Measuring the amount of genomic instability during early development of Human tumors

Mathieu Emily Olivier François

April 6, 2005

TIMC-TIMB Department
Institut de l'Ingénierie de l'Information de Santé
Faculty of Medicine
38706 La Tronche
France

RH: Measuring the amount of genomic instability

Corresponding author:

TIMC-TIMB Department

Institut de l'Ingénierie de l'Information de Santé

Faculty of Medicine

38706 La Tronche

France

phone: +33 (0)4 56 52 00 25

email: olivier.francois@imag.fr

Keywords: Genomic instability, conditional coalescent genealogies, nucleotide polymorphism, nucleotide diversity.

Abstract

Humans have invested several genes in DNA repair and fidelity replication. To account for the disparity between the rarity of mutations in normal cells and the large numbers of mutations present in cancer, an hypothesis is that cancer cells must exhibit a mutator phenotype (genomic instability) during tumor progression with the initiation of abnormal mutation rates caused by the loss of mismatch repair. In this study we introduce a stochastic model of mutation in tumor cells with the aim of estimating the amount of genomic instability due to the alteration of DNA repair genes. Our approach took into account the difficulties generated by sampling within tumoral clones, and the fact that these clones must be difficult to isolate. We provide corrections to two classical statistics in order to obtain unbiased estimators of the raised mutation rate, and we show that large statistical errors may be associated with such estimators. The power of these new statistics to reject genomic instability is assessed and proved to increase with the intensity of mutation rates. In addition we show that genomic instability can hardly be detected unless the raised mutation rates exceed the normal rates by a factor at least equal to one thousand.

INTRODUCTION

DNA replication in normal human cells is an extremely accurate process. During the cell division cycle, copy errors occur with probabilities less than $10^{-9} - 10^{-10}$ per nucleotide. In contrast, the malignant cells that constitute cancer tissues are markedly heterogeneous and exhibit alterations in the nucleotide sequence of DNA (e.g., Bielas and Loeb, 2005). To account for the disparity between the rarity of mutations in normal cells and the large numbers of mutations present in cancer, Loeb *et al.* (1974) hypothesized that during tumor progression, cancer cells must exhibit a *mutator phenotype* (see the review by Loeb *et al.*, 2003). It is still a matter of debate to know exactly which event initiates tumorigenesis. But one hypothesis for the initiation of abnormal mutation rates in tumors is the loss of mismatch repair (MMR).

For instance, this phenomenon may follow from the inactivation of the genes hMSH2 and hMLH1 involved in hereditary nonpolyposis colorectal cancers (HNPCC) (Fishel *et al.*, 1993; Leach *et al.*, 1993; Lindblom *et al.*, 1993). In normal conditions, the MMR repair system involves a complex interaction among the protein products of hMSH2 and hMLH1 genes. The result is to eliminate about 99.9% of the errors in DNA replication reducing errors to a rate of about 1 per 10^{12} bp in genes that regulate the apoptosis or the cell cycle duration. HNPCC is inherited in an autosomal dominant fashion. One copy of the mutant allele is defective, and is inherited in the germline. The loss of MMR may start when the second mutation occurs somatically as a consequence of the two-hits theory (Moolgavkar and Knudson, 1981).

Widespread genomic instability seems associated with MMR defective genes. For instance microsatellite instability is associated with HNPCC (Ionov *et al.*, 1993; Peltomaki *et al.*, 1993; Thibodeau *et al.*, 1993). Detection of DNA instability

is therefore a crucial step in view of non-invasive diagnosis of such forms of cancer. Because numerous mutations are required for the full development of cancer, inactivation of *caretaker* genes can greatly accelerate its development (KINZLER and VOGELSTEIN, 2002). For an account of the etiology and genetic epidemiology of cancer with a statistical perspective a major review is by THOMAS (2004).

This study introduces a two-rates model of DNA mutation based on the infinitely-many sites model (WATTERSON, 1975). We consider a sample of n sequences taken from a pretumoral tissue, and assume that loss of DNA repair has occurred once (and only once) during the history of the n sequences tracking back to their most recent common ancestor. We denote the mutational event by the formal symbol Δ. The event Δ is assumed to occur at a very low rate δ.

The loss of MMR (occurrence of Δ) may lead to a ten to thousand fold increase in the normal mutation rate μ_0 (SHIBATA *et al.*, 1994; BHATTACHARYYA *et al.*, 1994). However only the sequences that descend from Δ are concerned with such an increase in the mutation rate. Because heterogeneity prevails in cancer tissues and sampling from the tumor is difficult, we consider that an unknown random number of sequences among the sample descend from the mutation Δ.

The goal of this study is to provide statistical estimators for the raised mutation rate μ_1 under the assumption that the normal rate μ_0 is known, but the number of descendants of Δ is unknown. Two classical statistics will be studied (see HARTL and CLARK, 1997) for a review in a population genetics context). The first one is the *nucleotide polymorphism* computed as the number of segregating sites in the DNA. The second one is the *nucleotide diversity* computed as the number of pairwise nucleotide differences. Our main contribution is the calculation of corrections to the classical statistics that are needed because the increase in the mutation rate concerns only a random subgenealogy of the sample.

In our study, the clonal evolution of mitotic cell divisions is assumed to be neu-

tral. This hypothesis is often competing with the hypothesis of selective evolution (Cairn, 1975; Nowell, 1976). Under the neutral assumption, we shall model the genealogies of DNA sequences using conditional coalescent trees (Tavaré, 2004; Griffiths and Tavaré, 2003; Wiuf and Donnelly, 1999). This formalism has been developed for the primary purpose of estimating the age of an allele (Griffiths and Tavaré, 1998; Stephens, 2000). So far evolutionary models have been introduced for dating the loss of MMR (Tsao *et al.*, 2000; Calabrese *et al.*, 2004). Tsao *et al.* (2000) observed microsatellite alleles in noncoding regions assuming neutrality as well. However, the need for further mathematical studies has been emphasized in a recent review to better understand the influence of existing hypotheses in the evolution of cancer (Michor *et al.*, 2004).

In the next section, we recall notations and give an account on the existing results in the actual theory of conditional trees. In addition, we extend many results of the theory so as to encompass other times or ages useful in the context of genomic instability, and describe an efficient way for simulating conditional trees. In the Nucleotide polymorphism section and the Nucleotide diversity section, we introduce unbiased estimators of the raised mutation rate μ_1 based on the number of segregating sites and the number of pairwise differences within the sample. The statistical errors and the power of tests based on these estimators are then compared using Monte Carlo methods.

CONDITIONAL COALESCENT TREES

Model and Notations: We consider a sample of n copies of a gene at a particular DNA locus taken from a pretumoral tissue, and assume that the loss of MMR (event Δ) occurred once in the sample history. However the date and place at which this event occurred in the sample genealogy are unknown. Mathematically, we consider taking the limit as the rate of occurrence δ tends to zero conditional on Δ having

occurred. In further statements the symbol = will therefore often replace the limit symbol as δ goes to zero.

The sample is divided in two random complementary subsamples $\mathcal{B}$ and $\mathcal{C}$. The cardinality of $\mathcal{B}$ is a random variable denoted by B. Given the number $B = b$ of sequences in $\mathcal{B}$, the number of sequences in $\mathcal{C}$ is then equal to $c = n - b$. As usual in studies of conditional coalescent trees, the analysis requires two levels of conditioning. At the first level, the sample has the property that all sequences in $\mathcal{B}$ are descendants of the particular mutation Δ while none of those in $\mathcal{C}$ are. This property is called the *topological event*, and is denoted by E. At the secont level, we assume that the mutation Δ arised only once in the history of the sample. We denote this event by M. Conditioning on E impacts the random topology of the tree, while conditioning on M affects branch lengths. In the terminology of Tavaré (2004), conditioning on $E \cap M$ amounts to considering an unique event polymorphism in the tree. Stephens (2000) showed that the probability distribution of B can be described as

$$\mathrm{P}(B = b \mid E \cap M) = \frac{1}{bH_{n-1}}, \quad b = 1, \ldots, n-1,$$

where H_{n-1} denotes the $(n-1)$th harmonic number. This is a *Yule* distribution also called the *frequency spectrum* (see Griffiths and Tavaré, 2003).

