An information-gain approach to detecting three-way epistatic interactions in genetic association studies

Ting Hu,[1,2] Yuanzhu Chen,[1,3] Jeff W Kiralis,[1] Ryan L Collins,[1] Christian Wejse,[4] Giorgio Sirugo,[5] Scott M Williams,[2] Jason H Moore[1,2]

[1]Computational Genetics Laboratory, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA
[2]Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, New Hampshire, USA
[3]Department of Computer Science, Memorial University, St. John's, Newfoundland, Canada
[4]Center for Global Health, School of Public Health, Aarhus University, Skejby, Denmark
[5]Centro di Genetica, Centro di Ricerca Scientifica, Ospedale San Pietro FBF, Rome, Italy

Correspondence to
Jason H Moore, HB 7937, One Medical Center Drive, Lebanon, NH 03756, USA;
Jason.h.moore@dartmouth.edu

Received 26 November 2012
Revised 26 November 2012
Accepted 5 January 2013
Published Online First
8 February 2013

ABSTRACT

Background Epistasis has been historically used to describe the phenomenon that the effect of a given gene on a phenotype can be dependent on one or more other genes, and is an essential element for understanding the association between genetic and phenotypic variations. Quantifying epistasis of orders higher than two is very challenging due to both the computational complexity of enumerating all possible combinations in genome-wide data and the lack of efficient and effective methodologies.

Objectives In this study, we propose a fast, non-parametric, and model-free measure for three-way epistasis.

Methods Such a measure is based on information gain, and is able to separate all lower order effects from pure three-way epistasis.

Results Our method was verified on synthetic data and applied to real data from a candidate-gene study of tuberculosis in a West African population. In the tuberculosis data, we found a statistically significant pure three-way epistatic interaction effect that was stronger than any lower-order associations.

Conclusion Our study provides a methodological basis for detecting and characterizing high-order gene-gene interactions in genetic association studies.

INTRODUCTION

Understanding the mapping from genetic variation to phenotypic variation has a great potential helping us understand, predict, diagnose, and treat common human diseases. However, existing main-effect-centered methodologies and techniques that depend on fundamental assumptions about a simple genetic architecture can only find very limited individual associations with disease risks. Genome-wide association studies[1–3] and next generation sequencing[4] make millions of single nucleotide polymorphisms (SNPs) in the human genome available for testing associations with phenotypic traits. These developments call for new methodologies that embrace the complex genetic architecture of diseases.[5 6] The non-additive effects of gene–gene interactions, that is, epistasis, are believed to be an important contributor to the complex relationship between genetic and phenotypic variations.[7–11] The focus of recent disease association research is shifting from identifying single locus susceptibility to quantifying interaction effects between multiple candidate loci throughout the human genome.[8 9 12]

Detecting and quantifying epistasis is a very challenging task. First, the epistatic interactions could involve multiple genes from a pair to a large set, and this undetermined order of interactions imposes enormous computational complexity of enumerating all possible combinations of genetic attributes for varying orders.[6 13] Second, as the order of interacting genetic attributes goes beyond two, it becomes mathematically difficult to separate the additive lower-order effects and the pure higher-order synergy, that is, the extreme case in which the association can only be observed when all attributes are considered together. Those are also the major reasons why most existing epistasis studies are limited to pairwise interactions on genetics data of moderate sizes.

Information-theoretic measures have emerged as a very useful tool to quantify synergistic interactions among multiple genetic attributes.[14–21] These measures are based on the Shannon entropy, which quantifies the amount of information, or uncertainty, of a random variable.[22] By considering genetic attributes and phenotypic traits as random variables, entropy-based information-theoretic measures can be used to quantify the shared information between one gene and a trait, i.e., the main effect, as well as the gained extra information about a trait obtained from combining multiple genes, i.e., the synergistic effect or epistasis. However, as discussed previously, due to the mathematical complexity, the application of information-theoretic measures in disease association studies is mainly limited to pairwise epistasis between two genetic attributes.

In this study, we propose a new measure to quantify the synergistic effects among three genetic attributes that contribute to disease susceptibility by extending the information-theoretic measure for pairwise synergy. In particular, we first measure the total amount of information that three attributes together can provide about the phenotypic status, and then subtract all lower-order effects including the main effects of the three attributes and all pairwise synergies between them. This yields a very strict measure of pure three-way epistasis. There have been previous attempts at extending information-theoretic measures on three-way and higher-order synergies.[14 17 23 24] However, most existing measures are not able to decouple lower-order interaction effects from the higher-order effects and the formalization of higher-order synergy is still debatable. We compared our new measure to those existing ones and were able to show that our measure performed the best at

To cite: Hu T, Chen Y, Kiralis JW, *et al*. *J Am Med Inform Assoc* 2013;**20**: 630–636.

separating all lower-order effects from the pure three-way epistasis by applying to both synthetic data containing artifact epistasis models and real human disease data from a candidate-gene association study of tuberculosis in a population from West Africa.[25] Of particular interest, we identified a statistically significant three-way epistasis model in the tuberculosis data, in which the three-way synergy is stronger than all the main effects and pairwise-interaction effects combined. With further biological verification and interpretation, this model could be very valuable in advancing tuberculosis research.

