BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Accurate and fast methods to estimate the population mutation rate from error prone sequences

Bjarne Knudsen[1] and Michael M Miyamoto*[2]

Address: [1]CLC bio, 8200 Århus N, Denmark and [2]Department of Biology, Box 118525, University of Florida, Gainesville, Florida 32611-8525, USA

Email: Bjarne Knudsen - bknudsen@clcbio.com; Michael M Miyamoto* - miyamoto@ufl.edu

* Corresponding author

Published: 11 August 2009

BMC Bioinformatics 2009, **10**:247 doi:10.1186/1471-2105-10-247

Received: 25 February 2009
Accepted: 11 August 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/247

Abstract

Background: The population mutation rate (θ) remains one of the most fundamental parameters in genetics, ecology, and evolutionary biology. However, its accurate estimation can be seriously compromised when working with error prone data such as expressed sequence tags, low coverage draft sequences, and other such unfinished products. This study is premised on the simple idea that a random sequence error due to a chance accident during data collection or recording will be distributed within a population dataset as a singleton (i.e., as a polymorphic site where one sampled sequence exhibits a unique base relative to the common nucleotide of the others). Thus, one can avoid these random errors by ignoring the singletons within a dataset.

Results: This strategy is implemented under an infinite sites model that focuses on only the internal branches of the sample genealogy where a shared polymorphism can arise (i.e., a variable site where each alternative base is represented by at least two sequences). This approach is first used to derive independently the same new Watterson and Tajima estimators of θ, as recently reported by Achaz [1] for error prone sequences. It is then used to modify the recent, full, maximum-likelihood model of Knudsen and Miyamoto [2], which incorporates various factors for experimental error and design with those for coalescence and mutation. These new methods are all accurate and fast according to evolutionary simulations and analyses of a real complex population dataset for the California seahare.

Conclusion: In light of these results, we recommend the use of these three new methods for the determination of θ from error prone sequences. In particular, we advocate the new maximum likelihood model as a starting point for the further development of more complex coalescent/mutation models that also account for experimental error and design.

Background

The population mutation rate (θ) remains one of the most fundamental parameters in genetics, ecology, and evolutionary biology [3-5]. This interest in θ derives from the fact that this parameter measures the effective size (N_e) and whole-locus mutation rate (μ) of a population, which are of great importance in understanding its demography and history. Specifically, θ is a compound parameter that is calculated as the product of $2pN_e\mu$ (with $p = 1$ or 2 for haploids and diploids, respectively). Correspondingly, a number of alternative methods are available to estimate θ from a population sample of allelic sequences [6-8]. These alternative methods range from relatively simple summary statistics (moment methods) to full coalescent/

mutation models. Indeed, the estimation of θ remains central to even the most complex coalescent/mutation models that are otherwise concerned with the determination of other population genetic parameters (e.g., for growth, migration, and recombination).

A population sample of sequences is obtained from interbreeding or potentially interbreeding individuals and is therefore usually associated with a small number of mutations [9-12]. Thus, when estimating θ from a population sample, sequence errors can pose a real problem, since their numbers can begin to approach or even surpass those for the mutations [13-19]. This problem becomes particularly acute when working with error prone data such as expressed sequence tags (EST), low coverage draft sequences, and other such unfinished products [20,21]. For example, an error rate of one mistake per every 500 nucleotides (e.g., as for an EST dataset obtained from single sequencing passes) will make a significant contribution to the observed variation among sequences that differ because of mutations by <1 to 2%. If uncorrected, such errors can lead to an inflated estimate of θ and even erroneous conclusions about the biology of their population [1,2,18,22].

Many sequence errors arise as random accidents that occur during the nucleic acid isolation, cloning/amplification, sequencing, and recording phases of a DNA sequencing study [11,23,24]. As chance events that are rare (even for error prone data), each of these random mistakes will most likely be limited to a single sequence, rather than repeated among two or more different ones within the population sample [1,15]. Thus, these random errors will most likely inflate the number of singletons within the dataset (i.e., polymorphic positions where one sampled sequence exhibits a unique base relative to the shared nucleotide of the others) (Figure 1). In contrast, these rare chance mistakes will make a much smaller contribution to the shared polymorphisms (i.e., variable sites where each alternative base is common to at least two different sampled sequences).

This study relies on the simple premise that the random mistakes of error prone sequences can be avoided by ignoring the singletons within their population sample. This strategy is first used to obtain independently the same new Watterson [25] and Tajima [26] estimators of θ, which were recently reported by Achaz [1] for error prone sequences. This approach is then implemented in the recent maximum likelihood (ML) model of Knudsen and

Sampled	Alignment position									
sequence	1	2	3	4	5	6	7	8	9	10
I	A	C	A	G	T	C	G	A	T	C
II	.	.	.	.	C	.	.	.	.	T
III	.	T	G	.	.	.	.	.	.	T
IV	.	.	G	.	.	T	.	.	.	T
V	.	.	G	.	.	T	.	.	-	T
Singleton (*) or shared polymorphism (+)		*	+		*	+				*

Figure 1
Multiple sequence alignment for five hypothetical sequences as sampled from a single population. Periods in sequences II-V refer to the same base as in I. The dash refers to a gap. Variable sites corresponding to singletons are marked with asterisks, whereas those representing shared polymorphisms are denoted with pluses.

Miyamoto [2], which incorporates various factors of experimental error and design with those for coalescence and mutation. By relying on only shared polymorphisms, these three new approaches allow for more accurate estimates of θ. However, this greater accuracy comes with a cost as singletons due to actual mutations are ignored along with those due to random errors. To assess this tradeoff, the three new methods are tested against each other and their original predecessors that count singletons with evolutionary simulations and/or analyses of a real population dataset for the California seahare (*Aplysia californica*). These tests document that these new approaches offer reliable and fast alternatives for the determination of θ from error prone sequences.

Results and discussion

Three new methods for estimating θ from error prone sequences

Infinite sites model

As in their original versions, the new Watterson, Tajima, and Knudsen/Miyamoto methods rely on the infinite sites model to accommodate a neutral mutation process [27,28]. The infinite sites model assumes that only a single mutation can occur at any homologous position of the population sample. Thus, each variable site will be represented by only two bases that subdivide the sampled sequences into two non-overlapping subsets consisting of those with the first nucleotide versus the remainder with the second base. Correspondingly, each mutation will map to a specific branch within the sample genealogy, which thereby partitions the sequences into their two non-overlapping subsets. For example, the shared polymorphism at position 3 in Figure 1 is attributable to a unique mutation along the internal branch that partitions sequences I and II from III, IV, and V.

