2870–2878 Nucleic Acids Research, 2014, Vol. 42, No. 5 Published online 25 December 2013
doi:10.1093/nar/gkt1320

A phylogenetic model for understanding the effect of gene duplication on cancer progression

Qin Ma[1], Jaxk H. Reeves[2], David A. Liberles[3], Lili Yu[4], Zheng Chang[5], Jing Zhao[2], Juan Cui[6], Ying Xu[1,7,8,*] and Liang Liu[1,2,*]

[1]Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA, [2]Department of Statistics, University of Georgia, Athens, GA 30602, USA, [3]Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA, [4]Department of Biostatistics, Georgia Southern University, Statesboro, GA 30458, USA, [5]School of Mathematics, Shandong University, Jinan 250100, China, [6]Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 65888, USA, [7]BioEnergy Science Center, Oak Ridge, TN 37830, USA and [8]College of Computer Science and Technology, Jilin University, Changchun, Jilin, China

Received July 23, 2013; Revised November 22, 2013; Accepted November 26, 2013

ABSTRACT

As biotechnology advances rapidly, a tremendous amount of cancer genetic data has become available, providing an unprecedented opportunity for understanding the genetic mechanisms of cancer. To understand the effects of duplications and deletions on cancer progression, two genomes (normal and tumor) were sequenced from each of five stomach cancer patients in different stages (I, II, III and IV). We developed a phylogenetic model for analyzing stomach cancer data. The model assumes that duplication and deletion occur in accordance with a continuous time Markov Chain along the branches of a phylogenetic tree attached with five extended branches leading to the tumor genomes. Moreover, coalescence times of the phylogenetic tree follow a coalescence process. The simulation study suggests that the maximum likelihood approach can accurately estimate parameters in the phylogenetic model. The phylogenetic model was applied to the stomach cancer data. We found that the expected number of changes (duplication and deletion) per gene for the tumor genomes is significantly higher than that for the normal genomes. The goodness-of-fit test suggests that the phylogenetic model with constant duplication and deletion rates can adequately fit the duplication data for the normal genomes. The analysis found nine duplicated genes that are significantly associated with stomach cancer.

INTRODUCTION

Cancer is one of the leading causes of death in Americans (1). Cancer research has led to a variety of effective treatments and diagnostic techniques for cancers. Yet, the fundamental genetic mechanisms that turn normal cells into tumors remain mysterious. Advances in the biotechnology field have provided an unprecedented opportunity for understanding the origin and progression of cancer (2–4). The availability of genetic data ignites the hope that we may discover the genetic mechanisms of cancer by examining the genetic differences between normal and cancer genomes (5). It is, however, a challenging task to effectively analyze such genetic data by modeling the genetic variation observed within and between the normal and cancer groups (6). Previous studies have demonstrated that cancer progression is an evolutionary process in which mutation and natural selection are two key factors (7,8). Mutation causes genetic variation among normal cells that can trigger cancer (9). On the other hand, selection plays an important role in therapeutic resistance (10–12) and in the birth and death process of cancer cells, as cancer cells vary and the fittest ones survive after competition (13).

In the last few decades, theory from evolution and ecology has been adapted in cancer studies to investigate the genetic mechanisms of cancer (14–18). Muto *et al*. (19) studied colon cancers and found that most colon cancers have evolved from adenomatous polyps known as polyp-cancer sequences. Evolutionary ideas have been explored in many cancer analyses (20–22). Nowell (23) proposed a landmark colonial evolution model for tumor progression, which assumes that most neoplasms originate from a single cell. Gillies *et al*. (24) proposed an evolutionary

*To whom correspondence should be addressed. Tel: +1 706 542 3309; Fax: +1 706 542 3391; Email: lliu@uga.edu
Correspondence may also be addressed to Ying Xu. Tel: +1 706 542 7783; Fax: +1 706 542 9751; Email: xyn@bmb.uga.edu

model for malignant cancers, in which the micro-environmental forces such as hypoxia can stimulate genetic instability and impose selection pressures on cancer cells. Recently, Wu (25) investigated the evolution of cancer cells after the primary tumor had spread to secondary sites (26). Ultimately, cancer evolution within an individual can be viewed genetically as adaptation to a new lifestyle and ecologically in the context of the other cell types and resources available in an individual.

Heterogeneity of cancer caused by genetic instability is the main challenge in the process of understanding cancer evolution and in the process of identifying driver genes (8). Due to this challenge, cancer data sets often lack signal regarding the evolutionary process of cancer. It is difficult to find genomic mutations/events that trigger cancer, especially those that trigger the early-stage cancer. High throughput technologies, particularly next generation sequences (NGS) provide researchers with new opportunities to understand the evolutionary process of cancer development at a single cell nucleotide level (16,27–29). NGS technology is able to identify alterations in the genome, e.g. chromosomal rearrangement and copy number variation, rather than point mutations; and can sequence genetic material from lower-frequency samples (30). Because of the advantages of NGS data, the NGS technology has been extensively used in cancer studies to examine genetic mechanisms that cause cancer progression.

Gene duplication is believed to play an important role in tumor progression (31). Duplicated genes have been frequently observed in the genomes of cancer patients. Waris and Ahsan (32) suggested that gene duplication and other changes in DNA may be involved in the initiation of various cancers. Previous studies found that there is a strong correlation between gene duplication and large tumor size, indicating that gene duplication may play a critical role in tumor progression (33). However, the information at early stages of cancer is usually unavailable and little is known about the relationship between gene duplication and early-stage cancer.

The primary goal of the study is to investigate the effects of gene duplication and deletion on the incidence and progression of cancers. Specifically, this study aims to estimate the duplication and deletion rates on normal and tumor genomes, and to identify duplicated genes that are highly associated with stomach cancer. We have developed a probabilistic model in the context of coalescent trees of normal genomes attached with five tumor genomes for understanding how gene duplication and deletion are related to different stages of cancer as cancer progresses. A maximum likelihood approach is adopted to estimate model parameters, including duplication and deletion rates. This approach can identify duplicated genes that are significantly associated with stomach cancer.

