Published in final edited form as:
Epigenomics. 2015 ; 7(8): 1239–1243. doi:10.2217/epi.15.88.

Mendelian randomization: applications and limitations in epigenetic studies

Caroline L Relton and **George Davey Smith**
MRC Integrative Epidemiology Unit, University of Bristol, Bristol, BS8 2BN, UK

Keywords

causal inference; DNA methylation; epidemiology; epigenetics; mediation; SNP

> "…extensions of the Mendelian randomization approach offer a potentially fruitful method for strengthening causal inference in epigenetic studies."

There is a rapidly growing body of literature to demonstrate that a range of environmental, behavioral and social exposures can alter the epigenome [1,2]. The plastic nature of epigenetic patterns in the face of exposures means that although epigenetic variation may be associated with phenotypic traits, it can be difficult to disentangle cause from consequence. DNA methylation and other epigenetic marks can be considered as intermediate phenotypes and like many other intermediate phenotypes, they are vulnerable to confounding by the 'usual' factors; age, sex, socioeconomic position, diet, smoking, alcohol intake, etc.

Epigenetic variation is highly likely to play a role in a range of traits and diseases, with emerging evidence for a role in cancer [3], neurological diseases (Parkinson's disease, Alzheimer's disease, bipolar disorder), obesity, atopy and other diseases [4–6]. In many instances the associations observed between epigenetic variation and disease are correlations without robust evidence of causality. Indeed, in many situations epigenetic variation may be a consequence of disease rather than a cause.

Mendelian randomization (MR) is a method that can be applied to strengthen causal inference [7,8]. The MR approach is predicated on the principle that if a genetic variant (e.g., *FTO*) either alters the level of, or mirrors the biological effects of, an environmentally modifiable exposure (e.g., obesity) that itself alters disease risk (e.g., blood pressure), then this genetic variant should also be related to disease risk to the degree predicted by the joint effects of the genetic variant on the modifiable exposure and of the modifiable exposure with the outcome. Instrumental variable methods of analysis [9] can be applied in the MR setting to produce quantitative estimates of the magnitude (with confidence intervals) of the causal influence of the modifiable exposure on a health outcome. Genetic variants that have a well-

Author for correspondence: Tel.: +44 0117 331 4028, caroline.relton@bristol.ac.uk.

Financial & competing interests disclosures
The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.
No writing assistance was utilized in the production of this manuscript.

characterized biological function (or are markers for such variants) can, therefore, be utilized to estimate the causal effect of a suspected exposure on disease risk [10]. The variants should not have an association with the disease outcome except through their link to the risk process of interest. The advantages of adopting a MR approach are documented in detail elsewhere [7,8] but apply equally in the context of epigenetic traits, namely that this method overcomes confounding, and reverse causation. The latter is of particular relevance to epigenetic studies where reverse causation has the potential to be a major issue (i.e., the trait or disease state itself alters the epigenome, not *vice versa*).

Analogous to MR are approaches that have been termed ‘genetical genomics’, which have utilized *cis* genetic variation related to transcription abundance (i.e., levels of mRNA) to identify which RNAs are causally related to disease [11]. In this setting RNAs are treated as an intermediate phenotype that lie between genetic variation and disease-related phenotype [12]. As with intermediate phenotypes in an MR framework, RNA levels can also be associated with confounding factors and suffer from reverse causation, being phenotypic rather than genotypic. It is on this foundation that the use of genetic variation as a causal anchor has been extended to epigenetic studies. Thus, *cis* genetic variation related to DNA methylation levels can be used to establish whether DNA methylation levels at a particular site in the genome are causally related to disease. Other close relations of MR have been recognized and applied in epigenetic studies, including the causal inference test [13], although this does not provide a quantified estimate of the causal effect. This particular test is also vulnerable to measurement error in the mediator, which can lead to incorrect inference [14], in contrast to MR-based approaches to mediation, and should be interpreted with caution.