As random errors are treated as rare chance events, they are also modeled in this study along with the mutations by the infinite sites process. Thus, only a single random error or mutation is allowed at any site of the sampled sequences. In turn, each random error is limited to a single sequence (and therefore to a particular singleton) in contrast to a mutation that can also result in a shared polymorphism (Figure 1). The reason is that random errors arise during the experimental determination and recording of individual sequences, whereas mutations occur at specific points within the sample genealogy. Thus, a mutation along an internal branch of the genealogy will result in a new base that will be shared by two or more of its descendant sequences.

New Watterson estimator (θ'_W)

Define T_i as the length of time (as scaled by N_e generations) during which there are exactly i ancestors for n sampled sequences. Standard coalescent theory tells us that:

$$T_i \sim Exp(\frac{i(i-1)}{2}) \qquad (1)$$

and

$$E[T_i] = \frac{2}{i(i-1)} \qquad (2)$$

[4,29,30]. The expected total branch length of the genealogy for n sequences (as measured in units of scaled coalescent time) can now be calculated as:

$$E[L] = \sum_{i=2}^{n} iE[T_i] = 2\sum_{i=1}^{n-1}\frac{1}{i}. \qquad (3)$$

Let n_s be the observed number of segregating (polymorphic) sites in the dataset. Under the infinite sites model, n_s also counts the number of mutations, since each observed variable site is attributable to a single mutation. The expected number of mutations per locus per unit of branch length is θ/2. Thus, an estimate of θ can be obtained as:

$$\theta_W = \frac{2n_S}{E[L]} \qquad (4)$$

[25].

A new Watterson estimator that avoids singletons (and thereby random sequence errors, θ'_W) can now be derived from Equation (4) by adjusting both its numerator and denominator. The numerator is easily adjusted by counting only the shared polymorphic sites in the original dataset (n'_s). Although more complicated, the denominator can also be readily adjusted by including in its calculation only the lengths of the internal branches where shared polymorphisms can arise (see below).

Let us again consider a point in the genealogy where there are exactly i ancestors for the n sampled sequences. Looking forward in time, the probability that a particular branch of the genealogy is not chosen for the next split (leading to i + 1 lineages) is [(i - 1)/i]. Thus, the probability that this branch remains unbroken to the present is:

$$\frac{i-1}{i} \times \frac{i}{i+1} \times \ldots \times \frac{n-2}{n-1} = \frac{i-1}{n-1}. \qquad (5)$$

By combining Equations (3) and (5), the total length of the external branches where singletons can occur can now be calculated as:

$$E[L_1] = \sum_{i=2}^{n} \frac{i-1}{n-1} iE[T_i] = 2. \quad (6)$$

Our Equation (6) is equivalent to equation (10) of Fu and Li [31], apart from our use of different symbols and terms and of time as scaled by N_e generations (rather than generations alone). Thus, as previously noted by them, the total length of the external branches where a singleton can occur is independent of the original number of sampled sequences. Fu and Li [31] also obtained the variance for the total length of the external branches as their equation (14).

In an asymmetrical genealogy with a basal split of one versus (n - 1) sampled sequences, a mutation in the internal branch leading to the common ancestor of the (n - 1) group will result in a singleton within the dataset (i.e., a variable site where the single nonmember sequence exhibits the unique base) (Figure 2). Thus, the length of this (n - 1) basal branch (as weighted by its probability of occurrence) must also be accounted for in the adjustment of the denominator for θ'_W. The weighted length of this (n - 1) basal branch can be calculated as follows. First, the chance of this branch is determined as the probability of an (n - 1) asymmetrical topology or equivalently as the probability that either one of the two basal branches for the genealogy remains unbroken to the present. According to Equation (5), the latter probability is 2 [1/(n - 1)]. Second, the unweighted length of this branch is then calculated as the length of the time interval T_2, which is 1 [Equation (2)]. Thus, the weighted length of the (n - 1) basal branch is:

$$E[L_{n-1}] = \frac{2}{n-1}. \quad (7)$$

By combining Equations (4), (6), and (7), θ'_W can now be obtained as:

$$\theta'_W = \frac{2n'_s}{E[L]-E[L_1]-E[L_{n-1}]} = n'_s(\sum_{i=2}^{n-2}\frac{1}{i})^{-1}. \quad (8)$$

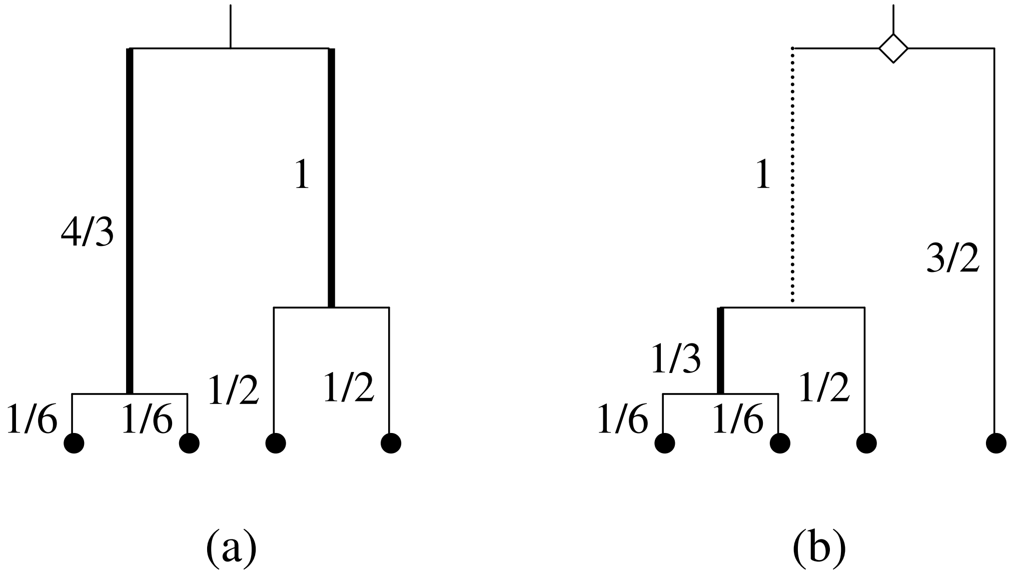

Figure 2

Representative symmetrical (a) and asymmetrical (b) genealogies for four sampled sequences. In all, there are six symmetrical and 12 asymmetrical labeled genealogies for these four sequences [8,45]. This figure illustrates how the shape of a genealogy affects whether a mutation will lead to a singleton or shared polymorphism. This heterogeneity contributes to the variance of θ for the three new methods. This contribution is in addition to the heterogeneity of the genealogical branch lengths and Poisson mutation process [4,5]. The diamond highlights the common ancestor of the (n - 1) basal group of the asymmetrical genealogy. The dotted and thin solid lines mark the basal branch leading to this ancestor and the external branches, respectively, where a mutation will result in a singleton. The thick solid lines denote the other internal branches where a mutation will lead to a shared polymorphism. Expected branch lengths are given in units of scaled coalescent time next to each internode (those of the symmetrical genealogy are specific for its particular labeled history). Although both genealogies have an expected overall length of 11/3, the total length of the internal branches where a shared polymorphism can arise is 7/3 for (a) but only 1/3 for (b).

