigation revealed many known or highly plausible associations, such as the significant effect of BMI on *ALDH1A1* ($\alpha_{\text{revTWMR}} = -0.17$, $P_{\text{revTWMR}} = 2.2 \times 10^{-06}$, Supplementary Fig. 1), an enzyme that converts retinaldehyde to retinoic acid³⁵. Retinoids have long been implicated in adipogenesis^{36,37} and *ALDH1A1* expression in visceral adipose tissue was shown to positively correlate with BMI³⁸.

Despite strong indications of functional relevance, most revTWMR-implicated genes fall into genomic regions completely missed by GWAS, as is illustrated by the fact that revTWMR p values are completely uncorrelated ($r < 0.05$) with those obtained by classical gene-based GWAS test performed using PASCAL³⁹ (See Methods; Supplementary Fig. 2). In line with this observation, only one out of the 46 revTWMR-identified genes were significant for TWMR: *FDFT1* shows a negative causal feedback loop between its expression and TG ($\alpha_{\text{TWMR}} = -0.04$, $P_{\text{TWMR}} = 1.6 \times 10^{-31}$ and $\alpha_{\text{revTWMR}} = 0.15$, $P_{\text{revTWMR}} = 1.3 \times 10^{-09}$, Supplementary Fig. 1).

To test the robustness of revTWMR, we performed MR analysis using two alternative approaches allowing the presence of invalid instruments: a weighted median method that assumes that a majority of genetic variants are valid instruments⁴⁰ and a weighted mode-based estimation method that assumes a plurality of genetic variants are valid instruments⁴¹. Results strongly supported the robustness of the IVW-based findings as 47 out of the 51 trait-gene revTWMR associations were significant in at least one of these methods ($P_{\text{MR}} < 0.05/51$; Supplementary Data 6). This analysis revealed 32 additional genes significantly

affected by RA (30 genes), TC (1), and TG (1) (Supplementary Data 7). Of note, the additional 30 genes associated with RA, strengthened the previously detected enrichment for the immunoglobulin InterPro functional groups (Supplementary Data 8).

Pleiotropic SNPs lead to biased causal effect estimates. The validity of revTWMR, as any MR approach, relies on three assumptions about the instruments: (i) they must be sufficiently strongly associated with the exposure; (ii) they should not be associated with any confounder of the exposure-outcome relationship; and (iii) they should be associated with the outcome only through the exposure. The third assumption is crucial as MR causal estimates will be biased in the presence of pleiotropy^{42,43}. Accordingly, revTWMR assumes that all genetic variants used as instrumental variables affect the gene expression only through the phenotype under scrutiny and not through independent biological pathways.

To test for the presence of pleiotropy, we used a similar approach to MR-PRESSO global test^{43,44}, performing Cochran's Q test. Under the assumption that the majority of SNPs influence gene expression only through the phenotype tested in the model, SNPs violating the third MR assumption would significantly increase the Cochran's heterogeneity Q statistic (see Methods), allowing their detection and exclusion. This was the case for 16 of the 52 originally significant trait \rightarrow gene associations. Out of these 16 associations, nine passed the heterogeneity test after removing pleiotropic SNPs from the instrumental variables. Moreover, this procedure led to the identification of six additional associations initially masked by heterogeneity, bringing the final number of robust associations to 51 (Supplementary Data 1). Importantly, revTWMR, like other MR methods, discriminates likely causal effects from pleiotropy, as illustrated by the example of *STX1B*, a gene that was found to be associated with EDU through a PRS approach ($P_{\text{PRS}} = 1.3 \times 10^{-20}$)¹⁴. Applying revTWMR, we did not observe an association between EDU and *STX1B* ($\alpha_{\text{revTWMR}} = 0.03$, $P_{\text{revTWMR}} = 0.83$) and detected a highly pleiotropic variant, rs2456973, strongly associated with hematological and anthropometric traits⁴⁵ (Supplementary Data 9 and Supplementary Fig. 3).

Trait correlation. Exploring the shared effect of complex traits and diseases on transcriptional programs can provide useful etiological insights. Hence, for every phenotype-pair (P_i, P_j) we computed the gene expression perturbation correlation between the respective causal effect estimates of each phenotype on the gene expression ($\hat{\rho}_{P(i,j)} = \text{corr}(\alpha_{P_i \rightarrow E}, \alpha_{P_j \rightarrow E})$) across a subset of 2974 independent genes across the genome¹⁵. Among the 55 pairs of traits, we found 21 significant correlations (FDR $< 1\%$). We compared these results with the genetic correlation ($\hat{\rho}_G$) between traits estimated by LD score regression⁴⁶ and found a remarkable concordance between the two estimates ($r = 0.84$). On average, $\hat{\rho}_P$ represents 56% of $\hat{\rho}_G$. Although $\hat{\rho}_G$ having smaller variance may explain part of this attenuation, we think that the main reason behind this observation is that only a part of $\hat{\rho}_G$ translates into consequences on gene expression level in whole blood (Supplementary Fig. 4). In particular, nine pairs of traits showed significance for both $\hat{\rho}_P$ and $\hat{\rho}_G$, whereas 12 were significant only for $\hat{\rho}_P$, and seven only for $\hat{\rho}_G$. Among the significant correlations not identified by LD score regression $\hat{\rho}_G$, we found that HDL and LDL are negatively correlated ($\hat{\rho}_P = -0.13$, FDR = 3.1×10^{-09}) and that RA positively correlated with several traits: CD ($\hat{\rho}_P = 0.08$, FDR = 4.0×10^{-04}), SCZ ($\hat{\rho}_P = 0.14$, FDR = 1.5×10^{-10}), height ($\hat{\rho}_P = 0.09$, FDR = 6.0×10^{-05}), TC ($\hat{\rho}_P = 0.08$, FDR = 4.0×10^{-04}), and TG ($\hat{\rho}_P = 0.12$, FDR = 4.8×10^{-08}) (Supplementary Data 10).

Partitioning the observational correlation. As a proof-of-concept, we asked how highly revTWMR-identified causal genes would rank in a DEG analysis. To address this question, we collected the observational correlation estimates between whole blood gene expression levels and the quantitative traits in three independent European cohorts (EGCUT ($N = 488$), InChianti ($N = 609$), and SHIP-Trend ($N = 991$)).