MATERIALS AND METHODS

Genome annotation and duplication data

The genomic data was obtained from five stomach cancer patients (34). Two samples (tumor and normal) were taken from each patient; tumor tissues were surgically removed from part of the patients' stomachs, while blood samples were extracted as normal tissues from the same patients (34). Determination of pathologic stages of tumor tissues is based on the standards recommended by World Health Organization (WHO). Pathological examination suggested that two patients were in stage II of stomach cancer, while remaining three patients were in stages I, III and IV (Table 1). Stomach cancer has two subtypes in terms of the genome instability—micro satellite instability (35) and chromosome instability (36). However, the subtypes of the stomach cancer for five patients in this study are not available in (34). The genomes were sequenced for each of the two samples (normal and tumor). Both the normal and tumor genomes were compared with the human reference genome to identify duplicated genes. As the human reference genome is a haploid sequence, it may result in underestimation of duplication events. High-confidence duplication events were identified if a junction in the genomic data satisfied all of the following criteria: (i) at least 10 mate-pairs in cluster for its junction, (ii) successful *de novo* assembly of the junction, (iii) high mapping diversity with both left length and right length no less than 70 and (iv) absence of specific repeat sequences on left and right side of junction. As it is assumed that duplication events occur independently among genes, the junctions that covered more than two genes were excluded. With these criteria, we identified 210 genes on which duplication occurred for at least one of the 10 genomes (5 normal and 5 tumors). We use '1' to denote duplication and '0' to denote no duplication. Cui *et al.* (34) did not estimate the total number of genes in the genomes of five patients. We used the estimate from ENCODE (37) that the total number of genes in the human genome is 21 000. Because the most significant inferences are based on the relative duplication and deletion rates, uncertainty in the total number of human genes does not affect the major conclusions of the data analysis. In summary, the data matrix D has 21 000 rows and 10 columns; each row represents a gene and each column represents a genome (normal or tumor). The entries in the matrix D are either 0 (no duplication) or 1 (duplication).

Table 1. The number of duplicated genes on the genomes of five stomach cancer patients

Subject	Stage	No of duplicated genes in tumor genomes
S1	I	64
S2	II	84
S3	II	57
S4	III	75
S5	IV	72

A phylogenetic model for gene duplication

The stochastic process of gene duplication and deletion

The process of gene duplication and deletion is a continuous time Markov process with two states 1 (duplication)

and 0 (no duplication). Let $d(t)$ denote the Markov process on states 0 and 1. We assume that transition probabilities $P_{ij}(t)$ are stationary and the infinitesimal duplication and deletion rates are a and b, respectively. The probability of a duplication event (and a deletion event) during a time period of duration Δt, as $\Delta t \to 0$, is (38,39)

$$\begin{aligned} &P(d(t+\Delta_t) = 1 | d(t) = 0) = a\Delta_t + o(\Delta_t) \text{ and} \\ &P(d(t+\Delta_t) = 0 | d(t) = 1) = b\Delta_t + o(\Delta_t) \end{aligned} \quad (1)$$

The notation $o(\Delta t)$ indicates $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$. The probability distribution $P(t)$ of $d(t)$ can be derived from theory of Markov process. Let $T = (a+b)t$ and $m = a/(a+b)$. The transition probabilities $P_{ij}(t)$ are given as follows:

$$\begin{aligned} P_{d_1,d_2}(T) = &|d_1 + d_2 - 1| + (1 - 2d_1)(2d_2 - 1)m^{1-x_1} \\ &\times (1-m)^{x_1}(1 - e^{-T}), \text{ for } d_1, d_2 = \{0,1\}. \end{aligned} \quad (2)$$

As $T \to \infty$,

$$P_{0,0}(T) \to 1 - m, P_{0,1}(T) \to m, P_{1,0}(T) \to 1 - m, \text{ and } P_{1,1}(T) \to m \quad (3)$$

Thus the limiting distribution as $T \to \infty$ is $P(d(T) = 0) = 1 - m$ and $P(d(T) = 1) = m$. This model is time reversible, in the sense that $P(d(t) = 0, d(t + T) = 1) = P(d(t) = 1, d(t + T) = 0)$.

The likelihood function under the phylogenetic model

A phylogenetic tree with five extended branches describes the history of 10 genomes of five cancer patients. The five normal genomes are at the tips of the tree, which are attached with five extended branches leading to the tumor genomes (Figure 1). The tree without the extended branches represents the history of five normal genomes, while the extended branches represent the duplication/deletion process that leads to the tumor genomes. Each tumor progression was treated as an independent process, even though the tumors progressed through the same stages phenotypically. As described above, the normal and tumor sequences are coded as binary data (0: no duplication or 1: duplication). Each site of the sequences contains duplication status of a gene across five patients. Parameters of the phylogenetic model include the topology of the tree, branch lengths T_i, and parameter $m = a/(a+b)$. Let M be the number of branches in the tree and W be the number of extended branches. We assume that m is constant on the main branches of the tree, but the extended branches have variable (relative) duplication rates $\{m^e_k, k = 1, \ldots, W\}$. Given a phylogenetic tree S (topology, branch lengths T and parameter m) and the extended branches (branch lengths T^e and m^e), the probability distribution of data matrix D can be derived from the transition probability function in (2). Let d_{ij} and d^*_{ij} be the duplication status of gene j at the two ends of branch i (with length T_i) in tree S with topology τ.

