Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits

Eleonora Porcu[1,2,*], Sina Rüeger[2,3], eQTLGen Consortium, Federico A. Santoni[4], Alexandre Reymond[1,5], Zoltán Kutalik[2,3,5,*]

[1] Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, Lausanne Switzerland
[3] Institute of Social and Preventive Medicine, CHUV and University of Lausanne, Switzerland
[4] Endocrine, Diabetes, and Metabolism Service, CHUV and University of Lausanne, Switzerland
[5] These authors jointly supervised this work: Alexandre Reymond and Zoltán Kutalik

[*] Corresponding authors (eleonora.porcu@unil.ch or zoltan.kutalik@gmail.com)

Abstract

Genome-wide association studies (GWAS) identified thousands of variants associated with complex traits, but their biological interpretation often remains unclear. Most of these variants overlap with expression QTLs (eQTLs), indicating their potential involvement in the regulation of gene expression.
Here, we propose an advanced summary statistics-based Mendelian Randomization approach that uses multiple SNPs jointly as instruments and multiple gene expression traits as exposures, simultaneously.
When applied to 43 human phenotypes it uncovered 2,277 putative genes whose blood expression is causally associated with at least one phenotype resulting in 5,009 gene-trait associations; of note, 55% of them had no genome-wide significant SNP nearby in previous GWAS analysis. Using independent association summary statistics (UKBiobank), we confirmed that the majority of these loci were missed by conventional GWAS due to power issues. Noteworthy among these novel links are height- and intelligence-associated *PEX19* and *CDC42*, respectively known to carry mutations leading to short stature and Takenouchi-Kosaki syndrome. We similarly unraveled novel pleiotropic causal effects suggestive of mechanistic connections, e.g. the shared genetic effects of *TSPAN14* in rheumatoid arthritis, Crohn's and inflammatory bowel disease. Finally, we show that causal genes can be highly tissue-specific.
Our advanced Mendelian Randomization unlocks hidden value from published GWAS through higher power in detecting associations. It better accounts for pleiotropy and unravels new biological mechanisms underlying complex and clinical traits.

Introduction

Genome-wide association studies (GWAS) have identified tens of thousands of common genetic variants associated with hundreds of complex traits [1]. The identification of causal genes using GWAS results is difficult, however, because this

approach only highlights fine-mapped intervals of associated variants in linkage disequilibrium (LD) with the causal marker [2-4]. Linking the effect of those to gene function is not straightforward without additional data, especially, since the majority of these trait-associated variants fall into non-coding regions of the genome [1] with no direct influence on protein structure or function.
Recently, it has been shown that trait-associated SNPs are three times more likely to be associated with gene expression, i.e. expression quantitative trait loci (eQTLs) [5-8], therefore, understanding the biological mechanism behind SNP-trait associations requires a focus on investigating their potential role in gene expression regulation.
Transcriptome-wide association studies (TWAS) integrating GWAS and eQTLs data have been proposed to unravel gene-trait associations [7, 9, 10]. However, while these studies aim to identify genes whose (genetically determined) expression is significantly associated to 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 influences multiple phenotypes).

Here, we integrate summary-level data from GWAS and eQTLs studies in a Mendelian Randomization (MR) framework to estimate such causal effect of gene expression on several human phenotypes. Of note, through the manuscript the term "causal" is used for simplicity to mean "found to be causal by the MR method". A recent work [11] implementing a MR approach based on single genetic variants was not designed to distinguish causality from pleiotropy. However, a more recent work from the same authors [12] proposed an approach performing a MR with multiple SNPs simultaneously. Since trait-associated variants are often eQTLs for multiple genes [13], we introduce here a multiple-instrument, multiple-exposure MR approach that estimates joint causal effects of gene expression levels on outcome traits. Our method requires summary-level data only (along with pair-wise SNP LD estimated from e.g. the 1000 Genomes [14] or the UK10K reference panel [15]) allowing data integration from different studies.
We apply our method to the largest publicly available GWAS summary statistics (based on sample sizes ranging from 20,883 to 339,224 individuals) and combine them with eQTL data from GTEx (Genotype Tissue Expression Project)[16] and the eQTLGen Consortium (n=14,115, unpublished) to provide an atlas of putative functionally relevant genes for 43 complex human traits.

Results

Overview of the approach

Mendelian Randomization 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. Violation of any of these assumptions often through pleiotropy would lead to biased estimates of the causal effect and potential false positives [17]. The second assumption is the most difficult to verify because confounding factors are often unknown, whereas the last assumption can only be partially verified at best if SNPs, exposures and outcome data are available from the same (large) sample, a condition not fulfilled for two-sample MR.

Recently, Zhu et al [11] applied a summary data-based MR (SMR) approach to test whether the effects of genetic variants on a phenotype are mediated by gene expression. Since their original method, based on a single genetic variant, was particularly vulnerable to pleiotropy, they extended it to a more general form (GSMR: generalized SMR) [12] that performs a MR analysis with multiple independent SNPs as instrumental variables. They designed a heterogeneity test (HEIDI) to identify pleiotropic SNPs, which could bias the causal estimate. We noticed that many pleiotropic effects are simply mediated through the expression levels of neighbouring genes; hence including other genes as exposures should reduce MR assumption violation, while improving power. Furthermore, such approach might be able to better distinguish the causal effects of genes with correlated expression levels. For this reason, we propose a multi-gene approach (**Fig. 1a**) that should specifically reduce the bias due to pleiotropic effects.
For a set of k genes, using inverse-variance weighted method for summary statistics [18, 19], we estimated the multivariate causal effect of the expression levels for several genes at a locus on an outcome trait as

$$\hat{\alpha}=(\mathbf{E}'\mathbf{C}^{-1}\mathbf{E})^{-1}(\mathbf{E}'\mathbf{C}^{-1}\mathbf{G}),$$

where **E** is a $n\times k$ matrix that contains the univariate effect size of n SNPs on k gene expressions (these estimates come from an eQTL study); **G** is a vector of length n that contains the univariate effect sizes of the same n SNPs on the phenotype (these estimates come from publicly available GWAS summary statistics) and **C** is the pair-wise correlation (LD) matrix between n SNPs (estimated from the UK10K panel).

Applying multi-instrument, multi-exposure Mendelian Randomization to GWAS and eQTL summary statistics

We applied our multiple-instrument, multiple-exposure MR approach to summary data from an eQTL meta-analyses performed in blood samples from >14K individuals (eQTLGen Consortium, unpublished data) and the largest publicly available GWAS data to assess causal associations between gene expression and 43 complex traits.

Note that the aim of our paper is not to confirm the validity of a well-established method in simulation studies, because such MR approaches have been extensively tested in various simulation setups [18, 20]. Our focus is rather to demonstrate the utility of such approaches when applied to real data.

