Journal of Computer Science & Systems Biology - Open Access
Research Article JCSB/Vol.2 January-February 2009

Modeling Host-Cancer Genetic Interactions with Multilocus Sequence Data

Yao Li [1] and Rongling Wu [1,2]*

[1]Department of Statistics, University of Florida, Gainesville, FL 32611 USA
[2]Departments of Public Health Sciences and Statistics, Pennsylvania State University, Hershey, PA 17033

*Corresponding author: Rongling Wu, Department of Public Health Sciences, Pennsylvania State College of Medicine, Hershey, PA 17033 USA, E-mail: rwu@hes.hmc.psu.edu

Received January 06, 2009; Accepted February 04, 2009; Published February 05, 2009

Citation: Yao L, Rongling W (2009) Modeling Host-Cancer Genetic Interactions with Multilocus Sequence Data. J Comput Sci Syst Biol 2: 024-043. doi:10.4172/jcsb.1000015

Abstract

Cancer susceptibility may be controlled not only by host genes and mutated genes in cancer cells, but also by the epistatic interactions between genes from the host and cancer genomes. We derive a novel statistical model for cancer gene identification by integrating the gene mutation hypothesis of cancer formation into the mixture-model framework. Within this framework, genetic interactions of DNA sequences (or haplotypes) between host and cancer genes responsible for cancer risk are defined in terms of quantitative genetic principle. Our model was founded on a commonly used genetic association design in which a random sample of patients is drawn from a natural human population. Each patient is typed for single nucleotide polymorphisms (SNPs) on normal and cancer cells and measured for cancer susceptibility. The model is formulated within the maximum likelihood context and implemented with the EM algorithm, allowing the estimation of both population and quantitative genetic parameters. The model provides a general procedure for testing the distribution of haplotypes constructed by SNPs from host and cancer genes and the linkage disequilibria of different orders among the SNPs. The model also formulates a series of testable hypotheses about the effects of host genes, cancer genes, and their interactions on cancer susceptibility. We carried out simulation studies to examine the statistical properties of the model. The implications of this model for cancer gene identification are discussed.

Keywords: Cancer genome; EM algorithm; Haplotype; Host genome; Single nucleotide polymorphism

Introduction

Since the recognition of cancer as a genetic disease, a number of familial cancer genes with high-penetrance mutations, such as oncogenes and the tumor suppressors, have been chromosomally identified, isolated or cloned (Balman et al., 2003; Rand et al., 2008). However, growing evidence shows that most cancer is the result of an intricate interaction of lowpenetrance genetic variants with environmental exposures that humans experience (Brennan, 2002). These low-penetrance cancer genes, each usually with a minor effect and cooperating with others in a complicated web, are difficult to detect and, therefore, their contribution to the risk of cancer development remains unclear. There is a pressing demand on the development of powerful statistical models and computational algorithms for identifying and mapping specific DNA sequence variants that regulate cancer susceptibility.

Human cancer cells frequently possess large-scale chromosomal rearrangements due to chromosomal instability (CIN) (Stock and Bialy 2002; Thompson and Compton, 2008) or gene mutation (Jallepalli and Lengauer, 2001; Greenman et al., 2007). CIN makes whole chromosomes or large frac-

tions of chromosomes gained or lost during cell division, resulting in an imbalance in the number of chromosomes per cell (aneuploidy) and an enhanced rate of loss of heterozygosity. Thus, the "aneuploidy hypothesis of cancer" (Stock and Bialy, 2002) proposes that the main differences between normal and abnormal (cancer) cells result from the number of genes rather than the types of genes differentially expressed, as opposed to the "gene-mutation hypothesis" (Jallepalli and Lengauer, 2001). In general, cancer incidence and development are not only affected by the host genes, but also by genes derived from the cancer cells themselves. Given strong mechanistic interactions between the host and cancer tissues (Araujo and McElwain, 2006), these two different systems of genes operate interactively or epistatically to alter the course of cancer progression. Thus, it can be well anticipated that any statistical model for gene detection that incorporates these two hypotheses are likely to make groundbreaking discoveries of cancer genes.

Genetic mapping has proven to be a powerful approach for detecting quantitative trait loci (QTLs) for complex traits. But a QTL may contain multiple genes that operate in a collective way. It is not possible to study the DNA structure, organization and function of a QTL detected from a mapping approach. A more accurate and useful approach for the characterization of genetic variants contributing to quantitative variation is to directly analyze DNA sequences, known as haplotypes, associated with a particular disease (Liu et al., 2004; Lin and Wu, 2006). If a string of DNA sequence is known to increase disease risk, this risk can be prevented by inhibiting the expression of this string using a specialized drug. The control of this disease can be made more efficient if all possible DNA sequences determining its variation are identified in the entire genome. The elucidation of the entire human genome has been accelerated by the haplotype map, or HapMap, constructed by SNPs (The International HapMap Consortium, 2003). More recently, the marvelous plans of sequencing the cancer genomes (Kaiser, 2005) will provide unprecedented fuel for studying the genetic architecture of cancer risk.

In this article, we will derive a statistical model for detecting the actions and interactions of haplotypes derived from the host and cancer genomes for cancer susceptibility. We will incorporate the "gene-mutation hypothesis" into the model. The "aneuploidy hypothesis of cancer" will be considered in a next paper. Through the release of software to the public, our statistical model will serve as a routine means for the genetic diagnosis of cancer risk. Results from the model will provide scientific guidance for clinical doctors to design an optimal treatment scheme in terms of cancer genes and patient's genes.

Design

Sampling Strategies

Suppose there is a natural human population at Hardy–Weinberg equilibrium (HWE) from which a random sample is drawn for cancer gene identification. In order to identify DNA sequences responsible for cancer susceptibility, we genotype SNPs from the entire host genome and also SNPs from the cancer genome for the same patient. We assume that the cancer genome is a diploid, whose difference from the normal genome is due to the "genemutation hypothesis". Recent molecular surveys suggest that the human genome contains many discrete haplotype blocks that are sites of closely located SNPs (Daly et al., 2001; Patil et al., 2001; Gabriel et al., 2002). Each block may have a minimal subset of SNPs, i.e., "tagging" SNPs, that can characterize the most common haplotypes. Our model will be based on tagging SNPs within each haplotype block. Although no detailed information about the structure of the cancer genome is available, we can assume that a particular set of SNPs may contribute to cancer formation at the haplotype level. The tenet of our epistatic model is that the effect of a given DNA sequence in the host genome on cancer is masked or enhanced by one or more sequences in the cancer genome.

Genetic Models

Population Genetic Model: Consider a set of *R* tagging SNPs from a haplotype block of the host genome and a set of *S* SNPs from the cancer genome. We denote two alleles of SNP *r* from the host genome by $H^r_{k_r}(k_r = 1,0; r = 1,\cdots R)$ and two alleles of SNP s from the cancer genome by $C^s_{l_s}(l_s = 1,0; s = 1,\cdots S)$. Let $p^{\mathbf{H}}_{k_r}$ and $p^{\mathbf{C}}_{l_s}$ denote allele frequencies at the corresponding SNP from the host and cancer genomes, respectively. All the SNPs considered from the host and cancer genomes form 2^{R+S} possible joint haplotypes expressed as $(H^1_{k_1}H^2_{k_2}\cdots H^R_{k_R})$ $(C^1_{l_1}C^2_{l_2}\cdots C^S_{l_S})$. The corresponding haplotype frequencies are denoted by $p_{(k_1k_2\cdots k_R)(l_1l_2\cdots l_S)}$, which are composed of allele frequencies at each SNP and linkage disequilibria of different orders among SNPs within and between the genomes (Wu and Lin, 2008). A general expression for the relationships

between haplotype frequencies and allele frequencies and linkage disequilibria was originally given by Bennett (1954). Table 1 lists the compositions of the frequency of a haplotype constructed jointly by two SNPs from the host genome and two SNPs from the cancer genome in which linkage disequilibria are specified with two, three, and four sites. From these compositions, linkage disequilibria are expressed as

$$D_{H_1H_2}$$

$$= (p(11)(11) + p(11)(10) + p(11)(01) + p(11)(00))(p(00)(11) + p(00)(10) + p(00)(01) + p(00)(00))$$

$$- (p(10)(11) + p(10)(10) + p(10)(01) + p(10)(00))(p(01)(11) + p(01)(10) + p(01)(01) + p(01)(00)) \quad (1)$$

$$D_{C_1C_2}$$

$$= (p(11)(11) + p(10)(11) + p(01)(11) + p(00)(11))(p(11)(00) + p(10)(00) + p(01)(00) + p(00)(00))$$

$$- (p(11)(10) + p(10)(10) + p(01)(10) + p(00)(10))(p(11)(01) + p(10)(01) + p(01)(01) + p(00)(01)) \quad (2)$$

$$D_{H_2C_1}$$

$$= (p(11)(11) + p(11)(10) + p(01)(11) + p(01)(10))(p(10)(01) + p(10)(00) + p(00)(01) + p(00)(00))$$

$$- (p(11)(01) + p(11)(00) + p(01)(01) + p(01)(00))(p(10)(11) + p(10)(10) + p(00)(11) + p(00)(10)) \quad (3)$$

$$D_{H_2C_2}$$

$$= (p(11)(11) + p(11)(01) + p(01)(11) + p(01)(01))(p(10)(10) + p(10)(00) + p(00)(10) + p(00)(00))$$

$$- (p(11)(10) + p(11)(00) + p(01)(10) + p(01)(00))(p(10)(11) + p(10)(01) + p(00)(11) + p(00)(01)) \quad (4)$$

$$D_{H_1C_1}$$

$$= (p(11)(11) + p(11)(10) + p(10)(11) + p(10)(10))(p(01)(01) + p(01)(00) + p(00)(01) + p(00)(00))$$

$$- (p(11)(01) + p(11)(00) + p(10)(01) + p(10)(00))(p(01)(11) + p(01)(10) + p(00)(11) + p(00)(10)) \quad (5)$$

$$D_{H_1C_2}$$

$$= (p(11)(11) + p(11)(01) + p(10)(11) + p(10)(01))(p(11)(10) + p(11)(00) + p(10)(10) + p(10)(00))$$

$$- (p(11)(10) + p(11)(00) + p(10)(10) + p(10)(00))(p(01)(11) + p(01)(01) + p(00)(11) + p(00)(01)) \quad (6)$$

$$D_{H_1H_2C_1}$$

$$= [(p(11)(11) + p(11)(10))(p(10)(01) + p(10)(00)) + (p(01)(01) + p(01)(00))(p(00)(11) + p(00)(10))]$$

$$- [(p(00)(01) + p(00)(00))(p(01)(11) + p(01)(10)) + (p(11)(01) + p(11)(00))(p(10)(11) + p(10)(10))] \quad (7)$$

$$D_{H_1H_2C_2}$$

$$= [(p(11)(11) + p(11)(01))(p(01)(10) + p(01)(00)) + (p(01)(10) + p(01)(00))(p(00)(11) + p(00)(01))]$$

$$- [(p(00)(10) + p(00)(00))(p(01)(11) + p(01)(01)) + (p(11)(10) + p(11)(00))(p(10)(11) + p(10)(01))] \quad (8)$$

$$D_{H_1C_1C_2}$$

$$= [(p(11)(11) + p(10)(11))(p(11)(00) + p(10)(00)) + (p(01)(10) + p(00)(10))(p(01)(01) + p(00)(01))]$$

$$- [(p(01)(00) + p(00)(00))(p(01)(11) + p(00)(11)) + (p(11)(10) + p(10)(10))(p(11)(01) + p(10)(01))] \quad (9)$$

$$D_{H_2C_1C_2}$$

$$= [(p(11)(11) + p(01)(11))(p(11)(00) + p(01)(00)) + (p(10)(10) + p(00)(10))(p(10)(01) + p(00)(01))]$$

$$- [(p(10)(00) + p(00)(00))(p(10)(11) + p(00)(11)) + (p(11)(10) + p(01)(10))(p(11)(01) + p(01)(01))] \quad (10)$$

$$
\begin{aligned}
& (-1)^4(-1)^{k_1+k_2+l_1+l_2} D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2} \\
= \; & p_{(k_1k_2)(l_1l_2)} - p^{\mathbf{H}}_{k_1} p^{\mathbf{H}}_{k_2} p^{\mathbf{C}}_{l_1} p^{\mathbf{C}}_{l_2} \\
- \; & (-1)^2(-1)^{k_1+k_2} p^{\mathbf{C}}_{l_1} p^{\mathbf{C}}_{l_2} D_{\mathbf{H}_1\mathbf{H}_2} - (-1)^2(-1)^{l_1+l_2} p^{\mathbf{H}}_{k_1} p^{\mathbf{H}}_{k_2} D_{\mathbf{C}_1\mathbf{C}_2} \\
- \; & (-1)^2(-1)^{k_2+l_1} p^{\mathbf{H}}_{k_1} p^{\mathbf{C}}_{l_2} D_{\mathbf{H}_2\mathbf{C}_1} - (-1)^2(-1)^{k_2+l_2} p^{\mathbf{H}}_{k_1} p^{\mathbf{C}}_{l_1} D_{\mathbf{H}_2\mathbf{C}_2} \qquad (11) \\
- \; & (-1)^2(-1)^{k_1+l_1} p^{\mathbf{H}}_{k_2} p^{\mathbf{C}}_{l_2} D_{\mathbf{H}_1\mathbf{C}_1} - (-1)^2(-1)^{k_1+l_2} p^{\mathbf{H}}_{k_2} p^{\mathbf{C}}_{l_1} D_{\mathbf{H}_1\mathbf{C}_2} \\
- \; & (-1)^3(-1)^{k_1+k_2+l_1} p^{\mathbf{C}}_{l_2} D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1} - (-1)^3(-1)^{k_1+k_2+l_2} p^{\mathbf{C}}_{l_1} D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_2} \\
- \; & (-1)^3(-1)^{k_1+l_1+l_2} p^{\mathbf{H}}_{k_2} D_{\mathbf{H}_1\mathbf{C}_1\mathbf{C}_2} - (-1)^3(-1)^{k_2+l_1+l_1} p^{\mathbf{H}}_{k_1} D_{\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2}
\end{aligned}
$$

