bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Quantifying mediation between omics layers and complex traits

Marie C. Sadler [1,2,3,*], Chiara Auwerx [1,2,3,4], Eleonora Porcu [1,2,3,4,5], Zoltán Kutalik [1,2,3,5,*]

[1]University Center for Primary Care and Public Health, Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[4] Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

[5]Authors jointly supervised this work

[*]Corresponding author: marie.sadler@unil.ch, zoltan.kutalik@unil.ch

Abstract

Background: High-dimensional omics datasets provide valuable resources to determine the causal role of molecular traits in mediating the path from genotype to phenotype. Making use of quantitative trait loci (QTL) and genome-wide association studies (GWASs) summary statistics, we developed a multivariable Mendelian randomization (MVMR) framework to quantify the connectivity between three omics layers (DNA methylome (DNAm), transcriptome and proteome) and their cascading causal impact on complex traits and diseases.

Results: Evaluating 50 complex traits, we found that on average 37.8% (95% CI: [36.0%-39.5%]) of DNAm-to-trait effects were mediated through transcripts in the *cis*-region, while only 15.8% (95% CI: [11.9%-19.6%]) are mediated through proteins in *cis*. DNAm sites typically regulate multiple transcripts, and while found to predominantly decrease gene expression, this was only the case for 53.4% across $\approx$ 47,000 significant DNAm-transcript pairs. The average mediation proportion for transcript-to-trait effects through proteins (encoded for by the assessed transcript or located in *trans*) was estimated to be 5.27% (95%CI: [4.11%-6.43%]). Notable differences in the transcript and protein QTL architectures were detected with only 22% of protein levels being causally driven by their corresponding transcript levels. Several regulatory mechanisms were hypothesized including an example where cg10385390 (chr1:8'022'505) increases the risk of irritable bowel disease by reducing *PARK7* transcript and protein expression.

Conclusions: The proposed integrative framework identified putative causal chains through omics layers providing a powerful tool to map GWAS signals. Quantification of causal effects between successive layers indicated that molecular mechanisms can be more complex than what the central dogma of biology would suggest.

Keywords: multi-omics, multivariable Mendelian randomization, omics QTL, GWAS, complex traits, molecular mechanisms, bioinformatics

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

In the past decade, genome-wide association studies (GWASs) have identified thousands of genetic variants associated to complex traits [1], however mapping these variants to molecular processes and pathways still remains challenging [2]. A first step towards interpreting GWAS signals is to map trait-associated single nucleotide polymorphisms (SNPs) to genes. Naive approaches based on physical distance attribute SNPs to their closest gene [3] and many of them additionally take into account the linkage disequilibrium (LD) structure and GWAS association strengths (*i.e.* p-values) to compute gene scores [4, 5]. In a second step, scores from several genes can be combined and mapped to biological pathways by incorporating knowledge from external databases such as KEGG [6], Gene Ontology [7], WikiPathways [8], Reactome [9], or MSigDB [10], and making use of pathway enrichment analysis tools [5, 11, 12].

GWAS signals of common diseases predominantly fall into the non-coding genome [13] and both their enrichment in regulatory elements (e.g. quantitative trait loci (QTL) [13, 14]), as well as advances in omics technology [15], has motivated the establishment of large-scale consortia providing publicly available QTL datasets for molecular phenotypes such as DNA methylation (DNAm) [16], as well as transcript [17, 18], protein [19, 20, 21] or metabolite [22, 23] levels. Consequently, a next step in interpreting GWAS findings has been to integrate this new type of data, allowing to find diverse mediators of SNP-trait associations in a high-throughput, data-driven fashion. Integrative statistical methods combining GWAS and omics QTL summary data include colocalization tests [24, 25], summary versions of transcriptome-wide association studies (TWAS) [26, 27] and Mendelian randomization (MR) studies [28, 29]. Their application to a wide variety of GWAS datasets has resulted in the identification of many putative molecular trait-disease associations confirming known and highlighting potential new molecular mechanisms [30]. Colocalization methods identify shared QTL and GWAS signals, and while this might indicate causality between the molecular and GWAS trait, shared signals can also arise due to reverse causality (*i.e.* causal effect of the GWAS trait on the molecular trait [31]) or horizontal pleiotropy (*i.e.* the identified shared genetic variant drives the molecular and trait perturbation independently). In comparison, MR studies, which are conceptually similar to TWAS, that use multiple genetic variants as instrumental variables (IVs) are less prone to reverse causality and artefacts arising from LD patterns [32] - although horizontal pleiotropy can never be ruled out entirely. An important advantage of MR methods is that they allow the detection and elimination of pleiotropic markers. In addition, MR analyses allow the quantification - direction and magnitude - of the causal effect of the omic on the outcome trait.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

With the advent of QTL datasets with increased sample sizes [16, 18], opportunities to integrate GWAS data with multiple molecular traits are no longer hampered by low statistical power. Previous efforts integrating multiple QTL omics data either adopted colocalization strategies [33, 34] or combined pairwise MR associations (two-step MR) [35, 36] to predict molecular mechanisms of the following scheme: omics trait 1 → omics trait 2 → outcome trait. While these approaches provide evidence for regulatory pathways, ascertaining their robustness can be difficult, since often only a single causal variant underlying these multiple associations was assessed [33, 35, 36]. As a consequence, the control over horizontal pleiotropy remained limited, although it was usually mitigated by the HEIDI (heterogeneity in dependent instruments) test statistic [28]. Furthermore, combining pairwise associations can lack the ability of inferring directionality between the different traits involved, an issue that can be identified by comparing the magnitude of QTL and GWAS effects [37]. Overall, while current integration methods test genetic downstream effects through omics traits, they often only accommodate the testing of a single molecular mediator.

Multivariable MR (MVMR) approaches have been proposed to identify multiple mediators of exposure-outcome relationships [38, 39]. These approaches enable the dissection of the total causal effect of an exposure on an outcome into a direct and indirect effect measured via mediators. Similar to MR, the use of genetic instruments allows for robust causal inference and MVMR has proven as an unbiased approach for mediation analyses, even in the presence of confounders [38, 39]. Hence, in addition to identifying causal effects through multiple layers, MVMR allows the quantification of mediation effects. Although not yet widely implemented on high-dimensional omics data, they provide great opportunities in the study of molecular mediation [40].

Here, we proposed a three-sample MVMR (3S-MVMR) framework to quantify the role of molecular mediators (omics trait 2) on a molecular exposure (omics trait 1) - complex trait relationship (Figure 1). We integrated methylomic, transcriptomic and proteomic QTL (mQTL, eQTL and pQTL, respectively) with GWAS summary data of 50 clinically relevant traits to perform mediation analyses and to estimate global mediation proportions (MPs). Three different combinations of exposure-mediator molecular traits were analysed: DNAm regulating transcripts in *cis*, DNAm regulating proteins in *cis*, and transcripts regulating their encoded protein in addition to proteins in *trans*. We performed simulation studies to estimate the bias of the defined MP under various parameter settings. In addition to quantifying the regulatory connectivity between each of these molecular layers, we investigated underlying factors driving high MPs, and hypothesized several mechanistic pathways between DNAm, gene expression and complex traits.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

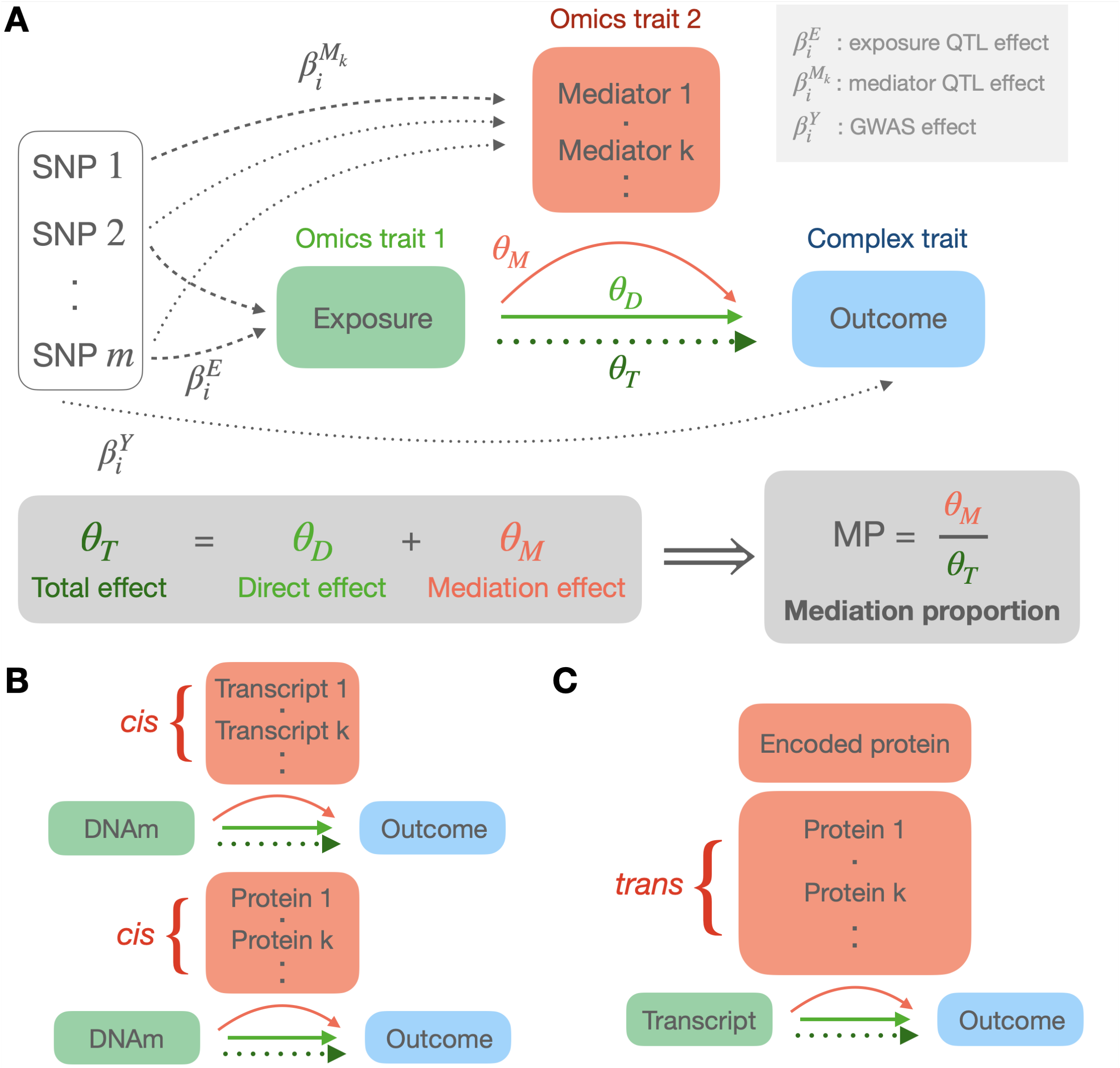

Figure 1: Overview of the MVMR design to quantify mediation of complex traits through DNAm, transcripts and proteins. **A)** General MVMR model: genetic instruments (SNPs) are selected to be directly associated (dashed arrow) with either the exposure (omics trait 1) or any mediator k (omics trait 2). The total effect θ_T (dotted arrow) of the exposure on the outcome (complex trait) is estimated in a univariable MR analysis based on exposure-associated SNPs only. The direct effect θ_D is estimated in a MVMR analysis on all valid instruments. The mediation effect θ_M results from the difference between θ_T and θ_D, and allows to calculate the mediation proportion (MP). The genetic effect sizes β on the exposure, mediator and outcome come from m/e/pQTL and GWAS summary statistics, respectively. **B)** DNAm-to-complex trait effects were mediated once through transcripts in *cis* and once through proteins in *cis*. **C)** Transcript-to-complex trait effects were mediated through the protein the transcript is encoding for (encoded protein; if available in the dataset), as well as through proteins in *trans*. Except for the encoded protein, mediators were required to be causally associated to the exposure in both DNAm- and transcript-exposure settings.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Overview of the method

We performed univariable and multivariable MR to estimate total and direct effects, $\hat{\theta}_T$ and $\hat{\theta}_D$, respectively, of molecular exposures on 50 outcomes through various molecular mediators (Figure 1; Equation 1 and 3). MP estimates were then calculated as the ratio of the indirect effect through the molecular mediators to the total effect of the exposure on the outcome trait [41] and computed only for exposure-outcome pairs with significant Bonferroni-corrected $\hat{\theta}_T$ effects, grouped by trait, trait category and all pairs combined. We further filtered out exposure-outcome pairs whose exposures have no significant causal effect on any potential mediator as the link between those pairs cannot be mediated. For completeness, however, we also present results for the scenario when the last filtering step is omitted.

While weak genetic instruments in univariable MR analyses can introduce a bias towards the null [42], it has been shown that this bias can be in any direction in MVMR studies [43]. Both the sample size and the choice of instruments and mediators can contribute to biases in various directions [43], leading to under- or over-estimations of the MP. To quantify this bias and assess the sensitivity and robustness of estimated $\widehat{\text{MP}}$s, we conducted simulation studies mimicking the settings that emerge in real data applications for either DNAm or transcript levels as exposure (Methods) (Figure S1).