In words, θ'_W ignores the singletons (and thereby random errors) in the original dataset, while restricting the total branch length calculation to only those internal internodes where a shared polymorphism can arise.

Apart from our use of different symbols and terms for calculating the denominator, our Equation (8) is equivalent to equation (13) of Achaz [1]. Achaz [1] also derived his equation B22 for the calculation of the associated variance for n'_s (i.e., his Var [$S_{-\eta 1}$]).

New Tajima estimator (θ'_T)

Let π_{ij} denote the number of observed pairwise differences between sampled sequences i and j (for $i \neq j$). Standard coalescent theory tells us that the expected waiting time for these two sequences to coalesce is ~ N_e generations. Thus, the expected value of each π_{ij} is θ, whereas that for their sum is:

$$E[\sum_{i=1}^{n}\sum_{j=i+1}^{n}\pi_{ij}] = \frac{n(n-1)}{2}\theta. \tag{9}$$

Rearranging Equation (9) leads to:

$$\theta_T = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\pi_{ij} \tag{10}$$

[26].

The Tajima and Watterson estimators have the same expected value of θ, even though the former is based on the total differences between each sampled sequence pair whereas the latter is obtained from the total number of segregating sites within the sample. Their different summaries of the observed variation are the basis of Tajima's *D* that tests for departures from the standard neutral model [32] (see also [31,33,34]).

To avoid random errors, Equation (10) can be adjusted in a manner similar to that used to modify Equation (4) of θ_W. First, let π'_{ij} count the number of observed pairwise differences between sequences i and j after the removal of all singletons from their dataset. Then, the expected number of singletons can be calculated from Equations (6) and (7) as the total length of the external branches and (n - 1) basal branch (as weighted by its probability of occurrence) multiplied by θ/2 {i.e., [1 + (1/n - 1)] θ}. The removal of each singleton from the original dataset reduces the expected sum in Equation (9) by (n - 1). Combining these results, we obtain:

$$\begin{aligned} E[\sum_{i=1}^{n}\sum_{j=i+1}^{n}\pi'_{ij}] &= E[\sum_{i=1}^{n}\sum_{j=i+1}^{n}\pi_{ij}] - (n-1)(1+\frac{1}{n-1})\theta \\ &= (\frac{n(n-1)}{2} - n)\theta = \frac{n(n-3)}{2}\theta. \end{aligned} \tag{11}$$

Rearranging Equation (11) then leads to:

$$\theta'_T = \frac{2}{n(n-3)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\pi'_{ij}. \tag{12}$$

Our Equation (12) for θ'_T is equivalent to equation (15) of Achaz [1], except for our use of different symbols and total π'_{ij} (rather than average π'_{ij}) for the population sample. Achaz [1] also derived his equation B23 for the calculation of the associated variance for average π'_{ij} [i.e., his Var[$\pi_{-\eta 1}$]).

New Knudsen/Miyamoto model (θ'_{KM})

Knudsen and Miyamoto ([2], hereafter referred to as "KM") developed a full coalescent/mutation model that accounts for three specific factors of experimental error and design: (a) For random sequence errors, (b) For unobserved polymorphisms due to missing data, and (c) For the uncertain assignment of the multiple sequencing reads for a diploid or polyploid individual to its two or more homologues. Their KM model uses recursion to calculate an exact probability of the population sample under the standard Fisher [35] and Wright [36] model for reproduction (hereafter, referred to as "FW") and the infinite sites process for both mutations and random sequence errors [37,38]. Their model relies on ML to estimate θ and the expected number of errors per full length sequence (ε).

At the heart of the KM model is equation (1) of Knudsen and Miyamoto [2], which is reproduced here as:

$$\begin{aligned} P_c(S) &= \frac{2}{|S|(|S|-1)+\theta\sum_i|s_i|}\sum_i\sum_{j>i:s_j\sim s_i}P_c(\beta_{ij}(S)) \\ &+ \frac{\theta}{|S|(|S|-1)+\theta\sum_i|s_i|}\sum_{i:\sigma_i>0}\frac{|s_i|\sigma_i}{m_s}P_c(\alpha_i(S)). \end{aligned} \tag{13}$$

The parameters and factors of this equation are defined in Table 1. As one works backwards in time, the probability that the next observed event is a specific coalescence is given by the first term in the top line before the double sum. As indicated by this double sum, if indeed a coalescent event occurs, then it will happen between two sampled and/or ancestral alleles with compatible (if not

Table 1: Definitions of the parameters and factors used in Equation (13) of the KM model

Parameter or factor	Description
m_s	The number of segregating sites within the current set of sampled and/or ancestral sequences
σ_i	Counts the number of "singletons" for sampled or ancestral sequence *i*. Here, "singleton" refers both to the derived mutations of the shared polymorphisms for the sampled sequences as well as to those of the observed singletons (in the strict sense) within the original dataset (Figure 3).
$P_c(\alpha_i(S))$	Probability of $\alpha_i(S)$, which is the current set of sampled and/or ancestral sequences prior to a mutation in sequence *i*
$P_c(\beta_{ij}(S))$	Probability of $\beta_{ij}(S)$, which is the current set of sampled and/or ancestral sequences after the coalescence of combinable sequences *i* and *j* ($s_i \sim s_j$; see below)
$P_c(S)$	Probability of *S*, which is the current ordered set of *n* sampled and/or ancestral sequences (s_1, s_2, ..., s_n) during a particular coalescent interval in the genealogy
s_i, s_j	Sampled and/or ancestral sequences *i* and *j* (where $i \neq j$)
$s_i \sim s_j$	Signifies that the available regions of sampled and/or ancestral sequences *i* and *j* are at least compatible and that the two are therefore combinable (i.e., can coalesce)
$\|s_i\|$	Measures the relative degree to which sampled or ancestral sequence *i* is a complete or partial sequence. Thus, $\Sigma_i \|s_i\|$ summarizes the total available length of all sampled and/or ancestral sequences during a particular coalescent interval.
$\|S\|$	Summarizes the current number of sampled and/or ancestral sequences during a particular coalescent interval in the genealogy

identical) sequences. Such sequences are referred to as combinable ($s_i \sim s_j$).

