Review

Strengthening Causal Inference for Complex Disease Using Molecular Quantitative Trait Loci

Sonja Neumeyer[1], Gibran Hemani[2], Eleftheria Zeggini[1*]

[1]Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.

[2]MRC Integrative Epidemiology Unit (IEU), Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom.

***Correspondence:** eleftheria.zeggini@helmholtz-muenchen.de (E. Zeggini).

Keywords: Mendelian randomization, QTL, complex trait, GWAS, genome-wide association study, gene expression

Abstract

Large genome-wide association studies have identified loci associated with complex traits and diseases, but often index variants are not causal and reside in non-coding regions of the genome. To gain a better understanding of the relevant biological mechanisms, intermediate traits such as gene expression or protein levels are increasingly being investigated, as these are likely mediators between genetic variants and disease outcome. Genetic variants associated with intermediate traits, termed molecular quantitative trait loci (molQTLs), can then be used as instrumental variables in a Mendelian randomization approach to identify causal features and mechanisms of complex traits. Challenges such as pleiotropy and non-specificity of molQTLs remain and further approaches and methods need to be developed.

Genome-Wide Association Studies

Genome-wide association studies **(GWAS, Box 1)** have identified thousands of sequence variants that contribute to the genetic architecture of complex diseases and medically-relevant quantitative traits. This endeavour has been fuelled by two major ambitions: creating genetic predictors for disease; and identifying the genomic regions responsible for the disease to gain a better understanding of the relevant biological mechanisms [1, 2]. The latter objective is the focus of this review.

Typically, associated variants individually account for a very small proportion of phenotypic variation. This is common for quantitative or “complex” traits which are usually influenced by a large number of genes with small effects on the trait [3]. There is no simple Mendelian inheritance pattern but random sampling of alleles at each associated gene results in a normally distributed phenotype in the population [4]. Functional information on the underlying mechanisms of genetic variants identified by GWAS is often unclear, i.e. it is challenging to identify effector genes based on the observed association summary statistics only [3, 5]. The majority of complex trait variants reside in noncoding regions of the genome [6, 7] and it is possible that they confer their effect through modulating gene expression levels [8]. In their second decade of existence, GWAS are showing signs of maturity, with increasing diversity in populations studied [9], inclusion of low frequency and rare variants, and finer definition of phenotypic traits examined.

In this review we will describe how molecular traits are also being assayed and analysed for genetic associations, and how the understanding of complex disease aetiology is improving through combining genetic analysis of both the disease and molecular traits. The presiding manner in which these relationships are constructed is using a causal inference method known as Mendelian randomization (MR) which capitalizes on the abundance of GWAS results now available. We will describe MR in terms of both its current implementation and the future developments that are needed to address known limitations.

Molecular Quantitative Trait Loci

The influence of a genetic variant associated with a disease is likely to be mediated via molecular traits (Figure 1), which themselves are often complex. Quantitative molecular traits, such as gene expression or protein abundance, are frequently dysregulated in disease and can act as intermediate phenotypes, affording greater power to detect association compared to the dichotomous definition of a disease endpoint, which is the culmination of multiple biological processes being perturbed [10].

Multiple studies have investigated mRNA levels combined with genome-wide genotype information to identify expression quantitative trait loci (eQTLs), i.e. genetic variants associated with gene expression levels [11]. The first studies to investigate molecular quantitative trait loci (molQTLs) started out with small sample sizes. Due to challenges associated with collecting human biospecimens using invasive procedures, analyses initially focussed on using the most accessible tissues [12]. Today, sample sizes used for molQTL investigation in blood have grown very large [13]. MolQTLs are generally classified into cis-acting, which is typically defined as regulation of genes within 1Mb, or trans-acting, defined as molQTLs affecting genes further away or on different chromosomes [14]. Whereas detected cis-effects have generally been large and easily found using small sample sizes, trans effects tend to be much smaller and larger sample sizes are required. Large studies such as the eQTLGen Consortium [13] or GoDMC (http://www.godmc.org.uk/) are emerging to identify these small effects that might play central roles in disease etiology. Molecular trait loci seem to be highly tissue dependent [15, 16]. However, tissue-sharing of cis-eQTLs seems to be bimodal. Either cis-eQTLs seem to be shared across many tissues or they are very specific to only a small subset of tissues [17].To provide a resource which enables the systematic study of genetic variation on regulation of gene expression in multiple human tissues, the Genotype-Tissue Expression (GTEx) project was initiated a decade ago [18]. The current GTEx release provides a total of 11688 samples and 53 tissues across 714 donors (current release V7, dbGaP accession phs000424.v7.p2). Sample sizes of other studies have also largely increased [19-21] and a variety of tissues have been studied. The picture is far from complete, but has been massively enhanced since the inception of these studies.

