European Journal of Human Genetics (2008) 16, 644–651

www.nature.com/ejhg

ARTICLE

Composite measure of linkage disequilibrium for testing interaction between unlinked loci

Xuesen Wu[1,2], L Jin[1,3] and Momiao Xiong*[,1,4]

[1]*School of Life Science, Fudan University, Shanghai, China;* [2]*Department of Epidemiology and Statistics, Bengbu Medical College at Bengbu, Anhui, China;* [3]*CAS-MPG Partner Institute of Computational Biology, SIBS, CAS, Shanghai, China;* [4]*Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA*

Widely used statistical interaction models essentially treated the interaction effect as a residual term and hence are likely to limit the power to detect interaction. Alternatively, interactions between two loci can be understood as irreducible dependencies between loci causing disease or viewed as the linkage disequilibrium (LD) between them. This motivated the development of LD-based statistics for the detection of interaction between two loci. Although LD-based statistics have demonstrated high power to detect interaction between two loci, in general, linkage phase information of marker loci for unrelated individuals is unknown. To overcome this limitation, we classify the interaction between two loci into intragametic interaction that characterizes interaction of two alleles from different loci on the same haplotype and intergametic interaction that characterizes the interaction of two alleles from different loci on different haplotypes. Then we show that intragametic and intergametic interaction will lead to the corresponding intragametic and intergametic LD. This stimulates the use of composite measure of LD for developing statistics to detect interaction between two unlinked loci. To study the validity of the composite LD-based statistic for testing interaction, we estimate its type 1 error rates by simulation. To evaluate the performance of the composite LD-based statistic for detection of interaction between two loci, we compare its power with logistic regression and apply it to two real examples. The preliminary results demonstrate that the composite LD-based statistic is a strong alternative to the logistic regressions and the intragametic LD-based statistic for the detection of interaction between two unlinked loci.
European Journal of Human Genetics (2008) **16,** 644–651; doi:10.1038/sj.ejhg.5202004; published online 23 January 2008

Keywords: composite measure of linkage disequilibrium; interaction; association; complex diseases and logistic regressions

Introduction

It is increasingly recognized that common diseases are not consequences of independent actions of the genes, but are caused by complex joint actions of multiple genetic and environmental risk factors. Gene–gene interactions play an essential role in the ignition and development of the diseases.[1]

Despite current enthusiasm for investigation of interactions between genes, the essential issue of how to define and detect gene–gene interaction remains unresolved.[2] In the past, statistical and biological interactions are often defined separately. As Rothman *et al*[3] pointed out, 'The term statistical interaction is intended to denote the interdependence between the effects of two or more factors

*Correspondence: Dr M Xiong, School of Life Science, Fudan University, Shanghai 200433, China or Human Genetics Center, University of Texas Health Science Center at Houston, PO Box 20334, Houston, Texas 77225, USA.
Tel: +1 713 500 9894; Fax: +1 713 500 0900;
E-mail: Momiao.Xiong@uth.tmc.edu
Received 3 April 2007; revised 10 December 2007; accepted 13 December 2007; published online 23 January 2008

within the confines of a given model of risk', and 'Biological interaction may be defined as the interdependent operation of two or more causes to produce disease'. A core part of statistical interaction is to specify statistical models. Most popular models for statistical interactions between genes are additive models that defined the effect of gene interactions as a statistical deviance from the additive effects of single genes in the linear models (or logistic regression for qualitative traits) and were originally proposed by Fisher[4,] and further developed into their modern representations by Cockerham[5] and Kempthorne.[6] Statistical interaction models essentially treated the interaction effect as a residual term in genetic analysis, and hence are likely to limit the power to detect the interaction.

As an alternative to statistical interaction models, interactions between two loci (or genes) can be understood as irreducible dependencies between loci causing disease.[7] The purpose of a new definition of interaction is to develop a mathematical representation of biological interaction, which is close to the true biological interaction. We use penetrance of the risks to measure the degrees of the risks in causing diseases. In a broad sense, the interaction corresponds to the situation in which the effect of one locus (gene) is affected by the presence or absence of the other.[8–10] The presence of interaction between two loci implies that the two loci share something in common to cause diseases (or phenotype). The shared common features or information lead to the association of two loci in the disease population, that is, high dependency or correlation between two loci in the disease population. In the language of population genetics, the dependency between two loci corresponds to the linkage disequilibrium (LD) between two loci. In other words, although LD between two loci is not the interaction of the effects of those alleles on a disease, LD can be used to detect interaction. If we assume that the controls are sampled from a single isolated population, two unlinked loci are in linkage equilibrium in controls. However, the interaction between two loci will generate LD in disease population.[11] Therefore, we can use the difference in LD between controls and cases to assess whether the interaction between two unlinked loci is present or not. If we assume that two loci are unlinked in the controls, in the presence of interaction, we observe LD between two loci in the cases. The level of LD due to interaction in the disease population depends on the magnitude of interaction between two loci. This motivated the development of statistics based on deviations from linkage equilibrium in the cases for detection of interaction between two loci.