Under the neutral hypothesis, we assume that lineages coalesce at random, and time is rescaled so that the unit of time corresponds to N generations with N the total cell population size (Kingman, 1982). In this setting, the normal mutation rate is usually rescaled so that to $\theta_0/2 = 2N\mu_0$ and the raised mutation rate is $\theta_1/2 = 2N\mu_1$. Conditioning on $B = b$ leads to a model of genealogies which we refer to as the *conditional coalescent tree* (Wiuf and Donnelly, 1999, see Figure 1). All subsequent results will be established conditional on the event $E \cap M$, but at the exception of the appendix section we omit this condition in order to alleviate

notations in long formulae.

Recalls: In order to state results about conditional coalescent times, some additional recalls are needed. As far as possible, we use notations similar to those of Tavaré (2004) and Wiuf and Donnelly (1999). For $r = 1, \ldots, b-1$, we define J_r to be the total number of ancestors at the time the subsample $\mathcal{B}$ first has r ancestors. This definition implies that J_r ranges between $(r+1)$ and $(n-b+r)$. In addition, we denote by J_0 the number of ancestors in the sample at the time the $\mathcal{B}$ lineages first coalesce with the rest of the sample. This means that we have

$$1 \leq J_0 < J_1 < \ldots < J_{b-1} < J_b \equiv n\,.$$

Similarly, we consider K_r to be the total number of ancestors at the time the subsample $\mathcal{C}$ first has r ancestors. We have

$$K_1 < K_2 < \ldots < K_{c-1} < K_c \equiv n$$

where the subset $\mathcal{B}$ is replaced by $\mathcal{C}$ in the previous definition, and the K_r's are complementary to the $J_r's$ in the set of labels $[1, n]$. Note that conditional on $J_0 = j$, we have $K_r = r$ for all $r < j$ and $j + 1 \leq K_j$. To finish, we denote by J_Δ the total number of ancestors in the sample at the time the mutation Δ occurs. This implies that J_Δ takes its values between 2 and $n-b+1$. A picture of a tree with a summary of notations is displayed in Figure 2.

The conditional joint distributions of the J_r's given the events E or $E \cap M$ were described in Tavaré (2004, Chapter 8, p. 106-109) for which we shall refer to when necessary. For example, we easily deduce that

$$\mathrm{P}(J_r = j_r\,;\, r = 1, \ldots, b-1 \mid J_0 = j\,;\, E \cap M) = \left(\begin{array}{c} n-j-1 \\ b-1 \end{array} \right)^{-1} \qquad (1)$$

for all $j < j_1 < \ldots < j_{b-1} < n$. This result will be useful in the NUCLEOTIDE DIVERSITY section. Similar properties will be stated without proofs when they are direct consequences of Tavaré's notations.

Another useful result concerns the number of ancestors in the sample at the time when the mutation Δ occurs. Recall that we have

$$p_k^\Delta \equiv \mathrm{P}(J_\Delta = k \mid E \cap M) = \frac{\left(\begin{array}{c} n-k \\ b-1 \end{array} \right)}{\left(\begin{array}{c} n-1 \\ b \end{array} \right)} \tag{2}$$

for all $k = 2, \ldots, n-b+1$.

The age of the mutation Δ has been studied by STEPHENS (2000), WIUF and DONNELLY (1999), GRIFFITHS and TAVARÉ (1998). Conditional on $B = b$, the expected age is given by

$$\tau_\Delta = 2 \sum_{k=2}^{n-b+1} \frac{n-k+1}{n(k-1)} p_k^\Delta \,. \tag{3}$$

The distribution of intercoalescence times: In the standard coalescent, the durations X_ℓ that separate coalescence events backward in time are independent random variables, and have exponential distribution of rate $\lambda_\ell = \ell(\ell-1)/2$ where ℓ is the number of ancestors just before the event. In this section, we show how the conditioning on $B = b$ and the existence of an unique event polymorphism $E \cap M$ further modify the shape of the genealogy by lengthening the intercoalescence times.

Theorem 1 *Assume that the mutation Δ has $B = b$ descendants. The joint probability distribution of $(X_2, \ldots, X_n)$ conditional on the event $E \cap M$ has density equal to*

$$f(x_2, \ldots, x_n) = \sum_{k=2}^{n-b+1} p_k^\Delta \lambda_k x_k \prod_{\ell=2}^{n} f_\ell(x_\ell) \tag{4}$$

where $f_\ell(x_\ell)$ is the probability density function of the exponential distribution of rate λ_ℓ.

As a consequence of Theorem 1 we have the following result.

Corollary 1 *Assume that the mutation Δ has $B = b$ descendants. Let $\ell = 2, \ldots, n$. Then we have*

$$\mathbf{E}[X_\ell \mid E \cap M] = \begin{cases} \left(1 + p_\ell^\Delta\right)/\lambda_\ell & \text{if } \ell \leq n - b + 1 \\ 1/\lambda_\ell & \text{otherwise.} \end{cases} \tag{5}$$

As a consequence of Theorem 1, note that conditional on the event $E \cap M$ the X_ℓ's are no longer independent random variables. However Theorem 1 has the nice interpretation that once we know that the number of ancestors is equal to k at the time Δ occurs, then X_k has gamma $G(2, \lambda_k)$ distribution, the other X_ℓ have exponential $G(1, \lambda_\ell)$ distribution, and the variables are mutually independent. This remark is useful for simulating conditional trees given that $B = b$. The algorithm is as follows.

1. Draw $J_\Delta = k$ according to the distribution (p_k^Δ) for $k = 2, \ldots, n - b + 1$
2. Draw J_0 from the conditional distribution
$$\mathrm{P}(J_0 = j \mid J_\Delta = k\,;\, E \cap M) = \frac{2j}{k(k-1)}, \quad j = k-1, \ldots, 1.$$
3. Draw an ordered sequence $k \leq J_1 < \ldots < J_{b-1} < n$ uniformely from the set of ordered integral sequences $\mathcal{I}_b(k, n) = \{k \leq j_1 < \ldots < j_{b-1} < n\}$
4. Fill the holes left in $[1, n]$ by the J_r's with the K_r's
5. Sample X_k from the gamma $G(2, \lambda_k)$ distribution, otherwise sample X_ℓ from the exponential distribution $G(1, \lambda_\ell)$, for $\ell \neq k$.

Testing for the absence of Δ: This paragraph presents a brief study of the power of a rather "abstract" test to reject the null hypothesis H_0 of absence of the mutation Δ against the alternative hypothesis H_1 of its existence. The test is abstract because it assumes the knowledge of the sample genealogy, and the dataset consists of the all the intercoalescence times (X_k). Under the null hypothesis we assume that the property E holds for a specific subsample of b sequences. In the alternative hypothesis we assume that the mutation Δ has $B = b$ descendants as well. The test statistic consists of the ratio of likelihoods which is believed to behave optimally for reasonably large sample sizes. It can be described as

$$r = \frac{L(x, \mathrm{H}_1)}{L(x, \mathrm{H}_0)} = \sum_{k=2}^{n-b+1} \lambda_k p_k^{\Delta} x_k \, .$$

Under H_0, we see that this ratio has the same distribution as a sum of independent exponential random variables of rates $\nu_k = 1/p_k^{\Delta}$

$$Y = \sum_{k=2}^{n-b+1} \mathcal{E}(\nu_k) \, , \tag{6}$$

whereas under H_1 it is distributed as Y plus a sum of independent exponential random variables of rates ν_k^2,

$$Z = Y + \sum_{k=2}^{n-b+1} \mathcal{E}(\nu_k^2). \tag{7}$$

The criteria for rejection is r greater than the 0.95th percentile from neutral data sets (Equation 6). The power of the test was studied numerically from 10,000 replicates of Y and Z. We found that the power did not exceed a value close to 0.2 for $n = 10, 20, 50, 100$, and $b \approx n$. For smaller b's, the lack in power was even more striking. For example the power dropped to ≈ 0.1 for $b/n \approx 0.5$.

Because we assume the ideal knowledge of tree topologies and branch lengths, the interest in these power calculations is more theoretical than directed toward applications. However these results put some limitations in testing for the occurrence

of the mutation Δ. They are evidence that the occurrence of Δ alone conveys too weak information for being detected by any kind of statistical testing even if the full genealogy were observed. This could be explained as the shapes of such trees do not undergo significant changes under the occurrence of Δ.

