Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 38

A Statistical Model for Genetic Mapping of Viral Infection by Integrating Epidemiological Behavior

Yao Li[*] Arthur Berg[†] Myron N. Chang[‡]
Ping Du[**] Kwangmi Ahn[††] David Mauger[‡‡]
Duanping Liao[§] Rongling Wu[¶]

[*]West Virginia University, yli@stat.wvu.edu
[†]Pennsylvania State University, berg@psu.edu
[‡]University of Florida, mchang@cog.ufl.edu
[**]Pennsylvania State University, ping.du@psu.edu
[††]Pennsylvania State University, kahn@hes.hmc.psu.edu
[‡‡]Pennsylvania State University, dmauger@psu.edu
[§]Pennsylvania State University, dliao@psu.edu
[¶]Pennsylvania State University, rwu@hes.hmc.psu.edu

A Statistical Model for Genetic Mapping of Viral Infection by Integrating Epidemiological Behavior*

Yao Li, Arthur Berg, Myron N. Chang, Ping Du, Kwangmi Ahn, David Mauger, Duanping Liao, and Rongling Wu

Abstract

Large-scale studies of genetic variation may be helpful for understanding the genetic control mechanisms of viral infection and, ultimately, predicting and eliminating infectious disease outbreaks. We propose a new statistical model for detecting specific DNA sequence variants that are responsible for viral infection. This model considers additive, dominance and epistatic effects of haplotypes from three different genomes, recipient, transmitter and virus, through an epidemiological process. The model is constructed within the maximum likelihood framework and implemented with the EM algorithm. A number of hypothesis tests about population genetic structure and diversity and the pattern of genetic control are formulated. A series of closed forms for the EM algorithm to estimate haplotype frequencies and haplotype effects in a network of genetic interactions among three genomes are derived. Simulation studies were performed to test the statistical properties of the model, recommending necessary sample sizes for obtaining reasonably good accuracy and precision of parameter estimation. By integrating, for the first time, the epidemiological principle of viral infection into genetic mapping, the new model shall find an immediate application to studying the genetic architecture of viral infection.

KEYWORDS: genetic mapping, haplotype, epidemiological model, viral infection, higher-order genetic interaction

---

*This work is partially supported by Joint NSF/NIH grant DMS/NIGMS-0540745.

1 Introduction

Several serious human diseases, such as AIDS, hepatitis B, and meningitis, are caused by viruses. On exposure to a pathogen, some people may resist infection, some become subclinically infected, whereas others progress through several stages from mild to severe infection. Although such interpersonal variability may be due to environmental risk factors, strong evidence suggests that it may also involve a genetic component from the host genome [1, 2, 3, 4, 5, 6, 7]. A growing number of molecular genetic studies provide new evidence that the spread of many infectious diseases is determined by genetic interactions of loci from the host and viral genomes [8, 9, 10]. This can be exemplified by the fact that the transmitted virus differs from the virus that predominates in the transmitter [11]. Recent work identifies the control mechanism of intricate host-pathogen interaction networks through the effectors of microRNAs (miRNAs), a new class of 18–23 nucleotide long non-coding RNAs [12, 13].

At present, genetic mapping aimed to detect genes for infection diseases is limited to testing the association between genotypes at particular candidate genes and disease progression with a familial design or a random set of patients from a natural population [1], although the availability of whole-genome polymorphic data allows a genome-wide screen of genes. It is likely that such a mapping design is too simple to extract precise information from the data about the genetic control of the disease given its transmission complexities. It has been recognized that the occurrence, spread and outbreak of infectious diseases are affected by the ways diseases are transmitted. Diseases are caused by germs, which are transmitted from one person to another through direct and indirect contacts. The integration of the transmission mechanisms of diseases into an epidemiological model can help not only to construct a network of pathogenic infection and predict the path of disease spread, but also to simulate various intervention strategies to determine the one that might be most effective.

In this article, we develop a statistical model for mapping genes and genetic interactions for infection diseases by incorporating the epidemiological principle of viral infection. We jointly model the genetic segregation of the virus and the host. In particular, the host is considered in terms of both the transmitter and recipient by assuming that the germs carrying pathogens spread through the exchange of body fluids from sexual contact or a blood transfusion. Other transmissions via indirect contacts by air, water and insects, can also be modeled, but will not be considered in this study. The genetic information of the virus and host will be modeled

at the haplotype level constructed by high-throughput single nucleotide polymorphisms (SNPs) [14, 15]. The haplotype model, proven to be powerful for documenting, mapping and understanding the structure and patterns of the human genome linked to a complex phenotype [16, 17, 18], can characterize concrete nucleotides or their combinations that are associated with viral infection. The model developed provides a quantitative framework for testing the genetic control of disease infection through additive, dominance, and different kinds of epistatic interactions. Computer simulation was conducted to examine the statistical properties of the model and its utilization.

2 Method

2.1 Genetic Design

Suppose there is a natural population at Hardy-Weinberg equilibrium consisting of patients who are infected by a type of virus. From this diseased population, we randomly sample a set of patients which are sorted into two groups, recipients and transmitters. The recipients receive the virus because of their close contacts with the transmitters. We assume that the transmitter of each recipient can be traced, allowing two or more recipients to have a common transmitter. The virus of a transmitter was given by its preceding transmitters. In this study, we will not consider the preceding transmitters. Now, viral loads in both the recipients and transmitters are measured as a phenotypic trait. The demographic factors of the patients, such as sex, age, race, and ethnicity, are also recorded.

The host genes of these recipients and transmitters as well as the genes of the virus these host carry are typed for SNPs genome-wide or at particular candidate regions. Let us first consider two SNPs **A** (with two alleles A and a) and **B** (with two alleles B and b) for the diploid host and two SNPs **C** (with an allele C) and **D** (with an allele D) for the haploid virus. We use p (and $1-p$) to denote the frequency of allele A (and a) and q (and $1-q$) to denote the frequency of allele B (and b) in the host population, and r (and $1-r$) to denote the frequency of allele C (and c) and s (and $1-s$) to denote the frequency of allele D (d) in the virus population. The two SNPs for the host and virus are associated with linkage disequilibrium D_H and D_V, respectively.

DOI: 10.2202/1544-6115.1475

The frequencies of four haplotypes for the two host SNPs are denoted as

$$\begin{array}{lcll} p_{11} & = & pq + D_H & \text{for } AB \\ p_{10} & = & p(1-q) - D_H & \text{for } Ab \\ p_{01} & = & (1-p)q - D_H & \text{for } aB \\ p_{00} & = & (1-p)(1-q) + D_H & \text{for } ab. \end{array} \quad (1)$$

The frequencies of four haplotypes for the two virus SNPs are denoted as

$$\begin{array}{lcll} q_{11} & = & rs + D_V & \text{for } CD \\ q_{10} & = & r(1-s) - D_V & \text{for } Cd \\ q_{01} & = & (1-r)s - D_V & \text{for } cD \\ q_{00} & = & (1-r)(1-s) + D_V & \text{for } cd. \end{array} \quad (2)$$

The four host haplotypes derived from the maternal and paternal parents are combined at random to generate 9 observable genotypes, $AABB$ (coded as 1), $AABb$ (coded as 2), ..., $aabb$ (coded as 9). Each of these genotypes may carry one of four possible virus genotype CD (coded as 1), Cd (coded as 2), cD (coded as 3), and cd (coded as 4), thus leading to 36 joint host-virus genotypes. Let $N_{j_r j_t/j_v}$ denote the observation of a joint genotype $j_r j_t/j_v$ where j_r and j_t ($j_r, j_t = 1, ..., 9$) is a genotype of recipients and transmitters, respectively, and j_v ($j_v = 1, ..., 4$) is a virus genotype. With the independence assumption, the frequencies of joint host-virus genotypes are expressed as the products of the host and virus genotype frequencies. Table 1 tabulates the joint host-virus genotype frequencies expressed in terms of haplotype frequencies $\mathbf{\Omega}_H = (p_{11}, p_{10}, p_{01}, p_{00})$ for the host and $\mathbf{\Omega}_V = (q_{11}, q_{10}, q_{01}, q_{00})$ for the virus. Three-way composite diplotype frequencies for the recipients, transmitters, and virus can be obtained by multiplying two-way composite diplotype frequencies (Table 1) by virus haplotype frequencies q_{11}, q_{10}, q_{01}, and q_{00}. The expression of genotypic value for a three-way composite diplotype is given by assuming that AB is the risk haplotype for both the recipients and transmitters and that CD is the risk haplotype for the virus.