Simulation results

Simulations showed that the bias in the estimated MPs ($\widehat{\text{MP}}$) is minimal for the settings most relevant for real data we explored (Figures S2-3; Table S2; Methods). A determining factor in accurately estimating MPs was the sample size of the mediator QTL effects. Low sample sizes resulted in significant underestimations of the MP, with sample sizes of 3,000 compared to 30,000 resulting in a 20% relative decrease (6% in absolute values) of the estimated $\widehat{\text{MP}}$ in the DNAm-exposure simulation settings (Figure 2A). The reason for this significant underestimation was the omission of relevant mediators with 0.45/3 (15%) being missed at a sample size of 3,000 in the DNAm-exposure simulation settings (Figure 2B). We further tested the robustness of the $\widehat{\text{MP}}$ with respect to the number of included mediators by varying the mediator selection threshold P_{EM} (Methods). At more stringent threshold, relevant mediators were more likely to be missed resulting in an underestimation of the $\widehat{\text{MP}}$ (Figure S4). Importantly, including irrelevant mediators at more lenient thresholds did not bias the $\widehat{\text{MP}}$, although a critical point was reached upon the inclusion of >10 irrelevant mediators where the estimated $\widehat{\text{MP}}$ started to become underestimated in the transcript-exposure setting (Figure S4). The used transcript and DNAm QTL datasets provide SNP effect sizes in *cis* of the assessed transcript and probe, respectively, and were primarily restricted to significant mQTLs for the latter. Thus, SNP-exposure effects for SNPs serving as mediator instruments are often missed and set to zero. However, our simulation studies, which mimicked this scenario by

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

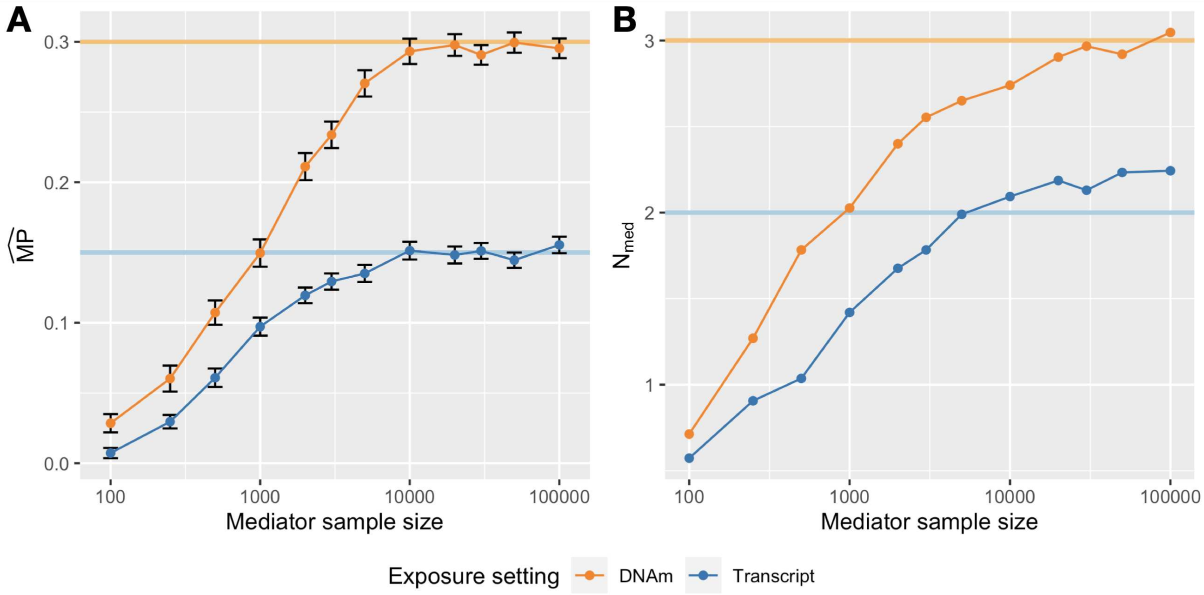

Figure 2: Simulation results in DNAm- (orange) and transcript- (blue) exposure settings to assess the impact of the mediator sample size on the A) estimated $\widehat{\text{MP}}$ and B) number of selected mediators. For a given mediator sample size, 300 exposure-outcome pairs were simulated on which an $\widehat{\text{MP}}$ and 95% CI (error bars) were estimated. The true MP of the model was 0.3 and 0.15, and the true number of relevant mediators was 3 and 2 in the DNAm-exposure and transcript-exposure setting, respectively, as indicated by the solid horizontal lines.

setting non-significant effects to zero (Methods), showed that this did not induce any bias.

DNAm-to-complex trait effects mediated by gene expression in *cis*

Across 50 traits (Table S1), we evaluated the mediation of 2,069 DNAm-trait causal pairs by transcripts in *cis*. The $\widehat{\text{MP}}$ for each of the 41 traits influenced by at least 10 DNAm probes ranged from 18.0 to 78.0% (mean: 36.9%, 95% CI: [13.5%-60.3%]) (Figure 3A). Regressing $\hat{\theta}_D$ against $\hat{\theta}_T$ for all pairs combined and accounting for regression dilution bias (Equation 4) yielded an $\widehat{\text{MP}}$ of 37.8% (95% CI: [36.0%-39.5%]) (Figure 3B). Grouping the traits into 10 physiological categories (Table S1) showed that the $\widehat{\text{MP}}$ was highest for hepatic biomarkers (mean: 46.6%, 95%CI: [41.5%-51.7%]), followed by renal biomarkers (mean: 43.5%, 95%CI: [37.5%-49.5%]). In contrast, adiposity-related and hormonal traits exhibited the lowest $\widehat{\text{MP}}$ (Figure 3B, Figure S5).

The average number of mediator transcripts was 3.3 per methylation-trait pair, indicating that the impact of methylation is not mediated by a single transcript. To further explore this observation, we assessed the extent to which DNAm→trait effects were mediated by the single most significantly DNAm-associated transcript ("top" transcript; Methods), as opposed to all transcripts in *cis*. This resulted in an $\widehat{\text{MP}}_{\text{top}}$ of 26.0% (range: [13.0%-46.8%]) averaged across the 41 traits, and an $\widehat{\text{MP}}_{\text{top}}$ of 26.6% (95% CI:

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[25.1%-28.1%]) when aggregating the 2,069 DNAm-trait pairs. This significant drop in the $\widehat{\text{MP}}$ ($P_{\text{diff}} <$ 5e-21) corroborates our initial hypothesis that DNAm sites regulate the expression of multiple transcripts in the *cis* region.

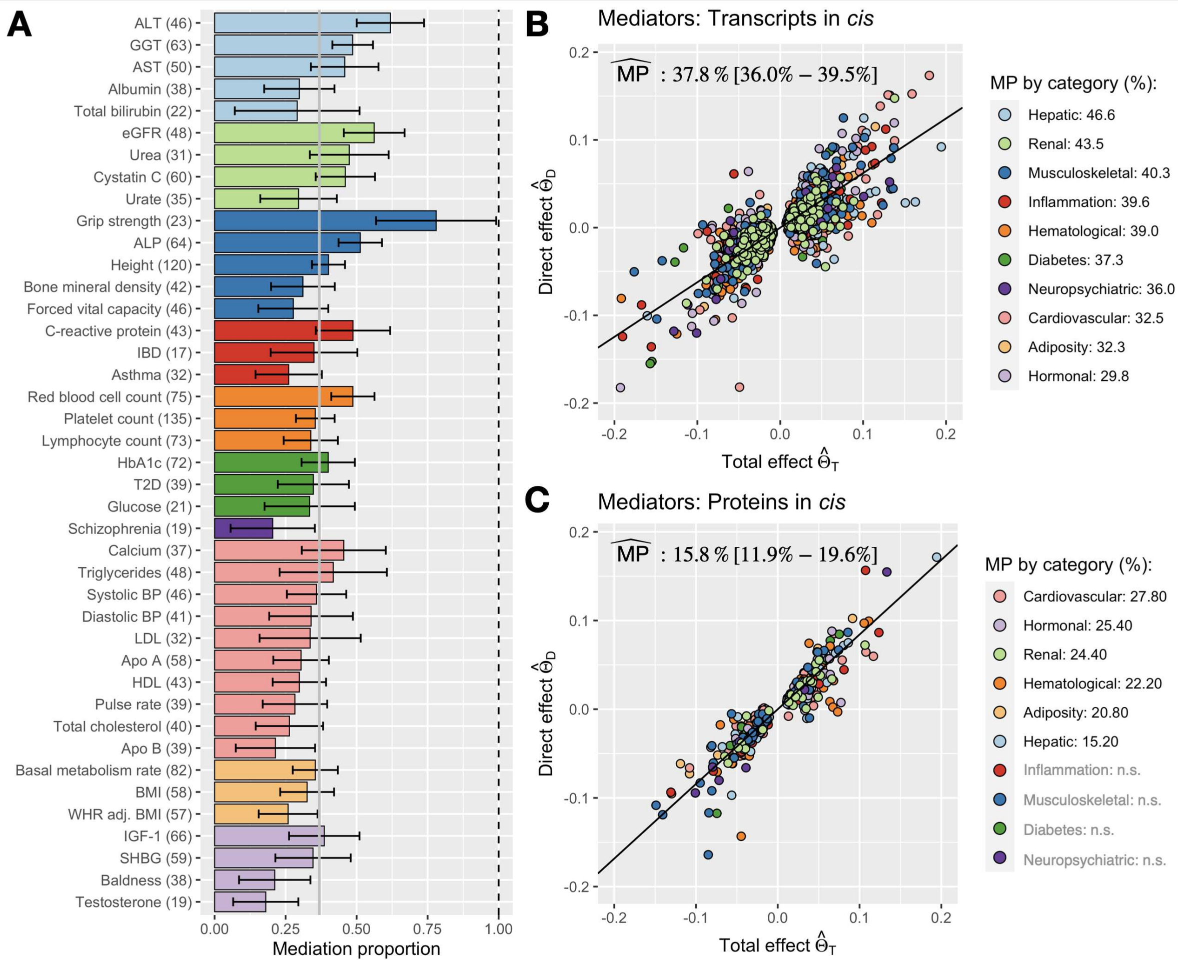

Figure 3: $\widehat{\text{MP}}$s for transcripts and proteins in *cis* mediating DNAm-to-trait effects. **A)** $\widehat{\text{MP}}$s by trait in the DNAm-to-trait via transcripts in *cis* analysis. Error bars denote the 95% CI, and the grey vertical bar shows the mean $\widehat{\text{MP}}$ across the traits. Only traits with $\geq$ 10 DNAm-trait pairs are displayed (41 traits with the exact number of evaluated pairs indicated in parentheses), colour-coded by their physiological category as defined in the legends of B) and C). **B)** All DNAm-trait pairs with traits being grouped into 10 physiological categories. The global $\widehat{\text{MP}}$ in % with 95% CI is shown in the plotting area and individual category $\widehat{\text{MP}}$s in the legend. **C)** Same analysis as in B), but with mediators being proteins in *cis*. Category $\widehat{\text{MP}}$s not significantly different from zero (n.s.) are written in grey.

Including DNAm-trait pairs with testable transcripts in the *cis* region, but not causally linked to the assessed DNAm site (2,623 DNAm-trait pairs, Methods: adjusted MP calculation; Table S3), decreased the overall $\widehat{\text{MP}}$ to 28.3% (95% CI: [26.9%-29.8%]) (Figure S7). While it may seem to be a more objective measure of the importance of the transcriptome in mediating DNAm-to-phenotype effects, it is overly conservative since the set of testable transcript mediators (N = 19,250 [18]) is a magnitude lower than

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that of the whole transcriptome [44]. A distribution of the number of times no mediation analysis could be conducted due to the absence of (causally associated) transcripts in the region or insufficient (exposure-associated) IVs is shown in Figure S24.

Transcripts levels are under tighter DNAm control than protein levels

Next, we investigated the role of protein levels as mediators. Assessing the same DNAm-trait pairs as previously, we performed mediation analyses based on a potential mediator set of 2,838 proteins in total (INTERVAL pQTL dataset [19]). The estimated $\widehat{\text{MP}}$ equalled 15.8% (95% CI: [11.9%-19.6%]) across 328 DNAm-trait pairs with at least 1 mediator protein (Figure 2C). Highest $\widehat{\text{MP}}$s were obtained for cardiovascular traits (mean: 27.8%, 95% CI: [19.9%-35.8%]) (Figure S6). Given the lower sample size of the pQTL dataset and the results from the simulation studies, a drop in MP was expected. Not only the lower sample size, but also the lower number of testable proteins contributed to this decrease. To compare the difference in MP due to the mediators being transcripts instead of proteins, we repeated the analysis on the common set of transcripts and their encoded proteins (N = 2,145). We observed a drop in the adjusted $\widehat{\text{MP}}$ from 28.3% (95% CI: [26.9%-29.8%], 2,623 DNAm-trait pairs) to 8.15% (95% CI: [7.11%-9.19%], 2,111 DNAm-trait pairs) for transcripts and from 1.24% (95%CI: [0.66%-1.83%], 2,380 DNAm-trait pairs) to 0.85% (95%CI: [0.30%-1.39%], 2,111 DNAm-trait pairs) for proteins (Figure S8). A key difference in the two mediation analyses was the number of mediators (N_{med}) found to be causally associated to the DNAm site and subsequently included in the mediation analysis (mean $N_{med,transcript}$ = 0.48 and mean $N_{med,protein}$ = 0.12). Restricting MP calculations to the same DNAm-trait pairs with at least one transcript and protein mediator, no statistical difference between the two MPs could be detected (P_{diff} = 0.28; Figure S9). Besides differences in sample size, a previous pQTL study of larger sample size (N = 30,931) reported strong differences in the underlying genetic architecture of transcript and protein levels, with less than a third of pQTLs being also eQTLs [21]. Accordingly, we found that only 333 out of 1,510 transcripts (*i.e.* those with a corresponding protein product present in the INTERVAL dataset and having at least 3 independent eQTLs to be used as IVs) could explain the levels of their encoded protein at a nominal significance threshold (Methods; Table S6). Focusing on DNAm-trait pairs where the top transcript mediator was the same than the top protein mediator (N = 106), the proportion of protein levels causally linked to their transcript levels increased to 72%. While both mediation analyses yielded similar results for transcript and protein levels with the same QTL structure, the findings suggest that overall, the genetic architecture of mQTLs is more similar to the one of eQTLs than to the one of pQTLs, which translates to a stronger DNAm-trait mediation through transcripts than through protein levels.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Determining factors of mediation proportions

We further explored underlying factors driving high MPs through transcript levels (Figure 4A). $\widehat{\text{MP}}_{\text{top}}$ decreased with increased distances between the DNAm site and the gene transcription start site (TSS) of the top transcript (ρ = -0.076, P = 5.2e-4; Figure 4B). Further investigations revealed that this distance is negatively correlated to the DNAm-to-transcript MR squared effect size, α^2_{EM}, (ρ = -0.13, P = 3.1e-19; Figure 4C), which in turn is a good predictor for high MPs (ρ = 0.39, P = 2.5e-75; Figure 4D). The mediation proportion was the highest for DNAm sites residing in the first exon, followed by those in the 5'UTR, within 200bp of the TSS and finally lowest for those within 1500bp and in the gene body (Figure S10).