NUCLEOTIDE POLYMORPHISM

Corrected estimator: We now take account of the mutations that are superimposed to the conditional coalescent trees. Mutations on the tree branches are distributed according to Poisson processes of rates $\theta_0/2$ or $\theta_1/2$ depending on where Δ takes place. Assuming the infinitely-many sites model of the DNA molecule, we introduce an unbiased estimator of θ_1 based on the number of segregating sites S. This variable equals the number of mutations that occurred during the sample history back to the most recent common ancestor of the sample. In the classical coalescent, S has Poisson distribution of parameter $L_n\theta/2$ where θ is the mutation rate, and L_n is the length of the genealogy. The nucleotide polymorphism or Watterson's estimator is defined as $\hat{\theta} = S/H_{n-1}$ (Watterson, 1975). It is an unbiased estimator of θ with the property that

$$\mathrm{Var}[\hat{\theta}] = \frac{1}{H_{n-1}^2} \sum_{i=1}^{n-1} \left(\frac{\theta^2}{i^2} + \frac{\theta}{i} \right).$$

In analogy with the classical approach, we denote by L_n^{Δ} the length of the genealogy of the full sample and by L_n^1 the length of the subgenealogy of $\mathcal{B}$. Borrowing the notation from Wiuf and Donnelly (1999), we also denote by η_n the time separating the root of the subgenealogy from the mutation Δ. In the two-rates model, the number of segregating sites can be splitted in two independent terms

$$S = S^0 + S^1,$$

where S^1 has Poisson distribution of rate $(L_n^1 + \eta_n)\theta_1/2$ and S^0 has Poisson distribution of rate $(L_n^\Delta - L_n^1 - \eta_n)\theta_0/2$. Taking expectations, we obtain the expected number of segregating sites as

$$\mathbf{E}[S] = A_n\theta_0 + B_n\theta_1$$

where

$$B_n = \frac{1}{2}\left(\mathbf{E}[L_n^1] + \mathbf{E}[\eta_n]\right),$$

and

$$A_n = \frac{1}{2}\mathbf{E}[L_n^\Delta] - B_n.$$

Accordingly, an unbiased estimator of θ_1 can be defined as follows

$$\hat{\theta_1} = \frac{S - A_n\theta_0}{B_n}.$$

Table 1 and Figures 3-4 provide numerical values for A_n and B_n with sample sizes in the range $5-50$. Exact formulae are derived afterwards. First of all the expectation $\mathbf{E}[L_n^\Delta]$ results from corollary 1 as follows

$$\frac{1}{2}\mathbf{E}[L_n^\Delta] = H_{n-1} + \frac{1}{H_{n-1}}\sum_{b=1}^{n-1}\sum_{k=2}^{n-b+1}\frac{p_k^\Delta}{b(k-1)}$$

Given that the mutation Δ has b descendants $(B = b)$, the conditional expectations involved in the computation of A_n and B_n can be obtained thanks to the results of Griffiths and Tavaré (2003) and Wiuf and Donnelly (1999). On the one hand Griffiths and Tavaré (2003) proved that

$$\mathbf{E}[L_n^1 \mid B = b] = \sum_{j=2}^{n-b+1} p_j^\Delta \sum_{k=j+1}^{n}\frac{2}{k(k-1)}c_{jk}$$

where

$$c_{jk} = b - (b-1)\frac{n-k}{n-j} - \frac{(n-k)!(n-j-b+1)!}{(n-j)!(n-k-b+1)!}$$

for $j = 2, \ldots, n-b+1$ and $k = j+1, \ldots, n$. On the other hand Wiuf and Donnelly (1999) showed that

$$\mathbf{E}[\eta_n \mid B = b] = 2 \sum_{k=2}^{n-b+1} \frac{p_k^{\Delta}}{k}, \quad b = 1, \ldots, n-1.$$

The values of A_n and B_n can then be computed by integration over all b's.

Statistical errors and power of tests: In the first half of this paragraph, we evaluate the standard deviation (SD) of the estimator $\hat{\theta}_1$. The exact computation of $\mathrm{Var}[\hat{\theta}_1]$ appears intricate enough so that we resort to Monte Carlo methods. In a second half, we evaluate the power of the statistic $\hat{\theta}_1$ to reject the hypothesis that the mutation rate increases simultaneously with the occurrence of the mutation Δ. Simulations were performed using the R statistical programming language (R Development Core Team, 2004).

Statistical errors: For evaluating statistical errors, the following experimental design was used. The parameter θ_0 was set equal to the value $\theta_0 = 1$. Roughly, this corresponded to a normal mutation rate per mitotic division equal to $\mu_0 \approx 10^{-10}$, while the total number of cells N in the tissue approximated 2.5 billions. We considered three different values for the raised mutation rate $\theta_1 = 10, 10^2, 10^3$, and the sample sizes were taken in the range $n = 10 - 50$. Simulations were performed using the method described in the previous section. Table 2 gives the bias and the standard deviation computed over 10,000 replicates. These results confirms that $\hat{\theta}_1$ was indeed unbiased. Nevertheless, the standard deviations were rather high. This could be explained as the empirical distributions exhibited strong positive skew. In addition, most of the error was contributed by a term which seemed proportional to θ_1^2. For $n = 20$, we adjusted a regression model of the form $a_n\theta_1 + b_n\theta_1^2$ to the variance, and an almost perfect fit was obtained as $\mathrm{Var} = 1.47\theta_1^2$ ($R^2 = 0.999$, $P < 10^{-12}$). For $n = 40$, we obtained $\mathrm{Var} = 1.68\theta_1^2$ ($R^2 = 0.997$, $P < 10^{-12}$).

Apparently, SDs did not exhibit fast decrease as sample sizes increased. This might be due to a strong correlation of data within the subsample $\mathcal{B}$, and the fact that the most recent ancestor of this subsample is expected to be recent. Note that the shape of the correcting constant B_n suggested a logarithmic rate of decrease of errors toward zero. This was actually hardly sensitive for the sample sizes used there.

Powers: A fundamental assumption through this work is that the mutation Δ has occurred once in the history of the sample. Assuming a normal mutation rate θ_0, we report results regarding the power of the test based on $\hat{\theta_1}$ to reject the null hypothesis of absence of Δ against the alternative of its existence together with an increase in mutation rate $\theta_1 > \theta_0$. Results for $\theta_0 = 1$ and $\theta_1 = 10 - 10^3$ are given in Table 3. Power values ranged from ≈ 0.06 to ≈ 0.90. Reasonable powers were obtained for θ_1 greater than $10^3\theta_0$. No significant improvements were observed when the sample sizes varied from $n = 10$ to $n = 50$.

In a second step we reverted the role of the null and alternative hypotheses, and used a test based on $\hat{\theta}$. The results are reported in Table 4. In this Table, powers range from ≈ 0.43 to ≈ 0.90. For θ_1 lower than 10, the test exhibited performances similar to those presented in the previous section where the simultaneous rise in mutation rate was ignored. Significant gains in power were obtained for $\theta_1 = 10^3\theta_0$. Increasing the sample sizes did not provide additional benefit. Table 4 indicates that the event Δ was more easily detected when associated with large mutation rates and small sample sizes. However the power to detect Δ remains small for θ_1 lower than $1,000\theta_0$.

NUCLEOTIDE DIVERSITY

Corrected estimator: This section introduces an unbiased estimator of θ_1 based on the nucleotide diversity Π. In the infinitely many sites model the nucleotide diversity is defined as the mean number of pairwise differences between nucleotides.

Let $\Pi(i,j)$ be the number of sites at which the sequence i differs from the sequence j, for $1 \leq i \leq n$ and $1 \leq j \leq n$. The nucleotide diversity is the average value of $\Pi(i,j)$. It can be computed as follows

$$\Pi = \frac{1}{n(n-1)} \sum_{i \neq j} \Pi(i,j)$$

In the unconditional coalescent, we have $\mathbf{E}[\Pi(1,2)] = \theta \mathbf{E}[X_2]$, and Π is an unbiased estimator of θ. The variance of Π is equal to $\mathrm{Var}[\Pi] = (n+1)\theta/3(n-1) + 2(n^2 + n + 3)\theta^2/9n(n-1)$ (Tajima, 1983).