The first term in the bottom line before the single sum is then for the probability that the next observed event is instead a mutation. As indicated by this sum, if indeed a mutation occurs, then it will happen to a sampled or ancestral sequence with at least one "singleton" ($\sigma_i > 0$). Here, the definition of a "singleton" is expanded to include the derived mutations of the common ancestors for the different groups of related sampled sequences as well as those for the singletons (in the strict sense) of the original dataset (Table 1). This expanded use of the term acknowledges that a shared polymorphism under the infinite sites model is due to a unique mutation within the common ancestor of those sampled sequences sharing the derived base (Figure 3). Thus, even though they result in shared polymorphisms among the sampled sequences, these derived mutations are ultimately counted as "singletons" as one works backwards in time. This expanded definition of a "singleton" allows for the economical use of σ_i alone to track the mutations of both the original shared polymorphisms and singletons.

In the KM model, the available region(s) of each sampled sequence is summarized as a closed interval(s) that is scored over the range of (0:1). The $|s_i|$ factor then quantifies the total amount of sequence available for this sampled allele. For example, the (0.2:1.0) score for the leftmost (first) sampled sequence in Figure 4a indicates that it is lacking the initial 20% of the full multiple alignment. Thus, $|s_i| = 0.8$ for this partial, leftmost, sampled sequence. In turn, as one works backwards in time, the closed intervals for the available regions of the sequences for common ancestors are calculated as the union of the known lengths for their two immediate descendants. Thus, the closed interval of the common ancestor for the two leftmost sampled sequences in Figure 4a is (0.2:1.0) ∪ (0.0:1.0) = (0.0:1.0), with $|s_i| = 1.0$.

The purpose of $\Sigma_i |s_i|$ in Equation (13) is to track the total available length of all sampled and/or ancestral sequences for the detection of mutations as one works backwards in time. The corresponding use of $|s_i|$ in the bottom line of Equation (13) allows for the adjustment of σ_i due to unobserved polymorphisms resulting from missing data.

The KM model uses only Equation (13) when working with error free sequences. In turn, this model also relies on equation (4) of Knudsen and Miyamoto [2] when dealing with error prone sequences. This additional equation of the KM model assumes that the random errors are uniformly distributed along the sampled sequences and that their total number is Poisson distributed with an intensity of $\lambda = \varepsilon \Sigma_i |s_i|$. The inclusion of $\Sigma_i |s_i|$ in the calculation of λ allows for the adjustment of this error rate to account for incomplete sampled sequences.

In contrast to its predecessor, the KM' model uses only Equation (13) when working with either error prone or error free sequences. As for the new Watterson and Tajima estimators, the KM' model operates by counting only those internal branches of the genealogy where a mutation will result in a shared polymorphism (Figure 4b). The sampled sequences are rescored as unknowns with empty intervals (∅), as is the sequence of the common ancestor for the (n - 1) basal group of each asymmetrical genealogy (Figure 2). Correspondingly, $|s_i|$ is then reset to 0.0 for these sequences. The KM' model ignores the external and basal branches of the genealogy, where a mutation will result in a singleton, by multiplying σ_i for each sampled and/or (n - 1) ancestral sequence by $|s_i| = 0.0$ in Equation

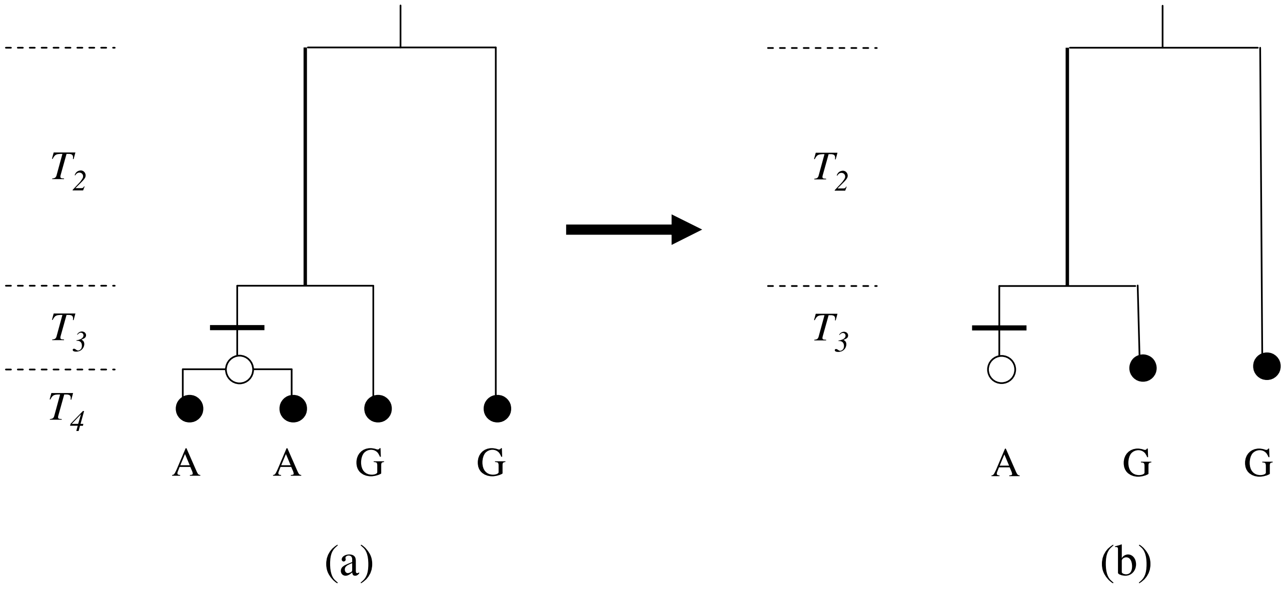

Figure 3

Illustration of how a shared polymorphism among the sampled sequences becomes counted as a "singleton" as one works backwards in time. In (a), the A/G bases for a shared polymorphism are shown for four sampled sequences. The three coalescent time intervals of this asymmetrical genealogy (T_j, with j = 4, 3, or 2 sampled and/or ancestral sequences) are labeled on the left. The solid bar marks the G to A mutation in the common ancestor of the two leftmost sampled sequences (open dot), which results in the A/G shared polymorphism. The first event is the coalescence of the two leftmost sampled sequences in T_4 (a). This coalescence reduces the number of alleles to three as the two leftmost sampled sequences in T_4 are replaced by their common ancestor in T_3 (b). In the process, their shared polymorphism is replaced by the "singleton" of their common ancestor. This replacement is the basis of the expanded definition for a "singleton" (Table 1).