METHODS

Information-theoretic measures

In information theory,[22] entropy is a measure of the uncertainty of a random variable. It can be explained as the amount of information required on average to describe a random variable. For a discrete variable X with alphabet $\mathcal{X}$ and probability mass function $p(x)$, its entropy $H(X)$ is defined as

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x).$$

When there are more than one random variable, the definition of entropy can be extended as follows. The joint entropy of two discrete random variables X and Y with a joint distribution $p(x,y)$ is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y),$$

and the conditional entropy of X given the knowledge of Y can be obtained by the chain rule as

$$H(X|Y) = H(X,Y) - H(Y).$$

The dependency between two random variables can be described using mutual information.[22] This is a measure of the amount of information that one random variable contains about the other, or can be thought of as the reduction of uncertainty of one random variable given the knowledge of the other. In the context of genetic association studies, mutual information can be very useful to quantify how much of a phenotypic status is explained by genotypic variations. We consider a genetic attribute G_1 and the phenotypic class C, for example, case or control, are both discrete random variables. The mutual information $I(G_1;C)$ measures the reduction in the uncertainty of the class C due to the knowledge about the genotype of G_1 (figure 1A), defined as

$$I(G_1;C) = H(C) - H(C|G_1).$$

Intuitively, $I(G_1;C)$ can be used as a measure of the main effect of the genetic attribute G_1 on the class C.

Mutual information can also be extended to measure the epistatic interaction effect between two attributes. Given two genetic attributes G_1 and G_2, the mutual information

$$I(G_1,G_2;C) = H(C) - H(C|G_1,G_2)$$

measures how much of the phenotypic class joining G_1 and G_2 together can explain (figure 1B). By subtracting the individual main effects of G_1 and G_2 from their joint effect $I(G_1,G_2;C)$, that is,

$$IG(G_1;G_2;C) = I(G_1,G_2;C) - I(G_1;C) - I(G_2;C),$$

the information gain $IG(G_1;G_2;C)$ is the gain of mutual information of knowing both G_1 and G_2 with respect to the class C. A positive value of $IG(G_1;G_2;C)$ indicates the synergy between G_1 and G_2, while a negative value indicates the redundancy or correlation between them. The synergy can be well explained using the epistatic interactions between two genetic attributes. As discussed previously, this pairwise information-gain measure has been successfully applied in many epistasis studies thanks to its model-free, non-parametric, and fast implementation.

Further extension of the information-gain measure on more than two genetic attributes is non-trivial for epistasis studies because many complex human diseases could very likely involve genetic interactions of orders higher than two way. There is no widely accepted formal definition of information gain including genetic attributes higher than two. Here, we make an effort measuring three-way synergistic interactions using information gain.

In a previous attempt, Anastassiou[14] and Varadan *et al*[24] proposed to define the three-way information gain by comparing the integrated joint mutual information to the best-achieved subsets mutual information after breaking the whole into partitions, mathematically written as

$$IG_{partition}(G_1;G_2;G_3;C) = I(G_1,G_2,G_3;C) - \max\begin{cases} I(G_1,G_2;C)+I(G_3;C) \\ I(G_1,G_3;C)+I(G_2;C) \\ I(G_2,G_3;C)+I(G_1;C) \\ I(G_1;C)+I(G_2;C)+I(G_3;C). \end{cases}$$

The partition of the set $\{G_1,G_2,G_3\}$ chosen in this formula is the one that maximizes the sum of the amounts of mutual information connecting the subsets with the phenotypic class. This is referred to as 'maximum-information partition' of the set $\{G_1,G_2,G_3\}$ with respect to the class C. This three-way $IG_{partition}(G_1;G_2;G_3;C)$ quantifies the information that can be gained by combining G_1,G_2 and G_3 together comparing to its maximum-information partition. Although technically sound, this formula might include false-positive errors of pure three-way epistasis. For instance, assuming $I(G_1,G_2;C)$ and $I(G_3;C)$ is the maximum-information partition, after combining G_3 with $\{G_1,G_2\}$, the gained information could be the result of either the pure three-way epistasis, or the pairwise epistasis between G_3 and one (or both) of $\{G_1,G_2\}$, or the mixture of all above.

A more strict alternative measure[23] was proposed as follows

$$\begin{aligned} IG_{alternative}(G_1;G_2;G_3;C) = {} & I(G_1,G_2,G_3;C) \\ & -IG(G_1;G_2;C) - IG(G_1;G_3;C) \\ & -IG(G_2;G_3;C) - I(G_1;C) - I(G_2;C) \\ & -I(G_3;C), \end{aligned}$$

where all the lower-order effects are subtracted. However, as reviewed and pointed out in Anastassiou,[14] this formula fails in the extreme redundancy case where all G_1, G_2, and G_3 provide the same full amount of information on C, that is, $G_1 = G_2 = G_3 = C$. In this case, $I(G_i;C) = I(G_i,G_j;C) = I(G_i,G_j,G_k;C) = H(C)$, where i,j,k are different values taken from $\{1,2,3\}$. Therefore $IG_{alternative}(G_1;G_2;G_3;C) = H(C)$, which indicates the contradictory extreme synergy.