Correlating revTWMR effects to observational correlations (equivalent to DEG analysis), we found a significant agreement for all the traits (Table 1). We reestimated these correlations accounting for the error in the compared estimates (regression dilution bias) (see Methods). No significant correlation between observational correlations and the causal effects of the gene expression on phenotypes estimated by TWMR was observed (Table 1). Of note, when we correlated the P values of the observational correlations with those obtained by conventional gene-based tests using GWAS results, we detected a significant concordance only for HDL ($r = 0.05$, $P = 1.3 \times 10^{-10}$) and TG ($r = 0.03$, $P = 5.7 \times 10^{-04}$) (Supplementary Data 11).

As we previously showed that causal feedback loops are rare (i.e., $\alpha_{TWMR} * \alpha_{revTWMR} = 0$), the observational correlation (r) can be approximated as the sum of the bidirectional effects estimated by TWMR and revTWMR plus the contribution of the

confounding factors (see Methods). Hence, we calculated the proportion of correlation due to confounders. For each gene we calculated the contribution of TWMR and revTWMR as $\frac{\alpha_{TWMR}}{r}$ and $\frac{\alpha_{revTWMR}}{r}$, respectively. Consequently, the contribution of confounders is $1 - \frac{\alpha_{TWMR}}{r} - \frac{\alpha_{revTWMR}}{r}$. In each correlation bin (Fig. 2) we combined such contributions using inverse-variance meta-analysis and revealed that the observed correlation between gene expression and phenotype is mainly driven by confounders. For example, for genes correlated ($|r| > 0.1$) with BMI, 83% ($P < 5.0 \times 10^{-324}$) of the correlation is due to the confounders, 17% ($P = 6.7 \times 10^{-45}$) to the effect of BMI on gene expression and 0% ($P = 0.67$) to the forward effect (Fig. 2 and Supplementary Data 12). A similar scenario was observed for TG: 90% ($P < 5.0 \times 10^{-324}$) of the correlation is due to confounders and only 10% ($P = 2.9 \times 10^{-35}$) and 0% ($P = 0.98$) are due to reverse and forward effect of the gene expression on TG, respectively. For HDL we observed a stronger effect due to confounders (94%, $P < 5.0 \times 10^{-324}$) and a mild reverse effect (6%, $P = 3.9 \times 10^{-15}$) (Fig. 2).

Genes affected by lipid traits are linked to drug targets. It is important to note that since GWAS findings point to loci underlying disease susceptibility, changes in expression detected by revTWMR do not necessarily represent the consequence of the disease but can also reflect the consequences of a genetic predisposition to that disease. Therefore, identified genes might represent early biomarkers of disease (predisposition) and modulation of their expression could be a promising therapeutical strategy. For this reason, we assessed whether the protein products of the transcripts identified by our revTWMR analysis are targets of drugs used to treat the disease in question. We started by defining a set of drugs relevant to the traits under investigation according to DrugBank¹³. Next, we retrieved high confidence interactions (confidence score > 0.7) involving these drugs, from STITCH, a manually curated database of predicted and experimental chemical-protein interactions¹². We then searched for proteins that were (a) identified as dysregulated by revTWMR and (b) targeted by a drug indicated for the treatment of a given trait.

The gene product of 4 out of the 8 genes detected by revTWMR for HDL-cholesterol met these criteria: phospholipid-transporting ATPase ABCA1 (*ABCA1*), squalene synthase (*FDFT1*), low-density lipoprotein receptor (*LDLR*), and sterol regulatory element-binding protein 1 (*SREBF1*) which interact with atorvastatin, lovastatin, pravastatin, and simvastatin. We

Table 1 Correlation between observational phenotype-gene expression correlation and revTWMR and TWMR effects.

Trait	revTWMR		TWMR	
	correlation (adjusted)	P value	correlation	P value
BMI	0.11 (0.37)	1.97E-51	0	0.75
EDU	0.04 (0.29)	1.50E-08	0.02	0.04
Fasting glucose	0.08 (0.24)	2.56E-17	0.01	0.40
HDL	0.10 (0.27)	1.86E-43	0.01	0.18
height	0.09 (0.38)	3.26E-37	0.01	0.33
LDL	0.02 (0.09)	5.36E-04	0.02	0.05
TC	0.04 (0.13)	5.30E-08	0.03	0.01
TG	0.13 (0.32)	1.11E-68	0	0.73
WHR	0.02 (0.14)	1.74E-04	0.01	0.40

For each phenotype available in at least two cohorts, we calculated the correlation between the observational correlation estimates and the revTWMR and TWMR effects. The P value indicates the significance of the correlation coefficient calculated using a two-sided t -test. For significant correlations, we computed the adjusted correlation correcting for regression dilution bias.

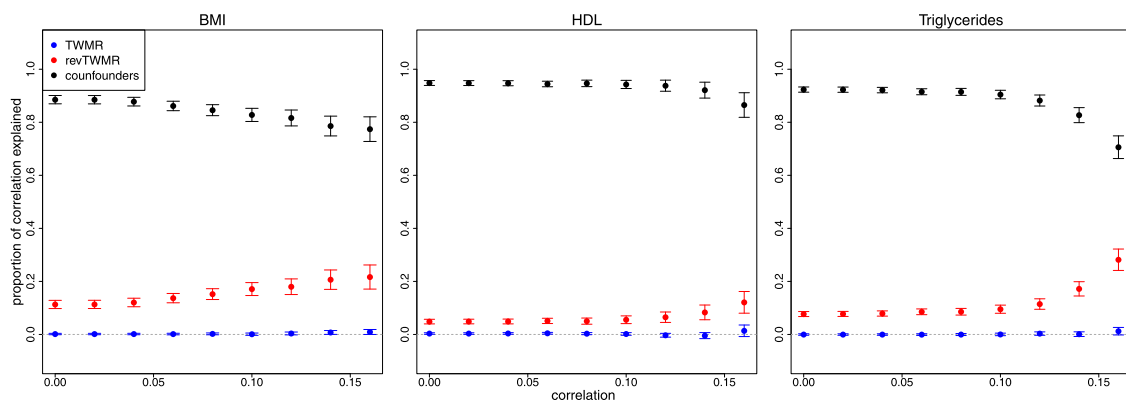


Fig. 2 Partitioning the gene expression-trait observational correlation for BMI, HDL, and triglycerides. Using all the genes tested by TWMR and revTWMR ($N = 10,395$ for BMI, $N = 10,391$ for HDL, and $N = 10,390$ for TG), for each bin of correlation (absolute value) we plotted the combined contributions of the forward (TWMR, blue dots) and reverse (revTWMR, red dots) effect of the gene expression on the trait, the contribution of confounders (black dots). Data were presented as estimated contributions and 95% confidence intervals.

found that *ABCA1*, *FDFT1*, and *SREBF1* are also dysregulated by high triglyceride levels. We did not find drug targets among the genes significantly dysregulated by RA (Supplementary Data 13).