The first expression phenotypes to be studied were gene transcript levels. They are highly heritable [22]. It is estimated that around 88% of all genes have at least one eQTL [13]. To date, many different molecular traits with a potential influence on gene regulation have been investigated [23]. They range from influencing the epigenome such as DNA methylation (meQTL), histone modification (hQTL) or chromatin accessibility (caQTL) to alternative splicing (sQTL), protein levels (pQTL), microRNA expression (mirQTL) or ribosome occupancy (rQTL) [23]. In addition, higher level intermediate phenotypes such as metabolites have been investigated and QTLs for metabolites such as carbohydrates, amino acids or fatty acids identified [24].

In an effort to find the molecular pathways that connect genetic variants to complex traits, overlapping/colocalisation methods between GWAS and molQTL signals have been developed. Colocalisation of an eQTL with a GWAS signal suggests that the eQTL target gene could be involved in the molecular pathway of the complex disease under investigation [25]. Several studies already discovered GWAS signals enriched for molQTLs in a tissue dependent-manner [26]. For example, the myocardial infarction and high LDL cholesterol-associated **1p13 locus (see Glossary)** had been fine mapped to the ***CELSR2*** gene. Using eQTL analyses, it was discovered that actually the expression of ***SORT1*** was influenced by this variant [27].

MolQTLs are being used as instrumental variables for molecular traits in a variety of ways: to infer the relative importance of different classes of molecular features on variation in complex traits; to identify the causal gene for a particular complex trait [23]; to identify the causal tissue for a complex trait [28] and to estimate causal relationships between different molecular traits [29]. In this review, we will focus on their use for identifying causal features of complex traits.

Mendelian Randomization Studies Strengthen Causal Inference

Mendelian randomization (**MR, Box 2**) studies use genetic variants as proxies for modifiable risk factors to test whether the risk factor is causally relevant to an outcome of interest [30, 31]. The advantage of such an approach is that unmeasured confounding, an issue of observational studies, and reverse causation can be minimized. It is, therefore, possible to use genetic information to draw causal inferences.

Early MR studies mainly used one-sample approaches, where the exposure and outcome phenotypes along with the genetic variants that were being used to instrument the exposure were available for all samples in a single dataset. Nowadays, when many large-scale GWASs are conducted, it is much more powerful to use published **SNP (single nucleotide polymorphism)** -trait associations from large consortia. It is, therefore, common to use two-sample MR approaches where SNP-exposure and SNP-outcome associations are estimated in different studies and subsequently combined [32]. When using genome-wide significant SNPs as instrumental variable for an exposure, the first MR assumption should be verified.

For two-sample MR methods, only summary statistics are required (per allele regression coefficients, standard errors and effect allele) which are typically obtained from published GWAS of the largest possible datasets [33]. The causal effect can be estimated using the Wald ratio estimate, which is the ratio of SNP-outcome association and SNP-exposure association.

SNP-exposure and SNP-outcome association statistics should ideally be obtained from studies of non-overlapping individuals (two-sample MR). When using summary statistics from only one sample or from partially overlapping samples, results might be biased in the direction of the observational estimate, especially if the genetic effects on the exposure are weak [34]. When several independent genetic variants are known to be associated with the exposure of interest, these can be combined into a single MR estimate using inverse variance weighted meta-analysis of the single Wald ratio estimates [32]. In doing so, the MR framework can then be viewed as a meta-analysis problem which itself has a rich set of tools to evaluate and correct for bias [35]. One issue that has been of particular concern in MR is in proving that violation of the third assumption, i.e. that the genetic instrument influences the outcome only through the exposure, does not induce bias [36]. A suite of sensitivity analyses [37-41] are now routinely implemented in MR studies that use multiple independent instruments to model pleiotropy [42].

Mendelian Randomization Studies Using Molecular QTLs as Instrumental Variables

Whole genome approaches have indicated that the causal variants influencing complex traits are overrepresented by those that are also associated with eQTLs [43, 44]. This supports the notion that disease biology could be unravelled by mapping the causal path from genetic variant through the use of intermediate molQTLs [45]. At its most basic implementation, a Mendelian randomization framework for evaluating the causal influence of a molecular trait on a complex trait would be to test if a known molQTL is also associated with the complex trait (Key Figure, Figure 2). The Wald ratio of SNP-complex trait and SNP-molecular trait effects can then be obtained as an estimate of the causal effect. This simple method suffers from a number of potential pitfalls and is often performed as an initial screen to find, from amongst many molecular phenotypes (e.g. hundreds of thousands of DNA methylation levels), a few putative causal molecular phenotypes for more detailed follow up and sensitivity analysis [46-48]. Here we describe some of these approaches.

Linkage disequilibrium links a causal variant for one trait with a different causal variant for another trait.

