frontiers in Genetics

REVIEW
published: 10 July 2018
doi: 10.3389/fgene.2018.00238

Application of Causal Inference to Genomic Analysis: Advances in Methodology

Pengfei Hu[1], *Rong Jiao*[2], *Li Jin*[3,4]* *and Momiao Xiong*[2]*

[1] Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China, [2] Department of Biostatistics and Data Science, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, United States, [3] State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China, [4] Human Phenome Institute, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Mariza De Andrade,
Mayo Clinic, United States

Reviewed by:
Kui Zhang,
Michigan Technological University, United States
Yuehua Cui,
Michigan State University, United States

***Correspondence:**
Li Jin
lijin@fudan.edu.cn
Momiao Xiong
momiao.xiong@uth.tmc.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 15 March 2018
Accepted: 14 June 2018
Published: 10 July 2018

Citation:
Hu P, Jiao R, Jin L and Xiong M (2018) Application of Causal Inference to Genomic Analysis: Advances in Methodology. Front. Genet. 9:238. doi: 10.3389/fgene.2018.00238

The current paradigm of genomic studies of complex diseases is association and correlation analysis. Despite significant progress in dissecting the genetic architecture of complex diseases by genome-wide association studies (GWAS), the identified genetic variants by GWAS can only explain a small proportion of the heritability of complex diseases. A large fraction of genetic variants is still hidden. Association analysis has limited power to unravel mechanisms of complex diseases. It is time to shift the paradigm of genomic analysis from association analysis to causal inference. Causal inference is an essential component for the discovery of mechanism of diseases. This paper will review the major platforms of the genomic analysis in the past and discuss the perspectives of causal inference as a general framework of genomic analysis. In genomic data analysis, we usually consider four types of associations: association of discrete variables (DNA variation) with continuous variables (phenotypes and gene expressions), association of continuous variables (expressions, methylations, and imaging signals) with continuous variables (gene expressions, imaging signals, phenotypes, and physiological traits), association of discrete variables (DNA variation) with binary trait (disease status) and association of continuous variables (gene expressions, methylations, phenotypes, and imaging signals) with binary trait (disease status). In this paper, we will review algorithmic information theory as a general framework for causal discovery and the recent development of statistical methods for causal inference on discrete data, and discuss the possibility of extending the association analysis of discrete variable with disease to the causal analysis for discrete variable and disease.

Keywords: causal inference, genomic analysis, additive noise models for discrete variables, association analysis, entropy

INTRODUCTION

By February 6th, 2017, a catalog of published genome-wide association studies (GWAS) had reported significant association of 26,791 SNPs with more than 1,704 traits in 2,337 publications (A catalog of Published Genome-Wide Association Studies, 2017)[1]. Many of these associated SNPs are non-coding variants (Timpson et al., 2017). Despite significant progress in dissecting the genetic architecture of complex diseases by GWAS, understanding the etiology and mechanism of complex diseases remains elusive. It is known that significant findings of association analysis are[1] lacking in consistency and often proved to be controversial (Valente et al., 2015; Wakeford, 2015). Complex diseases are often caused by different genetic mutations, have complex and multimodal genetic etiology, and show substantial phenotype heterogeneity in morphology, physiology and behavior (Brookes and Robinson, 2015). There are multiple steps between genes and phenotypes. Each step may be influenced by genomic variation and can weaken links between genes and phenotypes. As a consequence, this will obscure the causal mechanisms of disease. The recent study finding "association signals tend to be spread across most of the genome" again shows that association signals provide limited information on causes of disease, which calls the future of the GWAS into question (Boyle et al., 2017; Callaway, 2017). Association and causation are different concepts (Altman and Krzywinski, 2015). Association of a genetic variant with the disease is to characterize the dependence between the genetic variant and disease, while causation from the genetic variant to the disease is to indicate that the presence of the genetic variant will produce effect and cause disease. Observed association may not infer a causal relationship and the lack of association may not be necessary to imply the absence of a causal relationship (Wakeford, 2015). Finding causal SNPs only by searching the set of associated SNPs may miss many causal variants and may not be an effective research direction in genetics. The dominant use of association analysis for genetic studies of complex diseases is a key issue that hampers the theoretical development of genomic science and its applications to discovery of mechanisms of diseases in practice. Causality shapes how we view and understand mechanism of complex diseases (Gottlieb, 2017). In addition, causal models can also be used to directly predict the results of intervention, but association usually cannot.

It is time to develop a new generation of genetic analysis to shift the current paradigm of genetic analysis from association analysis to causal inference. To make the shift feasible, we need (1) to develop novel causal inference methods for genome-wide and epigenome-wide causal studies of complex disease; (2) to develop unified frameworks for systematic casual analysis of integrated genomic, epigenomic, imaging and clinical phenotype data and to infer multilevel omics and image causal networks for the discovery of paths from genetic variants to diseases. The focus of this paper is to survey statistical methods for causal inference and explore the potential to use modern causal inference theory for developing statistics for genome-wide causal studies (GWCS) of complex diseases.

[1]A catalog of Published Genome-Wide Association Studies. 2017. Available from: https://www.genome.gov/page.cfm?pageid=26525384&clearquery=1#result_table

CAUSAL MARKOV CONDITIONS

The widespread view about causation in genetic epidemiology field is that only interventions to a system can discover causal relations. Many epidemiologists and statistical geneticists have doubt about the possibility of using observational data to identify disease-causing variants and "shied away" from causal inference based on observation data (Janzing et al., 2016). In the past decade, great progress in causal inference has been made. In contrast to classical statistics where the relationships between random variables are measured by statistical dependence or association, the algorithms for causal inference that are designed to discover the data generating processes based on statistical observations have been developed.