Now consider the occurrence of Δ and the two rates of mutation θ_0 and θ_1. Again, we assume that the mutation Δ has $B = b$ descendants. Consider two arbitrary sequences labelled 1 and 2. In the classical coalescent, $\mathbf{E}[X_2]$ is the expected coalescence time of sequences 1 and 2. In analogy with this, the computation of $\mathbf{E}[\Pi(1,2)]$ requires distinguishing three cases. In the first case, sequences 1 and 2 both belong to $\mathcal{B}$, and we have

$$\mathbf{E}[\Pi(1,2)] = \tau_{\mathcal{B}} \theta_1$$

where $\tau_{\mathcal{B}}$ is the expected coalescence time within $\mathcal{B}$. This case occurs with a probability equal to $(b/n)^2$. In the second case, one sequence is in $\mathcal{B}$ while the other belongs to $\mathcal{C}$. This event occurs with probability $2b(n-b)/n^2$, and we have

$$\mathbf{E}[\Pi(1,2)] = (2\tau_{\mathcal{B},\mathcal{C}} - \tau_\Delta)\theta_0 + \tau_\Delta \theta_1$$

where $\tau_{\mathcal{B},\mathcal{C}}$ is the expected coalescence time of sequence 1 and sequence 2, and τ_Δ is the age of Δ given in Equation (3). The third case occurs with probability $(1-b/n)^2$. It corresponds to the situation where both sequence 1 and sequence 2 are in $\mathcal{C}$. Then we have

$$\mathbf{E}[\Pi(1,2)] = \tau_{\mathcal{C}} \theta_0$$

where $\tau_{\mathcal{C}}$ is the corresponding expected coalescence time. Taking expectation with respect to B, we deduce that

$$\mathbf{E}[\Pi(1,2)] = C_n\theta_0 + D_n\theta_1$$

where the constants C_n and D_n can be computed from the above defined coalescence times. Therefore, an unbiased estimator Π_1 of θ_1 is of the form

$$\Pi_1 = \frac{\Pi - C_n\theta_0}{D_n}$$

Table 5 and Figures 5-6 give numerical values for C_n and D_n for n in the range $10 - 50$. The next section explains the way by which the exact computations of all coalescence times can be achieved.

Coalescence times: In this section, we provide explicit ways of computing the coalescence times $\tau_{\mathcal{B}}$, $\tau_{\mathcal{B},\mathcal{C}}$, and $\tau_{\mathcal{C}}$. As a consequence, we are able to give formal expressions for the correcting constants C_n and D_n. Because the formal expressions are somewhat ugly, the following results should be more considered as recipes for computing expressions than immediate closed mathematical formulae. The strategy for establishing these exact formulae is rather simple and replicable with slight variations in the three cases.

Case 1: Coalescence within $\mathcal{B}$. Let $T_{j+1} = X_n + \ldots + X_{j+1}$ denote the time at which the sample first has j ancestors. A basic argument shows that if a node has J ancestors, then its expected age is $\mathbf{E}[T_{J+1}]$. Therefore, the coalescence time of two individuals in a subsample of size b for which the total number of ancestors at each node are $J_1 < \ldots < J_{b-1}$ is given by

$$\tau_{\mathcal{B}} = \frac{b+1}{b-1}\sum_{r=1}^{b-1}\frac{2}{(r+1)(r+2)}\mathbf{E}[T_{J_r+1}]$$

which writes as

$$\tau_{\mathcal{B}} = \frac{b+1}{b-1}\sum_{r=1}^{b-1}\frac{2}{(r+1)(r+2)}\sum_{k=2}^{n-b+1}\sum_{j=k+r-1}^{c+r}\mathrm{P}(J_r = j \mid J_\Delta = k)\mathbf{E}[T_{j+1} \mid J_\Delta = k]p_k^\Delta.$$

In this expression, we used corollary 1

$$\mathbf{E}[T_{j+1} \mid J_\Delta = k] = \frac{2(n-j)}{jn}, \quad \text{for } j \geq k,$$

and the result stated in Lemma 2 (Appendix).

Case 2: Coalescence between $\mathcal{B}$ and $\mathcal{C}$. The average coalescence time for two sequences one within $\mathcal{B}$ and one within $\mathcal{C}$ is straightforward from the conditioning on J_Δ. We obtain that

$$\tau_{\mathcal{B},\mathcal{C}} = 2 \sum_{k=2}^{n-b+1} \frac{(k+1)}{(k-1)} \phi(n,k) p_k^\Delta \,, \tag{8}$$

where

$$\phi(n,k) = \sum_{j=2}^{k} \mathbf{E}[T_j \mid J_\Delta = k]/j(j+1), \qquad k = 2, \ldots, n-b+1.$$

Because $j \leq k$ in the above summation, we obtain from corollary 1 that

$$\mathbf{E}[T_j \mid J_\Delta = k] = \frac{2(n-j+1)}{(j-1)n} + \frac{2}{k(k-1)} \,.$$

The expression of $\tau_{\mathcal{B},\mathcal{C}}$ has a simple interpretation in terms of the age of Δ. It can be reduced using a symbolic computing language such as MAPLE$^{\text{TM}}$. Because the gamma distribution $G(2, \lambda_k)$ is the sum of two independent exponential, we find that the coalescence time $\tau_{\mathcal{B},\mathcal{C}}$ is equal to τ_Δ (age of Δ) plus the coalescence time of two ancestors among the k present at the occurrence of Δ. According to Theorem 1, the second coalescence time has exponential $G(1,1)$ distribution. Hence, we have

$$\tau_{\mathcal{B},\mathcal{C}} = 1 + \tau_\Delta,$$

which corresponds to Equation (8) exactly.

Case 3: Coalescence within $\mathcal{C}$. The third case concerns the average coalescence time formula for two sequences within $\mathcal{B}$. This leads to the most terrific (unexplicit) formula which uses a series of probabilistic results stated in Lemmas 3-4 (see APPENDIX). The average coalescence time for two individuals within $\mathcal{C}$ can be obtained

from conditioning on $J_0 = j$, and the observation that we have $K_r = r$ for $r < j$ given that $J_0 = j$. We have

$$\tau_{\mathcal{C}} = 2\frac{(c+1)}{(c-1)}\sum_{r=1}^{c-1}\frac{\mathbf{E}[T_{K_r+1}]}{(r+1)(r+2)}, \qquad r = 1, \ldots, c-1.$$

Now we use the fact

$$\mathbf{E}[T_{K_r+1}] = \sum_{j=1}^{c}\mathbf{E}[T_{K_r+1} \mid J_0 = j]\, P(J_0 = j).$$

For $j = 1, \ldots, c$ and $r < j$, we have

$$\mathbf{E}[T_{K_r+1} \mid J_0 = j] = \frac{2(n-r)}{rn} + \epsilon_j,$$

with

$$\epsilon_j = \sum_{\ell=j+1}^{n-b+1}\frac{2}{\ell(\ell-1)}\, P(J_\Delta = \ell \mid J_0 = j).$$

Otherwise, we have $r \geq j$ and

$$\mathbf{E}[T_{K_r+1} \mid J_0 = j] = \sum_{k=r+1}^{b+r}\mathrm{P}(K_r = k \mid J_0 = j)\left(\frac{2(n-k)}{nk} + \epsilon_{jk}\right)$$

where

$$\epsilon_{jk} = \sum_{\ell=j+1}^{k}\frac{2}{\ell(\ell-1)}\, P(J_\Delta = \ell \mid J_0 = j), \qquad k = r+1, \ldots, b+r.$$

For all $\ell = j+1, \ldots, n-b+1$, the conditional probabilities $\mathrm{P}(J_\Delta = \ell \mid J_0 = j)$ can be obtained from the Bayes formula.

Statistical errors and power of tests: In this paragraph we report numerical estimates of the standard deviations of Π_1, and we study the power of this statistic to reject the hypothesis that the mutation rate raised simultaneously with the occurrence of the mutation Δ. The same experimental design was used as regards the statistic $\hat{\theta_1}$ defined in the previous section. The results closely parallel those

obtained for $\hat{\theta_1}$ (see Tables 6 - 7 - 8). The estimator appears to be unbiased. The standard deviations are of the same order than those computed for $\hat{\theta_1}$ although they seems slighly higher. Using Π_1 instead of $\hat{\theta_1}$ to reject the existence of Δ leads to a 12 or 13 percent loss in power when $\theta_1 = 100$ or $\theta_1 = 10^3$. Reverting the two hypotheses and using Π yields the same conclusions as for $\hat{\theta}$.

DISCUSSION

Genetic information must tightly regulated, and its faithful replication and repair is the highest imperative. To this end humans have invested more than 130 genes in DNA repair, and this number is even greater if genes dedicated to fidelity replication were included (Wood *et al.*, 2001; Anderson 2001). In this article we introduced a stochastic model of mutation in tumor cells with the aim of estimating the amount of genomic instability in cancer tissues due to the alteration of DNA repair genes. Our approach took into account the difficulties generated by sampling within tumoral clones, and the fact that these clones must be difficult to isolate (Anderson *et al.*, 2001). We provided unbiased estimators of the normalized raised mutation rates. These quantities can be interpreted as the mean numbers of new mutations present in daughter cells after each mitotic generation (this corresponds to an evaluation of $\theta_1/2 = 2\mu_1 N$). The power of these statistics to reject genomic instability was assessed and proved to increase with the intensity of mutation. However, we showed that large statistical errors may be associated with such estimates. Conditional on the presence of loss of MMR within a sample of cells, no significant benefit would be expected from large sample sizes. In addition we proved that genomic instability can hardly be detected unless the raised mutation rates exceed the normal rates by a factor greater than 10^3. These results suggest monitoring several loci in order to increase power and reliability of tests, and give theoretically supported grounds to current clinical guidelines (Boland *et al.*, 1998).

