Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3

Mira V. Han,*1,2 Gregg W.C. Thomas,2 Jose Lugo-Martinez,2 and Matthew W. Hahn*2,3

1National Evolutionary Synthesis Center, Durham, North Carolina
2School of Informatics and Computing, Indiana University
3Department of Biology, Indiana University

*Corresponding authors: E-mail: mira.han@nescent.org; mwh@indiana.edu.

Abstract

Current sequencing methods produce large amounts of data, but genome assemblies constructed from these data are often fragmented and incomplete. Incomplete and error-filled assemblies result in many annotation errors, especially in the number of genes present in a genome. This means that methods attempting to estimate rates of gene duplication and loss often will be misled by such errors and that rates of gene family evolution will be consistently overestimated. Here, we present a method that takes these errors into account, allowing one to accurately infer rates of gene gain and loss among genomes even with low assembly and annotation quality. The method is implemented in the newest version of the software package CAFE, along with several other novel features. We demonstrate the accuracy of the method with extensive simulations and reanalyze several previously published data sets. Our results show that errors in genome annotation do lead to higher inferred rates of gene gain and loss but that CAFE 3 sufficiently accounts for these errors to provide accurate estimates of important evolutionary parameters.

Key words: duplication, gene family, adaptive evolution.

Introduction

Genome sequencing projects have revealed large and frequent changes between species in the size of gene families (e.g., Gibbs et al. 2004, 2007; Drosophila 12 Genomes Consortium 2007; Li et al. 2009; Floudas et al. 2012). These changes may underlie many important morphological, physiological, and behavioral differences between species and contribute much of the genetic and genomic diversity observed in nature (reviewed in Demuth and Hahn 2009). Recent work on diversity within species has also revealed surprising numbers of polymorphic gene duplications and losses (e.g., Sebat et al. 2004; Emerson et al. 2008; Kidd et al. 2008; Schrider et al. 2011), variation that contributes to long-term differences in the size of gene families between species. To further understand the importance of these changes, researchers must be able to accurately estimate the rate at which gene families evolve over time.

Our previous approach to estimating this rate modeled the gain and loss of genes within a gene family using a birth-and-death stochastic process (Hahn et al. 2005). (This probability distribution should not be confused with the birth-and-death conceptual model of gene family evolution of Nei and Rooney [2005].) Given input data on the size of gene families across multiple species and an ultrametric phylogenetic tree describing relationships among these species, the original CAFE software package (De Bie et al. 2006) can estimate the maximum likelihood value of the rate parameter, λ, and the maximum likelihood estimates (MLEs) of the size of each gene family at ancestral nodes of the tree. These MLEs can then be used to infer expansions and contractions of individual gene families on any lineage (e.g., Demuth et al. 2006). Updated versions of this software (CAFE 2; Hahn, Demuth, et al. 2007; Hahn, Han, et al. 2007) allowed for separate λ values on different branches of the tree, as well as several other novel features. A number of other programs using the birth-and-death model—or related models—have also appeared and offer similar as well as additional features (e.g., Liu et al. 2011; Ames et al. 2012; Librado et al. 2012).

A major concern when studying changes in gene family size is the quality of the underlying genome assembly and genome annotation. Low sequencing coverage in genome assemblies can lead to both the erroneous addition and subtraction of genes. Genes can be missing because there is incomplete coverage of the entire genome, with whole or parts of genes falling in unassembled portions of the genome; genes can also be missing because base-calling errors mistakenly indicate frameshifts or nonsense mutations (e.g., Hubisz et al. 2011). Extra gene copies can be inserted into the assembly if allelic diversity is incorrectly assembled as duplicated loci (e.g., Holt et al. 2002; Colbourne et al. 2011) or if a single multiexon gene is split among multiple scaffolds or contigs—in which case multiple gene models may be predicted from a single gene (e.g., Colbourne et al. 2011). Similar problems can arise even in “finished” genomes (such as Drosophila melanogaster), as gene annotation software can often miss short open-reading frames or can cleave a single gene into multiple predicted genes (e.g., Hahn, Han, et al. 2007; Stark et al. 2007; Costello et al. 2008).