The classical causal inference theory assumes the Causal Markov conditions to infer causal relationships among multiple variables and connect causality with statistics (Janzing and Schölkopf, 2010). Consider a set of variables $V = \{X_1, \ldots, X_n\}$. Let $G = \{V, E\}$ be a directed acyclic graph (DAG) where nodes in the DAG represent the random variables X_i and an arrow from node X_i to node X_j denotes a causal direction. Let $P(X_1, \ldots, X_n)$ be a joint probability distribution of variables $X_1, \ldots, X_n$. The Markov condition describes the causal structure of the DAG G. The DAG can also be characterized by the concept of parent sets. Let pa_i and ND_i be the set of parents of the node X_i, respectively. The Markov conditions can be formulated as the following three forms (Janzing and Schölkopf, 2010):

1. **Factorization of the joint distribution**:
   The joint distribution $P(X_1, \ldots, X_n)$ can be factorized into the conditional distribution:
   $$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|pa_i), \quad (1)$$
   where $P(X_i|pa_i)$ is the conditional probability of X_i, given the values of all parents of X_i.
2. **Local markov condition**:
   Every node is conditionally independent of its non-descendants, given its parents, i.e.,
   $$X_i \perp\!\!\!\perp ND_i|pa_i,$$
   where ND_i denotes the non-descendent of the node i .
3. **Global markov condition:**
   Consider three datasets X, Y and Z. If X and Y are d-separated by Z then we have
   $$X \perp\!\!\!\perp Y \,|Z.$$
   To assess conditional independence in a more general case, we introduce a useful concept, d-separation which associates "correlation" with "connectedness" and independence with "separation". Two sets of variables X and Y are d-separated by a third set of variables Zif and only if
   (1) there is a path $i \to m \to j$ where $i \in X$, $m \in Z$ and $j \in Y$ (**Figure 1A**) or

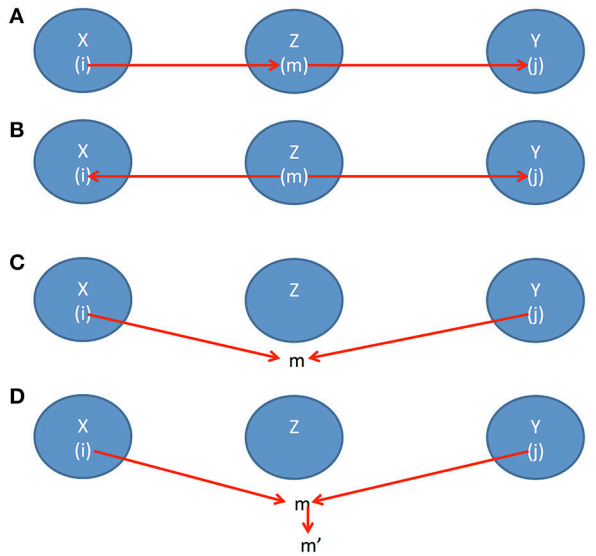

FIGURE 1 | (A) Causal direction 1, **(B)** Causal direction 2, **(C)** Causal direction 3, and **(D)** Causal direction 4.

(2) there is a path $i \leftarrow m \rightarrow j$ where $i \in X$, $m \in Z$ and $j \in Y$ (**Figure 1B**) or
(3) there is path $i \rightarrow m \leftarrow j$ where $i \in X$, $j \in Y$, but m is not in Z (**Figure 1C**) or
(4) there is path $i \rightarrow m \leftarrow j$ where $i \in X$, $j \in Y$, but m and its descendants are all not in Z (**Figure 1D**).

In the DAG G, we defined the conditional densities $P(x_i|pa_i)$ as the Markov kernels. The set of Markov kernels defines a Markovian density. However, in general, the Markov conditions cannot uniquely determine causal graphs. Many different DAGs satisfy the same set of independence relations. For example, consider three simple DAGs: $x \rightarrow y \rightarrow z$, $x \leftarrow y \leftarrow z$ and $x \leftarrow y \rightarrow z$. Three variables x, y and $z$in all three DAGs satisfy the same conditional independent distribution: x and z are independent, given$y$. **Figure 2** shows three different DAGs that share the same conditional independent distributions: $B \perp\!\!\!\perp C|A$ and $A \perp\!\!\!\perp D|(B, C)$. These examples show that multiple graphs may satisfy the Markov conditions. Except for Markov conditions, we need other constraints for learning causal structure.

The faithfulness condition is another constraint that helps to infer causal structure from observational data. The faithfulness condition requires that conditional independences in the distribution correspond to the d-separation in the DAG one by one. In other words, faithfulness condition requires that every conditional independence in the distribution must correspond to the Markov condition that is applied to the DAG (Peters et al., 2017; Xiong, 2018).

ALGORITHMIC COMPLEXITY FOR CAUSAL INFERENCE

Markov condition approaches to causal inference have two serious limitations. First, the approaches based on Markov condition and faithfulness assumption can only identify the graph up to its Markov equivalence class (Janzing and Schölkopf, 2010). Second, the Markov condition approaches need sampling. However, some cases require causal inference for single observation where the Markov condition approaches cannot be applied.

To overcome these limitations, we need to develop methods for causal inference without resorting to probability theory. One approach is to infer causation via algorithmic information theory. We begin with introduction of algorithmic information theory. Consider two 20-bit binary strings:

$$\mathrm{S} = 10101010101010101010$$
$$\mathrm{S} = 00011101001000101101,$$

which are equally likely to represent the results of 20 flips of coin. However these two strings have large difference in the complexity between them. The first string has a short description: a 20-bit string with 1 in position in odd number, or can be described as "10 10 times," which consisting of 11 characters. The second string has no obvious simple description. Therefore, the length of the shortest description of the first and second strings are 11 and 20, respectively.

For any string x, we define the Kolmogorov complexity $K(x)$ as the length of the shortest program that generates x, denoted as x^*, using universal prefix Turing machine (Janzing and Schölkopf, 2010). The conditional Kolmogorov complexity $K(x|y)$ of a string x given another string y is defined as the length of the shortest program that can generate x from y. The joint Kolmogorov complexity $K(x, y)$ is defined as the complexity of concatenation $x'y$ where x' denotes a prefix code of x. Mutual information and Markov conditions can be extended to algorithmic mutual information and Algorithmic Markov conditions (Supplementary Note A).

ADDITIVE NOISE MODELS

In the previous section, we introduce the Markov conditions and faithfulness, which are the constrained-based approaches. These constrained-based approaches cannot identify the unique causal solution or make causal inference between two variables. However, the genome wide association studies (GWAS) test the association of single SNP with the disease. GWAS investigates dependence between two variables. Similar to GWAS, the genome-wide causation studies (GWCS) needs to discover the SNP that causes disease. We need to develop statistical methods for bivariate causal discovery. The DAG approach requires at least variables for causal inference. Therefore, the classical Bayesian networks and DAG-based causal inference methods cannot be applied to the GWCS. To overcome these limitations, some authors (Kano and Shimizu, 2003; Shimizu et al., 2006; Peters et al., 2011) proposed additive noise models (ANMs):

$$Y = f(X) + N,\ N \perp\!\!\!\perp X, \tag{2}$$

where f is a function, and N is noise that is independent of the cause X. It is clear that the conditional distribution $P(Y|X)$ of the