DNAm inhibiting the binding of transcription factors (TFs) and thus repressing gene expression is often alluded to as the classical mechanism of action for DNAm [45] and might be driving this observation. However, many other mechanisms have been hypothesized [46] and many might still be unknown [47]. From the 1,066,307 unique DNAm-to-transcript causal effects assessed, 47,445 were significant at P $<$ 4.7e-8. Although negative effects had a larger magnitude than positive ones (two-sided t-test: P = 0.0082) only 53.4% of DNAm→transcript causal effects were negative. Stratifying DNAm sites with respect to their location on the assessed transcript, we found that DNAm sites situated in the first exon and nearby the TSS were enriched for negative effects (P = 2.7e-3, 1.2e-5 and 3.8e-4 for 1st exon, TSS1500 and TSS200, respectively), whereas those in the gene body were enriched for positive ones (P = 2.2e-10; Table S4). These observations are in line with previous studies that only showed a slight trend for negative methylation-gene expression correlations [46, 48, 49, 47]. We further tested whether the MR DNAm-to-transcript causal effects correlated with reported methylation-transcript correlations [48] and found a strong agreement (ρ = 0.39, P = 2.6e-18, 471 DNAm-transcript pairs).

Consistent with higher MPs when mediating through multiple transcripts, we found a strong correlation between the number of mediators and the MP (ρ = 0.39, P = 4.4e-75; Figure 4E). Many of these mediators were correlated amongst each other, which in theory should be accounted for via the multivariable Mendelian randomisation. To ensure that this was the case, we repeated the mediation analysis with uncorrelated mediators ($\text{R}_{\text{med}} < 0.3$; Methods). The mean number of selected mediators dropped by more than half, from 3.3 to 1.2 (Figure S11), and the $\widehat{\text{MP}}$ across all the DNAm-trait pairs decreased ($\widehat{\text{MP}}_{\text{uncorrelated}}$ = 30.5% (95% CI [28.8%-32.1%])), while remaining significantly higher than $\widehat{\text{MP}}_{\text{top}}$ (P_{diff} = 6.6e-4). Decreasing the R_{med} threshold to 0.2 and 0.1 did not significantly decrease $\widehat{\text{MP}}_{\text{uncorrelated}}$ (P_{diff} $>$ 0.05), which stabilized at 29.2% (95% CI: [27.5%-30.8%] for $\text{R}_{\text{med}} < 0.1$ (Figure S11).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

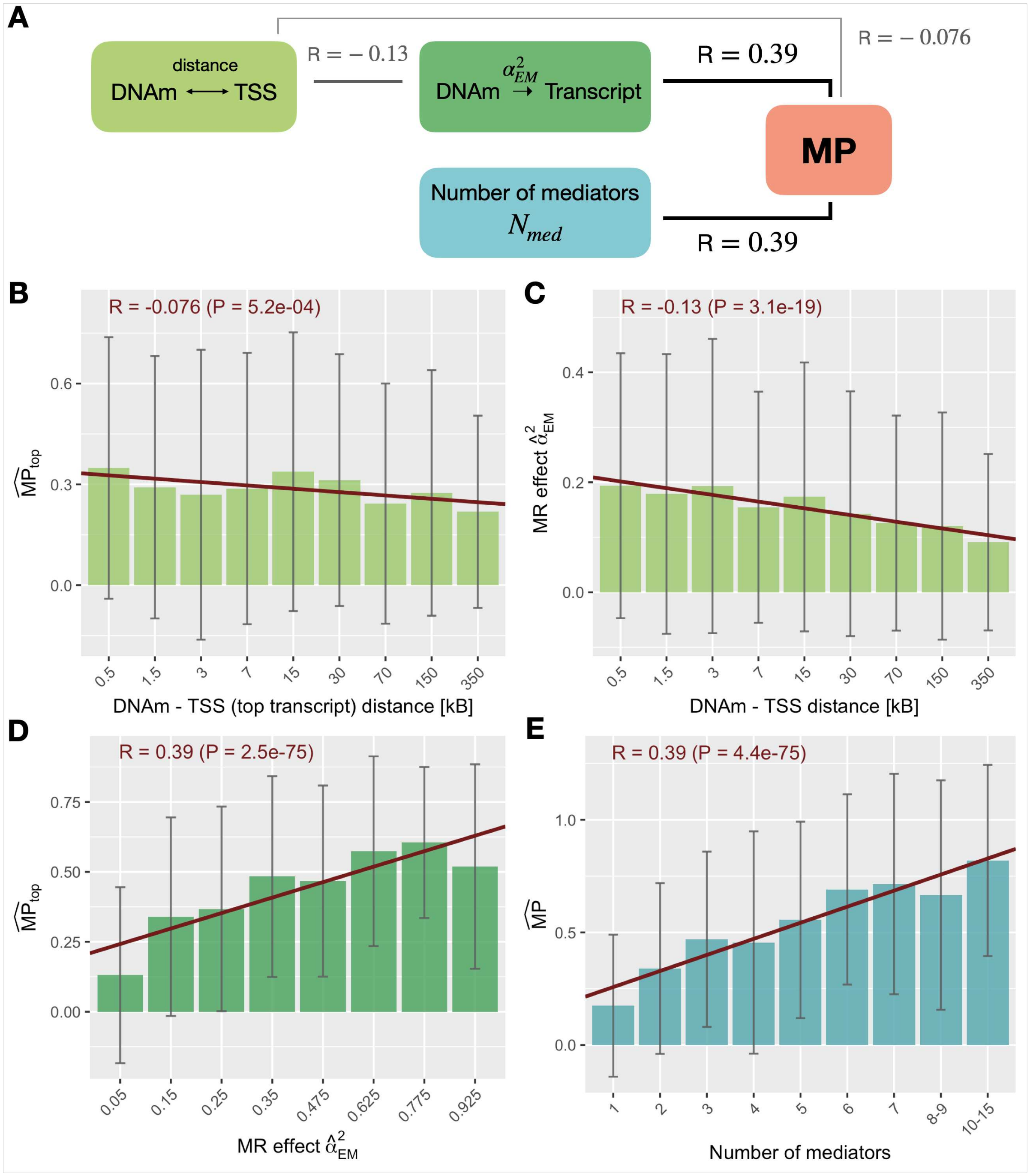

Figure 4: Exposure-to-mediator regulatory strength and number of mediators explaining MPs. **A)** Summary of the correlations (R) between MP and DNAm-to-transcript causal MR effects ($\hat{\alpha}^2_{EM}$; dark green), distance between the DNAm site and transcription start site (TSS; light green) and number of mediators (N_{med}; blue). **B)** Average $\widehat{\text{MP}}_{\text{top}}$ of DNAm-transcript pairs stratified according to the distance between the DNAm site and the TSS of the top transcript. All DNAm-trait pairs with at least one mediator were included. **C)** Average MR causal effects ($\hat{\alpha}^2_{EM}$) of DNAm-transcript pairs stratified according to the distance between the DNAm site and the TSS. Unique DNAm-transcript mediator pairs across all DNAm-trait pairs were included. **D)** Average top $\widehat{\text{MP}}$s ($\widehat{\text{MP}}_{\text{top}}$) of DNAm-trait pairs stratified according to DNAm-to-top transcript MR causal effect size $\hat{\alpha}^2_{EM}$. All DNAm-trait pairs with at least one mediator were included. **E)** Average $\widehat{\text{MP}}$s of DNAm-trait pairs stratified according to the number of mediators. All DNAm-trait pairs with at least one mediator were included. In every calculation, Pearson correlations and corresponding p-values (P) between the two respective quantities were calculated on DNAm-trait/DNAm-transcript pairs prior stratification. Error bars represent standard deviations and the red slope represents the regression fit between the bin's positions and heights, and serves merely for visualization purposes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we assessed the influence of the p-value threshold P_{EM} to select mediators based on the exposure-to-mediator causal effect (default P_{EM} = 0.01 for which N = 2,069 DNAm-trait pairs with at least 1 mediator were found). With a more lenient threshold (P_{EM} = 0.05), more DNAm-trait pairs with mediators emerged (N = 2,189). Conversely, with a more stringent threshold (P_{EM} = 0.001), less pairs were detected (N = 1,881). No differences in MPs between the three settings were found ($P_{diff} > 0.05$; Figure S12), but when calculating the adjusted MP (inclusion of all DNAm-trait pairs with potential transcript mediators in the *cis*-region) on a common set of DNAm-trait pairs (N = 2,543, $\widehat{MP}_{adj,P01}$ = 27.6% (95% CI: [26.1%-29.2%])), a significantly higher MP for the more lenient threshold ($\widehat{MP}_{adj,P05}$ = 32.0% (95% CI: [30.4%-33.6%]); P_{diff} = 1.1e-4), and significantly lower MP for the more stringent threshold were observed ($\widehat{MP}_{adj,P001}$ = 24.6% (95% CI: [23.2%-26.1%]; P_{diff} = 4.8e-3; Figure S13). Overall, these sensitivity analyses showed that the estimated MPs remain robust with respect to correlations within the selected mediator set, while also suggesting that the P_{EM} may be threshold-sensitive and mediators selected at the 0.01 p-value threshold may lead to conservative MP estimates.

Transcript-to-complex trait effects mediated by proteins

Next, we quantified the role of proteins in mediating transcript-to-trait causal effects. First, we identified 3,848 significant transcript-trait pairs ($P_T < 5e\text{-}5$, $\geq$ 5 IVs; Table S3) across the 50 traits and performed mediation analyses through the protein encoded by the transcript (if present in the INTERVAL pQTL dataset [19]), in addition to any other protein in *trans* (*i.e.* any protein which is not encoded by the investigated transcript) causally associated to the transcript (P_{EM} = 1e-3; Figure 5A). The estimated $\widehat{MP}$ for the 1,577 transcript-trait pairs with at least 1 mediator was 5.27% (95%CI: [4.11%-6.43%]) and significantly higher than average for cardiovascular traits (P_{diff} = 8.7e-3; mean = 9.62%, 95% CI: [6.58%-12.7%]; Figure 5B). A distribution of the number of times transcript-trait pairs could be assessed in a mediation analysis through proteins is shown in Figure S25. When further restricting the mediation analysis to only those transcript-trait pairs for which the encoded protein was present, we observed an $\widehat{MP}_{encoded}$ of 5.08% (95% CI: [2.62%-7.53%], 333 transcript-trait pairs) which increased, albeit not significantly, when additionally considering mediator proteins in *trans* ($\widehat{MP}_{trans}$ = 6.89%; 95% CI: [4.17%-9.61%]; Figure S14). As mentioned previously, less than a quarter of the causal MR effects of transcripts on their encoded proteins were nominally significant and since we found exposure-mediator effects to be driving the MP (Figure 4B), we next focused on the 93 transcript-trait pairs for which the encoded protein was nominally significantly associated to the transcript (Figure S16). We observed a non-significant increase in the $\widehat{MP}_{encoded}$ to 13.7% (95% CI: [4.34%-23.0%]) and in the $\widehat{MP}_{trans}$ to 16.6% (95% CI: [8.19%-25.1%]). Stratifying traits by broad categories (e.g. metabolite, protein, physical measurement; Table S1), highest MPs were achieved for protein outcome traits (e.g. apolipoprotein B, alkaline phosphatase; Figure S15 and S17).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