Computations were conducted under the assumptions of selective neutrality. Tumors of clonal origin have long lifespan with evolutionary history that may last over 10 or 20 years, and exhibits multistep progression. At least in the early stages of tumor progression selective neutrality is yet compatible with Loeb's theory of cancerogenesis. Evidence are lacking that the initiating events are neither highly advantageous nor highly deleterious. A competing assumption explains that a cell must exhibit a selective advantage to be converted into a pretumoral cell. Then by a selective clonal expansion the cell becomes malignant (Tomlinson *et al.* 1996, Cairns, 1975, Nowell 1976). The material presented in this article may serve as a basis for testing such kinds of assumption. A classical way of doing so is by computing Tajima's D statistic (Tajima 1989). In our framework this statistic can be defined as the difference $\hat{\theta}_1 - \Pi_1$. In order to apply the test, p-values can be obtained from Monte Carlo replicates using the new simulation procedure described in the Conditional coalescent trees section.

Genomic instability particularly affects DNA repeat sequences. It has even been calculated to affect hundred of thousands of such sequences in each tumor cell but very few of these events are within coding sequences (Perucho, 1996). It is widely argumented that stepwise mutation models might be more appropriate for such DNA sequences than the infinitely-many sites model used in this work. However, genomic instability is not restricted to repeat sequences, and even not limited to the nucleus. Mitochondrial DNA may also be mutated in a process that involves clonal expansion (Polyak *et al.* 1998). Infinitely-many site models may thus be acceptable in several situations.

Anderson *et al.* (2001) reported several difficulties with measuring the amount of instability in cancer cell genomes. The ideal measurement would be how many genomic events occur per cell generation because this number would allow evaluating the rate of tumor progression. Regardless the fact that this seems yet difficult to

approach in clinical application, a rigorous way of calculating unbiased estimates of the amount of genomic instability in pretumoral tissues would nevertheless require the correction coefficients described in this article.

LITERATURE CITED

Anderson, G. R., 2001 Genomic instability in cancer. Current Science **81**: 501-507.

Anderson, G. R., D. L. Stoler, and B. M. Brenner, 2001 Cancer: the evolved consequence of a destabilized genome. BioEssays **23**(11): 1037-1046.

Bhattacharyya, N.P., A. Skandalis, A. Ganesh, J. Groden, and M. Meuth, 1994 Mutator phenotypes in human colorectal carcinoma cell lines. Proceedings of the National Academy of Sciences **91**(14): 6319-6323.

Bielas, J. H., and L. A. Loeb, 2005 Mutator phenotype in cancer: Timing and perspectives. Environmental and Molecular Mutagenesis, **45**(2-3): 143-149.

Boland, C. R., S. N. Thibodeau, S. R. Hamilton, D. Sidransky, J. R. Eshleman, R. W. Burt, S. J. Meltzer, M. A. Rodriguez-Bigas, R. Fodde, G. N. Ranzani, and S. Srivastava, 1998 A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Research **58**(22): 5248-57.

Cairns, J., 1975 Mutation selection and the natural history of cancer. Nature **255**(5505): 197-200.

Calabrese, P., J. P. Tsao, Y. Yatabe, R. Salovaara, J. P. Mecklin, H. J. Järvinen, L. A. Aaltonen, S. Tavaré, and D. Shibata, 2004 Colorectal pretumor progression before and after DNA mismatch repair. American Journal of Pathology **164**(4): 1447-1453.

Fishel, R., M. K. Lescoe, M. R. Rao, N. G. Copeland, N. A. Jenkins, J. Garber, M. Kane, and R. Kolodner, 1993 The human mutator gene homolog MSH2 and its

association with hereditary nonpolyposis colorectal cancer. Cell **75**(5), 1027-1038 Erratum in Cell **77**(1): 167.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stochastic Models **14**: 273-295.

Griffiths, R. C., and S. Tavaré, 2003 The genealogy of a neutral mutation, in P. Green, N. Hjort, and S. Richardson, editors. Highly Structured Stochastic Systems 393-412.

Hartl, D., and A. Clark, 1997 *Principles of Population Genetics, Third Edition*, MA: Sinauer Associates, Sunderland.

Ionov, Y., M. A. Peinado, S. Malkhosyan, D. Shibata, and M. Perucho, 1993 Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature **363**(6429): 558-561.

Kingman, J. F. C., 1982 The Coalescent. Stochastic Processes and their Applications **13**: 235-248.

Kinzler, K. W., and B. Vogelstein, 2002 Familial cancer syndromes: the role of caretakers and gatekeepers. pp. 209-210 in *The Genetic Basis of Human Cancer (2nd ed.)*, edited by B. Vogelstein and Kinzler K. W., McGraw-Hill, New York.

Leach, F. S., N. C. Nicolaides, N. Papdopulos, B. Liu, J. Jen, R. Parsons, P. Peltomaki, P. Sistonen, L. A. Aaltonen, M. Nystrom-Lathi, *et al.*, 1993 Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. Cell **75**(6): 1215-1225.

Lindblom, A., P. Tannergard, B. Werelius, and M. Nordenskjold, 1993 Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. Nature Genetics **5**(3): 279-282.

Loeb, L. A., B. N. Springgate, and N. Battula, 1974 Errors in DNA replication as a basis of malignant changes. Cancer Research **34**(9): 2311-2321.

Loeb, L. A., K. R. Loeb, and J. P. Anderson, 2003 Multiple mutations and cancer. Proceedings of the National Academy of Sciences **100**(3): 776-781.

Michor, F., Y. Iwasa, and M. A. Nowak, 2004 Dynamics of Cancer Progression, *Nature Reviews Cancer*, **4**(3): 197-205.

Moolgavkar, S. H., and A. G. Jr. Knudson, 1981 Mutation and cancer: a model for human carcinogenesis. Journal of National Cancer Institute **66**(6): 1037-1052.

Nowell, P. C., 1976 The clonal evolution of tumor cell populations. Science **194**: 23-28.

Peltomaki, P., L. Aaltonen, P. Sistonen, L. Pylkkanen, J. P. Mecklin, H. Jarvinen, J. S. Green, J. R. Jass, J. L. Weber, F. S. Leach, G. M. Petersen, S. R. Hamilton, A. de la Chapelle, and B. Vogelstein, 1993 Genetic mapping of a locus predisposing to human colorectal cancer. Science **260**(5109): 751-752.

Perucho, M. 1996 Cancer of the microsatellite mutator phenotype. Journal of Biological Chemistry **377**(11): 675-684.

Polyak, K., Y. Li, H. Zhu, C. Lengauer, J. K. Willson, S. D. Markowitz, M. A. Trush, K. W. Kinzler, and B. Vogelstein, 1998 Somatic mutations of the mitochondrial genome in human colorectal tumours. Nature Genetics **20**(3): 291-293.

R Development Core Team 2004 R: A language and development for statistical computing *R foundation for Statistical Computing Vienna, Austria.* ISBN 3-900051-00-3, URL http://www.R-project.org

Shibata, D., M. A. Peinado, Y. Ionov, S. Malkhosyan, and M. Perucho, 1994 Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. Nature Genetics **6**(3): 273-281.

Stephens, M., 2000 Times on tree, and the Age of en Allele. Theoretical Population Biology **57**: 109-119.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**(2): 437-460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**(3): 585-595.

Tavaré, S., 2004 Ancestral inference in population genetics, pp. 1-188 in *Lectures on Probability Theory and Statistics. Ecole d'Eté de Probabilité de St Flour XXXI–2001.*,

edited by J. Picard. Springer Verlag. New York.

Thibodeau, S. N., G. Bren, and D. Schaid, 1993 Microsatellite instability in cancer of the proximal colon. Science **260**(5109): 816-819.

Thomas, D. C., 2004 *Statistical Methods in Genetic Epidemiology.* Oxford University Press. New York.

Tomlinson, I., M. Novelli, and W. Bodmer, 1996 The mutation rate and cancer. Proceedings of the National Academy of Sciences **93**(25): 14800-14803.