A major lesson from GWAS is that complex traits follow a polygenic architecture [49, 50]. As a consequence, finding that a chosen SNP happens to show an association with a complex trait might not be surprising because many non-causal common variants are likely to be in **linkage disequilibrium (LD)** with a causal variant for a complex trait (Figure 3a). Colocalisation techniques seek to analyse specific genomic regions, determining whether the pattern of test statistics for one trait are concordant with the pattern from another, often with respect to the underlying LD structure. Evidence for shared causal variants at a locus is determined by the extent to which the test statistic patterns are shared between the two traits. An important recent finding is that the majority of genes that colocalise with a trait are not the genes that are closest to the biggest signal for the trait [11].

Typically, the proportion of overlapping signals between molecular and complex traits that appear to be due to LD is high. For example in [29] it was shown that two thirds of putative expression-trait MR relationships were due to LD, with a similar proportion being found for DNA methylation-trait MR relationships. Nevertheless, when assessed across hundreds of complex traits, there are now tens of thousands of examples of colocalisation between gene expression levels and complex traits [51]. It remains important to note that there are many colocalisation techniques [11, 52-54] and there is not always strong agreement between them [54].

The association is reverse causal

One of the purported advantages of MR is that it protects against reverse causation. This is true to the extent that the instrument is known to primarily influence the hypothesised exposure. However it is conceivable that a molQTL arises because a complex trait influences it. Mediation-based methods exist that require individual-level data to orient the causal direction [55-57], but are susceptible to making the wrong orientation under specific patterns of confounding or measurement error [58]. An alternative approach is to perform MR in the reverse direction [47], identifying SNPs that instrument the complex trait and testing for its association on the molecular trait. Typically however, one would not expect reverse causal relationships to explain a molQTL associated with a complex trait because in order for the molQTL to have been detected in a small sample size it will necessarily be a large effect, which is impossible if it were mediated through a polygenic trait [29].

The instrumenting SNP is non-specific to the hypothesised exposure

Often a single SNP is detected as an instrument for multiple molecular phenotypes. For example, a SNP could be strongly associated with more than one gene expression level, or the same gene expression level in different tissues or time points, or both a gene expression level and a DNA methylation level (Figure 3). This is not necessarily a problem, as all the molecular phenotypes that are associated with the trait could be on the same causal pathway to the disease, and indeed it could be advantageous as it presents us with multiple points of intervention. Non-specificity of genetic associations is classically known as pleiotropy though care should be taken in using the term. MR assumes a ‘vertically’ pleiotropic relationship, where the genetic instrument is associated with the outcome because it is mediated by the exposure. By contrast, ‘horizontal’ pleiotropy is a source of problems in MR, inducing bias or false causal inference if the SNP influences the outcome through a pathway other than the hypothesised exposure [59]. Proving that a putative MR finding is due to vertical and not horizontal pleiotropy is far from trivial [36].

There are vastly more molecular phenotypes than independent genetic regions, especially when temporal- and tissue-specific measurements are possible [60]. By definition it is expected that many molQTL will not be specific to a particular molecular trait. Therefore, it is difficult to prove which, from amongst the set of molecular traits that are influenced by the molQTL, is the causal factor [51].

One approach is to focus on the use of cis-acting molQTLs, with the rationale that they are biologically ‘closer’ to the intended molecular trait. Trans-acting QTLs are likely to only influence the molecular trait because they are mediated by other molecular traits, opening up a greater possibility that the instrument is non-specific to the intended target (Figure 3b). Testing explicitly if the molQTL is associated with other molecular traits is also sensible, as this can be used to (de-)prioritise a putative association depending on how much evidence there is for (non-)specificity [2]. Methods are now arising that attempt to model the MR estimates of multiple molecular exposures simultaneously, thereby adjusting for potential horizontal pleiotropy [61]. While a useful tool, interpretation remains difficult as the use of multivariable MR [62] requires that there are marked differences in the genetic signatures across the exposures [63]. It also requires measurement of all possible exposures that could be inducing the pleiotropy, which is a similar assumption to observational study designs that prompted the development of MR in the first place.

There are more standard MR sensitivity analyses that can be applied in the event that multiple independent causal variants are available [42]. However, this typically requires introducing trans-QTLs into the analysis which may not bring clarity, as they could have systematically different properties to cis-QTLs. At this stage, if a molecular trait colocalises with a complex trait, and doesn’t appear to be reverse causal, it is still extremely difficult to prove that it is causal and not simply one of many traits that are all influenced by the same molQTL.

In the GoDMC study, which used 30k samples to discover instruments for DNA methylation levels, multiple cis and trans instruments were used to model causal relationships between DNA methylation levels and complex traits. It was found that, while there were many putative colocalising signals with complex traits, there was almost no agreement between the causal effect estimated using primary and secondary molQTLs, implying that the majority of colocalising signals were due to horizontal pleiotropy.