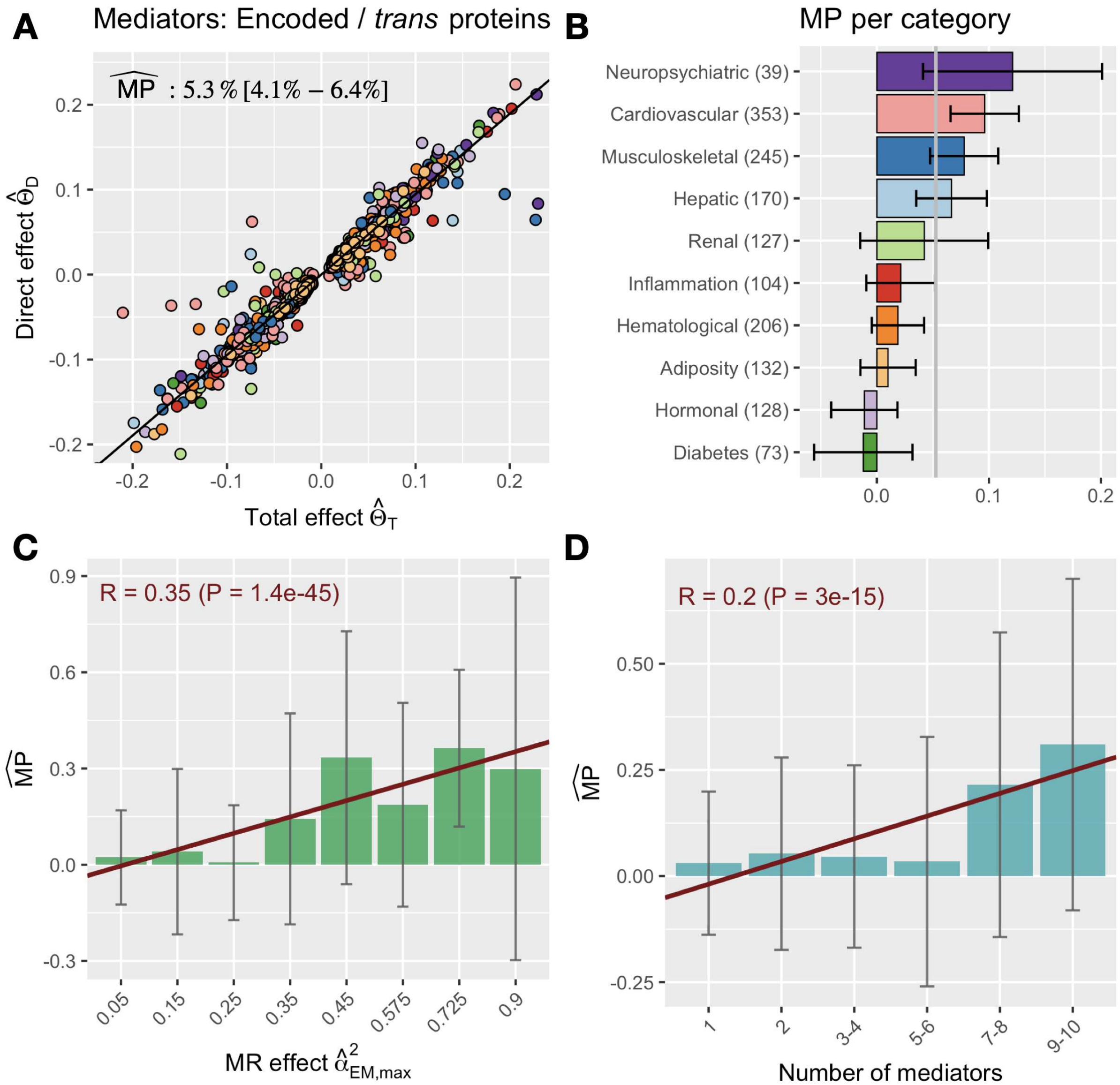

Figure 5: Proteins mediating transcript-to-trait causal effects. **A)** All transcript-trait pairs with with at least one mediating protein (encoded or/and in *trans*). Pairs are colour-coded by physiological categories as defined in B) and overall $\widehat{MP}$ in % with 95% CI is shown. **B)** $\widehat{MP}$s by trait category for the same pairs. Error bars denote the 95% CI, and the grey vertical line shows the mean $\widehat{MP}$ across transcript-trait pairs. The number of evaluated transcript-trait pairs in each category is indicated in parentheses. **C)** $\widehat{MP}$s of transcript-trait pairs stratified according to the maximum transcript-to-protein MR causal effect size within the set of included mediator proteins. **D)** $\widehat{MP}$s of transcript-trait pairs stratified according to the number of mediators. Correlations (R) and corresponding p-values (P) in C) and D) were calculated on transcript-trait pairs (same ones as in A) and B)) prior to stratification. Error bars represent standard deviations and the red slope represents the regression fit between the bin's positions and heights, and serves merely for visualization purposes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As for the DNAm-trait mediation analysis, we confirmed that strong exposure-mediator effect size (α^2_{EM}) was the major driver of high MPs (Figure 5C). We calculated the correlation between MP and the maximum causal effect size squared between the transcript and any of its mediator proteins ($\alpha^2_{EM,max}$), yielding ρ = 0.35 (P = 1.4e-45, N_{pairs} = 1,577, mean N_{med} = 2.15). Additionally, we found a significant correlation between N_{med} and MP (ρ = 0.20, P = 3.0e-15, N_{pairs} = 1,577). We performed further sensitivity analyses to assess the influence of the P_{EM} threshold (default P_{EM} = 1e-3). Considering all transcript-trait pairs (including those with no encoded protein), choosing a more lenient threshold (P_{EM} = 0.01) resulted in more transcript-trait pairs to be evaluated (N_{pairs} = 2,820), but not in a significant change in MP ($\widehat{\text{MP}}$ = 4.08%, 95% CI: [3.14%-5.03%]; $\text{P}_{\text{diff}} > 0.05$; Figure S18). On the other hand, a more stringent threshold (P_{EM} = 1e-4) resulted in fewer transcript-trait pairs (N_{pairs} = 758) and a significantly higher MP ($\widehat{\text{MP}}$ = 10.9%, 95% CI: [9.02%-12.9%]; P_{diff} = 7.2e-7), as consequence of selecting transcript-trait pairs with higher α^2_{EM} (Figure S21).

Finally, we aimed at validating our results using a different protein dataset. While there are publicly available pQTL datasets of larger sample size, the number of tested proteins in these studies is orders of magnitude lower (e.g. 71 proteins (N = 6,861) in the Framingham Heart Study [20]; 90 proteins (N = 30,000) in the SCALLOP consortium [21]). We used pQTL summary statistics of 41 cardiovascular proteins released by the SCALLOP Consortium [21] which overlapped with our main pQTL dataset from the INTERVAL Consortium [19], as well as with our main eQTL dataset of the corresponding transcripts (eQTLGen Consortium [18]). In a first step, we tested the agreement of MR causal effects of the transcripts on their encoded protein between the two protein datasets. Among the 38 tested transcript-protein pairs ($\geq$ 1 eQTL), we observed a very strong correlation of causal effect sizes (ρ = 0.74, P = 9.4e-8) and no difference in their magnitude (two-sided t-test: P $>$ 0.05). Given the larger sample size of the SCALLOP dataset, standard errors of the effect estimates were on average three times smaller (Figure S19). Due to the small number of overlap between the proteins and transcripts in the three datasets, there were only 6 transcript-trait pairs that could be compared in the mediation analysis through encoded protein, and for these, there was no significant difference in the direct effects $\hat{\theta}_D$ (Figure S20). Overall, while we could replicate causal effect sizes between the transcript and protein levels, the small number of available proteins did not allow to reliably quantify the bias in our estimated MP caused by the comparably small sample size of the pQTL dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DNAm-to-complex traits mechanisms of action

In addition to providing insights into global patterns governing the mediation between different intermediate phenotypic layers and functional traits, our analyses generated plausible hypotheses regarding specific biological pathways. Recently, the involvement of the anti-oxidant and anti-inflammatory protein PARK7 in inflammatory bowel disease (IBD) has been brought to light [50, 51, 52, 53]. While the exact role of the protein in the disease remains debated, reduced intestinal expression of *PARK7* was observed in patients and mouse models for IBD [53]. Moreover, *Park7* knockout mice were shown to have increased levels of pro-colitis bacterial species in their microbiome [54, 52] and experience aggravated symptoms of experimental-induced colitis [53]. In line with these observations, DNAm of the *PARK7* promoter probe cg10385390 (chr1:8'022'505) decreased both *PARK7* transcript ($\hat{\alpha}_{EM,T}$ = -0.675, P = 2.7e-4) and protein ($\hat{\alpha}_{EM,P}$ = -0.193, P = 2.0e-3) expression (Figure 6A). High transcript ($\hat{\alpha}_{MY,T}$ = -0.131, P = 1.7e-7) and protein ($\hat{\alpha}_{EM,P}$ = -0.193, P = 0.31) levels decrease IBD risk, resulting in an overall increased IBD risk upon DNAm ($\hat{\theta}_T$ = 0.114, P = 8.2e-9). Interestingly, early GWAS identified the region as a susceptibility locus for IBD, listing *TNFRSF9* as the top candidate gene [55, 56], thereby exemplifying how the integration of multiple omics layers can help to identify further causal genes.

Despite often being associated with decreased expression [45], our data provides examples of methylation boosting expression. For instance, DNAm of cg13428477 (chr3:122'748'086) increased *PDIA5* expression ($\hat{\alpha}_{EM,T}$ = 0.333, P = 7.3e-11; $\hat{\alpha}_{EM,P}$ = 0.931, P = 7.1e-63), whose levels subsequently increased platelet count ($\hat{\alpha}_{MY,T}$ = 0.062, P = 0.018; $\hat{\alpha}_{MY,P}$ = 0.058, P = 3.8e-24), so that DNAm resulted in increased platelet count ($\hat{\theta}_T$ = 0.056, P = 1.3e-43) (Figure 6B). Association between the *PDIA5* locus and platelet count was reported through GWAS [57]. Platelets are small cell fragments produced by megakaryocytes, which themselves are derived from hematopoietic stem cells. Accordingly, *PDIA5* has a binding site for the hematopoietic stem and progenitor cell TF MEIS1 [58] and is overexpressed in megakaryocytes as compared to other blood cell types [59]. Further studies showed that *pdia5* protein knockdown in zebrafish resulted in strongly decreased platelet count [60], matching our findings and confirming the role of *PDIA5* in thrombopoiesis. Additional putative regulatory mechanisms of DNAm-to-complex traits through transcript and protein levels are shown in Table S8-9, respectively, and were selected based on $|\hat{\theta}_T| > 0.02$ and $\widehat{MP} > 0.2$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

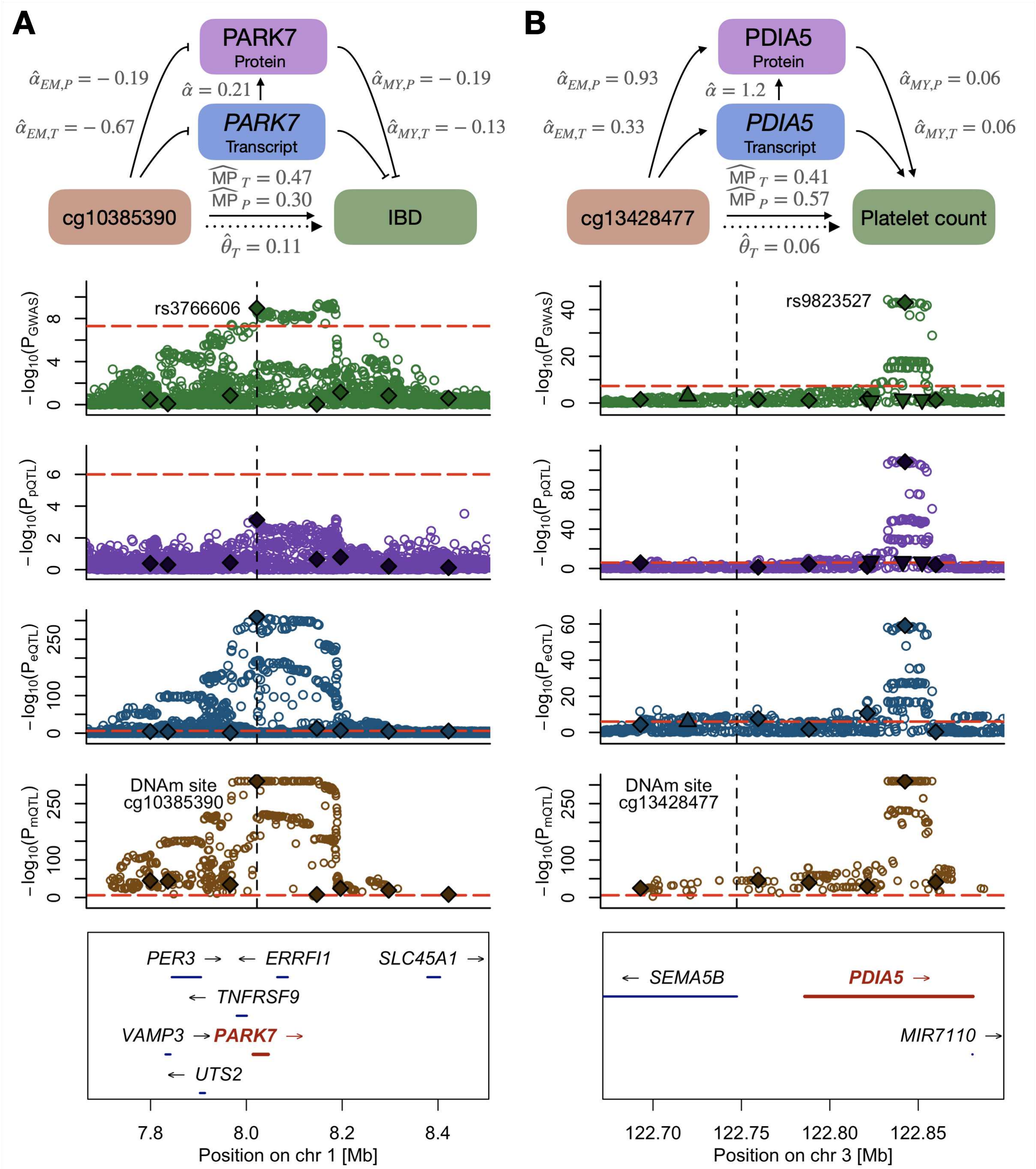

Figure 6: Plausible DNAm-transcript/protein-trait regulatory mechanisms between **A)** *PARK7* and irritable bowel disease (IBD) and **B)** *PDIA5* and platelet count. The top row displays a schematic of the mechanism with calculated univariable and multivariable MR effects. The four following rows show the regional SNP associations (-$\log_{10}$(p-values)) with the trait (green), encoded protein (purple), transcript (blue) and DNAm (brown) probe, respectively. Solid diamonds represent DNAm-associated instruments used in the univariable (for $\hat{\theta}_T$ calculation) and multivariable (for $\hat{\theta}_D$ calculation) MR analyses. Upwards and downwards pointing triangles are transcript- and protein-associated SNPs, respectively, that were additionally included in the MVMR instrument set. Red dashed lines indicate the significance thresholds of the respective SNP associations and the vertical black dashed line represents the DNAm probe position. Bottom row illustrates the positions and strand direction of the genes in the locus.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

While the aforementioned DNAm → gene expression → trait mechanisms were supported by both differential transcript and protein levels, other examples for which protein expression could not be assessed due to lack of pQTL data still reflect highly plausible mechanisms. For instance, we observed that DNAm of cg09070378 (chr1:161'183'762) decreased asthma risk ($\hat{\theta}_T$ = -0.031, P = 8.1e-11) by reducing *FCER1G* expression ($\hat{\alpha}_{EM,T}$ = -1.0, P = 3.5e-18), a gene listed in the KEGG pathway for asthma (hsa05310) and whose expression associated with an increased risk for asthma ($\hat{\alpha}_{MY,T}$ = 0.019, P = 3e-12) (Figure S21). The *FCER1G* promoter was found to be hypomethylated in patients with atopic dermatitis, with DNAm levels correlating negatively with the gene's expression [61], suggesting a broad role of *FCER1G* in allergic disorders. Our data also supports and provides a mechanistical explanation for the recent finding that reduced *IFNAR2* expression causally decreases the odds of severe coronavirus disease 2019 (COVID-19) [62, 63], which was later supported by the increased susceptibility for severe COVID-19 in individuals with rare loss-of-function mutations in *IFNAR2* [64]. Indeed, we found that DNAm of the *IFNAR2* promoter probe cg13208562 (chr21:34'603'264) decreased the gene's expression ($\hat{\alpha}_{EM,T}$ = -0.446, P = 2.4e-19) (Figure S22). As *IFNAR2* expression protects against hospitalization following COVID-19 infection ($\hat{\alpha}_{MY,T}$ = -0.090, P = 4.2e-6), DNAm of the locus increased the risk of severe infection ($\hat{\theta}_T$ = 0.064, P = 8.5e-13).