For studies focusing on gene family size change, errors in genome assembly and annotation will result in biased
estimates of the rate of change. Because a higher rate of evolution must be proposed in the presence of errors—whether additional or missing genes—estimates that do not account for errors are likely to have been upwardly biased. Indeed, we have previously found that *Drosophila* species represented by the lowest quality assemblies in comparative analyses using CAFE also appear to evolve at the highest rates (Hahn, Han, et al. 2007). Therefore, to estimate accurate rates of gene family evolution, we must be able to account for the error present in all current genome annotations. In this article, we present one such method and implement it in a new version of CAFE. Our method accounts for errors in gene family sizes by explicitly modeling the uncertainty associated with observed family sizes at the tips of a tree. We show that, given a known error distribution for each genome, we can recover accurate estimates of the true rate of gene family evolution. In addition, we present multiple methods for estimating error rates from the data when they are not known in advance and show how these can be used to provide more accurate values of evolutionary parameters.

**New Approaches**

We assume a random variable X that is a true count of homologous members of a gene family within a single lineage. In theory, X can be from 0 to infinite size, but for ease of computation, we limit it to be at most M. A whole genome can then be thought of as a random sample of size N, where each gene family within a genome corresponds to each observation, and N is the total number of gene families found in the genome. Each gene family size in a genome is assumed to be independent and identically distributed.

We also consider the error-prone measure of gene family size W, W = w (w = 0,1,2,3 \ldots M). W represents our observation for each gene family size that is affected by errors in the genome assembly and errors in the gene annotations. The measurement-error model, which describes the behavior of W given X = x, is specified by the error probabilities:

\[ \theta_{w|X} = P(W = w | X = x), \]

that is, the probability of observing w when the true gene family size is x. The error probabilities can be represented as a matrix, Θ:

\[ \Theta = \begin{bmatrix} \theta_{1|1} & \cdots & \theta_{1|M} \\ \vdots & \ddots & \vdots \\ \theta_{M|1} & \cdots & \theta_{M|M} \end{bmatrix}, \]

where the item in the ith row and jth column represents the probability of observing i when the true gene family size is j. Note that the rows of the matrix do not have to sum to 1 but the columns do. We also define the probability \( \theta_x \) as:

\[ \theta_x = P(X = x), \]

that is, the probability of a true gene family size of x found in the genome. The lower case \( \theta \) denotes the discrete probability distribution \( \theta_x \) for \( x = 0 \ldots M \).

In cases where there is no measurement error, it is known that we can estimate the rate of change in gene family size across the phylogeny by specifying a transition matrix based on the rate parameters \( \lambda \) and \( \mu \) and fitting the model to the observed (=true) counts (X) at the tips of the tree and the time between the nodes (described by branch lengths, T). Under a birth-and-death process, the probability of going from s number of genes to c number of genes in time t is given by (Bailey 1964):

\[ P(X(t) = c | X(0) = s) = \sum_{j=0}^{\min(s,c)} \left( \begin{array}{c} s + c - j - 1 \\ s - 1 \end{array} \right) \alpha^{s-

where \( \lambda \) is the rate of gene gain and \( \mu \) is the rate of gene loss. If the rates of gain and loss are equal, that is, \( \lambda = \mu \), the above probability is as given in equation 1 of Hahn et al. (2005). Here, we focus on cases with \( \lambda = \mu \), but the updated version of CAFE can also estimate separate rates of gain and loss (as can BadiRate; Librado et al. 2012).

For multiple species, \( S = (1 \ldots s) \), we define a tree with ultrametric branch lengths, T, that has the set of species \( S \) as the leaves and a set of ancestral nodes, \( U = (1 \ldots u) \). We define \( X_n \) as the vector \( X_n = (X_{n1}, X_{n2}, \ldots X_{nm}) \), in which each item \( X_{ni} \) describes the size of the nth gene family in each genome of species i (\( i \in S \)). Similarly, \( Z_n = (Z_{n1}, Z_{n2}, \ldots Z_{nm}) \) is the vector in which each item \( Z_{ni} \) is the gene family size of the ancestral genome at the inner node \( j \) (\( j \in U \)). The actual calculation of the likelihood over the whole tree utilizes the "pruning algorithm" (Felsenstein 1973, 1981) to sum over the inner node values that we cannot observe:

\[ \hat{\lambda}, \mu = \arg\max_{\hat{\lambda}, \mu} \prod_{n=1}^{N} P(X_n | \hat{\lambda}, \mu, T) \]

\[ = \arg\max_{\hat{\lambda}, \mu} \prod_{n=1}^{N} P(X_n | Z_n, \hat{\lambda}, \mu, T) \cdot P(Z_n | \lambda, \mu, T) \]

\[ = \arg\max_{\hat{\lambda}, \mu} \prod_{n=1}^{N} \left( \sum_{z_{n1}=0}^{M} \sum_{z_{n2}=0}^{M} \cdots \sum_{z_{nm}=0}^{M} P(X_n | Z_n) \right) \]

\[ = \arg\max_{\hat{\lambda}, \mu} \left( \prod_{n=1}^{N} P(Z_n | \hat{\lambda}, \mu, T) \right) \]

With error in the measurements, a naïve inference based on use of the W’s instead of the X’s leads to:

\[ \hat{\lambda}, \mu_W = \arg\max_{\hat{\lambda}, \mu} \prod_{n=1}^{N} P(W_n | \hat{\lambda}, \mu, T) \]

where we define the vector \( W_n = (W_{n1}, W_{n2}, \ldots W_{nm}) \). Similar to \( X_n \), \( W_n \) is the error-prone measurement of the gene count for the nth gene family in species i.

To account for error, we add an additional layer of uncertainty on the true value X to the values at the leaf nodes of the phylogeny. This necessitates an additional summation of
likelihoods at each leaf over \( X \), in addition to all internal nodes, \( Z \). The only difference between the summation at the leaf nodes and the summation at the inner nodes is that the probability at the leaf nodes is defined by the error matrix, rather than following the transition probabilities derived from the rate matrix and the branch lengths:

\[
\lambda, \mu = \arg\max_{\lambda, \mu} \left\{ \prod_{n=1}^{N} \left[ \sum_{x_1=0}^{M} \sum_{x_2=0}^{M} \cdots \sum_{x_N=0}^{M} P(W_n | X_n) \right] \right\}
\]

\[
= \arg\max_{\lambda, \mu} \left\{ \prod_{n=1}^{N} \left[ \sum_{x_1=0}^{M} \sum_{x_2=0}^{M} \cdots \sum_{x_N=0}^{M} P(W_{n1} | X_{n1}) \right] \right\}
\]

The probability \( P(W_{n1} = w_{n1} | X_{n1} = x_{n1}) \) follows from the error matrix \( \Theta_{w_{n1}|x_{n1}} \).

Because we do not know the error matrix \( \Theta \), it becomes an additional set of parameters we need to estimate:

\[
(\lambda, \mu, \Theta) = \arg\max_{\lambda, \mu, \Theta} \left\{ \prod_{n=1}^{N} \left[ \sum_{x_1=0}^{M} \sum_{x_2=0}^{M} \cdots \sum_{x_N=0}^{M} \Theta_{w_{n1}|x_{n1}} \Theta_{w_{n2}|x_{n2}} \cdots \Theta_{w_{nN}|x_{nN}} P(X_{n1}, \lambda, \mu, T) \right] \right\}
\]

When the error probabilities are unknown, it is theoretically possible to estimate the whole set of parameters including the error matrix using maximum likelihood, but in practice the number of parameters to be estimated is too large unless the number of samples is extremely large. For example, even if we assume that the error matrix is the same across all families and all species, the number of parameters to be estimated is \( 2 \times M^2 \); that is, the entries of a full error matrix when \( M \) is the maximum possible size of a family. Instead, here we focus on cases where we have some information about the distribution of errors affecting measurement. In practice, this means that rather than estimating the joint distribution of \( \lambda \) and \( \Theta \), we estimate \( \lambda \) using external information about the distribution of error. If we assume a highly simplified error model, we can also estimate the error matrix using a pseudomaximum likelihood approach (Buonaccorsi 2010). Later, we present extensive simulation results that suggest that our method provides accurate estimates of all parameters.