Data from eQTLGen Consortium contain association summary statistics for 9,136,405 SNPs and 20,468 genes. In our analyses we included only cis-eQTLs at $P_{eQTL}<1\text{x}10^{-03}$ resulting in 3,520,820 eQTLs for 18,732 egenes (i.e. genes whose expression levels are associated with at least one genetic variant). We only used cis-eQTLs data since trans-eQTLs typically have weaker effect size and less direct effect, hence are more prone to violate MR assumptions. In total, we found 2,277 genes putatively causally associated with at least one phenotype giving rise to 5,009 gene-trait associations ($P_{MR}<5\text{x}10^{-07}$ = 0.05/(2306*43), where 2,306 and 43 correspond to the number of effective genes in terms of expression level across the genome and the number of

traits analyzed in our study, respectively)(**Supplementary Fig. 1** and **Supplementary Table 1 and 2**).

Pleiotropic SNPs lead to biased causal effect estimates

Using our approach we show how single-gene analyses can lead to biased causal effect estimates (**Supplementary Fig. 2**). Still, genetic variants can influence the outcome through other (not expression related) risk factors. Thus, we applied a heterogeneity test (see Methods) to detect pleiotropic SNPs, i.e. genetic variants whose effect on the outcome significantly differs from the expected effect estimated using their effect on the exposures and the estimated causal effects of the exposures on the outcome. Overall, we detected heterogeneity ($P_{HET}<1x10^{-04}$) for 2,072 of the 6,137 originally significant gene-trait associations. Out of these 2,072 associations, 822 passed the heterogeneity test after removing SNPs showing pleiotropic effects. Furthermore, removing pleiotropic SNPs led to the identification of 122 additional associations, giving the final number of 5,009 robust associations.

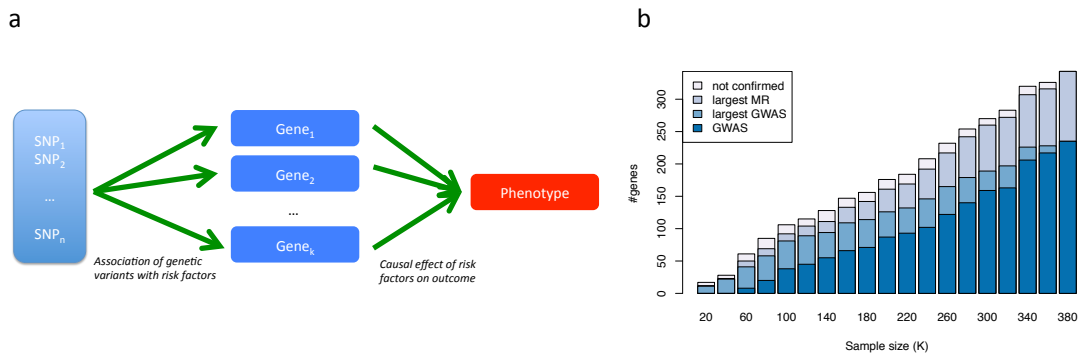

Fig. 1 a. Schematic representation of multi-instrument, multi-exposure MR to estimate the causal effect of multiple exposures (expression of genes) on a phenotype using multiple instrumental variables (SNPs). **b.** Validation of genes found by MR. Bars represent the number of genes found to be causal for BMI when carrying out the MR analysis in different sized subsets of the UKBB. Among them, those confirmed by GWAS in the same sample (i.e. that fall within 500kb of a GW significant SNP) are marked in dark blue; those confirmed only when running GWAS in the full 380K UKBB samples in light blue; those confirmed only running MR in the full UKBB dataset in grey; whereas the light grey bars on the top represent the number of genes found by MR, but not confirmed in the full UKBB dataset.

MR improves GWAS power to detect associated loci

Our method, incorporating eQTL information into GWAS analyses, has the potential to increase the power of GWAS in identifying loci associated with complex traits.
Conventional gene-based tests, usually based on physical distance or LD rarely lead to newly discovered loci [21, 22]. With such approaches, identifying new loci missed by GWAS is very difficult because a limited set of (independent) associated variants in/near a gene is diluted by the large number of null SNPs. Hence the combined

association signal for genes is typically weaker than the strongest SNP in the region. Furthermore, given that the of GWAS SNPs falls outside coding regions, they may be excluded from gene-centric analyses.

To evaluate the performance of our approach we conducted GWAS and MR analyses for BMI on several subsets of individuals from UKBiobank (UKBB). Using the full sample, we found 343 significant MR genes associated with BMI **(Fig. 1b)**. Among these, 108 are more than 500kb away from any GW significant SNP, hence potentially representing new loci that were missed by the conventional GWAS due to power issues. To assess if the additional genes implicate regions harbouring truly associated SNPs, we carried out the MR and GWAS analysis for 18 subsets of the UKBB with increasing sample size (from 20,000 to 360,000). We observed that most of MR identified genes in the small subsets are confirmed i.e. fall within 500kb vicinity of lead SNPs identified as GW significant in the GWAS using the full sample. For example, 68 genes found by MR and not found by GWAS (i.e. positioned >500kb away from any GW significant SNP) in the subset of 100K individuals overlap significantly (OR = 20, P=9.65x10^{-26}) with the 354 genes mapping within 500kb of the SNPs identified by the GWAS performed using the full dataset. We observed that the portion of loci missed by GWAS decreases with increasing sample size indicating a saturation effect. To support the consistency of our findings, >74% of MR genes identified in any of the 18 subsets of the full sample and missed by GWAS in the full dataset show significant association in the MR analyses performed in the full dataset.

New trait-associated genes

In total we found 2,277 putative genes causally associated with at least one phenotype (5,009 gene-trait associations at P_{MR}<5x10^{-07}) (**Supplementary Fig. 1, Supplementary Table 1 and 2**). Of these gene-trait associations, more than half (2,736) have been missed by previous GWASs, as no SNP reached genome-wide significance level within the gene +/- 500kb (**Supplementary Fig. 3**). For example we detected a causal effect of *CDC42* on intelligence (**Fig. 2a**). This is the only gene showing a significant causal effect in that region and its positive link with intelligence is consistent with the already known involvement of this gene in Takenouchi-Kosaki syndrome (OMIM:#616737)[23], which includes among its symptoms intellectual disability. Additionally, among other regions missed by conventional GWAS, we show that the association of *ALS2* (P_{MR}=1.58x10^{-07}) (**Fig. 2b**) and *ZNF565* (P_{MR}=2.65x10^{-07}) with amyotrophic lateral sclerosis [24] and schizophrenia [25], respectively, is not restricted to rare coding variants.