(13). In this way, the KM' model discounts the singletons in favor of the shared polymorphisms within the dataset without the use of equation (4) and its associated parameters (e.g., λ and ε) of the original KM model.

For sampled sequences without missing data, the previous explanation is complete as to how the KM' model corrects for the branches of the genealogy where a mutation will result in a singleton. However, for incomplete sampled sequences, the "above and below" test of the KM' model also becomes necessary to correct for the regions of each common ancestral sequence where a mutation in its internal branch will lead to a singleton, rather than to a shared polymorphism, because of this missing information (Figure 4b). In this test, "below" refers to those sampled sequences that belong to the monophyletic group of the common ancestor in question. "Above" then corresponds to the remaining, more distantly related, sampled sequences. In Figure 4b, the two leftmost sampled sequences are direct descendants of the first, leftmost, common ancestor, whereas the other three are not. Thus, these two sampled sequences lie "below," whereas the other three occur "above" the internal branch for this leftmost common ancestor.

The above and below test checks whether comparable information for a region is known for at least two, descendant, sampled sequences below and at least two, more distantly related, sampled sequences above an internal branch in the genealogy. If not, then a mutation within this region of the ancestral sequence that corresponds to this internal branch will result in a false singleton within the original dataset. For example, a mutation within the first 20% of the leftmost common ancestor in Figure 4 will result in a false singleton of the second sampled sequence rather than in a true shared polymorphism of the first and second. The problem is that the leftmost sampled sequence is missing comparable information for the detection of this mutation as a shared change in the leftmost common ancestor. The above and below test corrects for such regions of the common ancestral sequences during the rescoring of their closed intervals and $|s_i|$. As the first 20% of the leftmost common ancestor fails the below half of this test, the KM' model disregards this region during the recalculation of its closed interval as (0:2:1.0) and $|s_i|$ = 0.8 (Figure 4b).

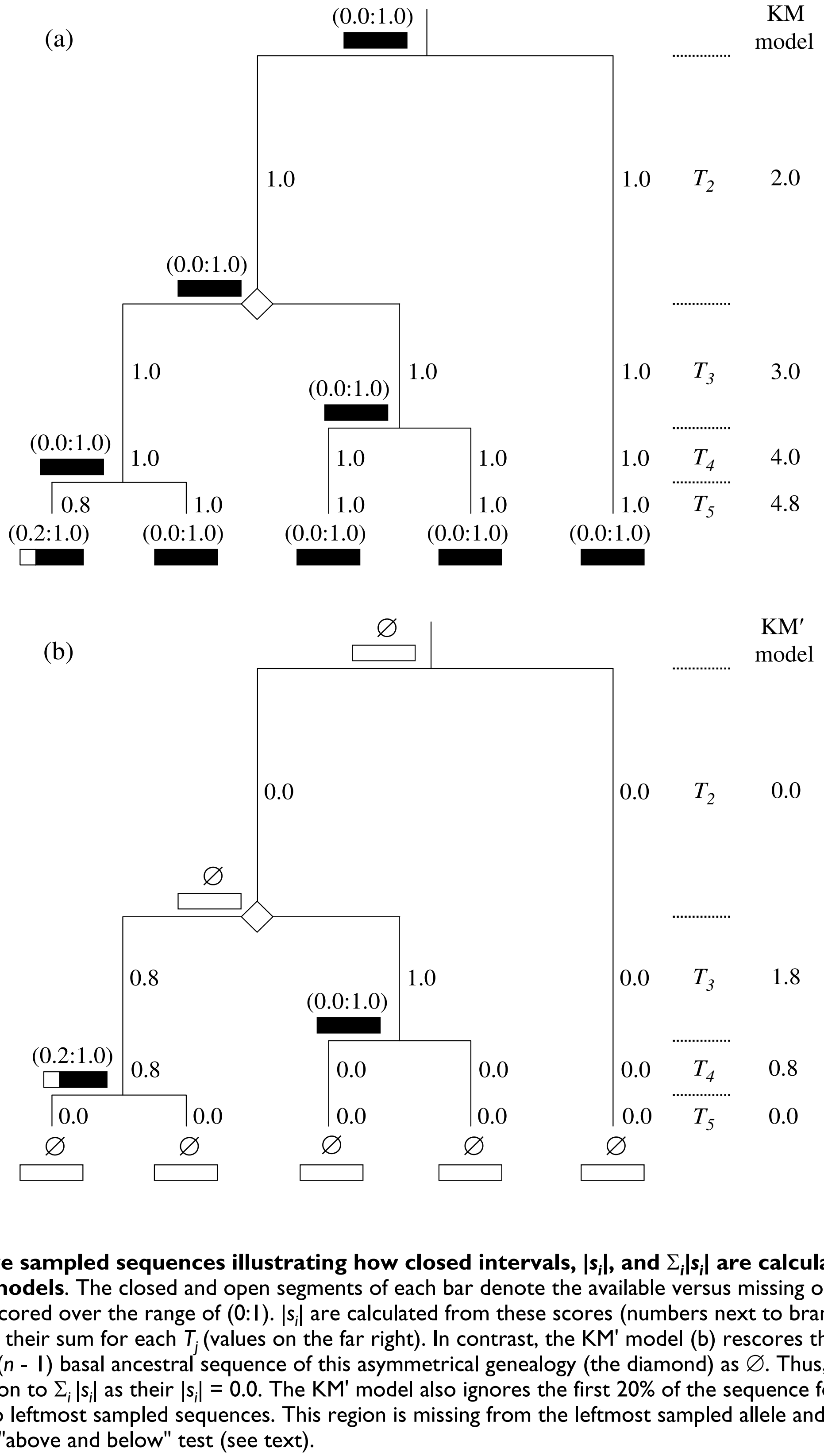

Figure 4

Genealogy for five sampled sequences illustrating how closed intervals, $|s_i|$, and $\Sigma_i|s_i|$ are calculated by the KM (a) and KM' (b) models. The closed and open segments of each bar denote the available versus missing or ignored regions of its sequence as scored over the range of (0:1). $|s_i|$ are calculated from these scores (numbers next to branches) and $\Sigma_i\,|s_i|$ is then determined as their sum for each T_j (values on the far right). In contrast, the KM' model (b) rescores the sampled sequences and the (*n* - 1) basal ancestral sequence of this asymmetrical genealogy (the diamond) as ∅. Thus, these sequences make no contribution to $\Sigma_i\,|s_i|$ as their $|s_i|$ = 0.0. The KM' model also ignores the first 20% of the sequence for the common ancestor of the two leftmost sampled sequences. This region is missing from the leftmost sampled allele and thereby fails the "below" half of the "above and below" test (see text).