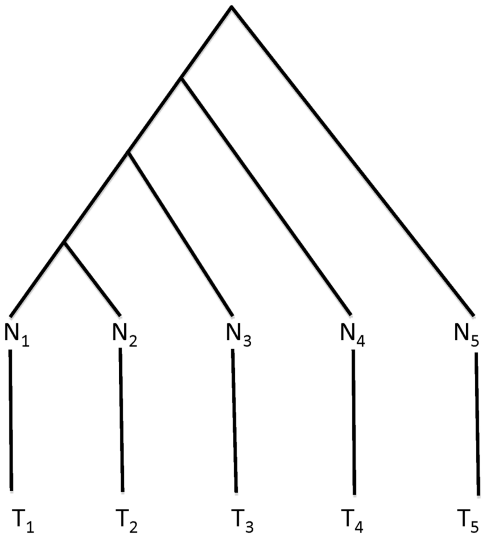

Figure 1. The tree in the phylogenetic model. The normal genomes (N) at the tips of the tree are attached with five extended branches leading to the tumor genomes (T). N_i and T_i are the normal and tumor genomes of patient i. The tree (above normal genomes) represents the history of five normal genomes, while the extended branches represent the process leading to the five tumor genomes. The normal genomes are the ancestral genomes of the tumor genomes.

It follows from (2) that the probability of d_{ij} and d^*_{ij}, given branch length T_i, parameter m, and tree topology τ, is

$$\begin{aligned} P(d_{ij}, d^*_{ij} | T_i, m, \tau) = &\left| d_{ij} + d^*_{ij} - 1 \right| \\ &+ (1 - 2d_{ij})(2d^*_{ij} - 1)m^{1-d_{ij}}(1-m)^{d_{ij}}(1 - e^{-T_i}). \end{aligned} \quad (4)$$

Given the states of the internal nodes, the Markov processes on different branches of the phylogenetic tree are independent of one another. The probability distribution function of duplication events on gene j (denoted by D_j) is the product of the probabilities for individual branches in (4), i.e.,

$$\begin{aligned} P(D_j, I | T, m, \tau, T^e, m^e) = &\left\{ \prod_{i=1}^{M} P(d_{ij}, d^*_{ij} | T_i, m, \tau) \right\} \\ &\times \left\{ \prod_{k=1}^{W} P(c_{kj}, c^*_{kj} | T^e_k, m^e_k) \right\}. \end{aligned} \quad (5)$$

In (5), I denotes the duplication status at the internodes of the tree. The first term in (5) is the probability of duplication events in normal genomes, given the phylogenetic tree without the extended branches. The second term is the probability of duplication events in tumor genomes, given the extended branches, in which c_{kj} and c^*_{kj} are the duplication status of gene j at the two ends of the extended branch k. The probability function $P(D_j, I \mid T, m, \tau, T^e, m^e_k)$ in (5) assumes that the duplication status at the internodes of the tree are given. Because in reality I is often not given, we calculate

$P(D_j \mid T, m, \tau, T^e, m^e)$, which is the sum over all possible realizations of I, i.e.,

$$P(D_j|T,m,\tau,T^e,m^e) = \sum_I P(D_j, I|T,m,\tau,T^e,m^e). \quad (6)$$

The probability distribution $P(D_j \mid T, m, \tau, T^e, m^e)$ in (6) can be efficiently calculated by a peeling technique described by Felsenstein (40).

In the context of population genetics, the phylogenetic tree of n individuals varies over different loci due to the coalescent. Let $t = \{t_j, j = 2, \ldots, n\}$ be the waiting times until the next coalescence event. Let $\theta = 4uN_e$ be the population size parameter, in which N_e is the effective population size and u is the change (duplication and deletion) rate per gene. According to the coalescent theory, the waiting times t_j's are independently distributed with the exponential density

$$f(t_j|\theta) = \frac{j(j-1)}{\theta} e^{-\frac{j(j-1)t_j}{\theta}}. \quad (7)$$

The expected coalescent time for a haploid genome from two individuals chosen at random from the human population is $E(t_2) = \theta/2$, which indicates that if the sequences of a gene are sampled from one of the genomes of two individuals chosen at random from the human population, the expected duplication probability $E(P_{0,I}(\mathrm{T}))$ is equal to the probability $P_{0,I}(\mathrm{T})$, averaging over coalescence time T, which has an exponential density described in (7), i.e.,

$$E(P_{0,1}(T)) = \int_0^\infty m(1-e^{-2T})\frac{2}{\theta}e^{-2T/\theta}\,dt = \frac{m\theta}{\theta+1}. \quad (8)$$

Similarly, the expected deletion probability is $E(P_{1,0}(T)) = \frac{(1-m)\theta}{\theta+1}$. The expected number of changes per gene between two genomes chosen at random from the human population is

$$E(P(x_1 = 0, x_2 = 1|T)) = \int_0^\infty (1-m)m(1-e^{-2T})\frac{2}{\theta}e^{-2T/\theta}dT = \frac{(1-m)m\theta}{\theta+1}, \quad (9)$$

in which x_1 represents the duplication status of a gene in one of the two genomes, and x_2 represents the duplication status of the same gene in the other genome.

The parameters θ and m are estimated by averaging over gene trees, in which branch lengths T are the sum of a set of coalescence waiting times t with a density function described in (7). The probability of observing certain duplication states (D for current individuals and I for their ancestors) of a gene is then equal to the likelihood in (5) (without the extended branches) averaging over coalescence waiting times t i.e.,

$$P(D_j, I|m,\theta) = \int_t \prod_{i=1}^M P(d_{ij}, d_{ij}^*|T_i, m, \tau) f(t|\theta)dt.$$