Tsao, J. L., Y. Yatabe, R. Salovaara, H. J. Järvinen, J. P. Mecklin, L. A. Aaltonen, S. Tavaré, and D. Shibata, 2000 Genetic reconstruction of individual colorectal tumor histories. Proceedings of the National Academy of Sciences **97**(3): 1236-1241.

Watterson, G.A., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **2**: 256-276.

Wiuf, C., and P. Donnelly, 1999 Conditional genealogies and the age of a neutral mutant. Theoretical Population Biology **56**: 183-201.

Wood, A., 2001 Racial differences in the response to drugs - pointers to genetic differences. The New England Journal of Medicine **344**: 1393-1395.

APPENDIX

Proof of Theorem 1. The proof follows the same lines as TAVARÉ (2004, Chap. 8, p. 110). Let $s = (s_2, \ldots, s_n)$ and $X = (X_2, \ldots, X_n)$, we have

$$s.X = \sum_{i=2}^{n} s_i X_i$$

Conditional on E, the multidimensional Laplace transform of X is equal to

$$\begin{aligned}
\mathbf{E}[e^{-s.X} \mid E] &= \sum_{k=2}^{n-b+1} \mathbf{E}[e^{-s.X} \mathbb{1}_{J_\Delta = k} \mid E] \\
&= \sum_{k=2}^{n-b+1} \mathbf{E}[\mathbf{E}[e^{-s.X} \mathbb{1}_{J_\Delta = k} \mid X, E]] \\
&= \sum_{k=2}^{n-b+1} \mathbf{E}[e^{-s.X} \mathrm{P}(J_\Delta = k \mid X, E)] \\
&= \sum_{k=2}^{n-b+1} \mathbf{E}[e^{-s.X} X_k \frac{\delta}{2} e^{-\frac{L_n \delta}{2}}] \lambda_k \begin{pmatrix} n-k \\ b-1 \end{pmatrix} \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1} \\
&= \frac{\delta}{2} \sum_{k=2}^{n-b+1} \mathbf{E}[e^{s.X} X_k] \lambda_k \begin{pmatrix} n-k \\ b-1 \end{pmatrix} \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1} + o(\delta) \\
&= \frac{\delta}{2} \sum_{k=2}^{n-b+1} \mathbf{E}[e^{-s.X}] \frac{\lambda_k}{s_k + \lambda_k} \begin{pmatrix} n-k \\ b-1 \end{pmatrix} \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1} + o(\delta)
\end{aligned}$$

Further conditioning on M, and taking the limit as δ tends to zero leads to

$$\mathbf{E}[e^{-s.X} \mid E \cap M] = \sum_{k=2}^{n-b+1} E[e^{-s.X}] \frac{\lambda_k}{s_k + \lambda_k} \frac{\begin{pmatrix} n-k \\ b-1 \end{pmatrix}}{\begin{pmatrix} n-1 \\ b \end{pmatrix}}$$

The Theorem follows from the fact that the multidimensional Laplace transform of the density given in Equation (4) coincides with the above formula. ∎

Proof of Corollary 1. Let $n \geq 2$. Assuming that Δ has b descendants ($1 \leq b \leq n-1$) and using Theorem 1 we obtain the marginal distribution of each inter-coalescence time. For $\ell = 2, \ldots, n$ we have

$$f(x_\ell) = \left(\sum_{k=2,\, k\neq\ell}^{n-b+1} p_k^\Delta + p_\ell^\Delta \lambda_\ell x_\ell \right) f_\ell(x_\ell)$$

if $\ell \leq n-b+1$, otherwise, it is equal to

$$f(x_\ell) = f_\ell(x_\ell)$$

where f_ℓ is the density of the exponential $G(1, \lambda_\ell)$ distribution. Taking expectations it comes

$$\mathbf{E}[X_\ell \mid E \cap M] = \begin{cases} \left(1 + p_\ell^\Delta\right)/\lambda_\ell & \text{if } \ell \leq n-b+1 \\ 1/\lambda_\ell & \text{otherwise.} \end{cases}$$

■

Lemma 1 *Let $n \geq 2$. We have*

$$\frac{1}{2}\mathbf{E}[L_n^\Delta] = H_{n-1} + \frac{1}{H_{n-1}} \sum_{b=1}^{n-1} \sum_{k=2}^{n-b+1} \frac{p_k^\Delta}{b(k-1)}$$

Proof. Let $b = 1, \ldots, n-1$. From corollary 1 we have

$$\begin{aligned} \mathbf{E}[L_n^\Delta | B = b] &= \sum_{k=2}^{n} k\mathbf{E}[X_k] \\ &= 2H_{n-1} + 2 \sum_{k=2}^{n-b+1} \frac{p_k^\Delta}{k-1} \end{aligned}$$

Then

$$\begin{aligned} \mathbf{E}[L_n^\Delta] &= \sum_{b=1}^{n-1} \frac{1}{bH_{n-1}} \mathbf{E}[L_n^\Delta | B = b] \\ &= 2\left(H_{n-1} + \frac{1}{H_{n-1}} \sum_{b=1}^{n-1} \sum_{k=2}^{n-b+1} \frac{p_k^\Delta}{b(k-1)} \right) \end{aligned}$$

■

Lemma 2 *Let $n \geq 2$ and assume that Δ has b descendants. Let $r = 1, \ldots, b-1$ and $k \in [2, n-b+1]$. For $j \in [k+r-1, n-b+r]$, we have*

$$\mathrm{P}(J_r = j \mid J_\Delta = k\,;\, E \cap M) = \frac{\left(\begin{array}{c} j-k \\ r-1 \end{array}\right)\left(\begin{array}{c} n-j-1 \\ b-r-1 \end{array}\right)}{\left(\begin{array}{c} n-k \\ b-1 \end{array}\right)}$$

Proof. Let $k \in [2, n-b+1]$ and $r \in [1, b-1]$. For all $j \in [k+r-1, n-b+r]$ it is known that for $k \leq j_1 < \ldots < j_{r-1} < j$ we have (Tavaré, 2004)

$$\mathrm{P}(J_1 = j_1, \ldots, J_{r-1} = j_{r-1}, J_r = j \mid J_\Delta = k;\, E \cap M) = \left(\begin{array}{c} n-j-1 \\ b-r-1 \end{array}\right)\left(\begin{array}{c} n-k \\ b-1 \end{array}\right)^{-1}$$

Note that the above formula is independent on $j_1, \ldots, j_{r-1}$. We have

$$\begin{aligned} &\mathrm{P}(J_r = j \mid J_\Delta = k\,;\, E \cap M) \\ &= \sum_{k \leq j_1 < \ldots < j_{r-1} < j} \mathrm{P}(J_1 = j_1, \ldots, J_{r-1} = j_{r-1}, J_r = j \mid J_\Delta = k;\, E \cap M) \\ &= \left(\begin{array}{c} j-k \\ r-1 \end{array}\right)\left(\begin{array}{c} n-j-1 \\ b-r-1 \end{array}\right)\left(\begin{array}{c} n-k \\ b-1 \end{array}\right)^{-1} \end{aligned}$$

■

Lemma 3 *Let $n \geq 2$ and assume that Δ has b descendants. Let J_0 be defined as in the* conditional coalescent trees *section. For $j = 1, \cdots, n-b$, we have*

$$\mathrm{P}(J_0 = j \mid E \cap M) = 2j \sum_{k=j+1}^{n-b+1} \frac{p_k^\Delta}{k(k-1)}$$

Proof. Thanks to a straighforward combinatorial argument, for $j = 1, \cdots, n-b$ we have

$$\mathrm{P}(J_0 = j \mid J_\Delta = k\ ;\ E \cap M) = \frac{2j}{k(k-1)}$$

Then intregrating over J_Δ's implies that

$$\mathrm{P}(J_0 = j \mid E \cap M) = 2j \sum_{k=j+1}^{n-b+1} \frac{p_k^\Delta}{k(k-1)}, \qquad k = j+1, \ldots, n-b+1$$

■

Lemma 4 *Let $n \geq 2$, assume that Δ has b descendants and denote $c = n - b$. Let $r = j, \ldots, c-1$ and K_r be defined as in the* CONDITIONAL COALESCENT TREES *section. For $k \in [r+1, r+b]$, we have*

$$\mathrm{P}(K_r = k \mid J_0 = j\,;\, E \cap M) = \frac{\begin{pmatrix} k-j-1 \\ r-j \end{pmatrix} \begin{pmatrix} n-k-1 \\ c-r-1 \end{pmatrix}}{\begin{pmatrix} n-j-1 \\ b \end{pmatrix}}$$