Tissue-specific effects. Many traits and diseases manifest themselves only in certain tissues. For this reason, we performed tissue-specific revTWMR analyses using tissue-specific *trans*-eQTLs identified by the Genotype-Tissue Expression Project (GTEx)⁴⁷, which provides a unified view of genetic effects on gene expression across 49 human tissues. We tested the 51 previously identified significant trait \rightarrow gene associations found in whole blood and detected three genes showing tissue-specific associations ($P_{\text{revTWMR}} < 0.05/51$, Supplementary Data 14). These include a negative effect of RA on *MYO1B* in the kidney cortex ($\alpha_{\text{revTWMR}} = -1.71$, $P_{\text{revTWMR}} = 2.4 \times 10^{-04}$), as well as a positive effect on *TRBV19* in the small intestine terminal ileum ($\alpha_{\text{revTWMR}} = 1.14$, $P_{\text{revTWMR}} = 1.7 \times 10^{-04}$) and in esophagus mucosa ($\alpha_{\text{revTWMR}} = 0.63$, $P_{\text{revTWMR}} = 5.7 \times 10^{-04}$). In addition, we observed a negative effect of HDL on *MYLIP* ($\alpha_{\text{revTWMR}} = -0.98$, $P_{\text{revTWMR}} = 5.9 \times 10^{-04}$) in the brain spinal cord (cervical c-1).

Testing the reverse causal effects in mouse models. RevTWMR pointed to 26 genes affected by TG levels. To experimentally validate these causal effects, we analyzed how the hepatic expression of these genes and TG levels were co-affected in the mouse BXD genetic reference panel, a set of inbred mice strains generated by crossing the C57BL/6J and DBA/2 strains, upon switch to a high-fat diet (HFD)⁴⁸. We hypothesized that if HFD-induced changes in TG correlate with HFD-induced changes in the expression of these genes, then diet-induced changes in TG might indeed be causal to changes in the expression of selected genes. Importantly, this relies on the assumption that there is a reasonable correlation between the expression of these genes in human blood and mice livers and that TG \rightarrow gene expression mechanisms are conserved between the two species. Among the 19,942 genes tested in revTWMR, 10,841 had a detectable ortholog measured in the mouse samples. For each gene, we computed the Spearman correlation between HFD-induced expression fold change and the diet-induced TG differences as an indicator of genes perturbed by TG levels. Among the genes showing a significant ($P < 0.001$) correlation, we found an enrichment ($P_{\text{Fisher}} = 4.0 \times 10^{-04}$) of revTWMR genes ($P_{\text{revTWMR}} < 2.5 \times 10^{-06}$). Performing the same analysis on significant TWMR genes ($P_{\text{TWMR}} < 3 \times 10^{-06}$) did not yield an enrichment ($P_{\text{Fisher}} = 0.7$), confirming that correlations are mainly driven by the effect of TG on gene expression (Supplementary Data 15).

Discussion

We presented a Mendelian randomization approach to study the impact of human phenotypes on the transcriptome. When calculating the reverse effect of phenotypes on gene expression it is important to note that findings from GWAS provide a measure of the genetic liability to develop a disease. In fact, using such genetic liability as exposure, the association with the gene expression does not necessarily reflect the consequence of the fully developed disease but might reflect a consequence of an early asymptomatic stage of the disease or a mere genetic predisposition⁴⁹. Hence, these changes in the expression of revTWMR-implicated genes may occur before the disease manifest itself. As such, revTWMR results should not exclusively interpreted as markers of downstream mechanisms post-disease onset, but as potential early biomarkers.

Across the 46 genes identified by revTWMR, we observed a clear trend for functional relevance. Genes perturbed by complex diseases seem to confirm several previously reported associations between immune-related genes (*TRBV*) and RA²⁷. In addition, revTWMR allowed gaining insight into the regulatory mechanisms controlling biological pathways, as illustrated with serum lipid levels. We observed that high HDL-cholesterol lowers the expression of genes involved in cholesterol biosynthesis (*FDFT1*) and cellular cholesterol uptake (*LDLR*), while it increases the expression of genes responsible for the degradation of *LDLR* (*MYLIP*) and cholesterol efflux (*ABCA1* and *ABCG1*). Together, this suggests that high HDL levels prevent intracellular cholesterol overload, which could explain its known cardioprotective effects⁵⁰. However, TG levels, which were shown to independently increase risk of coronary artery disease (CAD)⁵¹, impact the expression of the same genes in the opposite direction. Hence, high TG levels might increase CAD risk through an intracellular accumulation of cholesterol. The biological relevance of our findings is further supported by our drug target analysis, which found that four genes (*SREBF1*, *FDFT1*, *LDLR*, and *ABCA1*) whose expression was perturbed by serum lipid traits were targets of statins, a category of drugs aiming at regulating the very same traits. Lipids are major modulators of CAD risk^{50,51} and established regulators of gene expression⁵². Hence, drugs targeting these downstream genes might modulate CAD risk, even though mediation analysis is warranted to support this hypothesis.