Table 1: The frequencies of two-way composite diplotypes (and genotypes) at a pair of SNPs from the recipient and transmitter genomes expressed by host haplotype frequencies.

			Transmitter									
			$AABB$	$AABb$	$AAbb$	$AaBB$	$AaBb$		$Aabb$	$aaBB$	$aaBb$	$aabb$
							$AB\|ab$	$Ab\|aB$				
Recipients			R_TR_T	$R_T\bar{R}_T$	$\bar{R}_T\bar{R}_T$	$R_T\bar{R}_T$	$R_T\bar{R}_T$	$\bar{R}_T\bar{R}_T$	$\bar{R}_T\bar{R}_T$	$\bar{R}_T\bar{R}_T$	$\bar{R}_T\bar{R}_T$	$\bar{R}_T\bar{R}_T$
$AABB$		$R_R R_R$	p_{11}^4 $\mu_{22/1}$	$2p_{11}^3p_{10}$ $\mu_{21/1}$	$p_{11}^2p_{10}^2$ $\mu_{20/1}$	$2p_{11}^3p_{01}$ $\mu_{21/1}$	$2p_{11}^3p_{00}$ $\mu_{21/1}$	$2p_{11}^2p_{10}p_{01}$ $\mu_{20/1}$	$2p_{11}^2p_{10}p_{00}$ $\mu_{20/1}$	$p_{11}^2p_{01}^2$ $\mu_{20/1}$	$2p_{11}^2p_{01}p_{00}$ $\mu_{20/1}$	$p_{11}^2p_{00}^2$ $\mu_{20/1}$
$AABb$		$R_R\bar{R}_R$	$2p_{11}^3p_{10}$ $\mu_{12/1}$	$4p_{11}^2p_{10}^2$ $\mu_{11/1}$	$2p_{11}p_{10}^3$ $\mu_{10/1}$	$4p_{11}^2p_{10}p_{01}$ $\mu_{11/1}$	$4p_{11}^2p_{10}p_{00}$ $\mu_{11/1}$	$4p_{11}p_{10}^2p_{01}$ $\mu_{10/1}$	$4p_{11}p_{10}^2p_{00}$ $\mu_{10/1}$	$2p_{11}p_{10}p_{01}^2$ $\mu_{10/1}$	$4p_{11}p_{10}p_{01}p_{00}$ $\mu_{10/1}$	$2p_{11}p_{10}p_{00}^2$ $\mu_{10/1}$
$AAbb$		$\bar{R}_R\bar{R}_R$	$p_{11}^2p_{10}^2$ $\mu_{02/1}$	$2p_{11}p_{10}^3$ $\mu_{01/1}$	p_{10}^4 $\mu_{00/1}$	$2p_{11}p_{10}^2p_{01}$ $\mu_{01/1}$	$2p_{11}p_{10}^2p_{00}$ $\mu_{01/1}$	$2p_{10}^3p_{01}$ $\mu_{00/1}$	$2p_{10}^3p_{00}$ $\mu_{00/1}$	$p_{10}^2p_{01}^2$ $\mu_{00/1}$	$2p_{10}^2p_{01}p_{00}$ $\mu_{00/1}$	$p_{10}^2p_{00}^2$ $\mu_{00/1}$
$AaBB$		$R_R\bar{R}_R$	$2p_{11}^3p_{01}$ $\mu_{12/1}$	$4p_{11}^2p_{10}p_{01}$ $\mu_{11/1}$	$2p_{11}p_{10}^2p_{01}$ $\mu_{10/1}$	$4p_{11}^2p_{01}^2$ $\mu_{11/1}$	$4p_{11}^2p_{01}p_{00}$ $\mu_{11/1}$	$4p_{11}p_{10}p_{01}^2$ $\mu_{10/1}$	$4p_{11}p_{10}p_{01}p_{00}$ $\mu_{10/1}$	$2p_{11}p_{01}^3$ $\mu_{10/1}$	$4p_{11}p_{01}^2p_{00}$ $\mu_{10/1}$	$2p_{11}p_{01}p_{00}^2$ $\mu_{10/1}$
$AaBb$	$AB\|ab$	$R_R\bar{R}_R$	$2p_{11}^3p_{00}$ $\mu_{12/1}$	$4p_{11}^2p_{10}p_{00}$ $\mu_{11/1}$	$2p_{11}p_{10}^2p_{00}$ $\mu_{10/1}$	$4p_{11}^2p_{01}p_{00}$ $\mu_{11/1}$	$4p_{11}^2p_{00}^2$ $\mu_{11/1}$	$4p_{11}p_{10}p_{01}p_{00}$ $\mu_{10/1}$	$4p_{11}p_{10}p_{00}^2$ $\mu_{10/1}$	$2p_{11}p_{01}^2p_{00}$ $\mu_{10/1}$	$4p_{11}p_{01}p_{00}^2$ $\mu_{10/1}$	$2p_{11}p_{00}^3$ $\mu_{10/1}$
	$Ab\|aB$	$\bar{R}_R\bar{R}_R$	$2p_{11}^2p_{10}p_{01}$ $\mu_{02/1}$	$4p_{11}p_{10}^2p_{01}$ $\mu_{01/1}$	$2p_{10}^3p_{01}$ $\mu_{00/1}$	$4p_{11}p_{10}p_{01}^2$ $\mu_{01/1}$	$4p_{11}p_{10}p_{01}p_{00}$ $\mu_{01/1}$	$4p_{10}^2p_{01}^2$ $\mu_{00/1}$	$4p_{10}^2p_{01}p_{00}$ $\mu_{00/1}$	$2p_{10}p_{01}^3$ $\mu_{00/1}$	$4p_{10}p_{01}^2p_{00}$ $\mu_{00/1}$	$2p_{10}p_{01}p_{00}^2$ $\mu_{00/1}$
$Aabb$		$\bar{R}_R\bar{R}_R$	$2p_{11}^2p_{10}p_{00}$ $\mu_{02/1}$	$4p_{11}p_{10}^2p_{00}$ $\mu_{01/1}$	$2p_{10}^3p_{00}$ $\mu_{00/1}$	$4p_{11}p_{10}p_{01}p_{00}$ $\mu_{01/1}$	$4p_{11}p_{10}p_{00}^2$ $\mu_{01/1}$	$4p_{10}^2p_{01}p_{00}$ $\mu_{00/1}$	$4p_{10}^2p_{00}^2$ $\mu_{00/1}$	$2p_{10}p_{01}^2p_{00}$ $\mu_{00/1}$	$4p_{10}p_{01}p_{00}^2$ $\mu_{00/1}$	$2p_{10}p_{00}^2$ $\mu_{00/1}$
$aaBB$		$\bar{R}_R\bar{R}_R$	$p_{11}^2p_{01}^2$ $\mu_{02/1}$	$2p_{11}p_{10}p_{01}^2$ $\mu_{01/1}$	$p_{10}^2p_{01}^2$ $\mu_{00/1}$	$2p_{11}p_{01}^3$ $\mu_{01/1}$	$2p_{11}p_{01}^2p_{00}$ $\mu_{01/1}$	$2p_{10}p_{01}^3$ $\mu_{00/1}$	$2p_{10}p_{01}^2p_{00}$ $\mu_{00/1}$	p_{01}^4 $\mu_{00/1}$	$2p_{01}^3p_{00}$ $\mu_{00/1}$	$p_{01}^2p_{00}^2$ $\mu_{00/1}$
$aaBb$		$\bar{R}_R\bar{R}_R$	$2p_{11}^2p_{10}p_{00}$ $\mu_{02/1}$	$4p_{11}p_{10}^2p_{00}$ $\mu_{01/1}$	$2p_{10}^3p_{00}$ $\mu_{00/1}$	$4p_{11}p_{10}p_{01}p_{00}$ $\mu_{01/1}$	$4p_{11}p_{10}p_{00}^2$ $\mu_{01/1}$	$4p_{10}^2p_{01}p_{00}$ $\mu_{00/1}$	$4p_{10}^2p_{00}^2$ $\mu_{00/1}$	$2p_{10}p_{01}^2p_{00}$ $\mu_{00/1}$	$4p_{10}p_{01}p_{00}^2$ $\mu_{00/1}$	$2p_{10}p_{00}^2$ $\mu_{00/1}$
$aabb$		$\bar{R}_R\bar{R}_R$	$p_{11}^2p_{00}^2$ $\mu_{02/1}$	$2p_{11}p_{10}p_{00}^2$ $\mu_{01/1}$	$p_{10}^2p_{00}^2$ $\mu_{00/1}$	$2p_{11}p_{01}p_{00}^2$ $\mu_{01/1}$	$2p_{11}p_{00}^3$ $\mu_{01/1}$	$2p_{10}p_{01}p_{00}^2$ $\mu_{00/1}$	$2p_{10}p_{00}^3$ $\mu_{00/1}$	$p_{01}^2p_{00}^2$ $\mu_{00/1}$	$2p_{01}p_{00}^3$ $\mu_{00/1}$	p_{00}^4 $\mu_{00/1}$