The application of MR in an epigenetic context relies (in most scenarios) upon the identification of *cis* genetic variants that proxy for DNA methylation levels [15]. It has been demonstrated that DNA methylation patterns often correlate closely with local genetic variants. Studies of human brain tissue have demonstrated that a large proportion of interindividual variation in DNA methylation is associated with common *cis*-acting genetic variation [16]. This was corroborated in an extensive genomic, epigenomic and transcriptomic analysis of HapMap cell line DNA which reported a predominance of *cis*-acting SNPs with respect to DNA methylation levels, as opposed to more distal *trans* effects [17]. More recently, genome-wide association studies of peripheral blood DNA in large sample series have been reported.

MR can be applied in multiple different contexts to interrogate causal relationships in epigenetic studies. Firstly, many (but not all) of the genetic proxies, or instrumental variables, validated to date could be applied in a conventional MR approach, where an epigenetic trait is considered as the outcome. For example, genetic proxies for smoking behavior, alcohol consumption, BMI, or lipid profiles have been used widely in MR studies and could be applied to assess whether these environmentally modifiable factors impact upon the epigenome [18–21]. If multiple genetic variants are known to proxy for an exposure of interest, then these can be combined into an allele score and implemented as a combined instrumental variable.

The first task in assessing the relationship between an exposure and DNA methylation is identifying which part(s) of the epigenome are associated with the exposure of interest. An epigenome-wide association study (EWAS) approach can be applied to interrogate potential associations between a given environmental exposure and DNA methylation [22]. This approach is increasingly commonly applied and many examples of EWAS with respect to a wide variety of traits or exposures can be found in the literature [5–6,13]. Alternatively, a candidate region of the epigenome could be selected based on other evidence. An extension to the application of conventional MR, which aims to explore the causal relationship between a modifiable exposure and the epigenome, is a bidirectional approach. This involves instrumental variable analysis from exposure to epigenome and from epigenome to exposure.

A major advantage of MR is that it can be executed in large datasets that have genotype data available without the necessity to have measured the exposure or trait for which the genetic proxy is known. Indeed, the discovery of a genetic proxy for use in MR ought to be established in an independent dataset prior to its application in instrumental variable analysis. This has useful implications in the field of epigenetics where the generation of DNA methylation can be costly and sometimes impractical. For example, the association between DNA methylation and BMI was interrogated using a *cis*-SNP that proxied for methylation at the *HIF3A* locus; the association between this *cis*-SNP and BMI was analyzed using data from the publicly available GIANT consortium where 123,791 individuals had both genotype and outcome (BMI) data available [5].

A substantial challenge facing the field of epigenetics is the issue of tissue specificity, that is, the observation that DNA methylation patterns differ between tissues and for obvious reasons large scale population-based studies rely heavily on DNA collected from readily accessible, minimally invasive collection methods (usually being peripheral blood or saliva). In what will generally be small studies, in some cases it will be possible to show that the genetic proxy identified demonstrates the same degree of association with DNA methylation in the tissue type of interest. For example, in the study of BMI and *HIF3A* by Dick *et al.* the *cis*-SNP proxy for methylation identified in peripheral blood was also associated with DNA methylation in adipose tissue and skin [5]. In another study of genetic association of DNA methylation in lung tissue, a high degree of concordance in *cis*-SNPs was observed across other tissues (skin and peripheral blood DNA) [23]. This suggests that MR may be applicable across multiple tissue types, or at least inferences can be made regarding causal relationships across tissues; that is, this may inform about causal mechanisms common to more than one tissue but not tissue specific causality. This requires robust assessment in a suitable experimental setting.