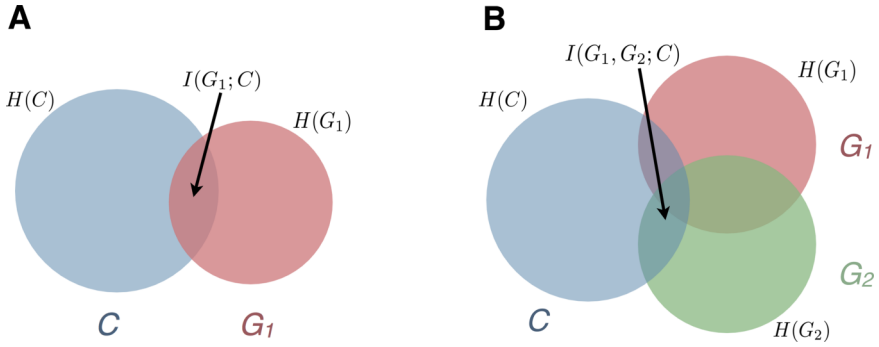

Figure 1 Venn diagrams showing the entropy and mutual information of genetic attributes G and the phenotypic class C. (A) For one attribute G_1 and the phenotypic class C, their entropies $H(G_1)$ and $H(C)$ are indicated as the two colored sets. The mutual information $I(G_1;C)$ is defined as the intersection of the two sets, and the joint entropy $H(G_1,C)$ is the reunion of the two sets. (B) For two genetic attributes G_1, G_2 and the phenotypic class C, the mutual information $I(G_1,G_2;C)$ is the intersection of entropy $H(C)$ and joint entropy $H(G_1,G_2)$.

In our study, we propose a new strict measure by modifying $IG_{alternative}$ as

$$IG_{strict}(G_1;G_2;G_3;C) = I(G_1,G_2,G_3;C) - \max\begin{Bmatrix} IG(G_1;G_2;C) \\ 0 \end{Bmatrix} - \max\begin{Bmatrix} IG(G_1;G_3;C) \\ 0 \end{Bmatrix} - \max\begin{Bmatrix} IG(G_2;G_3;C) \\ 0 \end{Bmatrix} - I(G_1;C) - I(G_2;C) - I(G_3;C).$$

We only subtract pairwise synergies, that is, positive information gain, because the failure of $IG_{alternative}$ is due to the fact that it adds back information by subtracting negative information gain. By subtracting all lower-order effects and synergies, IG_{strict} measures the pure three-way synergy that is observable only by considering three attributes together.

Also note that, when applying to genetics data, the above mutual-information and information-gain measures can be normalized by dividing the class entropy $H(C)$. The normalized measures range from $[-1,1]$, and provide the percentage of explaining the phenotypic class C by giving the knowledge of one or multiple genetic attributes.

Datasets

We applied both three-way synergy measures $IG_{partition}$ and IG_{strict} to synthetic datasets and a real dataset on pulmonary tuberculosis from a West African population. The synthetic datasets were generated using genetic architecture model emulator for testing and evaluating software (GAMETES),[26 27] a direct approach to simulating bi-allelic n-locus epistatic models. In GAMETES, an n-locus epistasis model is generated deterministically with specified genetic constraints such as heritability and minor allele frequencies, and then a population of samples can be generated for that model. Using GAMETES, we can have synthetic data of models with epistasis at desired orders, which are ideal for testing and comparing the two three-way synergy measures. We first generated a pure three-way epistasis model $\{P_1,P_2,P_3\}$, where the association to phenotypic status was only observable when all three SNPs were considered together, that is, no main effects and no pairwise interactions. The total heritability of combining three SNPs was set to 0.27, and the minor allele frequencies of all three SNPs were set to 0.2. A corresponding dataset was generated by including 100 SNPs in total with 97 randomized SNPs, $\{N_1,N_2,N_3,\ldots,N_{97}\}$, to provide a null distribution. This synthetic dataset had 400 cases and 400 controls. Then we used GAMETES to generate a collective pairwise epistasis model $\{P_1',P_2',P_3'\}$, in which any two of the three attributes had strong pairwise interactions but there was no epistasis at the three-way level. Again there was no main effect for each attribute. The heritability of combining three SNPs was set