Although LD-based statistics have demonstrated high power to detect interaction between two loci, in general, linkage phase information of marker loci for unrelated individuals is unknown; only genotype data are available. Experiments for generation of haplotype data are expensive and time consuming. Estimation of haplotypes based on genotype data inevitably incurs errors, which in turn will lead to increasing false interaction positive interaction in detection of interaction between two loci.[12] The main purpose of this paper is to directly use unphased genotypes to develop statistics for the detection of interaction between two unlinked loci. Similar to the Hardy–Weinberg disequilibrium at marker loci, which can be used to develop an association test,[13] the composite measure of LD[14–16] that uses the genotype data to estimate the nonrandom association of alleles from different loci on the chromosomes, which are from the same parent (intragametic LD) and on the chromosomes, which are fromdifferent parents (intergametic LD), was used to design association tests allowing unknown linkage phase.[17–19] We extend the composite measure of LD to test the interaction between two unlinked loci when only genotype data are available. To achieve this, we first develop a general theory to study intragametic and intergametic LD patterns under two-locus disease models. Then we develop a novel definition and measure of intragametic interaction, which is caused by two interacted alleles from unlinked loci on the same haplotype and intergametic interaction, which is caused by two interacted alleles from unlinked loci on different haplotypes. The pattern of intragametic and intergametic LD between two unlinked loci due to gene–gene interaction provides a foundation for developing statistics for the detection of interaction between two loci using genotype data. This motivates us to develop the composite LD-based statistics for testing interactions between two unlinked loci. To study the validity of the composite LD-based statistic for testing interaction, we estimate type 1 error rates of the test statistic using simulation. To evaluate the performance of the composite LD-based statistic for detection of interaction between two loci, we compare its power with logistic regression and apply it to two real examples.

Methods

Measure of interaction between two loci

Let D_1 and d_1 be the two alleles at the first disease locus with frequencies P_{D_1} and P_{d_1}, respectively. Let D_2 and d_2 be the two alleles at the second disease locus with frequencies P_{D_2} and P_{d_2}, respectively. Alleles D_1 and d_1 can be indexed by 1 and 2, respectively. At the first disease locus, let D_1D_1 be genotype 11, D_1d_1 be genotype 12 (or d_1D_1 be genotype 21) and d_1d_1 be genotype 22. Thus, the genotypes at the first disease locus can be indexed by *ij*. The genotypes at the second disease locus are similarly defined and can be indexed by *kl*. Two-locus genotypes are simply denoted by *ijkl* for individuals carrying the genotype *ij* at the first disease locus and *kl* at the second disease locus. Let f_{ijkl} be the penetrance of the individuals with genotype

ijkl. Let P_{11}, P_{12}, P_{21}, and P_{22} be the frequencies of haplotypes $H_{D_1D_2}$, $H_{D_1d_2}$, $H_{d_1D_2}$ and $H_{d_1d_2}$ in the general population, respectively. Let P^A_{11}, P^A_{12}, P^A_{21}, and P^A_{22} be their corresponding haplotype frequencies in the disease population. Let $P_{1/1}$, $P_{1/2}$, $P_{2/1}$ and $P_{2/2}$ be the frequencies of H_{D_1/D_2}, H_{D_1/d_2}, H_{d_1/D_2} and H_{d_1/d_2}, respectively, where the slash denotes the two chromosomes in the individual, which are from different parents. Let $P^A_{1/1}$, $P^A_{1/2}$, $P^A_{2/1}$, and $P^A_{2/2}$ be their corresponding frequencies of H_{D_1/D_2}, H_{D_1/d_2}, H_{d_1/D_2} and H_{d_1/d_2} in the disease population. Let $P^A_{D_1}$, $P^A_{d_1}$, $P^A_{D_2}$ and $P^A_{d_2}$ be the frequencies of the alleles D_1, d_1, D_2, and d_2 in the disease population, respectively.

In general, it is genotypes that have penetrances. For ease of discussion, we introduce a concept of haplotype penetrance. Consider a haplotype with two alleles at the different loci on the same chromosome. Then, the penetrance of haplotype $H_{D_1D_2}$ is defined as

$$h_{11} = [P^{D_1D_2}_{D_1D_2} f_{1111} + \frac{1}{2}(P^{D_1D_2}_{D_1d_2} f_{1112} + P^{D_1D_2}_{d_1D_2} f_{1211} + P^{D_1D_2}_{d_1d_2} f_{1212})]/P_{11}$$

In other words, the penetrance of haplotype $H_{D_1D_2}$ is defined as the probability that individual with the haplotype $H_{D_1D_2}$ is affected. It is a weighted sum of the penetrances that contain haplotype $H_{D_1D_2}$. The penetrance h_{12}, h_{21}, and h_{22} is similarly defined.

The penetrance of two alleles at different loci on different chromosomes H_{D_1/D_2} can be defined as

$$h_{1/1} = [P^{D_1D_2}_{D_1D_2} f_{1111} + \frac{1}{2}(P^{D_1D_2}_{D_1d_2} f_{1112} + P^{D_1D_2}_{d_1D_2} f_{1211} + P^{d_1D_2}_{D_1d_2} f_{2112})]/P_{1/1}.$$

It is a weighted sum of genotypic penetrances. Similarly, we can define the penetrance $h_{1/2}$, $h_{2/1}$, and $h_{2/2}$. If we assume the Hardy–Weinberg equilibrium and genotypic equilibrium in general population, then we have $h_{11} = h_{1/1}$, $h_{12} = h_{1/2}$, $h_{21} = h_{2/1}$, and $h_{22} = h_{2/2}$. Let $\delta_{D_1D_2} = P_{11} - P_{D_1}P_{D_2}$ be the measure of intragametic LD that measures the association of alleles from different loci on the same haplotype[17] and $\delta_{D_1/D_2} = P_{1/1} - P_{D_1}P_{D_2}$ be the measure of intergametic LD that measures the association of two alleles from different loci on different haplotypes[17] in the general population. We can show that haplotype frequencies in disease population can be expressed as

$$P^A_{11} = \frac{P_{11}h_{11}}{P_A}, P^A_{12} = \frac{P_{12}h_{12}}{P_A}, P^A_{21} = \frac{P_{21}h_{21}}{P_A}, P^A_{22} = \frac{P_{22}h_{22}}{P_A}, \quad (1)$$

and

$$P^A_{1/1} = \frac{P_{1/1}h_{1/1}}{P_A}, P^A_{1/2} = \frac{P_{1/2}h_{1/2}}{P_A}, P^A_{2/1} = \frac{P_{2/1}h_{2/1}}{P_A}, P^A_{2/2} = \frac{P_{2/2}h_{2/2}}{P_A}, \quad (2)$$

where P_A denotes disease prevalence.