Combining results of DEG analysis and bidirectional TWMR allowed decomposing the observational correlation between whole blood gene expression and complex traits. This analysis showed that DEGs often reflect disease-induced changes in the transcriptome rather than disease-causing ones. Importantly, we observed that most of the correlation between gene expression and complex traits is due to confounders, which could partially be explained by age and sex being important determinants of both. The remaining correlation can almost entirely be explained by the trait-to-gene expression causal effects. Just like single SNPs, the individual expression of most genes has only a minute contribution to the phenotype, even if cumulatively their effect can be substantial. Diseases, however, represent a major burden for the organism, which can lead to drastic changes in the transcriptome program. In light of these considerations, one would expect that the correlation between a gene's expression level and a complex trait is reflecting disease status rather than an expression-to-trait link. Validating revTWMR requires large cohorts in which gene expression is measured before and after trait-modifying interventions. As conducting such studies in humans imposes serious logistic and ethical hurdles, we turned to mice studies to assess the impact of diet-induced changes in TG on gene expression and found that genes detected by TG-revTWMR are enriched among mouse orthologs whose HFD-induced changes in expression correlate with HFD-induced changes in TG.

Our approach has its limitations, which need to be considered when interpreting results. First, our results are mainly focused on gene expression levels in whole blood. This is primarily due to the reduced power resulting from the small sample sizes when conducting tissue-specific analyses. However, because gene regulation is tissue-specific and many diseases manifest themselves only in certain tissues, future possibilities to interrogate larger and more diverse tissue-specific *trans*-eQTL datasets could unravel causative disease-gene links for genes not differentially expressed in blood. This speculation is supported both by the fact that the effect sizes of the few tissue-specific associations we detected were more than fivefold larger than those estimated using whole blood data, as well as recent reports suggesting that *trans*-eQTLs are particularly cell type-specific⁵³.

Another caveat lies in the fact that differences in power makes it difficult to compare the results of TWMR and revTWMR. One of the most important determinants of statistical power for MR is the sample size available for the outcome, thus revTWMR is less powered, picking up mostly strong effects. Still, another factor influencing power is the number and strength of instruments. Hence, TWMR results will be more accurate once larger eQTL datasets become available, which will in turn increase the number of testable genes (currently 16 K). Finally, as with every MR approach, revTWMR is at risk of violating the MR assumptions. In particular, horizontal pleiotropy and indirect effects of the instruments on the exposures can substantially bias causal effect estimates. RevTWMR assumes that the top GWAS SNPs have a direct effect on the phenotype. In particular, correlated pleiotropy can lead to biased causal effect estimates and currently available methods that attempt to tackle such MR violations (e.g., CAUSE⁵⁴, LHC-MR⁵⁵, MR-APSS⁵⁶) require genome-wide summary statistics, which is not yet available for transcripts in a large enough sample size. However, many SNPs show indirect or pleiotropic effects. We, therefore, mitigate the influence of these potential biases by excluding pleiotropic SNPs failing the heterogeneity test. Further gain in robustness should be obtained by integrating additional phenotypes as exposures through which instruments may act, as accounting for pleiotropy is a better approach than excluding violating instruments. Such a multi-phenotype revTWMR approach will be possible only once genome-wide *trans*-eQTLs summary statistics will become available.

A very exciting perspective is that revTWMR can theoretically be extended to other types of omics data, e.g., integrating methylomics data, as alterations in DNA methylation are more often the consequence rather than the cause of diseases⁵⁷. One could apply the approach to protein levels (revPWMR) to gain further insights into the effects of complex traits on biomarkers but the sample size of proteomics datasets are currently too small.

In conclusion, our bidirectional analysis disentangles the causes and consequences of gene expression for complex traits and reveals that complex traits have a more pronounced impact on gene expression than the reverse. Therefore, studies comparing gene expression levels of diseased and healthy subjects may still point to useful biomarkers of disease predisposition or severity, but interventions that restore levels of the biomarker to normal levels will not necessarily be disease-modifying.

Methods

Reverse transcriptome-wide Mendelian randomization (revTWMR).

RevTWMR is a multi-instrument MR approach designed to estimate the causal effect of the phenotypes (exposure) on gene expression (outcome). For each gene, using an inverse-variance weighted method for summary statistics¹⁶, we define the joint causal effect of the phenotypes on the outcome as

$$\hat{\alpha} = (\hat{\beta}'C^{-1}\hat{\beta})^{-1}(\hat{\beta}'C^{-1}\hat{\gamma}) \quad (2)$$

Here β is an n -vector that contains the standardized effect size of n independent SNPs on the phenotype, derived from GWAS. γ is a vector of length n that contains the standardized effect size, in *trans*-, of each SNP on the gene expression. C is the pair-wise LD matrix between the n SNPs.

As instrumental variables, we used independent ($r^2 < 0.01$) significant ($P_{\text{GWAS}} < 5 \times 10^{-8}$) SNPs chosen among the 10 K preselected trait-associated SNPs included in a *trans*-eQTL dataset from eQTLGen Consortium (31,684 whole blood samples). As we are using only strongly independent SNPs, we use the identity matrix to approximate C . The SNPs with larger effects on the outcome than on the exposure were removed, as these would indicate a violation of MR assumptions (likely reverse causality and/or confounding).

The variance of α can be calculated approximately by the Delta method

$$\text{var}(\hat{\alpha}) = \left(\frac{\partial \hat{\alpha}}{\partial \hat{\beta}}\right)^2 * \text{var}(\hat{\beta}) + \left(\frac{\partial \hat{\alpha}}{\partial \hat{\gamma}}\right)^2 * \text{var}(\hat{\gamma}) + \left(\frac{\partial \hat{\alpha}}{\partial \hat{\beta}}\right) * \left(\frac{\partial \hat{\alpha}}{\partial \hat{\gamma}}\right) * \text{cov}(\hat{\beta}, \hat{\gamma}) \quad (3)$$

where $\text{cov}(\beta, \gamma)$ is 0 if β and γ are estimated from independent samples. We defined the causal effect Z-statistic for gene i as $\hat{\alpha}_i/\text{SE}(\hat{\alpha}_i)$, where $\text{SE}(\hat{\alpha}_i) = \sqrt{\text{var}(\hat{\alpha})_{i,i}}$.