Term.	Composition	Remark
(1)	$p^{\mathbf{H}}_{k_1} p^{\mathbf{H}}_{k_2} p^{\mathbf{C}}_{l_1} p^{\mathbf{C}}_{l_2}$	No LD
(2)	$(-1)^2(-1)^{k_1+k_2} p^{\mathbf{C}}_{l_1} p^{\mathbf{C}}_{l_2} D_{\mathbf{H}_1\mathbf{H}_2}$	Digenic LD within the host genome (**H**)
(3)	$(-1)^2(-1)^{l_1+l_2} p^{\mathbf{H}}_{k_1} p^{\mathbf{H}}_{k_2} D_{\mathbf{C}_1\mathbf{C}_2}$	Digenic LD within the cancer genome (**C**)
(4)	$(-1)^2(-1)^{k_2+l_1} p^{\mathbf{H}}_{k_1} p^{\mathbf{C}}_{l_2} D_{\mathbf{H}_2\mathbf{C}_1}$	Digenic LD between SNP 2 of **H** and SNP 1 of **C**
(5)	$(-1)^2(-1)^{k_2+l_2} p^{\mathbf{H}}_{k_1} p^{\mathbf{C}}_{l_1} D_{\mathbf{H}_2\mathbf{C}_2}$	Digenic LD between SNP 2 of **H** and SNP 2 of **C**
(6)	$(-1)^2(-1)^{k_1+l_1} p^{\mathbf{H}}_{k_2} p^{\mathbf{C}}_{l_2} D_{\mathbf{H}_1\mathbf{C}_1}$	Digenic LD between SNP 1 of **H** and SNP 1 of **C**
(7)	$(-1)^2(-1)^{k_1+l_2} p^{\mathbf{H}}_{k_2} p^{\mathbf{C}}_{l_1} D_{\mathbf{H}_1\mathbf{C}_2}$	Digenic LD between SNP 1 of **H** and SNP 2 of **C**
(8)	$(-1)^3(-1)^{k_1+k_2+l_1} p^{\mathbf{C}}_{l_2} D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1}$	Trigenic LD between **H** and SNP 1 of **C**
(9)	$(-1)^3(-1)^{k_1+k_2+l_2} p^{\mathbf{C}}_{l_1} D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_2}$	Trigenic LD between **H** and SNP 2 of **C**
(10)	$(-1)^3(-1)^{k_1+l_1+l_2} p^{\mathbf{H}}_{k_2} D_{\mathbf{H}_1\mathbf{C}_1\mathbf{C}_2}$	Trigenic LD between SNP 1 of **H** and **C**
(11)	$(-1)^3(-1)^{k_2+l_1+l_1} p^{\mathbf{H}}_{k_1} D_{\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2}$	Trigenic LD between SNP 2 of **H** and **C**
(12)	$(-1)^4(-1)^{k_1+k_2+l_1+l_2} D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2}$	Quadrigenic LD between **H** and **C**

Table 1: Disequilibrium compositions of four-SNP haplotype frequencies derived from the host and cancer genomes.

The random combination of maternal and paternal haplotypes generates $2^{(R+S-1)}(2^{R+S}+1)$ *diplotypes* expressed as $(H^1_{k_1} H^2_{k_2} \cdots H^R_{k_R})(C^1_{l_1} C^2_{l_2} \cdots C^S_{l_S})|(H^1_{k'_1} H^2_{k'_2} \cdots H^R_{k'_R})(C^1_{l'_1} C^2_{l'_2} \cdots C^S_{l'_S})$ $(k_1 \geq k'_1, k_2 \geq k'_2, ..., k_R \geq k'_R = 1,0; l_1 \geq l'_1, k_2 \geq l'_2, ..., l_S \geq l'_S = 1,0)$. We use a vertical line to separate two haplotypes derived from the maternal and paternal parents, respectively, for a given diplotype. Under the HWE assumption, diplotype frequencies are expressed as the products of the frequencies of the two haplotypes that constitute the diplotype, i.e.,

$$
p_{(k_1k_2\cdots k_R)(l_1l_2\cdots l_S)|(k'_1k'_2\cdots k'_R)(l'_1l'_2\cdots l'_S)} = \begin{cases} p^2_{(k_1k_2\cdots k_R)(l_1l_2\cdots l_S)} & k_1 = k'_1, k_2 = k'_2, \cdots, k_R = k'_R; \\ & l_1 = l'_1, l_2 = l'_2, \cdots, l_S = l'_S \\ 2p_{(k_1k_2\cdots k_R)(l_1l_2\cdots l_S)} p_{(k'_1k'_2\cdots k'_R)(l'_1l'_2\cdots l'_S)} & \text{Otherwise.} \end{cases}
$$

In practice, diplotypes cannot be observed, although observable *zygotic genotypes* will be the same as diplotypes when at most one SNP is heterozygous. Thus, the numbers of zygotic genotypes, 3^{R+S}, will be less than the number of diplotypes.

Quantitative genetic model: Our model will be derived to characterize haplotypes that are responsible for complex traits because the association between haplotype diversity and phenotypic variation has been detected by several genetic studies (Judson, 2000; Bader, 2001; Rha et al., 2007). Among all possible haplotypes are there some particular haplotypes, called the *risk haplotype* (A), that perform differently than the rest of the haplotypes, called the *non-risk haplotype* ($\bar{A}$). The combinations between the risk and non-risk haplotypes, $AA, A\bar{A}$, and $\bar{A}\bar{A}$, are called the *composite diplotype* (Liu et al., 2004; Wu and Lin, 2008). Thus, by testing the differences in the genotypic value of a trait among composite diplotypes, we can estimate the genetic effects of haplotypes on the trait. It is also feasible to detect epistatic interactions between haplotypes from different genomes.

Let $A, \bar{A}$ and $B, \bar{B}$ denote the risk haplotypes and non-risk haplotypes for a series of SNPs genotyped from the host and cancer genomes, respectively. These two genomes form nine different composite diplotypes expressed as $AABB$, $AAB\bar{B}$, $AA\bar{B}\bar{B}$, $A\bar{A}BB$, $A\bar{A}B\bar{B}$, $A\bar{A}\bar{B}\bar{B}$, $\bar{A}\bar{A}BB$, $\bar{A}\bar{A}B\bar{B}$, and $\bar{A}\bar{A}\bar{B}\bar{B}$. We will use Mather and Jinks's (1982) formulation for genetic epistasis between different loci (Table 2) to model the genetic effects of the composite diplotypes. The genotypic value ($\mu_{j_1 j_2}$) of a joint composite diplotype from the two genomes can be decomposed into nine different components as follows:

$$\begin{aligned}
\mu_{j_1 j_2} = \quad & \mu && \text{Overall mean} \\
& +(j_1-1)a_{\mathbf{H}} + (j_2-1)a_{\mathbf{C}} && \text{Additive effects} \\
& +j_1 d_{\mathbf{H}} + j_2 d_{\mathbf{C}} && \text{Dominant effects} \\
& +(j_1-1)(j_2-1)i_{aa} && \text{Additive} \times \text{additive effect} \\
& +(j_1-1)j_2 i_{ad} && \text{Additive} \times \text{dominance effect} \\
& +j_1(j_2-1)i_{da} && \text{Dominance} \times \text{additive effect} \\
& +(1-j_1)(1-j_2)i_{dd} && \text{Dominance} \times \text{dominance effect},
\end{aligned} \tag{12}$$

where

$$j_1, j_2 = \begin{cases} 2 & \text{for } AA \text{ or } BB \\ 1 & \text{for } A\bar{A} \text{ or } B\bar{B} \\ 0 & \text{for } \bar{A}\bar{A} \text{ or } \bar{B}\bar{B} \end{cases}$$

stand for the composite diplotypes from the host and cancer genomes, respectively, μ is the overall mean; $a_{\mathbf{H}}$ and $a_{\mathbf{C}}$ are the additive effects of haplotypes from the host and cancer genome; $d_{\mathbf{H}}$ and $d_{\mathbf{C}}$, the dominance effects of haplotypes; and i_{aa}, i_{ad}, i_{da}, and i_{dd}, the additive × additive, additive × dominance, dominance × additive, and dominance × dominance epistatic effects between the haplotypes from the two different genomes (Table 2).

	BB	$B\bar{B}$	$\bar{B}\bar{B}$
AA	$\mu_{AABB} =$	$\mu_{AAB\bar{B}} =$	$\mu_{AA\bar{B}\bar{B}} =$
	$\mu + a_{\mathbf{H}} + a_{\mathbf{C}} + i_{aa}$	$\mu + a_{\mathbf{H}} + d_{\mathbf{C}} + i_{ad}$	$\mu + a_{\mathbf{H}} - a_{\mathbf{C}} - i_{aa}$
$A\bar{A}$	$\mu_{A\bar{A}BB} =$	$\mu_{A\bar{A}B\bar{B}} =$	$\mu_{A\bar{A}\bar{B}\bar{B}} =$
	$\mu + d_{\mathbf{H}} + a_{\mathbf{C}} + i_{da}$	$\mu + d_{\mathbf{H}} + d_{\mathbf{C}} + i_{dd}$	$\mu + d_{\mathbf{H}} - a_{\mathbf{C}} - i_{da}$
$\bar{A}\bar{A}$	$\mu_{\bar{A}\bar{A}BB} =$	$\mu_{\bar{A}\bar{A}B\bar{B}} =$	$\mu_{\bar{A}\bar{A}\bar{B}\bar{B}} =$
	$\mu - a_{\mathbf{H}} + a_{\mathbf{C}} - i_{aa}$	$\mu - a_{\mathbf{H}} + d_{\mathbf{C}} - i_{ad}$	$\mu - a_{\mathbf{H}} - a_{\mathbf{C}} + i_{aa}$

Table 2: Additive, dominance, and epistatic compositions of the genotypic value of a composite diplotype constructed with haplotypes from the host and cancer genomes.