One potentially powerful approach for assessing the role of epigenetic processes as mediators is to implement a two-step epigenetic MR strategy [15]. In step 1, the causal impact of an exposure on epigenetic signatures is established. In the second step the causal nature of these epigenetic markers on a health-related outcome is interrogated. In the first step of a two-step epigenetic MR approach the use of a genetic variant as an instrumental variable follows the very same principles as conventional MR. The difference here is solely the consideration of DNA methylation as the trait of interest rather than another trait or disease endpoint. In the second step, a genetic proxy for methylation levels is required. This may take the form of a *cis*-SNP, that is, a SNP in the vicinity of the CpG site that correlates with methylation levels. It is possible to locate such potential SNPs or proxies by interrogating the SNP architecture flanking the CpG site or differentially methylated region of interest and assessing the correlation between methylation levels at the site of interest and genotype. However, data from methylation quantitative trait locus (mQTL) mapping will be readily accessible in the near future, making available a catalogue of SNPs associated with methylation at specific CpG sites across the genome, so it should be possible simply to 'look up' *cis*-SNPs in the genomic region of interest to identify potential genetic variants that can be used as instrumental variables. A dual-step instrumental variables (IV) analysis is required to execute the two-step MR approach the details of which are beyond the scope of this article and can be found elsewhere [7–8,10,15].

It is widely acknowledged that MR has certain limitations and these apply equally to the application of MR in an epigenetic context [15]. It is recognized that the application of MR requires large sample sizes, which is due to the fact that in many instances a genetic variant proxying for an exposure or trait may only explain a very small proportion of variance in that exposure or trait. Therefore, in order to acquire precise risk estimates, sample sizes of the order of magnitude of thousands are generally required. This is often not the case with regard to *cis*-SNPs that tag CpG sites, although it remains an issue where a trait is the focus of MR in conventional approaches.

MR can only be executed if a genetic proxy can be identified. Although GWAS studies are rapidly increasing the number of genetic variants associated with traits of interest and, therefore, potential proxies for MR, there remain situations where genetic variants cannot be identified. This issue is as relevant to the identification of *cis*-SNPs to tag DNA methylation as it is to SNPs that proxy for other traits or exposures. It is also noteworthy that to avoid overfitting of data, an instrumental variable must be established in an independent sample before being applied in the main study. This requirement for separate datasets can be limiting especially if DNA methylation data are not readily available from more than one source.

It is well documented that some of the more popular platforms used to quantify DNA methylation (namely, the Illumina Infinium HumanMethylation450 BeadChip) include a considerable number of CpG sites that are polymorphic. Furthermore, many of the probes that are used to recognize and anneal to DNA sequence using this method also harbor polymorphisms in their sequence. This can lead to bias in measurement of DNA methylation. Filtering methods are usually implemented to remove many of these methylation SNPs (i.e., SNPs that directly create or ablate a methylation site) and polymorphisms in probes (PiPs) from datasets. This has potentially important implications for the choice of a genetic proxy for MR, as unfiltered data may highlight spurious genetic associations. Neither *cis*-SNPs that are obligatory methylation SNPs nor PiPs should be selected for use as instrumental variables in MR.

In epidemiological studies the identification and analysis of mediation is often a key focus. For example, higher BMI is associated with elevated risk of coronary heart disease (CHD), and some of this association may reflect a causal influence of BMI on blood pressure, which, in turn, influences CHD risk. In this situation blood pressure would be a partial mediator of the influence of BMI on CHD, with the important implication that therapeutically modifying blood pressure could break this link. It should be noted that particular sources of bias and confounding can occur in such mediation analyses and measurement error in the mediator that will distort interpretations of such data [24]. These issues are relevant to the consideration of DNA methylation as a mediating mechanism.

Although predicated on the now well established MR framework, concrete examples of the application of MR in epigenetics are still relatively limited [6,25–27]. There is, however, evidence to suggest that there are many potential applications; for example, there is support for gene-specific DNA methylation being associated with exposures including benzene, air pollution, arsenic, cigarette smoking and alcohol drinking [28]. These exposure–methylation associations could be interrogated utilizing conventional MR. It must be recognized, however, that reported associations between environmental factors and DNA methylation are often modest in size and lack the robustness of equivalent contemporary genetic association studies. For the second step in a two-step framework, robust genetic instruments for DNA methylation, such as the CpG sites at the *AHRR* locus associated with smoking exposure, for example [29], will be required. The relationship of variably methylated regions with underlying DNA sequence requires more detailed interrogation to elicit a greater number of *cis*-acting variants robustly associated DNA methylation levels.