We applied revTWMR across the human genome for a causal association between a set of 12 phenotypes and the expression levels of 19,942 genes using summary statistics from GWAS and eQTLs studies. The analysed traits include BMI, CD¹⁷, EDU²⁰, FG¹⁸, HDL-cholesterol, height, LDL-cholesterol, RA¹⁹, SCZ²¹, TC, TG, and WHRadjBMI. While for CD, EDU, FG, RA, SCZ, and WHRadjBMI summary statistics (estimated univariate effect size and standard error) originate from the most recent meta-analysis and were downloaded from the publicly available NIH Genome-wide Repository of Associations Between SNPs and Phenotypes (<https://grasp.nhlbi.nih.gov/>), for the other traits the GWAS were performed in UKBiobank and the summary statistics are from the Neale Lab (<http://www.nealelab.is/uk-biobank/>) (Supplementary Data 16). We only used SNPs on autosomal chromosomes and were available in the UK10K reference panel, which allowed estimating the LD among these SNPs and prune them. Strand ambiguous SNPs were removed.

Heterogeneity test. The validity of all MR approaches, such as revTWMR, relies on three assumptions. The third assumption (no pleiotropy) is crucial as MR causal estimates will be biased if the genetic variants (IVs) have pleiotropic effects⁴³. Hence, revTWMR assumes that all genetic variants used as instrumental variables affect the outcome only through gene expression and not through independent biological pathways. To test for the presence of pleiotropy, we used Cochran's Q test^{42,44}. In brief, we tested whether there is a significant difference between the revTWMR-effect of an instrument (i.e., $\alpha\beta_i$) and the estimated effect of that instrument on the gene expression (γ_i). We defined

$$d_i = \hat{\gamma}_i - \hat{\alpha}\hat{\beta}_i \quad (4)$$

and its variance as

$$\text{var}(d_i) = \text{var}(\hat{\gamma}_i) + (\hat{\beta}_i)^2 * \text{var}(\hat{\alpha}) + \text{var}(\hat{\gamma}_i) * (\alpha)^2 + \text{var}(\hat{\beta}_i) * \text{var}(\hat{\alpha}) \quad (5)$$

Next, we tested the deviation of each SNP using the following test statistic

$$T_i = \frac{d_i^2}{\text{var}(d_i)} \sim \chi^2_1 \quad (6)$$

In case where $P < 1 \times 10^{-4}$, we removed the SNP with largest $|d_i|$ and then repeated the test.

Transcriptome-wide Mendelian randomization (TWMR). In order to test the presence of a feedback loop of association, we ran TWMR¹⁵ for all the significant revTWMR genes. To make TWMR and revTWMR results comparable, we ran a univariable TWMR where for each gene we estimated its total effect on the phenotype. The associations between the instrumental variables and the exposure (gene expression) and the outcome (complex traits) are estimated from the same studies used for revTWMR.

Gene-based test. To compare GWAS and revTWMR results, we performed gene-based test for association summary statistics using PASCAL³⁹. PASCAL assesses the total contribution of all SNP within close physical proximity to a given gene by combining SNP association Z-statistics into gene-based P values while accounting for local LD structure.

Replication cohorts

EGCUT

Study population. The Estonian Genome Center, University of Tartu (EGCUT) cohort denotes the Estonian Biobank sample of more than 200,000 individuals or about 20% of the Estonian adult population. All Biobank participants have been genotyped and linked to electronic health records (EHR) of the Health Insurance Fund, national registries, and major hospitals. The EHR linkage captures the participants' medical history together with demographics, lifestyle information, and laboratory measurements; additional information is provided by self-completed questionnaires. Disease diagnoses are in the form of ICD-10 codes. RNA-seq data is available on 491 unrelated individuals. All Biobank participants have signed a broad informed consent to allow using their genetic and medical information for research purposes.

Whole-blood-transcriptome analysis. The preparation of RNA-seq data has been described in detail elsewhere⁵⁸. RNA-seq reads were trimmed of adapters together with low-quality leading and trailing bases using Trimmomatic (version 0.36)⁵⁹. Additional quality control was performed with FastQC (version 0.11.2). The final set of reads were mapped to a human genome reference version GRCh37.p13 using STAR (version 2.4.2a)⁶⁰. Sample mix-ups were tested and corrected for using MixupMapper⁶¹. Principal component analysis on RNA-seq read counts revealed a batch of outlying samples which was uncovered to be due to a technical problem in library preparation—affected samples were discarded. Data were normalized using the weighted trimmed mean of M values⁶² and used as log₂-transformed counts per million. To account for (hidden) batch effects, the sequencing batch date together with the first gene expression principal components were used in all subsequent analyses.

InChianti

Study population. The InCHIANTI study is a population-based sample that includes 298 individuals of age <65 years and 1155 individuals of age ≥65 years. The study design and protocol have been described in detail previously⁶³. The data collection started in September 1998 and was completed in March 2000. The INRCA Ethical Committee approved the entire study protocol.

Whole-blood-transcriptome analysis. Peripheral blood specimens were collected from 712 individuals using the PAXgene tube technology to preserve levels of mRNA transcripts. RNA was extracted from peripheral blood samples using the PAXgene Blood mRNA kit (Qiagen, Crawley, UK) according to the manufacturer’s instructions.

RNA was biotinylated and amplified using the Illumina® TotalPrep(tm) –96 RNA Amplification Kit and directly hybrid with Human HT-12_v3 Expression BeadChips that include 48,803 probes. Image data were collected on an Illumina iScan and analysed using Illumina GenomeStudio software. These experiments were performed as per the manufacturer’s instructions and as previously described⁶⁴. Quality-control analysis of gene expression levels were previously described⁶⁵.

SHIP-Trend

Study population. The Study of Health in Pomerania (SHIP-Trend) is a longitudinal population-based cohort study in West Pomerania, a region in the northeast of Germany, assessing the prevalence and incidence of common population-relevant diseases and their risk factors. Baseline examinations for SHIP-Trend were carried out between 2008 and 2012, comprising 4420 participants aged between 20 and 81 years. Study design and sampling methods were previously described⁶⁶. The medical ethics committee of the University of Greifswald approved the study protocol, and oral and written informed consents were obtained from each of the study participants.