The probability $P(D_j \mid m, \theta)$ for a single locus is equal to the probability in (5) summing over all possible duplication states at the internal nodes of the phylogenetic tree,

$$p(D_j|m,\theta) = \sum_I P(D_j, I|m,\theta)$$

Since probability $P(D_j \mid m, \theta)$ under the coalescent model is invariant to the order of the duplication states of individuals, the relevant random variable here is the number of duplications across individuals. When there are n individuals, the number of individuals who have duplication for a particular gene could be 0 up to n. Let $\{x_i, i = 0, \ldots, n\}$ be the number of genes for which i individuals have duplications. The sum of x_i's (N) is the total number (21 000) of genes considered in this study. We use $\{p_i, i = 0, \ldots, n\}$ to denote the probability of observing i individuals with duplication. Thus $\{x_i, \mathrm{i} = 0, \ldots, \mathrm{n}\}$ follows a multinomial distribution, i.e.,

$$P(x \mid m, \theta) = \frac{N!}{x_0! \ldots x_n!}\prod_{i-0}^n P_i^{x_i},$$

Because the multinomial coefficient does not involve model parameters, we delete this term and write the log-likelihood function as

$$l(m,\theta) = \sum_{i-0}^n x_i \log p_i \quad (10)$$

In this equation, p_i is a function of m and θ, which will be derived as follows under the coalescent theory. Considering a simple case of two individuals, the coalescence time t has an exponential distribution with density $2e^{-2t/\theta}/\theta$. Let y be the duplication state at the root, and z_1 and z_2 be the duplicate states of two individuals at the tips of the tree. As there are only two states for y, z_1 and z_2, the domain of y, z_1 and z_2 has only two values, 0 and 1. The goal is to derive the probabilities of ($z_1 = 0$, $z_2 = 0$), ($z_1 = 0$, $z_2 = 1$), ($z_1 = 1$, $z_2 = 0$) and ($z_1 = 1$, $z_2 = 1$). We assume that the states at the root have the equilibrium distribution with probability mass function $P(y = 0) = 1 - m$ and $P(y = 1) = m$. The probability of ($z_1 = 0$, $z_2 = 0$) is

$$\begin{aligned} P(z_1 = 0, z_2 = 0) &= P(z_1 = 0, z_2 = 0|y = 0) \\ &\quad \times P(y = 0) + P(z_1 = 0, z_2 = 0|y = 1) \times P(y = 1) \\ &= (1-m)\int_{t=0}^\infty 2(1-m+me^{-t})^2 e^{-2t/\theta}/\theta dt \\ &\quad + m\int_{t=0}^\infty 2(1-m)^2(1-e^{-t})^2 e^{-2t/\theta}/\theta dt \\ &= (1-m)^2 + m(1-m)/(\theta+1). \end{aligned}$$

Similarly, $P(z_1 = 0, z_2 = 1) = P(z_1 = 1, z_2 = 0) = \theta m(1-m)/(\theta+1)$, and $P(z_1 = 1, z_2 = 1) = m^2 + m(1-m)/(\theta+1)$. Thus we have $p_0 = (1-m)^2 + m(1-m)/(\theta+1)$,

$p_1 = 2\theta m(1-m)/(\theta+1)$, and $p_2 = m^2 + m(1-m)/(\theta+1)$. The log-likelihood function becomes

$$l(m,\theta) = x_0 log\{(1-m)^2 + m(1-m)/(\theta+1)\} + x_1 log\{2\theta m(1-m)/(\theta+1)\} + x_2 log\{m^2 + m(1-m)/(\theta+1)\}. \quad (11)$$

For an arbitrary number of individuals, we use an iterative algorithm (Supplementary Material S1) to calculate probability p_i. The maximum likelihood estimates of θ and m are obtained by using the L-BFGS-B algorithm (41) implemented in an R optimization function *optim*. In addition to the estimates of model parameters, function *optim* outputs the hessian matrix (also called observed Fisher information matrix) that can be used to calculate the variances of the estimates.

Additionally, equation (5) implies that duplication processes occurring on the extended branches are conditionally independent of those occurring on the other branches of the tree. Thus, parameters on the extended branches can be estimated separately. Let $\{a_k; k = 1, \ldots, W\}$ and $\{b_k; k = 1, \ldots, W\}$ be the duplication and deletion rates on the W extended branches. The ratio parameter is $m_k = a_k/(a_k + b_k)$ and the branch length is $T_k^e = (a_k + b_k)t_k$. Parameters $\{m_k, T_k^e; k = 1, \ldots, W\}$ on the extended branches can be estimated from the empirical frequencies of observations 00, 01, 10 and 11 on the normal and tumor genomes of each patient. The two digits are the duplication status of the genes on the normal and tumor genomes, respectively, from the same patient. Let n_{00}, n_{01}, n_{10}, n_{11} be the count of the genes with pattern 00, 01, 10 and 11, respectively. The count n_{01} has binomial distribution with $p_{01} = m^e(1 - e^{-T^e})$ and $n_0 = n_{00} + n_{01}$, in which T^e is the length of the extended branch. Similarly, the count n_{10} has binomial distribution with probability $p_{10} = (1 - m^e)(1 - e^{-T^e})$ and $n_1 = n_{10} + n_{11}$. The maximum likelihood estimate of p_{01} is n_{01}/n_0, i.e., $p_{01} = m^e(1 - e^{-T^e}) \approx n_{01}/n_0$ and $p_{10} = (1 - m^e)(1 - e^{-T^e}) \approx n_{10}/n_1$. Thus, the estimates of m^e and T^e are given by $\hat{m}^e = \frac{n_{01}n_1}{n_{01}n_1 + n_{10}n_0}$ and $\hat{T}^e = -log\left\{1 - \frac{n_{01}}{\hat{m}^e n_0}\right\}$.

Simulation study

To evaluate the performance of the phylogenetic model developed in the previous section, duplication and deletion events were simulated from the phylogenetic model. The values of parameters (m, θ) were set to (0.01, 0.01), (0.01, 0.1), (0.3, 0.01), (0.3, 0.1), respectively. For each parameter setting, we simulated duplication and deletion events for 1000, 5000 and 10 000 genes. The simulated data were then used to estimate parameters (m, θ) in the phylogenetic model. Each simulation was repeated 10 times, and the square root of mean square error (RMSE) between the estimate and the true value of the model parameter was calculated. Let $\hat{\theta}$ be the estimate of parameter θ. The RMSE is $\sqrt{\frac{1}{g}\sum_{i=1}^{g}(\hat{\theta}_i - \theta)^2}$, where g is the number of simulations and $\hat{\theta}_i$ is the estimate of θ for the i-th simulation. Overall, the results show that the RMSEs of parameters m and θ decrease as the number of genes increases for all parameter settings (Figure 2). The RMSE of m depends on not only the value of m, but also the value of θ and the number of genes. It appears that m has a smaller RMSE when θ is large ($\theta = 0.1$) at 1000 genes (Figure 2a). But this pattern is reversed for $m = 0.3$ at 5000 and 10 000 genes. In contrast, θ consistently has a smaller RMSE when m is large (Figure 2b). This may be caused by the fact that a large m tends to generate more duplications in the simulated data. Thus it is relatively straightforward to estimate θ when m is large. For all parameter settings, the RMSEs of m and θ decrease to a reasonable level (<0.008), when the number of genes reaches 10 000.

