Harnessing the power of twins in epigenetic association studies: causal inference and more

Qihua Tan*,^{1,2}

¹Epidemiology & Biostatistics, Department of Public Health, University of Southern Denmark, Odense, Denmark

²Unit of Human Genetics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark

*Author for correspondence: Tel.: 0045 65503536; qtan@health.sdu.dk

First draft submitted: 26 November 2019; Accepted for publication: 27 November 2019; Published online: 13 December 2019

The literature of epigenome-wide association studies (EWAS) is growing rapidly in the field of biomedical research. Lots of sites have been identified in significant association with clinical and health traits of interest, with the number of publications increasing nearly exponentially since 2010 [1]. The reported sites are usually claimed as epigenetic biomarkers of biological significance or as potential targets for intervention. It is, however, unfortunate that one important conceptual issue has been missing in making these claims: a significant association does not guarantee a causal effect but just a correlation. Although the hypothesis-free EWASs enable us comprehensive and perhaps unbiased analysis of the epigenome, they are observational studies by nature. In epidemiology, observational studies are considered to have less probative force due to inherent limitation in controlling confounding factors that influence both explanatory (exposure) and (clinical or health) outcome variables, a situation that can result in biased, confusing and even misleading results [2]. Indeed, it is true that association does not prove causation but it is not true that association refutes causation. Lots of efforts have been taken in inferring causality from observational studies [3] and some of them have been proven effective and valuable. For example, the propensity score matching which uses fitted logistic probability for matching samples between comparison groups to approximate randomized controlled trial [4], which is the gold standard for assessing causality in biomedicine.

Among the different causal inference methods, the genetically informed method is a promising approach that fosters efficient causal assessment [5]. By engaging genetically related individuals such as twins, siblings and family members, unobserved genetic and familiar environmental confounding can be controlled to achieve or approach exchangeability, which is essential in ensuring consistent causal inference. In the case of using identical or monozygotic (MZ) twin design [6,7], exchangeability can be sufficiently approximated, an advantage of their perfect sharing of DNA sequence variations and rearing environments. As a matter of fact, the efficient control of genetic and common environmental factors in the MZ twin design enables significant enrichment of statistical power in EWAS, as revealed by our recent computer simulation study [8]; one good reason for the popularity of using disease-discordant MZ twin pairs in epigenetic association studies. Here, it is necessary to point out that the matching-out of genetic and nongenetic (common-exposure) variables in the discordant MZ twin design does not mean that the effects of such variables on disease can no longer be assessed, which is described in the current literature as a limitation of the design [5]. By simple mathematics, we show that the inclusion of pair-specific or common-exposure variables in the analysis of MZ twin data allows estimation of valuable interaction effects with the disease. For example, taking a pair-specific variable age, if there is an age-specific effect of DNA methylation (*Me*) on the diseases (i.e. an interaction effect), we have, for the healthy (-) twin,

$$\log Me(-) = \alpha_0 + \beta_1 age + \varepsilon_-.$$

For the affected (+) twin,

$$\log Me(+) = \alpha_0 + \alpha_+ + (\beta_1 + \beta_{1+})age + \varepsilon_+.$$

Here, ε is a random error term, α_0 is the mean methylation level in healthy controls, α_+ is the mean methylation difference between affected and unaffected healthy twins; β_1 is the main effect of age on DNA methylation and β_{1+} represents additional age effect on DNA methylation, specifically in the affected twins, in other words, an interaction effect. Taking the intrapair difference, we have:

$$\log \left[\frac{Me(+)}{Me(-)} \right] = \alpha_+ + \beta_{1+}age + \varepsilon.$$

In the above model, an estimated β_{1+} significantly >0 or <0 indicates that the differential DNA methylation between disease and control twins increases or decreases with increasing age. Another simple example, the male twin pairs can be, on average, more discordant than the female pairs if coefficient of sex (male = 0, female = 1) is significantly <0 , an indication of sex-dependent effect of DNA methylation on the disease. Even more interestingly, when a genotype at a specific locus is available and included as a pair-specific variable, its significant estimate could suggest a methylation quantitative trait locus associated with the disease.