Current Challenges and Issues

The prospects of finding new drug targets has propelled forwards the data acquisition and methodological development for mapping the pathways between molecular and complex traits.

Genetic variation is finite, and though molecular traits are often polygenic the use of more than the cis-region for instrumentation is currently not fully understood. This incurs a limit on the extent to which current tools designed to protect against incorrect causal inference due horizontal pleiotropy can be used. Conceptually, here we use genetic instruments as a proxy for molecular phenotypes. However, molecular phenotypic variation dwarfs the cis-genetic resource that is available for instrumentation. Hence, the ubiquitous non-specificity of any molQTL makes it very difficult to determine which molecular feature is actually mediating the genetic effect on a trait. This could be because inference is for the wrong developmental time point (e.g. genetic effects are very consistent over time [64] for DNA methylation) or the wrong tissue (cis-QTLs are strongly shared across tissues [17]). Alternatively, it could be that it was an entirely different molecular feature (e.g. gene expression, DNA methylation and histone variation often share similar cis-regulatory features [65]).

Coupled with this problem of non-specificity, is the emerging evidence supporting a model of ubiquitous horizontal pleiotropy [40, 66], in which any particular genetic variant potentially influences a particular complex trait through multiple independent pathways. The omnigenic model offers an extreme viewpoint on this problem, in which polygenic architecture arises because every gene is related to every trait through an underlying dense gene regulatory network [3].

Making meaningful inference from such an under-specified model requires a departure from current practices of treating molecular features singly, and reliably incorporating trans-instruments, which may exhibit tissue specificity [17]. Though any one instrument might be non-specific, it is seldom the case that the genetic correlation of complex traits is 1 [67], meaning that there are potentially combinations of instruments that together provide some specificity. Large-scale pleiotropy maps are beginning to be produced [40, 68, 69], and may provide an avenue into constructing instrument combinations conditional on a background of complex pleiotropy.

Concluding Remarks

Many genetic variants associated with complex traits and diseases have been discovered, but often there is a lack of knowledge about mechanisms involved **(see Clinician`s Corner)**. Investigation of intermediate traits and associated molQTLs has been very helpful, as these better explain how genetic variants influence complex traits. Using molQTLs combined with an MR approach, causal features of a complex trait can be revealed. Challenges, such as the model of ubiquitous horizontal pleiotropy and, therefore, a non-specificity of molQTLs to a particular molecular trait, remain **(see Outstanding Questions)**. Therefore, new methods need to be developed, including for example those that reliably incorporate trans-molQTLs, which have a greater possibility for non-specificity of the instrument.

Despite our growing understanding of the limitations of MR, the current data resources and statistical frameworks for MR can be viewed as a resource with tremendous utility. Most directly, using MR to support a negative association could be less prone to some of the issues described. Of growing importance in causal inference is the concept of triangulation, where information from orthogonal experimental designs are integrated together to obtain a more reliable conclusion [70]. There are now open source data and software repositories (including those that can be used in web browsers [42]) that automate MR analyses. The inclusion of genetic evidence through MR should be a natural part of any causal inquiry [71].

BOX 1: Genome-wide association studies

Genome-wide association studies (GWAS) *compare large numbers of affected with unaffected individuals to identify sequence variants that are associated with risk of complex diseases, or at the population-level to identify associations with quantitative traits. The foundation for GWAS was laid by the sequencing of the human genome [72], characterization of the correlation patterns between pairs of variants genome-wide [73], development of high-throughput genotyping platforms, and the availability of large-scale sample sizes. Millions of single nucleotide polymorphisms (**SNPs**) have been mapped [74]. For several reasons it has been difficult to elucidate the underlying mechanism between associated genetic variant and disease trait. One reason is the co-inheritance of many genetic variants with the disease-associated variant (linkage disequilibrium (**LD**)) [75]. Due to this complicated correlation structure of human genome, the most strongly associated GWAS signal (index variant) is often not causal [76]. Similarly, compounded by complex regulatory mechanisms, the nearest gene to the top GWAS signal is not necessarily the causal gene [11].*

Box 2: Mendelian randomization studies

*Due to the laws of Mendelian inheritance, alleles are assigned at conception to individuals independent of environmental risk factors and confounders. To obtain valid estimates using Mendelian randomization (**MR**), three assumptions have to be met: firstly, the genetic variants need to be sufficiently associated with the exposure of interest; secondly, the genetic variants should not be associated to any confounder of the risk factor – outcome relationship; finally there should not be any other pathway from genetic variants to outcome except through the exposure of interest. Except for the first assumption, which can be tested, the other two assumptions can only be addressed by sensitivity analyses [77].*