Now we calculate the measures of intragametic and intergametic LD in disease population under a general two-locus disease model. The measures of intragametic and intergametic LD in disease population are denoted by $\delta^A_{D_1D_2}$ and $\delta^A_{D_1/D_2}$, respectively. We can show that they can be given by

$$\delta^A_{D_1D_2} = \frac{\delta_{D_1D_2}}{P_A}h_{11} + \frac{P_{D_1}P_{D_2}}{P_A}(h_{11} - \frac{h_{D_1}h_{D_2}}{P_A}) \quad (3)$$

and

$$\delta^A_{D_1/D_2} = \frac{\delta_{D_1/D_2}}{P_A}h_{1/1} + \frac{P_{D_1}P_{D_2}}{P_A}(h_{1/1} - \frac{h_{D_1}h_{D_2}}{P_A}) \quad (4)$$

where $h_{D_1} = P(\text{Affected}|D_1)$ and $h_{D_2} = P(\text{Affected}|D_2)$. We define a measure of intragametic interaction that measures the interaction of two alleles from different loci on the same haplotype as $I_{\text{int }ra} = h_{11} - \frac{h_{D_1}h_{D_2}}{P_A}$ and a measure of intergametic interaction that measures the interaction of two alleles from different alleles on the different haplotypes as $I_{\text{int }er} = h_{1/1} - \frac{h_{D_1}h_{D_2}}{P_A}7$. Then a measure of total interaction between two loci, which consists of intragametic and intergametic interaction is given by

$$I = I_{\text{int }ra} + I_{\text{int }er} \quad (5)$$

Equation clearly shows that the interaction between two loci is defined by the penetrance of the two loci. Although the penetrance of the risks is not directly related to the biological process, it is related to the causes of the disease. Therefore, the above definition of interaction may have something to do with biological interaction. It follows from equations (3–5) that the composite measure of LD, $\Delta^A_{D_1D_2}$ (Weir 1996) in disease population is given by

$$\Delta^A_{D_1D_2} = \delta^A_{D_1D_2} + \delta^A_{D_1/D_2} = \frac{\delta_{D_1D_2}}{P_A}h_{11} + \frac{\delta_{D_1/D_2}}{P_A}h_{1/1} + \frac{P_{D_1}P_{D_2}}{P_A}I \quad (6)$$

Absence of interaction between two loci is then defined as

$$h_{11} = \frac{h_{D_1}h_{D_2}}{P_A} \text{ or } \frac{h_{11}}{P_A} = \frac{h_{D_1}}{P_A}\frac{h_{D_2}}{P_A}, h_{1/1} = \frac{h_{D_1}h_{D_2}}{P_A} \text{ or } \frac{h_{1/1}}{P_A} = \frac{h_{D_1}}{P_A}\frac{h_{D_2}}{P_A} \quad (7)$$

equation (7) indicate that similar to linkage equilibrium where frequency of a haplotype is equal to the product of the frequencies of the component alleles of the haplotype, absence of interaction between two loci implies that the proportion of individuals carrying two alleles (either in the same chromosome or in the different chromosome) in the disease population is equal to the product of proportions of individuals carrying single allele in the disease population, if we assume that the disease is caused by only two investigated disease loci. In other words, the interaction

between two disease susceptibility loci occurs when the contribution of one locus to the disease depends on another locus. In contrast to additive model for interaction, which was introduced by Fisher[4], the interaction model defined by equations (5 and 7) are referred as to a multiplicative interaction model.

Indirect interaction between two unlinked marker loci

In the previous section, we studied interaction between two unlinked disease loci. Now we consider two marker loci, each of which is in LD with either of the two interacting loci. Assume marker M_1 is in LD with disease locus D_1 and marker M_2 is in LD with disease locus D_2. Furthermore, we assume that two disease loci D_1 and D_2 are unlinked. Let $\delta^A_{M_1M_2}$ and $\delta^A_{M_1/M_2}$ be the measures of intragametic and intergametic LD between two marker loci in the disease population, respectively. We denote the composite measure of LD between two marker loci by $\Delta^A_{M_1M_2}$. Let δ_i be the LD measure between marker M_i and disease locus D_i ($i=1,2$) in the general population. Then, we can show that (Appendix A)

$$highlight7Delta^A_{M_1M_2} = \delta^A_{M_1M_2} + \delta^A_{M_1/M_2} = \frac{\delta_1\delta_2}{P_{D_1}P_{D_2}P_{d_1}P_{d_2}}\Delta^A_{D_1D_2} \quad (8)$$

It is clear that when the marker loci are the disease loci themselves, $\Delta^A_{M_1M_2}, \delta^A_{M_1M_2}$ and $\delta^A_{M_1/M_1}$ are reduced to $\Delta^A_{D_1D_2}$, $\delta^A_{D_1D_2}$ and $\delta^A_{D_1/D_2}$. equation (8) can also be written in terms of the measure of interaction between two unlinked loci

$$\Delta^A_{M_1M_2} = \frac{\delta_1\delta_2}{P_AP_{d_1}P_{d_2}}I \quad (9)$$

Since $\delta_i \le P_{D_i}P_{d_i}$, the absolute value of the LD measure between two unlinked marker loci in the disease population, for example, the composite measure of LD between two marker loci $|\Delta^A_{M_1M_2}|$, will be less than or equal to the absolute value of the composite measure of LD between two unlinked disease loci in the disease population.