2.2 Haplotype Effects

We will model the effects of haplotypes, constructed by alleles at different SNPs on the same chromosome, on the phenotypic trait of viral infection. Among the four haplotypes, one is assumed to be the risk haplotype, denoted by R_R for the recipients, R_T for the transmitters, and R_V for the virus, with the rest as the non-risk haplotype, denoted by r_R for the recipients, r_T for the transmitters, and r_V for the virus. This will form three composite diplotypes R_RR_R, R_Rr_R, and r_Rr_R for the recipients and R_TR_T, R_Tr_T, and r_Tr_T for the transmitters. We use l_rl_t/l_v (l_r = 2 for R_RR_R, 1 for R_Rr_R, 0 for r_Rr_R, l_t = 2 for R_TR_T, 1 for R_Tr_T, 0 for r_Tr_T and l_v = 1 for R_V and 0 for r_V) to denote a composite diplotype constructed by the recipient, transmitter, and virus. According to the quantitative genetic principle [19], we partition the genotypic value of composite diplotype into different additive, dominant, and epistatic components as follows:

$$
\begin{aligned}
\mu_{22/1} &= \mu + a_r + a_t + i_{a_ra_t} + a_v + i_{a_ra_v} + i_{a_ta_v} + i_{a_ra_ta_v} \\
\mu_{21/1} &= \mu + a_r + d_t + i_{a_rd_t} + a_v + i_{a_ra_v} + i_{d_ta_v} + i_{a_rd_ta_v} \\
\mu_{20/1} &= \mu + a_r - a_t - i_{a_ra_t} + a_v + i_{a_ra_v} - i_{a_ta_v} - i_{a_ra_ta_v} \\
\mu_{12/1} &= \mu + d_r + a_t + i_{d_ra_t} + a_v + i_{d_ra_v} + i_{a_ta_v} + i_{d_ra_ta_v} \\
\mu_{11/1} &= \mu + d_r + d_t + i_{d_rd_t} + a_v + i_{d_ra_v} + i_{d_ta_v} + i_{d_rd_ta_v} \\
\mu_{10/1} &= \mu + d_r - a_t - i_{d_ra_t} + a_v + i_{d_ra_v} - i_{a_ta_v} - i_{d_ra_ta_v} \\
\mu_{02/1} &= \mu - a_r + a_t - i_{a_ra_t} + a_v - i_{a_ra_v} + i_{a_ta_v} - i_{a_ra_ta_v} \\
\mu_{01/1} &= \mu - a_r + d_t - i_{a_rd_t} + a_v - i_{a_ra_v} + i_{d_ta_v} - i_{a_rd_ta_v} \\
\mu_{00/1} &= \mu - a_r - a_t + i_{a_ra_t} + a_v - i_{a_ra_v} - i_{a_ta_v} + i_{a_ra_ta_v} \\
\mu_{22/0} &= \mu + a_r + a_t + i_{a_ra_t} - a_v - i_{a_ra_v} - i_{a_ta_v} - i_{a_ra_ta_v} \\
\mu_{21/0} &= \mu + a_r + d_t + i_{a_rd_t} - a_v - i_{a_ra_v} - i_{d_ta_v} - i_{a_rd_ta_v} \\
\mu_{20/0} &= \mu + a_r - a_t - i_{a_ra_t} - a_v - i_{a_ra_v} + i_{a_ta_v} + i_{a_ra_ta_v} \\
\mu_{12/0} &= \mu + d_r + a_t + i_{d_ra_t} - a_v - i_{d_ra_v} - i_{a_ta_v} - i_{d_ra_ta_v} \\
\mu_{11/0} &= \mu + d_r + d_t + i_{d_rd_t} - a_v - i_{d_ra_v} - i_{d_ta_v} - i_{d_rd_ta_v} \\
\mu_{10/0} &= \mu + d_r - a_t - i_{d_ra_t} - a_v - i_{d_ra_v} + i_{a_ta_v} + i_{d_ra_ta_v} \\
\mu_{02/0} &= \mu - a_r + a_t - i_{a_ra_t} - a_v + i_{a_ra_v} - i_{a_ta_v} + i_{a_ra_ta_v} \\
\mu_{01/0} &= \mu - a_r + d_t - i_{a_rd_t} - a_v + i_{a_ra_v} - i_{d_ta_v} + i_{a_rd_ta_v} \\
\mu_{00/0} &= \mu - a_r - a_t + i_{a_ra_t} - a_v + i_{a_ra_v} + i_{a_ta_v} - i_{a_ra_ta_v},
\end{aligned}
\tag{3}
$$

where

(1) μ is the overall mean,

(2) a_r, a_t, and a_v are the additive genetic effects of risk haplotypes expressed in the recipients, transmitters, and virus,

(3) d_r and d_t are the dominance genetic effects due to interactions between risk haplotypes and non-risk haplotypes in the recipients and transmitters,

(4) $i_{a_r a_t}$, $i_{a_r d_t}$, $i_{d_r a_t}$, and $i_{d_r d_t}$ are the additive × additive, additive × dominance, dominance × additive, and dominance × dominance epistatic effects between the recipients and transmitters,

(5) $i_{a_r a_v}$ and $i_{d_r a_v}$ are the additive × additive and dominance × additive epistatic effects between the recipients and virus,

(6) $i_{a_t a_v}$ and $i_{d_t a_v}$ are the additive × additive and dominance × additive epistatic effects between the transmitters and virus,

(7) $i_{a_r a_t a_v}$, $i_{a_r d_t a_v}$, $i_{d_r a_t a_v}$, and $i_{d_r d_t a_v}$ are the additive × additive × additive, additive × dominance × additive, dominance × additive × additive, and dominance × dominance × additive epistatic effects among the recipients, transmitters, and virus, respectively.

All these genetic effect parameters are arrayed in $\mathbf{\Omega}_E$.

Different from traditional quantitative genetic models, equation (3) includes epistatic interactions between genes from different individuals. Individual-individual interactions have been thought to contribute to the phenotypic variation of a trait through indirect ways. The model proposed in this article will provide a procedure for estimating these interaction effects and testing their roles in the genetic control of complex traits.

2.3 Likelihood and Estimation

To simplify the description of the model being developed, we will ignore covariate effects in the model although it is straightforward to incorporate these effects. Thus, the phenotypic value of viral infection for recipient i is expressed as the sum of genetic effects and residual errors, i.e.,

$$y_i = \sum_{j_r=2}^{0} \sum_{j_t=2}^{0} \sum_{j_v=2}^{0} \xi_i \mu_{l_r l_t / l_v} + e_i, \tag{4}$$

where ξ_i is the indicator variable defined as 1 if recipient i with composite diplotype l_r is given a virus of risk haplotype l_v by a transmitter of composite diplotype l_t, and 0 otherwise, $\mu_{l_r l_t / l_v}$ is the genotypic value explained in equation (3), and e_i is the residual error which is independently and identically distributed with mean 0 and variance σ^2. Because viral loads tend to have a skewed distribution, the log-transformation of the data may be more suitable

DOI: 10.2202/1544-6115.1475

for this model which assumes normality of residual error. To preserve the biological feature of parameters, a transforms-both-sides approach as used by Wu et al. [20] may be used.