Clinician`s Corner

- *Poor efficacy and poor safety are the two major reasons for the very high failure rate of drug trials, ultimately driving up the cost of drugs and their development times. This can be partly framed as a causal inference problem, where the objective is to identify which molecular targets are causal for the disease of interest and filter out those that are likely to fail prior to initiating trials.*
- *Randomized controlled trials (RCTs) are ideal for making causal inference but are expensive, slow and often impracticable for a particular causal enquiry. The Mendelian randomization statistical framework leverages genetic associations to mimic randomized control trials. The potential of this strategy is increasingly being exploited due to the ready availability of data to quickly and cheaply evaluate the causal importance for thousands of molecular features on complex diseases.*
- *To interrogate the causal influence of a particular molecular trait on a particular disease, knowledge of robust genetic factors for the molecular trait, and the corresponding effect of those factors on the disease, are both required. Thanks to over a decade of genome-wide association studies and the recent emergence of national genetic biobanks, most complex diseases have genome-wide genetic associations from large sample sizes made publicly available. In addition, the genetic influences on a range of molecular features such as protein levels, gene expression levels, DNA methylation levels, metabolites etc are being mapped and made publicly available.*
- *Though it is impossible to mimic an RCT perfectly using such observational data, statistical techniques and data continue to improve, and Mendelian randomization is poised to further help make causal claims about a molecular trait on complex disease.*

Glossary

1p13 locus: GWAS analysis in humans demonstrated that this locus on chromosome 1 is strongly associated with plasma low-density lipoprotein cholesterol (LDL-C) levels, which in turn is a major risk factor for myocardial infarction. SNPs (see below) in this locus have also been linked to coronary artery disease. This locus alters the expression of SORT1 (see below) in the liver.

***CELSR2*:** Cadherin EGF LAG seven-pass G-type receptor 2, a receptor with possible role in cell/cell signaling during nervous system formation. *CELSR2* is physically linked to the 1p13 locus. Because of this, CELSR2 expression was thought to be controlled by the 1p13 locus until eQTL analysis showed that this was not the case.

LD: linkage disequilibrium, the non-random association of alleles at different loci. Based on the assumption that over time recombination events will result in a random association of alleles at two loci, linkage disequilibrium is defined as the difference between the observed frequency of a particular combination of alleles at two loci compared to the frequency expected at random. When analyzing causal SNPs in GWAS analysis, special care must be taken to not wrongly interpret a non-causal SNP that is in LD with a causal SNP.

SNP: single nucleotide polymorphism, a DNA sequence variant within a population. SNPs can be linked to disease development and response to pathogens or medication in humans, which makes them invaluable in personalized medicine. Comparison of SNP composition in genomic regions between different cohorts (e.g. with and without disease) is of great importance in biomedical research on a larger scale (e.g. GWAS).

***SORT1*:** Sortilin, which is localized in intracellular compartments, notably the Golgi apparatus. It is involved in endocytosis and functions as a sorting receptor in the Golgi compartment and clearance receptor on the cell surface. SORT1 expression is modulated by the 1p13 locus (see above). In liver cells of mouse models, LDL-C levels are significantly decreased by SORT1 overexpression whereas SORT1 knockdown resulted in an increase of LDL-C levels.

Acknowledgements

GH is funded by the Wellcome Trust and Royal Society [208806/Z/17/Z]. We thank Tom Richardson and Iris Fischer for valuable input.

References

1. Nelson, M.R. et al. (2015) The support of human genetic evidence for approved drug indications. Nat. Genet. 47, 856-860.

2. Zheng, J. et al. (2019) Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. bioRxiv, 627398.

3. Boyle, E.A. et al. (2017) An expanded view of complex traits: From polygenic to omnigenic. Cell 169, 1177-1186.

4. Fisher, R.A. (1919) XV.—The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinburgh 52, 399-433.

5. Ruiz-Narváez, E.A. (2011) What is a functional locus? Understanding the genetic basis of complex phenotypic traits. Med. Hypotheses 76, 638-642.

6. Edwards, Stacey L. et al. (2013) Beyond GWASs: Illuminating the dark road from association to function. Am. J. Hum. Genet. 93, 779-797.

7. Farh, K.K.-H. et al. (2014) Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337-343.

8. Maurano, M.T. et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190-1195.

9. Gurdasani, D. et al. (2019) Genomics of disease risk in globally diverse populations. Nat. Rev. Genet., doi: 10.1038/s41576-019-0144-0.

10. Civelek, M. and Lusis, A.J. (2013) Systems genetics approaches to understand complex traits. Nat. Rev. Genet. 15, 34-48.