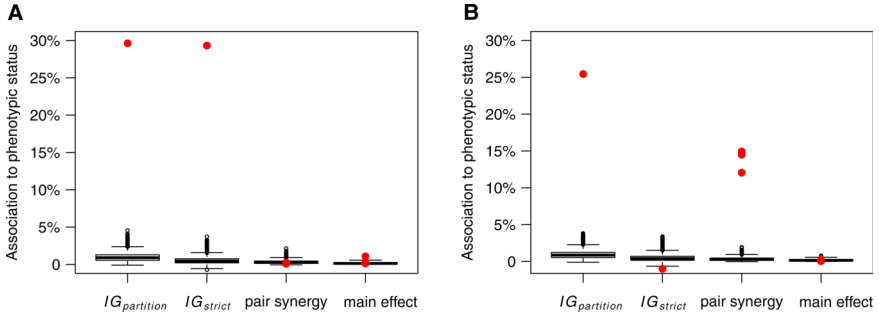

Figure 2 Information-theoretic measures (normalized information gain and mutual information representing the genetic associations to the case–control status) of the synthetic datasets generated by GAMETES. (A) The first synthetic dataset that has a pure three-way epistasis model $\{P_1,P_2,P_3\}$ with no main effects and no pairwise epistasis. (B) The second synthetic dataset with a collective pairwise epistasis model $\{P_1',P_2',P_3'\}$, in which there are no main effects and no three-way epistasis but all the pairs $\{P_1',P_2'\}$, $\{P_1',P_3'\}$, and $\{P_2',P_3'\}$ have interaction effects. Points in red represent the observed values of the model $\{P_1,P_2,P_3\}$ in (A) and $\{P_1',P_2',P_3'\}$ in (B). Box plots in black show the null distributions of randomized single nucleotide polymorphisms.

to 0.32, and the heritability of combining any two SNPs, that is, $\{P_1', P_2'\}, \{P_1', P_3'\}$, and $\{P_2', P_3'\}$, was about 0.1. The minor allele frequencies for all three SNPs were again set to 0.2. The second synthetic dataset also included 100 SNPs in total with 97 randomized SNPs and had 400 cases and 400 controls.

The pulmonary tuberculosis data are from a case–control study[25] conducted at the Bandim Health Project in Bissau, the capital city of Guinea-Bissau. This area has a high prevalence of pulmonary tuberculosis and tuberculosis symptoms.[28] Tuberculosis is one of the highest mortality diseases due to the infection of *Mycobacterium tuberculosis*. However, the majority of infected individuals keep the bacterium under control and never develop a clinical disease. Genetic variation among individuals is a promising direction to look into the factors that could influence the susceptibility to develop tuberculosis. Tuberculosis patients included in this study were residents or long-term guests of Bissau aged 15 years or older. From November 2003 to November 2005, 438 tuberculosis patients were screened at local health centers. A total of 344 subjects met the inclusion criteria and accepted participation, and DNA samples were successfully collected from 321 of them. Healthy controls were recruited from the study area with certain exclusion criteria, such as history of tuberculosis and household tuberculosis records within the past 2 years. Three hundred and forty-seven DNA samples were obtained from the healthy control group. All DNA samples were extracted using a standard salting-out procedure. DNA purities were estimated spectrophotometrically, and final concentrations were determined by PicoGreen. Samples (4 ng of DNA) were genotyped by TaqMan SNP assays (ABI; Applera International Inc, Foster City, California, USA) in 10 µl reaction volume, using the Rotor-Gene 3000 (Corbett Robotics Pty Ltd, Brisbane, Queensland, Australia) and the ABI 7500 real-time PCR systems. Fluorescence curves were analyzed with the Rotor-Gene Software V.6 and the 7500 Sequence Detection Software V.1.2.1 for allelic discrimination. The data include 19 SNPs from innate immunity genes, DC-SIGN (CD209), long pentraxin 3 (PTX3), toll-like receptors (TLRs), and vitamin D receptor (VDR), which are relevant to the defense against *M tuberculosis*. The missing genotypes (<5%) were imputed using a frequency-based method. The missing value of a sample was filled using the most common genotype of the corresponding SNP in the population.

RESULTS

Information-gain measurements on synthetic data

Figure 2A shows the results of applying information-theoretic measures to the first synthetic dataset that had a pure three-way epistasis model. The points in red represent the observed values of the model $\{P_1, P_2, P_3\}$, and the box plots in black show the measures for the randomized SNPs $\{N_1, N_2, N_3, \ldots, N_{97}\}$. Neither main effect nor pairwise synergy was found using the information-theoretic measures as the observed data points do not distinguish from the null distributions. In addition, both three-way synergy measurements $IG_{partition}$ and IG_{strict} successfully captured the pure three-way epistasis in the model.

The results of the second synthetic dataset with the collective pairwise epistasis model are shown in figure 2B. Again, the points in red represent the observed values of the model $\{P_1', P_2', P_3'\}$, and the box plots in black show the null distributions of 97 randomized SNPs. As we can see, all three pairwise synergies were detected. No three-way epistasis was detected by IG_{strict}, but $IG_{partition}$ reported a strong three-way epistasis among P_1', P_2', and P_3' by including some portions of their pairwise synergies.

Information-gain measurements on real data

We used information-theoretic measures to quantify the main effects, two-way and three-way synergies on all possible combinations of attributes in the tuberculosis data. In addition, to show the collective and neighborhood structures of strong synergistic pairs, we built a pairwise epistasis network[18] (figure 3, rendered by software Cytoscape[29]). The network was constructed through an incrementing process as follows. An edge and its two end vertices were added to the network only if their pairwise epistasis strength was greater than a given threshold. When we gradually decreased the threshold from its maximal observed value to its minimal value, the network started from a single edge with two vertices and eventually became a complete graph, that is, every single vertex is directly connected to every other vertex. We chose the highest pairwise synergy threshold, that is, 0.71%, when all 19 attributes were included in the