Test statistic

In the previous section, we showed that under the multiplicative disease model, interaction between unlinked loci will create LD. Intuitively, we can test interaction by comparing the difference in the composite genotypic disequilibrium between two unlinked loci between cases and controls. Precisely, if we denote the estimators of the composite LD measures in cases and controls by $\hat\Delta_A$ and $\hat\Delta_N$, respectively, then the test statistic can be defined as

$$T_I = \frac{(\hat\Delta_A - \hat\Delta_N)^2}{\mathrm{Var}(\hat\Delta_A) + \mathrm{Var}(\hat\Delta_N)} \quad (10)$$

where

$$\begin{aligned}
\hat\Delta_A &= \hat P^A_{11} + \hat P^A_{1/1} - 2\hat P^A_{D_1}\hat P^A_{D_2},\\
\hat\Delta_N &= \hat P^N_{11} + \hat P^N_{1/1} - 2\hat P^N_{D_1}\hat P^N_{D_2},\\
\mathrm{Var}(\hat\Delta_A) &= \frac{1}{n_A}[\hat\pi^A_{D_1} + \hat\delta^A_{D_1})(\hat\pi^A_{D_2} + \hat\delta^A_{D_2})\\
&\quad + \frac{1}{2}\hat\tau^A_{D_1}\hat\tau^A_{D_2}\hat\Delta_A + +\hat\tau^A_{D_1}\hat\delta^A_{D_1D_2D_2} + \hat\tau^A_{D_2}\hat\delta^A_{D_1D_1D_2} + \hat\Delta^A_{D_1D_1D_2D_2}],\\
\mathrm{Var}(\hat\Delta_N) &= \frac{1}{n_G}[\hat\pi^N_{D_1} + \hat\delta^N_{D_1})(\hat\pi^N_{D_2} + \hat\delta^N_{D_2}) + \frac{1}{2}\hat\tau^N_{D_1}\hat\tau^N_{D_2}\hat\Delta_N + +\hat\tau^N_{D_1}\hat\delta^N_{D_1D_2D_2}\\
&\quad + \hat\tau^N_{D_2}\hat\delta^N_{D_1D_1D_2} + \hat\Delta^N_{D_1D_1D_2D_2}]\\
\hat\pi^A_{D_1} &= \hat P^A_{D_1}(1-\hat P^A_{D_1}), \hat\pi^A_{D_2} = \hat P^A_{D_2}(1-\hat P^A_{D2}),\\
\hat\delta^A_{D_1} &= \hat P^A_{D_1D_1} - (\hat P^A_{D_1})^2, \hat\delta^A_{D_2} = \hat P^A_{D_2D_2} - (\hat P^A_{D_2})^2,\\
\hat\tau^A_{D_1} &= (1-2\hat P^A_{D_1}), \hat\tau^A_{D_2} = (1-2\hat P^A_{D_2}),\\
\hat\delta^A_{D_1D_1D_2} &= \hat P^A_{D_1D_1D_2} - \hat P^A_{D_1}\hat\Delta_A - \hat P_{D_2}\hat\delta^A_{D_1} - (\hat P^A_{D_1})^2\hat P^A_{D_2}\\
\hat\delta^A_{D_1D_2D_2} &= \hat P^A_{D_1D_2D_2} - \hat P^A_{D_2}\hat\Delta_A - \hat P_{D_1}\hat\delta^A_{D2} - (\hat P^A_{D_2})^2\hat P^A_{D_1},\\
\hat\Delta^A_{D_1D_1D_2D_2} &= \hat P^A_{D_1D_1/D_2D_2} - 2\hat P^A_{D_1}\hat\delta^A_{D_1D_2D_2} - 2\hat P^A_{D_2}\hat\delta_{D_1D_1D_2} - 2\hat P^A_{D_1}\hat P^A_{D_2}\hat\Delta_A\\
&\quad -(\hat\Delta_A)^2 - (\hat P^A_{D_1})^2\hat\delta^A_{D_2} - (\hat P^A_{D_2})^2\hat\delta^A_{D_1} - \hat\delta^A_{D_1}\hat\delta^A_{D_2} - (\hat P^A_{D_1}\hat P^A_{D_2})^2
\end{aligned}$$

$\hat\pi^N_{D_1}, \hat\pi^N_{D_2}, \hat\tau^N_{D_1}, \hat\tau^N_{D_2}, \hat\delta^N_{D_1}, \hat\delta^N_{D_2}, \hat\delta^N_{D_1D_2D_2}, \hat\delta^N_{D_1D_1D_2}$ and $\hat\Delta^N_{D_1D_1D_2D_2}$ are similarly defined for controls, the formula for calculations of the composite measure of LD in cases and controls is given in Weir (1996),[15] $P^A_{11}, P^A_{1/1}, P^A_{D_1}, P^A_{D_2}, P^N_{11}P^N_{1/1}, P^N_{D_1}$ and $P^N_{D_2}$ are defined as before, $\hat P^A_{11}, \hat P^A_{1/1}, \hat P^A_{D_1}, \hat P^A_{D_2}, \hat P^N_{11}\hat P^N_{1/1}, \hat P^N_{D_1}$ and $\hat P^N_{D_2}$ are their estimators, the quantities n_A and n_G denote the number of sampled individuals in cases and controls, respectively; the variance of the composite LD measure was the large-sample variance.[15] Under the null hypothesis and assumption of the Hardy–Weinberg equilibrium, the variance of the composite measure of LD in cases and controls becomes $\mathrm{Var}(\hat\Delta_A) = \frac{\hat\pi^A_{D_1}\hat\pi^A_{D_2}}{n_A}$ and $\mathrm{Var}(\hat\Delta_N) = \frac{\hat\pi^N_{D_1}\hat\pi^N_{D_2}}{n_G}$. When sample size is large enough to ensure application of large sample theory, test statistic T_I is asymptotically distributed as a central $\chi^2_{(1)}$ distribution under the null hypothesis of no interaction (both intragametic and intergametic interactions) between two unlinked loci and assumption of the Hardy–Weinberg equilibrium.

In theory, we can use case only design to study interaction between two loci. However, in practice, background LD between two unlinked loci may exist in the population due to many unknown factors. Therefore, the test statistic based on case–control design is more robust than the statistic based on case only design.