11. Zhu, Z. et al. (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48, 481-487.

12. Emilsson, V. et al. (2008) Genetics of gene expression and its effect on disease. Nature 452, 423-428.

13. Võsa, U. et al. (2018) Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv, 447367.

14. Westra, H.-J. et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. 45, 1238-1243.

15. Brown, C.D. et al. (2013) Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genet. 9, e1003649.

16. Sun, B.B. et al. (2018) Genomic atlas of the human plasma proteome. Nature 558, 73-79.

17. The GTEx Consortium et al. (2017) Genetic effects on gene expression across human tissues. Nature 550, 204-213.

18. Lonsdale, J. et al. (2013) The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580-585.

19. Hannon, E. et al. (2018) Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. Am. J. Hum. Genet. 103, 654-665.

20. Yao, C. et al. (2018) Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat. Commun. 9, 3268.

21. Taylor, K. et al. (2019) Prioritizing putative influential genes in cardiovascular disease susceptibility by applying tissue-specific Mendelian randomization. Genome Med. 11, 6.

22. Morley, M. et al. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430, 743-747.

23. Vandiedonck, C. (2018) Genetic association of molecular traits: A help to identify causative variants in complex diseases. Clin. Genet. 93, 520-532.

24. Kastenmüller, G. et al. (2015) Genetics of human metabolism: an update. Hum. Mol. Genet. 24, R93-R101.

25. Wen, X. et al. (2017) Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet. 13, e1006646.

26. Nicolae, D.L. et al. (2010) Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. PLoS Genet. 6, e1000888.

27. Musunuru, K. et al. (2010) From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. Nature 466, 714-719.

28. Ongen, H. et al. (2017) Estimating the causal tissues for complex traits and diseases. Nat. Genet. 49, 1676-1683.

29. Taylor, D.L. et al. (2019) Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. Proc. Natl. Acad. Sci. 116, 10883-10888.

30. Davey Smith, G. and Ebrahim, S. (2003) ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. 32, 1-22.

31. Ebrahim, S. and Davey Smith, G. (2008) Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? Hum. Genet. 123, 15-33.

32. Pierce, B.L. and Burgess, S. (2013) Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am. J. Epidemiol 178, 1177-1184.

33. Hartwig, F.P. et al. (2017) Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. Int. J. Epidemiol. 45, 1717-1726.

34. Burgess, S. et al. (2016) Bias due to participant overlap in two-sample Mendelian randomization. Genet. Epidemiol. 40, 597-608.

35. Bowden, J. and Holmes, M.V. (2019) Meta-analysis and Mendelian randomization: A review. Res. Synth. Methods, doi: 10.1002/jrsm.1346.

36. Hemani, G. et al. (2018) Evaluating the potential role of pleiotropy in Mendelian randomization studies. Hum. Mol. Genet. 27, R195-R208.

37. Bowden, J. et al. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int. J. Epidemiol. 44, 512-525.

38. Bowden, J. et al. (2016) Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet. Epidemiol. 40, 304-314.

39. Hartwig, F.P. et al. (2017) Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int. J. Epidemiol. 46, 1985-1998.

40. Hemani, G. et al. (2017) Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. bioRxiv, 173682.

41. Zhu, Z. et al. (2018) Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat. Commun. 9, 224.

42. Hemani, G. et al. (2018) The MR-Base platform supports systematic causal inference across the human phenome. eLife 7, e34408.

43. Gusev, A. et al. (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. 95, 535-552.

44. Banos, D.T. et al. (2018) Bayesian reassessment of the epigenetic architecture of complex traits. bioRxiv, 450288.

45. Schadt, E.E. et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297-302.

46. Hannon, E. et al. (2015) Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. Nat. Neurosci. 19, 48-54.

47. Richardson, T.G. et al. (2017) Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk. Am. J. Hum. Genet. 101, 590-602.

48. Richardson, T.G. et al. (2018) Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease. Hum. Mol. Genet. 27, 3293-3304.

49. The International Schizophrenia Consortium et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-752.

50. Yang, J. et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565-569.

51. Richardson, T.G. et al. (2019) A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. bioRxiv, 563379.

52. Giambartolomei, C. et al. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383.

53. Hormozdiari, F. et al. (2016) Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. 99, 1245-1260.

54. Barbeira, A.N. et al. (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. 9, 1825.

55. Aten, J.E. et al. (2008) Using genetic markers to orient the edges in quantitative trait networks: The NEO software. BMC Syst. Biol. 2, 34.

56. Millstein, J. et al. (2009) Disentangling molecular relationships with a causal inference test. BMC Genet. 10, 23.

57. Waszak, Sebastian M. et al. (2015) Population variation and genetic control of modular chromatin architecture in humans. Cell 162, 1039-1050.

58. Hemani, G. et al. (2017) Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet. 13, e1007081.

59. Davey Smith, G. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. 23, R89-R98.

60. Houle, D. et al. (2010) Phenomics: the next challenge. Nat. Rev. Genet. 11, 855-866.

61. Porcu, E. et al. (2019) Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. bioRxiv, 377267.

62. Burgess, S. and Thompson, S.G. (2015) Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. Am. J. Epidemiol 181, 251-260.