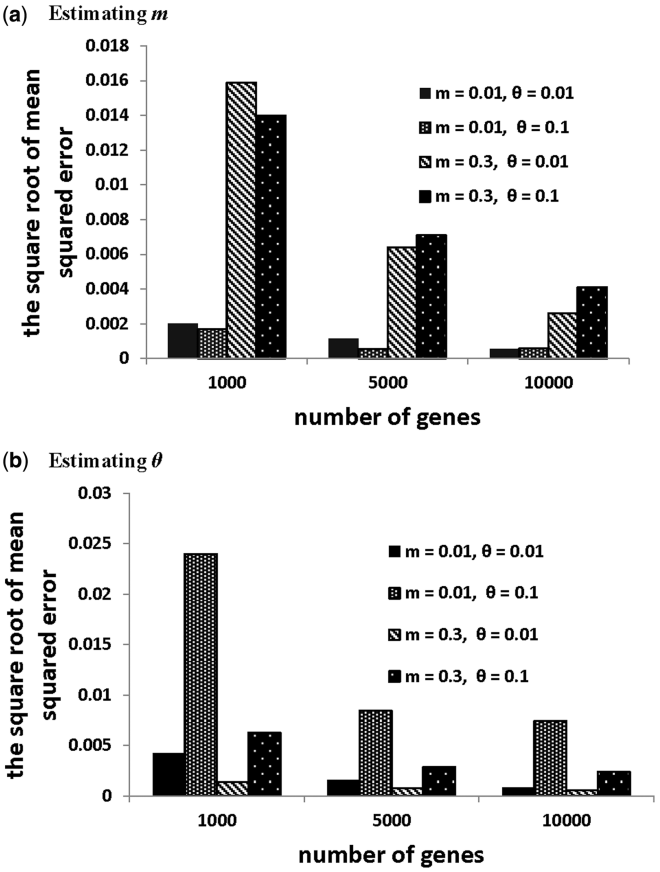

Figure 2. Simulation results. The square root of the mean square error for (**a**) estimating parameter m, and for (**b**) estimating parameter θ.

RESULTS

In the stomach cancer data set, there are 210 genes on which duplication has occurred for at least one of the 10 genomes. The remaining 20 790 genes have the pattern of (0,0,0,0, 0,0,0,0,0,0), given that the total number of genes on the human genomes is 21 000 (37). The number of duplicated genes in the tumor genomes varies across the stages of stomach cancer (Table 1). There is a high degree of individual variation in this data, although this might suggest that duplication and deletion rates may vary across different stages of stomach cancer.

We use the phylogenetic model to estimate and compare the overall changes (duplications and deletions) on the normal genomes and tumor genomes. We expect that the overall changes in the tumor genome are significantly higher than those in the normal genomes. We also investigate the pattern (increasing or decreasing) of the duplication and deletion rates across cancer stages. Finally, we identify significant duplication and deletion events associated with tumor genomes. As duplication and deletion rates depend on the total number of genes in the human genomes, the duplication and deletion rates estimated from the stomach cancer data set are relative rates. Moreover, the strategies of identifying duplication events on the genomes of five patients may underestimate the number of duplications. Because underestimation occurs for both normal and tumor genomes, it will not affect the gross conclusions based on the comparison of the relative duplication and deletion rates on normal and tumor genomes, or among the tumor genomes in different cancer stages.

We first assumed a fix tree for all genes (described in Figure 1), and used a Bayesian approach (Supplementary Material S2) to estimate the phylogenetic tree and the duplication and deletion rates. The Bayesian estimate of the phylogenetic tree is poorly supported with all posterior probabilities <0.4 (Supplementary Figure S3c). The low posterior probabilities for the nodes in the Bayesian tree, despite such a large number of observations, suggest that there is not a single tree generating the empirical data. Thus, we modeled the gene trees in the context of population genetics using the coalescent theory as described above, and calculated the maximum likelihood estimates of model parameters m and θ.

The maximum likelihood estimates of model parameters

The model parameters m and θ were estimated by maximizing the log-likelihood function described in equation (10). We used the L-BFGS-B algorithm (41) implemented in the R optimization function *optim* to maximize the log-likelihood function in equation (10). The estimate of θ for tumor genomes is twice as high as that for normal genomes (Table 2). The expected number of changes per gene for the tumor genomes is $m(1-m)\theta/(1+\theta) = 0.0022$ (see equation (9)), which is significantly higher than that (0.0012) for the normal genomes. This result suggests that the number of changes (duplications and deletions) in the tumor genomes is significantly greater than the number of changes in the normal genomes.