Results

Type 1 error rates of test statistics

To examine the validity of the statistic for testing interaction, we performed a series of simulation studies. The computer program SNaP[20] was used to generate two-locus genotype data of the sample individuals. A total of 20 000 individuals, who were equally divided into cases and controls were generated in the general population,

assuming genotypic equilibrium (both intragametic and intergametic equilibria) between two loci. We randomly sampled 100–400 individuals from each of the cases and controls for the calculation of the type I error rates. A total of 10 000 simulations were repeated. Table 1 shows that the estimated type I error rates of the statistic T_I for testing the interaction between two unlinked loci were not appreciably different from the nominal levels $\alpha = 0.05$, $\alpha = 0.01$ and $\alpha = 0.001$.

Power evaluation

To evaluate the performance of the composite LD-based statistic in testing gene–gene interaction, we compared the power of the statistic employing composite measure of the LD to that of the logistic model. We use the genotype coding scheme in QUANTO[21] for power calculations. Specifically, we considered two types of genotype coding (genetic covariate variables). For a dominant model, homozygous wild type, heterozygous, and homozygous mutant genotypes were coded as 0, 1, and 1, respectively. For an additive model, they were coded as 0, 1, and 2, respectively. We considered two loci, denoted as G and H, respectively. We assume the following logistic model:

$$P(D = 1|G, H) = \frac{e^{\alpha+\beta_g G+\beta_h H+\beta_{gh} GH}}{1 + e^{\alpha+\beta_g G+\beta_h H+\beta_{gh} GH}}$$

where $OR_b = \frac{e^{\alpha}}{1+e^{\alpha}}$ is the baseline probability of disease in the population,

$$OR_G = e^{\beta_g}, OR_H = e^{\beta_h} \text{ and } OR_{GH} = e^{\beta_{gh}}$$

are the odds ratios for G when H = 0, H when G = 0 and interaction G × H, respectively.[21] Power for both composite LD-based statistic and logistic regression[22] was calculated by simulation. The computer program SNaP[20] was used to generate 10 000 cases and 10 000 controls with unlinked two-locus genotype data. Two-locus interaction effect were simulated for two-locus dominant and additive models with penetrance functions as given in Gauderman (2002).[21] Five hundred individuals were randomly sampled from each of the cases and controls. A total of 10 000 simulations were repeated. Figures 1a and 1b present the power comparisons between the logistic regression model and the composite LD-based statistic under the following two genetic interaction models: dominance × dominance and additive × additive. Figures 1a and 1b show that the power of both logistic regression and the composite LD-based statistic in detecting gene–gene interaction was an increasing monotonic function of the interaction odds ratio, a widely used measure in quantifying the strength of interaction between two loci. This implies that the proposed new interaction measure and test statistic are closely related to the traditional interaction measure. We can also see that the power of the composite LD-based

Table 1 Type 1 error rates of the test statistic T_I to test interaction between two unlinked loci in a homogenous population

	Nominal levels		
Sample size	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
100	0.0545	0.0133	0.0012
150	0.0509	0.0105	0.0011
200	0.0511	0.0112	0.0013
250	0.0480	0.0101	0.0008
300	0.0530	0.0100	0.0011
350	0.0510	0.0110	0.0010
400	0.0472	0.0091	0.0012

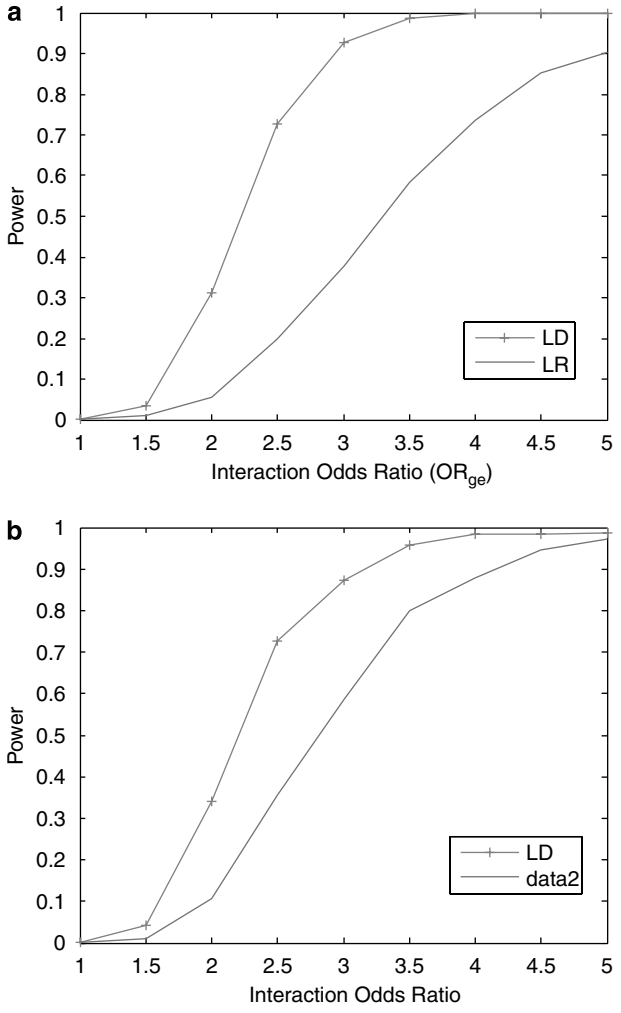

Figure 1 (**a**) Power of the test statistic T_I and logistic regression analysis as a function of interaction odds ratio ($OR_{GH} = e^{\beta_{gh}}$) under a dominance × dominance model, assuming risk allele frequencies at both loci G and H are 0.2, number of individuals in both cases and controls are 500, population risk is 0.001, significance level is 0.001, and genetic odds ratios $R_G = 1$ and $R_H = 1$. (**b**) Power of the test statistic T_I and logistic regression analysis as a function of interaction odds ratio ($OR_{GH} = e^{\beta_{gh}}$) under an additive × additive model, assuming risk allele frequencies at both loci G and H are 0.2, number of individuals in both cases and controls are 500, population risk is 0.001, significance level is 0.001, and genetic odds ratios $R_G = 1$ and $R_H = 1$.

statistic T_I is higher than that of the logistic regression model.