Proof. Note that the vector $(J_0, \ldots, J_{b-1}, K_0, \ldots, K_{c-1})$ is obtained from a permutation of the labels $(1, 2, \ldots, n-1)$, where J_r's and K_r's are defined as in the CONDITIONAL COALESCENT TREES section. Conditional on $J_0 = j$, the vector $(J_1, \ldots, J_{b-1}, K_j, \ldots, K_{c-1})$ is also a permutation of the labels $(j+1, \ldots, n-1)$. Then Equation 1 implies that for $j < k_j < \ldots < k_{c-1} < n$, we have

$$\mathrm{P}(K_r = k_r,\, r = 1, \ldots, c-1 \mid J_0 = j\,;\, E \cap M) = \begin{pmatrix} n-j-1 \\ b \end{pmatrix}^{-1}$$

Note that the above formula is independent on $k_1, \ldots, k_{c-1}$. We have

$$\begin{aligned}
&\mathrm{P}(K_r = k \mid J_0 = j\,;\, E \cap M) \\
&= \sum_{j<k_j<\ldots<k_{r-1}<r} \sum_{r<k_{r+1}<\ldots<k_{c-1}<n} \mathrm{P}(K_r = k_r,\, r = 1, \ldots, c-1 \mid J_0 = j\,;\, E \cap M) \\
&= \frac{\begin{pmatrix} k-j-1 \\ r-j \end{pmatrix} \begin{pmatrix} n-k-1 \\ c-r-1 \end{pmatrix}}{\begin{pmatrix} n-j-1 \\ b \end{pmatrix}}.
\end{aligned}$$

■

Table 1: **Correction coefficients for** $\hat{\theta}_1$. Numerical values for the correcting coefficients A_n and B_n in the statistic $\hat{\theta}_1 = (S - A_n\theta_0)/B_n$ for n in the range $5 - 50$.

n	5	10	15	20	25	30	35	40	45	50
A_n	2.171	2.693	3.024	3.265	3.455	3.612	3.747	3.864	3.967	4.061
B_n	0.595	0.68	0.713	0.732	0.746	0.756	0.764	0.771	0.776	0.781

Table 2: **Statistical errors for $\hat{\theta_1}$.** Bias and standard deviation for the estimator $\hat{\theta_1}$ for sample size $n = 10 - 50$. The normal rate was set to the value $\theta_0 = 1$ and the raised rates varied from $\theta_1 = 10$ to $\theta_1 = 1,000$.

	$\theta_1 = 10$		$\theta_1 = 100$		$\theta_1 = 1,000$	
n	**E**	SD	**E**	SD	**E**	SD
10	9.9	12.0	97.4	112.4	947.5	1109.7
20	10.3	12.5	99.7	122.4	991.9	1211.1
30	10.2	12.8	102.9	126.1	1060.3	1286.1
40	10.2	13.2	100.9	128.9	1018.2	1286.2
50	10.4	13.5	102.0	131.7	1045.7	1235.9

Table 3: **Powers for $\hat{\theta}_1$.** Power of the test based on the statistic $\hat{\theta}_1$ where the null hypothesis H_0 is the existence of Δ and $\theta_1 > \theta_0$ whereas the alternative hypothesis H_1 is the absence of Δ. The normal rate was set to the value $\theta_0 = 1$ and the raised rates varied from $\theta_1 = 10$ to $\theta_1 = 1,000$.

n	$\theta_1 = 10$	$\theta_1 = 100$	$\theta_1 = 1,000$
10	0.10	0.29	0.90
20	0.06	0.18	0.70
30	0.13	0.29	0.65
40	0.11	0.24	0.59
50	0.09	0.21	0.55

Table 4: **Powers for $\hat{\theta}$.** Power of the test based on the statistic $\hat{\theta}$ where the null hypothesis H_0 is the absence of Δ whereas the alternative hypothesis H_1 is the existence of Δ associated with $\theta_1 > \theta_0$. The normal rate was set to the value $\theta_0 = 1$ and the raised rates varied from $\theta_1 = 10$ to $\theta_1 = 1,000$.

n	$\theta_1 = 10$	$\theta_1 = 100$	$\theta_1 = 1,000$
10	0.44	0.75	0.93
20	0.44	0.74	0.90
30	0.48	0.75	0.89
40	0.42	0.73	0.88
50	0.43	0.72	0.87

Table 5: **Correction coefficients for** Π_1. Numerical values for the correcting coefficients C_n and D_n in the statistic $\Pi_1 = (\Pi - C_n\theta_0)/D_n$ for n in the range $5-50$.

n	5	10	15	20	25	30	35	40	45	50
C_n	0.996	1.019	1.021	1.02	1.02	1.019	1.019	1.018	1.018	1.018
D_n	0.253	0.218	0.199	0.187	0.178	0.171	0.166	0.161	0.156	0.154

Table 6: **Statistical errors for** Π_1. Bias and standard deviation for the estimator Π_1 for sample size $n = 10 - 50$. The normal rate was set to the value $\theta_0 = 1$ and the raised rates varied from $\theta_1 = 10$ to $\theta_1 = 1,000$.

	$\theta_1 = 10$		$\theta_1 = 100$		$\theta_1 = 1,000$	
n	**E**	SD	**E**	SD	**E**	SD
10	9.9	13.7	107.342	133.9	1006.2	1243.5
20	10.2	14.7	100.91	136.2	1030.5	1458.9
30	9.5	15.5	100.875	147.9	1040.0	1589.5
40	10.7	17.8	95.763	159.0	998.4	1538.1
50	10.3	17.6	106.478	164.6	1039.7	1598.1

Table 7: **Powers for Π_1.** Power of the test based on the statistic Π_1 where the null hypothesis H_0 is the existence of Δ and $\theta_1 > \theta_0$ whereas the alternative hypothesis H_1 is the absence of Δ. The normal rate was set to the value $\theta_0 = 1$ and the raised rates varied from $\theta_1 = 10$ to $\theta_1 = 1,000$.

n	$\theta_1 = 10$	$\theta_1 = 100$	$\theta_1 = 1000$
10	0.09	0.32	0.72
20	0.12	0.29	0.54
30	0.14	0.24	0.44
40	0.12	0.19	0.35
50	0.13	0.20	0.40

Table 8: **Powers for** Π. Power of the test based on the statistic Π where the null hypothesis H_0 is the absence of Δ whereas the alternative hypothesis H_1 is the existence of Δ associated with $\theta_1 > \theta_0$. The normal rate was set to the value $\theta_0 = 1$ and the raised rates varied from $\theta_1 = 10$ to $\theta_1 = 1,000$.

n	$\theta_1 = 10$	$\theta_1 = 100$	$\theta_1 = 1000$
10	0.44	0.73	0.91
20	0.44	0.69	0.84
30	0.39	0.64	0.80
40	0.34	0.64	0.79
50	0.34	0.62	0.76

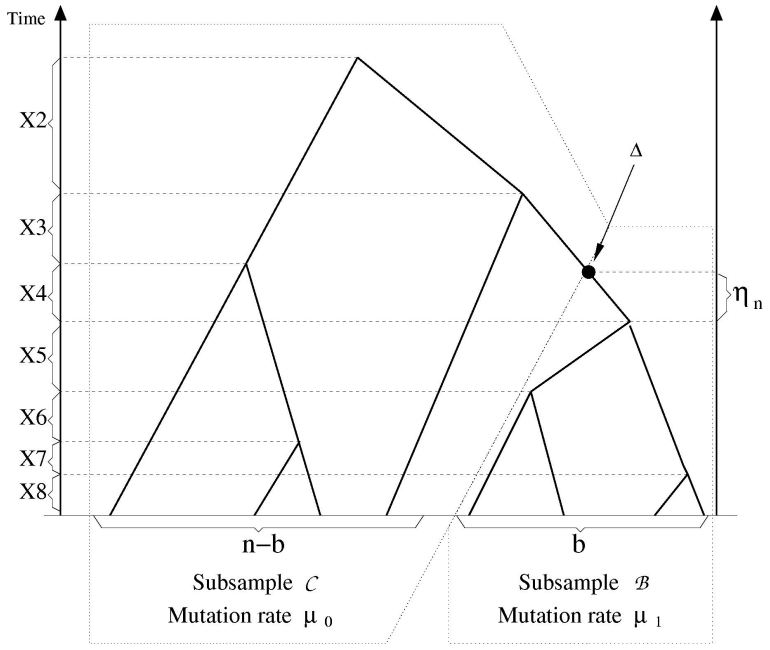

Figure 1: Conditional coalescent tree with $n = 8$ leaves. The mutation Δ has $B = 4$ descendants.