The frequent use of twins – especially MZ twins – in epigenetic association studies indeed helps to control unobserved confounding factors and minimizes false positive findings [9], but it does not guarantee causality as in randomized controlled trial because of potential confounding factors beyond the control of the MZ twin design, for example, nonshared or individual environmental factors. Moreover, even if a causal relationship is established, the statistical models for association analysis do not provide direction of causation – an important consideration due to possible reverse causation. The latter can be avoided when the exposure variable is free from reverse like a germline genetic variant in genome-wide association study, but unfortunately, not in EWAS. In fact, the specific genetic variants a person is born with serve as instrumental variables in Mendelian randomization, a well-known approach for causal inference from observational studies [10]. Determining the direction of causation is of high clinical and etiological relevance. For example, in aging research, one could easily ask if an identified age-dependent methylation change at a genomic site is the cause of aging or just in response to aging. Through examining the cross-trait cross-pair correlation, Hopper and colleagues performed inference on causation from examination of familial confounding (ICE FALCON) using regression analysis assigning the co-twin as a ‘negative control’ [11]. If the associations between the outcome of twin A and the predictors of both the twin A and co-twin B remain unchanged before and after adjusting for each other, then no evidence of causal relationship is given. On the other hand, if there is a significant attenuation of the cross-trait cross-twin association after conditioning on twin A or self, there is an evidence ‘consistent with’ some causation. ICE FALCON is analogous to Mendelian randomization in causal assessment. The latter requires genotype data but can be applied to unrelated individuals. ICE FALCON can use twin (monozygotic or dizygotic), sibling or other relative pairs with efficiency of inference intuitively decreasing with reduced level of relatedness. The method has been recently applied to EWASs and reported causal effects of smoking [12] and body mass index [13] on site-specific DNA methylation variations. We have applied ICE FALCON to infer causality between DNA methylation and gene expression in MZ twins and discovered a large number of genes with promoter methylation displaying causal effects on expression activity.

Given the frequent use of the twin design in EWAS, it is highly recommended that, by taking further advantage of using twins, causal inference be performed on important sites to establish causal relationships before claiming them as molecular targets because noncausal markers are meaningless for clinical or preventive intervention, although they could be used for prediction.

Financial & competing interests disclosure

The author has no 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. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

1. Li M, Zou D, Li Z *et al.* EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.* 47(D1), D983–D988 (2019).
2. Rush CJ, Campbell RT, Jhund PS, Petrie MC, McMurray JJV. Association is not causation: treatment effects cannot be estimated from observational data in heart failure. *Eur. Heart J.* 39(37), 3417–3438 (2018).

3. Listl S, Jürges H, Watt RG. Causal inference from observational data. *Community Dent. Oral Epidemiol.* 44(5), 409–415 (2016).
4. Nguyen MH, Yang HI, Le A *et al.* Reduced incidence of hepatocellular carcinoma in cirrhotic and noncirrhotic patients with chronic hepatitis B treated with Tenofovir – a propensity score-matched study. *J. Infect. Dis.* 219(1), 10–18 (2019).
5. Pingault JB, O'Reilly PF, Schoeler T, Ploubidis GB, Rijdsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* 19(9), 566–580 (2018).
6. McGue M, Osler M, Christensen K. Causal inference and observational research: the utility of twins. *Perspect. Psychol. Sci.* 5(5), 546–556 (2010).
7. Tan Q. The epigenome of twins as a perfect laboratory for studying behavioural traits. *Neurosci. Biobehav. Rev.* 107, 192–195 (2019).
8. Li W, Christiansen L, Hjelmborg J, Baumbach J, Tan Q. On the power of epigenome-wide association studies using a disease-discordant twin design. *Bioinformatics* 34(23), 4073–4078 (2018).
9. Tan Q, Christiansen L, von Bornemann Hjelmborg J, Christensen K. Twin methodology in epigenetic studies. *J. Exp. Biol.* 218, 134–139 (2015).
10. Koellinger PD, de Vlaming R. Mendelian randomization: the challenge of unobserved environmental confounds. *Int. J. Epidemiol.* 48(3), 665–671 (2019).
11. Dite GS, Gurrin LC, Byrnes GB *et al.* Predictors of mammographic density: insights gained from a novel regression analysis of a twin study. *Cancer Epidemiol. Biomarkers Prev.* 17(12), 3474–3481 (2008).
12. Li S, Wong EM, Bui M *et al.* Causal effect of smoking on DNA methylation in peripheral blood: a twin and family study. *Clin. Epigenetics* 10, 18 (2018).
13. Li S, Wong EM, Bui M *et al.* Inference about causation between body mass index and DNA methylation in blood from a twin family study. *Int. J. Obes. (Lond.)* 43(2), 243–252 (2019).