Different types of genetic actions and interactions can be expressed in terms of genotypic values by solving a group of regular equations (12). This lets us describe the overall mean, additive, dominance, and four kinds of epistatic effects between the two genomes by

$$\begin{aligned}
\mu &= \tfrac{1}{4}(\mu_{\bar{A}\bar{A}\bar{B}\bar{B}} + \mu_{AA\bar{B}\bar{B}} + \mu_{\bar{A}\bar{A}BB} + \mu_{AABB}) \\
a_{\mathbf{H}} &= \tfrac{1}{4}(\mu_{AABB} - \mu_{\bar{A}\bar{A}\bar{B}\bar{B}} + \mu_{AA\bar{B}\bar{B}} - \mu_{\bar{A}\bar{A}BB}) \\
a_{\mathbf{C}} &= \tfrac{1}{4}(\mu_{\bar{A}\bar{A}BB} - \mu_{\bar{A}\bar{A}\bar{B}\bar{B}} - \mu_{AA\bar{B}\bar{B}} + \mu_{AABB}) \\
d_{\mathbf{H}} &= \tfrac{1}{4}(2\mu_{A\bar{A}\bar{B}\bar{B}} - \mu_{\bar{A}\bar{A}\bar{B}\bar{B}} - \mu_{AA\bar{B}\bar{B}} - \mu_{\bar{A}\bar{A}BB} - \mu_{AABB} + 2\mu_{A\bar{A}BB}) \\
d_{\mathbf{C}} &= \tfrac{1}{4}(2\mu_{\bar{A}\bar{A}B\bar{B}} - \mu_{\bar{A}\bar{A}\bar{B}\bar{B}} - \mu_{AA\bar{B}\bar{B}} - \mu_{\bar{A}\bar{A}BB} - \mu_{AABB} + 2\mu_{AA\bar{B}B}) \\
i_{aa} &= \tfrac{1}{4}(\mu_{AABB} - \mu_{AA\bar{B}\bar{B}} - \mu_{\bar{A}\bar{A}BB} + \mu_{\bar{A}\bar{A}\bar{B}\bar{B}}) \\
i_{ad} &= \tfrac{1}{4}(2\mu_{AAB\bar{B}} - \mu_{AABB} - 2\mu_{\bar{A}\bar{A}B\bar{B}} + \mu_{\bar{A}\bar{A}\bar{B}\bar{B}} - \mu_{AA\bar{B}\bar{B}} + \mu_{\bar{A}\bar{A}BB}) \\
i_{da} &= \tfrac{1}{4}(2\mu_{A\bar{A}BB} - 2\mu_{A\bar{A}\bar{B}\bar{B}} + \mu_{\bar{A}\bar{A}\bar{B}\bar{B}} + \mu_{AA\bar{B}\bar{B}} - \mu_{\bar{A}\bar{A}BB} - \mu_{AABB}) \\
i_{dd} &= \tfrac{1}{4}(4\mu_{A\bar{A}B\bar{B}} + \mu_{\bar{A}\bar{A}\bar{B}\bar{B}} + \mu_{AA\bar{B}\bar{B}} + \mu_{\bar{A}\bar{A}BB} + \mu_{AABB} - 2\mu_{A\bar{A}\bar{B}\bar{B}} \\
&\quad -2\mu_{A\bar{A}BB} - 2\mu_{\bar{A}\bar{A}B\bar{B}} - 2\mu_{AAB\bar{B}}).
\end{aligned} \tag{13}$$

Thus, by testing the significance of i_{aa}, i_{ad}, i_{da}, and i_{dd}, we can judge whether there is epistasis and how the epistasis affects a phenotypic trait.

Estimation Procedures

We will estimate two types of parameters for gene cancer identification. First is the population genetic parameters (Ω_p) that describe the distribution and diversity of haplotypes in the sampled population quantified by haplotype frequencies for multiple SNPs from the host and cancer genomes, allele frequencies at these SNPs and their linkage disequilibria. Second is the quantitative genetic parameters (Ω_q) that describe the genotypic values of composite diplotypes, specified by the action and interaction effects of haplotypes on cancer susceptibility, and residual variance. Given observed phenotypic (y) and marker data from the host (**H**) and cancer (**C**) genomes, we construct a likelihood and factorize it to two components:

$$\begin{aligned}
\log L(\Omega_p, \Omega_q|\mathbf{y}, \mathbf{H}, \mathbf{C}) &= \log L(\Omega_p|\mathbf{H}, \mathbf{C}) \\
&\quad + \log L(\Omega_q|\mathbf{y}, \mathbf{H}, \mathbf{C}, \Omega_p),
\end{aligned} \tag{14}$$

where the first component is related to haplotype frequencies and the second component related to haplotype effects and variance. Thus, maximizing the likelihood (14) is equivalent to maximizing its two components separately.

Estimating Across-Genome Haplotype Frequencies: Because the same genotype may be formed from multiple diplotypes, we need to incorporate the EM algorithm to estimate the unknown diplotype of a genotype, which is statistically viewed as a missing data problem. Excoffier and Slatkin, (1995) provided a simple approach for estimating haplotype frequencies, without specifying the configuration of haplotype. We provide a similar estimating algorithm for specifying the haplotype configuration. An observed zygotic genotype is generally expressed as $H^1_{k_1}H^1_{k'_1}/H^2_{k_2}H^2_{k'_2}/\cdots/H^R_{k_R}H^R_{k'_R})\ (C^1_{l_1}C^1_{l'_1}/C^2_{l_2}C^2_{l'_2}/\cdots/C^S_{l_S}C^S_{l'_S})$, where the slashes are used to separate genotypes at different SNPs. Let $n^{(k_1k'_1/k_2k'_2/\cdots/k_Rk'_R)(l_1l'_1/l_2l'_2/\cdots/l_Sl'_S)}$ (which sums to a total sample size of n subjects) be the observation of a typical joint-SNP genotype from the host and cancer genomes. Table 3 is an example of data structure for genotypic observations of two SNPs derived from the host genome and two SNPs from the cancer genome. The table also provides the expected frequencies of different genotypes in terms of haplotype frequencies.

Based on the information about observed data, it is not difficult to construct a multinomial likelihood, log $L(\Omega_p|\mathbf{H},\mathbf{C})$, in which a mixture model is incorporated for those genotypes that are heterozygous at two more SNPs. By maximizing the observed data likelihood, the EM algorithm is derived. In the E step, we calculate the expected number of a particular across-genome haplotype $(H^1_{k_1}H^2_{k_2}\cdots H^R_{k_R})$ $(C^1_{l_1}C^2_{l_2}\cdots C^S_{l_S})$ within the mixture of diplotypes that form the same genotypes. For example, such an expected number is calculated for two SNPs from the host genome and two SNPs from the cancer genome using the following formulas:

$$\phi_1 = \frac{p_{(k_1k_2)(l_1l_2)}p_{(k_1k_2)(l'_1l'_2)}}{p_{(k_1k_2)(l_1l_2)}p_{(k_1k_2)(l'_1l'_2)} + p_{(k_1k_2)(l_1l'_2)}p_{(k_1k_2)(l'_1l_2)}},$$

$$\phi_2 = \frac{p_{(k_1k_2)(l_1l_2)}p_{(k_1k_2')(l_1l_2')}}{p_{(k_1k_2)(l_1l_2)}p_{(k_1k_2')(l_1l_2')} + p_{(k_1k_2)(l_1l_2')}p_{(k_1k_2')(l_1l_2)}},$$

$$\phi_3 = \frac{p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2)(l_1l_2')}}{p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2)(l_1l_2')} + p_{(k_1k_2)(l_1l_2')}p_{(k_1'k_2)(l_1l_2)}},$$

$$\phi_4 = \frac{p_{(k_1k_2)(l_1l_2)}p_{(k_1k_2')(l_1'l_2)}}{p_{(k_1k_2)(l_1l_2)}p_{(k_1k_2')(l_1'l_2)} + p_{(k_1k_2)(l_1'l_2)}p_{(k_1k_2')(l_1l_2)}},$$

$$\phi_5 = \frac{p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2)(l_1'l_2)}}{p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2)(l_1'l_2)} + p_{(k_1k_2)(l_1'l_2)}p_{(k_1'k_2)(l_1l_2)}},$$

$$\phi_6 = \frac{p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2')(l_1l_2)}}{p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2')(l_1l_2)} + p_{(k_1k_2')(l_1l_2)}p_{(k_1'k_2)(l_1l_2)}},$$

$$\psi_1 = [p_{(k_1k_2)(l_1l_2)}p_{(k_1k_2')(l_1'l_2')}]/[p_{(k_1k_2)(l_1l_2)}p_{(k_1k_2')(l_1'l_2')} + p_{(k_1k_2)(l_1l_2')}p_{(k_1k_2')(l_1'l_2)} + p_{(k_1k_2)(l_1'l_2)}p_{(k_1k_2')(l_1l_2')} + p_{(k_1k_2)(l_1'l_2')}p_{(k_1k_2')(l_1l_2)}],$$

$$\psi_2 = [p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2)(l_1'l_2')}]/[p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2)(l_1'l_2')} + p_{(k_1k_2)(l_1l_2')}p_{(k_1'k_2)(l_1'l_2)} + p_{(k_1k_2)(l_1'l_2)}p_{(k_1'k_2)(l_1l_2')} + p_{(k_1k_2)(l_1'l_2')}p_{(k_1'k_2)(l_1l_2)}],$$

$$\psi_3 = [p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2')(l_1l_2')}]/[p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2')(l_1l_2')} + p_{(k_1k_2)(l_1l_2')}p_{(k_1'k_2')(l_1l_2)} + p_{(k_1k_2)(l_1l_2')}p_{(k_1'k_2')(l_1l_2)} + p_{(k_1k_2')(l_1l_2)}p_{(k_1'k_2)(l_1l_2')}],$$

$$\psi_4 = [p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2')(l_1'l_2)}]/[p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2')(l_1'l_2)} + p_{(k_1k_2)(l_1'l_2)}p_{(k_1'k_2')(l_1l_2)} + p_{(k_1k_2')(l_1'l_2)}p_{(k_1'k_2)(l_1l_2)} + p_{(k_1k_2)(l_1'l_2)}p_{(k_1'k_2')(l_1l_2)}],$$

$$\varphi = [p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2')(l_1'l_2')}]/p_\varphi, \quad (15)$$

where

$$p_\varphi = p_{(k_1k_2)(l_1l_2)}p_{(k_1'k_2')(l_1'l_2')} + p_{(k_1k_2)(l_1l_2')}p_{(k_1'k_2')(l_1'l_2)} + p_{(k_1k_2')(l_1l_2)}p_{(k_1'k_2)(l_1'l_2')} + p_{(k_1'k_2)(l_1l_2)}p_{(k_1k_2')(l_1'l_2')} + p_{(k_1k_2)(l_1'l_2')}p_{(k_1'k_2')(l_1l_2)} + p_{(k_1k_2')(l_1l_2')}p_{(k_1'k_2)(l_1'l_2)} + p_{(k_1'k_2)(l_1l_2')}p_{(k_1k_2')(l_1'l_2)} + p_{(k_1'k_2)(l_1'l_2)}p_{(k_1k_2')(l_1l_2')}].$$

In the M step, we estimate a haplotype frequency with the expected number of haplotypes calculated above and the observations given in Table 3 by

$$\begin{aligned}
p_{(k_1k_2)(l_1l_2)} = {} & \frac{1}{2n}[2n_{(k_1k_1/k_2k_2)(l_1l_1/l_2l_2)}] \\
& + n_{(k_1k_1/k_2k_2)(l_1l_1/l_2l_2')}, \quad l_2' < l_2 \\
& + n_{(k_1k_1/k_2k_2)(l_1l_1'/l_2l_2)}, \quad l_1' < l_1 \\
& + n_{(k_1k_1/k_2k_2')(l_1l_1/l_2l_2)}, \quad k_2' < k_2 \\
& + n_{(k_1k_1'/k_2k_2)(l_1l_1/l_2l_2)}, \quad k_1' < k_1 \\
& + \phi_1 n_{(k_1k_1/k_2k_2)(l_1l_1'/l_2l_2')}, \quad l_1' < l_1, l_2' < l_2 \\
& + \phi_2 n_{(k_1k_1/k_2k_2')(l_1l_1/l_2l_2')}, \quad k_2' < k_2, l_2' < l_2 \\
& + \phi_3 n_{(k_1k_1'/k_2k_2)(l_1l_1/l_2l_2')}, \quad k_1' < k_1, l_2' < l_2 \\
& + \phi_4 n_{(k_1k_1/k_2k_2')(l_1l_1'/l_2l_2)}, \quad k_2' < k_2, l_1' < l_1 \\
& + \phi_5 n_{(k_1k_1'/k_2k_2)(l_1l_1'/l_2l_2)}, \quad k_1' < k_1, l_1' < l_1 \\
& + \phi_6 n_{(k_1k_1'/k_2k_2')(l_1l_1/l_2l_2)}, \quad k_1' < k_1, k_2' < k_2 \\
& + \psi_1 n_{(k_1k_1/k_2k_2')(l_1l_1'/l_2l_2')}, \quad k_2' < k_2, l_1' < l_1, l_2' < l_2 \\
& + \psi_2 n_{(k_1k_1'/k_2k_2)(l_1l_1'/l_2l_2')}, \quad k_1' < k_1, l_1' < l_1, l_2' < l_2 \\
& + \psi_3 n_{(k_1'k_1/k_2'k_2')(l_1l_1/l_2l_2')}, \quad k_1' < k_1, k_2' < k_2, l_2' < l_2 \\
& + \psi_4 n_{(k_1'k_1/k_2'k_2')(l_1l_1'/l_2l_2)}, \quad k_1' < k_1, k_2' < k_2, l_1' < l_1 \\
& + \varphi n_{(k_1k_1'/k_2k_2')(l_1l_1'/l_2l_2')}, \quad k_1' < k_1, k_2' < k_2, l_1' < l_1, l_2' < l_2.
\end{aligned} \quad (16)$$

Both the E and M steps are iterated between equations (15) and (16) until the estimates converge to stable values. The estimates at convergence are the maximum likelihood estimates (MLEs) of haplotype frequencies. The MLEs of allele frequencies at different SNPs and their linkage disequilibria of different orders can be solved from these estimated haplotype frequencies using a system of equations given in Table 1.