To test whether our putative causal genes are functionally relevant, we overlapped the genes significantly associated with height with a list of genes involved in abnormal skeletal growth syndrome [26]. We observed a trend for enrichment of 1.3-fold (P>0.05) suggesting that our prioritized genes have additional supporting evidence. For example, within the *VANGL2* locus, which harbours height-associated SNPs identified by Wood et al [26] (topSNP: rs6688100 P=2.30x10^{-08}), our results suggest that *VANGL2* is not causal (P_{MR}=0.72) (**Fig. 2c**) and reveal that high expression of *NCSTN* (P_{MR}=2.30x10^{-18}), *COPA* (P_{MR}=4.48x10^{-11}) and *PEX19* (P_{MR}=8.02x10^{-11}) is causally linked to high stature. While it is unclear which of the three significant MR genes is the most functionally relevant in this region, mutations

in *PEX19* are known to be associated with short stature (Zellweger Syndrome, OMIM:#614886)[27], making it the strongest candidate.

We also found 2,626 regions where MR identified only one putative causal gene. In one region for height, our analysis pointed exclusively to *DYM* (P_{MR}=3.42x10^{-44}), whose high expression is associated with high stature. *DYM* is already known to be associated with Smith-McCort dysplasia (OMIM:#607326) and short stature in particular [28]. In another example, for rheumatoid arthritis, *CTLA4* emerged as the only causal gene at its locus: its low expression is significantly associated with the risk of rheumatoid arthritis in the general population (P_{MR}=3.04x10^{-22}). *CTLA4* was associated with autoimmune lymphoproliferative syndrome (OMIM:#616100) [29] presenting autoimmune arthritis among its symptoms.

With the assumption that a coding variant is likely the causal variant in a given locus, we tested the 70 genes harbouring at least one coding variant associated with height [30] for causal association in our MR results. Only four out of the fraction of 44 testable genes showed significant effect on height, suggesting that when the association is driven by coding variant and not by gene expression, our MR approach correctly does not point to any causal gene.

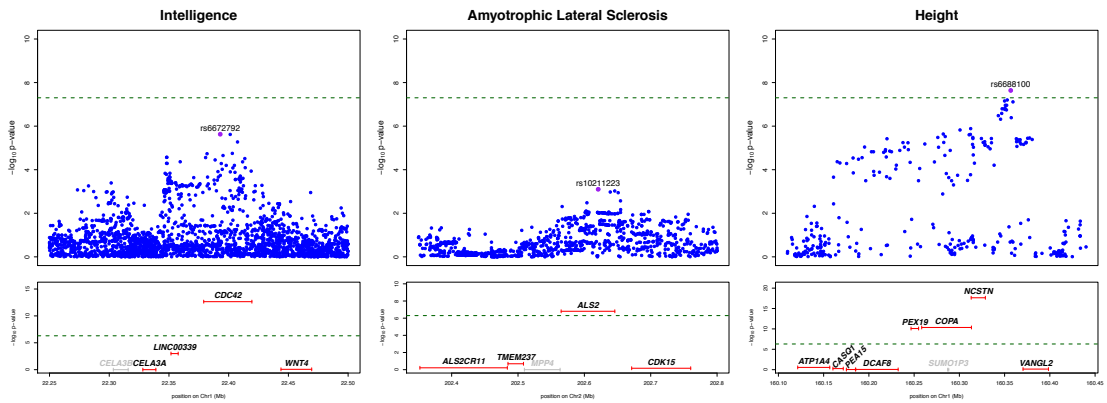

Fig. 2 Regional association plot for GWAS and MR analyses. Top panels show single-SNP association strengths (y axis shows the –log$_{10}$(p-value)) versus genomic position (on hg19/GRCh37 genomic build; x axis) around the most significant SNP, indicated with a purple dot. Lower panels illustrate gene causal association strengths (y axis shows the –log$_{10}$(p-value)) versus genomic position. Genes highlighted in grey were not tested (i.e.: they do not pass the heterogeneity test or not have significant eQTLs). Green dotted lines represent GWAS and MR significance thresholds. Note that for intelligence and amyotrophic lateral sclerosis the most significant SNP did not reach GW significance (top), whereas MR pinpoints to putative causal genes (bottom).

Closest genes are often not causal

Recent studies suggested that the gene closest to the GWAS top hits is often not the causal one [11, 31, 32]. Consistent with these findings, among the 1,423 MR significant regions harbouring at least one genome-wide significant SNP, we found that 63% of the genes closest to the top SNP in the region do not show any significant association with the phenotype (**Supplementary Fig. 4**). One of the numerous such examples is the significant causal association between educational attainment and *ERCC8* (P_{MR}=4.82x10^{-07}), a gene previously linked with the monogenic Cockayne

Syndrome A (OMIM#216400) [33] in the *ELOVL7* region (topSNP: rs61160187, P=5.93x10^{-13}).

Genes with pleiotropic effect

Genes often play a role in numerous and independent biological processes leading to different outcomes. We thus investigated the degree of pleiotropy and identified 1,870 genes associated with multiple phenotypes (**Supplementary Fig. 1**). Among them, *XKR6* and *BLK* turned out to be the most pleiotropic, showing causal association with >20 phenotypes including autoimmune diseases, neurological, metabolic, hematopoietic and anthropometric traits. Consistent with this result, they were "identified" in many GWASs [34-41]. One such pleiotropic gene is *TSPAN14*, which displays strong associations with Crohn's disease (CD)(P_{MR}=6.93x10^{-09}), inflammatory bowel disease (IBD)(P_{MR}=2.78x10^{-07}), rheumatoid arthritis (RA)(P_{MR}=8.92x10^{-10}), monocytes (P_{MR}=2.04x10^{-08}) and reticulocytes (P_{MR}=2.80x10^{-09}). This result pinpoints shared effects across CD, IBD and RA (**Fig. 3**). Of note, we observed significant causal associations despite the fact that this locus was missed by the GWASs we used in our MR analysis for the three diseases [42] (CD: N=5,956 cases / 14,927 controls; IBD: N=12,882/21,770; RA: N=14,361/43,923). Supporting our findings, the locus was later confirmed by a larger meta-analyses that combined GWAS data with ImmunoChip data [42] (CD: N=20,550/41,642; IBD: N=38,155/48,485), leading to GW significant association found for CD and IBD (**Supplementary Fig. 5**).
We also found 19 genes showing significant causal effects on BMI and age at menarche confirming the common genetic basis of these traits [43]. 16 of these genes have opposite effects as expected by the negative genetic and phenotypic correlations between these traits [44]. Within these 16 genes, two (*CADM1* and *ZC3H4*) were also detected by previous GWAS for both traits [45, 46], whereas seven (*BNIPL*, *C17orf89*, *GGT7*, *MAPK3*, *PM20D1*, *PPM1F* and *PPP4C*) and five (*APOBR*, *C1QTNF4*, *GTF3A*, *NUP88* and *SULT1A1*) were known to be exclusively associated with age at menarche [46] or BMI [45] , respectively. Interestingly, we identified two genes (*DCAF12* and *DNAJA4*) as new putative causal genes for both traits.