**Results and Discussion**

The Effect of Error on Inferred Rates of Gene Family Evolution

To examine the effect of error in the gene family size taken from suboptimal genome annotations, we simulated gene families under a model with known error. These data were simulated using the phylogeny of 12 *Drosophila* species (supplementary fig. S1, Supplementary Material online) and the distribution of sizes among 11,434 gene families previously analyzed in these species (Hahn, Han, et al. 2007), with a true rate parameter of \( \lambda = \mu = 0.0012 \). A simulation consists of generating a data set using CAFE’s genfamily command and adding error to the data set as specified. In the simplest simulations, a known amount of error (\( \varepsilon \)) was added to each data set by randomly adding or subtracting genes from \( \varepsilon \) percent of the gene families, with all gene family sizes having the same error distribution (i.e., the same error matrix). Error can be added to all species or to a subset of species, effectively modeling heterogeneous assembly and annotation quality among genomes. The error distributions added to our simulated data were either \( \varepsilon = 0.1 \) or \( \varepsilon = 0.4 \) and consisted of an addition or subtraction of one to three genes in a family per species independently (fig. 1). An error value of \( \varepsilon = 0.1 \) means that in 90% of gene families, the observed size is equal to the true size, whereas in 10% of gene families, the observed size is either too large or too small (fig. 1A, C, and E). These error distributions approximate the range and distributions of error that we observe in several published genomes (see later).

To first assess the effect of error on inferred rates of gene family evolution, CAFE 3 was run on each simulated data set with standard settings—that is, with no error model incorporated. Estimating \( \lambda \) from these error-prone data sets gave values of 0.0027 and 0.0085 when adding \( \varepsilon = 0.1 \) and \( \varepsilon = 0.4 \), respectively (table 1). As expected, the more error contained in each data set, the higher value of \( \lambda \) we inferred; this is expected because higher rates of gene family evolution must be proposed to account for greater disparities in gene family size. Even when only 10% of families have an incorrect size (in each of the 12 genomes), the rate of gene family evolution is more than twice its true value (\( \lambda = 0.0012 \)). Although adding symmetric error does not change the mean family size across species, it does change the variance in size: from a variance equal to 0.519 in the true data (mean = 1.097), adding \( \varepsilon = 0.1 \) changed the variance to 0.609 and adding \( \varepsilon = 0.4 \) gave 0.894. Adding asymmetric error did change the mean family size but only very slightly (data not shown); the variances were the same as in the symmetric case.

There did not appear to be a clear effect of asymmetry in the error model on the absolute values of \( \lambda \), as putting more of the mass of the error distribution in either gains or losses did not significantly affect the estimated parameter value (compare results from error models 1A to 1C, and 1B to 1D in table 1). However, we did observe a small but substantial increase in \( \lambda \) when errors involving larger changes in family size (e.g., \( \pm 3 \)) were included (compare results from error models 1A to 1E, and 1B to 1F in table 1).

Accounting for Errors in Gene Family Size Using CAFE 3

We have observed how error in the observed gene family sizes can lead to an overestimation of the rates of gene gain and loss. We were therefore interested in whether the error model
described earlier could be used to account for this error to correctly infer the true value of $\lambda$. We apply the error model in two cases: first, when we assume that we know the correct error distribution and second, when we purposefully use an incorrect error model. Because the correct level of error and the exact distribution of error will often not be known, these two cases allow us to assess the consistency of our method. In both cases, the new “errormodel” function is used in CAFE 3, along with a specified error distribution.

In our ideal test case, we again simulated data with $\lambda = \mu = 0.0012$ across the phylogeny, and added a proportion of error, $\varepsilon$, equal to either 0.1 or 0.4 to all species. In both cases, the size of gene families with error was either incremented or decremented by a count of 1, with equal probability. In the case of using the same error distribution to correct for error as was added to the data set, CAFE correctly recovers the true $\lambda$ with high precision. The data sets with either $\varepsilon = 0.1$ or $\varepsilon = 0.4$ had the corresponding 0.1 and 0.4 error models applied to all species, and the $\lambda$ value inferred was again very close 0.0012 (table 1). To ensure that these results are not unique to a single rate parameter, we repeated the above simulations with $\lambda = \mu = 0.01$ and 0.0001 and the same error values. For $\varepsilon = 0.1$, application of the error model gave $\lambda = 0.00996$ and 0.00009, respectively. For $\varepsilon = 0.4$, application of the error model gave $\lambda = 0.01085$ and 0.000083, respectively. We can see that for $\lambda = 0.01$, we were able to infer the correct value