Estimating across-genome haplotype interactions: To detect how haplotypes or diplotypes are associated with phenotypic variation in a trait (y), we will first assume a risk haplotype from the host genome and a risk haplotype from the cancer genome. These two types of risk haplotypes will generate composite diplotypes. As an example shown in Table 3, in which there are two SNPs from each genome,

	Genotype			
No.	Host	Cancer	Observation	Frequency
1	$H_1^1H_1^1/H_1^2H_1^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(11/11)(11/11)}$	$p^2_{(11)(11)}$
2	$H_1^1H_1^1/H_1^2H_1^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(11/11)(11/10)}$	$2p_{(11)(11)}p_{(11)(10)}$
3	$H_1^1H_1^1/H_1^2H_1^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(11/11)(11/00)}$	$p^2_{(11)(10)}$
4	$H_1^1H_1^1/H_1^2H_1^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(11/11)(10/11)}$	$2p_{(11)(11)}p_{(11)(01)}$
5	$H_1^1H_1^1/H_1^2H_1^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(11/11)(10/10)}$	$2p_{(11)(11)}p_{(11)(00)} + 2p_{(11)(10)}p_{(11)(01)}$
6	$H_1^1H_1^1/H_1^2H_1^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(11/11)(10/00)}$	$2p_{(11)(10)}p_{(11)(00)}$
7	$H_1^1H_1^1/H_1^2H_1^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(11/11)(00/11)}$	$p^2_{(11)(01)}$
8	$H_1^1H_1^1/H_1^2H_1^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(11/11)(00/10)}$	$2p_{(11)(01)}p_{(11)(00)}$
9	$H_1^1H_1^1/H_1^2H_1^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(11/11)(00/00)}$	$p^2_{(11)(00)}$
10	$H_1^1H_1^1/H_1^2H_0^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(11/10)(11/11)}$	$2p_{(11)(11)}p_{(10)(11)}$
11	$H_1^1H_1^1/H_1^2H_0^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(11/10)(11/10)}$	$2p_{(11)(11)}p_{(10)(10)} + 2p_{(10)(11)}p_{(11)(10)}$
12	$H_1^1H_1^1/H_1^2H_0^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(11/10)(11/00)}$	$2p_{(11)(10)}p_{(10)(10)}$
13	$H_1^1H_1^1/H_1^2H_0^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(11/10)(10/11)}$	$2p_{(11)(11)}p_{(10)(01)} + 2p_{(10)(11)}p_{(11)(01)}$
14	$H_1^1H_1^1/H_1^2H_0^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(11/10)(10/10)}$	$2p_{(11)(11)}p_{(10)(00)} + 2p_{(11)(10)}p_{(10)(01)}$ $+2p_{(10)(11)}p_{(10)(00)} + 2p_{(10)(10)}p_{(11)(01)}$
15	$H_1^1H_1^1/H_1^2H_0^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(11/10)(10/00)}$	$2p_{(11)(10)}p_{(10)(00)} + 2p_{(11)(00)}p_{(10)(10)}$
16	$H_1^1H_1^1/H_1^2H_0^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(11/10)(00/11)}$	$2p_{(11)(01)}p_{(10)(01)}$
17	$H_1^1H_1^1/H_1^2H_0^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(11/10)(00/10)}$	$2p_{(11)(01)}p_{(10)(00)} + 2p_{(11)(00)}p_{(10)(01)}$
18	$H_1^1H_1^1/H_1^2H_0^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(11/10)(00/00)}$	$2p_{(11)(00)}p_{(10)(00)}$
19	$H_1^1H_1^1/H_0^2H_0^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(11/00)(11/11)}$	$p^2_{(10)(11)}$
20	$H_1^1H_1^1/H_0^2H_0^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(11/00)(11/10)}$	$2p_{(10)(11)}p_{(10)(10)}$

21	$H_1^1H_1^1/H_0^2H_0^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(11/00)(11/00)}$	$p^2_{(10)(10)}$
22	$H_1^1H_1^1/H_0^2H_0^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(11/00)(10/11)}$	$2p_{(10)(11)}p_{(10)(01)}$
23	$H_1^1H_1^1/H_0^2H_0^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(11/00)(10/10)}$	$2p_{(10)(11)}p_{(10)(00)} + 2p_{(10)(10)}p_{(10)(01)}$
24	$H_1^1H_1^1/H_0^2H_0^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(11/00)(10/00)}$	$2p_{(10)(10)}p_{(10)(00)}$
25	$H_1^1H_1^1/H_0^2H_0^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(11/00)(00/11)}$	$p^2_{(10)(01)}$
26	$H_1^1H_1^1/H_0^2H_0^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(11/00)(00/10)}$	$2p_{(10)(01)}p_{(10)(00)}$
27	$H_1^1H_1^1/H_0^2H_0^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(11/00)(00/00)}$	$2p^2_{(10)(00)}$
28	$H_1^1H_0^1/H_1^2H_1^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(10/11)(11/11)}$	$2p_{(11)(11)}p_{(01)(11)}$
29	$H_1^1H_0^1/H_1^2H_1^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(10/11)(11/10)}$	$2p_{(11)(11)}p_{(01)(10)} + 2p_{(01)(11)}p_{(11)(10)}$
30	$H_1^1H_0^1/H_1^2H_1^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(10/11)(11/00)}$	$2p_{(11)(10)}p_{(01)(10)}$
31	$H_1^1H_0^1/H_1^2H_1^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(10/11)(10/11)}$	$2p_{(11)(11)}p_{(01)(01)} + 2p_{(01)(11)}p_{(11)(10)}$
32	$H_1^1H_0^1/H_1^2H_1^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(10/11)(10/10)}$	$2p_{(11)(11)}p_{(01)(00)} + 2p_{(11)(10)}p_{(01)(01)}$ $+2p_{(01)(11)}p_{(11)(00)} + 2p_{(01)(10)}p_{(11)(01)}$
33	$H_1^1H_0^1/H_1^2H_1^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(10/11)(10/00)}$	$2p_{(11)(10)}p_{(00)(10)} + 2p_{(10)(10)}p_{(01)(10)}$
34	$H_1^1H_0^1/H_1^2H_1^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(10/11)(00/11)}$	$2p_{(11)(01)}p_{(01)(01)}$
35	$H_1^1H_0^1/H_1^2H_1^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(10/11)(00/10)}$	$2p_{(11)(01)}p_{(01)(00)} + 2p_{(01)(01)}p_{(11)(00)}$
36	$H_1^1H_0^1/H_1^2H_1^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(10/11)(00/00)}$	$2p_{(11)(00)}p_{(01)(00)}$
37	$H_1^1H_0^1/H_1^2H_0^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(10/10)(11/11)}$	$2p_{(11)(11)}p_{(00)(11)} + 2p_{(10)(11)}p_{(01)(11)}$
38	$H_1^1H_0^1/H_1^2H_0^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(10/10)(11/10)}$	$2p_{(11)(11)}p_{(00)(10)} + 2p_{(10)(11)}p_{(01)(10)}$ $+2p_{(11)(10)}p_{(00)(11)} + 2p_{(10)(10)}p_{(01)(11)}$
39	$H_1^1H_0^1/H_1^2H_0^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(10/10)(11/00)}$	$2p_{(11)(10)}p_{(00)(10)} + 2p_{(10)(10)}p_{(01)(10)}$
40	$H_1^1H_0^1/H_1^2H_0^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(10/10)(10/11)}$	$2p_{(11)(11)}p_{(00)(01)} + 2p_{(10)(11)}p_{(01)(01)}$ $+2p_{(11)(01)}p_{(00)(11)} + 2p_{(10)(01)}p_{(01)(11)}$

41	$H_1^1H_0^1/H_1^2H_0^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(10/10)(10/10)}$	$2p_{(11)(11)}p_{(00)(00)} + 2p_{(11)(10)}p_{(00)(01)}$ $+2p_{(10)(10)}p_{(01)(01)} + 2p_{(10)(10)}p_{(01)(01)}$ $+2p_{(01)(11)}p_{(10)(00)} + 2p_{(01)(10)}p_{(10)(01)}$ $+2p_{(00)(10)}p_{(11)(01)} + 2p_{(00)(10)}p_{(11)(01)}$
42	$H_1^1H_0^1/H_1^2H_0^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(10/10)(10/00)}$	$2p_{(11)(10)}p_{(00)(00)} + 2p_{(10)(10)}p_{(01)(00)}$ $+2p_{(00)(10)}p_{(11)(00)} + 2p_{(01)(10)}p_{(10)(00)}$
43	$H_1^1H_0^1/H_1^2H_0^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(10/10)(00/11)}$	$2p_{(11)(01)}p_{(00)(01)} + 2p_{(10)(01)}p_{(01)(01)}$
44	$H_1^1H_0^1/H_1^2H_0^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(10/10)(00/10)}$	$2p_{(11)(01)}p_{(00)(00)} + 2p_{(11)(00)}p_{(00)(01)}$ $+2p_{(10)(01)}p_{(01)(00)} + 2p_{(10)(00)}p_{(01)(01)}$
45	$H_1^1H_0^1/H_1^2H_0^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(10/10)(00/00)}$	$2p_{(11)(00)}p_{(00)(00)} + 2p_{(10)(00)}p_{(01)(00)}$
46	$H_1^1H_0^1/H_0^2H_0^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(10/00)(11/11)}$	$2p_{(10)(11)}p_{(00)(11)}$
47	$H_1^1H_0^1/H_0^2H_0^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(10/00)(11/10)}$	$2p_{(10)(11)}p_{(00)(10)} + 2p_{(00)(11)}p_{(10)(10)}$
48	$H_1^1H_0^1/H_0^2H_0^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(10/00)(11/00)}$	$2p_{(10)(10)}p_{(00)(10)}$
49	$H_1^1H_0^1/H_0^2H_0^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(10/00)(10/11)}$	$2p_{(10)(11)}p_{(00)(01)} + 2p_{(00)(11)}p_{(10)(01)}$
50	$H_1^1H_0^1/H_0^2H_0^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(10/00)(10/10)}$	$2p_{(10)(11)}p_{(00)(00)} + 2p_{(10)(10)}p_{(00)(01)}$ $+2p_{(00)(11)}p_{(10)(00)} + 2p_{(10)(10)}p_{(00)(01)}$
51	$H_1^1H_0^1/H_0^2H_0^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(10/00)(10/00)}$	$2p_{(10)(10)}p_{(00)(00)} + 2p_{(10)(00)}p_{(00)(10)}$
52	$H_1^1H_0^1/H_0^2H_0^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(10/00)(00/11)}$	$2p_{(10)(01)}p_{(00)(01)}$
53	$H_1^1H_0^1/H_0^2H_0^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(10/00)(00/10)}$	$2p_{(10)(01)}p_{(00)(00)} + 2p_{(10)(00)}p_{(00)(01)}$
54	$H_1^1H_0^1/H_0^2H_0^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(10/00)(00/00)}$	$2p_{(10)(00)}p_{(00)(00)}$
55	$H_0^1H_0^1/H_1^2H_1^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(00/11)(11/11)}$	$p^2_{(01)(11)}$
56	$H_0^1H_0^1/H_1^2H_1^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(00/11)(11/10)}$	$2p_{(01)(11)}p_{(01)(10)}$
57	$H_0^1H_0^1/H_1^2H_1^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(00/11)(11/00)}$	$p^2_{(01)(10)}$
58	$H_0^1H_0^1/H_1^2H_1^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(00/11)(10/11)}$	$2p_{(01)(11)}p_{(01)(01)}$
59	$H_0^1H_0^1/H_1^2H_1^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(00/11)(10/10)}$	$2p_{(01)(11)}p_{(01)(00)} + p_{(01)(10)}p_{(01)(01)}$