Application to real data examples

To further evaluate its performance for detecting interaction between two unlinked loci, the proposed test statistic T_I was applied to two real examples. The first example was a breast cancer case–control study. A total of 398 Caucasian breast cancer cases and 372 matched controls were sampled from the Ontario Familial Breast Cancer Registry (OFBCR).[23] Nineteen SNPs from 18 key genes in DNA repair, cell cycle, carcinogen/estrogen metabolism, and immune system were typed. All SNPs were in Hardy–Weinberg equilibrium. Using multivariate logistic analysis under the codominant models, four pairs of genes: XPD and IL10, GSTP1 and COMT, COMT and CCND1, and BARD1 and XPD showed significant interactions.[23] We used the statistic T_I to test interactions between these four pairs of genes. The test results are summarized in Table 2, where the crude *P*-values were from the Table 4 in the paper by Onay *et al* (2006).[23] The crude *P*-values were obtained from multivariate logistic regression analysis that includes all main effects and only the interaction of interest under the codominant models. As shown in Table 2, logistic regression analysis interactions between XPD-(Lys751Gln) and IL10-(G(-1082)A), BARD-(Pro24Ser) and XPD-(Lys751Gln), COMT-(Met108/158Val) and CCND1-(Pro24Pro) and GSTP1-(Ile241Val), and COMT-(Met108/158Val) were identified. But after the more conservative Bonferroni adjustment, none of these interactions were significant.[23] Table 2 demonstrated that the *P*-values based on the test statistic T_I were smaller than those based on the traditional logistic regression analysis for the XPD-(Lys751Gln) and IL10-(G(1082)A), but larger for the BARD1-(Pro24Ser)and XPD (Lys751Gln), and COMT-(Met108/158Val) and CCND1 (Pro241Pro).

A popular point of view is that the statistics using haplotype data usually have smaller *P*-values than the statistics using genotype data. To examine this statement, the second example is coronary heart disease study in Shaghai, China in which 812 SNPs in 176 genes were typed for 1320 cases and 1129 controls. Atherosclerosis is the primary cause of coronary heart disease.[24,25] Although a majority of the results demonstrate that the *P*-values of intragametic LD-based statistic is smaller than that of composite LD-based statistic (data not shown), we can still find many pairs of SNPs for which the *P*-values of the composite LD-based statistic are smaller than that of the intragametic LD-based statistic and logistic regression. Here, we report the results of the detected interactions between 10 pairs of SNPs in Table 3. In Table 3, we can see that for all 10 pairs of SNPs, the *P*-values of the composite LD-based statistic are smaller than those of the intragametic LD-based statistic. This indirectly shows that there may exist intragametic and intergametic interactions, which generate intragametic and intergametic LD, respectively. The composite LD is the summation of intragametic and intergametic LD. When both intragametic and intergametic LD have the same sign, the absolute value of composite LD is larger than that of its component.

Table 2 Comparison of *P*-values for testing gene-gene interactions (example 1)

Pair of interactions	P-*values obtained by* *Logistic regression*[a]	*Composite LD-based statistic*
XPD-(Lys751Gln) and IL10-(G(1082)A)	0.035	0.0046
BARD1-(Pro24Ser)and XPD (Lys751Gln)	0.024	0.7038
COMT-(Met108/158Val) and CCND1 (Pro241Pro)	0.010	0.9945
GSTP1-(Ile105Val) and COMT (Met108/158Val)	0.036	0.0075

[a]*P*-values reported by Onay *et al* (2006).[23]

Table 3 *P*-values for testing interaction between unlinked loci in CAD study (example 2)

SNP1	*Gene*	*SNP2*	*Gene*	P-*values* T_I *haplotype*	T_I *genotype*	*Logistic*
rs1511024	FABP2	rs10916683	PLA2G2A	8.07E−02	2.02E−05	5.07E−01
rs1267857	F13A1	rs2071397	GLA	6.72E−04	7.54E−05	3.28E−01
rs2612103	ENOS	rs2515901	GLA	1.73E−04	3.57E−05	9.30E−02
rs2479412	PCSK9	rs2071228	GLA	9.94E−03	6.77E−05	3.69E−01
rs5194	AGTR2	rs4149026	SLCO1B1	5.77E−04	9.80E−05	6.61E−01
rs17014553	GSTM5	rs2515901	GLA	1.27E−04	2.28E−05	2.88E−04
rs3829462	LIPC	rs1126535	CD40L	1.60E−04	1.08E−05	1.95E−02
rs3821664	P2RY12	rs2280964	CXCR3	1.60E−04	5.02E−05	2.51E−02
rs4963516	GNB3	rs1126535	CD40L	1.3E−05	1.44E−06	1.13E−03
rs12990449	LRP1B	rs12840631	AGTR2	3.66E−04	2.60E−05	7.82E−01

Therefore, in this case the *P*-values of the composite LD-based statistic will be smaller than those of the intragametic LD-based statistic.

Discussion

For almost a century, interaction between loci is defined as a deviance from the summation of their genetic main effects of individual locus. As an alternative to additive model of interaction, we have shown that the interaction between loci can be interpreted as irreducible dependencies between them. In genetics, dependencies between loci can be understood as LD If two loci in the general population are in linkage equilibrium (or independent), their departure from equilibrium in the disease population is often attributed to the interaction between them. Therefore, the LD due to interaction between two loci can be used to measure the magnitude of interaction.