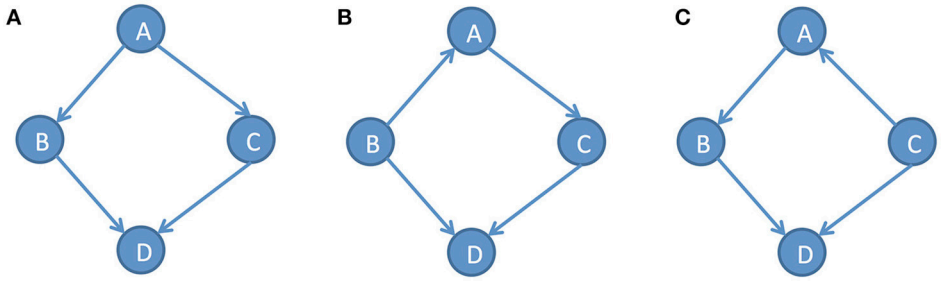

FIGURE 2 | Different DAGs can have same set of conditional independence distributions. **(A)** DAG example A, **(B)** DAG example B, **(C)** DAG example C.

response , given X is equal to the noise distribution $P_N(Y-f(X))$. If $X \to Y$ then we should have (Janzing and Steudel, 2010)

$$I(P(X) : P_N\left(Y - f(X)\right) = 0. \tag{3}$$

Consider two ANMs: integer models and cyclic models (Peters et al., 2011).

Integer Models

Consider two random variables X and Y that take integer values ($\mathbb{Z}$). The support can be either infinite or finite. Consider an ANM from $\to Y$:

$$Y = f(X) + N,\ N \perp\!\!\!\perp X, \tag{4}$$

where f is a function $f : \mathbb{Z} \to \mathbb{Z}$ and N is a noise variable. Let $n\left(l\right) = P(N = l)$. In (4), we further assume $n(0) \geq n(j)$ for all $j \neq 0$. If there is also an ANM to fit the data, then the ANM is referred to as reversible.

Cyclic Models

We first define the ring $\mathbb{Z}/m\mathbb{Z}$ as the set of remainders modulo m, i.e., $\mathbb{Z}/m\mathbb{Z} = \{[0], [1], \ldots, [m-1]\}$.

Now we consider random variables that can take values in a periodic domain. Formally, we define m-cyclic random variable as the variable that takes values in $\mathbb{Z}/m\mathbb{Z}$. Let X and Y be m and k-cyclic random variables, respectively. Define an ANM from X to Y:

$$Y = f(X) + N,\ N \perp\!\!\!\perp X, \tag{5}$$

where f is a function $f : \mathbb{Z}/m\mathbb{Z} \to \mathbb{Z}/k\mathbb{Z}$, N is a k-cyclic noise. We assume $n(0) \geq n(j)$ for all $j \neq 0$.

Similarly, we can define an ANM from Y to :

$$X = g(Y) + N_x,\ N_x \perp\!\!\!\perp Y, \tag{6}$$

where $g : \mathbb{Z}/k\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$ and N_x is a m-cyclic noise.

If both the ANM from $X \to Y$ and ANM from Y to X hold then the ANM is called reversible. The selection of the integer ANM or cyclic ANM mainly depends on the target Y domain.

The cyclic models can be used for genetic studies of complex disease. Let Y be a binary trait to indicate the disease status of the individual. Thus, Y is a 2-cyclic variable. Thepotential cause variable X denotes the genotype of the individual. Thus, X is a 3-cyclic variable.

IDENTIFIABILITY

In most cases, nature selects one causal direction. The causal inference principle states that if Y satisfies the ANM from X to Y, but does not satisfy the ANM from Y to X, then we infer X to be the cause for Y, which is denoted by $X \to Y$. In some cases, Y satisfies ANMs in two directions and hence causation does not exist. The question is under what conditions a joint distribution admits an ANM in at most one direction.

The causal inference principle implies that if $X \to Y$, then the conditional probability $P(Y|X)$ does not depend on the cause X. Several examples to illustrate identifiability are given in Supplementary Note B.

GENETIC ASSOCIATION ANALYSIS AND CAUSATION ANALYSIS

It is well known that correlation (association) and causation are different concepts. Correlation does not imply causation and conversely, causation also does not imply correlation. Our experiences demonstrated that large proportions of causal loci cannot be discovered by association analysis. This explains why a large number of genetic variants still remain hidden. The observed association may be in part due to chance, bias and confounding. Furthermore, a recent study found that "association signals tend to be spread across most of the genome," by the debatable omnigenic model, which again shows that association signals provide limited information on causes of disease, calling the future of the GWAS into question (Boyle et al., 2017; Callaway, 2017). An observed association may not lead to inferring a causal relationship and the lack of association may not be necessary to imply the absence of a causal relationship. Finding causal SNPs only via searching the set of associated SNPs may not be an effective research direction in genetics. The dominant use of association analysis for genetic studies of complex diseases hampers the theoretical development of genomic science and its application in practice. The examples showing that association and causation are difference concepts are given in Supplementary Note C.

Algorithm 1 for Causal Genetic Analysis Using ANM:

1. Fit the following nonlinear integer regression to the data.

$$Y = f(X) + N_Y.$$

Calculate the residuals $\widehat{N_Y} = Y - \hat{f}(X)$.

2. Fit the following nonlinear integer regression to the data.

$$X = g(Y) + N_X.$$

Calculate the residuals $\widehat{N_x} = X - \hat{g}(Y)$.

3. Test for independence.

If $\widehat{N_Y}$ and X are independent ($\widehat{N_Y} \perp\!\!\!\perp X$), and $\widehat{N_X}$ and Y are not independent, then X is causing Y ($X \rightarrow Y$)

If both $\widehat{N_Y}$ and X, and $\widehat{N_X}$ and Y are not independent or if both $\widehat{N_Y}$ and X, and $\widehat{N_X}$ and Y are independent, then no causation conclusion can be made.

CAUSAL GENETIC ANALYSIS

The ANMs can be used for casual genetic analysis of complex disease. The procedures are summarized as follows Peters et al. (2011).

The ordinary egression usually minimizes the sum of square of errors. However, here we evaluate the proposed nonlinear function by checking the independence of regressor and the residuals. Therefore, Peters et al. (2011) suggested using a dependence measure (DM) between regressor and residuals as a lost function. If we simultaneously consider multiple SNPs, the ANMs with a variate (SNP) can be extended to the ANMs with multivariate variables (multiple SNPs). We assume that W is a q dimensional vector of variables. An ANM with multiple SNPs for causal genetic analysis is given as follows:

$$Y = f(W) + N,\ W \perp\!\!\!\perp N, \tag{7}$$

where $W = [W_1, \ldots, W_q]$.

Algorithm 1 in Peters et al. (2011) can be adopted to the regression with multiple regressor. The following algorithm for discrete regression with multiple regressor takes the discrete regression with one regressor as its special case.

where ε and T are pre-specified.