60	$H_0^1H_0^1/H_1^2H_1^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(00/11)(10/00)}$	$2p_{(01)(10)}p_{(01)(00)}$
61	$H_0^1H_0^1/H_1^2H_1^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(00/11)(00/11)}$	$p^2_{(01)(01)}$
62	$H_0^1H_0^1/H_1^2H_1^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(00/11)(00/10)}$	$2p_{(01)(01)}p_{(01)(00)}$
63	$H_0^1H_0^1/H_1^2H_1^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(00/11)(00/00)}$	$p^2_{(01)(00)}$
64	$H_0^1H_0^1/H_1^2H_0^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(00/10)(11/11)}$	$2p_{(01)(11)}p_{(00)(11)}$
65	$H_0^1H_0^1/H_1^2H_0^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(00/10)(11/10)}$	$2p_{(01)(11)}p_{(00)(10)} + 2p_{(00)(11)}p_{(01)(10)}$
66	$H_0^1H_0^1/H_1^2H_0^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(00/10)(11/00)}$	$2p_{(01)(10)}p_{(00)(10)}$
67	$H_0^1H_0^1/H_1^2H_0^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(00/10)(10/11)}$	$2p_{(01)(11)}p_{(00)(01)} + 2p_{(00)(11)}p_{(01)(01)}$
68	$H_0^1H_0^1/H_1^2H_0^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(00/10)(10/10)}$	$2p_{(01)(11)}p_{(00)(00)} + 2p_{(01)(10)}p_{(00)(01)}$ $+2p_{(00)(11)}p_{(01)(00)} + 2p_{(00)(10)}p_{(01)(01)}$
69	$H_0^1H_0^1/H_1^2H_0^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(00/10)(10/00)}$	$2p_{(01)(10)}p_{(00)(00)} + 2p_{(01)(00)}p_{(00)(10)}$
70	$H_0^1H_0^1/H_1^2H_0^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(00/10)(00/11)}$	$2p_{(01)(01)}p_{(00)(01)}$
71	$H_0^1H_0^1/H_1^2H_0^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(00/10)(00/10)}$	$2p_{(01)(01)}p_{(00)(00)} + 2p_{(01)(00)}p_{(00)(01)}$
72	$H_0^1H_0^1/H_1^2H_0^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(00/10)(00/00)}$	$2p_{(01)(00)}p_{(00)(00)}$
73	$H_0^1H_0^1/H_0^2H_0^2$	$C_1^1C_1^1/C_1^2C_1^2$	$n_{(00/00)(11/11)}$	$p^2_{(00)(11)}$
74	$H_0^1H_0^1/H_0^2H_0^2$	$C_1^1C_1^1/C_1^2C_0^2$	$n_{(00/00)(11/10)}$	$2p_{(00)(11)}p_{(00)(10)}$
75	$H_0^1H_0^1/H_0^2H_0^2$	$C_1^1C_1^1/C_0^2C_0^2$	$n_{(00/00)(11/00)}$	$p^2_{(00)(10)}$
76	$H_0^1H_0^1/H_0^2H_0^2$	$C_1^1C_0^1/C_1^2C_1^2$	$n_{(00/00)(10/11)}$	$2p_{(00)(11)}p_{(11)(01)}$
77	$H_0^1H_0^1/H_0^2H_0^2$	$C_1^1C_0^1/C_1^2C_0^2$	$n_{(00/00)(10/10)}$	$2p_{(00)(11)}p_{(00)(00)} + 2p_{(00)(10)}p_{(00)(01)}$
78	$H_0^1H_0^1/H_0^2H_0^2$	$C_1^1C_0^1/C_0^2C_0^2$	$n_{(00/00)(10/00)}$	$2p_{(00)(10)}p_{(00)(00)}$
79	$H_0^1H_0^1/H_0^2H_0^2$	$C_0^1C_0^1/C_1^2C_1^2$	$n_{(00/00)(00/11)}$	$p^2_{(00)(01)}$
80	$H_0^1H_0^1/H_0^2H_0^2$	$C_0^1C_0^1/C_1^2C_0^2$	$n_{(00/00)(00/10)}$	$2p_{(00)(01)}p_{(01)(00)}$
81	$H_0^1H_0^1/H_0^2H_0^2$	$C_0^1C_0^1/C_0^2C_0^2$	$n_{(00/00)(00/00)}$	$p^2_{(00)(00)}$

Table 3: Observed 81 joint host-cancer SNP genotypes and their frequencies described in terms of their haplotype/ diplotype compositions.

we assume $H_1^1 H_1^2$ from the host genome and $C_1^1 C_1^2$ from the cancer genome as two risk haplotypes. This leads to nine different across-genome composite diplotypes. A mixture-based likelihood for quantitative genetic parameters (Ω_q) is formulated as

$$
\begin{aligned}
&\log L(\Omega_q|y,\mathbf{H},\mathbf{C},\hat{\mathbf{\Omega}}_\mathbf{p}) = \\
&\sum_{i=1}^{n_{(11/11)(11/11)}} \log f_{AABB}(y_i) + \sum_{i=1}^{n_{(11/11)(\bullet)}} \log f_{AAB\bar{B}}(y_i) + \sum_{i=1}^{n_{(11/11)(\bullet\bullet)}} \log f_{AA\bar{B}\bar{B}}(y_i) \\
&+ \sum_{i=1}^{n_{(11/11)(10/10)}} \log[\omega_\mathbf{C} f_{AAB\bar{B}}(y_i) + (1-\omega_\mathbf{C}) f_{AA\bar{B}\bar{B}}(y_i)] \\
&+ \sum_{i=1}^{n_{(\bullet)(11/11)}} \log f_{A\bar{A}BB}(y_i) + \sum_{i=1}^{n_{(\bullet)(\bullet)}} \log f_{A\bar{A}B\bar{B}}(y_i) + \sum_{i=1}^{n_{(\bullet)(\bullet\bullet)}} \log f_{A\bar{A}\bar{B}\bar{B}}(y_i) \\
&+ \sum_{i=1}^{n_{(\bullet)(10/10)}} \log[\omega_\mathbf{C} f_{A\bar{A}B\bar{B}}(y_i) + (1-\omega_\mathbf{C}) f_{A\bar{A}\bar{B}\bar{B}}(y_i)] \\
&+ \sum_{i=1}^{n_{(\bullet\bullet)(11/11)}} \log f_{\bar{A}\bar{A}BB}(y_i) + \sum_{i=1}^{n_{(\bullet\bullet)(\bullet)}} \log f_{\bar{A}\bar{A}B\bar{B}}(y_i) + \sum_{i=1}^{n_{(\bullet\bullet)(\bullet\bullet)}} \log f_{\bar{A}\bar{A}\bar{B}\bar{B}}(y_i) \\
&+ \sum_{i=1}^{n_{(\bullet\bullet)(10/10)}} \log[\omega_\mathbf{C} f_{\bar{A}\bar{A}B\bar{B}}(y_i) + (1-\omega_\mathbf{C}) f_{\bar{A}\bar{A}\bar{B}\bar{B}}(y_i)] \\
&+ \sum_{i=1}^{n_{(10/10)(11/11)}} \log[\omega_\mathbf{H} f_{A\bar{A}BB}(y_i) + (1-\omega_\mathbf{H}) f_{\bar{A}\bar{A}BB}(y_i)] \\
&+ \sum_{i=1}^{n_{(10/10)(\bullet)}} \log[\omega_\mathbf{H} f_{A\bar{A}B\bar{B}}(y_i) + (1-\omega_\mathbf{H}) f_{\bar{A}\bar{A}B\bar{B}}(y_i)] \\
&+ \sum_{i=1}^{n_{(10/10)(\bullet\bullet)}} \log[\omega_\mathbf{H} f_{A\bar{A}\bar{B}\bar{B}}(y_i) + (1-\omega_\mathbf{H}) f_{\bar{A}\bar{A}\bar{B}\bar{B}}(y_i)] \\
&+ \sum_{i=1}^{n_{(10/10)(10/10)}} \log[\omega_{A\bar{A}B\bar{B}}^{(10/10)(10/10)} f_{A\bar{A}B\bar{B}}(y_i) + \omega_{A\bar{A}\bar{B}\bar{B}}^{(10/10)(10/10)} f_{A\bar{A}\bar{B}\bar{B}}(y_i) \\
&\qquad + \omega_{\bar{A}\bar{A}B\bar{B}}^{(10/10)(10/10)} f_{\bar{A}\bar{A}B\bar{B}}(y_i) + \omega_{\bar{A}\bar{A}\bar{B}\bar{B}}^{(10/10)(10/10)} f_{\bar{A}\bar{A}\bar{B}\bar{B}}(y_i)],
\end{aligned}
\tag{17}
$$

where

$$
\begin{aligned}
n_{(11/11)(\bullet)} &= n_{(11/11)(11/10)} + n_{(11/11)(10/11)}, \\
n_{(11/11)(\bullet\bullet)} &= n_{(11/11)(11/00)} + n_{(11/11)(10/00)} + n_{(11/11)(00/11)} + n_{(11/11)(00/10)} + n_{(11/11)(00/00)}, \\
n_{(\bullet)(11/11)} &= n_{(11/10)(11/11)} + n_{(10/11)(11/11)}, \\
n_{(\bullet\bullet)(11/11)} &= n_{(11/00)(11/11)} + n_{(10/00)(11/11)} + n_{(00/11)(11/11)} + n_{(00/10)(11/11)} + n_{(00/00)(11/11)}, \\
n_{(\bullet)(\bullet)} &= n_{(11/10)(11/10)} + n_{(11/10)(10/11)} + n_{(10/11)(11/10)} + n_{(10/11)(10/11)}, \\
n_{(\bullet)(\bullet\bullet)} &= n_{(11/10)(11/00)} + n_{(11/10)(10/00)} + n_{(11/10)(00/11)} + n_{(11/10)(00/10)} + n_{(11/10)(00/00)} \\
&\quad + n_{(10/11)(11/00)} + n_{(10/11)(10/00)} + n_{(10/11)(00/11)} + n_{(10/11)(00/10)} + n_{(10/11)(00/00)},
\end{aligned}
$$

$$
\begin{aligned}
n_{(\bullet\bullet)(\bullet)} &= n_{(11/00)(11/10)} + n_{(10/00)(11/10)} + n_{(00/11)(11/10)} + n_{(00/10)(11/10)} + n_{(00/00)(11/10)} \\
&\quad + n_{(11/00)(10/11)} + n_{(10/00)(10/11)} + n_{(00/11)(10/11)} + n_{(00/10)(10/11)} + n_{(00/00)(10/11)}, \\
n_{(\bullet\bullet)(\bullet\bullet)} &= n_{(11/00)(11/00)} + n_{(11/00)(10/00)} + n_{(11/00)(00/11)} + n_{(11/00)(00/10)} + n_{(11/00)(00/00)} \\
&\quad + n_{(10/00)(11/00)} + n_{(10/00)(10/00)} + n_{(10/00)(00/11)} + n_{(10/00)(00/10)} + n_{(10/00)(00/00)} \\
&\quad + n_{(00/11)(11/00)} + n_{(00/11)(10/00)} + n_{(00/11)(00/11)} + n_{(00/11)(00/10)} + n_{(00/11)(00/00)} \\
&\quad + n_{(00/10)(11/00)} + n_{(00/10)(10/00)} + n_{(00/10)(00/11)} + n_{(00/10)(00/10)} + n_{(00/10)(00/00)} \\
&\quad + n_{(00/00)(11/00)} + n_{(00/00)(10/00)} + n_{(00/00)(00/11)} + n_{(00/00)(00/10)} + n_{(00/00)(00/00)}, \\
n_{(\bullet)(10/10)} &= n_{(11/10)(10/10)} + n_{(10/11)(10/10)}, \\
n_{(10/10)(\bullet)} &= n_{(10/10)(11/10)} + n_{(10/10)(10/11)}, \\
n_{(\bullet\bullet)(10/10)} &= n_{(11/00)(10/10)} + n_{(10/00)(10/10)} + n_{(00/11)(10/10)} + n_{(00/10)(10/10)} + n_{(00/00)(10/10)}, \\
n_{(10/10)(\bullet\bullet)} &= n_{(10/10)(11/00)} + n_{(10/10)(10/00)} + n_{(10/10)(00/11)} + n_{(10/10)(00/10)} + n_{(10/10)(00/00)},
\end{aligned}
$$

$$
\begin{aligned}
\omega_{\mathbf{H}} &= \frac{\hat{p}^{\mathbf{H}}_{11}\hat{p}^{\mathbf{H}}_{00}}{\hat{p}^{\mathbf{H}}_{11}\hat{p}^{\mathbf{H}}_{00} + \hat{p}^{\mathbf{H}}_{10}\hat{p}^{\mathbf{H}}_{01}}, \ \hat{p}^{\mathbf{H}}_{k_1k_2} = \sum_{l_1=0}^{1}\sum_{l_2=0}^{1}\hat{p}_{(k_1k_2)((l_1l_2)}, \\
\omega_{\mathbf{C}} &= \frac{\hat{p}^{\mathbf{C}}_{11}\hat{p}^{\mathbf{C}}_{00}}{\hat{p}^{\mathbf{C}}_{11}\hat{p}^{\mathbf{C}}_{00} + \hat{p}^{\mathbf{C}}_{10}\hat{p}^{\mathbf{C}}_{01}}, \ \hat{p}^{\mathbf{C}}_{l_1l_2} = \sum_{k_1=0}^{1}\sum_{k_2=0}^{1}\hat{p}_{(k_1k_2)((l_1l_2)}, \\
\omega^{(10/10)(10/10)}_{A\bar{A}B\bar{B}} &= \frac{p_{(11)(11)}p_{(00)(00)} + p_{(11)(00)}p_{(00)(11)}}{p_{\varphi}}, \\
\omega^{(10/10)(10/10)}_{A\bar{A}\bar{B}\bar{B}} &= \frac{p_{(11)(10)}p_{(00)(01)} + p_{(11)(01)}p_{(00)(10)}}{p_{\varphi}}, \\
\omega^{(10/10)(10/10)}_{\bar{A}\bar{A}B\bar{B}} &= \frac{p_{(10)(11)}p_{(01)(00)} + p_{(01)(11)}p_{(10)(00)}}{p_{\varphi}}, \\
\omega^{(10/10)(10/10)}_{\bar{A}\bar{A}\bar{B}\bar{B}} &= \frac{p_{(10)(10)}p_{(01)(01)} + p_{(10)(01)}p_{(01)(10)}}{p_{\varphi}}.
\end{aligned}
$$