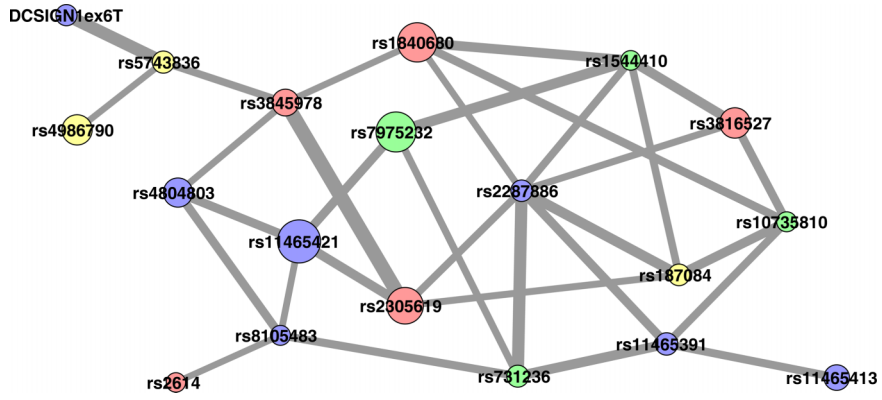

Figure 3 The pairwise statistical epistasis network of the tuberculosis data. The network has 32 edges (pairwise interactions) and 19 vertices (single nucleotide polymorphisms; SNPs). The size of a vertex indicates its main effect, and edge width indicates the pairwise synergy. The color of a vertex denotes the gene that a SNP belongs to, with blue representing CD209, green representing VDR, pink representing PTX3, and yellow representing TLRs. The network was built by incrementally adding pairwise interactions, ranked by their strength (pairwise information gain), and their two end SNPs. This network construction process completed when all 19 SNPs were included. It thus shows the strongest pairwise interactions and the neighborhood structures for all 19 SNPs.

network. Therefore, we can have a map of the strongest pairwise interactions showing the neighborhood structure of every attribute.

Figure 3 has 19 vertices and 32 edges that represent the top 32 strongest pairwise interactions out of all $\binom{19}{2}=171$ pairs. Using this network, we can easily identify three-way models that have strong collective pairwise synergies. In graph theory,[30] the distance $d(v_1, v_2)$ of a pair of vertices v_1 and v_2 is defined as the minimal number of edges for one vertex to reach the other. Given three vertices v_1, v_2, and v_3, we define their trio distance $d_{trio}(v_1, v_2, v_3)$ as the sum of all pairwise distances, that is, $d_{trio}(v_1, v_2, v_3) = d(v_1, v_2) + d(v_1, v_3) + d(v_2, v_3)$. Therefore, for trios with $d_{trio} = 3$, any two of them are directly joined by an edge, which indicates three strong collective pairwise interactions in a three-way model. If a trio has $d_{trio}=4$, one vertex is directly connected to the other two but the other two are not joined by an edge. A trio with $d_{trio}>4$ does not have strong collective pairwise epistasis.

Figure 4 shows the comparison of the results of both three-way synergy measures. In general, $IG_{partition}$ and IG_{strict} are positively correlated (Spearman's rank correlation $\rho=0.8448$, $p<2.2\times10^{-16}$). All the data points, that is, three-way models, are positioned on one side of the $x=y$ line, which indicates that the IG_{strict} measure is always less than or equal to the $IG_{partition}$ measure. This is intuitive because IG_{strict} subtracts more terms than $IG_{partition}$ does from $I(G_1, G_2, G_3; C)$. Colors indicate whether a trio has strong collective pairwise epistasis. As seen in the figure, the discrepancies (away from the $x=y$ line) between $IG_{partition}$ and IG_{strict} are the most distinguishing for red data points, that is, trios that have either two or three strong collective pairwise interactions. These results also verify our previous discussion on these two three-way synergy measures using synthetic data. That is, $IG_{partition}$ probably includes pairwise synergies into its three-way epistasis measure when there are more than one pairwise synergies in a three-way model, and IG_{strict} only captures the pure three-way epistasis that are beyond any lower-order synergies or effects.

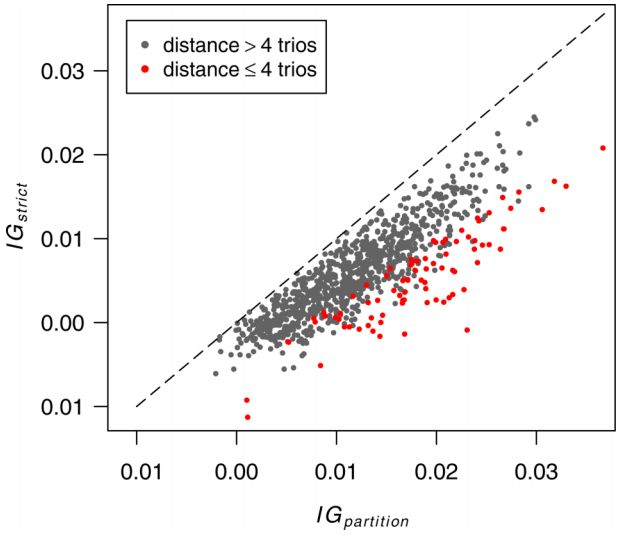

Figure 4 Comparison of two three-way synergy measurements in the tuberculosis data. Each data point represents a three-way model and its color indicates whether the three-way model has strong collective pairwise epistasis. If a trio's distance is less than or equal to 4 in the pairwise epistasis network (figure 3), this three-way model has strong collective pairwise interactions among its three attributes (in red). Otherwise a trio does not possess strong collective pairwise epistasis (in black). The dashed curve represents the $x=y$ line.