The joint likelihood of parameters given marker data for the recipients ($\mathbf{M}_R$), transmitters ($\mathbf{M}_T$), and virus ($\mathbf{M}_V$) and phenotypic data for the recipients (y) is constructed as

$$\begin{aligned} & L(\mathbf{\Omega}_H, \mathbf{\Omega}_V, \mathbf{\Omega}_E, \sigma^2 | \mathbf{M}_R, \mathbf{M}_T, \mathbf{M}_V, y) \\ = \; & L(\mathbf{\Omega}_H, \mathbf{\Omega}_V | \mathbf{M}_R, \mathbf{M}_T, \mathbf{M}_V) + L(\mathbf{\Omega}_E, \sigma^2 | \mathbf{M}_R, \mathbf{M}_T, \mathbf{M}_V, y; \mathbf{\Omega}_H, \mathbf{\Omega}_V) \end{aligned} \quad (5)$$

where the first term is the likelihood of haplotype frequencies for the hosts and virus and the second term is the likelihood of haplotype effects and variation constructed on the haplotype frequencies. To maximize the likelihood (5), we can maximize its two terms individually.

The estimates of haplotype frequencies are obtained via maximizing $L(\mathbf{\Omega}_H, \mathbf{\Omega}_V | \mathbf{M}_R, \mathbf{M}_T, \mathbf{M}_V)$ from which virus haplotype frequencies can be estimated directly using

$$q_{j_v} = \frac{1}{N} \sum_{j_r=1}^{9} \sum_{j_t=1}^{9} N_{j_t j_v / j_v}, \quad (6)$$

where $N = \sum_{j_r=1}^{9} \sum_{j_t=1}^{9} \sum_{j_v=1}^{4} N_{j_t j_v / j_v}$, and host haplotype frequencies are estimated by implementing the EM algorithm. In the E step, the proportion of a diplotype in a double heterozygote is calculated using

$$\pi = \frac{p_{11} p_{00}}{p_{11} p_{00} + p_{10} p_{00}}. \quad (7)$$

In the M step, the calculated proportion is used to estimate host haplotype frequencies by

$$\begin{aligned} p_{11} = \; & \frac{1}{4N} \sum_{j_v=1}^{4} \Big[4N_{11/j_v} + 3(N_{12/j_v} + N_{21/j_v} + N_{14/j_v} + N_{41/j_v}) + \\ & + 2(N_{22/j_v} + N_{44/j_v} N_{13/j_v} + N_{31/j_v} + N_{16/j_v} + N_{61/j_v} + N_{17/j_v} \\ & + N_{71/j_v} + N_{18/j_v} + N_{81/j_v} + N_{19/j_v} N_{91/j_v} + N_{24/j_v} + N_{42/j_v}) \\ & + N_{23/j_v} + N_{32/j_v} + N_{26/j_v} + N_{62/j_v} + N_{27/j_v} + N_{72/j_v} + N_{28/j_v} \\ & + N_{82/j_v} + N_{29/j_v} + N_{92/j_v} + N_{34/j_v} + N_{43/j_v} + N_{46/j_v} + N_{64/j_v} \\ & + N_{47/j_v} + N_{74/j_v} + N_{48/j_v} + N_{84/j_v} + N_{49/j_v} + N_{94/j_v} \\ & + \pi(3N_{15/j_v} + 3N_{51/j_v} + 2N_{25/j_v} + 2N_{52/j_v} + N_{35/j_v} + N_{53/j_v} \\ & + 2N_{45/j_v} + 2N_{54/j_v} + N_{56/j_v} + N_{65/j_v} + N_{57/j_v} + N_{75/j_v} \\ & + N_{85/j_v} + N_{58/j_v} + N_{59/j_v} + N_{95/j_v}) + (1 - \pi)\big(2N_{15/j_v} \\ & + 2N_{51/j_v} + N_{25/j_v} + N_{52/j_v} + N_{45/j_v} + N_{54/j_v}\big) + 2\pi N_{55/j_v} \Big], \end{aligned} \quad (8)$$

$$
\begin{aligned}
p_{10} \;=\; & \frac{1}{4n}\sum_{k=0}^{1}\Big[4N_{33/j_v}+3(N_{23/j_v}+N_{32/j_v}+N_{36/j_v}+N_{63/j_v}) \\
& +2(N_{22/j_v}+N_{66/j_v}+N_{13/j_v}+N_{31/j_v}+N_{26/j_v}+N_{62/j_v}+N_{34/j_v} \\
& +N_{43/j_v}+N_{37/j_v}+N_{73/j_v}+N_{38/j_v}+N_{83/j_v}+N_{39/j_v}+N_{93/j_v}) \\
& +N_{12/j_v}+N_{21/j_v}+N_{16/j_v}+N_{61/j_v}+N_{24/j_v}+N_{42/j_v}+N_{27/j_v} \\
& +N_{72/j_v}+N_{28/j_v}+N_{82/j_v}+N_{29/j_v}+N_{92/j_v}+N_{46/j_v}+N_{64/j_v} \\
& +N_{67/j_v}+N_{76/j_v}+N_{68/j_v}+N_{86/j_v}+N_{69/j_v}+N_{96/j_v} \\
& +(1-\pi)\big(N_{15/j_v}+N_{51/j_v}+2N_{25/j_v}+2N_{52/j_v}+3N_{35/j_v}+3N_{53/j_v} \\
& +N_{45/j_v}+N_{54/j_v}+2N_{56/j_v}+2N_{65/j_v}+N_{57/j_v}+N_{75/j_v}+N_{85/j_v} \\
& +N_{58/j_v}+N_{59/j_v}+N_{95/j_v}\big)+\pi\big(N_{25/j_v}+N_{52/j_v}+2N_{35/j_v} \\
& +2N_{53/j_v}+N_{56/j_v}+N_{65/j_v}\big)+2(1-\pi)N_{55/j_v}\Big],
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
p_{01} \;=\; & \frac{1}{4N}\sum_{k=0}^{1}\Big[4N_{77/j_v}+3(N_{47/j_v}+N_{74/j_v}+N_{78/j_v}+N_{87/j_v}) \\
& +2\big(N_{44/j_v}+N_{88/j_v}+N_{17/j_v}+N_{71/j_v}+N_{27/j_v}+N_{72/j_v}+N_{37/j_v} \\
& +N_{73/j_v}+N_{67/j_v}+N_{76/j_v}+N_{48/j_v}+N_{84/j_v}+N_{79/j_v}+N_{97/j_v}\big) \\
& +N_{14/j_v}+N_{41/j_v}+N_{24/j_v}+N_{42/j_v}+N_{34/j_v}+N_{43/j_v}+N_{46/j_v} \\
& +N_{64/j_v}+N_{18/j_v}+N_{81/j_v}+N_{28/j_v}+N_{82/j_v}+N_{38/j_v}+N_{83/j_v} \\
& +N_{68/j_v}+N_{86/j_v}+N_{49/j_v}+N_{94/j_v}+N_{89/j_v}+N_{98/j_v} \\
& +(1-\pi)\big(N_{15/j_v}+N_{51/j_v}+N_{25/j_v}+N_{52/j_v}+N_{35/j_v}+N_{53/j_v} \\
& +2N_{45/j_v}+2N_{54/j_v}+N_{56/j_v}+N_{65/j_v}+3N_{57/j_v}+3N_{75/j_v} \\
& +2N_{85/j_v}+2N_{58/j_v}+N_{59/j_v}+N_{95/j_v}\big)+\pi\big(N_{45/j_v}+N_{54/j_v} \\
& +2N_{57/j_v}+2N_{75/j_v}+N_{58/j_v}+N_{85/j_v}\big)+2(1-\pi)N_{55/j_v}\Big],
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
p_{00} \;=\; & \frac{1}{4N}\sum_{k=0}^{1}\Big[4N_{99/j_v}+3(N_{69/j_v}+N_{96/j_v}+N_{89/j_v}+N_{98/j_v}) \\
& +2\big(N_{66/j_v}+N_{88/j_v}+N_{19/j_v}+N_{91/j_v}+N_{29/j_v}+N_{92/j_v}+N_{39/j_v} \\
& +N_{93/j_v}+N_{49/j_v}+N_{94/j_v}+N_{79/j_v}+N_{97/j_v}+N_{69/j_v}+N_{96k}\big) \\
& +N_{16/j_v}+N_{61/j_v}+N_{26/j_v}+N_{62/j_v}+N_{36/j_v}+N_{63/j_v}+N_{46/j_v} \\
& N_{64/j_v}+N_{67/j_v}+N_{76/j_v}+N_{18/j_v}+N_{81/j_v}+N_{28/j_v}+N_{82/j_v} \\
& +N_{38/j_v}+N_{83}+N_{48/j_v}+N_{84/j_v}+N_{78/j_v}+N_{87/j_v} \\
& +\pi\big(N_{15/j_v}+N_{51/j_v}+N_{25/j_v}+N_{52/j_v}+N_{35/j_v}+N_{53/j_v}+N_{45/j_v} \\
& +N_{54/j_v}+2N_{56/j_v}+2N_{65/j_v}+N_{57/j_v}+N_{75/j_v}+2N_{85/j_v}+2N_{58/j_v} \\
& +3N_{59/j_v}+3N_{95/j_v}\big)+(1-\pi)\big(N_{65/j_v}+N_{56/j_v}+N_{85/j_v}+N_{58/j_v} \\
& +2N_{95/j_v}+2N_{59/j_v}\big)+2\pi N_{55/j_v}\Big].
\end{aligned}
\tag{11}
$$