For inferring causation involving one vector of variables, distance correlation will be used as *DM*. The problem for testing the independence between cause (regressor) and residuals can be formulated as a $2 \times q$ contingence table (**Table 1**). Let n_0 and n_1 be number of individuals with $N = 0$ and $N = 1$, respectively. Let $n = n_0 + n_1$. Let $g_{j_1 \ldots j_q}$ denote the genotype of q SNPs, $a_{j_1 \ldots j_q}$ and $b_{j_1 \ldots j_q}$ be the number of individuals with genotype $g_{j_1 \ldots j_q}$, and $N = 0$ and $N = 1$, respectively. Define the marginal frequencies: $\frac{n_0}{n}$, $\frac{n_1}{n}$ and $\frac{a_{j_1 \ldots j_q} + b_{j_1 \ldots j_q}}{n}$. Then, we obtain

Algorithm 2 (Distance Regression With Dependence Measure)

Step 1: Calculate the sampling distribution $\hat{P}(W, Y)$.

Step 2: Initialization.

$f^{(0)}\left(w^i\right) = argmax_y \hat{P}\left(W = w^i, Y = y\right),\ t = 0,$

where $w^i = [w_1^i, \ldots,\ w_q^i]$.

Step 3: Repeat

t =t+1;

Step 4: for $i = 1, ..., n$ do

Step 5: $f^{(t)}\left(w^i\right) = argmin_Y DM(W,\ Y - f_{w_i \rightarrow y}^{(t-1)}(W))$

end for

Step 6: until $\left|\left|f^{(t)} - f^{(t-1)}\right|\right|_2 < \varepsilon$ or $\left(\widehat{N} = Y - f^{(t)}(W)\right) \perp\!\!\!\perp W$, or $t = T$,

TABLE 1 | Contingency table for testing independence.

	Genotype $g_{1\,1\,\ldots\,1}$	Genotype ⋯	$g_{j_1\,\ldots j_q}$	Genotype ⋯	Genotype $g_{3.3\,\ldots\,3}$	
$N = 0$	$a_{1\,1\ldots\,1}$	⋯	$a_{j_1 \ldots j_q}$	⋯	$a_{3.3\ldots\,3}$	n_0
$N = 1$	$b_{1\,1\ldots\,1}$	⋯	$b_{j_1 \ldots j_q}$	⋯	$b_{3.3\ldots\,3}$	n_1
	$a_{1\,1\ldots\,1}$ $+b_{1\,1\ldots\,1}$	⋯	$a_{j_1 \ldots j_q}+$ $b_{j_1 \ldots j_q}$	⋯	$a_{3.3\ldots\,3}+$ $b_{3.3\ldots\,3}$	$n = n_0 + n_1$

$E[a_{j_1 \ldots j_q}] = \frac{1}{n} n_0 (a_{j_1 \ldots j_q} + b_{j_1 \ldots j_q})$ and $\mathrm{E}[b_{j_1 \ldots j_q}] = \frac{n_1}{n}(a_{j_1 \ldots j_q} + b_{j_1 \ldots j_q})$. Then, the test statistic for testing independence is defined as

$$T = \sum_{j_1 j_2 \ldots j_q} \left[\frac{\left(\hat{a}_{j_1 \ldots j_q} - E[a_{j_1 \ldots j_q}]\right)^2}{E[a_{j_1 \ldots j_q}]} + \frac{\left(\hat{b}_{j_1 \ldots j_q} - E[b_{j_1 \ldots j_q}]\right)^2}{E[b_{j_1 \ldots j_q}]} \right], \tag{8}$$

where $\hat{a}_{j_1 \ldots j_q}$ and $\hat{b}_{j_1 \ldots j_q}$ are observed values of $a_{j_1 \ldots j_q}$ and $b_{j_1 \ldots j_q}$, respectively. Under the null hypothesis of independence, the test statistic T is distributed as a central $\chi^2_{(3^q - 1)}$ distribution with degrees of freedom $3^q - 1$.

If SNPs involve rare variants or number of SNPs increases, the expected counts of many cells will be small. Fisher's exact test should be used to test for independence.

CAUSATION IDENTIFICATION USING ENTROPY

The ANM assumes that noise or any outside factor (exogenous variable) affects the effect variable additively, i.e.,

$$Y = f(X) + E,\ E \perp\!\!\!\perp X.$$

The ANM can be extended to more general nonlinear model (Kocaoglu et al., 2016). We assume that variable Y is an arbitrary function of X and an exogenous variable E:

$$Y = f(X, E),\ X \perp\!\!\!\perp E, \tag{9}$$

where we assume that the exogenous variable E has low Renyi entropy. Renyi entropy is defined as

$$H_0(E) = \log k, \tag{10}$$

where k is the number of states of the variable E.

In other words, if the model in the wrong causal direction: $X = g(Y, \tilde{E})$, then the exogenous variable $\tilde{E}$ has higher Renyi entropy than that of E, i.e., $H_0(E) \leq H_0(\tilde{E})$.

Since Renyi entropy H_0 is difficult to compute, we can replace Renyi entropy by Shannon entropy H_1. An algorithm for inferring the true causal direction is to find the exogenous variable E with the smallest $H_1(E)$ such that

$$Y = f(X, E),\ X \perp\!\!\!\perp E.$$

Kocaoglu et al. (2016) proposed an algorithm to calculate the Sharon entropy $H_1(E)$. They make argument that

$$P\left(Y = y \middle| X = x\right) = P(f(x, E) = y). \tag{11}$$

If we define $f_x(E) = f(x, E)$, then it follows from Equation (11) that we can use the conditional distribution $P(Y|X)$ to calculate the distribution of E: $P\left(Y = y \middle| X = x\right) = P(f_X(E) = y)$. Algorithm 1 in Kocaoglu et al. (2016) assumed that the number of states of the variables X and Y is equal, which prevents application of Algorithm 1 to causal genetic analysis for genotype data. However, if we consider allele distribution data, Algorithm 1 in Kocaoglu et al. (2016) can still be applied to the causal genetic analysis. Consider the matrix

$$M = \begin{bmatrix} P_y(0) & P_x(0) \\ P_y(1) & P_x(1) \end{bmatrix},$$

where

$$x = \begin{cases} 0 & a \\ 1 & A \end{cases},\ y = \begin{cases} 0 & Normal \\ 1 & disease \end{cases}.$$

Algorithm 1 in Kocaoglu et al. (2016) can then be written as follows.