Table 1 Spearman's rank correlation of association synergies or effects at different model orders in the tuberculosis dataset

	Three-way $IG_{partition}$	Three-way IG_{strict}
Main effect	$\rho=0.0588$ $p=0.0015$	$\rho=0.0508$ $p=0.0062$
Pairwise synergy	$\rho=0.3278$ $p=<2.2\times10^{-16}$	$\rho=0.0565$ $p=0.0023$

We also looked into the correlations between the three-way synergies and lower-order synergies or effects (table 1). Both three-way synergy measures do not correlate with individual main effects. However, the three-way $IG_{partition}$ shows a correlation with pairwise synergy but IG_{strict} does not. This further confirms our previous discussion on the differences between these two three-way synergy measures.

The best three-way models using those two synergy measures are reported in figure 5. The figure shows all the individual main effects, pairwise synergies, three-way synergies, and the total mutual information for a three-way model. The best $IG_{partition}$ model (figure 5A) includes SNPs *rs11465421* from CD209, *rs1544410* from VDR, and *rs7975232* from VDR. It has the total mutual information 6.695% and three-way $IG_{partition}$ 3.668%. This model is clearly a mixed epistasis model by possessing both strong collective pairwise interactions and a pure three-way interaction. The best IG_{strict} model (figure 5B) involves SNPs from three different genes, *rs5743836* from TLR9, *rs2305619* from PTX3, and *rs4804803* from CD209. All three SNPs have very limited main effects and pairwise synergies. However, the strict three-way information gain is stronger than all other lower-order effects combined together, and contributes $\left(\frac{IG_{strict}}{I}\right)=54.23\%$ to the total association. An explicit test of epistasis[31] was used to assess the statistical significance of those observations. Instead of shuffling the case–control class in a standard permutation test, the explicit test randomly shuffled genotypes of samples within each class. Therefore, the genotype frequencies within each class remained fixed. This preserved the independent main effects while randomizing any non-linear interactions, and provided the null hypothesis that the only genetic effects in the data were linear and additive. We performed the explicit test 1000 times for each observation, and both best models were statistically significant ($p=0.001$ for the best $IG_{partition}$ and $p=0.008$ for the best IG_{strict}).

DISCUSSION

Epistasis is recognized as playing an important role in the genetic architecture of complex traits such as common human diseases. Quantifying interaction effects among multiple loci throughout the human genome has become the major focus of current research for understanding the complex relationship between genetic variations and phenotypic traits.[8 9 12] However, this task is challenging due to the fact that first enumerating all possible combinations of genetic variants in a dataset of moderate size is computationally infeasible, and second it is difficult to separate the additive effects and the synergistic effects among multiple genetic factors. The majority of epistasis studies focus on pairwise interactions and rarely look beyond interactions of orders higher than two. It is computationally intensive to enumerate three-way combinations in human genome data and,

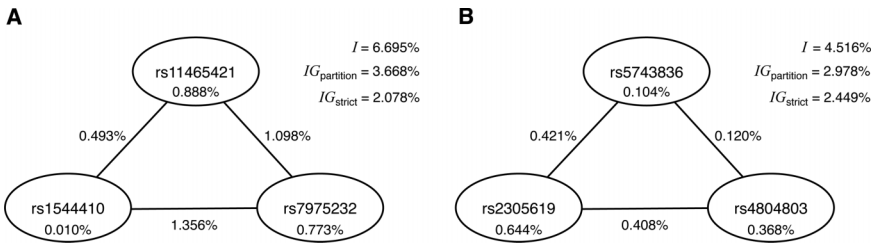

Figure 5 The models with the highest three-way information gain in the tuberculosis data. (A) The best model of $IG_{partition}$ with permutation testing significance $p=0.001$, and (B) the best model of IG_{strict} with permutation testing significance $p=0.008$. Each circle is a single nucleotide polymorphism (SNP) with its name and main effect strength. An edge represents a pairwise epistasis with its strength labeled. I is the total mutual information between combining three SNPs and the case–control status, and IG is the three-way information gain. All measures are normalized by dividing the entropy of the case–control status H(C), and thus give the percentage of predicting information on the phenotype status.

furthermore, not many good methods of quantifying three-way epistasis are proposed in the literature.[14]

In the present study, we proposed a fast, model-free, and non-parametric measure to detect and characterize three-way epistatic interactions. It is a natural extension of the pairwise synergy measure using information gain by measuring the total three-way mutual information and then subtracting three main effects and three pairwise synergies. Our approach was shown to be able to detect pure three-way epistasis, that is, no observable effect at either the two-way or one-way level, in both synthetic and real population-based genetics data. Our study is not the first attempt to quantify three-way epistasis using information-theoretic measures. We compared our measure IG_{strict} to a previously published one $IG_{partition}$.[32] Both measures exclude main effects. However, $IG_{partition}$ is biased towards three-way models that possess strong pairwise interactions. Our IG_{strict} excludes all one-way and two-way effects and is able to detect pure three-way synergy that is only observable when all three attributes are considered together.