In likelihood (17), we model $f_{\ldots.}(y_i)$ by a normal distribution with diplotype-specific mean $\mu_{\ldots.}$ and variance σ^2. The EM algorithm is implemented to estimate these means and variance that maximize the likelihood. In the E step, we calculate the posterior probability of a particular diplotype within a genotype for SNPs across the genomes using

$$
\begin{aligned}
\Phi^{\mathbf{C}}_{(11/11)(10/10)i} &= \frac{\omega_{\mathbf{C}} f_{AA\bar{B}\bar{B}}(y_i)}{\omega_{\mathbf{C}} f_{AA\bar{B}\bar{B}}(y_i) + (1-\omega_{\mathbf{C}}) f_{A\bar{A}\bar{B}\bar{B}}(y_i)} \\
\Phi^{\mathbf{C}}_{(\bullet)(10/10)i} &= \frac{\omega_{\mathbf{C}} f_{A\bar{A}B\bar{B}}(y_i)}{\omega_{\mathbf{C}} f_{A\bar{A}B\bar{B}}(y_i) + (1-\omega_{\mathbf{C}}) f_{A\bar{A}\bar{B}\bar{B}}(y_i)} \\
\Phi^{\mathbf{C}}_{(\bullet\bullet)(10/10)i} &= \frac{\omega_{\mathbf{C}} f_{\bar{A}\bar{A}B\bar{B}}(y_i)}{\omega_{\mathbf{C}} f_{\bar{A}\bar{A}B\bar{B}}(y_i) + (1-\omega_{\mathbf{C}}) f_{\bar{A}\bar{A}\bar{B}\bar{B}}(y_i)} \\
\Phi^{\mathbf{H}}_{(10/10)(11/11)i} &= \frac{\omega_{\mathbf{H}} f_{A\bar{A}BB}(y_i)}{\omega_{\mathbf{H}} f_{A\bar{A}BB}(y_i) + (1-\omega_{\mathbf{H}}) f_{\bar{A}\bar{A}BB}(y_i)} \\
\Phi^{\mathbf{H}}_{(10/10)(\bullet)i} &= \frac{\omega_{\mathbf{H}} f_{A\bar{A}B\bar{B}}(y_i)}{\omega_{\mathbf{H}} f_{A\bar{A}B\bar{B}}(y_i) + (1-\omega_{\mathbf{H}}) f_{\bar{A}\bar{A}B\bar{B}}(y_i)} \\
\Phi^{\mathbf{H}}_{(10/10)(\bullet\bullet)i} &= \frac{\omega_{\mathbf{H}} f_{A\bar{A}\bar{B}\bar{B}}(y_i)}{\omega_{\mathbf{H}} f_{A\bar{A}\bar{B}\bar{B}}(y_i) + (1-\omega_{\mathbf{H}}) f_{\bar{A}\bar{A}\bar{B}\bar{B}}(y_i)} \\
\Phi^{A\bar{A}B\bar{B}}_{(10/10)(10/10)i} &= \frac{\omega^{(10/10)(10/10)}_{A\bar{A}B\bar{B}} f_{A\bar{A}B\bar{B}}(y_i)}{\Delta_i}
\end{aligned}
\qquad (18)
$$

$$\Phi^{A\bar{A}\bar{B}\bar{B}}_{(10/10)(10/10)i} = \frac{\omega^{(10/10)(10/10)}_{A\bar{A}\bar{B}\bar{B}} f_{A\bar{A}\bar{B}\bar{B}}(y_i)}{\Delta_i}$$

$$\Phi^{\bar{A}A\bar{B}\bar{B}}_{(10/10)(10/10)i} = \frac{\omega^{(10/10)(10/10)}_{\bar{A}A\bar{B}\bar{B}} f_{\bar{A}A\bar{B}\bar{B}}(y_i)}{\Delta_i}$$

$$\Phi^{\bar{A}\bar{A}B\bar{B}}_{(10/10)(10/10)i} = \frac{\omega^{(10/10)(10/10)}_{\bar{A}\bar{A}B\bar{B}} f_{\bar{A}\bar{A}B\bar{B}}(y_i)}{\Delta_i}$$

with $\Delta_i = \omega^{(10/10)(10/10)}_{A\bar{A}B\bar{B}} f_{A\bar{A}B\bar{B}}(y_i) + \omega^{(10/10)(10/10)}_{A\bar{A}\bar{B}B} f_{A\bar{A}\bar{B}B}(y_i) + \omega^{(10/10)(10/10)}_{\bar{A}AB\bar{B}} f_{\bar{A}AB\bar{B}}(y_i) + \omega^{(10/10)(10/10)}_{\bar{A}A\bar{B}B} f_{\bar{A}A\bar{B}B}(y_i)$.

In the M step, the quantitative genetic parameters are estimated by

$$\mu_{AABB} = \frac{\sum_{i=1}^{n_{(11/11)(11/11)}} y_i}{n_{(11/11)(11/11)}}$$

$$\mu_{AAB\bar{B}} = \frac{\sum_{i=1}^{n_{(11/11)(\bullet)}} y_i + \sum_{i=1}^{n_{(11/11/)(10/10)}} \Phi^{\mathbf{C}}_{(11/11)(10/10)i} y_i}{n_{(11/11)(\bullet)} + \sum_{i=1}^{n_{(11/11/)(10/10)}} \Phi^{\mathbf{C}}_{(11/11)(10/10)i}}$$

$$\mu_{AA\bar{B}\bar{B}} = \frac{\sum_{i=1}^{n_{(11/11)(\bullet\bullet)}} y_i + \sum_{i=1}^{n_{(11/11/)(10/10)}} (1-\Phi^{\mathbf{C}}_{(11/11)(10/10)i}) y_i}{n_{(11/11)(\bullet)} + \sum_{i=1}^{n_{(11/11/)(10/10)}} (1-\Phi^{\mathbf{C}}_{(11/11)(10/10)i})}$$

$$\mu_{A\bar{A}BB} = \frac{\sum_{i=1}^{n_{(\bullet)(11/11)}} y_i + \sum_{i=1}^{n_{(10/10/)(11/11)}} \Phi^{\mathbf{H}}_{(10/10)(11/11)i} y_i}{n_{(\bullet)(11/11)} + \sum_{i=1}^{n_{(10/10/)(11/11)}} \Phi^{\mathbf{H}}_{(10/10)(11/11)i})}$$

$$\mu_{A\bar{A}B\bar{B}} = \frac{\sum_{i=1}^{n_{(\bullet)(\bullet)}} y_i + \sum_{i=1}^{(\bullet)n_{(10/10/)}} \Phi^{\mathbf{H}}_{(\bullet)(10/10)i} y_i}{n_{(\bullet)(\bullet)} + \sum_{i=1}^{n_{(\bullet)(10/10/)}} \Phi^{\mathbf{H}}_{(\bullet)(10/10)i}}$$

$$\mu_{A\bar{A}\bar{B}\bar{B}} = \frac{\sum_{i=1}^{n_{(\bullet)(\bullet\bullet)}} y_i + \sum_{i=1}^{n_{(\bullet)(10/10/)}} (1-\Phi^{\mathbf{H}}_{(\bullet)(10/10)i}) y_i + \sum_{i=1}^{n_{(10/10)(\bullet\bullet)}} \Phi^{\mathbf{H}}_{(10/10)(\bullet\bullet)i} y_i + \sum_{i=1}^{n_{(10/10/)(10/10)}} \Phi^{A\bar{A}\bar{B}\bar{B}}_{(10/10)(10/10)i} y_i}{n_{(\bullet)(\bullet\bullet)} + \sum_{i=1}^{(\bullet)n_{(10/10/)}} (1-\Phi^{\mathbf{H}}_{(\bullet)(10/10)i}) + + \sum_{i=1}^{n_{(10/10)(\bullet\bullet)}} (1-\Phi^{\mathbf{H}}_{(10/10)(\bullet\bullet)i}) + \sum_{i=1}^{n_{(10/10/)(10/10)}} \Phi^{A\bar{A}\bar{B}\bar{B}}_{(10/10)(10/10)i}}$$

$$\mu_{\bar{A}\bar{A}BB} = \frac{\sum_{i=1}^{n_{(\bullet\bullet)(11/11)}} y_i + \sum_{i=1}^{n_{(10/10/)(11/11)}} (1-\Phi^{\mathbf{H}}_{(10/10)(11/11)i}) y_i}{n_{(\bullet\bullet)(11/11)} + \sum_{i=1}^{n_{(10/10/)(11/11)}} (1-\Phi^{\mathbf{H}}_{(10/10)(11/11)i})}$$

$$\mu_{\bar{A}\bar{A}B\bar{B}} = \frac{\sum_{i=1}^{n_{(\bullet\bullet)(\bullet)}} y_i + \sum_{i=1}^{n_{(\bullet\bullet)(10/10/)}} \Phi^{\mathbf{C}}_{(\bullet\bullet)(10/10)i} y_i + \sum_{i=1}^{n_{(10/10)(\bullet)}} (1-\Phi^{\mathbf{H}}_{(10/10)(\bullet)i}) y_i + \sum_{i=1}^{n_{(10/10)(10/10)}} \Phi^{\bar{A}\bar{A}B\bar{B}}_{(10/10)(10/10)i} y_i}{n_{(\bullet\bullet)(\bullet)} + \sum_{i=1}^{n_{(\bullet\bullet)(10/10/)}} \Phi^{\mathbf{C}}_{(\bullet\bullet)(10/10)i} + \sum_{i=1}^{n_{(10/10)(\bullet)}} (1-\Phi^{\mathbf{H}}_{(10/10)(\bullet)i}) + \sum_{i=1}^{n_{(10/10)(10/10)}} \Phi^{\bar{A}\bar{A}B\bar{B}}_{(10/10)(10/10)i}}$$

$$\mu_{\bar{A}\bar{A}\bar{B}\bar{B}} = \frac{\sum_{i=1}^{n_{(\bullet\bullet)(\bullet\bullet)}} y_i + \sum_{i=1}^{n_{(\bullet\bullet)(10/10/)}} (1-\Phi^{\mathbf{H}}_{(\bullet\bullet)(10/10)i}) y_i + \sum_{i=1}^{n_{(10/10)(\bullet\bullet)}} (1-\Phi^{\mathbf{H}}_{(10/10)(\bullet\bullet)i}) y_i + \sum_{i=1}^{n_{(10/10/)(10/10)}} \Phi^{\bar{A}\bar{A}\bar{B}\bar{B}}_{(10/10)(10/10)i} y_i}{n_{(\bullet\bullet)(\bullet\bullet)} + \sum_{i=1}^{n_{(\bullet\bullet)(10/10/)}} (1-\Phi^{\mathbf{H}}_{(\bullet\bullet)(10/10)i}) + \sum_{i=1}^{n_{(10/10)(\bullet\bullet)}} (1-\Phi^{\mathbf{H}}_{(10/10)(\bullet\bullet)i}) + \sum_{i=1}^{n_{(10/10/)(10/10)}} \Phi^{\bar{A}\bar{A}\bar{B}\bar{B}}_{(10/10)(10/10)i}}$$