Step 1. Input matrix $M = \left(P_i(j)\right)_{2\times 2}$.
Step 2. Define $e = [\vdash]$.
Step 3. Sort each row of M in decreasing order. $p_i(1) = \max_j\{p_i(j)\}$.
Step 4. Search minimum of maximum of each row: $\alpha\ \min_i(p_i(1))$.
Step 5. While $\alpha > 0$ do
$e \leftarrow [e, \alpha]$,
Remove α from maximum of each row: $p_i(1)\ p_i(1) - \alpha$, for all i
Sort each row of updated M in decreasing order.
$\alpha \leftarrow \min_i\{P_i(1)\}$.
Step 6. End while
Step 7. Output e.

For the genotype data, we suggest to use Sharron entropy $H_1(E)$ and $H_1(\tilde{E})$ as a DM in the ANM and compare $H_1(E)$ with $H_1(\tilde{E})$. If $H_1(E) < H_1(\tilde{E})$ then $X \to Y$; if $H_1(\tilde{E}) < H_1(E)$ then $Y \to X$.

DISTANCE CORRELATION FOR CAUSAL INFERENCE WITH DISCRETE VARIABLES

In previous sections, we introduce the basis principal for assessing causation $X \to Y$ that the distribution $P(X)$ of causal X is independent of the causal mechanism or conditional distribution $P(Y|X)$ of the effect Y, given causal X. Now the question is how to assess their independence. Recently, distance correlation is proposed to measure dependence between random vectors which allows for both linear and nonlinear dependence (Sze'kely et al., 2007; Sze'kely and Rizzo, 2009). Distance correlation extends the traditional Pearson correlation in two remarkable directions:

(1) Distance correlation extends the Pearson correlation defined between two random variables to the correlation between two sets of variables with arbitrary numbers;
(2) Zero of distance correlation indicates independence of two random vectors.

Discretizing distributions $P(X)$ and $P(Y|X)$, and viewing their discretized distributions as two vectors $P(X)$ and $P(Y|X)$, the distance correlation between $P(X)$ and $P(Y|X)$ can be used to assess causation between X and Y.

Consider two vectors of random variables: p- dimensional vector X and q- dimensional vector Y. Let $P(x)$ and $P(y)$ be density functions of the vectors X and Y, respectively. Let $P(x, y)$ be the joint density function of X and Y. There are two ways to define independence between two vectors of variables: (1) density definition and (2) characteristic function definition. In other words, if X and Y are independent then either

(1) $P(x, y) = P(x)P(y)$ or
(2) $f_{X,Y}(t, s) = f_X(t)f_Y(s)$,

where $f_{X,Y}(t, s) = E[e^{i(t^T x + s^T y)}]$, $f_X(t) = E[e^{it^T x}]$ and $f_Y(s) = E[e^{is^T y}]$ are the characteristic functions of (X, Y), X and Y, respectively. Therefore, we can use both distances $||P(x, y) - P(x)P(y)||$ and $||f_{X,Y}(t, s) - f_X(t)f_Y(s)||$ to measure dependence between two vectors X and Y. Distance correlation (Sze'kely et al., 2007) uses distance between characteristic function to define the dependence measure.

Assume that pairs of $(X_k, Y_K), k = 1, ..., n$ are sampled. Calculate the Euclidean distances:

$$a_{kl} = ||X_K - X_l||_p, b_{kl} = ||Y_k - Y_l||_q, k = 1, ..., n, l = 1, ..., n.$$

Define

$$\bar{a}_{k.} = \frac{1}{n}\sum_{l=1}^{n} a_{kl},\ \bar{a}_{.l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl},$$
$$\bar{a}_{..} = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} a_{kl},$$

$$\bar{b}_{k.} = \frac{1}{n}\sum_{l=1}^{n} b_{kl},\ \bar{b}_{.l} = \frac{1}{n}\sum_{k=1}^{n} b_{kl} \text{ and}$$
$$\bar{b}_{..} = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} b_{kl}.$$

Define two matrices:

$$A = (A_{kl})_{n\times n} \text{ and } B = (B_{kl})_{n\times n},$$

where

$$A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..},$$
$$B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..},\ k,l = 1,...,n.$$

Finally, the sampling distance covariance$V_n(X,Y)$, variance $V_n(X)$ and correlation $R_n(X,Y)$ are defined as

$$V_n^2(X,Y) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} A_{kl}B_{kl}, \tag{12}$$

$$V_n^2(X) = V_n^2(X,X) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} A_{kl}^2,$$
$$V_n^2(Y) = \sum_{k=1}^{n}\sum_{l=1}^{n} B_{kl}^2,$$

$$R_n^2(X,Y) = \begin{cases} \frac{V_n^2(X,Y)}{\sqrt{V_n^2(X)V_n^2(Y)}}, & V_n^2(X)V_n^2(Y) > 0 \\ 0 & V_n^2(X)V_n^2(Y) = 0, \end{cases} \tag{13}$$

respectively.

Independence can be formally tested by statistics based on distance correlation. The null hypothesis is defined as $H_0 : X$ and Y are independent.

We can use Equation (12) to define a test statistic:

$$T_{IND} = n\frac{V_n^2(X,Y)}{\bar{a}_{..}\bar{b}_{..}}. \tag{14}$$

Since distribution of test statistics is difficult to compute, we often use permutations to calculate P-values. Specifically, we can permutate X and Ymillions of times. For each permutation, we compute test statistic T_{IND}. Therefore, via permutations we can calculate the empirical distribution of T_{IND}. Using an empirical distribution, we can calculate the P-value as

$$P - value = P(T_{IND} > T_{IND0}),$$

where T_{IND0} is the observed value of T_{IND} in real data.

Distance correlation can be used to test independence between causal and causal generating mechanisms (Liu and Chan, 2016). Consider p- dimensional random vector$X$ and q- dimensional random vector$Y$. Let $P(X,Y)$ be their joint distribution. Let $P(X)$ and $P(Y|X)$ be the density function of X and conditional density function of Y, given$X$, respectively. Similarly, we can define $P(Y)$ and $P(X|Y)$. Unlike association analysis where dependence