63. Sanderson, E. et al. (2018) An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int. J. Epidemiol., doi: 10.1093/ije/dyy262.

64. Gaunt, T.R. et al. (2016) Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 17, 61.

65. Grubert, F. et al. (2015) Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell 162, 1051-1065.

66. Jordan, D.M. et al. (2019) The landscape of pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. bioRxiv, 311332.

67. Bulik-Sullivan, B. et al. (2015) An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236-1241.

68. Cortes, A. et al. (2017) Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. Nat. Genet. 49, 1311-1318.

69. Cho, Y. et al. (2018) MR-TRYX: Exploiting horizontal pleiotropy to infer novel causal pathways. bioRxiv, 476085.

70. Lawlor, D.A. et al. (2017) Triangulation in aetiological epidemiology. Int. J. Epidemiol. 45, 1866-1886.

71. Walker, V. et al. (2019) Using the MR-Base platform to investigate risk factors and drug targets for thousands of phenotypes [version 1; peer review: 2 approved]. Wellcome Open Research 4.

72. Lander, E.S. (2011) Initial impact of the sequencing of the human genome. Nature 470, 187-197.

73. Thorisson, G.A. et al. (2005) The International HapMap Project Web site. Genome Res. 15, 1592-1593.

74. Wang, D.G. et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the Human genome. Science 280, 1077-1082.

75. Altshuler, D. et al. (2005) A haplotype map of the human genome. Nature 437, 1299-1320.

76. Wu, Y. et al. (2017) Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. Genome Biol. 18, 86.

77. Burgess, S. et al. (2017) Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. Epidemiology 28, 30-42.

Figure legends:

Figure 1: Molecular quantitative trait loci influencing intermediate traits. Left graph: Molecular quantitative trait loci (molQTL) are genetic variants associated to a molecular trait and have an influence on intermediate traits (genotypes AA, AG, GG). Right graph: The GG genotype (blue) is associated with higher expression levels of the molecular quantitative trait compared to the AG (yellow) and AA (rose) genotype. These molecular traits can modulate the expression of further target genes (green).

Key Figure, Figure 2: Schematic representation of a Mendelian randomization study using quantitative trait loci as instrumental variables. Due to random distribution of alleles at conception, genetic variants are unrelated to environmental confounders. If genetic variants are sufficiently associated with the modifiable exposure of interest (here: methylation levels, RNA expression levels or protein levels) and not associated to the outcome by a different pathway, then they can be used as instrumental variable for the exposure.

Figure 3: Simplified directed acyclic graphs of possible systems that would lead to an apparent causal effect of gene expression on a trait. Gene regulation may be regulated by several elements. In all the situations depicted, a naïve Mendelian randomization (MR) analysis would return a causal signal for any of the regulatory elements though most often they are not on the causal pathway. A) Three scenarios for cis molecular quantitative trait loci (molQTL) regulation are presented. Vertical: Both gene expression and DNA methylation (DNAm) are on the causal pathway, hence MR using the cis-genetic variant will give valid causal estimates whether it is used to instrument either of these elements. Horizontal: Using the instrument for DNAm will be invalid due to horizontal pleiotropy. Different causal variants: The molQTL is in linkage disequilibrium (LD) with another variant that influences the trait, hence neither regulatory element is causally influenced though naïve MR could indicate otherwise. B) Four scenarios for molQTL regulation are similar to A) except the molQTL for DNAm is on a different chromosome. There are now more opportunities for horizontal pleiotropy as there needs to be a longer path from the trans chromosome to the DNA methylation level.