Evolutionary simulations and A. californica dataset

Evolutionary simulations

To test the three new procedures, evolutionary simulations were conducted according to standard methods [4]. Two hundred datasets apiece were simulated for eight or 16 sequences of length 500 from a single population under the baseline conditions of the standard FW and infinite sites models with θ = 1, 2, 4, or 8. In addition to these baseline conditions, sequence errors were introduced as four or eight randomly placed changes among the eight and 16 sequences, respectively, for an expected ε of 0.5. Estimates of θ were then obtained for the 200 datasets of each tested combination with the FW model and the original and new Watterson estimators, Tajima estimators, and KM and KM' models.

As expected, the FW model and original Watterson and Tajima estimators overestimate θ when the datasets contain random errors (Table 2). In these cases, the random errors are counted as mutations, thereby inflating their estimates of θ. Furthermore, these overestimations are greater for the Watterson estimator than for the Tajima estimator. The reason is that each singleton makes the same contribution as a shared polymorphism to the number of segregating sites (n_s) in Equation (4) of θ_W, but a smaller one to the sum of the pairwise differences in Equation (10) of θ_T. Specifically, each singleton is limited in Equation (10) to the (n - 1) pairwise comparisons of the unique sequence to the (n - 1) other sequences. Thus, the Tajima estimator is less vulnerable than the Watterson estimator to random errors even though its vulnerability is still significant.

In contrast, the KM model underestimates θ when the sequences are errorless (Table 2). Furthermore, this tendency to underestimate θ is also evident (although not significant) when the sequences contain errors (simulations A2, A4, B2, and B4; see also [2] for additional cases). In these situations, mutations are sometimes counted as random errors, thereby deflating their estimates of θ. In cases where few to no random errors are expected (i.e., finished sequences), one can first perform a likelihood ratio test to determine if ε = 0 [39]. If this null hypothesis cannot be rejected, then the KM analysis should be restricted to only Equation (13). However, if ε > 0 according to this likelihood ratio test, then equation (4) of Knudsen and Miyamoto [2] is also needed and the user of the KM model must remain aware that her/his estimate of θ most likely includes some slight downward bias.

Unlike their original versions, the new Watterson estimator, Tajima estimator, and KM' model all consistently recover the true θ in the simulations both with and without errors (Table 2). Furthermore, these three methods also avoid the tendency of the KM model to underestimate θ due to unnecessary or "greedy" parameters for random errors. In particular, the reliance of the KM' model on one less equation [(4)] and fewer parameters (e.g., λ and ε) makes it much simpler and less prone to overparameterizations than its predecessor.

Table 2: Results of the evolutionary simulations

Evolutionary simulations			Coalescent/mutation models			Watterson and Tajima estimators			
θ	*n*	ε	θ_{FW}	θ_{KM}	θ'_{KM}	θ_W	θ'_W	θ_T	θ'_T
A1: 1	8	0.0	0.96 ± 0.10	**0.76 ± 0.10**	1.10 ± 0.15	0.98 ± 0.10	1.12 ± 0.16	1.02 ± 0.12	1.10 ± 0.15
A2: 1	8	0.5	*3.49 ± 0.23*	0.90 ± 0.14	0.96 ± 0.15	*2.63 ± 0.15*	0.94 ± 0.15	*2.04 ± 0.13*	0.95 ± 0.16
A3: 1	16	0.0	1.02 ± 0.10	**0.84 ± 0.09**	1.01 ± 0.12	1.00 ± 0.09	0.99 ± 0.11	0.97 ± 0.10	0.96 ± 0.12
A4: 1	16	0.5	*5.10 ± 0.28*	0.99 ± 0.12	1.02 ± 0.13	*3.43 ± 0.16*	1.03 ± 0.13	*2.01 ± 0.13*	1.01 ± 0.13
B1: 2	8	0.0	1.92 ± 0.17	**1.51 ± 0.17**	1.80 ± 0.25	1.92 ± 0.18	1.81 ± 0.25	1.89 ± 0.20	1.82 ± 0.26
B2: 2	8	0.5	*4.38 ± 0.26*	1.80 ± 0.20	1.92 ± 0.26	*3.44 ± 0.19*	1.91 ± 0.26	*2.91 ± 0.20*	1.92 ± 0.25
B3: 2	16	0.0	2.06 ± 0.16	**1.78 ± 0.15**	2.02 ± 0.19	2.04 ± 0.16	2.04 ± 0.20	2.09 ± 0.19	2.10 ± 0.22
B4: 2	16	0.5	*6.19 ± 0.33*	1.88 ± 0.17	1.92 ± 0.19	*4.32 ± 0.20*	1.92 ± 0.19	*2.95 ± 0.20*	1.96 ± 0.22
C1: 4	8	0.0	-	-	4.05 ± 0.44	4.27 ± 0.35	4.27 ± 0.53	4.31 ± 0.39	4.32 ± 0.54
C2: 4	8	0.5	-	-	3.89 ± 0.46	*5.62 ± 0.33*	4.09 ± 0.53	*5.12 ± 0.38*	4.13 ± 0.54
C3: 4	16	0.0	-	-	4.00 ± 0.31	4.12 ± 0.28	4.21 ± 0.38	4.24 ± 0.36	4.28 ± 0.41
C4: 4	16	0.5	-	-	3.82 ± 0.31	*6.23 ± 0.27*	3.87 ± 0.35	*4.84 ± 0.32*	3.86 ± 0.38
D1: 8	8	0.0	-	-	7.88 ± 0.71	7.75 ± 0.49	7.87 ± 0.78	7.78 ± 0.55	7.85 ± 0.77
D2: 8	8	0.5	-	-	7.90 ± 0.85	*9.67 ± 0.63*	8.38 ± 1.03	*9.22 ± 0.73*	8.39 ± 1.03
D3: 8	16	0.0	-	-	7.82 ± 0.54	8.08 ± 0.49	8.20 ± 0.72	8.26 ± 0.69	8.32 ± 0.82
D4: 8	16	0.5	-	-	8.00 ± 0.58	*10.73 ± 0.52*	8.31 ± 0.71	*9.57 ± 0.68*	8.60 ± 0.81

These results are for the standard FW model (θ_{FW}) and the original and new Watterson estimators (θ_W and θ'_W), Tajima estimators (θ_T and θ'_T), and KM and KM' models (θ_{KM} and θ'_{KM}). They are summarized as the means ± twice their standard deviations for 200 simulated datasets apiece for each of the 16 different combinations of θ, *n* (number of sampled sequences), and ε (A1 to D4). Mean estimates that are significantly greater than or less than the true θ are highlighted in italics and boldface, respectively. No results are provided for the FW and KM models with θ = 4 or 8 (lower left corner), since these calculations are too computationally intensive (see text).