The goodness of fit of the phylogenetic model was evaluated by the chi-square goodness-of-fit test implemented in an R function *chisq.test*. The observed counts of genes for which 0 up to 5 individuals have duplication were calculated for the normal and tumor genomes (Table 3). Moreover, the expected count of genes for which i individuals have duplication equals $21\,000 \times p_i$, in which the probability p_i of observing i individuals with duplication was obtained from the L-BFGS-B algorithm described in the previous section. The probabilities $\{p_0, p_1, p_2, p_3, p_4, p_5\}$ for the normal genomes are 0.9942, 0.0021, 0.0011, 0.0008, 0.0007, 0.0008, respectively. The chi-square test cannot reject the phylogenetic model for the normal genomes, with P-value = 0.828 (Table 3). In contrast, the probabilities $\{p_0, p_1, p_2, p_3, p_4, p_5\}$ for the tumor genomes are 0.9903, 0.0049 0.0022, 0.0012, 0.0007, 0.0004, respectively. The phylogenetic model for tumor genomes is strongly rejected by the chi-square test, with P-value $< 10^{-6}$ (Table 3). The phylogenetic model assumes constant duplication and deletion rates across branches of the tree. However, duplication and deletion rates may be highly variable in different stages of stomach cancer, and thus the assumption of constant duplication and deletion rates may be seriously violated when modeling tumor genomes in different stages of cancer. To take into account variable duplication and deletion rates, we separately fit the two-states duplication and deletion model to each of the external branches. As we expected, duplication and deletion rates vary across external branches (Table 4). Overall, the deletion rates are much higher than the duplication rates on the extended branches (Table 4), suggesting that deletion occurred more often than duplication in tumor genomes. The duplication and

Table 2. The maximum likelihood estimates of m and θ for normal and tumor genomes

	m (SE)	θ (SE)
Normal	0.0028 (0.0002)	0.7134 (0.1024)
Tumor	0.0037 (0.0003)	1.4750 (0.1881)

The values within parentheses are standard errors of the estimates.

Table 3. The chi-square goodness-of-fit test for the phylogenetic model

No of patients with duplication	Normal		Tumor	
	Observed counts	Expected counts	Observed counts	Expected counts
0	20 885	20 880	20 803	20 798
1	46	44.6	129	103.2
2	23	24.1	26	46.4
3	18	17.8	16	26.9
4	10	15.6	7	16.4
5	18	17.6	19	8.9
	P-value = 0.8283		P-value = 7.3e-07	

Table 4. The estimates of relative duplication and deletion rates on the extended branches

	Duplication rate	Deletion rate
T1	0.0013	0.2626
T2	0.0024	0.2079
T3	0.0007	0.2412
T4	0.0009	0.1443
T5	0.0017	0.2817

deletion rates appear not to have either an increasing or decreasing pattern associated with cancer stages.

Identifying cancer-related duplicated genes

Let x denote the duplication status of a gene, with $x = 1$ referring to the cases where the gene collected from the tumor tissue is duplicated, while the gene collected from the normal tissue of the same patient is not duplicated. If duplication of the same gene is observed on a large number of tumor genomes, it is strong evidence that the duplicated gene is associated with tumor. We call this type of duplication 'cancer-related duplication' (occurring on the tumor genome, but not on the normal genome). Let y be the number of cancer-related duplications for a particular gene observed in the genomes of five patients. For the stomach cancer data, the value of y can be 0 up to 5. Under the null hypothesis that the duplication of a particular gene in the tumor genome is normal, we expect that the observed cancer-related duplication probability of a gene will be similar to the duplication probability in normal genomes. Thus, a duplicated gene is associated with cancer if the observed probability is significantly higher than the duplication probability in normal genomes. Given that $m = 0.0028$ and $\theta = 0.7134$, the average duplication probability in normal genomes is $p = m\theta/(1+\theta) = 0.0012$ (see equation (8)). Under the null hypothesis, the random variable y (number of duplications) has a binomial distribution with $P = 0.0012$ and $n = 5$. The null hypothesis was rejected for nine duplicated genes (CDH4, CLPS, CLSTN2, EML5, NPEPL1, SENP5, SPTB, VAMP7, XAGE-4), with the overall P-value < 0.05 adjusted by Bonferroni correction for multiple comparisons (Table 5). The same list of genes was identified when the duplication probability p was calculated from the two ends of the 95% confidence interval (mean $\pm$ 2SE) of m and θ. Similarly, the frequencies of deleted genes on the tumor genomes are compared with the deletion probability in normal genomes. If the observed frequency of deletions is significantly higher than the deletion probability in the normal genomes, we conclude that the deletion is significantly associated with cancer. We did not find any deletion that is significantly associate with cancer.

The functional annotation of nine significantly duplicated genes (Supplementary Table S1) was conducted by the DAVID web server (42). The analyses generated two significant annotation clusters (Supplementary Table S2). The first annotation cluster includes three genes (CDH4, CLSTN2 and NPEPL1), which are related to ion binding, specifically metal ion binding. Metal ion binding has been found to play an important role in the anticancer activity of UK-1 analogs (43). The four genes (CDH4, VAMP7, CLSTN2 and SPTB) in the second annotation cluster are mainly related to membrane or transmembrane proteins, which function as gateways to link inside and outside of a cell. Previous cancer studies suggest that membrane proteins are related to cancer progression (44), and transmembrane genes are usually quite important in drug design (45).

Table 5. Identification of duplicated genes associated with cancer

Number of duplications	Number of genes	P-value	Cumulative P-value	Significance
0	20 853	1.0	1.0	
1	109	0.006	>0.5	
2	7	1.4e-05	1e-04	*
3	1	1.7e-8	1.7e-08	*
4	0	1e-11	2.4e-15	*
5	1	2.4e-15	2.4e-15	*

The significant genes are indicated by *. The cumulative P-value was adjusted with Bonferroni correction for multiple comparisons.

DISCUSSION

Genomic data have become one of the most valuable resources of information for understanding the genetic mechanisms of cancer (22). Due to the complexity of the genomic data, it is challenging to develop a probabilistic model that can effectively extract useful information from genomic data. The genome-wide association study (GWAS) is a powerful approach for identifying cancer-related genes, based on comparison of single-nucleotide polymorphisms (SNP) in the normal and cancer genomes (46–48). The phylogenetic model developed in this article is based on the same principle to identify cancer-related duplications by comparing the normal and tumor genomes. Additionally, the phylogenetic model adds a layer of biological realism to the analysis that was otherwise not present in the GWAS analysis.