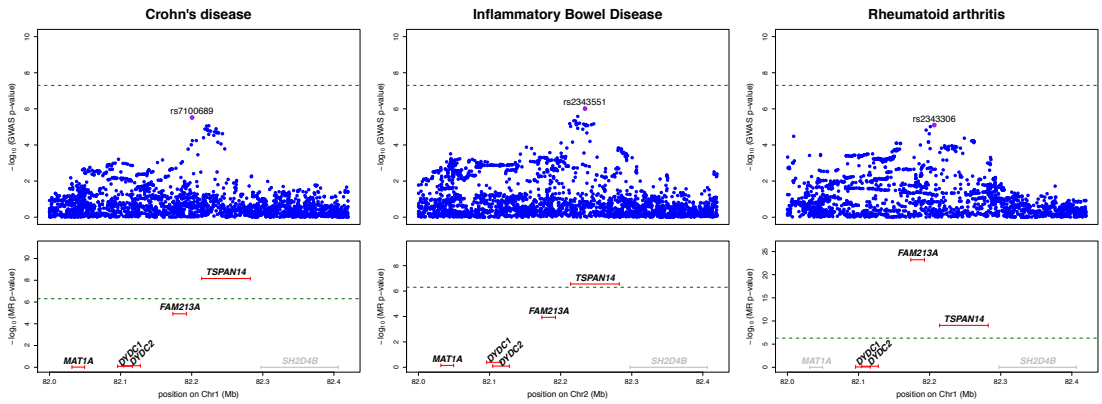

Fig. 3 Regional association plot for GWAS and MR analyses for *TSPAN14* locus associated with Crohn's disease, inflammatory bowel disease and rheumatoid arthritis. Top panels show single-SNP association strengths (y axis shows the –log$_{10}$(p-value)) versus genomic position (on hg19/GRCh37 genomic build; x axis) around the most significant SNP, indicated with a purple dot. Lower panels illustrate gene causal association strengths (y axis shows the –log$_{10}$(p-value)) versus genomic position. Genes highlighted in grey were not tested (i.e.: they do not

pass the heterogeneity test or not have significant eQTLs). Green dotted lines represent GWAS and MR significance thresholds.

Trait correlation

To evaluate the shared causal effects of gene expression on each pair of traits we computed the correlation ($\hat{\rho}_E$) between the causal effect estimates of the gene expression (or equivalently the Z-scores from our MR analysis) across a subset of 2,203 independent genes (including those that were not significant for any trait; see Methods). Among the 903 pairs of traits, we found several significant correlations in line with previous epidemiological observations (**Fig. 4a**). For example, for age at menarche we observed negative ($\hat{\rho}_E$=-0.17 P_{FDR}=2.25x10^{-15}) and positive correlations ($\hat{\rho}_E$=0.11, P_{FDR}=1.04x10^{-06}) with BMI and height, respectively [47, 48]. As expected, we observed a negative correlation ($\hat{\rho}_E$=-0.10, P_{FDR}=1.26x10^{-05}) between coronary artery disease and educational attainment [49]. Out of the 43 traits included in our analyses, 17 were included in a previous study reporting genetic correlation between traits estimated by LD score regression ($\hat{\rho}_G$) [50]. Comparing the expression- ($\hat{\rho}_E$) with genetic correlation ($\hat{\rho}_G$) estimates for 136 common pairs of traits, we found a remarkable concordance between the two estimates (r=0.84). Of note, the expression correlation seems to be 44% of the genetic correlation on average. While the genetic correlation estimate having smaller variance may explain part of this attenuation, we think that the main reason is that about half of the observed genetic correlation propagates to gene-expression level in whole blood (**Fig. 4b**). In particular we observed 30 pairs of traits showing significance for $\hat{\rho}_E$ and $\hat{\rho}_G$, whereas seven were significant only for $\hat{\rho}_E$, 12 only for $\hat{\rho}_G$ and 87 not significant for either. Among the significant correlations not identified by LD score regression we found positive correlation between schizophrenia and ulcerative colitis ($\hat{\rho}_E$=0.07, P_{FDR}=4.45x10^{-03}) in directional agreement with genetic correlation reported in a previous study [51] and supporting the molecular evidence for an autoimmune etiology for a fraction of schizophrenia cases. We also observed positive correlation between coronary artery disease and Crohn's disease ($\hat{\rho}_E$=0.08, P_{FDR}=8.09x10^{-04}), confirming the results of a recent study [52].

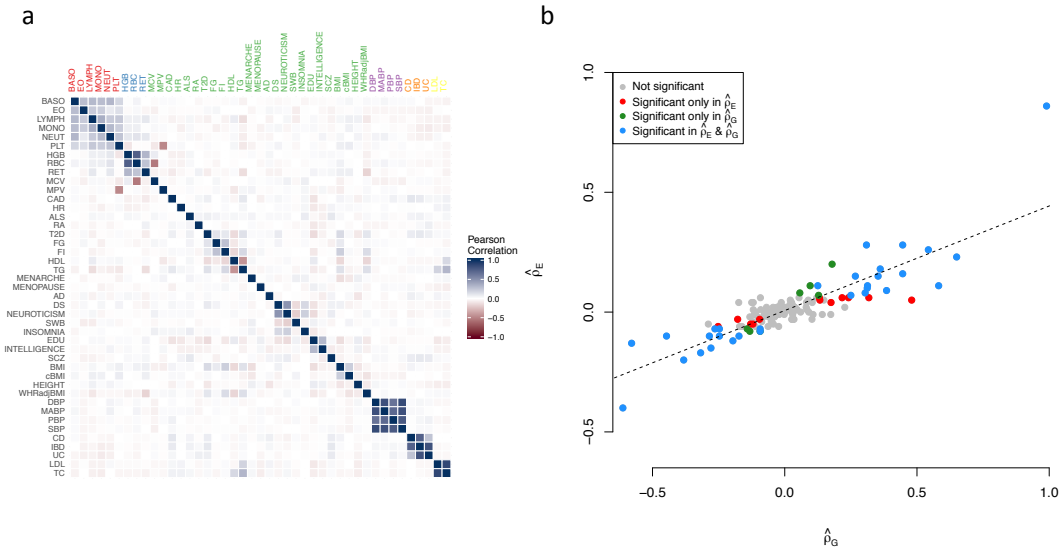

Fig. 4 a. Expression correlation from MR results among 43 traits. For the 2,203 independent genes and each pair of traits we calculated the Pearson's correlation ($\hat{\rho}_E$) between the Z-scores (proportional to the standardised causal effects). Darker colours represent higher correlations, with blue and red corresponding to positive and negative associations, respectively. **b.** Linear relationship between the expression correlation ($\hat{\rho}_E$) and the genetic correlation ($\hat{\rho}_G$) obtained from LD Score Regression. We selected the traits analysed by our study and Bulik-Sullivan *et al* [50] and for each pair of traits we compared the two correlations. Grey dots represent non-significant trait pairs, blue dots represent trait pairs significant for both correlation and red and green ones correspond to pairs of traits significant only in ($\hat{\rho}_G$) or ($\hat{\rho}_E$), respectively. The dotted line represents the regression line.