DOI: 10.2202/1544-6115.1475

The E and M steps are iterated between equations (7) and (8)–(11) until the estimates are stable. The estimated haplotype frequencies can be used to solve the linkage disequilibrium between two SNPs in the host and virus using equations (1) and (2).

The construction of likelihood $L(\mathbf{\Omega}_E, \sigma^2|\mathbf{M}_R, \mathbf{M}_T, \mathbf{M}_V, y; \mathbf{\Omega}_H, \mathbf{\Omega}_V)$ needs the assumption of a risk haplotype (and therefore composite diplotypes). Table 1 shows the distribution of composite diplotypes by assuming that AB is the risk haplotype for both the recipients and transmitters and that CD is the risk haplotype for the virus. Based on this information, we construct the mixture model-based likelihood as follows:

$$
\begin{aligned}
& L(\mathbf{\Omega}_E, \sigma^2|\mathbf{M}_R, \mathbf{M}_T, \mathbf{M}_V, y; \mathbf{\Omega}_H, \mathbf{\Omega}_V) \\
= & \left\{ \prod_{i=1}^{n_{11/1}} f_{22/1}(y_i) \prod_{i=1}^{n_{12/1}} f_{21/1}(y_i) \cdots \prod_{i=1}^{n_{15/1}} \Big[\pi f_{21/1}(y_i) + (1-\pi) f_{20/1}(y_i)\Big] \cdots \right. \\
\times & \prod_{i=1}^{n_{55/1}} \Big[\pi^2 f_{11/1}(y_i) + \pi(1-\pi) f_{10/1}(y_i) + (1-\pi)\pi f_{01/1}(y_i) \\
& \qquad +(1-\pi)^2 f_{00/1}(y_i)\Big] \cdots \\
\times & \left. \prod_{i=1}^{n_{95/1}} [\pi f_{01/1}(y_i) + (1-\pi) f_{00/1}(y_i)] \cdots \prod_{i=1}^{n_{99/1}} f_{00/1}(y_i) \right\} \\
\times & \prod_{j_v=2}^{4} \left\{ \prod_{i=1}^{n_{11/j_v}} f_{22/0}(y_i) \prod_{i=1}^{n_{12/j_v}} f_{21/0}(y_i) \cdots \prod_{i=1}^{n_{15/j_v}} [\pi f_{21/0}(y_i) + (1-\pi) f_{20/0}(y_i)] \cdots \right. \\
\times & \prod_{i=1}^{n_{55/j_v}} [\pi^2 f_{11/0}(y_i) + \pi(1-\pi) f_{10/0}(y_i) + (1-\pi)\pi f_{01/0}(y_i) \\
& \qquad +(1-\pi)^2 f_{00/0}(y_i)] \cdots \\
\times & \left. \prod_{i=1}^{n_{95/j_v}} [\pi f_{01/0}(y_i) + (1-\pi) f_{00/0}(y_i)] \cdots \prod_{i=1}^{n_{99/j_v}} f_{00/0}(y_i) \right\}, \qquad (12)
\end{aligned}
$$

where $f_{l_r l_t/l_v}(y_i)$ is assumed to be a normal distribution with mean $\mu_{l_r l_t/l_v}$ and variance σ^2.

The EM algorithm is implemented to solve the likelihood (12). In the E step, the posterior probability with which recipient i carries a particular

composite diplotype is calculated using the following formulas:

$$\begin{aligned}
\Pi_{21/1|i} &= \frac{\pi f_{21/1}(y_i)}{\pi f_{21/1}(y_i) + (1-\pi) f_{20/1}(y_i)}, \\
\Pi_{11/1|i} &= \frac{\pi f_{11/1}(y_i)}{\pi f_{11/1}(y_i) + (1-\pi) f_{10/1}(y_i)}, \\
\Pi_{01/1|i} &= \frac{\pi f_{01/1}(y_i)}{\pi f_{01/1}(y_i) + (1-\pi) f_{00/1}(y_i)}, \\
&\cdots \\
\Phi_{1_i} &= \frac{\pi^2 f_{11}(y_i)}{\pi^2 f_{11}(y_i) + \pi(1-\pi) f_{10}(y_i) + (1-\pi)\pi f_{01}(y_i) + (1-\pi)^2 f_{00}(y_i)}, \\
\Phi_{2_i} &= \frac{\pi(1-\pi) f_{10}(y_i)}{\pi^2 f_{11}(y_i) + \pi(1-\pi) f_{10}(y_i) + (1-\pi)\pi f_{01}(y_i) + (1-\pi)^2 f_{00}(y_i)}, \\
\Phi_{3_i} &= \frac{(1-\pi)\pi f_{01}(y_i)}{\pi^2 f_{11}(y_i) + \pi(1-\pi) f_{10}(y_i) + (1-\pi)\pi f_{01}(y_i) + (1-\pi)^2 f_{00}(y_i)}, \\
\Phi_{4_i} &= \frac{(1-\pi)^2 f_{00}(y_i)}{\pi^2 f_{11}(y_i) + \pi(1-\pi) f_{10}(y_i) + (1-\pi)\pi f_{01}(y_i) + (1-\pi)^2 f_{00}(y_i)}.
\end{aligned} \tag{13}$$

In the M step, the genotypic values of each composite diplotype and residual variance are estimated with the calculated posterior probabilities using the log-likelihood equations as follows:

$$\begin{aligned}
\hat{\mu}_{22/1} &= \frac{\sum_{i=1}^{N_{11/1}} y_i}{N_{11/1}}, \\
\hat{\mu}_{21/1} &= \frac{\sum_{i=1}^{N_{12/1}} y_i + \sum_{i=1}^{N_{14/1}} y_i + \sum_{i=1}^{N_{15/1}} \Pi_{21/1|i} y_i}{N_{12/1} + N_{14/1} + \sum_{i=1}^{N_{15/1}} \Pi_{21/1|i}}, \\
&\vdots \\
\hat{\mu}_{00/0} &= \frac{\sum_{i=1}^{\dot{N}} y_i + \sum_{i=1}^{\ddot{N}} \Pi_{21/1|i} y_i + \sum_{i=1}^{N_{55/0}} \Phi_{4_i} y_i}{\dot{N} + \sum_{i=1}^{\ddot{N}} \Pi_{21/1|i} + \sum_{i=1}^{N_{55/0}} (1 - \Pi_{21/1|i})^2}, \\
\hat{\sigma}^2 &= \frac{1}{N}\Big[\sum_{i=1}^{N_{11/1}} (y_i - \mu_{22/1})^2 + \sum_{i=1}^{N_{12/1}} (y_i - \mu_{21/1})^2 + \sum_{i=1}^{N_{14/1}} (y_i - \mu_{21/1})^2 \\
&\quad + \sum_{i=1}^{N_{15/1}} \Pi_{21/1|i} (y_i - \mu_{21/1})^2 + \cdots + \sum_{i=1}^{\dot{N}} (y_i - \mu_{00/0})^2 \\
&\quad + \sum_{i=1}^{\ddot{N}} \Pi_{21/1|i} (y_i - \mu_{00/0})^2 + \sum_{i=1}^{N_{55/0}} \Phi_{4_i} (y_i - \mu_{00/0})^2 \Big],
\end{aligned} \tag{14}$$

DOI: 10.2202/1544-6115.1475

where $\dot{N} = N_{33/0} + N_{36/0} + N_{63/0} + N_{37/0} + N_{73/0} + N_{38/0} + N_{83/0} + N_{39/0} + N_{93/0} + \sum_{i=6}^{9}\sum_{j=6}^{9} N_{ij/0}$, $\ddot{N} = N_{35/0} + N_{53/0} + N_{56/0} + N_{65/0} + N_{57/0} + N_{75/0} + N_{58/0} + N_{85/0} + N_{59/0} + N_{95/0}$.