Whole-blood-transcriptome analysis. Blood sample collection, as well as RNA preparation, were described in detail elsewhere⁶⁷. Briefly, whole-blood samples of a subset of SHIP-TREND were collected from the participants after overnight fasting (≥10 h) and stored in PAXgene Blood RNA Tubes (BD). Subsequently, RNA was prepared using the PAXgeneTM Blood miRNA Kit (QIAGEN, Hilden, Germany). The purity and concentration of RNA were determined using a NanoDrop ND-1000 UV-Vis Spectrophotometer (Thermo Scientific). To ensure a constantly high quality of the RNA preparations, all samples were analyzed using RNA 6000 Nano LabChips (Agilent Technologies, Germany) on a 2100 Bioanalyzer (Agilent Technologies, Germany) according to the manufacturer’s instructions. Samples exhibiting an RNA integrity number (RIN) less than seven were excluded from further analysis. The Illumina TotalPrep-96 RNA Amplification Kit (Ambion, Darmstadt, Germany) was used for reverse transcription of 500 ng RNA into double-stranded (ds) cDNA and subsequent synthesis of biotin-UTP-labeled antisense-cRNA using this cDNA as the template. Finally, in total 3000 ng of cRNA were hybridized with a single array on the Illumina Human HT-12 v3 BeadChips, followed by washing and detection steps in accordance with the Illumina protocol. Processing of the SHIP-Trend RNA samples was performed at the Helmholtz Zentrum München. BeadChips were scanned using the Illumina Bead Array Reader. The Illumina software GenomeStudio V 2010.1 was used to read the generated raw data, for imputation of missing values and sample quality control. Subsequently, raw gene expression data were exported to the statistical environment R, version 2.14.2 (R Development Core Team 2011). Data were normalized using quantile normalization and log₂-transformation using the lumi 2.8.0 package from the Bioconductor open-source software (<http://www.bioconductor.org/>). Finally, 991 samples were available for gene expression analysis. Technical covariates used in all statistical models included RNA amplification batch, RNA quality (RIN), and sample storage time. The SHIP-Trend expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accession GSE 36382: 991 samples were available for analysis.

Phenotype-gene expression correlation. To calculate the correlation between the phenotypes and the gene expression levels, we asked each cohort to run the following analysis. First, the inverse normal transformation was applied to phenotypes and gene expression. Next, transformed phenotypes were adjusted only for sex, age, and age², while gene expression was also corrected for other known relevant covariates. Finally, Pearson’s correlation was calculated between the adjusted trait and the adjusted expression. Finally, correlations from single cohorts were combined using inverse-variance meta-analysis, where weights are proportional to the squared standard error of the correlation estimates, as implemented in METAL⁶⁸.

Observed and true correlation between gene expression and traits. The correlation between the effects estimated by revTWMR ($\alpha_{revTWMR}$) and the observational correlation ($corr(E, T)$) measured in the individual data from EGCUT, InChianti, and SHIP-Trend was calculated using Pearson’s correlation. As such correlation does not consider the error of the estimations, for the significant correlations we used the linear errors-in-variables models to compute the potential

true correlation using the following equation

$$corr_{obs} = corr_{true} * \sqrt{1 - \frac{\sum_{j=1}^{N_{Genes}} SE(\alpha_{revTWMR})^2}{\sum_{j=1}^{N_{Genes}} \alpha_{revTWMR}^2}} * \sqrt{1 - \frac{\sum_{j=1}^{N_{Genes}} SE(corr(E, T))^2}{\sum_{j=1}^{N_{Genes}} corr(E, T)^2}} \tag{7}$$

Proportion of observational correlation explained by bidirectional causal effects.

Let E and T denote the gene expression and the trait, respectively. In addition, there may exist a confounding factor U causally impacting both of them. We can express E and T as:

$$T = \alpha_{TWMR} * E + q_T * U + \epsilon_T \tag{8}$$

And

$$E = \alpha_{revTWMR} * T + q_E * U + \epsilon_E \tag{9}$$

where α_{TWMR} and $\alpha_{revTWMR}$ are the causal effects of E on T and of T on E estimated by TWMR and revTWMR respectively; q_T and q_E are the causal effects of the confounders on T and E ; and $\epsilon_T \sim N(0, \sigma_T)$ and $\epsilon_E \sim N(0, \sigma_E)$ represent uncorrelated errors. More specifically, ϵ_T , ϵ_E , and U are all independent of each other, because all dependence between T and E are due to bidirectional causal effects and the confounder U , the residual noises are independent of each other and of the confounder.

For simplicity, we assume that E , T , and U have zero mean and unit variance, so that the correlation between E and T can be expressed as

$$corr(E, T) = cov(E, T) = E(E * T) = \alpha_{TWMR} + \alpha_{revTWMR} - \alpha_{TWMR} * \alpha_{revTWMR} * E(E * T) + q_T * q_E \tag{10}$$

Equivalently,

$$corr(E, T) = \frac{\alpha_{TWMR} + \alpha_{revTWMR} + q_T * q_E}{1 + \alpha_{TWMR} * \alpha_{revTWMR}} \tag{11}$$

As we know the correlation, the bidirectional causal effects estimated by TWMR and revTWMR, we can estimate the contribution of the confounders ($q_T * q_E$) to the observed correlation. Since the magnitude of $\alpha_{TWMR} * \alpha_{revTWMR}$ is negligible, we replaced the denominator with 1.

To avoid the recursive equations expressing the forward and reverse causal effects of E on T , we can substitute T into the equation for E and obtain

$$\begin{aligned} E &= \alpha_{revTWMR} * (\alpha_{TWMR} * E + q_T * U + \epsilon_T) + q_E * U + \epsilon_E \\ E &= \alpha_{revTWMR} * \alpha_{TWMR} * E + (\alpha_{revTWMR} * q_T + q_E) * U + \alpha_{revTWMR} * \epsilon_T + \epsilon_E \\ (1 - \alpha_{revTWMR} * \alpha_{TWMR}) * E &= (\alpha_{revTWMR} * q_T + q_E) * U + \alpha_{revTWMR} * \epsilon_T + \epsilon_E \\ E &= \frac{(\alpha_{revTWMR} * q_T + q_E) * U + \alpha_{revTWMR} * \epsilon_T + \epsilon_E}{1 - \alpha_{revTWMR} * \alpha_{TWMR}} \end{aligned} \tag{12}$$