Tissue specificity is clearly an important aspect of epigenetic investigation as genotype must only be partially correlated with DNA methylation patterns in any one tissue to allow tissue-specific methylation signatures to exist on a background of uniform genotype. Therefore, genetic proxies for methylation levels may be tissue specific and ought to have tissue specific validation if being used in samples from tissues other than peripheral blood DNA. Detailed information of DNA methylation patterns across multiple tissues is gradually becoming available as initiatives to sequence reference methylomes gain momentum [30]. It is possible to assess the relationship between genetic variation and DNA methylation in a tissue-specific manner and this will be facilitated by the availability of openly accessible data sources. Limitations include that these data are often generated on diseased tissue and that they usually have very little information available on environmental exposures or other relevant covariates.

Pleiotropy, where a genetic variant has more than one direct correlate that would invalidate conclusions based on the assumption of a single pathway, is an important issue in MR. One strategy to overcome this is the use of multiple genetic instruments (including potentially many combinations of independent instruments). In some cases, it may be possible to identify two separate genetic variants, which are not in linkage disequilibrium with each other, but which both serve as proxies for the exposure of interest. If both variants are related to the outcome of interest and point to the same underlying association, then it becomes much less plausible that confounding explains the association, since it would have to be acting in the same way for these two unlinked variants. Methods that utilize multiple instruments in a more involved manner, and allow relaxation of the MR assumptions, are also being developed [31]. The same principles can in theory be applied to the selection of multiple *cis*-SNPs tagging a particular CpG site. This will be contingent on a better understanding of the correlation structure of the methylome and requires rigorous exploration.

In summary, extensions of the MR approach offer a potentially fruitful method for strengthening causal inference in epigenetic studies and these tools are beginning to be applied in contemporary large scale epigenetic studies. There are various options available in terms of which variable is subjected to genetic proxy (exposure, intermediate phenotype or outcome) and all have a place in epigenetic studies, although limitations of MR must be recognized. As epigenetic data continue to be generated the opportunities for the application of MR will increase as will the clarity with which causal inferences can be made.

Acknowledgements

The authors would like to thank members of the MRC Integrative Epidemiology Unit for their contribution in developing this area of research.

The authors' work related to the application of Mendelian randomization to epigenetics is supported by the Medical Research Council Integrative Epidemiology Unit at the University of Bristol (MC_UU_12013/1 and MC_UU_12013/2). This work was also supported by CRUK (grant number C18281/A19169).

References

Papers of special note have been highlighted as:

• of interest; •• of considerable interest

1. Mathers JC, Strathdee G, Relton CL. Induction of epigenetic alterations by dietary and other environmental factors. Adv Genet. 2010; 71:3–39. [PubMed: 20933124]

2. Cortessis VK, Thomas DC, Levine AJ, et al. Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. Hum Genet. 2012; 131:1565–1589. [PubMed: 22740325]

3. Barrow TM, Michels KB. Epigenetic epidemiology of cancer. Biochem Biophys Res Commun. 2014; 455:70–83. [PubMed: 25124661]

4. Urdinguio RG, Sanchez-Mut JV, Esteller M. Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. Lancet Neurol. 2009; 8:1056–1072. [PubMed: 19833297]

5. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: a genome-wide analysis. Lancet. 2014; 383:1990–1998. [PubMed: 24630777]

6. Liang L, Willis-Owen SA, Laprise C, et al. An epigenome-wide association study of total serum immunoglobulin E concentration. Nature. 2015; 520(7549):670–674. [PubMed: 25707804] [**• The first major epigenome-wide association study to implement formal Mendelian randomization.**]

7. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003; 32:1–22. [PubMed: 12689998] [**•• Details the application of Mendelian randomization.**]

8. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014; 23(R1):R89–R98. [PubMed: 25064373] [**•• Recent developments and refinements in the application of Mendelian randomization.**]

9. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol. 2000; 29:722–729. [PubMed: 10922351]