Transcript-to-complex traits mechanisms of action

Next, we focused on the results of the transcript-to-complex traits analysis to identify examples of transcriptome changes that mediate their phenotypic effect through the proteome. As a first example, we focused on *MANBA* (Figure 7A). After establishing that variants decreasing the gene's expression colocalized with risk variants for chronic kidney disease [65], the deleterious impact of decreased *MANBA* expression on renal health was recently confirmed in both humans with common expression-altering or rare loss-of-function variants, as well as *Manba* knockout mice [66]. Accordingly, we found that increased *MANBA* transcript had a beneficial impact on kidney damage biomarkers, as it decreased serum urea ($\hat{\theta}_T$ = -0.066, P = 1.2e-5) and cystatin C ($\hat{\theta}_T$ = -0.103, P = 1.6e-12), while increasing estimated glomerular filtration rate (eGFR; $\hat{\theta}_T$ = 0.052, P = 1.4e-6). Importantly, our data shows that these effects are mediated ($\widehat{\text{MP}}_{\text{urea}}$ = 120% ; $\widehat{\text{MP}}_{\text{cystatin C}}$ = 100% ; $\widehat{\text{MP}}_{\text{eGFR}}$ = 76%) through increased MANBA protein levels ($\hat{\alpha}_{EM,P}$ = 1.0, P = 5.2e-10), which in turn affected the aforementioned traits (serum urea: $\hat{\alpha}_{MY,P}$ = -0.022, P = 3.2e-5; cystatin C: $\hat{\alpha}_{MY,P}$ = -0.043, P = 1.9e-9; eGFR: $\hat{\alpha}_{MY,P}$ = 0.019, P = 6.3e-9). Furthermore, transcript levels of 3 pseudogenes overlapping the first intron of *MANBA* (RP11-10L12.1 (ENSG00000251288), KRT8P46 (ENSG00000248971); LRRC37A15P (ENSG00000230069)), as well as levels of the adjacent *UBE2D3* antisense RNA RP11-10L12.4 (ENSG00000246560), mediated their phenotypic impact on alkaline phosphatase, cystatin C, diastolic blood pressure, eGFR, and serum urea through decreased MANBA protein levels (Table S7). Overall, this suggests a complex gene-to-phenotype regulation of *MANBA* influenced by nearby non-coding elements.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

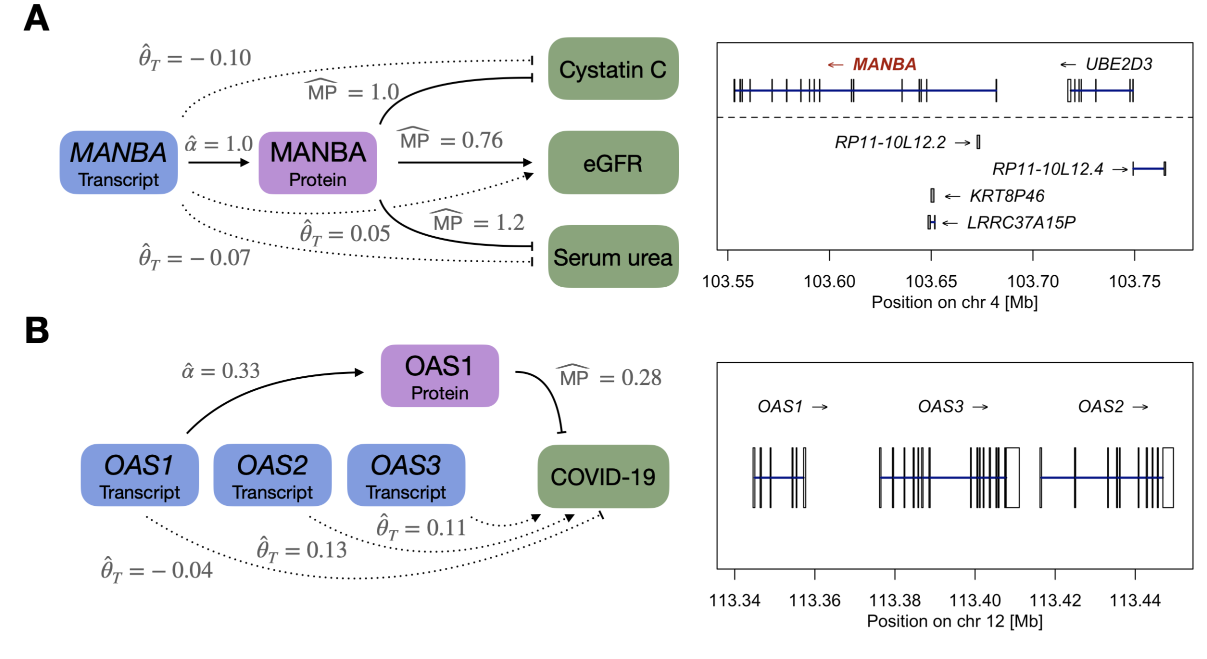

Figure 7: Plausible transcript-protein-trait regulatory mechanisms. **A)** Left: Impact of differential *MANBA* expression on kidney biomarkers through the regulation of its encoded protein. Annotated are the total effect $\hat{\theta}_T$ of the *MANBA* transcript levels on the respective outcomes, as well as $\widehat{\text{MP}}$s through the encoded protein. Right: Zoom on the *MANBA* region; transcripts below the dashed lines are non-coding and putative negative regulators of MANBA protein levels. **B)** Scheme and locus zoom of the effect of *OAS1/OAS2/OAS3* transcript levels on severe COVID-19 disease. Mediation through the encoded protein could only be tested for *OAS1*.

In contrast, non-coding elements were also found to exert their phenotypic effects through distantly encoded proteins, as illustrated by the transcript originating from the U6 small nuclear RNA 516 pseudogene ENSG00000223313 on chromosome 15, which decreased insulin-like growth factor 1 levels (IGF-1; $\hat{\theta}_T$ = -0.029, P = 4.0e-7) by decreasing the protein levels of IGF binding protein 3 (IGFBP3; $\hat{\alpha}_{EM,P}$ = -0.154, P = 7.6e-4), a well-known regulator of IGF-1's bioavailability and half-life [67] encoded on chromosome 7 ($\hat{\alpha}_{MY,P}$ = 0.115, P = 4.0e-7). Alternatively, we observed several cases of protein-coding transcripts affecting traits through proteins in *trans*. For instance, transcript levels of *SUOX*, encoding for a mitochondrial sulfite oxidase, increased lymphocyte count ($\hat{\theta}_T$ = 0.027, P = 4.7e-5) by positively affecting tyrosylprotein sulfotransferase 2 protein levels (TPST2; $\hat{\alpha}_{EM,P}$ = 0.206, P = 4.0e-4). In turn, TPST2 increased lymphocyte count ($\hat{\alpha}_{MY,P}$ = 0.029, P = 0.02). Both *SUOX* and *TPST2* belong to the KEGG sulfur metabolism pathway (hsa00920). Sulfite oxidase catalyzes the oxidation of sulfite to sulfate [68]. In contrast, sulfate is used by PAPSS1/PAPSS2 to generate 3'-phosphoadenosine-5'-phosphosulfate (PAPS) [69], the main cosubstrate of the sulfotransferase reactions catalyzed by TPST2 [70]. Sulfation of chemokine receptors, which play a critical role in immune function and are widely expressed on lymphocytes, can modulate a receptor's affinity and/or selectivity for cognate chemokines, as well as mediate pathogen entry [71], establishing the importance of sulfur metabolism for lymphocyte

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

function. Another immune-related example involves the recently established link between the interferon-induced antiviral OAS gene cluster (*OAS1*, *OAS2*, *OAS3*) and severe COVID-19 [62, 63]. In line with reports highlighting the protective effect of a Neandertal haplotype associating with increased OAS1 [72, 73], we found that the protective effect against COVID-19 of increased *OAS1* transcript levels ($\hat{\theta}_T$ = -0.038, P = 6.9e-8) was mediated ($\widehat{\text{MP}}$ = 28%) by increased levels of the encoded protein ($\hat{\alpha}_{EM,P}$ = 0.334, P = 2.0e-23) (Figure 7B). Of note, while protein levels were only available for OAS1, our MR analysis indicated that adjacent and related transcripts *OAS3* ($\hat{\theta}_T$ = 0.105, P = 6.8e-8) and *OAS2* ($\hat{\theta}_T$ = 0.133, P = 5.2e-3) exerted opposite effects on COVID-19 severity. The opposite effect of *OAS1* and *OAS3* on the outcome reflect previous findings [73] and highlight the complex role of the locus in mediating immunity. Further putative regulatory mechanisms of transcript-to-complex traits through protein levels are shown in Table S10 and were selected based on $|\hat{\theta}_T| > 0.02$, $\widehat{MP} > 0.1$ and $\text{P}_{\text{MY,k}} < 0.05$. Taken together, these examples illustrate how both protein-coding and non-coding transcripts can exert phenotypic changes through modulation of encoded, as well as *trans* protein levels, suggesting new biological mechanisms.

Discussion

We presented a framework to quantify mediation of complex trait-impacting effects through multiple omics layers, unravelling nuanced patterns in gene and protein expression regulation. First, we assessed the extent to which DNAm-to-trait effects were mediated by *cis*-transcripts and compared this proportion to the mediation through *cis*-proteins. Evaluating 50 complex traits, the overall adjusted $\widehat{\text{MP}}$ (*i.e.*including DNAm-trait pairs with testable mediators in *cis* not under DNAm regulation) through *cis*-transcripts and *cis*-proteins was estimated to be 28.3% and 1.2%, respectively. Simulation studies indicated that the lower sample size of the pQTL dataset ($\text{N}_{\text{pQTL}} \approx$ 3,300 vs $\text{N}_{\text{eQTL}} \approx$ 30,000) was estimated to result in a relative decrease of 20% in MP, in line with the fact that exposures/mediators with more precise genetic effect estimates are prioritized by MVMR regression models [43]. Despite the fact that $\approx$6.8x lower number of proteins present in the pQTL dataset (*i.e.* fewer testable indirect pathways) than transcripts in the eQTL data, it was not the main reason for the striking difference in the MPs. We demonstrated this by repeating the analysis on a common set of 2,145 transcripts and their encoded proteins, where the adjusted MP through proteins was still $\approx$10x lower than through transcripts (8.15% vs 0.85%). We suspect that this difference was mainly due to the fact that, on average, proteins were four times less likely to be causally linked to the investigated DNAm site than transcripts, suggesting a tighter link between DNAm and transcript expression than between DNAm and protein levels. This implies a moderate similarity between eQTLs and pQTLs which we confirmed when testing for causal effects between transcript and encoded protein levels: The fraction of testable transcripts linked to their respective protein (when available) at a nominal significance threshold was found to be only 22%. While some of the missing links might be due to the lack of statistical power, it indicates that the transcript to

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

protein regulation is more nuanced than the central dogma of biology would imply, whereby a straightforward translation from transcripts to proteins by ribosomes is assumed. As a consequence of these weak transcript-to-protein effects, the mediation of transcript-to-trait effects through the encoded protein yielded relatively low MPs (mean = 5.1%). Previous studies reported discrepancies in transcript and protein abundances with explained variances of protein levels by transcript levels ranging from 40 to 85% [74, 75], as well as in eQTL and pQTL co-analyses where only 12 to 40% of the signals were found to be shared [19, 21]. Mechanisms explaining why protein abundance cannot be entirely predicted from transcript levels include protein synthesis delay, transport, degradation, post-transcriptional changes, but also technical variation attributable to measurement instruments [74, 75].

Noteworthily, MR analyses provide directions of estimated causal effects, and two, rather counter-intuitive, observations were made: i) 46.6% of significant DNAm-to-transcript effects were of positive sign (*i.e.* DNAm increases transcription) and ii) 20% of significant transcript-to-protein effects were of negative sign (*i.e.* high transcript levels decrease protein levels). The first observation is in line with previous genome-wide methylation and gene expression association studies which reported high fractions of positive correlations (30-35%) [48, 46]. While poorly understood [47], several mechanisms have been proposed to explain the phenomenon: preferential binding of some transcription factors to methylated DNA [76, 77], prevention of repressor binding indirectly leading to increased expression through looping DNA [78, 35], or DNAm in the gene body provoking elongation efficiency and preventing spurious initiation of transcription [79]. As to the negative transcript-to-protein effects, which were consistent in the direction when computed with either the INTERVAL or SCALLOP pQTL datasets, literature is more sparse. While negatively correlating gene products have been reported previously [80, 81], this has, to the best of our knowledge, not yet been studied in the context of QTL analyses and remains the topic of future investigations. Finally, MP estimates indicate that DNAm sites typically regulate multiple transcripts in *cis*. Average MPs of 37% suggest that phenotypic DNAm effects are largely mediated through pathways other than local gene expression regulation, especially when the DNAm site is located further away from the TSS of the main transcript mediator. Collectively, these results describe a more diverse picture of the transcription and translation machinery, challenging the classical views of DNAm solely reducing gene expression, and this in the TF region, as well as mRNA levels being a good proxy for protein abundance.

Mapping genetic variants identified in GWAS analyses to biological processes is notoriously difficult [2]. However, systems genetics approaches that integrate multiple omics datasets as a way of leveraging GWAS summary data have proven successful in providing a more complete picture of the path from genotype to phenotype [82]. Here, we demonstrated that our multi-omics framework was able to attribute GWAS signals to biological pathways in loci harbouring multiple genes (e.g. *PARK7*-IBD and

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FCERG1-asthma). A challenge in identifying causal chains through omics layers is the attenuation in the genetic association strengths when moving up along layers. In a linear model, the genetic effect on the phenotype is assumed to be the product of causal effects between the preceding layers and it was previously shown that the variance explained by the top associated QTL of the first layer decreases with each successive omics layer [35]. In line with this observation, the biological examples depicted in Figure 6 visualize the decrease in the genetic associations from the DNAm to the complex trait level. Importantly, integration of both eQTL and pQTL data represent orthogonal approaches in corroborating mediators of DNAm-to-trait or transcript-to-trait effects. Current pQTL datasets lack the sample size and number of proteins to systematically validate regulatory mechanisms found through eQTL integration (e.g. *OAS1*/*OAS2*/*OAS3*-COVID-19). In the future, we expect larger datasets to become available and here presented a proof of concept of how protein-level data can either support mechanistic findings resulting from transcript data or warrant future investigations leading to the discovery of potential new mechanisms of action, implicating other genes.