Tissue-specific effects

Since many traits manifest themselves only in certain tissues, it is important to integrate data from the tissue of interest for the studied phenotype when trying to interpret GWAS results using gene expression as an intermediate phenotype. For this reason we performed tissue-specific MR analyses using the eQTLs identified by GTEx (Genotype Tissue Expression Project) [16], which provides a unified view of genetic effects on gene expression across 48 human tissues. Despite the fact that sharing eQTLs (in consistent effect direction) across tissues is very common [53], there are many tissue-specific eQTLs.

For practical reasons we performed the tissue-specific analyses only for four phenotypes for which the key tissue is well known: CAD (artery), CD (intestine), LDL (liver) and T2D (pancreas). Among our results (**Supplementary Tables 3-6**) we found *MRAS* and *PHACTR1*, associated with coronary artery disease (CAD), showing significant association in arterial tissues, which contribute most to the genetic causality of this trait [54] (*MRAS:* P_{MR}=4.89x10^{-09} in artery coronary and P_{MR}=5.28x10^{-09} in artery tibial, *PHACTR1*: P_{MR}=4.66x10^{-38} in artery coronary, P_{MR}=1.02x10^{-36} in artery aorta and P_{MR}=3.39x10^{-30} in artery tibial)(**Fig. 5**). Interestingly, no gene showed a significant causal effect in the other tissues, including whole blood using the large dataset from eQTLGen Consortium. *MRAS* and *PHACTR1* have significant eQTLs in other tissues but none of these effects was associated with the diseases, confirming that the disease-relevant eQTLs were tissue-specific.

We also confirm *SORT1* being causal for LDL in line with previous findings [32]. Tissues that are not causal for LDL, like skin, pituitary and testis [54], incorrectly pointed to *PSRC1* as the most likely putative causal gene and only liver pointed to *SORT1* as strong candidate gene (**Supplementary Fig. 6 and Supplementary Table 5**). This result confirms, once again, the importance to identify the relevant tissue(s) for the studied phenotype before looking for the causal gene.

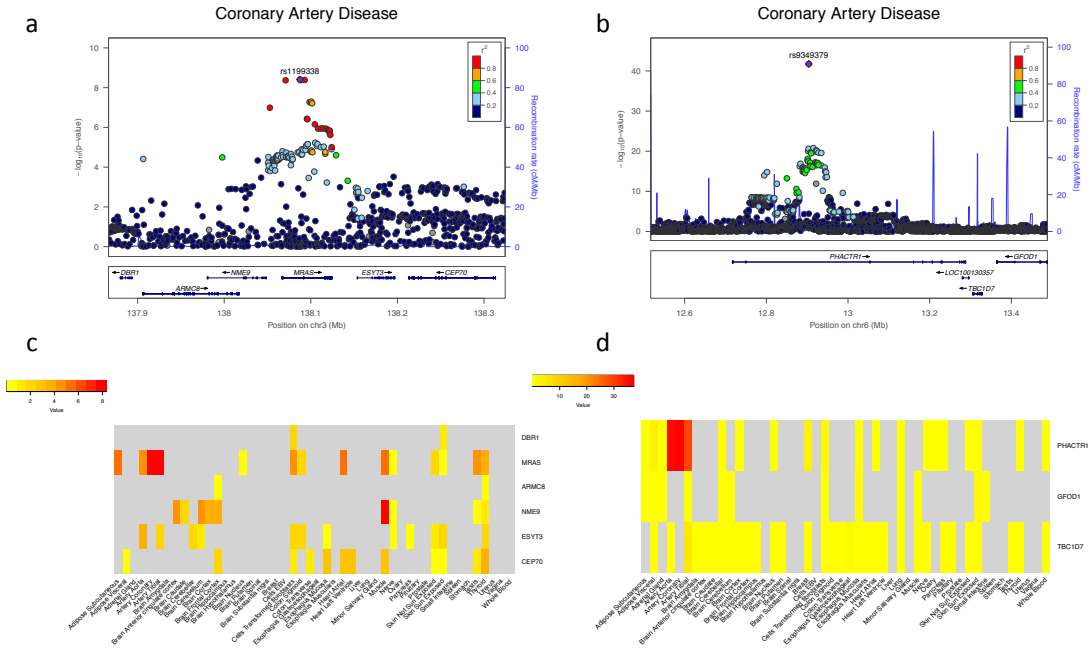

Fig. 5 Tissue-specific effects of *MRAS* and *PHACTR1* on coronary artery disease. **a-b.** Top. Association plot showing genome-wide significant loci for coronary artery disease in the *MRAS* (**a**) and *PHACTR1* (**b**) gene regions. Representation of single-SNP association strengths (y axis shows the –log10 p-value) versus genomic positions (on hg19/GRCh37 genomic build; x axis) around the most significant SNP, indicated with a purple dot. Other SNPs in the region are colour coded to reflect their LD with the top SNP (according to pairwise r^2 values from 1000 Genomes Project phase 3 haplotypes). Bottom. Genes and position of exons and respective transcriptional strands. **c-d.** Tissue-specific causal effects. Genes listed on y axis; tissues listed on x axis. Darker points correspond to stronger associations. Genes highlighted in grey were not tested. *MRAS* and *PHACTR1* show significant association only in arterial tissues.

Discussion

We presented a powerful approach to perform MR analysis with multiple instruments and multiple exposures to identify genes with expression causally associated with complex traits. By increasing the statistical power through the integration of gene-expression and GWAS data, this method enables the prioritization of genes in known or novel associated regions and the identification of loci missed by conventional GWAS. We showed its efficacy through extensive analysis performed in 43 phenotypes. Whereas we have space to present only selected interesting findings the readers will find the full results in the **Supplementary Tables 2-6**. Exploiting UKBB

data we showed that in most cases these new loci harbour true signals and will be eventually found by larger GWAS in the future.

Starting from our prioritized genes, we shed light on the shared genetic architecture of complex traits, and we estimated that about half of the genetic correlation is given by the gene-expression in whole blood, reinforcing the view that many risk variants affect complex traits through changes to gene regulation. Blood is, of course, not necessarily the causal tissue, more probably gene expression in blood may be an often good-enough proxy for the expression levels in more relevant tissues [53].