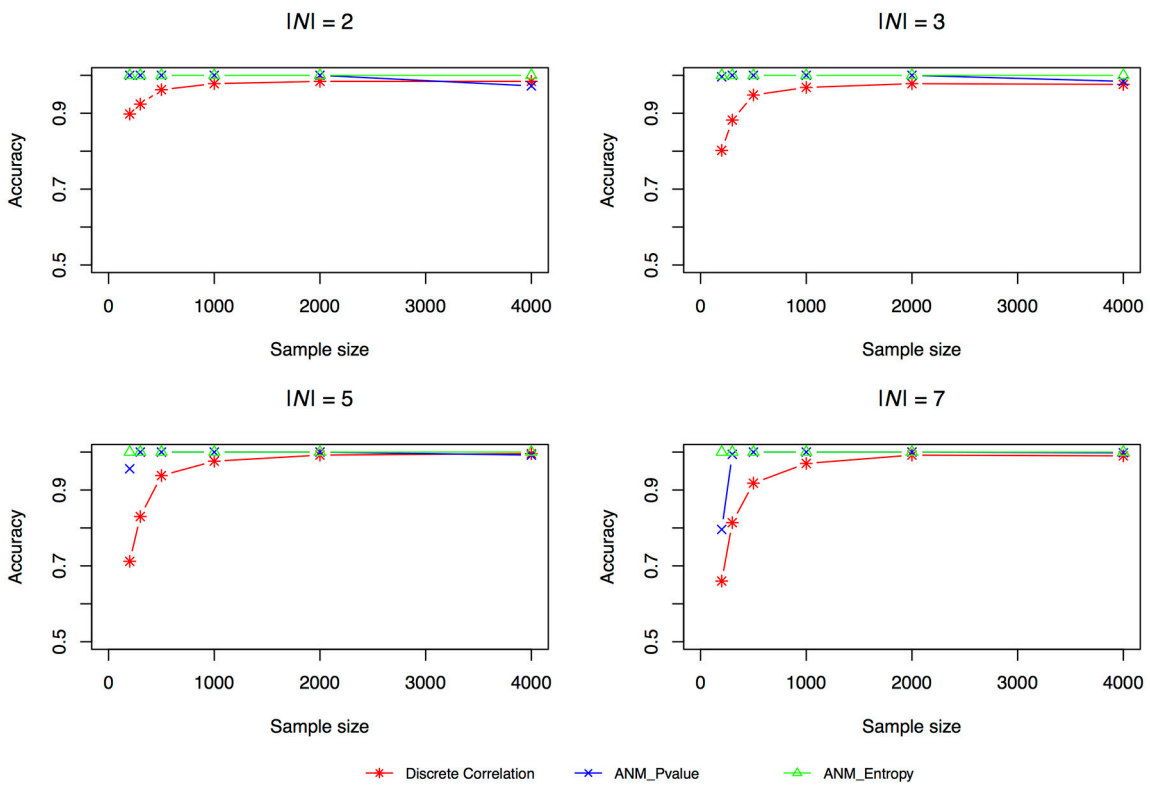

FIGURE 3 | Simulation scenario 1, |N| = 2. Simulation scenario 2, |N| = 3. Simulation scenario 3, |N| = 5. Simulation scenario 4, |N| = 7.

is measured between two random vectors, in causal analysis, dependence is measured between two distributions.

The distance correlation dependence measures between two distributions are defined as

$$\Delta_{X\to Y} = R(P(X), P(Y|X)), \tag{15}$$

$$\Delta_{Y\to X} = R(P(Y), P(X|Y)), \tag{16}$$

where $R(.,.)$ is a distance correlation measure between two vectors.

Suppose that X and Y are discretized (or divided) into m and k groups, respectively. Let m_i be the number of points X in the *ith* group and k_{ij} be the number of points (X, Y) where X is in the *ith* group and Y is in the *jth* group. Then, $n = \sum_{i=1}^{m} m_i$ and $m_i = \sum_{j=1}^{k} k_{ij}$. Let $X^{(i)}$ be the collection of all points X in the *ith* group and $Y^{(j)}$ be the collection of all points Y in the *jth* group. Then, the estimated density function $P(X^{(i)})$ is $P(X^{(i)}) = \frac{m_i}{n}$ and the conditional density function $P(Y^{(j)}|X^{(i)}) = \frac{k_{ij}}{m_i}$.
Let $S_{X\to Y} = a_{..}b_{..}$. Distance covariance is defined as

$$V_m^2(P(X), P(Y|X)) = \frac{1}{m^2}\sum_{i=}^{m}\sum_{j=1}^{m} A_{ij}B_{ij}. \tag{17}$$

Similarly, $V_m^2(P(Y), P(X|Y))$ and $S_{Y\to X}$ can be similarly defined. Define

$$\Delta_{X\to Y} = \frac{mV_m^2(P(X), P(Y|X))}{S_{X\to Y}}, \tag{18}$$

$$\Delta_{Y\to X} = \frac{mV_m^2(P(Y), P(X|Y))}{S_{Y\to X}}. \tag{19}$$

The null hypothesis for testing is

H_0 : no causation between two vectors X and Y.

The statistic for testing the causation between two vectors X and Y is defined as

$$T_C = |\Delta_{X\to Y} - \Delta_{Y\to X}|. \tag{20}$$

When T_C is large, either $\Delta_{X\to Y} > \Delta_{Y\to X}$ which implies Ycauses X, or $\Delta_{Y\to X} > \Delta_{X\to Y}$ which implies that X causes Y. When $T_C \approx 0$, this indicates that no causal decision can be made.

SIMULATIONS

We use simulation experiments that were presented in Liu and Chan (2016) to compare the performance of three methods: ANM, Distance correlation and entropy for causal inference with discrete variables. Consider two sets of data (1) dataset 1 and dataset 2 generated in section (Additive Noise Models) and section (Models with Randomly Generated P(X) and P(Y|X)) of the paper written by (Liu and Chan, 2016), respectively.

The accuracies of three methods for causation discovery in dataset 1 were shown in **Figure 3**. A total of 200, 300, 500, 1,000, 2,000, and 4,000 points for each model were sampled. **Figure 3** showed that the Entropy-based ANMs where the independence between cause X and residuals E are tested by entropy had the highest accuracies to infer cause-effect direction, the distance-correlation-based methods had the lowest accuracies, and the