The most popular measure of LD is the intragametic LD measure that quantifies nonrandom association of two alleles from different loci on the same haplotype. The major limitation of using the intragametic LD measure to test for interaction is that in practice, haplotype data are often unavailable. Although a number of algorithms for estimation of haplotypes have been developed, the errors of haplotype estimation are inevitable. This will lead to inaccuracy in the detection of interactions between loci. To overcome this limitation, we proposed to use the composite measure of LD based on genotype data for detection of interactions between loci.

To gain a deep understanding of intragametic and intergametic interactions, we first developed the general theory to study composite LD patterns in the disease population under two-locus disease models. We introduced a new concept of intragametic and intergametic penetrance and developed a measure of interaction between two unlinked loci, including both intragametic and intergametic interactions. The theoretic analysis of the intragametic and intergametic LD motivated us to use a composite measure of LD for developing statistics to test interactions.

We examined the distribution of the composite LD-based statistic under the null hypothesis of no interaction and calculated type 1 error rates of the proposed statistic by simulation. Our results showed that type 1 error rates were close to nominal significance levels. The composite LD-based statistic has two remarkable features. First, the calculation of the composite LD-based statistic does not require linkage phase information. Therefore, the results of the composite LD-based statistic are more reliable than that of the intragametic LD-based statistic. Second, the power of the composite LD-based statistic may not always be less than that of the intragametic LD-based statistic. Although by simulation we showed that in general, the composite LD-based statistic under the dominant and additive two-locus disease models has higher power than the logistic regressions, the critical question is whether there are situations where the composite LD-based statistic has higher power than the traditional LD (intragametic LD)-based statistic. The preliminary results of real data analysis showed that in some cases, *P*-values of the composite LD-based statistic may be smaller than those of the intragametic LD-based statistic. equations (6 and 10) show that in theory, the composite LD-based statistic varies from half of the intragametic LD-based statistic to two times of the intragametic LD-based statistic depending on the ratio of the intergametic LD over the intergametic LD. Therefore, when the intergametic LD is comparable with the intragametic LD, the composite LD-based statistic may have higher power than the intragametic LD-based statistic.

Although the composite LD-based statistic has merit, it also has potential limitations. First, in addition to interaction, HWD may also increase the composite LD. The small *P*-values of the composite LD-based statistic may be caused by HWD, not by the interaction. Although this will not be a problem for association studies of two loci with the disease, but it will be the problem for gene–gene interaction analysis. Second, like other population-based methods, the population substructure may generate LD and hence create spurious interactions. Third, the presented methods in this report require that the two loci are unlinked.

In summary, our results suggest that the composite LD-based statistic is an alternative to the traditional logistic regression or the haplotype-based LD statistics.

Acknowledgements

M Xiong is supported by NIH-NIAMS Grant P01 AR052915-01A1, NIH Grant HL74735, and ES09912 in the US and Shanghai Commission of Science and Technology Grant 04dz14003 in China. L Jin is supported by Shanghai Commission of Science and Technology Grant 04dz14003, China. X Wu is supported by Shanghai Commission of Science and Technology Grant 04dz14003 and postdoctoral fund 05R214115 from Shanghai, China.

References

1 Ay N: Locality of global stochastic interaction in directed acyclic networks. *Neural Comput* 2002; **14**: 2959–2980.
2 Clayton D, McKeigue PM: Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; **358**: 1356–1360.
3 Rothman KJ, Greenland S, Walker AM: Concepts of interaction. *Am J Epidemiol* 1980; **112**: 467–470.
4 Fisher RA: The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* 1918; **3**: 399–433.
5 Cockerham CC: An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics* 1954; **39**: 859–882.
6 Kempthorne O: The correlation between relatives in a random mating population. *Proc R Soc Lond B* 1954; **143**: 103–113.

7 Jakulin A, Bratko I: Analyzing attribute dependencies; in Lavrač N, Gamberger D, Blockeel H, Todorovski L (eds): *Proceedings of Principles of Knowledge Discovery in Data (PKDD)*; LNAI, 2003; **2838**: 229–240.
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**: 2463–2468.
9 Hansen TF, Wagner GP: Modeling genetic architecture: a multilinear theory of gene interaction. *Theor Popul Biol* 2001; **59**: 61–86.
10 Wagner GP, Laubichler MD, Bagheri-Chaichian H: Genetic measurement theory of epistatic effects. *Genetica* 1998; **102/103**: 569–580.
11 Zhao J, Jin L, Xiong MM: Test for interaction between two unlinked loci. *Am J Hum Genet* 2006; **79**: 831–845.
12 Fallin D, Schork NJ: Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 2000; **67**: 947–959.
13 Nielsen DM, Ehm MG, Weir BS: Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet* 1998; **63**: 1531–1540.
14 Weir B: Inferences about linkage disequilibrium. *Biometrics* 1979; **35**: 235–254.
15 Weir B: *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates, 1996.
16 Weir BS, Cockerham CC: Complete characterization of disequilibrium at two loci; in Feldman MW (ed): *Mathematical Evolutionary Theory*. Princeton, NJ: Princeton University Press, 1989, pp 86–110.
17 Schaid DJ: Linkage disequilibrium testing when linkage phase is unknown. *Genetics* 2004; **166**: 505–512.
18 Zaykin DV, Meng Z, Ehm MG: Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am J Hum Genet* 2006; **78**: 737–746.
19 Nielsen DM, Ehm MG, Zaykin DV, Weir BS: Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics* 2004; **168**: 1029–1040.
20 Nothnagel M: Simulation of LD block-structured SNP haplotype data and its use for the analysis of case–control data by supervised learning methods. *Am J Hum Genet* 2002; **4** (Suppl.): A2363.
21 Gauderman WJ: Sample size requirements for matched case–control studies of gene-gene interaction. *Am J Epidemiol* 2002; **155**: 478–484.
22 Millstein J, Conti DV, Gilliland FD, Gauderman WJ: A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet* 2006; **78**: 15–27.
23 Onay VU, Briollais L, Knight JA *et al*: SNP-SNP interactions in breast cancer susceptibility. *BMC Cancer* 2006; **6**: 114.
24 Lusis AJ, Mar R, Pajukanta P: Genetics of atherosclerosis. *Annu Rev Genomics Hum Genet* 2004; **5**: 189–218.
25 Libby P: Inflammation in atherosclerosis. *Nature* 2002; **420**: 868–874.