Like all methods, our approach has its limitations, which need to be considered when interpreting results. The putative causal associations reported in this study are not definitive. They provide a prioritized list of candidate genes for future follow-up studies and also shed light on possible biological mechanisms of complex traits. With our approach we can assess which trait-associated SNPs are likely to exert their effect through gene expression. For example, we interrogated the 697 GWAS-SNPs associated with height [26], and observed that 95 of them are likely to alter height through regulation of nearby gene expression in whole blood or in a tissue where the expression level is correlated. This fraction would certainly increase in the future as eQTL studies grow in both diversity and size. Increased sample size will be crucial to detect SNPs associated with the gene-expression, whereas the possibility to interrogate more tissues could unravel causative genes which expression is not well recapitulated in whole blood, the tissue in which the biggest human eQTL studies have been performed so far. Indeed, when conducting a tissue-specific analysis using GTEx data, we had considerably lower power to detect causal gene associations because for any given tissue, the sample size is >30 times lower than in the blood eQTL dataset, ranging from 80 (brain-substantia nigra) to 491 (muscle) individuals. This represents an insurmountable limitation upon trying to identify tissue-specific causal genes given the limited number of egenes shared between the tissues (**Supplementary Fig. 7**). In line with previous work [53] we observe that as long as a gene is expressed in blood, its eQTLs are fairly similar to its eQTLs in other tissues. We hypothesize that the tissue specificity of causal genes is mostly given whether the gene is expressed only in a given tissue. Moreover, our work provocatively suggests that large eQTL analyses such as GTEx with increased sample size and number of assessed tissues could pinpoint to causative genes more efficiently than increasing GWAS sampling.

Another shortcoming of our approach is that using current eQTL data only 15K egenes are testable, which substantially decreases power to detect enrichment of our prioritized gene-set in relevant pathways and regulatory networks. We expect that larger eQTLs studies will allow identifying additional egenes resulting in stronger enrichment when using causally associated gene-sets rather than selecting genes based on physical proximity to GWAS hits. Still, many genes lead to disease not through change in their gene expression, but via modification of the RNA or protein sequence, mechanisms our approach is blind to.

Further limitations of our study are the violations of the MR assumptions. In particular, horizontal pleiotropy and indirect effects of the instruments on the exposures can substantially bias causal effect estimates. Although *cis*-eQTLs are thought to have direct effect of gene expression, we made particular effort to protect our results from potential biases, such as using multiple instruments, testing for effect heterogeneity and including other gene expression traits as exposures through which potential pleiotropy may act. We have confidence that accounting for pleiotropy,

whenever the relevant trait is available, is a better approach than directly excluding violating instruments. In fact, with increasing GWAS study size, more SNPs will be excluded due to evidence for mild pleiotropy, thereby reducing MR power.

Eventually, it is important to mention that, while in this study we used eQTLs data from gene-expression, our MR approach can be applied to other "omics" (e.g. methylation, metabolomics, proteomics) data. Indeed our method only requires summary statistics from GWAS and any kind of exposure partnered with LD estimates, demonstrating once more the power of a carefully combined analysis of existing data to illuminate biological mechanism underlying complex traits and help the design of functional experiments.

Methods

Multiple-instrument, multiple-exposure Mendelian randomization analysis

Mendelian randomization (MR) is a statistical method that uses genetic variants, so-called instrumental variables (IVs), to estimate the causal effect of an exposure on an outcome.

The IVs used in MR must verify three strong assumptions:

i. They are associated with the exposure.
ii. They are independent of any confounder of the exposure-outcome association.
iii. They are conditionally independent of the output given the exposure.

These assumptions imply that the genetic variants (IVs) have a causal effect on the outcome only via the risk factor. While the first assumption can be easily verified, the second one is impossible to confirm as not all confounders are known, and the third requires instrument-exposure-outcome data measured in the same sample and is often violated by pleiotropy. One (ideal) way to guard against the violation of the third assumption is to use as many IVs possible, as the pleiotropic effect of each marker will cancel each other out under the INSIDE assumption (instrument strength independent of the strength of the pleiotropy).

In our study, we want to estimate the causal effect of the gene expression (exposure) on a complex trait (outcome). As many significant ($P_{eQTL}<1x10^{-03}$) eQTLs are in high LD with other nearby eQTLs, we pruned the eQTLs results using a stepwise selection procedure [55] on the basis of conditional P_{eQTL} to select, for each gene, independent eQTLs. For each gene, using inverse-variance weighted method for summary statistics [18], we defined the causal effect of the gene-expression on the outcome as

$$\hat{\boldsymbol{\alpha}}=(\mathbf{E}'\mathbf{C}^{-1}\mathbf{E})^{-1}(\mathbf{E}'\mathbf{C}^{-1}\mathbf{G})$$

Here **E** is a $n\times k$ matrix that contains the effect size of n SNPs on k gene expressions (these estimates come from an eQTL study). An individual SNP may affect a phenotype via different genes, therefore we estimated the causal effect jointly to allow for this correlation: starting from the IVs of the gene i, we included in the model all independent genes $e_2,\ldots,e_k$ for which the IVs are significant eQTLs and all the

independent significant eQTLs for those genes. **G** is a vector of length *n* that contains the effect size of each SNP on the phenotype (these estimates come from the publicly available GWAS summary statistics). **C** is the pair-wise LD matrix between the *n* SNPs

In order to avoid numerical instability in our multiple regression model, we pruned SNPs that are in high LD (r^2>0.7). The variance of $\boldsymbol{\alpha}$ can be calculated approximately by the Delta method [56] .

$$var(\hat{\alpha}) = \left(\frac{\partial \hat{\alpha}}{\partial \boldsymbol{E}}\right)^2 * var(\boldsymbol{E}) + \left(\frac{\partial \hat{\alpha}}{\partial \boldsymbol{G}}\right)^2 * var(\boldsymbol{G}) + \left(\frac{\partial \hat{\alpha}}{\partial \boldsymbol{E}}\right) * \left(\frac{\partial \hat{\alpha}}{\partial \boldsymbol{G}}\right) * cov(\boldsymbol{E}, \boldsymbol{G})$$

where $cov(\boldsymbol{E},\boldsymbol{G})$ is 0 if **E** and **G** are estimated from independent samples (or if **E** and **G** are independent).

We defined the causal effect Z-statistic for gene i as $\hat{\alpha}_i/SE(\hat{\alpha}_i)$, where $SE(\hat{\alpha}_i) = \sqrt{var(\hat{\alpha})_{i,i}}$.