Similarly for T

$$\begin{aligned} T &= \alpha_{TWMR} * (\alpha_{revTWMR} * T + q_E * U + \epsilon_E) + q_T * U + \epsilon_T \\ T &= \alpha_{TWMR} * \alpha_{revTWMR} * T + (\alpha_{TWMR} * q_E + q_T) * U + \alpha_{TWMR} * \epsilon_E + \epsilon_T \\ (1 - \alpha_{TWMR} * \alpha_{revTWMR}) * T &= (\alpha_{TWMR} * q_E + q_T) * U + \alpha_{TWMR} * \epsilon_E + \epsilon_T \\ T &= \frac{(\alpha_{TWMR} * q_E + q_T) * U + \alpha_{TWMR} * \epsilon_E + \epsilon_T}{1 - \alpha_{TWMR} * \alpha_{revTWMR}} \end{aligned} \tag{13}$$

GWAS hits trans-eQTL mapping in GTEx. Genotypes and gene expression quantifications from the GTEx project v8 dataset⁴⁷ were obtained via dbGaP accession number phs000424.v8.p1. This includes genotypes of 838 subjects, 85.3% of European American origin, 12.3% African American, and 1.4% Asian American. The phased version of the genotype files was used and the genotypes for 1078 out of 1093 GWAS hits used as instrument variables in revTWMR were retrieved, matching for chromosome, position, and reference/alternative allele, after conversion to GRCh38 coordinates using the UCSC liftOver tool⁶⁹. Gene expression quantification (TPM values) from RNA-seq experiments across 49 tissues (for which genotype data is also available for ≥70 individuals) processed and provided by the GTEx project v8 were also downloaded. These gene expression quantifications had been mapped to Gencode v26⁷⁰ gene annotations on GRCh38 and normalized by TMM between samples (as implemented in edgeR), and inverse normal transform across samples. Moreover, only genes passing an expression threshold of >0.1 TPM in ≥20% samples and ≥6 reads in ≥20% samples had been retained. The association between each of the 2177 GWAS hits genotyped in GTEx v8 and each gene expression (20,315 to 35,007 genes per tissue, all gene types) across 49 tissues of the GTEx v8 was computed using QTLtools v1.3.1 trans function⁷¹. This consists of more than 2 billion association tests performed. For this, the–nominal option for calculating nominal p values was used, as well as the–normal option, to enforce the gene expression phenotypes to match normal distributions $N(0,1)$. To include all associations, no *cis* window filtering was applied. Moreover, covariates provided by GTEx v8 for each tissue were regressed out of each expression matrix to account for potential confounding factors, by using the–covariate option on QTLtools. These included 15 to 60 PEER factors (depending on tissue sample size)⁷², five genotype PCA PCs as well as information

about the sequencing platform, PCR usage, and the sex of the samples provided by GTEx v8.

Analysis of mouse data. We used blood triglyceride and liver gene expression in a panel of BXD mice that were fed a chow or high-fat diet (CD and HFD⁴⁸). The study involves 52 strains, with five mice per strain per condition. Male mice were switched to an HFD diet at 8 weeks of age, subjected to extensive cardiometabolic phenotyping, and finally sacrificed at 29 weeks of age after an overnight fast. The blood and liver collection were performed simultaneously during tissue collection. Microarray data and triglyceride measurements in the two diets are available for a subset of 34 strains. Phenotype data, including blood triglyceride measurement, are deposited in the Mouse Phenome Database (<https://phenome.jax.org/projects/Auwerx1>) and the raw gene expression data in the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60149>). To calculate diet-induced TG change, we subtracted the strain average on CD from that on HFD and converted the resulting difference into a z-score. We normalized the Affymetrix Mouse Gene 1.0 ST Array data using the Affymetrix Power Tools software version 1.20.5 with GC correction (GCCN) and space transformation (SST). We removed the lowest quartile of genes based on average expression in all samples. For each BXD strain, we calculated strain-level HFD-induced fold change as the difference in the expression on HFD minus that of CD and then converted these values to z-scores. We then performed Spearman's correlation for each gene's HFD-induced fold change and the TG diet-induced differences. We used the biomaRt R package version 2.42.1⁷³ to convert between mouse and human gene Ensembl IDs.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The *trans*-eQTLs data used in this study are available on the eQTLGen Consortium website [<https://www.eqtlgen.org/trans-eqtl.html>]. The GWAS data are available in the NIH Genome-wide Repository of Associations Between SNPs and Phenotypes [<https://grasp.nhlbi.nih.gov/>] and in the Neal Lab website [<http://www.nealelab.is/uk-biobank/>]. For the mouse data, phenotype data, including blood triglyceride measurement, are deposited in the Mouse Phenome Database [<https://phenome.jax.org/projects/Auwerx1>] and the raw gene expression data in the Gene Expression Omnibus [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60149>]. Source data are provided with this paper.

Code availability

R-code for performing revTWMR analyses is available at <https://github.com/eleporcu/revTWMR> <https://doi.org/10.5281/zenodo.5119244>.

Received: 7 May 2021; Accepted: 24 August 2021;