Appendix A

Assume that marker locus M_1 has two alleles M_1 and m_1, and the marker locus M_2 has two alleles M_2 and m_2. Let $q^A_{M_1}$ *and* $q^A_{M_2}$ be the frequencies of the marker alleles M_1 and M_2 in the disease population, respectively. Let the frequencies of the haplotypes D_1M_1, D_1m_1, d_1M_1 and d_1m_1 be $P_{D_1M_1}, P_{D_1m_1}, P_{d_1M_1}$ *and* $P_{d_1m_1}$, respectively. The frequencies of the haplotypes D_2M_2, D_2m_2, d_2M_2, and d_2m_2 can be similarly defined. Let the frequencies of the haplotypes M_1M_2, M_1m_2, m_1M_2, and m_1m_2 in the disease population be q^A_{11}, q^A_{12}, q^A_{21}, and q^A_{22}, respectively. Then, we have

$$\begin{aligned} q^A_{11} &= P(M_1M_2|A) \\ &= \frac{P(M_1M_2, A)}{P_A} \\ &= \frac{P_{D_1M_1}P_{D_2M_2}h_{11} + P_{D_1M_1}P_{d_2M_2}h_{12} + P_{d_1M_1}P_{D_2M_2}h_{21} + P_{d_1M_1}P_{d_2M_2}h_{22}}{P_A} \\ &= P_{M_1}P_{M_2} + \frac{P_{M_2}(h_{D_1} - h_{d_1})\delta_1 + P_{M_1}(h_{D_2} - h_{d_2})\delta_2 + (h_{11} - h_{12} - h_{21} + h_{22})\delta_1\delta_2}{P_A} \end{aligned} \tag{A1}$$

Similarly, we have

$$q^A_{M_1} = P_{M_1} + \frac{h_{D_1} - h_{d_1}}{P_A}\delta_1$$

$$q^A_{M_2} = P_{M_2} + \frac{h_{D_2} - h_{D_1}}{P_A}\delta_2 \tag{A2}$$

Note that

$$\begin{aligned} h_{D_1} &= \frac{P(D_1D_2, Affected) + P(D_1d_2, Affected)}{P_{D_1}} \\ &= P_{D_2}h_{11} + P_{d_2}h_{12}, \\ h_{D_2} &= P_{D_1}h_{11} + P_{d_1}h_{21}, \\ h_{d_1} &= P_{D_2}h_{21} + p_{d_2}h_{22}, \\ h_{d_2} &= P_{D_1}h_{12} + P_{d_1}h_{22}. \end{aligned} \tag{A3}$$

It follows from equation (A3) that

$$\begin{aligned} h_{D_1}h_{D_2} &= P_{D_1}P_{D_2}h^2_{11} + P_{d_1}P_{D_2}h_{11}h_{21} + P_{D_1}P_{d_2}h_{11}h_{12} + P_{d_1}P_{d_2}h_{12}h_{21} \\ &= h_{11}P_A + P_{d_1}P_{d_2}(h_{12}h_{21} - h_{11}h_{22}) \end{aligned}$$

Similarly, we have

$$\begin{aligned} h_{D_1}h_{d_2} &= h_{12}P_A + P_{d_1}P_{D_2}(h_{11}h_{22} - h_{12}h_{21}) \\ h_{d_1}h_{D_2} &= h_{21}P_A + P_{D_1}P_{d_2}(h_{11}h_{22} - h_{12}h_{21}) \\ h_{d_1}h_{d_2} &= h_{22}P_A + P_{D_1}P_{D_2}(h_{12}h_{21} - h_{11}h_{22}) \end{aligned} \tag{A4}$$

From equations (A2–A4) we obtain that

$$\begin{aligned} q^A_{M_1}q^A_{M_2} = P_{M_1}P_{M_2} &+ \frac{P_{M_2}(h_{D_1} - h_{d_1})}{P_A}\delta_1 + \frac{P_{M_1}(h_{D_2} - h_{d_2})}{P_A}\delta_2 \\ &+ \left[\frac{h_{11} - h_{12} - h_{21} + h_{22}}{P_A} - \frac{h_{11}h_{22} - h_{12}h_{21}}{P^2_A}\right]\delta_1\delta_2 \end{aligned} \tag{A5}$$

Thus,

$$\begin{aligned} \delta^A_{M_1M_2} &= q^A_{11} - q^A_{M_1}q^A_{M_2} \\ &= \frac{1}{P_AP_{d_1}P_{d_2}}\left(h_{11} - \frac{h_{D_1}h_{D_2}}{P_A}\right)\delta_1\delta_2 \end{aligned} \tag{A6}$$

Similarly, we have

$$\begin{aligned} \delta^A_{M_1/M_2} &= q^A_{1/1} - q^A_{M_1}q^A_{M_2} \\ &= \frac{1}{P_AP_{d_1}P_{d_2}}\left(h_{1/1} - \frac{h_{D_1}h_{D_2}}{P_A}\right)\delta_1\delta_2 \end{aligned} \tag{A7}$$

Combining equations (A6 and A7) yields

$$\begin{aligned} \Delta^A_{M_1M_2} &= \frac{\delta_1\delta_2}{P_AP_{d_1}P_{d_2}}I \\ &= \frac{\delta_1\delta_2}{P_{D_1}P_{D_2}P_{d_1}P_{d_2}}\Delta^A_{D_1D_2}. \end{aligned}$$