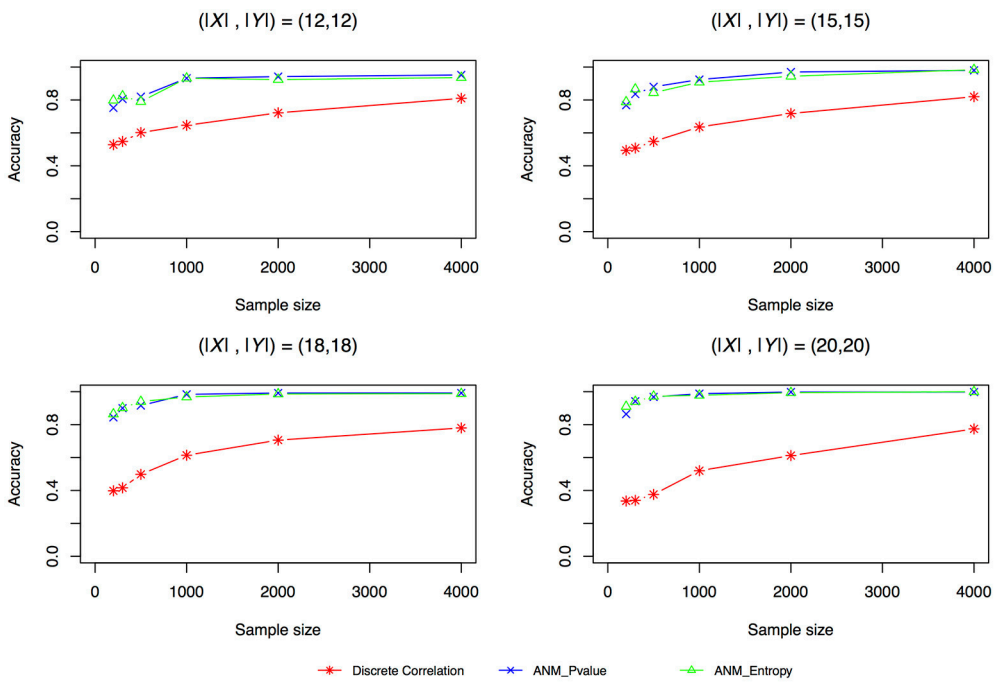

FIGURE 4 | Simulation scenario 1, (|X|,|Y|)=(12,12). Simulation scenario 2, (|X|,|Y|)=(15,15). Simulation scenario 3, (|X|,|Y|)=(18,18). Simulation scenario 4, (|X|,|Y|) = (20,20).

classical ANMs with discrete variables had the similar accuracies as that the entropy-based ANMs had. We observed that as sample sizes increased the accuracies increased and when the sample sizes are larger than 2,000 all three methods can accurately infer cause-effect directions.

Figure 4 plotted accuracy of three methods for inferring cause-effect direction as a function of sample sizes. Unlike the results in Liu and Chan (2016) where they showed that distance correlation had much higher accuracy to infer causal direction than the ANMs, we observed in **Figure 4** that both entropy-based ANMs and classical ANMs had similar accuracies and had higher accuracies than the distance correlation method. We also observed that even when sample sizes reached 4,000 the distance correlation method still could not reach accuracy beyond 80%.

To further evaluate their performance, we calculated the type I error rates three statistics for testing causation using simulations. We randomly selected 10 SNPs across the genome from 1,000 Genome Project data. A total of 1,000 simulations were conducted. We consider two scenarios: (1) no association and no causation and (2) presence of association, but no causation. **Tables 2**, **3** presented average type 1 error rates of three tests over 10 SNPs. **Tables 2**, **3** showed that type 1 error rates of the ANM based on permutation and DC method even in the presence of association were not significantly deviated from nominal levels, but the type 1 error rates of ANM based on entropy under association were significantly deviated from the Nominal levels. These results showed that the ANM with permutation tests and DC methods can be used for testing causation between SNP and disease, but the ANM based on entropy cannot be applied to test causation between SNP and disease.

To give some recommendations on when and which methods should be used, we conducted power simulations using the data for type 1 error calculation. We assume that both association and causation exist as described in Supplementary Note C. The results were summarized in **Table 4**. **Table 4** showed that in all scenarios, the ANM-based on permutation had the highest power among three statistical methods for testing causation.

TABLE 2 | Average type 1 error rates of three statistics for testing causation, assuming no association.

	Type I error rate			
Methods	α	$N = 500$	$N = 1{,}000$	$N = 2{,}000$
ANM Permutation	0.05	0.0431	0.0504	0.0505
	0.01	0.0090	0.0073	0.0081
ANM Entropy	0.05	0.0390	0.0392	0.0387
	0.01	0.0093	0.0080	0.0078
DC	0.05	0.0470	0.0461	0.0457
	0.01	0.0074	0.0091	0.0097

TABLE 3 | Average type 1 error rates of three statistics for testing causation, assuming presence of association.

	Type I error rate			
Methods	α	$N = 500$	$N = 1{,}000$	$N = 2{,}000$
ANM Permutation	0.05	0.0341	0.0355	0.0337
	0.01	0.0094	0.0117	0.0103
ANM Entropy	0.05	0.1962	0.2014	0.2169
	0.01	0.1666	0.1679	0.1705
DC	0.05	0.0507	0.0511	0.0508
	0.01	0.0103	0.0093	0.0099

TABLE 4 | Power of three statistical methods for testing causation in the presence of both association and causation.

		Number of samples				
		200	**500**	**1,000**	**2,000**	**5,000**
ANM *P*-value	0.05	0.5731	0.6983	0.7642	0.8127	0.8466
	0.01	0.5193	0.6830	0.7654	0.8239	0.8783
ANM Entropy	0.05	0.1290	0.1496	0.1611	0.1686	0.1747
	0.01	0.1120	0.1308	0.1470	0.1555	0.1771
DC	0.05	0.2549	0.3104	0.3458	0.3834	0.4175
	0.01	0.0954	0.1214	0.1476	0.1840	0.2400

REAL DATA ANALYSIS

Illustrate the application of causal inference to genetic analysis of complex diseases, three methods for causal inference were used to infer causal relationships between four SNPs in two genes with Alzheimer's disease (AD). Two SNPs in genomic positions 15528889 and 15530350 in gene *TRIM16,* two SNPs 15524749 and 15519576 in gene *CDRT1* and other two SNPs were types in 1,707 individuals (514 AD and 1,193 controls). The results were summarized in **Table 5**. Due to limitation of computer capability, only 1,000,000 permutations were carried out to compute *P*-values of classical ANM test statistics and entropy-based ANM test statistics. The threshold for declaring significance of association test was 1.14×10^{-8}. Since distance correlation test requires that the potential cause should take more than three values of states, in general, it cannot be used for causal genetic analysis. **Table 5** showed that *P*-values of the classical ANM and entropy-based ANM were close, and that four SNPs in two genes which showed strong association also demonstrated causation. The literature reported that gene *TRIM16* inhibited neuroblastoma cell proliferation through cell cycle regulation and dynamic nuclear localization and gene *CDRT1* was involved in frontotemporal dementia (Aronsson et al., 1998; Bell et al., 2013).

FUTURE PERSPECTIVE

Association analysis has been used as a major tool for dissecting genetic architecture and unraveling mechanisms of complex

TABLE 5 | *P*-values for testing the causation and association of four SNPs with AD.