$$\sigma^2 = \sum_{i=1}^{n_{(11/11)(11/11)}} (y_i - \mu_{AABB})^2 + \sum_{i=1}^{n_{(11/11)(\bullet)}} (y_i - \mu_{AAB\bar{B}})^2 + \sum_{i=1}^{n_{(11/11)(\bullet\bullet)}} (y_i - \mu_{AA\bar{B}\bar{B}})^2$$
$$+ \sum_{i=1}^{n_{(11/11)(10/10)}} [\Phi^{\mathbf{C}}_{(11/11)(10/10)i}(y_i - \mu_{AAB\bar{B}})^2 + (1 - \Phi^{\mathbf{C}}_{(11/11)(10/10)i})(y_i - \mu_{AA\bar{B}\bar{B}})^2]$$
$$+ \sum_{i=1}^{n_{(\bullet)(11/11)}} (y_i - \mu_{A\bar{A}BB})^2 + \sum_{i=1}^{n_{(\bullet)(\bullet)}} (y_i - \mu_{A\bar{A}B\bar{B}})^2 + \sum_{i=1}^{n_{(\bullet)(\bullet\bullet)}} (y_i - \mu_{A\bar{A}\bar{B}\bar{B}})^2$$
$$+ \sum_{i=1}^{n_{(\bullet)(10/10)}} [\Phi^{\mathbf{C}}_{(\bullet)(10/10)i}(y_i - \mu_{A\bar{A}B\bar{B}})^2 + (1 - \Phi^{\mathbf{C}}_{(\bullet)(10/10)i})(y_i - \mu_{A\bar{A}\bar{B}\bar{B}})^2]$$
$$+ \sum_{i=1}^{n_{(\bullet\bullet)(11/11)}} (y_i - \mu_{\bar{A}\bar{A}BB})^2 + \sum_{i=1}^{n_{(\bullet\bullet)(\bullet)}} (y_i - \mu_{\bar{A}\bar{A}B\bar{B}})^2 + \sum_{i=1}^{n_{(\bullet\bullet)(\bullet\bullet)}} (y_i - \mu_{\bar{A}\bar{A}\bar{B}\bar{B}})^2$$
$$+ \sum_{i=1}^{n_{(\bullet\bullet)(10/10)}} [\Phi^{\mathbf{C}}_{(\bullet\bullet)(10/10)i}(y_i - \mu_{\bar{A}\bar{A}B\bar{B}})^2 + (1 - \Phi^{\mathbf{C}}_{(\bullet\bullet)(10/10)i})(y_i - \mu_{\bar{A}\bar{A}\bar{B}\bar{B}})^2]$$
$$+ \sum_{i=1}^{n_{(10/10)(11/11)}} [\Phi^{\mathbf{H}}_{(10/10)(11/11)i}(y_i - \mu_{A\bar{A}BB})^2 + (1 - \Phi^{\mathbf{H}}_{(10/10)(11/11)i})(y_i - \mu_{\bar{A}\bar{A}BB})^2]$$
$$+ \sum_{i=1}^{n_{(10/10)(\bullet)}} [\Phi^{\mathbf{H}}_{(10/10)(\bullet)i}(y_i - \mu_{A\bar{A}B\bar{B}})^2 + (1 - \Phi^{\mathbf{H}}_{(10/10)(\bullet)i})(y_i - \mu_{\bar{A}\bar{A}B\bar{B}})^2]$$
$$+ \sum_{i=1}^{n_{(10/10)(\bullet\bullet)}} [\Phi^{\mathbf{H}}_{(10/10)(\bullet\bullet)}(y_i - \mu_{A\bar{A}\bar{B}\bar{B}})^2 + (1 - \Phi^{\mathbf{H}}_{(10/10)(\bullet\bullet)})(y_i - \mu_{\bar{A}\bar{A}\bar{B}\bar{B}})^2]$$
$$+ \sum_{i=1}^{n_{(10/10)(10/10)}} [\Phi^{A\bar{A}B\bar{B}}_{(10/10)(10/10)i}(y_i - \mu_{A\bar{A}B\bar{B}})^2 + \Phi^{A\bar{A}\bar{B}\bar{B}}_{(10/10)(10/10)i}(y_i - \mu_{A\bar{A}\bar{B}\bar{B}})^2$$
$$+ \Phi^{\bar{A}\bar{A}B\bar{B}}_{(10/10)(10/10)i}(y_i - \mu_{\bar{A}\bar{A}B\bar{B}})^2 + \Phi^{\bar{A}\bar{A}\bar{B}\bar{B}}_{(10/10)(10/10)i}(y_i - \mu_{\bar{A}\bar{A}\bar{B}\bar{B}})^2], \quad (19)$$

A loop of the E and M steps is formulated between equations (18) and (19) to obtain the MLEs of the genotypic values and variance.

For a practical data set, risk haplotypes are unknown. A combinatory approach is used to detect an optimal combination of risk haplotypes derived from the host and cancer genomes. This can be chosen from all 16 possible combinations. The combination that gives the largest likelihood is considered as the best risk-haplotype combination. Under such an optimal combination, we estimate genotypic values of the composite diplotypes including μ_{AABB}, $\mu_{AAB\bar{B}}$, $\mu_{AA\bar{B}\bar{B}}$, $\mu_{A\bar{A}BB}$, $\mu_{A\bar{A}B\bar{B}}$, $\mu_{A\bar{A}\bar{B}\bar{B}}$, $\mu_{\bar{A}\bar{A}BB}$, $\mu_{\bar{A}\bar{A}B\bar{B}}$, and $\mu_{\bar{A}\bar{A}\bar{B}\bar{B}}$. From the estimated μ.... values, we can then solve the additive, dominance, and epistatic effects between haplotypes from the two genomes using a system of equations (13).

Hypothesis Tests

It is imperative to know how different SNPs are associated within and between the host and cancer genomes and how haplotypes trigger cancer susceptibility singly or epistatically across different genomes. Two kinds of major hypotheses can be made to address these questions.

Linkage Disequilibrium Tests: The association between different SNPs within each genome and between two dif-

ferent genomes by testing their linkage disequilibria (LD). For example, the LD between four SNPs from the two genomes as shown in Table 1 can be tested using the two hypotheses as follows:

$$\begin{cases} H_0: & D_{\mathbf{H}_1\mathbf{H}_2} = D_{\mathbf{C}_1\mathbf{C}_2} = D_{\mathbf{H}_1\mathbf{C}_1} = D_{\mathbf{H}_1\mathbf{C}_2} = D_{\mathbf{H}_2\mathbf{C}_1} = D_{\mathbf{H}_2\mathbf{C}_2} = D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1} \\ & \quad = D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_2} = D_{\mathbf{H}_1\mathbf{C}_1\mathbf{C}_2} = D_{\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2} = D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2} = 0 \\ H_1: & \text{At least one of the LD above is not equal to zero.} \end{cases} \quad (20)$$

The log-likelihood ratio test statistic for the significance of LD is calculated by comparing the likelihood values under the H_1 (full model) and H_0 (reduced model) using

$$\mathrm{LR}_D = -2[\log L(\hat{p}^{\mathbf{H}}_{k_1}, \hat{p}^{\mathbf{H}}_{k_2}, \hat{p}^{\mathbf{C}}_{l_1}, \hat{p}^{\mathbf{C}}_{l_2}, \mathrm{LD} = 0 | \mathbf{H}, \mathbf{C}) - \log L(\hat{\Omega}_p | \mathbf{H}, \mathbf{C})] \quad (21)$$

where $p^{\mathbf{H}}_{k_1}$, $\hat{p}^{\mathbf{H}}_{k_2}$, $\hat{p}^{\mathbf{C}}_{l_1}$, and $\hat{p}^{\mathbf{C}}_{l_2}$ are the MLEs of allele frequencies at four SNPs from the two genomes. The LR_D calculated under the H_0 and H_1 hypotheses is considered to asymptotically follow a χ^2 distribution with the degrees of freedom (11) equal to the differences in the number of unknown parameters between the alternative and null hypotheses.

It is also interesting to test whether the linkage disequilibria of a different particular order are significant. This can be done by formulating the null hypotheses:

$$\begin{cases} H_0: & D_{\mathbf{H}_1\mathbf{H}_2} = D_{\mathbf{C}_1\mathbf{C}_2} = D_{\mathbf{H}_1\mathbf{C}_1} = D_{\mathbf{H}_1\mathbf{C}_2} = D_{\mathbf{H}_2\mathbf{C}_1} = D_{\mathbf{H}_2\mathbf{C}_2} = 0 \\ H_1: & \text{At least one of the LD above is not equal to zero,} \end{cases} \quad (22)$$

for the digenic LD,

$$\begin{cases} H_0: & D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1} = D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_2} = D_{\mathbf{H}_1\mathbf{C}_1\mathbf{C}_2} = D_{\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2} = 0 \\ H_1: & \text{At least one of the LD above is not equal to zero,} \end{cases} \quad (23)$$

for the trigenic LD, and

$$\begin{cases} H_0: & D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2} = 0 \\ H_1: & D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2} \neq 0, \end{cases} \quad (24)$$

for the quadrigenic LD.

Each LD can also be tested separately. Under the null hypothesis of no LD, haplotype frequencies are estimated with the same EM algorithm derived to estimate the frequency parameters under the alternative hypothesis, except for the constraint posed on the relationships of haplotype frequencies under the null hypothesis. Depending on which type of LD is tested, these constraints can be obtained from equations (1)–(11), respectively.

Genome-Genome Epistasis Tests: The significance of an assumed risk haplotype for its effect on cancer susceptibility should be tested by formulating the following hypotheses, expressed as

$$\begin{cases} H_0: & \mu_{AABB} = \mu_{AAB\bar{B}} = \mu_{AA\bar{B}\bar{B}} = \mu_{A\bar{A}BB} = \mu_{A\bar{A}B\bar{B}} \\ & \quad = \mu_{A\bar{A}\bar{B}\bar{B}} = \mu_{\bar{A}\bar{A}BB} = \mu_{\bar{A}\bar{A}B\bar{B}} = \mu_{\bar{A}\bar{A}\bar{B}\bar{B}} = \mu \\ H_1: & \text{At least one equality in } H_0 \text{ does not hold} \end{cases} \quad (25)$$

The log-likelihood ratio test statistic (LR_E) under these two hypotheses can be similarly calculated. The LR_E may asymptotically follow a χ^2 distribution with eight degrees of freedom. However, the approximation of a χ^2 distribution may be inappropriate when some regularity conditions, such as normality and uncorrelated residuals, are violated. The permutation test approach (Churchill and Doerge 1994), which does not rely upon the distribution of the LR_E, may be used to determine the critical threshold for determining the existence of risk haplotypes.

Different genetic effects, such as the additive, dominance, and additive × additive, additive × dominance, dominance × additive, and dominance × dominance effects between haplotypes from the host and cancer genomes can also be tested individually, with respective null hypotheses formulated as

$$H_0 : a_{\mathrm{H}} = 0, \quad (26)$$

$$H_0 : a_{\mathrm{C}} = 0, \quad (27)$$

$$H_0 : d_{\mathrm{H}} = 0, \quad (28)$$

$$H_0 : d_{\mathrm{C}} = 0, \quad (29)$$

$$H_0 : i_{aa} = 0, \quad (30)$$

$$H_0 : i_{ad} = 0, \quad (31)$$

$$H_0 : i_{da} = 0, \quad (32)$$

$$H_0 : i_{dd} = 0, \quad (33)$$

The parameter estimation under each of these null hypotheses can be obtained using the same EM algorithm as described for the alternative hypothesis (full model) of equation (25), with a constraint derived from a system of equations (13). The critical thresholds for these individual effects (26)–(33) can be determined on the basis of simulation studies.

Computer Simulation

The statistical behavior of the model proposed for cancer gene identification is investigated through simulation studies. We simulate a HWE population of cancer individuals in which a set of SNPs from the host genome are associated with a different set of SNPs from the cancer genome. The haplotypes of these SNPs within and across the genomes trigger main and interaction effects on the susceptibility of a cancer. The allele frequencies of two of the SNPs from each genome and their linkage disequilibria of different orders are given in Table 4. Four sample sizes from modest (200) to intermediate (400) to large (800) to very large (2000) are considered. These population genetic parameters that specify the distribution and diversity of haplotypes can be well estimated with the model. A modest sample size is adequate for estimating allele frequencies and digenic linkage disequilibria. An intermediate to large sample size is needed to estimate higher-order linkage disequilibria. Especially, to precisely estimate quadrigenic linkage disequilibrium, a sample of 2000 is recommended (Table 4).

		MLE			
Parameters	True	200	400	800	2000
$p^{\mathbf{H}}_{k_1}$	0.6	0.6004(0.0023)	0.6019(0.0017)	0.5995(0.0012)	0.5993(0.0007)
$p^{\mathbf{H}}_{k_2}$	0.5	0.5018(0.0025)	0.4987(0.0018)	0.4980(0.0013)	0.5003(0.0008)
$p^{\mathbf{C}}_{l_1}$	0.5	0.4981(0.0023)	0.5018(0.0018)	0.5017(0.0014)	0.5005(0.0007)
$p^{\mathbf{C}}_{l_2}$	0.4	0.3976(0.0024)	0.4033(0.0015)	0.3991(0.0015)	0.3998(0.0008)
$D_{\mathbf{H}_1\mathbf{H}_2}$	0.08	0.0778(0.0014)	0.0790(0.0010)	0.0797(0.0007)	0.0796(0.0005)
$D_{\mathbf{C}_1\mathbf{C}_2}$	0.08	0.0785(0.0016)	0.0809(0.0009)	0.0799(0.0007)	0.0808(0.0005)
$D_{\mathbf{H}_1\mathbf{C}_1}$	0.05	0.0504(0.0017)	0.0516(0.0012)	0.0494(0.0007)	0.0498(0.0005)
$D_{\mathbf{H}_1\mathbf{C}_2}$	0.05	0.0482(0.0016)	0.0499(0.0009)	0.0493(0.0008)	0.0499(0.0004)
$D_{\mathbf{H}_2\mathbf{C}_1}$	0.05	0.0468(0.0016)	0.0506(0.0010)	0.0505(0.0008)	0.0501(0.0005)
$D_{\mathbf{H}_2\mathbf{C}_2}$	0.05	0.0489(0.0017)	0.0493(0.0010)	0.0495(0.0009)	0.0496(0.0005)
$D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1}$	0.02	0.0193(0.0009)	0.0197(0.0006)	0.0195(0.0005)	0.0198(0.0003)
$D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_2}$	0.02	0.0194(0.0010)	0.0215(0.0006)	0.0206(0.0004)	0.0199(0.0003)
$D_{\mathbf{H}_1\mathbf{C}_1\mathbf{C}_2}$	0.02	0.0202(0.0009)	0.0201(0.0007)	0.0201(0.0004)	0.0206(0.0003)
$D_{\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2}$	0.02	0.0198(0.0009)	0.0195(0.0007)	0.0198(0.0004)	0.0200(0.0003)
$D_{\mathbf{H}_1\mathbf{H}_2\mathbf{C}_1\mathbf{C}_2}$	0.01	0.0098(0.0006)	0.0094(0.0004)	0.0096(0.0003)	0.0102(0.0001)

Table 4: The MLEs of population genetic parameters for two host SNPs and two cancer SNPs and the standard deviations of the estimates (in parantheses) in a simulated cancer population of varying sampling sizes. The results were obtained from 200 simulation replicates.