Applying advanced MR approach to GWAS and eQTL summary results

We applied our multi-instrument, multi-exposure MR approach to test each gene across the human genome for causal association between a phenotype and the expression level using summary statistics from GWAS and eQTLs studies.
We cover 43 traits by using publicly available GWAS summary statistics (i.e. for each SNP we extracted the estimated univariate effect size and its standard error) from the most recent meta-analyses. The traits analysed in this study are listed in **Supplementary Table 1**.
All summary statistics were downloaded from the NIH Genome-wide Repository of Associations Between SNPs and Phenotypes (https://grasp.nhlbi.nih.gov/)
We used only SNPs on autosomal chromosomes and available in the UK10K reference panel, in order to be able to estimate the LD among these SNPs. We removed SNPs that were strand ambiguous, as well as those in the major histocompatibility complex (MHC) region (chr6:26.2-33.8 Mb).
Cis-eQTL data were obtained from the eQTLGen Consortium (14,115 whole blood samples) and the Genotype-Tissue Expression Project (GTEx) [16], which includes 48 tissues collected from 11,688 post-mortem biopsies from 635 individuals (see **Supplementary Table 7** for sample size per tissue)

Heterogeneity test

The validity of the MR approach relies on three assumptions, of which the third assumption (no pleiotropy) is crucial as MR causal estimates will be biased if the genetic variants (IVs) have pleiotropic effects [17]. Our method assumes that all genetic variants used as instrumental variables affect the outcome only through gene expression and not through independent biological pathways.
In order to test for the presence of pleiotropy, we used Cochran's Q test [57] [20].

In brief, we tested whether there is a significant difference between the MR-effect of an instrument (i.e. $\sum_{k=1}^{K} \alpha_k E_{i,k}$) and the estimated effect of that instrument on the phenotype (G_i). We defined

$$d_i = G_i - \sum_{k=1}^{K} \alpha_k \, E_{i,k}$$

and its variance as

$$var(d_i) = var(G_i) + (E_i)^2 * var(\alpha) + \; var(E_i) * (\alpha)^2 + \\ +var(E_i) * var(\alpha)$$

We can test the deviation of each SNP using the following test statistic

$$\chi_K^2 = \frac{d_i^2}{var(d_i)}$$

In case of $P<10^{-4}$, we removed the SNP with largest $|d_i|$ and then repeated the test. Performing 20 iterations of the heterogeneity test for height, we observed that after 3 iterations we corrected more than one third of the genes and after 10 we reach a plateau. Thus for practical reasons we decided to perform maximum 3 iterations for each phenotype (**Supplementary Fig. 8**) and loci still showing heterogeneity were discarded from further analysis.

Correlation between genes

A common observation in the analysis of gene expression is that often physically close genes show similar expression patterns [58]. In order to avoid numerical instability caused by near-collinearity in our multiple regression model, we removed one gene from each pair of genes with $r^2 \geq 0.4$. The correlation r^2 was estimated as Pearson's correlation between the Z-scores of the shared, independent eQTLs.

Genetic correlation between traits at the level of gene expression

To estimate the phenotypic correlation between each pair of traits we calculated the Pearson's correlation between the Z-scores across the set of 2,203 independent genes extracted from the 12,759 genes analysed for all the 43 traits included in our analyses.

Genome-wide Association Study in UKBB

We ran genome-wide association study of BMI in 379,530 unrelated British individuals from UKBB using linear model association testing implemented in bgenie software [59].
We analysed only SNPs included in UK10K reference panel and with a rs name. In total we analysed 15,599,830 SNPs.
BMI was adjusted for sex, age, age^2 and 40 principal components.

Acknowledgments

This work was supported by grants from the Swiss National Science Foundation (31003A-143914 to ZK and 31003A_160203 to AR) and the Horizon2020 Twinning project ePerMed (692145 to AR). This research has been conducted using the UK Biobank Resource (#16389). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The members of the eQTLGen Consortium are (in alphabetical order): Mawussé Agbessi, Habibul Ahsan, Isabel Alves, Anand Andiappan, Philip Awadalla, Alexis Battle, Frank Beutner, Marc Jan Bonder, Dorret Boomsma, Mark Christiansen, Annique Claringbould, Patrick Deelen, Tõnu Esko, Marie-Julie Favé, Lude Franke, Timothy Frayling, Sina Gharib, Gregory Gibson, Gibran Hemani, Rick Jansen, Mika Kähönen, Anette Kalnapenkis, Silva Kasela, Johannes Kettunen, Yungil Kim, Holger Kirsten, Peter Kovacs, Knut Krohn, Jaanika Kronberg-Guzman, Viktorija Kukushkina, Zoltan Kutalik, Bernett Lee, Terho Lehtimäki, Markus Loeffler, Urko M. Marigorta, Andres Metspalu, Lili Milani, Martina Müller-Nurasyid, Matthias Nauck, Michel Nivard, Brenda Penninx, Markus Perola, Natalia Pervjakova, Brandon Pierce, Joseph Powell, Holger Prokisch, Bruce Psaty, Olli Raitakari, Susan Ring, Samuli Ripatti, Olaf Rotzschke, Sina Ruëger, Ashis Saha, Markus Scholz, Katharina Schramm, Ilkka Seppälä, Michael Stumvoll, Patrick Sullivan, Alexander Teumer, Joachim Thiery, Lin Tong, Anke Tönjes, Jenny van Dongen, Joyce van Meurs, Joost Verlouw, Peter Visscher, Uwe Völker, Urmo Võsa, Hanieh Yaghootkar, Jian Yang, Biao Zeng, Futao Zhang.

Contributions

Z.K. conceived and designed the study. Z.K., A.R and E.P. derived the theories. E.P. performed statistical analyses. Z.K. and E.P. designed the method. S.R. and F.A.S. contributed by providing statistical support. All the authors contributed by providing advice on interpretation of results. E.P, A.R. and Z.K. wrote the manuscript with the participation of all authors. eQTLGen Consortium provided eQTLs data.

References

1. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.
2. Flister, M.J., et al., *Identifying multiple causative genes at a single GWAS locus.* Genome Res, 2013. **23**(12): p. 1996-2002.
3. Benner, C., et al., *FINEMAP: efficient variable selection using summary data from genome-wide association studies.* Bioinformatics, 2016. **32**(10): p. 1493-501.
4. Hormozdiari, F., et al., *Identifying causal variants at loci with multiple signals of association.* Genetics, 2014. **198**(2): p. 497-508.