Throughout the manuscript, we highlighted multiple putative molecular mechanisms of action supported by high MPs through intermediate omics layer and strong literature evidence. More examples can be found in Tables S8-10, including some for which the putative mechanism of action remain strongly debated. For instance, our analyses implicated a DNAm site (cg15133208: chr4:90'757'351) in the TSS region of *SNCA* in Parkinson's disease (PD) (Figure S23). Many studies have investigated mechanisms involving DNAm, *SNCA* and PD, resulting in conflicting results as to the effect directions. Our results suggest a protective effect of that DNAm site on PD. While supported by studies in the field [83], the assumed DNAm effect on *SNCA* expression is different from our estimated MR effect. Both *SNCA* transcript and SNCA protein levels were estimated to be upregulated in the hypermethylated DNA state, with high *SNCA* levels calculated to decrease PD risk. It is generally assumed that increased *SNCA* expression contribute to PD pathogenesis [84], although blood and brain-specific *SNCA* expression pattern, as well as different isoforms, have been reported to correlate differently with PD [85]. A recent study showed positive correlations between *SNCA* levels and both PD and the related synucleinopathy of Lewy body dementia (LBD) in the temporal cortex, but negative and non-significant ones for LBD and PD in blood, respectively [85]. Another recent GWAS with integrative brain eQTL follow-up analyses indicated that high levels of *SNCA-AS1*, which regulates *SNCA* expression levels, might be protective against LBD [86], suggesting complex regulatory mechanisms governing the locus. Similarly, mechanisms involving proteins in *trans* mediating transcripts-to-trait effects were less straightforward to interpret. Several examples involved non-coding RNA for which functional information is sparse, complicating literature validation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

While our method highlights candidate pathways, several limitations have to be considered. First, like all MR-inverse variance weighting (IVW) analyses, our MR analyses assumed all genetic variants to be valid IVs. We applied Steiger filtering to mitigate the inclusion of pleiotropic IVs that violate independence of the outcome conditional on the exposure and mediators, as well as independence of the mediators conditional on the exposure in the case of variants associated with both the exposure and mediators (third MR assumption; Methods [37]). However, the presence of invalid IVs cannot be excluded and could therefore compromise causal effect estimates [40, 87]. In particular, since selected MR IVs are all in *cis* of the investigated molecular trait, they might be based on a single (pleiotropic) haplotype signal. Conversely, one might argue that the Steiger filter is too stringent if the reverse effect from the mediator on the exposure is biologically unlikely, so that it excludes IVs potentially important in accurately estimating causal effect sizes. Second, we select mediators based on their association to the exposure without taking into account their mediator potential, *i.e.* whether or not the mediator is additionally causally linked to the trait. Phrased differently, the selected mediators are simply candidates and such selection serves as a first filter to remove non-mediators. In line with our simulations, it has been shown that extremely large number of such mediator candidates that are not true mediators (92 candidates in total with 88 of them being false mediators) can cause MVMR regression models to fail [43], indicating that our framework is less suitable for large numbers of molecular mediators, unless the selection threshold P_{EM} is made more stringent. Third, our mediation model cannot completely exclude the possibility of reverse effects from the mediator(s) on the exposure. This concern especially applies when considering DNAm as exposure and *cis*-transcripts as mediator(s), since differential transcript levels have been suggested to modulate DNAm levels [35]. We use the largest publicly available mQTL dataset, however, it misses genetic effect sizes of the entire *cis*-region, which would be required to test for reverse or bi-directional effects of transcripts on DNAm. Fourth, with the exception of pQTLs [19], large-scale *trans*-QTL datasets are still lacking, prohibiting genome-wide assessment of mediation and restricting many analyses to *cis*-mediation. Finally, while molecular mechanisms ought to be tissue- or even cell type-specific, QTL data used in this study were all derived from whole blood. It is known that different tissues express different isoforms [88], with many splicing and expression QTLs shown to differ across tissues [89]. Accordingly, MPs for blood biomarkers were generally higher than those for diseases, for which blood might not be the most relevant tissue. Alternatively, this differences might also be due to the fact that indirect pathways, through unmeasured mediators, play a greater role for diseases than for biomarkers. Once tissue-stratified multi-omics datasets of larger sample size become available, more accurate, and potentially higher MPs will be obtained in trait-relevant tissues.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conclusion

We quantified the causal connectivity between three omics layers - DNAm, transcript and protein abundance - and their importance in shaping complex traits. We examined regulatory effects of DNAm on gene expression - assessed through both the transcriptome and proteome - which in its complementary use allowed for robust causal inference between molecular and complex traits. Overall, the results indicated that regulatory mechanisms can be more nuanced and complex than suggested by the central dogma of biology, leaving many open questions as to alternative transcription and translation processes. Our integrative omics framework can be extended to other omics-GWAS combinations using the software made available (`https://github.com/masadler/smrivw`), and provide a powerful tool for mapping GWAS signals to biological pathways and prioritizing functional follow-up experiments.

Methods

Univariable and multivariable Mendelian randomization

Univariable Mendelian randomization (MR) was applied to estimate the total causal effect (θ_T) and multivariable MR (MVMR) to estimate the direct causal effect (θ_D) of an exposure E on an outcome Y. The mediation proportion (MP) was defined as $1 - \theta_D/\theta_T$. Under the MR assumptions, genetic variants G used as instrumental variables (IVs) must be i) associated with E, ii) independent of any confounder of the E $-$ Y relationship, iii) conditionally independent of Y given E. Independent IVs ($r^2 <$ 0.05) associated with the molecular exposure (P $<$ 1e-6) and located in *cis* ($<$ 1 Mb) allowed the estimation of θ_T using an inverse-variance weighted (IVW) method assuming equal weights given the standardization of the data and accounting for correlated instruments [90]:

$$\hat{\theta}_T = (\boldsymbol{\beta}_E' \mathbf{C}^{-1} \boldsymbol{\beta}_E)^{-1} \boldsymbol{\beta}_E' \mathbf{C}^{-1} \boldsymbol{\beta}_Y \tag{1}$$

where $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_Y$ are vectors of genetic effect sizes obtained from summary statistics for E and Y, respectively. $\mathbf{C}$ is the linkage disequilibrium (LD) matrix with pairwise correlations between IVs estimated from the UK10K reference panel [91]. Prior to the causal effect calculation, IVs were filtered to fulfill the MR Steiger criterion of no larger Y than E genetic effects [37] and were thus required to pass a threshold $t_{rev} < \frac{|\beta_{E_i}|-|\beta_{Y_i}|}{\sqrt{var(\beta_{E_i})+var(\beta_{Y_i})}}$ with t_{rev} set at -2, equivalent to a one sided test p-value threshold of 0.023. IVs not passing this threshold are prone to violating the third MR assumption of horizontal pleiotropy since they are more directly linked to the outcome. As a result MR estimates including such IVs would potentially mix up forward and reverse causal effects. The standard error (SE) of θ_T can be approximated by the Delta method [92]:

$$\text{SE}(\hat{\theta}_T) = \sqrt{(\boldsymbol{\beta}_E' \mathbf{C}^{-1} \boldsymbol{\beta}_E)^{-1} \boldsymbol{\beta}_E' \mathbf{C}^{-1/2} \boldsymbol{\Sigma} \mathbf{C}^{-1/2} \boldsymbol{\beta}_E (\boldsymbol{\beta}_E' \mathbf{C}^{-1} \boldsymbol{\beta}_E)^{-1}} \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\mathbf{\Sigma}$ is a diagonal matrix with each diagonal element *i* equalling the maximum of the regression variance s^2 and $var(\beta_{Y_i})$ [93].

Through the inclusion of mediators M_k and their associated *cis* genetic variants ($r^2 < 0.05$, $\text{P} < \text{1e-6}$), θ_D can be estimated analogously to θ_T using a multivariable regression model [41] as the first element of $\boldsymbol{\theta_D}$:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{D}} = (\boldsymbol{B}'\mathbf{C}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\mathbf{C}^{-1}\boldsymbol{\beta}_Y \tag{3}$$

where $\boldsymbol{B}$ is a matrix with $k + 1$ columns containing the effect sizes of the IVs on the exposure in the first column and on each mediator in the subsequent columns. The remaining elements of $\boldsymbol{\theta_D}$ represent the direct effects of the mediators on the outcome and were referred to as $\alpha_{MY,k}$. In the estimation of MPs, we were not interested in $\alpha_{MY,k}$ values *per se*, but we took these effect sizes into account for inferring molecular mechanisms. If the number of mediator-associated instruments was sufficient ($\geq$ 3) to conduct a univariable MR from the mediator on the outcome, we estimated $\alpha_{MY,k}$ from this analysis instead, since computed on a single regressor, narrower CIs are obtained.

This MVMR model does not allow for the presence of a causal effect from the mediators on the outcome via the exposure, and we therefore conducted several Steiger filtering steps on the IVs. In addition to meeting the Steiger criterion described above, exposure-associated IVs were required to pass that same threshold t_{rev} of no larger mediator than exposure effects for each of the mediators M_k. Similarly, to mitigate reverse causal effects from the outcome on the mediators, mediator-associated instruments with larger Y than M effects were removed if not passing the t_{rev} threshold. The SE of $\hat{\theta}_D$ was derived analogously to the univariable form as shown in [29].

Omics and trait summary statistics

We used mQTL data from the GoDMC consortium (N = 32,851) [16], which contains $>$ 170,000 whole blood DNAm sites with at least one significant *cis*-mQTL (P $<$ 1e-6, $<$ 1 Mb from the DNAm site, N $>$ 5,000). *Cis*-eQTL data were taken from the eQTLGen consortium (N = 31,684) [18] which includes *cis*-eQTLs ($<$ 1 Mb from gene center, 2-cohort filter) for 19,250 transcripts (16,934 with at least one significant *cis*-eQTL at FDR $<$ 0.05 corresponding to P $<$ 1.8e-05). *Cis*- and *trans*-pQTL data were from the INTERVAL study (N = 3,301) [19]. SomaLogic aptamers (N_SOMAmers = 3,283) quantified the levels of 2,977 proteins and complexes with a UniProt ID. After removing protein complexes (N_SOMAmers = 42), sex chromosome encoded proteins (N_SOMAmers = 113), and UniProt IDs that could not be mapped to an EnsemblID (N_SOMAmers = 6), 2,838 proteins remained of which 696 had at least one significant *cis*-pQTL (P $<$ 1e-6, $<$ 1 Mb from the protein-encoding gene center, N $>$ 2,000). If two SOMAmers mapped to the same protein, the one with the strongest transcript-to-protein causal MR effect was retained (see omics-to-omics MR analysis). If the transcript was not available in the eQTLGen dataset or did not have any significant IVs, the SOMAmer with the highest number of significant *cis*-pQTLs (or *trans*-pQTLs if no *cis*-

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pQTLs were present) was chosen. Mapping from UniProt to Ensembl identifiers was done through the UniProt REST API [94] and genomic coordinates were retrieved from the Ensembl REST API (GRCh37 build) [95]. Exact mapping of SOMAmer-UniProt-Ensembl identifiers is provided in Table S5. A total of 2,145 transcript-encoded protein pairs were present in both the eQTL and pQTL datasets.

GWAS summary statistics for outcome traits came from the largest ($\mathrm{N}_{\mathrm{average}} > 320{,}000$), predominantly European-descent, publicly available studies, as listed in Table S1.

Prior to each mediation analysis, exposure and mediator omics, GWAS and the reference panel data were harmonized. The analysis was conducted on autosomal chromosomes, and palindromic single nucleotide variants (SNPs), as well as SNPs with an allele frequency difference > 0.05 between any pairs of datasets were removed. If allele frequencies were not reported by the GWAS summary statistics, allele frequencies from the UK Biobank were used. Z-scores of summary statistics (molecular and outcome GWAS) were standardized by the square root of the sample size to be on the same SD scale.

DNAm-to-trait mediation analysis

First, univariable MRs were conducted to estimate the total causal effect $\hat{\theta}_T$ of the DNAm sites on each trait, assessing $\sim$50,000 DNAm probes with $\geq$ 5 independent mQTLs after harmonization of the datasets ($r^2 < 0.05$). DNAm probes significantly associated to the outcome ($\mathrm{P}_\mathrm{T} < 0.05/50000 = 1\mathrm{e}{-6}$) were clumped based on the p-value of the total causal effect $\hat{\theta}_T$, P_T (distance-pruning at 1 Mb), to be independent of each other.

Second, MVMR analyses were performed to estimate the direct effect $\hat{\theta}_D$. Potential transcript mediators in *cis* of the DNAm exposure probe (± 500kb) were extracted and causal effects $\alpha_{EM,k}$ of the DNAm probe on these transcripts were assessed by univariable MR. Transcripts satisfying $\mathrm{P}_{\mathrm{EM,k}} < \mathrm{P}_{\mathrm{EM}}$ (default $\mathrm{P}_{\mathrm{EM}} = 0.01$, with 0.05 and 1e-3 being tested as well) were included as mediators, as well as their associated SNPs as additional instruments. Steiger filtering was applied as described previously and IVs were clumped based on a rank score determined as follows: 1) for each mediator, IVs were ranked according to their association p-value to the mediator and assigned an integer score, 2) for each IV, a final score was calculated as the sum of its individual mediator scores. Following the establishment of the $\boldsymbol{B}$ effect size matrix, $\hat{\theta}_D$ was calculated, as well as $\hat{\theta}_{D,top}$ which was estimated from a MVMR model that includes a single mediator, namely the transcript with the lowest $\mathrm{P}_{\mathrm{EM,k}}$. If no transcript causally associated to the DNAm probe, mediation is not detectable, hence $\hat{\theta}_D$ was set to $\hat{\theta}_T$ for that probe (inclusion of such probes in MP calculation was termed “adjusted mediation proportion”). As the Steiger filter removed exposure-associated instruments with larger mediator than exposure effects (see “Univariable and mul-

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tivariable Mendelian randomization"), the number of initial exposure-associated instruments ($m_E \geq 5$) could decrease. Therefore, to avoid scenarios of reverse causality where the mediator exerts an effect on the outcome through the exposure, we required ≥ 3 exposure-associated IVs.

We additionally conducted mediation analyses on independent mediators. To this end, selected mediators (those that passed P_{EM}) were clumped at various correlation thresholds R_{med} (default $\text{R}_{\text{med}} <$ 0.3, with 0.2 and 0.1 being tested as well). Correlations among the mediators were calculated based on QTL effect sizes of independent exposure and mediators IVs and priority was given to the mediator with the lowest $\text{P}_{\text{EM,k}}$.

The mediation proportion (MP) was calculated by regressing $\hat{\theta}_D$ on $\hat{\theta}_T$ to estimate for the unmediated proportion, $\hat{\gamma}$, which after correcting for regression dilution bias (Equation 4):

$$\hat{\gamma}_{cor} = \frac{\hat{\gamma}}{\sqrt{1 - \frac{\sum \text{se}^2(\hat{\theta}_T)}{\sum \hat{\theta}_T^2}}} \tag{4}$$

yielded $\widehat{\text{MP}} = 1 - \hat{\gamma}_{cor}$ for a defined set of DNAm-trait pairs. MVMR analyses were repeated on the selected DNAm-trait pairs through proteins in *cis* following the same mediator and IV filtering steps as described above.