The phylogenetic model developed in this article is designed for genome-wide duplication data analysis. It has been shown through simulation that the phylogenetic method can accurately estimate the model parameters, including duplication and deletion rates. Previous studies suggest that the mechanism of cancer is complex and may involve multiple biological processes (49). For those cases, the analysis based on the phylogenetic model in which only duplication and deletion events are considered may produce biased results. In the future, we will extend the current phylogenetic model by including more biological factors (see for example, 50 in an evolutionary context). In the phylogenetic model, we assume that genes are independent of each other. This assumption may not hold, because several genes might be in the same linkage block or under selection for functional purposes. Treating genes as independent samples, while they are not, may increase the effective sample size and thus produce an estimator with an artificially smaller variance. In addition, non-independent gene trees may bias the estimates of model parameters, especially when the recombination events are highly correlated with duplication and deletion events. The effect of non-independent gene trees depends on the recombination rate of human genomes. Non-independent gene trees have been modeled for a three-taxon case (51), but it is generally quite difficult to deal with non-independent gene trees due to linkage disequilibrium. Although we do not deal with non-independent gene trees in this article, this issue clearly needs more attention.

Despite the fact that genomic data from cancer patients will become increasingly available, the high cost of sequencing whole genomes significantly limits the size of such genomic data. The data set analyzed in this article contains genomes from only five patients, one or two patients for each stage of stomach cancer. We expect that the availability of multiple genomes from more patients (along with the actual number of gene copies for each gene) will significantly improve the estimation of model parameters and increase the power for testing relevant biological hypotheses about the mechanisms of cancer under the phylogenetic model.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Guojun Li for the helpful discussions on the phylogenetic model. We thank Prof. Madan Babu and two anonymous reviewers for the thoughtful comments.

FUNDING

National Science Foundation Grant [DMS-1222745] to Dr Liu and National Science Foundation Grant [DMS-1222940] to Dr Liberles. Funding for open access charge: National Science Foundation Grant [DEB-0830024] and the DOE BioEnergy Science Center [contract no. DE-PS02-717 06ER64304] [DOE 4000063512].

Conflict of interest statement. None declared.

REFERENCES

1. Siegel,R., Ward,E., Brawley,O. and Jemal,A. (2011) Cancer statistics. *A Cancer J. Clinicians*, **61**, 212–236.
2. Brosnan,J.A. and Iacobuzio-Donahue,C.A. (2012) A new branch on the tree: next-generation sequencing in the study of cancer evolution. *Semin. Cell Dev. Biol.*, **23**, 237–242.
3. Iacobuzio-Donahue,C.A. (2012) Genetic evolution of pancreatic cancer: lessons learnt from the pancreatic cancer genome sequencing project. *Gut*, **61**, 1085–1094.
4. Ma,Q.C., Ennis,C.A. and Aparicio,S. (2012) Opening Pandora's Box—the new biology of driver mutations and clonal evolution in cancer as revealed by next generation sequencing. *Curr. Opin. Genet. Dev.*, **22**, 3–9.
5. Jones,S., Chen,W.-d., Parmigiani,G., Diehl,F., Beerenwinkel,N., Antal,T., Traulsen,A., Nowak,M.A., Siegel,C., Velculescu,V.E. *et al.* (2008) Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA*, **105**, 4283–4288.
6. Prisman,E.Z., Gafni,A. and Finelli,A. (2011) Testing the evolution process of prostate-specific antigen in early stage prostate cancer: what is the proper underlying model? *Stat. Med.*, **30**, 3038–3049.
7. Ayala,F.J. (1977) 'Nothing in biology makes sense except in the light of evolution': Theodosius Dobzhansky: 1900-1975. *J. Hered.*, **68**, 3–10.
8. Greaves,M. and Maley,C.C. (2012) Clonal evolution in cancer. *Nature*, **481**, 306–313.
9. Erren,T.C. (2009) On the origin of cancer: evolution and a mutation paradox. *Med. Hypotheses*, **73**, 124–125.
10. Aktipis,C.A., Kwan,V.S.Y., Johnson,K.A., Neuberg,S.L. and Maley,C.C. (2011) Overlooking evolution: a systematic analysis of cancer relapse and therapeutic resistance research. *Plos One*, **6.**
11. Calcagno,A.M. (2011) Evolution of drug resistance in cancer: the emergence of unique mechanisms and novel techniques. *Mol. Pharmaceutics*, **8**, 1993–1993.
12. Gillies,R.J., Verduzco,D. and Gatenby,R.A. (2012) Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nat. Rev. Cancer*, **12**, 487–493.
13. Goymer,P. (2008) Natural selection: the evolution of cancer. *Nature*, **454**, 1046–1048.
14. Allred,D.C., Wu,Y., Mao,S., Nagtegaal,I.D., Lee,S., Perou,C.M., Mohsin,S.K., O'Connell,P., Tsimelzon,A. and Medina,D. (2008) Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. *Clin. Cancer Res.*, **14**, 370–378.
15. Ewald,P.W. and Ewald,H.A.S. (2012) Infection, mutation, and cancer evolution. *J. Mol. Med.*, **90**, 535–541.
16. Gerlinger,M., Rowan,A.J., Horswell,S., Larkin,J., Endesfelder,D., Gronroos,E., Martinez,P., Matthews,N., Stewart,A., Tarpey,P. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, **366**, 883–892.
17. Greaves,M. (2000) *Cancer: The Evolutionary Legacy*. Oxford University Press, Oxford.
18. Yu,K.-D. and Shao,Z.-M. (2012) Initiation, evolution, phenotype and outcome of BRCA1 and BRCA2 mutation-associated breast cancer. *Nat. Rev. Cancer*, **12**, 372–373.
19. Muto,T., Bussey,H.J. and Morson,B.C. (1975) The evolution of cancer of the colon and rectum. *Cancer*, **36**, 2251–2270.
20. Merlo,L.M.F., Pepper,J.W., Reid,B.J. and Maley,C.C. (2006) Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*, **6**, 924–935.
21. Otsuka,J. (2011) The large-scale evolution by generating new genes from gene duplication; similarity and difference between monoploid and diploid organisms. *J. Theor. Biol.*, **278**, 120–126.
22. Podlaha,O., Riester,M., De,S. and Michor,F. (2012) Evolution of the cancer genome. *Trends Genet.*, **28**, 155–163.
23. Nowell,P.C. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23–28.
24. Gillies,R.J., Verduzco,D. and Gatenby,R.A. (2012) Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nat. Rev. Cancer*, **12**, 487–493.
25. Wu,X., Northcott,P.A., Dubuc,A., Dupuy,A.J., Shih,D.J.H., Witt,H., Croul,S., Bouffet,E., Fults,D.W., Eberhart,C.G. *et al.* (2012) Clonal selection drives genetic divergence of metastatic medulloblastoma. *Nature*, **482**, 529–533.
26. Clifford,S.C. (2012) Cancer genetics: evolution after tumour spread. *Nature*, **482**, 481–482.
27. Hou,Y., Song,L., Zhu,P., Zhang,B., Tao,Y., Xu,X., Li,F., Wu,K., Liang,J., Shao,D. *et al.* (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, **148**, 873–885.
28. Xu,X., Hou,Y., Yin,X., Bao,L., Tang,A., Song,L., Li,F., Tsang,S., Wu,K., Wu,H. *et al.* (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**, 886–895.
29. Caldas,C. (2012) Cancer sequencing unravels clonal evolution. *Nat. Biotechnol.*, **30**, 408–410.
30. Thomas,R.K., Nickerson,E., Simons,J.F., Janne,P.A., Tengs,T., Yuza,Y., Garraway,L.A., LaFramboise,T., Lee,J.C., Shah,K. *et al.* (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat. Med.*, **12**, 852–855.
31. Ciullo,M., Debily,M.A., Rozier,L., Autiero,M., Billault,A., Mayau,V., El Marhomy,S., Guardiola,J., Bernheim,A., Coullin,P. *et al.* (2002) Initiation of the breakage-fusion-bridge mechanism through common fragile site activation in human breast cancer cells: the model of PIP gene duplication from a break at FRA7I. *Hum. Mol. Genet.*, **11**, 2887–2894.
32. Waris,G. and Ahsan,H. (2006) Reactive oxygen species: role in the development of cancer and various chronic conditions. *J. Carcinog.*, **5**, 14.
33. Moelans,C.B., de Weger,R.A., Monsuur,H.N., Vijzelaar,R. and van Diest,P.J. (2010) Molecular profiling of invasive breast cancer by multiplex ligation-dependent probe amplification-based