The E and M steps are iterated between equations (13) and (14) until stable estimates are obtained. The additive and dominance effects of risk haplotypes in the recipients, transmitters, and virus, as well as the epistatic effects of different kinds between these three different genomes, can be estimated with equation (3).

In practice, we do not know real risk haplotypes for viral loads. The most likely risk haplotypes for the recipients, transmitters, and virus are determined from the likelihoods estimated for all possible ($4 \times 4 \times 2 = 32$) combinations of risk haplotypes. The largest likelihood corresponds to the optimal combination of risk haplotypes that fit the data.

2.4 Hypothesis Tests

The model proposed can formulate a number of hypotheses about the detailed genetic control mechanisms of viral loads. The first hypothesis is about the significance of risk haplotypes derived from the three different genomes, recipients, transmitters, and virus. The null hypothesis for this test is

$$H_0 : \mu_{l_r l_t / l_v} \equiv \mu, \text{ for } l_r, l_t = 2, 1, 0; l_v = 1, 0 \tag{15}$$

The likelihood ratio (LR) for the null and alternative hypotheses is calculated and compared with the critical threshold determined from permutation tests [21]. By reshuffling the phenotypic values among subjects, a new data set is generated for which the LR value is calculated. This procedure is repeated 1000 times, obtaining the distribution of LR values under the null hypothesis (15). Thus, by comparing the LR value calculated from real data with this LR distribution, the empirical p-value can be determined. If significant risk haplotypes are found to exist, the next is to test how these risk haplotypes trigger an effect on viral loads. This can be done by testing individual genetic effects in the following sequence, three-genome interactions, two-genome interactions, and main effects (including the additive and dominance).

Although studies of three-way interactions have not received adequate attention, their role in the genetic control of complex diseases may be dramatic. The model developed allows the identification of three-way epistasis among the recipients, transmitters, and viruses. The null hypothesis for doing this can be formulated as

$$H_0: \quad i_{a_r a_t a_v} = i_{a_r d_t a_v} = i_{d_r a_t a_v} = i_{d_r d_t a_v} = 0. \tag{16}$$

Whether the risk haplotypes from the recipients interact with those from the transmitters can be tested using

$$H_0: \quad i_{a_r a_t} = i_{a_r d_t} = i_{d_r a_t} = i_{d_r d_t} = 0. \tag{17}$$

If each of these epistatic interactions between different genomes is tested, the model allows the characterization of specific genetic control mechanisms for viral infection. For example, among four kinds of epistatic interactions, additive × additive, additive × dominance, dominance × additive, and dominance × dominance, which one is the most important in affecting viral infection. Similarly, epistatic interactions between the recipients or transmitters and viruses can also be tested, respectively, using

$$H_0: \quad i_{a_r a_v} = i_{d_r v_v} = 0, \tag{18}$$
$$H_0: \quad i_{a_t a_v} = i_{d_t v_v} = 0. \tag{19}$$

The influences of risk haplotype derived from the recipients, transmitters, and virus can be tested, respectively, using

$$H_0: \quad a_r = d_r = 0, \tag{20}$$
$$H_0: \quad a_t = d_t = 0, \tag{21}$$
$$H_0: \quad a_v = 0. \tag{22}$$

The additive and dominance effects due to risk haplotypes of the recipients and transmitters can be tested individually.

Because the null hypotheses (16)–(22) are nested within their alternative, the likelihood ratios calculated can be thought to asymptotically follow a χ-distribution with the degree of freedom equal to the difference in the number of parameters between the alternative and null hypotheses. Also, the estimates of genotypic values for each composite diplotype under the null hypotheses can be obtained with a similar EM procedure derived under the alternative hypothesis, although a series of constraints that define relationships between specific genotypic values need to be incorporated.

DOI: 10.2202/1544-6115.1475

3 Monte Carlo Simulation

Simulation studies with the aid of computer were performed to investigate the statistical properties of the model proposed. A group of patients from a natural population were simulated as recipients and the people who transmit virus into these recipients are assumed to be known. We simulated marker data for the recipients, transmitters, and virus at a candidate gene for viral infection using given haplotype frequencies (calculated from allele frequencies at different SNPs and their linkage disequilibria in the hosts and virus; Table 2). Meanwhile, the phenotypic data of viral loads in the recipients were simulated by assuming that they follow a normal distribution. The simulation was conducted under several scenarios including different sample sizes of recipients (200 to 2000) and different heritabilities (0.1 and 0.4). The true additive, dominance and epistatic effects of different kinds and orders in the recipients, transmitters, and virus are given in Tables 3 and 4, which were used to simulate the phenotypic data under different heritabilities.

The model was used to analyze the simulated data, showing reasonably good estimates of all population and quantitative genetic parameters. Not surprisingly, haplotype frequencies in the hosts can be very well estimated even with a modest sample size (200), partly because we derived a closed form for the EM algorithm within the mixture model framework (Table 2). When sample sizes increases to 2000, a small linkage disequilibrium with a value being its lower bound can be precisely estimated (results not shown). Although our simulation considers highly heterozygous markers (for which the frequencies of two alternative alleles are similar), haplotype frequencies and linkage disequilibria for less heterozygous markers (for which the frequencies of two alternative alleles are very different) can also be reasonably estimated with a modest sample size (200) (results not shown). The estimates of haplotype frequencies in the virus are not shown since these estimates are obtained directly from an analytical expression.

Tables 3 and 4 give the results of estimation of haplotype effect parameters using the model. First, the additive and additive $\times$ additive epistatic effects can be very well estimated for the recipients, transmitters, and virus because these parameters hold a linear relationship with the number of risk haplotypes. As an interaction parameter between the risk haplotype and non-risk haplotype, the dominance effects are more difficult to estimate as compared to the additive and additive $\times$ additive epistatic effects. The estimates of the epistatic effects involving the dominance are also less precise; this is particularly for the dominance $\times$ dominance epistasis. Second, epistatic effects of higher orders among the recipients, transmitters, and virus show

similar estimation precision compared to those of lower order because only the additive effect is included for the virus gene. This results suggests that it is methodologically feasible to include the estimates of higher-order epistasis given its possible important role in disease infection. Third, sample size and heritability are two important factors for the precision of parameter estimation. For the additive and additive $\times$ additive epistatic effects, a modest sample size (200) and heritability level (0.1) would be sufficient for their precise estimation. Under such a modest heritability, the estimation of dominance effects needs a sample size of 400 or more, whereas a sample size of 800 or more is required for the estimation of dominance $\times$ dominance epistatic effects. When heritability increases to 0.4, a much smaller sample size is sufficient for parameter estimation.

We also examined the power of the model to detect epistatic effects (Table 4). It is found that the model has great power to detect epistatic effects between genes from different genomes even with a modest sample size and under a modest heritability level. Because risk haplotypes are not known for a practical data set, we tested the model's power to correctly identify risk haplotypes for the hosts and virus. One can always correctly detect the risk haplotype of the virus since no missing data exist. Our power analysis focuses on the detection of risk haplotypes for the hosts (Table 5). When heritability is 0.1, the sample size of 200 will have about 70% to correctly detect risk haplotypes. The power will increase to 99% if heritability is 0.4, or 0.86 if

Table 2: The means (upper) and standard errors (lower) of maximum likelihood estimates of haplotype frequencies for a pair of SNPs from the hosts obtained from 100 simulation replicates under different sample sizes.

n	p_a	p_b	D_H
200	0.601	0.598	0.079
	0.0018	0.0017	0.0011
400	0.601	0.601	0.080
	0.0012	0.0013	0.0007
800	0.601	0.600	0.0789
	0.0009	0.0009	0.0005
2000	0.600	0.599	0.080
	0.0005	0.0006	0.0003

DOI: 10.2202/1544-6115.1475

sample size is doubled. From the above precision and power analysis, we recommend that a sample size of 400 is required when the heritability of a viral infection trait is high (0.4), whereas at least 800 patients are needed if the trait has a modest heritability (0.1). Under each simulation scheme, type I error rates were calculated by analyzing the data simulated under the null hypothesis. Type I error rates are in a range of 0 to 0.10, suggesting that the size of the test is preserved.