Transcript-to-trait mediation analysis

MPs for transcript-to-trait mediation analyses were calculated similarly to DNAm-to-trait MPs. Briefly, we first computed total causal effects $\hat{\theta}_T$ of transcripts on traits for $\sim$ 11,000 transcripts with $\geq$ 5 independent ($r^2 <$ 0.05) and significant eQTLs (P $<$ 1e-6), $\sim$ 1,200 of which had an encoded protein in the pQTL dataset. For each trait, significant transcripts ($\text{P}_{\text{T}} <$ 0.05/1,000 = 5e-5) were selected. Second, MVMR analyses were conducted, where for each transcript, mediators were defined as i) the encoded protein or ii) the encoded protein plus any other protein in *trans* among a set of 696 proteins with $\geq$ 1 significant (P $<$ 1e-6) pQTL that satisfied $\text{P}_{\text{EM,k}} < \text{P}_{\text{EM}}$ (default P_{EM} = 1e-3, with 1e-2 and 1e-4 being tested as well). If more than 10 proteins satisfied the condition, the ten most strongly associated were retained. Associated pQTLs were included as IVs and following Steiger filtering, instruments were pruned as described in the DNAm-to-trait mediation analysis section. Effect sizes of mediator-associated IVs that were not significant (P $>$ 1e-6) for a given mediator were shrunk to 0 [96]. Direct effects $\hat{\theta}_D$ were calculated using encoded proteins (if available) as mediators in addition to selected *trans* proteins. Additionally, direct effects were calculated using only encoded proteins as mediators. Finally, $\widehat{\text{MP}}$s were calculated by aggregating all transcript-trait pairs as specified in each sub-analysis, and regressing $\hat{\theta}_D$ on $\hat{\theta}_T$ while accounting for regression dilution bias (Equation 4).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Omics-to-omics MR analysis

MR causal effects between two molecular traits were calculated following the same procedure than in the univariable MR to calculate total effects $\hat{\theta}_T$. First, independent ($r^2 <$ 0.05) and significant (P $<$ 1e-6) exposure IVs were selected and IVs not passing the aforementioned Steiger filter were discarded. MR causal effects were then computed based on Equation 1.

DNAm-to-transcript MR analysis

MR effects between DNAm sites and transcripts in *cis* ($\pm$ 500kb) with $\geq$ 3 exposure IVs were calculated. Pearson correlation coefficient with previously reported DNAm-transcript correlations [48] was calculated on common DNAm-transcript pairs. DNAm probe annotations with respect to the assessed transcript were from the IlluminaHumanMethylation450kanno.ilmn12.hg19 R package [97].

Transcript-to-encoded-protein MR analysis

On the common transcript-encoded protein pairs, causal effects were calculated for transcripts with $\geq$ 3 independent eQTLs (r^2 $<$ 0.05). When comparing causal effects obtained from the INTERVAL and SCALLOP pQTL dataset, we additionally included transcripts with a single eQTL.

Simulation studies

We conducted simulation studies to assess the robustness of our model and to identify sources of bias in the estimated MP. Two simulation settings were set up: one replicating the DNAm-to-trait via transcripts in *cis* mediation analysis and one replicating the transcript-to-trait via proteins in *trans* mediation analysis. Both scenarios were simulated under the same model, but with different parameter settings (Figure S1, Table S2).

We considered an exposure with heritability h^2_E and m_E independent IVs. Effect sizes β^E_i for m_E IVs were drawn from a normal distribution $\beta^E_i \sim \mathcal{N}(0, \sqrt{h^2_E/m_E})$ and rescaled to total h^2_E. N_{med} potential mediators were simulated, among which $N_{med,sig}$ were contributing to the indirect effect θ_M. Each mediator k associated with m_M IVs with direct effects $\beta^{M_k}_{direct,i} \sim \mathcal{N}(0, \sqrt{h^2_{M,direct,k}/m_M})$ rescaled to $h^2_{M,direct,k}$, the direct heritability of the mediator that does not take into account the additional heritability coming through the exposure. Direct heritabilities were sampled from a uniform distribution $h^2_{M,direct,k} \sim U(h^2_{M,low}, h^2_{M,high})$. Causal effects of the exposure on the mediator ($\alpha_{EM,k}$) and of the mediator on the outcome ($\alpha_{MY,k}$) for $N_{med,sig}$ mediators were drawn from a bivariate normal distribution $\alpha_{EM,k}, \alpha_{MY,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ the covariance matrix:

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\mathbf{\Sigma} = \begin{bmatrix} var(\alpha_{EM}) & \rho \cdot \sqrt{var(\alpha^{EM}) \cdot var(\alpha_{MY})} \\ \rho \cdot \sqrt{var(\alpha_{EM}) \cdot var(\alpha_{MY})} & var(\alpha_{MY}) \end{bmatrix}$$

where ρ is the correlation between $\alpha_{EM,k}$ and $\alpha_{MY,k}$. For the remaining N_{med} - $N_{med,sig}$ mediators, $\alpha_{EM,k}$ and $\alpha_{MY,k}$ causal effects were set to zero. The vector of effect sizes $\boldsymbol{\beta}^{M_k}$ of size $m_E + N_{med} \cdot m_M$ for each mediator k was constructed to have effect sizes equalling $\beta_i^E \cdot \alpha_{EM,k}$ for m_E exposure SNPs and effect sizes equalling $\beta_{direct,i}^{M_k}$ for m_M mediator-associated SNPs. The effect sizes of remaining IVs associated to mediators $i \neq k$ were set to zero. Likewise, effect sizes of the $N_{med} \cdot m_M$ IVs on the exposure in the $\boldsymbol{\beta}^E$ vector were set to zero.

The indirect effect θ_M, direct effect θ_D and total effect θ_T were calculated as:

$$\theta_M = \sum_k \alpha_{EM,k} \cdot \alpha_{MY,k} \quad ; \quad \theta_D = \theta_M(\frac{1}{\text{MP}} - 1) \quad ; \quad \theta_T = \theta_D + \theta_M$$

These quantities allowed to design the outcome effect size vector $\boldsymbol{\beta}^Y$:

$$\boldsymbol{\beta}^Y = \theta_D \cdot \boldsymbol{\beta}^E + \sum_k \alpha_{MY,k} \cdot \boldsymbol{\beta}^{M_k}$$

For each scenario, we simulated 300 data sets to each time get $\boldsymbol{\beta}^E$, $\boldsymbol{\beta}^{M_k}$ and $\boldsymbol{\beta}^Y$. Normally distributed noise, as a function of the sample size N, $\epsilon_i^E \sim \mathcal{N}(0, 1/N_E)$, $\epsilon_i^M \sim \mathcal{N}(0, 1/N_M)$ and $\epsilon_i^Y \sim \mathcal{N}(0, 1/N_Y)$ was added to each simulated vector. To approximate our real data, exposure effect sizes of SNPs serving as mediator instruments were set to zero again. We then estimated for each model $\hat{\theta}_T$ and $\hat{\theta}_D$ by including mediators that satisfied P_{EM} (p-value of the causal effect from the exposure on the mediator). Causal effects $\hat{\theta}_D$ were regressed on $\hat{\theta}_T$ to estimate the coefficient $\hat{\gamma}$ which after accounting for regression dilution (Equation 4) allowed to obtain the estimated $\widehat{\text{MP}}$.

Comparing mediation proportions

To test the statistical significance between $\widehat{\text{MP}}$s estimated on two different sets of exposure-trait pairs (e.g. $\widehat{\text{MP}}$ of a given physiological category vs all categories combined) or on the same exposure-trait pairs, but with different parameter settings (e.g. changing P_{EM}), we make use of $\hat{\gamma}$ and its corresponding standard error $se(\hat{\gamma})$ obtained from regressing $\hat{\theta}_D$ on $\hat{\theta}_T$ (both of which being corrected for regression dilution (Equation 4)) to yield $\hat{\gamma}_{cor}$ and $se(\hat{\gamma}_{cor})$. We then perform a two-sided z-test based on the following test statistic:

$$\frac{\hat{\gamma}_{cor,1} - \hat{\gamma}_{cor,2}}{\sqrt{se(\hat{\gamma}_{cor,1})^2 + se(\hat{\gamma}_{cor,2})^2}} \sim \mathcal{N}(0,1) \tag{5}$$

A significant difference between $\widehat{\text{MP}}$s was claimed if the two-sided p-value was below 0.05.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Availability of data and materials

QTL datasets can be downloaded at the following websites: mQTLs (`http://mqtldb.godmc.org.uk/downloads`), eQTLs (`https://www.eqtlgen.org/cis-eqtls.html`), pQTLs (`http://www.phpc.cam.ac.uk/ceu/proteins/`. The list of GWAS summary statistics used is in Table S1, all of which are all from the public domain.

Software to conduct univariable MR-IVW (molecular trait → outcome, molecular trait 1 → molecular trait 2) and multivariable MR-IVW (molecular trait 1 → molecular trait 2 → outcome) can be found at `https://github.com/masadler/smrivw`. Source code (C++, released under GPL v3 license) and executable file (for Linux platforms, released under MIT license) are provided which rely on functionalities and the data management architecture of the SMR software (`https://cnsgenomics.com/software/smr` [35]). The provided documentation hosted on the github repository guides users in reproducing the mediation results and conducting univariable and multivariable MR on their own combinations of QTL and GWAS datasets.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

M.C.S., E.P. and Z.K. conceived and designed the study. M.C.S. performed statistical analyses. E.P. provided guidance on statistical analyses. Z.K. supervised all statistical analyses. All the authors contributed by providing advice on interpretation of results. C.A. contributed with the biological interpretation of the results. M.C.S., E.P. and Z.K. drafted the manuscript. C.A. contributed to the writing of specific sections. All authors read, approved, and provided feedback on the final manuscript.

Acknowledgements

LD was calculated based on the UK10K data resource (EGAD00001000740, EGAD00001000741). Computations we performed on the JURA cluster of the University of Lausanne. We would also like to thank Kaido Lepik for critical comments on the manuscript.

References

[1] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[2] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

[3] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[4] Jimmy Z Liu, Allan F Mcrae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, Nicholas K Hayward, Grant W Montgomery, Peter M Visscher, Nicholas G Martin, et al. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139–145, 2010.

[5] David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS computational biology*, 12(1):e1004714, 2016.

[6] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[7] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[8] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles, et al. Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, 46(D1):D661–D667, 2018.

[9] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

[10] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[11] Da Wei Huang, Brad T Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W Baseler, H Clifford Lane, et al. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, 35(suppl_2):W169–W175, 2007.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[12] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one*, 5(11):e13984, 2010.

[13] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[14] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

[15] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[16] Josine L Min, Gibran Hemani, Eilis Hannon, Koen F Dekkers, Juan Castillo-Fernandez, René Luijk, Elena Carnero-Montoro, Daniel J Lawson, Kimberley Burrows, Matthew Suderman, et al. Genomic and phenotypic insights from an atlas of genetic effects on dna methylation. *Nature genetics*, 53(9):1311–1321, 2021.

[17] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[18] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Harm Brugge, et al. Large-scale cis-and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics*, pages 1–11, 2021.

[19] Benjamin B Sun, Joseph C Maranville, James E Peters, David Stacey, James R Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, et al. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018.

[20] Chen Yao, George Chen, Ci Song, Joshua Keefe, Michael Mendelson, Tianxiao Huan, Benjamin B Sun, Annika Laser, Joseph C Maranville, Hongsheng Wu, et al. Genome-wide mapping of plasma protein qtls identifies putatively causal genes and pathways for cardiovascular disease. *Nature communications*, 9(1):1–11, 2018.

[21] Lasse Folkersen, Stefan Gustafsson, Qin Wang, Daniel Hvidberg Hansen, Åsa K Hedman, Andrew Schork, Karen Page, Daria V Zhernakova, Yang Wu, James Peters, et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature metabolism*, 2(10):1135–1148, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[22] So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, et al. An atlas of genetic influences on human blood metabolites. *Nature genetics*, 46(6):543–550, 2014.

[23] Luca A Lotta, Maik Pietzner, Isobel D Stewart, Laura BL Wittemans, Chen Li, Roberto Bonelli, Johannes Raffler, Emma K Biggs, Clare Oliver-Williams, Victoria PW Auyeung, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nature Genetics*, 53(1):54–64, 2021.

[24] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*, 10(5):e1004383, 2014.

[25] Farhad Hormozdiari, Martijn Van De Bunt, Ayellet V Segre, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

[26] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, 2016.

[27] Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature communications*, 9(1):1–20, 2018.

[28] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, 48(5):481–487, 2016.

[29] Eleonora Porcu, Sina Rüeger, Kaido Lepik, Federico A Santoni, Alexandre Reymond, and Zoltán Kutalik. Mendelian randomization integrating gwas and eqtl data reveals genetic determinants of complex and clinical traits. *Nature communications*, 10(1):1–12, 2019.

[30] Masato Akiyama. Multi-omics study for interpretation of genome-wide association study. *Journal of Human Genetics*, 66(1):3–10, 2021.

[31] Eleonora Porcu, Marie C. Sadler, Kaido Lepik, Chiara Auwerx, Andrew R. Wood, Antoine Weihs, Maroun S. Bou Sleiman, Diogo M. Ribeiro, Stefania Bandinelli, Toshiko Tanaka, Matthias Nauck, Uwe Völker, Olivier Delaneau, Andres Metspalu, Alexander Teumer, Timothy Frayling, Federico A.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Santoni, Alexandre Reymond, and Zoltán Kutalik. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nature Communications*, 12(1):5647, September 2021.

[32] Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

[33] Claudia Giambartolomei, Jimmy Zhenli Liu, Wen Zhang, Mads Hauberg, Huwenbo Shi, James Boocock, Joe Pickrell, Andrew E Jaffe, CommonMind Consortium, Bogdan Pasaniuc, et al. A bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, 34(15):2538–2545, 2018.

[34] Kevin J Gleason, Fan Yang, Brandon L Pierce, Xin He, and Lin S Chen. Primo: integration of multiple gwas and omics qtl summary statistics for elucidation of molecular mechanisms of trait-associated snps and detection of pleiotropy in complex traits. *Genome biology*, 21(1):1–24, 2020.

[35] Yang Wu, Jian Zeng, Futao Zhang, Zhihong Zhu, Ting Qi, Zhili Zheng, Luke R Lloyd-Jones, Riccardo E Marioni, Nicholas G Martin, Grant W Montgomery, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nature communications*, 9(1):1–14, 2018.

[36] Eilis Hannon, Tyler J Gorrie-Stone, Melissa C Smart, Joe Burrage, Amanda Hughes, Yanchun Bao, Meena Kumari, Leonard C Schalkwyk, and Jonathan Mill. Leveraging dna-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *The American Journal of Human Genetics*, 103(5):654–665, 2018.

[37] Gibran Hemani, Kate Tilling, and George Davey Smith. Orienting the causal relationship between imprecisely measured traits using gwas summary data. *PLoS genetics*, 13(11):e1007081, 2017.

[38] Alice R Carter, Eleanor Sanderson, Gemma Hammerton, Rebecca C Richmond, George Davey Smith, Jon Heron, Amy E Taylor, Neil M Davies, and Laura D Howe. Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *European journal of epidemiology*, 36(5):465–478, 2021.