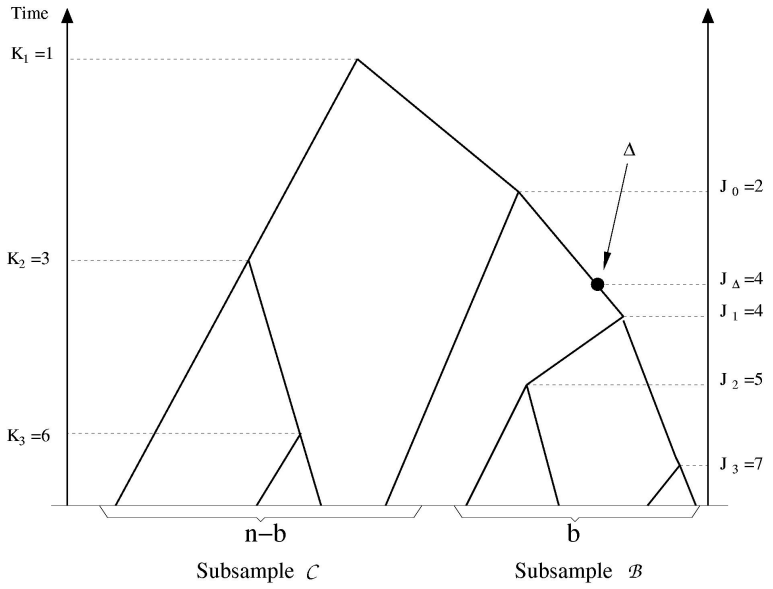

Figure 2: Coalescence levels in $\mathcal{B}$ and $\mathcal{C}$ with their notations J_r and K_r. Here we have $n = 8$, $B = 4$, $J_3 = 7, J_2 = 5, J_1 = 4, J_0 = 2$ and $K_3 = 6$, $K_2 = 3$, $K_1 = 1$.

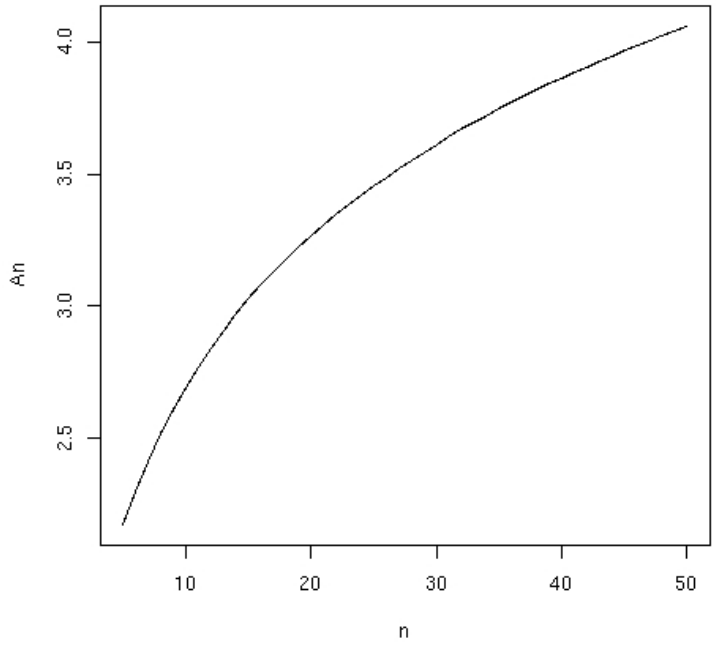

Figure 3: Numerical values for A_n for n in the range $5 - 50$.

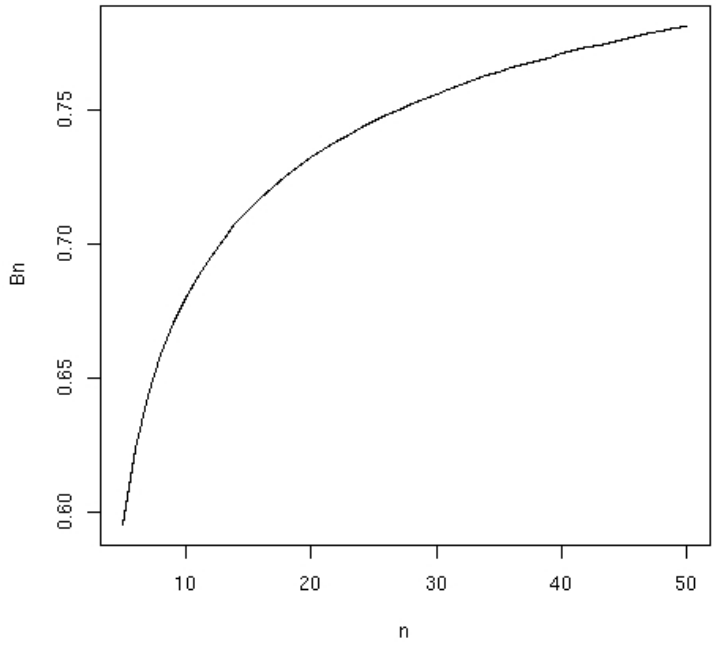

Figure 4: Numerical values for B_n for n in the range $5 - 50$.

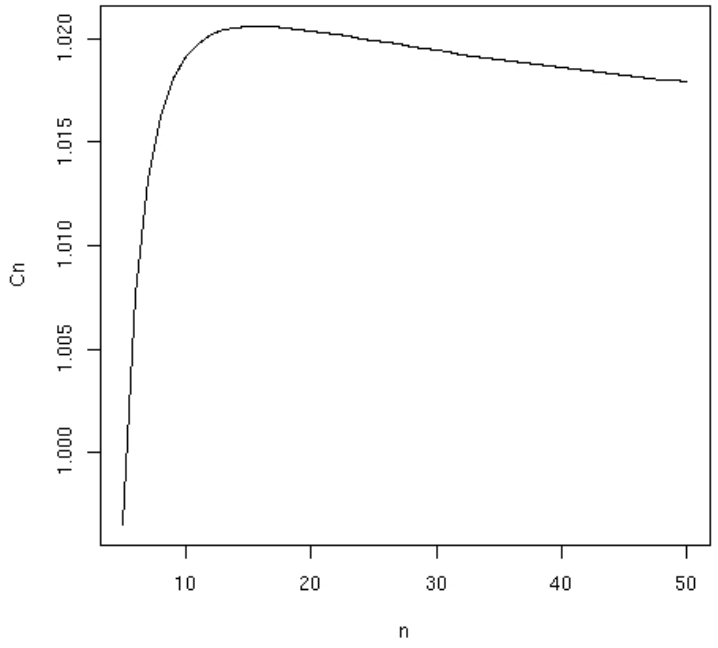

Figure 5: Numerical values for C_n for n in the range $5 - 50$.

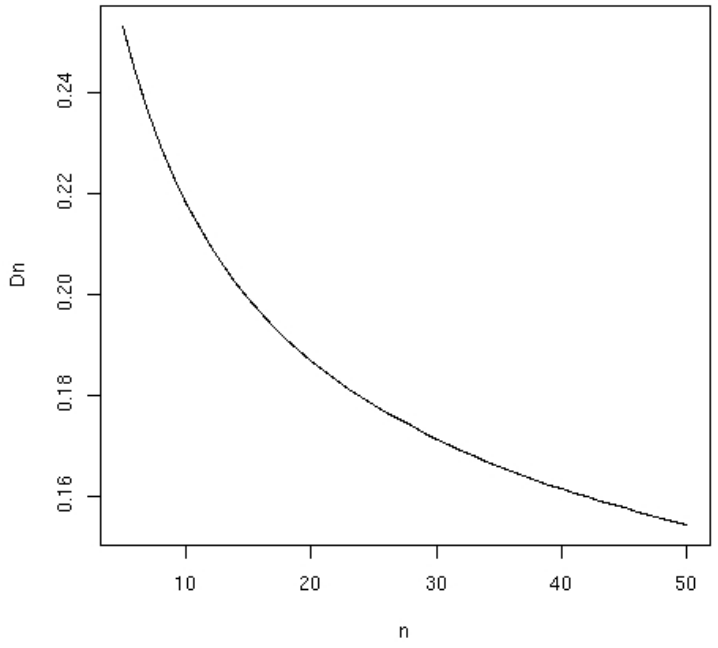

Figure 6: Numerical values for D_n for n in the range $5 - 50$.