copy number analysis of tumor suppressor and oncogenes. *Mod. Pathol.*, **23**, 1029–1039.
34. Cui,J., Yin,Y., Ma,Q., Wang,G., Olman,V., Zhang,Y., Chou,W.-C., Hong,C.S., Zhang,C., Cao,S. *et al.* (2013) Towards Understanding the Genomic Alterations in Human Gastric Cancer. *PLoS Genet.*, under review.
35. Kim,J.Y., Shin,N.R., Kim,A., Lee,H.J., Park,W.Y., Lee,C.H., Huh,G.Y. and Park do,Y. (2013) Microsatellite instability status in gastric cancer: a reappraisal of its clinical significance and relationship with mucin phenotypes. *Korean J. Pathol.*, **47**, 28–35.
36. Ottini,L., Falchetti,M., Lupi,R., Rizzolo,P., Agnese,V., Colucci,G., Bazan,V. and Russo,A. (2006) Patterns of genomic instability in gastric cancer: clinical implications and perspectives. *Ann. Oncol.*, **17(Suppl 7)**, vii97–vii102.
37. Pennisi,E. (2012) Genomics. ENCODE project writes eulogy for junk DNA. *Science*, **337**, 1159–1161.
38. Hahn,M.W., De Bie,T., Stajich,J.E., Nguyen,C. and Cristianini,N. (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.*, **15**, 1153–1160.
39. Liu,L., Yu,L., Kalavacharla,V. and Liu,Z. (2011) A Bayesian model for gene family evolution. *BMC Bioinformatics*, **12**, 426.
40. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
41. Byrd,R.H., Lu,P.H., Nocedal,J. and Zhu,C.Y. (1995) A limited memory algorithm for bound constrained optimization. *Siam. J. Sci. Comput.*, **16**, 1190–1208.
42. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
43. McKee,M.L. and Kerwin,S.M. (2008) Synthesis, metal ion binding, and biological evaluation of new anticancer 2-(2′-hydroxyphenyl)benzoxazole analogs of UK-1. *Bioorganic Med. Chem.*, **16**, 1775–1783.
44. Kampen,K.R. (2011) Membrane proteins: the key players of a cancer cell. *J. Membrane Biol.*, **242**, 69–74.
45. Zaman,G.J., Versantvoort,C.H., Smit,J.J., Eijdems,E.W., de Haas,M., Smith,A.J., Broxterman,H.J., Mulder,N.H., de Vries,E.G., Baas,F. *et al.* (1993) Analysis of the expression of MRP, the gene for a new putative transmembrane drug transporter, in human multidrug resistant lung cancer cell lines. *Cancer Res.*, **53**, 1747–1750.
46. Klein,R.J., Zeiss,C., Chew,E.Y., Tsai,J.Y., Sackler,R.S., Haynes,C., Henning,A.K., SanGiovanni,J.P., Mane,S.M., Mayne,S.T. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.
47. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
48. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.
49. Desmedt,C., Haibe-Kains,B., Wirapati,P., Buyse,M., Larsimont,D., Bontempi,G., Delorenzi,M., Piccart,M. and Sotiriou,C. (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.*, **14**, 5158–5165.
50. Konrad,A., Teufel,A.I., Grahnen,J.A. and Liberles,D.A. (2011) Toward a general model for the evolutionary dynamics of gene duplicates. *Genome Biol. Evol.*, **3**, 1197–1209.
51. Slatkin,M. and Pollack,J.L. (2006) The concordance of gene trees and species trees at two linked loci. *Genetics*, **172**, 1979–1984.