10. Sheehan NA, Didelez V, Burton PR, Tobin MD. Mendelian randomisation and causal inference in observational epidemiology. PLoS Med. 2008; 5:e177. [PubMed: 18752343]

11. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet. 2001; 17:388–391. [PubMed: 11418218]

12. Schadt EE, Lamb J, Yang X, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005; 37:710–717. [PubMed: 15965475]

13. Liu Y, Aryee MJ, Padyukov L, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013; 31:142–147. [PubMed: 23334450]

14. Blakely T, McKenzie S, Carter K. Misclassification of the mediator matters when estimating indirect effects. J Epidemiol Community Health. 2013; 67:458–466. [PubMed: 23386673]

15. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int J Epidemiol. 2012; 41:161–176. [PubMed: 22422451] [**•• Outlines the application of Mendelian randomization to epigenetics to strengthen causal inference.**]

16. Zhang D, Cheng L, Badner JA, et al. Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet. 2010; 86:411–419. [PubMed: 20215007]

17. Bell JT, Pai AA, Pickrell JK, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011; 12:R10. [PubMed: 21251332]

18. Timofeeva MN, McKay JD, Davey Smith G, et al. Genetic Polymorphisms in 15q25 and 19q13 loci, cotinine levels, and risk of lung cancer in EPIC. Cancer Epidemiol Biomarkers Prev. 2011; 20:2250–2261. [PubMed: 21862624]

19. Zuccolo L, Fitz-Simon N, Gray R, et al. A non-synonymous variant in ADH1B is strongly associated with prenatal alcohol use in a European sample of pregnant women. Hum Mol Genet. 2009; 18:4457–4466. [PubMed: 19687126]

20. Timpson NJ, Harbord R, Davey Smith G, Zacho J, Tybjaerg-Hansen A, Nordestgaard BG. Does greater adiposity increase blood pressure and hypertension risk? Mendelian randomization using the FTO/MC4R genotype. Hypertension. 2009; 54:84–90. [PubMed: 19470880]

21. Benn M, Tybjaerg-Hansen A, Stender S, Frikke-Schmidt R, Nordestgaard BG. Low-density lipoprotein cholesterol and the risk of cancer: a Mendelian randomization study. J Natl Cancer Inst. 2011; 103:508–519. [PubMed: 21285406]

22. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011; 12:529–541. [PubMed: 21747404]

23. Shi J, Marconett CN, Duan J, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun. 2014; 5:3365. [PubMed: 24572595]

24. Cole DA, Preacher KJ. Manifest variable path analysis: potentially serious and misleading consequences due to uncorrected measurement error. Psychol Methods. 2014; 19:300–315. [PubMed: 24079927]

25. Groom A, Potter C, Swan DC, et al. Postnatal growth and DNA methylation are associated with differential gene expression of the *TACSTD2* gene and childhood fat mass. Diabetes. 2012; 61:391–400. [PubMed: 22190649]

26. Allard C, Desgagne V, Patenaude J, et al. Mendelian randomization supports causality between maternal hyperglycaemia and epigenetic regulation of leptin gene in newborns. Epigenetics. 2015; 10(4):342–351. [PubMed: 25800063]

27. Kirkbride J, Susser E, Kundakovic M, Kresovich JK, Davey Smith G, Relton CL. Prenatal nutrition, epigenetics and schizophrenia risk: can we test causal effects? Epigenomics. 2012; 4(3): 303–315. [PubMed: 22690666]

28. Terry MB, Delgado-Cruzata L, Vin-Raviv N, Wu HC, Santella RM. DNA methylation in white blood cells: association with risk factors in epidemiologic studies. Epigenetics. 2011; 6:828–837. [PubMed: 21636973]

29. Zeilinger S, Kühnel B, Klopp N, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS ONE. 2013; 8:e63812. [PubMed: 23691101]

30. Maher B. ENCODE: the human encyclopaedia. Nature. 2012; 489:46–48. [PubMed: 22962707]

31. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015; 44(2):512–525. [PubMed: 26050253]