Both three-way synergy measures were applied to a pulmonary tuberculosis dataset from a West African population. Several potentially relevant innate immunity genes, CD209, PTX3, TLRs, and VDR, were included in these data to investigate their role in pulmonary tuberculosis susceptibility. The dendritic cell-specific intercellular adhesion molecule-3-grabbing non-integrin (DC-SIGN or CD209) is a crucial *M tuberculosis* receptor expressed on the surface of dendritic cells, and is involved in the initiation of innate immune response through identification of potential infectious agents. CD209 has previously been found to be associated with tuberculosis.[33 34] Long pentraxin 3 (PTX3) is produced by innate immunity cells and vascular cells in response to proinflammatory signals and TLR engagement.[35] PTX3 levels have been shown to be correlated to the degree of infection in lungs, and PTX3 plasma levels can be monitored to measure treatment efficacy because PTX3 concentration decreases as an infection is mitigated.[36] TLRs are a family of receptors that are a key component in the innate immune system. TLRs recognize pathogenic molecules and control host immune response against them, and have been extensively proved to impact susceptibility to infectious and inflammatory diseases.[37 38] The VDR has been shown to be linked to TLRs. It was found that TLR activation of human macrophages upregulated expression of the VDR and the vitamin D_1-hydroxylase genes.[39] This link suggests that the difference among human populations' ability in producing vitamin D may contribute to susceptibility to microbial infection. Furthermore, a case–control and family study reported the association between VDR polymorphisms and susceptibility to tuberculosis.[40]

The strongest three-way epistasis models we found (figure 5) could further extend and strengthen the understanding of how those relevant genes might work in a synergistic way. Although the synergistic effects were captured in a statistical manner, they may indicate either direct or indirect biological relationships among those genetic factors. In particular, the strongest $IG_{partition}$ model shows a 'nested' epistasis hierarchy with both strong pairwise interactions and a strong pure three-way interaction. These three SNPs are from VDR and CD209, which indicates a potential correlation between these two genes. More interestingly, the strongest IG_{strict} model shows a three-way synergy among TLR9, PTX3, and CD209. There have been studies reporting correlations and molecular interactions between TLRs and CD209.[41] However, no published research has indicated the three-way biological synergy among all three genes. Our findings may suggest that multiple pathways interweave in the innate immune system to defend the human body against *M tuberculosis*, and the deficiencies in all those three genetic factors greatly increase the risk of developing tuberculosis. We believe that with further biological validations, our findings could be very helpful in predicting high-risk *M tuberculosis*-infected individuals and preventing their tuberculosis clinical developments.

Investigating high-order genetic interactions is an arduous task, and yet essential for understanding the complex genetic architecture of human diseases. The effectiveness of our information-gain approach in detecting three-way interactions was verified using both synthetic and real genetics data. Note that our measurement is scalable regardless of the size of genetics data. However, when large-scale genome-wide data are considered and exhaustive enumeration of all possible three-way genetic attribute combinations is infeasible, data pre-screening using intelligent data-mining techniques or biology knowledge will be required. There are some interesting venues to extend our approach in future studies. First, it is important to study the genetic interactions on continuous traits. In that case, the probability density functions for continuous random

variables will be used to replace the probability mass function in current discrete information-theoretic measures. Second, it will be interesting to extend the measures on synergies higher than three way. However, as more attributes are involved, the interaction hierarchy gets more complicated. More carefully designed mathematical measures are required. We anticipate that designing powerful and efficient methods to quantify high-order epistasis has the great potential in improving disease treatment and healthcare by revealing the genetic complexity of common human diseases.

Correction notice This paper has been corrected since it was published Online First. A number of equations have been corrected, and the funding statement has been reworded.

Contributors TH designed the study, performed the analyses, and drafted the manuscript. YC participated in the design of the study and drafting the manuscript. JWK participated in the design of the study. RLC participated in the analyses. CW, GS, and SMW conducted the tuberculosis data collection. JHM conceived of the study and participated in its design.