The mean estimates of θ for the new Watterson estimator, Tajima estimator, and KM' model are also generally associated with greater standard deviations than those for their original versions (Table 2). The proximal reason for these increased standard deviations is that the elimination of singletons results in the loss of some actual mutations along with the random errors. However, this cost of the three new methods appears to be relatively small, given that their standard deviations are never twice as great as those for their original versions (with these discrepancies usually being much smaller). Indeed, the standard deviations for the three new methods are less than or equal to those of their counterparts in four cases (simulations A2, A4, and B4 for θ'_W and A4 for θ'_T in Table 2). Perhaps more surprising is that the standard deviations for the new Watterson and Tajima estimators are comparable to those for the KM' model, particularly when θ = 1 or 2. These similarities in their standard deviations are surprising given that the KM' model is a full ML model.

As summary statistics, the original and new Watterson and Tajima estimators are all computationally fast, requiring much less than 1 CPU second on a 2.4 GHz Pentium 4 CPU to analyze each of the simulated datasets. More importantly, the KM' model is much faster than the FW and KM models as documented by its completed analyses of the more complex datasets (simulations C1 to C4 and D1 to D4 in Table 2). In contrast, the FW and KM models fail to complete their analyses of these more complex datasets due to time and memory constraints. The KM' model is much faster than the FW and KM models, because it relies on only shared polymorphic sites. Less variable positions results in faster coalescences and fewer choices as one works back through the coalescent/mutation recursion of Equation (13).

Aplysia dataset

The *A. californica* dataset was the original motivating force behind the development of the KM model [2]. Thus, this real dataset was also analyzed with the KM' model to test further its performance against that of its predecessor. This dataset consists of 18 sequencing reads for six diploid individuals from a laboratory population of the California seahare at the Laboratory for Marine Bioscience, University of Florida (LL Moroz and AB Kohn, unpublished data). Three cloned inserts were sequenced from each individual as a pair of single sequencing passes starting from both ends of an internal segment of 1731 base pairs for the protein-coding region of the nuclear *FMRF* gene. These pairs of passes overlap in the middle for nine sequences, but at most by only 58 bases. Thus, these 18 essentially single-pass sequences contain many random errors and some missing data and their assignments to the two homologues of each diploid remain uncertain (even though their individual sources are known). Correspondingly, the KM analysis of this dataset with 44 singletons and 10 shared polymorphisms required the full use of its factors for experimental error and design (i.e., for random errors, missing data, and uncertain homologue assignments).

A full ML analysis of this complex dataset by the KM model proved too time and memory consuming and a two step procedure was therefore adopted instead whereby the number of errors was first ML estimated followed by the determination of θ for this ML value [2]. This heuristic approach resulted in an estimate of θ_{KM} = 6.32 for this sample, which still took more than 12 hours to complete on the same CPU as for the simulations (see above). In contrast, a full ML analysis of this dataset by the KM' model using the same factors for experimental design is much faster as it took less than 1 hour on this CPU to obtain a similar estimate of θ'_{KM} = 7.17. Correspondingly, nucleotide diversity (π) for this sample is calculated as (7.17/1731 positions) = 0.0041 mutations per site. To summarize, the KM and KM' models both support similar estimates of θ for this real dataset, but the latter (as in the evolutionary simulations in Table 2) is much faster as confirmed by its completed, full, ML analysis.

Conclusion

Expressed sequence tags, low coverage draft sequences, and other such unfinished products are known to contain random errors due to chance accidents during their data collection and recording [1,11,21,23]. However, such sequences often constitute the only available allelic information for a population and methods to deal with their random mistakes are therefore needed to obtain accurate estimates of θ from these error prone data. This study is based on the simple premise that random sequence errors are distributed as singletons. Thus, one can avoid the random mistakes of error prone sequences by ignoring their singletons in favor of their shared polymorphisms (Figures 2 and 4).

This strategy is implemented in the new Watterson estimator, Tajima estimator, and KM' model. These new methods are all accurate and fast according to their evolutionary simulations and analysis of the real complex dataset for *A. californica* (Table 2). These methods come with the cost of increased standard deviations, but this price appears small or even negligible compared to their advantages of significantly improved accuracy and/or computational speed. Obviously, additional evolutionary simulations and applications to real datasets are now needed to evaluate more fully under what conditions the removal of singletons is warranted in light of this tradeoff. Nevertheless, the current successes with the new Watterson estimator, Tajima estimator, and KM' model support our recommendation that these three methods be given serious consideration when estimating θ from error prone sequences.

Unlike the Watterson and Tajima estimators that represent summary statistics, the KM and KM' models both constitute full ML models that offer the framework for the further incorporation of other experimental and population genetic factors [2]. In particular, such further developments are encouraged for the KM' model in light of its current successes in the evolutionary simulations and *A. californica* analysis (Table 2). For example, biased errors within a sample due to the systematic misreading of specific bases during DNA sequencing, the postmortem biochemical degradation of ancient DNA, and/or other such sources of error can be accommodated by a finite sites process that allows for repeated mistakes as well as mutations at the same sequence positions [14-16,40]. Likewise, additional factors to account for other population genetic processes such as recombination (which is most likely the most important parameter overlooked in this study of the nuclear *FMRF* gene for *A. californica*) can be accommodated by the use of ancestral recombination graphs [41]. Inevitably, these more complex versions of the KM' model will require the use of sampling based procedures for their implementation (e.g., Markov chain Monte Carlo approximations), since the current use of direct ML evaluation will remain practical for only the smaller datasets and simpler models [14,42-44].

Authors' contributions

BK derived the equations, wrote the computer program for the estimation of θ, and conducted the evolutionary simulations. MMM provided biological interpretations about the results. Both authors contributed to the design and main ideas of this study, wrote and edited the manuscript, and read and approved its final draft.

Acknowledgements

We thank Leonid L Moroz and Andrea B Kohn for the use of their unpublished *A. californica* dataset; William Farmerie and Michele R Tennant for their suggestions about this research; the Carlsberg Foundation for its fellowship to BK; and the University of Florida for its support. A copy of our computer program to estimate θ'_W, θ'_T, and θ'_{KM} is available upon request from BK.