Figure 1

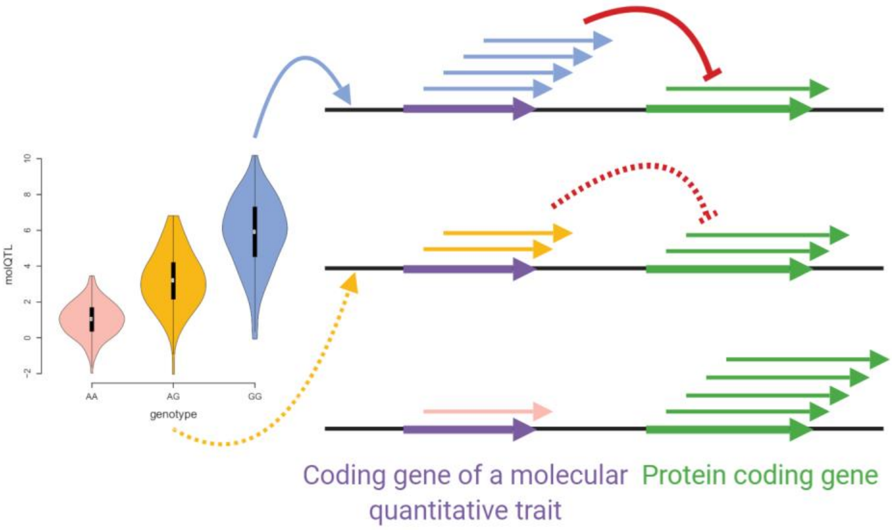


Figure 2

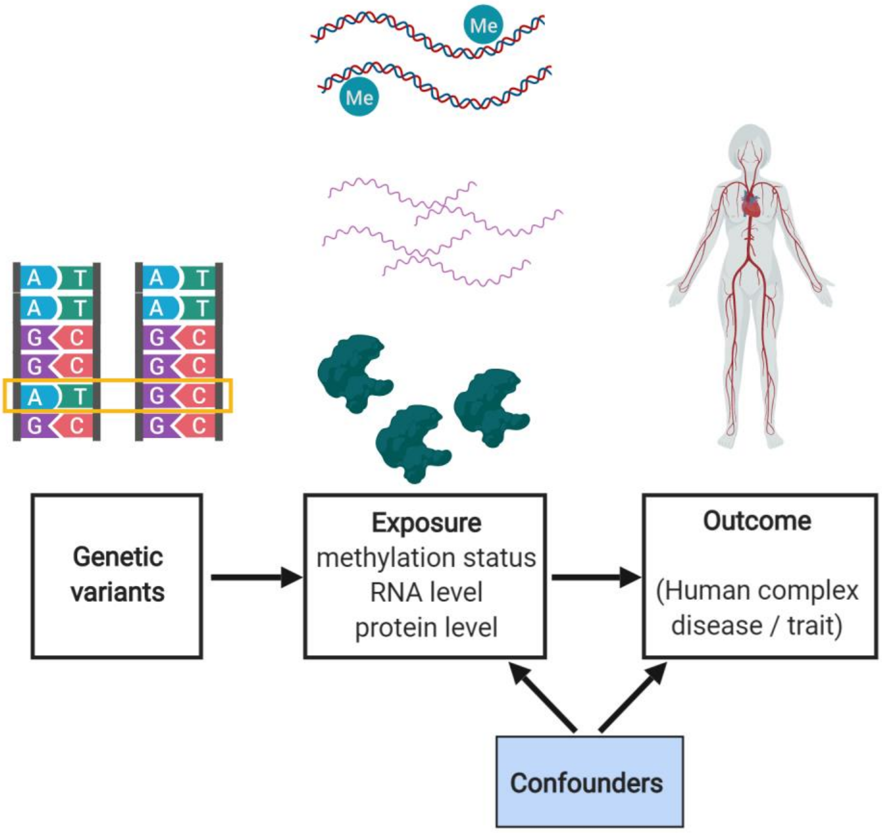
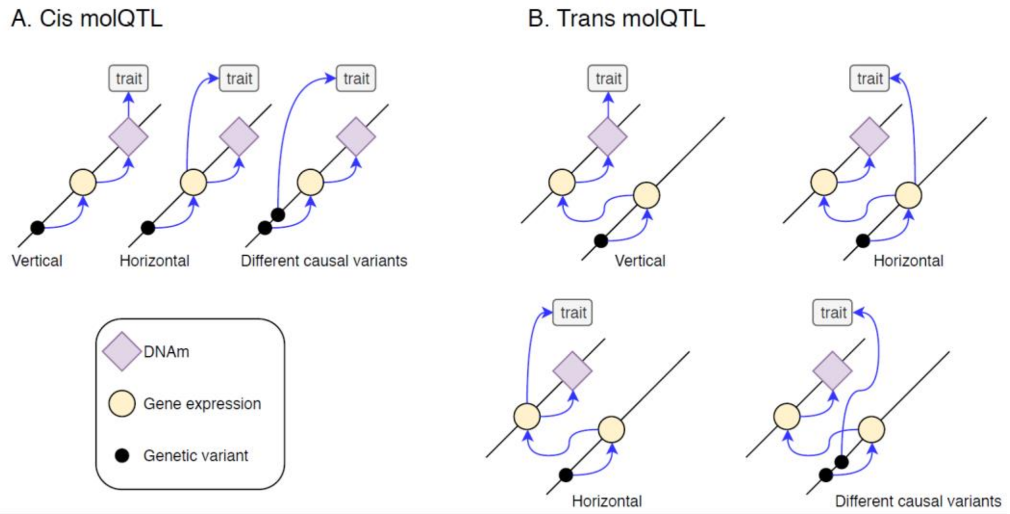


Figure 3


A. Cis molQTL
B. Trans molQTL
trait
Vertical
Horizontal
Different causal variants
DNAm
Gene expression
Genetic variant