Funding This work was supported by the National Institutes of Health (NIH) of the USA by grants R01-LM009012, R01-LM010098, R01-AI59694, and R01-EY022300 to JHM.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Hardy J, Singleton A. Genome-wide association studies and human disease. *N Engl J Med* 2009;360:1759–68.
2 Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;6:95–108.
3 Wang WYS, Barratt BJ, Clayton DG, *et al*. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;6:109–18.
4 Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;26:1135–45.
5 Clark AG, Boerwinkle E, Hixson J, *et al*. Determinants of the success of whole-genome association testing. *Genome Res* 2005;15:1463–7.
6 Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 2010;26:445–55.
7 Carlborg O, Haley CS. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* 2004;5:618–524.
8 Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002;11:2463–8.
9 Cordell HJ. Detecting gene–gene interactions that underlie human diseases. *Nat Rev Genet* 2009;10:392–404.
10 Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 2003;56:73–82.
11 Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 2005;27:637–46.
12 Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet* 2009;85:309–20.
13 Upstill-Goddard R, Eccles D, Fliege J, *et al*. Machine learning approaches for the discovery of gene–gene interactions in disease data. *Brief Bioinform* 2012. online first May 18, 2012 doi: 10.1093/bib/bbs024
14 Anastassiou D. Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol* 2007;3:83.
15 Chanda P, Sucheston L, Zhang A, *et al*. AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics* 2008;180:1191–210.
16 Chanda P, Sucheston L, Zhang A, *et al*. The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors. *Eur J Hum Genet* 2009;17:1274–86.
17 Fan R, Zhong M, Wang S, *et al*. Entropy-based information gain approaches to detect and to characterize gene–gene and gene–environment interactions/correlations of complex diseases. *Genet Epidemiol* 2011;35:706–21.
18 Hu T, Sinnott-Armstrong NA, Kiralis JW, *et al*. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics* 2011;12:364.
19 McKinney BA, Crowe JE, Guo J, *et al*. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genetics* 2009;5:e1000432.
20 Moore JH, Barney N, Tsai C-T, *et al*. Symbolic modeling of epistasis. *Hum Hered* 2007;63:120–33.
21 Moore JH, Gilbert JC, Tsai C-T, *et al*. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* 2006;241:252–61.
22 Cover TM, Thomas JA. *Elements of information theory*, 2nd edn. New York, NY: Wiley, 2006.
23 Chechik G, Globerson A, Tishby N, *et al*. Group redundancy measures reveal redundancy reduction in the auditory pathway. Dietterich TG, Becker S, Ghaharamani Z *Advances in neural information processing systems*. Cambridge, MA:MIT Press, 2002:173–80.
24 Varadan V, Miller DM III, Anastassiou D. Computational inference of the molecular logic for synaptic connectivity in C. elegans. *Bioinformatics* 2006;22:e497–506.
25 Olesen R, Wejse C, Velez DR, *et al*. DC-SIGN (CD209), pentraxin 3 and vitamin D receptor gene variants associate with pulmonary tuberculosis risk in West Africans. *Genes Immun* 2007;8:456–67.
26 Urbanowicz RJ, Kiralis J, Fisher JM, *et al*. Predicting the difficulty of pure, strict, epistatic models: metrics for simulated model selection. *BioData Mining* 2012;5:15.
27 Urbanowicz RJ, Kiralis JW, Sinnott-Armstrong NA, *et al*. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Mining* 2012;5:16.
28 Bjerregaard-Andersen M, da Silva ZJ, Ravn P, *et al*. Tuberculosis burden in an urban population: a cross sectional tuberculosis survey from Guinea Bissau. *BMC Infect Dis* 2010;10:96.
29 Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–504.
30 West DB. *Introduction to graph theory*, 2nd edn. Upper Saddle River, NJ:Prentice Hall, 2001.
31 Greene CS, Himmelstein DS, Nelson HH, *et al*. Enabling personal genomics with an explicit test of epistasis. *Pac Symp Biocomput* Stanford, CA 2010;327–36.
32 Varadan V, Anastassiou D. Inference of disease-related molecular logic from system-based microarray analysis. *PLoS Comput Biol* 2006;2:e68.
33 Barreiro LB, Neyrolles O, Babb CL, *et al*. Promoter variation in the DC-SIGN-encoding gene CD209 is associated with tuberculosis. *PLoS Med* 2006;3:e20.
34 Tailleux L, Pham-Thi N, Bergeron-Lafaurie A, *et al*. DC-SIGN induction in alveolar macrophages defines privileged target host cells for mycobacteria in patients with tuberculosis. *PLoS Med* 2005;2:e381.
35 Mantovani A, Garlanda C, Doni A, *et al*. Pentraxins in innate immunity: from C-reactive protein to the long pentraxin PTX3. *J Clin Immunol* 2008;28:1–13.
36 Azzurri A, Sow OY, Amedei A, *et al*. IFN-γ-inducible protein 10 and pentraxin 3 plasma levels are tools for monitoring inflammation and disease activity in *Mycobacterium tuberculosis* infection. *Microbes Infect* 2005;7:1–8.
37 Doherty TM, Arditi M. TB, or not TB: that is the question—does TLR signaling hold the answer? *J Clin Invest* 2004;114:1699–703.
38 Schroder NWJ, Schumann RR. Single nucleotide polymorphisms of Toll-like receptors and susceptibility to infectious disease. *Lancet Infect Dis* 2005;5:156–64.
39 Liu PT, Stenger S, Li H, *et al*. Toll-like receptor triggering of a vitamin D-mediated human antimicrobial response. *Science* 2006;311:1770–3.
40 Bornman L, Campbell SJ, Fielding K, *et al*. Vitamin D receptor polymorphisms and susceptibility to tuberculosis in West Africa: a case–control and family study. *J Infect Dis* 2004;190:1631–41.
41 Gringhuis SI, den Dunnen J, Litjens M, *et al*. C-type lectin DC-SIGN modulates toll-like receptor signaling via Raf-1 kinase-dependent acetylation of transcription factor NF-κB. *Immunity* 2007;26:605–16.