References

1. Achaz G: **Testing for neutrality in samples with sequencing errors.** *Genetics* 2008, **179:**1409-1424.
2. Knudsen B, Miyamoto MM: **Incorporating experimental design and error into coalescent/mutation models of population history.** *Genetics* 2007, **176:**2335-2342.
3. Fu YX: **Statistical properties of segregating sites.** *Theoretical Population Biology* 1995, **48:**172-197.
4. Hein J, Shierup M, Wiuf C: *Sequence Variation, Genealogies and Evolution* New York: Oxford University Press; 2005.
5. Wakeley J: *Coalescent Theory: An Introduction* Greenwood Village: Roberts & Company Publishers; 2008.
6. Hudson RR: **Gene genealogies and the coalescent process.** In *Oxford Surveys in Evolutionary Biology Volume 7*. Edited by: Antonovics J, Futuyma D. Oxford: Oxford University Press; 1990:1-44.
7. Fu YX: **Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences.** *Genetics* 1994, **138:**1375-1386.
8. Felsenstein J: *Inferring Phylogenies* Sunderland: Sinauer Associates; 2004.
9. Li WH, Sadler LA: **Low nucleotide diversity in man.** *Genetics* 1991, **129:**513-523.
10. Clark AG, Whittam TS: **Sequencing errors and molecular evolutionary analysis.** *Molecular Biology and Evolution* 1992, **9:**744-752.
11. Tiffin P, Gaut BS: **Molecular evolution of the wound-induced serine protease inhibitor *wip1* in *Zea* and related genera.** *Molecular Biology and Evolution* 2001, **18:**2092-2101.
12. Nabholz B, Jean-Francois M, Bazin E, Galtier N, Glemin S: **Determination of mitochondrial genetic diversity in mammals.** *Genetics* 2008, **178:**351-361.
13. Ho SY, Phillips MJ, Cooper A, Drummond AJ: **Time dependency of molecular rate estimates and systematic overestimation of recent divergence times.** *Molecular Biology and Evolution* 2005, **22:**1561-1568.
14. Ho SY, Heupink TH, Rambaut A, Shapiro B: **Bayesian estimation of sequence damage in ancient DNA.** *Molecular Biology and Evolution* 2007, **24:**1416-1422.
15. Johnson PL, Slatkin M: **Inference of population genetic parameters in metagenomics: A clean look at messy data.** *Genome Research* 2006, **16:**1320-1327.
16. Johnson PL, Slatkin M: **Accounting for bias from sequencing error in population genetic estimates.** *Molecular Biology and Evolution* 2008, **25:**199-206.
17. Axelsson E, Willerslev E, Gilbert MTP, Nielsen R: **The effect of ancient DNA damage on inferences of demographic histories.** *Molecular Biology and Evolution* 2008, **25:**2181-2187.
18. Burgess R, Yang Z: **Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors.** *Molecular Biology and Evolution* 2008, **25:**1979-1994.
19. Lynch M: **Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects.** *Molecular Biology and Evolution* 2008, **25:**2409-2419.
20. Lottaz C, Iseli C, Jongeneel CV, Bucher P: **Modeling sequencing errors by combining hidden Markov models.** *Bioinformatics* 2003, **19:**i103-i112.
21. Long AD, Beldade P, Macdonald SJ: **Estimation of population heterozygosity and library construction-induced mutation rate from expressed sequence tag collections.** *Genetics* 2007, **176:**711-714.
22. Slatkin M, Pollack JL: **Subdivision in an ancestral species creates asymmetry in gene trees.** *Molecular Biology and Evolution* 2008, **25:**2241-2246.
23. Wesche PL, Gaffney DJ, Keightley PD: **DNA sequence error rates in Genbank records estimated using the mouse genome as a reference.** *DNA Sequence* 2004, **15:**362-364.
24. Shendure JA, Porreca GJ, Church GM: **Overview of DNA sequencing strategies.** *Current Protocols in Molecular Biology* 2008, **81:**7.1.1-7.1.11.
25. Watterson GA: **On the number of segregating sites in genetical models without recombination.** *Theoretical Population Biology* 1975, **7:**256-276.
26. Tajima F: **Evolutionary relationship of DNA sequences in finite populations.** *Genetics* 1983, **105:**437-460.
27. Kimura M: **The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations.** *Genetics* 1969, **61:**893-903.
28. Kimura M: *The Neutral Theory of Molecular Evolution* Cambridge: Cambridge University Press; 1983.
29. Kingman JFC: **The coalescent.** *Stochastic Processes and their Applications* 1982, **13:**235-248.
30. Donnelly P, Tavaré S: **Coalescents and genealogical structure under neutrality.** *Annual Review of Genetics* 1995, **29:**401-421.
31. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133:**693-709.
32. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585-595.
33. Simonsen KL, Churchill GA, Aquadro CF: **Properties of statistical tests of neutrality for DNA polymorphism data.** *Genetics* 1995, **141:**413-429.

34. Fu YX: **Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection.** *Genetics* 1997, **147:**915-925.
35. Fisher RA: *The Genetical Theory of Natural Selection* Oxford: Clarendon Press; 1930.
36. Wright S: **Evolution in Mendelian populations.** *Genetics* 1931, **16:**97-159.
37. Griffiths RC: **Genealogical-tree probabilities in the infinitely-many-sites model.** *Journal of Mathematical Biology* 1989, **27:**667-680.
38. Griffiths RC, Tavaré S: **Unrooted genealogical tree probabilities in the infinitely-many-sites model.** *Mathematical Biosciences* 1995, **127:**77-98.
39. Huelsenbeck JP, Rannala B: **Phylogenetic methods come of age: Testing hypotheses in an evolutionary context.** *Science* 1997, **276:**227-232.
40. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: **Accuracy and quality of massively parallel DNA pyrosequencing.** *Genome Biology* 2007, **8:**R143.1-R143.9.
41. Griffiths RC: **Ancestral inference from gene trees.** In *Genes, Fossils, and Behaviour: An Integrated Approach to Human Evolution* Edited by: Donnelly P, Foley R. Amsterdam: IOS Press; 2001:137-172.
42. Kuhner MK, Felsenstein J: **Sampling among haplotype resolutions in a coalescent-based genealogy sampler.** *Genetic Epidemiology* 2000, **19:**S15-S21.
43. Nielsen R: **Estimation of population parameters and recombination rates from single nucleotide polymorphisms.** *Genetics* 2000, **154:**931-942.
44. Beerli P: **Comparison of Bayesian and maximum-likelihood inference of population genetic parameters.** *Bioinformatics* 2006, **22:**341-345.
45. Edwards AW: **Estimation of the branch points of a branching diffusion process.** *Journal of the Royal Statistical Society* 1970, **32:**155-174.