For the assumed population in which multiple SNPs are typed, a quantitative trait that describes cancer susceptibility was simulated, following a normal distribution with means depending on composite diplotypes of SNPs and a residual variance. The genotypic values of composite diplotype are determined by assuming specific values for the additive, dominance and epistatic effects of haplotypes on the cancer trait. Composite diplotypes are formed by assuming two risk haplotypes for the SNPs, one from the host genome $(H_1^1 H_1^2)$ and the second from the cancer genome $(C_1^1 C_1^2)$. The genotypic values of a total of 16 composite diplotypes, along with their probabilities calculated from haplotype frequencies, are used to compute the genetic variance. The residual variance is then determined by assuming different heritability levels (0.1 and 0.4).

	Cancer			
Host	$C_1^1C_1^2$	$C_1^1C_0^2$	$C_0^1C_1^2$	$C_0^1C_0^2$
$H_1^1H_1^2$	−561.96	−568.55	−571.42	−562.05
$H_1^1H_0^2$	−564.33	−574.81	−577.29	−570.54
$H_0^1H_1^2$	−564.82	−576.77	−579.90	−573.42
$H_0^1H_0^2$	−563.40	−577.65	−578.03	−564.63

Table 5: Log-likelihood values for different combinations of risk haplotypes from the host and cancer genomes in a simulated cancer population of 200 subjects with a heritability of 0.1.

Table 5 gives the log-likelihoods of population and quantitative genetic parameters by assuming different combinations of risk haplotypes from the host and cancer genomes for simulation data with sample size 200 and heritability 0.1. It can be seen that risk haplotype combination $(H_1^1 H_2^2)$ $(C_1^1 C_2^2)$ corresponds to the maximum likelihood among all possible combinations, suggesting that the model has correctly selected risk haplotypes. It appears that the power to correctly select the optimal combination of risk haplotypes is very high, even when a modest sample size and/or heritability is assumed (data not shown). The model provides reasonable estimates of quantitative genetic parameters (Table 6). The estimation precision of parameters increases with sample size and heritability. For the additive genetic effects, a modest sample size (200) is quite enough even when the heritability is low (0.1). The good estimation of dominance genetic effects needs an intermediately large sample size (400 or more) for a small heritability. Epistasis, especially the dominance × dominance interaction, can be reasonably estimated when a large sample size (2000) is used. This is especially true for a small heritability.

		MLE			
Parameters	True	200	400	800	2000
$H^2 = 0.1$					
$a_{\mathbf{H}}$	1.0	1.04(0.0799)	0.98(0.0509)	0.98(0.0198)	0.97(0.0079)
$a_{\mathbf{C}}$	0.8	0.76(0.0750)	0.73(0.0504)	0.82(0.0186)	0.81(0.0074)
$d_{\mathbf{H}}$	0.5	0.48(0.1017)	0.50(0.0641)	0.39(0.0231)	0.53(0.0098)
$d_{\mathbf{C}}$	0.8	0.92(0.1043)	0.95(0.0601)	0.73(0.0260)	0.77(0.0097)
i_{aa}	0.6	0.64(0.0827)	0.54(0.0532)	0.59(0.0195)	0.60(0.0086)
i_{ad}	0.5	0.55(0.0981)	0.38(0.0585)	0.54(0.0252)	0.58(0.0106)
i_{da}	0.5	0.66(0.0913)	0.47(0.0638)	0.41(0.0246)	0.51(0.0099)
i_{dd}	0.4	0.46(0.1356)	0.32(0.0815)	0.56(0.0338)	0.39(0.0123)

$H^2 = 0.4$					
$a_{\mathbf{H}}$	1.0	0.93(0.0354)	1.03(0.0213)	1.00(0.0075)	1.01(0.0032)
$a_{\mathbf{C}}$	0.8	0.83(0.0366)	0.79(0.0205)	0.79(0.0083)	0.80(0.0036)
$d_{\mathbf{H}}$	0.5	0.29(0.0508)	0.48(0.0224)	0.49(0.0110)	0.48(0.0041)
$d_{\mathbf{C}}$	0.8	0.70(0.0425)	0.81(0.0235)	0.79(0.0102)	0.76(0.0044)
i_{aa}	0.6	0.48(0.0390)	0.62(0.0203)	0.60(0.0083)	0.59(0.0032)
i_{ad}	0.5	0.60(0.0430)	0.46(0.0249)	0.50(0.0091)	0.50(0.0039)
i_{da}	0.5	0.45(0.0439)	0.53(0.0256)	0.49(0.0109)	0.49(0.0040)
i_{dd}	0.4	0.61(0.0501)	0.42(0.0281)	0.43(0.0128)	0.44(0.0053)

Table 6: The MLEs of quantitative genetic parameters of haplotypes for SNPs typed from the host and cancer genomes and the standard deviations of the estimates (in parantheses) in a simulated cancer population of varying sampling sizes and heritabilities. The results were obtained from 200 simulation replicates.

Discussion

Despite painstaking cumulative efforts to fight against cancer by researchers worldwide in the past five decades, we have still not achieved substantial progress in diagnosis, prevention and treatment of this disease. The latest research, however, has found a possibility to treat, control and prevent cancer by using gene therapy. The successful use of this technique relies on our profound understanding of the genetic architecture of cancer susceptibility and progression. When tremendous progress in genotyping and sequencing the human genome and cancer genome has taken place, we are now in a great position to study the genetic control mechanism of cancer. In this article, we have developed a statistical model for characterizing DNA sequence variants that encode cancer susceptibility.

The novelty of the model lies in three aspects. First, we incorporate the latest discovery of cancer genetics into the model that gene mutation cause cancer (Jallepalli and Lengauer, 2001; Stock and Bialy, 2002; Balman et al., 2003; Greenman et al., 2007). The model is not only able to characterize how gene mutation in the cancer genome acts to regulate cancer, but also can detect the genetic interactions between the host genes and cancer mutation. The model allows the test of haplotype distribution and diversity in the cancer population and patterns of genetic actions and interactions. Second, the model is integrated with multilocus SNP data, detecting cancer genes at the DNA sequence level (Liu et al., 2004; Wu and Lin, 2008). This will provide significant insights into the genetic regulation mechanisms of cancer and cloning of cancer genes. Third, the model was built on the interactions of genes between different genomes. Modeling genome-genome interactions has received an increasing interest in studying the genetic architecture of seed development (Cui and Wu, 2006) and pathogenesis (Foster et al., 2003; Wang et al., 2005).

The model was investigated in terms of its statistical behavior through simulation studies. Different schemes of simulation that consider varying sample sizes and heritabilities were used. The estimation precision of parameters and the power to detect genetic variants for cancer was explored under different schemes. The results from simulation provide scientific support for the model to be used for cancer gene identification in practical data sets. Although we did not include real data to validate our model, the statistical design and algorithm proposed in this work will help cancer geneticists and clinicians launch a novel experiment to test hypotheses about the genetic control of cancer.

This article presents a framework for haplotyping cancer genes, and its extension to including genes × environment interactions, haplotyping in a case-control study, genetic imprinting, and an arbitrary number of SNPs will be possible. As an inherited disease, genetic research of cancer is beneficial from an informative family-structured design in which one or both of the parents and offspring are sampled simultaneously. The general principle of haplotyping genome-genome interactions can be used for such a family design, facilitating our understanding of cancer genetics. Also, cancer can be better viewed as a dynamic trait which undergoes marked developmental transition. Functional mapping advocated by our group (Ma et al., 2002; Liu et al., 2005; Wu and Lin, 2006) can be implemented into the haplotyping model to explore the developmental change of genetic control of cancer in time course. In this article, we focus on the "gene-mutation hypothesis" of cancer formation when the epistatic model was derived. Other hypotheses, such as "aneuploidy hypothesis of cancer" (Stock and

Bialy, 2002), should also be integrated into the model, to better understand the genetic mechanisms of cancer formation and progression. The model that incorporate the aneuploidy control of cancer will be reported in other articles.

The preparation of this manuscript is partially supported by Joint NSF/NIH grant DMS/NIGMS-0540745.

References

1. Araujo RP, McElwain DLS (2006) The role of mechanical host-tumour interactions in the collapse of tumour blood vessels and tumour growth dynamics. J Theor Biol 238: 817- 827. » CrossRef » Pubmed » Google Scholar

2. Balman A, Gray J, Ponder B (2003) The genetics and genomics of cancer. Nature Genetics 33: 238-244. » CrossRef » Google Scholar

3. Brennan P (2002) Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it. Carcinogenesis 23: 381-387. » CrossRef » Pubmed » Google Scholar

4. Cui YH, Wu RL (2005) Mapping genome-genome epistasis: A multi-dimensional model. Bioinformatics 21: 2447-2455. » Google Scholar

5. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nature Genetics 29: 229-232. » CrossRef » Pubmed » Google Scholar

6. Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921-927. » CrossRef » Pubmed » Google Scholar

7. Foster JS, Palmer RJ Jr, Kolenbrander PE (2003) Human oral cavity as a model for the study of genome-genome interactions. Biol Bull 204: 200-204. » CrossRef » Pubmed » Google Scholar

8. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, et al. (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-2229. » CrossRef » Pubmed » Google Scholar

9. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153-158. » CrossRef » Pubmed » Google Scholar

10. Jallepalli PV, Lengauer C (2001) Chromosome segregation and cancer: Cutting through the mystery. Nat Rev Cancer 1: 109-117. » CrossRef » Pubmed » Google Scholar

11. Kaiser J (2005) Tackling the cancer genome. Science 309: 6.

12. Kolenbrander PE, Egland PG, Diaz PI, Palmer RJ Jr (2004) Genome-genome interactions: bacterial communities in initial dental plaque. Trends Microbiol 13: 11-15. » CrossRef » Pubmed » Google Scholar

13. Lin M, Wu RL (2006) Detecting sequence-sequence interactions for complex diseases. Current Genomics 7: 59-72. » Google Scholar

14. Liu T, Johnson JA, Casella G, Wu RL (2004) Sequencing complex diseases with HapMap. Genetics 168: 503-511. » CrossRef » Pubmed » Google Scholar

15. Liu T, Zhao W, Tian LL, Wu RL (2005) An algorithm for molecular dissection of tumor progression. Journal of Mathematical Biology 50: 336-354. » CrossRef » Pubmed » Google Scholar

16. Ma CX, Casella G, Wu RL (2002) Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. Genetics 161: 1751-1762. » CrossRef » Pubmed » Google Scholar

17. Patil N, Berno AJ, Hinds DA, Barrett WA, et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294: 1719-1723. » CrossRef » Pubmed » Google Scholar

18. Rand V, Prebble E, Ridley L, Howard M, Wei W, et al. (2008) Investigation of chromosome 1q reveals differential expression of members of the S100 family in clinical subgroups of intracranial paediatric ependymoma. Br J Cancer doi: 10.1038/sj.bjc.6604651. » CrossRef » Pubmed » Google Scholar

19. Stock RP, Bialy H (2002) The sigmoidal curve of cancer. Nat Biotech 21: 13-14.

20. The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789-794. » Pubmed » Google Scholar

21. Thompson SL, Compton DA (2008) Examining the link between chromosomal instability and aneuploidy in human cells. J Cell Biol 180: 665-672. » CrossRef » Pubmed » Google Scholar

22. Wang ZH, Hou W, Wu RL (2005) A statistical model to analyze quantitative trait locus interactions for HIV dynamics from the virus and human genomes. Stat Med 25: 495-511. » CrossRef » Google Scholar

23. Wu RL, Lin M (2006) Functional mapping C A new tool to study the genetic architecture of dynamic complex trait. Nat Rev Genet 7: 229-237.