Published online: 24 September 2021

References

- Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
- Gandal, M. J., et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat812 (2018).
- Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
- Uhlen, M. et al. A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017).
- MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
- Brem, R. B. et al. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
- Fehrmann, R. S. et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
- Hernandez, D. G. et al. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol. Dis.* **47**, 20–28 (2012).
- Nica, A. C. et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
- Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
- Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
- Mancuso, N. et al. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
- Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
- Vosa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *Nat. Genet.* **53**, 1300–1310 (2021).
- Porcu, E. et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300 (2019).
- Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Epidemiol.* **37**, 658–665 (2013).
- Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
- Manning, A. K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
- Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
- Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
- Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
- Shungin, D. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
- UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
- Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
- Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
- Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
- Jenkins, R. N. et al. T cell receptor V beta gene bias in rheumatoid arthritis. *J. Clin. Invest.* **92**, 2688–2701 (1993).
- Williams, P. T. et al. The associations of high-density lipoprotein subclasses with insulin and glucose levels, physical activity, resting heart rate, and regional adiposity in men with coronary artery disease: the Stanford Coronary Risk Intervention Project baseline survey. *Metabolism* **44**, 106–114 (1995).
- Tansey, T. R. & Shechter, I. Structure and regulation of mammalian squalene synthase. *Biochim. Biophys. Acta* **1529**, 49–62 (2000).
- Peltola, P. et al. Visceral obesity is associated with high levels of serum squalene. *Obesity* **14**, 1155–1163 (2006).
- Zelcer, N. et al. LXR regulates cholesterol uptake through idl-dependent ubiquitination of the LDL receptor. *Science* **325**, 100–104 (2009).
- Yvan-Charvet, L., Wang, N. & Tall, A. R. Role of HDL, ABCA1, and ABCG1 transporters in cholesterol efflux and immune responses. *Arterioscler. Thromb. Vasc. Biol.* **30**, 139–143 (2010).
- Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
- Afonso, M. S. et al. Molecular pathways underlying cholesterol homeostasis. *Nutrients* **10**, 760 (2018).
- Dockham, P. A., Lee, M. O. & Sladek, N. E. Identification of human liver aldehyde dehydrogenases that catalyze the oxidation of aldophosphamide and retinaldehyde. *Biochem. Pharmacol.* **43**, 2453–2469 (1992).
- Schwarz, E. J. et al. Retinoic acid blocks adipogenesis by inhibiting C/EBP beta-mediated transcription. *Mol. Cell. Biol.* **17**, 1552–1561 (1997).
- Ziouzenkova, O. et al. Retinaldehyde represses adipogenesis and diet-induced obesity. *Nat. Med.* **13**, 695–702 (2007).
- Kiefer, F. W. et al. Retinaldehyde dehydrogenase 1 regulates a thermogenic program in white adipose tissue. *Nat. Med.* **18**, 918–925 (2012).
- Lamparter, D. et al. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.* **12**, e1004714 (2016).
- Bowden, J. et al. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
- Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985–1998 (2017).

42. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
43. Verbanck, M. et al. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
44. Burgess, S. et al. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* **28**, 30–42 (2017).
45. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
46. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
47. Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
48. Williams, E. G. et al. Systems proteomics of liver mitochondria function. *Science* **352**, aad0189 (2016).
49. Holmes, M. V. & Davey Smith, G. Can Mendelian randomization shift into reverse gear? *Clin. Chem.* **65**, 363–366 (2019).
50. Gordon, T. et al. High density lipoprotein as a protective factor against coronary heart disease. The Framingham Study. *Am. J. Med.* **62**, 707–714 (1977).
51. Hokanson, J. E. & Austin, M. A. Plasma triglyceride level is a risk factor for cardiovascular disease independent of high-density lipoprotein cholesterol level: a meta-analysis of population-based prospective studies. *J. Cardiovasc. Risk* **3**, 213–219 (1996).
52. Jump, D. B. & Clarke, S. D. Regulation of gene expression by dietary fat. *Annu. Rev. Nutr.* **19**, 63–90 (1999).
53. Mortlock, S. et al. Tissue specific regulation of transcription in endometrium and association with disease. *Hum. Reprod.* **35**, 377–393 (2020).
54. Morrison, J. et al. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.* **52**, 740–747 (2020).
55. Darrous, L., Mounier, N. & Kutalik, Z. Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. Preprint at *medRxiv* 01.27.20018929 (2020).
56. Hu, X. et al. MR-APSS: a unified approach to Mendelian Randomization accounting for pleiotropy and sample structure using genome-wide summary statistics. Preprint at *bioRxiv* 03.11.434915 (2021).
57. Wahl, S. et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
58. Lepik, K. et al. C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLoS Comput. Biol.* **13**, e1005766 (2017).
59. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
60. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
61. Westra, H. J. et al. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**, 2104–2111 (2011).
62. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
63. Ferrucci, L. et al. Subsystems contributing to the decline in ability to walk: Bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *J. Am. Geriatrics Soc.* **48**, 1618–1625 (2000).
64. Gibbs, J. R. et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e100095 (2010).
65. Wood, A. R. et al. Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum. Mol. Genet.* **20**, 4082–4092 (2011).
66. Volzke, H. et al. Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).
67. Schurmann, C. et al. Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PLoS ONE* **7**, e50938 (2012).
68. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
69. Hinrichs, A. S. et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
70. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
71. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
72. Stegle, O. et al. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
73. Durinck, S. et al. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

Acknowledgements

This work was supported by grants from the Swiss National Science Foundation (310030-189147, 32473B-166450, 32003B-173092 to Z.K. and 31003A_182632 to A.R.) and Horizon2020 Twinning projects (ePerMed 692145 to A.R.). SHIP-Trend is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network “Greifswald Approach to Individualized Medicine (GANI_MED)” funded by the Federal Ministry of Education and Research (grant 03IS2061A). The University of Greifswald is a member of the Caché Campus program of the InterSystems GmbH. Research on BXD mice was supported by the Ecole Polytechnique Fédérale de Lausanne (EPFL), European Research Council (ERCAAdG-787702), and Swiss National Science Foundation (SNSF 310030B160318). We would like to thank Liza Darrous and Ninon Mounier for their valuable feedback and comments on this manuscript.

Author contributions

E.P. and Z.K. conceived and designed the study; E.P. and Z.K. contributed to the mathematical derivations of the research; E.P. performed statistical analyses; M.C.S. carried out drug target analyses; Z.K. supervised drug target analyses; K.L. has performed initial comparisons between gene expression-trait correlation and TWMR effects in the EGCUT cohort; C.A. and F.A.S. contributed with the biological interpretation of the results; E.P., A.R. and Z.K. drafted the manuscript; M.C.S. and C.A. contributed to the writing of specific sections; C.A. and F.A.S. revised the manuscript; K.L. performed statistical analyses on EGCUT cohort; A.M. oversaw the analysis in EGCUT; A.R.W. performed statistical analyses on InChianti cohort; S.B., T.T. and T.F. oversaw the analysis in InChianti; A.W. performed statistical analyses on SHIP-Trend cohort; U.V. and M.N. contributed to the data collection, quality control, and study design of SHIP-Trend; A.T. oversaw the analysis in SHIP-Trend; M.S.B.S. performed the analysis in mice data; D.M.R. performed *trans*-eQTLs analyses in GTEx dataset; O.D. designed and supervised *trans*-eQTLs analyses on GTEx dataset. All authors read the paper and contributed to its final form.

Competing interests

The authors declare no competing interests.

Additional information


Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-25805-y>.

Correspondence and requests for materials should be addressed to Eleonora Porcu.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contributions to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

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

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021