4 Discussion

We develop a new statistical model for mapping and identify genes for viral infection at the DNA sequence level. The model will find its immediate application to disease gene discovery given the increasing availability of single nucleotide polymorphisms (SNPs) stemming from a rapid development of high-throughput genotyping techniques. The model is particularly appealing to a genome-wide association study (GWAS), in which half a million or more SNPs are typed [22], aimed to uncover all of the DNA sequence variants that affect an individual's risk of disease. Different from existing mapping approaches, this model is characterized by several features. First, it integrates the epidemiological behavior of infectious disease into a statistical model for genetic mapping, thus allowing the test of genetic control of the spread of viral infection. According to mathematical analyses by Meyers and colleagues, the contact patterns of disease transmission in a community are central to understanding the outbreak or epidemics of an infectious disease [23, 24]. Thus, our model incorporating the genetic influence of contact people shall be more useful and relevant for mapping disease genes.

Second, the new model constructs a general framework for estimating and testing epistatic interactions between genes from different genomes including recipients, transmitters, and viruses. It has been recognized that genome-genome or individual-individual interactions play an important role in increasing genetic diversity and variation [25] and therefore organisms' adaptation to changing environments [26]. The model will provide a procedure for testing the effects of genome-genome interactions on viral dynamics and evolution. Also, the model allows the characterization of epistatic interactions of high orders. In the past, higher-order epistasis involving three genes or more has been generally ignored for simplifying analyses, but its importance should be re-appreciated given the complexity of a genetic network for a complex trait [27]. We provide a series of hypothesis tests about the effects of various epistatic interactions, providing an analytical approach for understanding the genetic control mechanisms of disease infection.

Table 3: The means (upper) and standard errors (lower) of maximum likelihood estimates of quantitative genetic parameters for three-way haplotypes derived from the recipients, transmitters, and virus obtained from 100 simulation replicates under different heritabilities (H^2) and sample sizes (N). True values for quantitative genetic parameters are given in parentheses. The second half of the model parameters are presented in Table 4.

H^2	N	μ (10)	a_r (1)	a_t (0.8)	d_r (0.5)	d_t (0.8)	$i_{a_r a_t}$ (0.6)	$i_{a_r d_t}$ (0.5)	$i_{d_r a_t}$ (0.5)	$i_{d_r d_t}$ (0.4)
0.1	200	10.124	0.939	0.671	0.427	0.598	0.826	0.646	0.654	0.534
		0.1011	0.0986	0.0853	0.1427	0.1458	0.1189	0.1348	0.01450	0.1931
	400	10.022	1.042	0.757	0.521	0.858	0.645	0.497	0.700	0.280
		0.0705	0.0689	0.0685	0.1052	0.0998	0.0674	0.0876	0.0967	0.1569
	800	9.960	0.937	0.866	0.574	0.889	0.541	0.514	0.397	0.275
		0.0436	0.0502	0.0421	0.0676	0.0688	0.0433	0.0778	0.0716	0.0930
	2000	10.036	1.022	0.818	0.484	0.761	0.573	0.463	0.473	0.409
		0.0312	0.0277	0.0265	0.0392	0.045	0.0281	0.0433	0.0416	0.0652
0.4	200	10.010	1.001	0.797	0.588	0.773	0.621	0.486	0.523	0.382
		0.0389	0.0408	0.0439	0.0573	0.0584	0.0416	0.0588	0.0669	0.0796
	400	10.033	0.997	0.805	0.483	0.728	0.600	0.475	0.484	0.455
		0.0289	0.0316	0.0312	0.0400	0.0426	0.0259	0.0432	0.0480	0.0602
	800	9.973	1.001	0.788	0.572	0.804	0.601	0.513	0.557	0.365
		0.0190	0.0202	0.0184	0.0247	0.0272	0.0196	0.0298	0.0253	0.0384
	2000	10.007	1.007	0.801	0.500	0.812	0.612	0.514	0.489	0.386
		0.0120	0.0120	0.0111	0.0188	0.0181	0.0118	0.0179	0.0169	0.0264

DOI: 10.2202/1544-6115.1475

Table 4: This is a continuation of Table 3. True values for quantitative genetic parameters are given in parentheses. The power to detect all genetic effects (including the main and across-genome epistatic of all orders) is given.

H^2	N	a_v (1)	$i_{a_ra_v}$ (0.4)	$i_{a_ta_v}$ (0.3)	$i_{d_ra_v}$ (0.15)	$i_{d_ta_v}$ (0.3)	$i_{a_ra_ta_v}$ (0.2)	$i_{a_rd_ta_v}$ (0.12)	$i_{d_ra_ta_v}$ (0.08)	$i_{d_rd_ta_v}$ (0.05)	Power
0.1	200	0.895	0.266	0.321	0.214	0.274	0.125	0.038	0.039	-0.010	0.95
		0.1049	0.1041	0.0987	0.1386	0.1285	0.1046	0.1336	0.1400	0.1991	
	400	0.990	0.359	0.313	0.067	0.293	0.140	0.206	-0.092	0.123	1
		0.0732	0.0669	0.0685	0.1045	0.1043	0.0720	0.0913	0.1085	0.1431	
	800	0.997	0.352	0.200	0.139	0.294	0.208	0.253	0.210	0.071	1
		0.0467	0.0497	0.0508	0.0714	0.0700	0.0479	0.0689	0.0791	0.1046	
	2000	0.998	0.399	0.322	0.118	0.296	0.183	0.129	0.105	0.038	1
		0.0295	0.0312	0.0311	0.0426	0.0469	0.0287	0.0470	0.0437	0.0582	
0.4	200	1.050	0.402	0.336	0.097	0.292	0.120	0.127	0.005	0.069	1
		0.0455	0.0411	0.0461	0.0664	0.0617	0.0406	0.0605	0.0630	0.0894	
	400	1.039	0.432	0.313	0.155	0.236	0.191	0.081	0.060	0.042	1
		0.0290	0.0292	0.0272	0.0392	0.0425	0.0305	0.0433	0.0402	0.0543	
	800	1.012	0.417	0.323	0.110	0.283	0.189	0.069	0.030	0.089	1
		0.0176	0.0171	0.0198	0.0274	0.0265	0.0183	0.0261	0.0287	0.0394	
	2000	1.008	0.399	0.277	0.154	0.293	0.206	0.120	0.099	0.050	1
		0.0128	0.0101	0.0123	0.0182	0.0182	0.0112	0.0178	0.0170	0.0280	

Although the model has a complex structure (see Table 1), parameter estimation is found to be efficient in terms of the choice of initial values and convergence rate because the estimation is based on the closed form solutions for the EM algorithm. We performed computer simulation to examine the precision of parameter estimation and power of the model by considering different sample sizes and heritabilities. In general, the model provides reasonably good estimates of all population and quantitative genetic parameters, although the influences of sample size and heritability are different, depending on the type of parameters. From simulation results, we recommend the use of appropriate sample sizes under different heritability

Table 5: The means of log-likelihood ratios for detecting significant haplotype effects from 100 simulated data sets by assuming that risk haplotype is any one of four possible haplotypes for the host. The true risk haplotype for the simulated data sets is haplotype AB. The power to correctly detect risk haplotype for the host is given.

		Assumed Risk Haplotype				
H^2	N	AB	Ab	aB	ab	Power
0.1	200	**29.43**	16.86	16.23	19.20	0.67
0.4	200	**98.05**	44.32	43.59	55.70	0.99
0.1	400	**46.27**	25.66	24.57	29.11	0.86
0.4	400	**181.15**	80.32	76.46	101.07	1
0.1	800	**80.69**	43.92	39.02	50.96	1
0.4	800	**352.41**	151.54	150.56	195.94	1
0.1	2000	**194.47**	97.84	91.20	117.09	1
0.4	2000	**862.63**	365.85	360.20	481.12	1

levels. If an infection trait has a high heritability (say 0.4), 400 patients will be sufficient to get reasonable parameter estimates and power for all genetic effect parameters including dominance $\times$ dominance epistasis. Yet, this number should increases to 800 or more with a modest heritability (say 0.1).