5. 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, 2011. **7**(8): p. e1002197.
6. 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, 2012. **47**(1): p. 20-8.
7. Nica, A.C., et al., *Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations.* PLoS Genet, 2010. **6**(4): p. e1000895.
8. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.
9. Gusev, A., et al., *Integrative approaches for large-scale transcriptome-wide association studies.* Nat Genet, 2016. **48**(3): p. 245-52.
10. Mancuso, N., et al., *Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits.* Am J Hum Genet, 2017. **100**(3): p. 473-487.
11. Zhu, Z., et al., *Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.* Nat Genet, 2016. **48**(5): p. 481-7.
12. Zhu, Z., et al., *Causal associations between risk factors and common diseases inferred from GWAS summary data.* Nat Commun, 2018. **9**(1): p. 224.
13. Westra, H.J., et al., *Systematic identification of trans eQTLs as putative drivers of known disease associations.* Nat Genet, 2013. **45**(10): p. 1238-1243.
14. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
15. Huang, J., et al., *Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel.* Nat Commun, 2015. **6**: p. 8111.
16. Consortium, G.T., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.
17. Verbanck, M., et al., *Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases.* Nat Genet, 2018. **50**(5): p. 693-698.
18. Burgess, S., A. Butterworth, and S.G. Thompson, *Mendelian randomization analysis with multiple genetic variants using summarized data.* Genet Epidemiol, 2013. **37**(7): p. 658-65.
19. McDaid, A.F., et al., *Bayesian association scan reveals loci associated with human lifespan and linked biomarkers.* Nat Commun, 2017. **8**: p. 15842.
20. Burgess, S., et al., *Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants.* Epidemiology, 2017. **28**(1): p. 30-42.
21. Liu, J.Z., et al., *A versatile gene-based test for genome-wide association studies.* Am J Hum Genet, 2010. **87**(1): p. 139-45.
22. Li, M.X., et al., *GATES: a rapid and powerful gene-based association test using extended Simes procedure.* Am J Hum Genet, 2011. **88**(3): p. 283-93.
23. Takenouchi, T., et al., *Macrothrombocytopenia and developmental delay with a de novo CDC42 mutation: Yet another locus for thrombocytopenia and developmental delay.* Am J Med Genet A, 2015. **167A**(11): p. 2822-5.

24. Hadano, S., et al., *A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2.* Nat Genet, 2001. **29**(2): p. 166-73.
25. Girard, S.L., et al., *Increased exonic de novo mutation rate in individuals with schizophrenia.* Nat Genet, 2011. **43**(9): p. 860-3.
26. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nat Genet, 2014. **46**(11): p. 1173-86.
27. Matsuzono, Y., et al., *Human PEX19: cDNA cloning by functional complementation, mutation analysis in a patient with Zellweger syndrome, and potential role in peroxisomal membrane assembly.* Proc Natl Acad Sci U S A, 1999. **96**(5): p. 2116-21.
28. Cohn, D.H., et al., *Mental retardation and abnormal skeletal development (Dyggve-Melchior-Clausen dysplasia) due to mutations in a novel, evolutionarily conserved gene.* Am J Hum Genet, 2003. **72**(2): p. 419-28.
29. Kuehn, H.S., et al., *Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4.* Science, 2014. **345**(6204): p. 1623-1627.
30. Marouli, E., et al., *Rare and low-frequency coding variants alter human adult height.* Nature, 2017. **542**(7640): p. 186-190.
31. Claussnitzer, M., et al., *FTO Obesity Variant Circuitry and Adipocyte Browning in Humans.* N Engl J Med, 2015. **373**(10): p. 895-907.
32. Musunuru, K., et al., *From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus.* Nature, 2010. **466**(7307): p. 714-9.
33. Henning, K.A., et al., *The Cockayne syndrome group A gene encodes a WD repeat protein that interacts with CSB protein and a subunit of RNA polymerase II TFIIH.* Cell, 1995. **82**(4): p. 555-64.
34. Lo, M.T., et al., *Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders.* Nat Genet, 2017. **49**(1): p. 152-156.
35. Astle, W.J., et al., *The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.* Cell, 2016. **167**(5): p. 1415-1429 e19.
36. Okbay, A., et al., *Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses.* Nat Genet, 2016. **48**(6): p. 624-33.
37. Wain, L.V., et al., *Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets From Blood and the Kidney.* Hypertension, 2017.
38. Hom, G., et al., *Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX.* N Engl J Med, 2008. **358**(9): p. 900-9.
39. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature, 2014. **506**(7488): p. 376-81.
40. Warren, H.R., et al., *Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk.* Nat Genet, 2017. **49**(3): p. 403-415.
41. Luciano, M., et al., *Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism.* Nat Genet, 2018. **50**(1): p. 6-11.

42. Liu, J.Z., et al., *Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.* Nat Genet, 2015. **47**(9): p. 979-986.
43. Kaprio, J., et al., *Common genetic influences on BMI and age at menarche.* Hum Biol, 1995. **67**(5): p. 739-53.
44. Day, F.R., et al., *Shared genetic aetiology of puberty timing between sexes and with health-related outcomes.* Nat Commun, 2015. **6**: p. 8842.
45. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.
46. Day, F.R., et al., *Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk.* Nat Genet, 2017. **49**(6): p. 834-841.
47. Parsons, T.J., et al., *Childhood predictors of adult obesity: a systematic review.* Int J Obes Relat Metab Disord, 1999. **23 Suppl 8**: p. S1-107.
48. Onland-Moret, N.C., et al., *Age at menarche in relation to adult height: the EPIC study.* Am J Epidemiol, 2005. **162**(7): p. 623-32.
49. Garrison, R.J., et al., *Educational attainment and coronary heart disease risk: the Framingham Offspring Study.* Prev Med, 1993. **22**(1): p. 54-64.
50. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nat Genet, 2015. **47**(11): p. 1236-41.
51. Duncan, L.E., et al., *Genetic Correlation Profile of Schizophrenia Mirrors Epidemiological Results and Suggests Link Between Polygenic and Rare Variant (22q11.2) Cases of Schizophrenia.* Schizophr Bull, 2017.
52. Wu, P., et al., *Risk of cardiovascular disease in inflammatory bowel disease.* Exp Ther Med, 2017. **13**(2): p. 395-400.
53. Qi, T., et al., *Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood.* Nat Commun, 2018. **9**(1): p. 2282.
54. Ongen, H., et al., *Estimating the causal tissues for complex traits and diseases.* Nat Genet, 2017. **49**(12): p. 1676-1683.
55. Yang, J., et al., *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits.* Nat Genet, 2012. **44**(4): p. 369-75, S1-3.
56. Lynch, M.W., B. , *Genetics and Analysis of Quantitative Traits*. 1998: Sinauer Associates.
57. Bowden, J., G. Davey Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* Int J Epidemiol, 2015. **44**(2): p. 512-25.
58. Ghanbarian, A.T. and L.D. Hurst, *Neighboring Genes Show Correlated Evolution in Gene Expression.* Mol Biol Evol, 2015. **32**(7): p. 1748-66.
59. Bycroft, C., et al., *Genome-wide genetic data on ~500,000 UK Biobank participants.* bioRxiv, 2017.