[39] Eleanor Sanderson. Multivariable mendelian randomization and mediation. *Cold Spring Harbor perspectives in medicine*, 11(2):a038984, 2021.

[40] Rebecca C Richmond, Gibran Hemani, Kate Tilling, G Davey Smith, and CL Relton. Challenges and novel approaches for investigating molecular mediation. *Human molecular genetics*, 25(R2):R149–R156, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[41] Stephen Burgess, Deborah J Thompson, Jessica MB Rees, Felix R Day, John R Perry, and Ken K Ong. Dissecting causal pathways using mendelian randomization with summarized genetic data: application to age at menarche and risk of breast cancer. *Genetics*, 207(2):481–487, 2017.

[42] Stephen Burgess and Simon G Thompson. Bias in causal estimates from mendelian randomization studies with weak instruments. *Statistics in medicine*, 30(11):1312–1323, 2011.

[43] Verena Zuber, Johanna Maria Colijn, Caroline Klaver, and Stephen Burgess. Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nature communications*, 11(1):1–11, 2020.

[44] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, et al. Ensembl 2021. *Nucleic acids research*, 49(D1):D884–D891, 2021.

[45] Adrian Bird. Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21, 2002.

[46] Jun Wan, Verity F Oliver, Guohua Wang, Heng Zhu, Donald J Zack, Shannath L Merbs, and Jiang Qian. Characterization of tissue-specific differential dna methylation suggests distinct modes of positive and negative gene expression regulation. *BMC genomics*, 16(1):1–11, 2015.

[47] Ieva Rauluseviciute, Finn Drabløs, and Morten Beck Rye. Dna hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC medical genomics*, 13(1):1–15, 2020.

[48] Elin Grundberg, Eshwar Meduri, Johanna K Sandling, Åsa K Hedman, Sarah Keildson, Alfonso Buil, Stephan Busche, Wei Yuan, James Nisbet, Magdalena Sekowska, et al. Global analysis of dna methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *The American Journal of Human Genetics*, 93(5):876–890, 2013.

[49] James R Wagner, Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen, and Mathieu Blanchette. The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome biology*, 15(2):1–17, 2014.

[50] Rita Lippai, Apor Veres-Székely, Erna Sziksz, Yoichiro Iwakura, Domonkos Pap, Réka Rokonay, Beáta Szebeni, Gábor Lotz, Nóra J. Béres, Áron Cseh, Attila J. Szabó, and Ádám Vannay. Immunomodulatory role of Parkinson's disease 7 in inflammatory bowel disease. *Scientific Reports*, 11(1):14582, July 2021.

[51] Antonio F Di Narzo, Carrie Brodmerkel, Shannon E Telesco, Carmen Argmann, Lauren A Peters, Katherine Li, Brian Kidd, Joel Dudley, Judy Cho, Eric E Schadt, et al. High-throughput identification

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the plasma proteomic signature of inflammatory bowel disease. *Journal of Crohn's and Colitis*, 13(4):462–471, 2019.

[52] Yogesh Singh, Christoph Trautwein, Achal Dhariwal, Madhuri S Salker, Md Alauddin, Laimdota Zizmare, Lisann Pelzl, Martina Feger, Jakob Admard, Nicolas Casadei, et al. Dj-1 (park7) affects the gut microbiome, metabolites and the development of innate lymphoid cells (ilcs). *Scientific reports*, 10(1):1–19, 2020.

[53] Jie Zhang, Min Xu, Weihua Zhou, Dejian Li, Hong Zhang, Yi Chen, Longgui Ning, Yuwei Zhang, Sha Li, Mengli Yu, et al. Deficiency in the anti-apoptotic protein dj-1 promotes intestinal epithelial cell apoptosis and aggravates inflammatory bowel disease via p53. *Journal of Biological Chemistry*, 295(13):4237–4251, 2020.

[54] Alexander R Moschen, Romana R Gerner, Jun Wang, Victoria Klepsch, Timon E Adolph, Simon J Reider, Hubert Hackl, Alexandra Pfister, Johannes Schilling, Patrizia L Moser, et al. Lipocalin 2 protects from inflammation and tumorigenesis associated with gut microbiota alterations. *Cell host & microbe*, 19(4):455–469, 2016.

[55] Carl A Anderson, Gabrielle Boucher, Charlie W Lees, Andre Franke, Mauro D'Amato, Kent D Taylor, James C Lee, Philippe Goyette, Marcin Imielinski, Anna Latiano, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature genetics*, 43(3):246–252, 2011.

[56] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 2012.

[57] Christian Gieger, Aparna Radhakrishnan, Ana Cvejic, Weihong Tang, Eleonora Porcu, Giorgio Pistis, Jovana Serbanovic-Canic, Ulrich Elling, Alison H. Goodall, Yann Labrune, Lorna M. Lopez, Reedik Mägi, Stuart Meacham, Yukinori Okada, Nicola Pirastu, Rossella Sorice, Alexander Teumer, Katrin Voss, Weihua Zhang, Ramiro Ramirez-Solis, et al. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208, December 2011.

[58] Sylvia T Nürnberg, Augusto Rendon, Peter A Smethurst, Dirk S Paul, Katrin Voss, Jonathan N Thon, Heather Lloyd-Jones, Jennifer G Sambrook, Marloes R Tijssen, HaemGen Consortium, et al. A gwas sequence variant for platelet volume marks an alternative dnm3 promoter in megakaryocytes near a meis1 binding site. *Blood, The Journal of the American Society of Hematology*, 120(24):4859–4868, 2012.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[59] Nicholas A Watkins, Arief Gusnanto, Bernard De Bono, Subhajyoti De, Diego Miranda-Saavedra, Debbie L Hardie, Will GJ Angenent, Antony P Attwood, Peter D Ellis, Wendy Erber, et al. A haematlas: characterizing gene expression in differentiated human blood cells. *Blood, The Journal of the American Society of Hematology*, 113(19):e1–e9, 2009.

[60] Ewa Bielczyk-Maczyńska, Jovana Serbanovic-Canic, Lauren Ferreira, Nicole Soranzo, Derek L Stemple, Willem H Ouwehand, and Ana Cvejic. A loss of function screen of identified genome-wide association study loci reveals new genes controlling hematopoiesis. *PLoS genetics*, 10(7):e1004450, 2014.

[61] Y Liang, P Wang, M Zhao, G Liang, H Yin, G Zhang, H Wen, and Q Lu. Demethylation of the fcer1g promoter leads to fcεri overexpression on monocytes of patients with atopic dermatitis. *Allergy*, 67(3):424–430, 2012.

[62] Erola Pairo-Castineira, Sara Clohisey, Lucija Klaric, Andrew D Bretherick, Konrad Rawlik, Dorota Pasko, Susan Walker, Nick Parkinson, Max Head Fourman, Clark D Russell, et al. Genetic mechanisms of critical illness in covid-19. *Nature*, 591(7848):92–98, 2021.

[63] COVID-19 Host Genetics Initiative et al. Mapping the human genetic architecture of covid-19. *Nature*, 2021.

[64] Sandra P. Smieszek and Mihael H. Polymeropoulos. Loss of Function Mutations in the IFNAR2 in COVID-19 Severe Infection Susceptibility. *Journal of Global Antimicrobial Resistance*, July 2021.

[65] Yi-An Ko, Huiguang Yi, Chengxiang Qiu, Shizheng Huang, Jihwan Park, Nora Ledo, Anna Köttgen, Hongzhe Li, Daniel J Rader, Michael A Pack, et al. Genetic-variation-driven gene-expression changes highlight genes with important functions for kidney disease. *The American Journal of Human Genetics*, 100(6):940–953, 2017.

[66] Xiangchen Gu, Hongliu Yang, Xin Sheng, Yi-An Ko, Chengxiang Qiu, Jihwan Park, Shizheng Huang, Rachel Kember, Renae L Judy, Joseph Park, et al. Kidney disease genetic risk variants alter lysosomal beta-mannosidase (manba) expression and disease severity. *Science Translational Medicine*, 13(576), 2021.

[67] Mayo Foundation for Medical Education and Research. Insulin-Like Growth Factor-Binding Protein 3 (IGFBP-3), Serum. `https://www.mayocliniclabs.com/test-catalog/Clinical+and+Interpretive/83300`, 2021. [Online; accessed August-2021].

[68] Changjian Feng, Gordon Tollin, and John H Enemark. Sulfite oxidizing enzymes. *Biochimica Et Biophysica Acta (BBA)-Proteins and Proteomics*, 1774(5):527–539, 2007.

[69] KV Venkatachalam. Human 3'-phosphoadenosine 5'-phosphosulfate (paps) synthase: Biochemistry, molecular biology and genetic deficiency. *IUBMB life*, 55(1):1–11, 2003.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[70] Martin J Stone, Sara Chuang, Xu Hou, Menachem Shoham, and John Z Zhu. Tyrosine sulfation: an increasingly recognised post-translational modification of secreted proteins. *New biotechnology*, 25(5):299–317, 2009.

[71] Justin P Ludeman and Martin J Stone. The structural role of receptor tyrosine sulfation in chemokine recognition. *British journal of pharmacology*, 171(5):1167–1179, 2014.

[72] Hugo Zeberg and Svante Pääbo. A genomic region associated with protection against severe covid-19 is inherited from neandertals. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[73] Sirui Zhou, Guillaume Butler-Laporte, Tomoko Nakanishi, David R Morrison, Jonathan Afilalo, Marc Afilalo, Laetitia Laurent, Maik Pietzner, Nicola Kerrison, Kaiqiong Zhao, et al. A neanderthal oas1 isoform protects individuals of european ancestry against covid-19 susceptibility and severity. *Nature medicine*, 27(4):659–667, 2021.

[74] Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews genetics*, 13(4):227–232, 2012.

[75] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the dependency of cellular protein levels on mrna abundance. *Cell*, 165(3):535–550, 2016.

[76] Heng Zhu, Guohua Wang, and Jiang Qian. Transcription factors as readers and effectors of dna methylation. *Nature Reviews Genetics*, 17(9):551–565, 2016.

[77] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, et al. Impact of cytosine methylation on dna binding specificities of human transcription factors. *Science*, 356(6337), 2017.

[78] Sean Whalen, Rebecca M Truty, and Katherine S Pollard. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature genetics*, 48(5):488–496, 2016.

[79] Daudi Jjingo, Andrew B Conley, V Yi Soojin, Victoria V Lunyak, and I King Jordan. On the presence and role of human gene-body dna methylation. *Oncotarget*, 3(4):462, 2012.

[80] Christian P Moritz, Timo Mühlhaus, Stefan Tenzer, Thomas Schulenborg, and Eckhard Friauf. Poor transcript-protein correlation in the brain: negatively correlating gene products reveal neuronal polarity as a potential cause. *Journal of neurochemistry*, 149(5):582–604, 2019.

[81] Rasmus Magnusson, Olof Rundquist, Min Jung Kim, Sandra Hellberg, Chan Hyun Na, Mikael Benson, David Gomez-Cabrero, Ingrid Kockum, Jesper Tegnér, et al. A validated strategy to infer protein biomarkers from rna-seq by combining multiple mrna splice variants and time-delay. *bioRxiv*, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[82] Eleonora Porcu, Jennifer Sjaarda, Kaido Lepik, Cristian Carmeli, Liza Darrous, Jonathan Sulc, Ninon Mounier, and Zoltán Kutalik. Causal inference methods to integrate omics and complex traits. *Cold Spring Harbor Perspectives in Medicine*, 11(5):a040493, 2021.

[83] Ullrich Wüllner, Oliver Kaut, Laura deBoni, Dominik Piston, and Ina Schmitt. Dna methylation in parkinson's disease. *Journal of neurochemistry*, 139:108–120, 2016.

[84] L Tagliafierro and O Chiba-Falek. Up-regulation of snca gene expression: implications to synucleinopathies. *Neurogenetics*, 17(3):145–157, 2016.

[85] Laura Marsal-García, Aintzane Urbizu, Laura Arnaldo, Jaume Campdelacreu, Dolores Vilas, Lourdes Ispierto, Jordi Gascón-Bayarri, Ramón Reñé, Ramiro Álvarez, and Katrin Beyer. Expression levels of an alpha-synuclein transcript in blood may distinguish between early dementia with lewy bodies and parkinson's disease. *International journal of molecular sciences*, 22(2):725, 2021.

[86] Ruth Chia, Marya S Sabir, Sara Bandres-Ciga, Sara Saez-Atienzar, Regina H Reynolds, Emil Gustavsson, Ronald L Walton, Sarah Ahmed, Coralie Viollet, Jinhui Ding, et al. Genome sequencing analysis identifies new loci associated with lewy body dementia and provides insights into its genetic architecture. *Nature genetics*, 53(3):294–303, 2021.

[87] Marie Verbanck, Chia-yen Chen, Benjamin Neale, and Ron Do. Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nature genetics*, 50(5):693–698, 2018.

[88] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, 2012.

[89] Diego Garrido-Martín, Beatrice Borsari, Miquel Calvo, Ferran Reverter, and Roderic Guigó. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nature communications*, 12(1):1–16, 2021.

[90] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665, 2013.

[91] UK10K consortium et al. The uk10k project identifies rare variants in health and disease. *Nature*, 526(7571):82, 2015.

[92] Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

[93] Gibran Hemani, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, Ryan Langdon, et al. The mr-base platform supports systematic causal inference across the human phenome. *elife*, 7:e34408, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.29.462396; this version posted October 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[94] Andrew Nightingale, Ricardo Antunes, Emanuele Alpi, Borisas Bursteinas, Leonardo Gonzales, Wudong Liu, Jie Luo, Guoying Qi, Edd Turner, and Maria Martin. The proteins api: accessing key integrated protein and genome information. *Nucleic acids research*, 45(W1):W539–W544, 2017.

[95] Andrew Yates, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham RS Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, and Paul Flicek. The ensembl rest api: Ensembl data for any language. *Bioinformatics*, 31(1):143–145, 2015.

[96] Aaron F McDaid, Peter K Joshi, Eleonora Porcu, Andrea Komljenovic, Hao Li, Vincenzo Sorrentino, Maria Litovchenko, Roel PJ Bevers, Sina Rüeger, Alexandre Reymond, et al. Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nature communications*, 8(1):1–11, 2017.

[97] KD Hansen. Illuminahumanmethylation450kanno. ilmn12. hg19: annotation for illumina's 450k methylation arrays. *R package version 0.6. 0*, 10:B9, 2016.