The model is founded on viral transmission from one person to other through sexual intercourse. This mechanism is thought to be a major cause of HIV transmissions worldwide [11]. The epidemiological mechanisms of virus transmission can be better understood by identifying the events that occur in genital or rectal mucosa during transmission. The model assumes the same extent of sexual transmission for all the transmitters, but new data indicate that the extent depends on the phase of the infection, and is much higher during acute infection [28, 29]. It is crucial to incorporate such information and other mechanisms of viral transmission via indirect contacts into the mapping model, making it clinically more relevant.

DOI: 10.2202/1544-6115.1475

There are a number of issues in both statistics and genetics which can be or need to be addressed. First, we consider a single phenotypic trait, but this is often insufficient to capture the comprehensive picture of genetic control of viral infection. Joint modeling of viral dynamics and human immune response traits, implemented with functional mapping [30, 31, 32, 33], will help to elucidate the dynamic and pleiotropic pattern of genes for infectious diseases. Also, when the model is applied in practice, we will need to incorporate covariate effects due to race, gender, age, body mass, etc. Second, the model considers haplotypes constructed by two SNPs, but the extension to including a large number of SNPs is technically straightforward although the computing requirements are much greater. All these haplotype models should be embedded into a general framework of GWAS. By scanning every set of SNPs across the genome, the genome-wide distribution, organization, and effects of significant haplotypes can be illustrated. But a critical issue of correcting the false-positive rate should be addressed here where corrections are ubiquitous between different pairs of SNPs through linkage disequilibria.

Third, the idea of the model proposed can be used in modeling the genetic interactions for malaria that include the co-evolution of three eukaryotic genomes, parasites, mosquitoes, and humans [34]. A recently launched global network for malaria genomic epidemiology is assembling genome data from these genomes [35], from which new discoveries for the genetic basis of malaria can be made with our model. In any case, a rapidly evolving body of knowledge about host genes and their interactions with viral genes and stochastic epidemiological models used to better describe infectious processes will strengthen our ability to understand the genetic control of infectious diseases. By integrating this knowledge into clinical and public health practice, we will be in a good position to control and prevent infectious disease in ways that provide the greatest benefit and least harm at a reasonable cost.

References

[1] Abel L, Dessein AJ. Genetic epidemiology of infectious diseases in humans: design of population-based studies. *Emerging Infectious Diseases* 1998; **4**:593–603.

[2] McNicholl JM, Cuenco KT. Host genes and infectious diseases HIV, other pathogens, and a public health perspective. *American Journal of Preventive Medicine* 1999; **16**:141–154.

[3] Michael NL. Host genetic influences on HIV-1 pathogenesis. *Current Opinion in Immunology* 1999; **11**:466–474.

[4] Martin MP, Gao X, Lee JH, Nelson GW, Detels R, Goedert JJ, Buchbinder S, Hoots K, Vlahov D, Trowsdale J, Wilson M, O'Brien SJ, Carrington M. Epistasis interaction between $KIR3DS1$ and HLA-B delays the progression to AIDS. *Nature Genetics* 2002; **31**:429-434.

[5] Tang JM, Wilson CM, Meleth S, Myracle A, Lobashevsky E, Mulligan MJ, Douglas SD, Korber B, Vermund SH, Kaslow RA. Host genetic profiles predict virological and immunological control of HIV-1 infection in adolescents. *AIDS* 2002; **16**:2275-2284.

[6] Fortin A, Abel L, Casanova JL, Gros P. Host genetics of mycobacterial diseases in mice and men: Forward genetic studies of BCG-osis and tuberculosis. *Annual Review of Genomics and Human Genetics* 2007; **8**:163–192.

[7] Yang XN, Yang HL, Zhou GQ, Zhao G-P. Infectious disease in the genomic era. *Annual Review of Genomics and Human Genetics* 2008; **9**:21–48.

[8] McNicholl JM, Downer MV, Udhayakumar V, Alper CA, Swerdlow DL. Host-pathogen interactions in emerging and re-emerging infectious diseases: A genomic perspective of tuberculosis, malaria, human immunodeficiency virus infection, hepatitis B, and cholera. *Annual Review of Public Health* 2000 **21**:15–46.

[9] Ameisen JC, Lelievre JD, Pleskoff O. HIV/host interactions: new lessons from the Red Queen's country. *AIDS* 2002; **16**:S25-S31.

[10] Asquith B, Bangharn CRM. How does HTLV-I persist despite a strong cell-mediated immune response? *Trends in Immunology* 2008; **29**:4–11.

[11] Gupta K, Klasse PJ. How do viral and host factors modulate the sexual transmission of HIV? Can transmission be blocked? *PLoS Medicine* 2006; **3**(2):e79.

[12] Scaria V, Hariharan M, Maiti S, Pillai B, Brahmachari SK. Host-virus interaction: a new role for microRNAs. *Retrovirology* 2006; **3**:68 doi:10.1186/1742-4690-3-68.

DOI: 10.2202/1544-6115.1475

[13] Scaria V, Hariharan M, Pillai B, Maiti S, Brahmachari SK. Host-virus genome interactions: macro roles for microRNAs. *Cell Microbiology* 2007; **9**:2784–2794.

[14] Liu T, Johnson JA, Casella G, Wu RL. Sequencing complex diseases with HapMap. *Genetics* 2004; **168**:503–511.

[15] Wu RL, Lin M. *Statistical and Computational Pharmacogenomics*. Chapman & Hall/CRC, London, 2008.

[16] Collins FS, Guyer MS, Charkravarti. A Variations on a theme: cataloging human DNA sequence variation. *Science* 1997; **278**:1580-1581.

[17] Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002; **23**:221-233

[18] Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002; **53**:79-91

[19] Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer; 1998.

[20] Wu, R. L., C.-X. Ma, M. Lin, Z. H. Wang and G. Casella, 2004 Functional mapping of growth quantitative trait loci using a transform-both-sides logistic model. *Biometrics* **60**:729-738.

[21] Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**:963–971.

[22] Welcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**:661–678.

[23] Meyers LA, Newman MEJ, Pourbohloul B. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology* 2006; **240**:400-418.

[24] Meyers LA. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society* 2007; **44**:63–86.

[25] Cui YH, Wu RL. Mapping genome-genome epistasis: A multi-dimensional model. *Bioinformatics* 2005; **21**:2447–2455.

[26] Wolf JB, Hager R. A maternal-offspring coadaptation theory for the evolution of genomic imprinting. *PLoS Biology* 2006; **4**(12):e380.

[27] Imielinski M, Belta C. Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Systems Biology* 2008; **2**:40.

[28] Wawer MJ, Gray RH, Sewankambo NK, Serwadda D, Li X. et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *Journal of Infectious Diseases* 2005; **191**:1403–1409.

[29] Galvin, SR, Cohen MS. The role of sexually transmitted diseases in HIV transmission. *Nature Reviews Microbiology* 2004; **2**:33–42.

[30] Wang ZH, Wu RL. A statistical model for high-resolution mapping of quantitative trait loci determining HIV dynamics. *Statistics in Medicine* 2004; **23**:3033-3051.

[31] Wang ZH, Hou W, Wu RL. A statistical model to analyze quantitative trait locus interactions for HIV dynamics from the virus and human genomes. *Statistics in Medicine* 2005; **25**:495–511.

[32] Wang ZH, Li Y, Wu RL. Joint functional mapping of quantitative trait loci for HIV-1 and CD4+ dynamics. *International Journal of Biostatistics* 2009; (accepted).

[33] Wu S, Yang J, Wu RL. Semiparametric functional mapping of quantitative trait loci governing long-term HIV dynamics. *Bioinformatics* 2007; **23**:i569-i576.

[34] Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics* 2005; **77**:171–190.

[35] The Malaria Genomic Epidemiology Network. A global network for investigating the genomic epidemiology of malaria. *Nature* 2008; **456**:732–737.

DOI: 10.2202/1544-6115.1475