Chr	Gene	Genomic position	*P*-values			
			Association	ANM	ANM-Entropy	Distance correlation
17	TRIM16	15528889	4.51E-09	<1.00E-06	<1.00E-06	0.5
17	TRIM16	15530350	1.01E-07	<1.00E-06	<1.00E-06	0.5
17	CDRT1	15524749	5.12E-08	<1.00E-06	<1.00E-06	0.5
17	CDRT1	15519576	5.49E-09	<1.00E-06	<1.00E-06	0.5
1	PYHIN1	158947655	0.00131	0.00011	0.00031	0.15831
4	AFAP1	7813044	0.00222	0.00083	0.00074	0.58278

diseases for more than a century (Fisher, 1918; Timpson et al., 2017). Although significant progress in dissecting the genetic architecture of complex diseases by genome-wide association studies (GWAS) has been made, the overall contribution of the new identified genetic variants to the diseases is small and a large fraction of disease risk genetic variants is still hidden. Understanding the etiology and causal chain of mechanism underlying many complex diseases remains elusive. The current approach to uncovering hidden genetic variants is (1) to increase sample sizes, (2) to study association of rare variants by next-generation sequencing and (3) to perform multi-omic analysis. Association and correlation analysis are the current paradigm of analysis for all these approaches. Our experiences in association analysis strongly demonstrate that association analysis lacks power to discover the mechanisms of the diseases for the two major reasons. The first reason is that association analysis cannot identify causal signals that are quite different from the association signals. The second reason is that the widespread networks that are constructed in integrated omic analysis are undirected graphs. Using undirected graphs, we are unable to infer direct cause-effect relations and hence cannot discover chain of causal mechanism from genetic variation to diseases via gene expressions, epigenetic variation, protein expressions, metabolism variation, and phenotype variations. The use of association analysis as a major analytical platform for genetic studies of complex diseases is a key issue that hampers the theoretical development of genomic science and its application in practice. Causal inference coupled with multiple omics, imaging, physiological and phenotypic data is an essential component for the discovery of disease mechanisms. It is time to develop a new generation of genetic analysis for shifting the current paradigm of genetic analysis from shallow association analysis to deep causal inference. This review paper introduced major statistical methods for inferring causal relationships between discrete variables and explored the potential roles causal inference may play in genetic analysis of complex diseases. Our purpose is to stimulate discussion about what research direction in genetic studies: causal analysis or association analysis should be taken in the future.

AUTHOR CONTRIBUTIONS

PH perform data analysis, write paper. RJ perform data analysis; LJ design project. MX design project and write paper.

FUNDING

PH and LJ were supported by National Natural Science Foundation of China (31521003), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), and the 111 Project (B13016) from Ministry of Education.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00238/full#supplementary-material

REFERENCES

Altman, N., and Krzywinski, M. (2015). Association, correlation and causation. *Nat. Methods* 12, 899–900. doi: 10.1038/nmeth.3587

Aronsson, F. C., Magnusson, P., Andersson, B., Karsten, S. L., Shibasaki, Y., Lendon, C. L., et al. (1998) The NIK protein kinase and C17orf1 genes: chromosomal mapping, gene structures and mutational screening in frontotemporal dementia and parkinsonism linked to chromosome 17. *Hum. Genet.* 103, 340–345.

Bell, J. L., Malyukova, A., Kavallaris, M., Marshall, G. M., and Cheung, B. B. (2013) TRIM16 inhibits neuroblastoma cell proliferation through cell cycle regulation and dynamic nuclear localization. *Cell Cycle* 12, 889–898. doi: 10.4161/cc.23825

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/j.cell.2017.05.038

Brookes, A. J., and Robinson, P. N. (2015). Human genotype-phenotype databases: aims, challenges and opportunities. *Nat. Rev. Genet*. 16, 702–715. doi: 10.1038/nrg3932

Callaway, E. (2017). Genome studies attract criticism: geneticists question ability of genome-wide association studies to find useful disease links. *Nature* 546:463. doi: 10.1038/nature.2017.22152

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Proc. Roy. Soc. Edinburgh.* 52, 99–433.

Gottlieb, S. S. (2017). Theory and fact: revisiting association and causation. *JACC Heart Fail.* 5, 327–328. doi: 10.1016/j.jchf.2017.03.005

Janzing, D., Chaves, R., and Schölkopf, B. (2016). Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New J. Phys.* 18:093052. doi: 10.1088/1367-2630/18/9/093052

Janzing, D., and Schölkopf, B. (2010). Causal inference using the algorithmic markov condition. *IEEE Trans. Inf. Theory* 56, 5168–5194. doi: 10.1109/TIT.2010.2060095

Janzing, D., and Steudel, B. (2010). Justifying additive-noise-model based causal discovery via algorithmic information theory. *Open Syst. Inf. Dyn.* 17, 189–212. doi: 10.1142/S1230161210000126

Kano, Y., and Shimizu, S. (2003). "Causal inference using nonnormality," in *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion* (Tokyo), 261–270.

Kocaoglu, M., Dimakis, A. G., Vishwanath, S., and Hassibi, B. (2016). Entropic causal inference. *arXiv [Preprint] arXiv* 1611.04035.

Liu, F., and Chan, L. (2016). Causal inference on discrete data via estimating distance correlation. *Neural Comput.* 28, 801–814. doi: 10.1162/NECO_a_00820

Peters, J., Janzing, D., and Schölkopf, B. (2011). Causal inference on discrete data using additive noise models. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 2436–2450. doi: 10.1109/TPAMI.2011.71

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference - Foundations and Learning Algorithms Adaptive Computation and Machine Learning Series*. Cambridge, MA: The MIT Press.

Shimizu, S., Hoyer, P. O., Hyv̈arinen, A., and Kerminen, A. J. (2006). A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* 7, 2003–2030.

Sze'kely, G. J., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat.* 35, 2769–2794. doi: 10.1214/009053607000000505

Sze'kely, G. J., and Rizzo, M. L. (2009) Brownian distance covariance. *Ann. Appl. Stat.* 3, 1236–1265. doi: 10.1214/09-AOAS312

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., and Richards, J. B. (2017). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* 19, 110–124. doi: 10.1038/nrg.2017.101

Valente, B. D., Morota, G., Peñagaricano, F., Gianola, D., Weigel, K., and Rosa, G. J. (2015). The causal meaning of genomic predictors and how it affects construction and comparison of genome-enabled selection models. *Genetics* 200, 483–494. doi: 10.1534/genetics.114.169490

Wakeford, R. (2015). Association and causation in epidemiology - half a century since the publication of Bradford Hill's interpretational guidance. *J. R. Soc. Med.* 108, 4–6. doi: 10.1177/0141076814562713

Xiong, M. M. (2018). *Big Data in Omics and Imaging: Integrated Analysis and Causal Inference*. New York, NY: CRC Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.