![Fig. 1. Error distributions used in simulations. These distributions include total errors of $\varepsilon = 0.1$ (A, C, E) and $\varepsilon = 0.4$ (B, D, F) but vary in the manner in which errors are spread across the error spectrum. In panels (A–D), only errors of $+1$ or $-1$ gene per family are considered, with (A) and (B) showing a symmetric spread of the total error between $+1$ and $-1$, whereas (C) and (D) show an asymmetric spread, skewed with 75% of the total error in the $-1$ category. The opposite skew, with 75% in the $+1$ category were also simulated, but is not shown here. Panels (E) and (F) show a symmetric distribution that extends to include the addition and subtraction of two or three genes.](http://mbe.oxfordjournals.org/doi/10.1093/molbev/mst100/fig1.png)
very accurately, whereas there was some inaccuracy for $\lambda = 0.0001$ (similar results held for both asymmetric and symmetric models). This latter result may be due to the small number of changes occurring across the tree with very low rates of change, but further simulations are needed to explore this effect.

We have also implemented the error model to allow for different error distributions among species. This corresponds to subsets of the input data coming from genomes of differential quality, as we have observed previously among the published Drosophila genomes (Drosophila 12 Genomes Consortium 2007; Hahn, Han, et al. 2007). To assess the accuracy of the $\lambda$ estimate when different amounts of error are applied to individual genomes, simulations were carried out as above with one difference: An error distribution of $\varepsilon = 0.4$ was applied to a data set simulated with $\varepsilon = 0.4$ (these distributions follow the same pattern as in figure 1C and D but with the asymmetry in the opposite direction). These distributions correspond to these two parts of the tree (supplementary fig. S1, Supplementary Material online).


d) These distributions follow the same pattern as in figure 1C and D but with the asymmetry in the opposite direction.

e) The simulated value without error in all cases is $\lambda = 0.0012$.

f) The incorrect error model for simulations with $\varepsilon = 0.4$ was applied to a data set simulated under $\varepsilon = 0.1$, the $\lambda$ value observed was 0.00045 (approximately one-third of the true value). Similarly, when using an error model with $\varepsilon = 0.1$ to correct for a data set simulated under $\varepsilon = 0.4$, the $\lambda$ value inferred was 0.0042. Although we have undercorrected in this case—the inferred value is more than three times higher than the simulated value—it is important to remember that the value that would have been obtained without the error model was twice as high again ($\lambda = 0.0086$). Overall, we conclude that our error model can recover the true value if the error distribution is known exactly, and if it is not known, the model performs according to expectations. The consistency of the error models suggests that we may be able to predict the error distribution when it is unknown; we demonstrate this feature of CAFE in the next section.

### Estimating the Error Distribution from External Data

Thus far, we have only considered the case where the error distribution for gene family size (i.e., the error matrix) is specified ahead of time. In this section and the next, we take two approaches to estimating the error distribution. In the first approach, we use external data from multiple genome assemblies, either two error-prone assemblies of the same genome or one high-quality assembly and one low-quality assembly.
Comparisons of these assemblies can be used to find MLEs of probabilities in the error matrix, as described in the Materials and Methods section. We demonstrate this approach using both simulated and real data sets. In the next section, we ask whether the likelihood reported by CAFE can be used to estimate the error distribution without additional external data.

To estimate the error distribution from external data, we first simulated gene family data as above. For each simulated data set without error, we generated two error-prone data sets by separately adding error according to the prespecified error distribution. By data set we mean a set of gene families having errors greater than ±3. When comparing estimated models from these data, we found better fits for an asymmetric model for the genomes of honey bee, guinea pig, and fugu, whereas the genomes of cow, zebrafish, fruit fly, human, and mouse were estimated to have a symmetric error model. Models with more parameters fit the data better in general (supplementary table S3, Supplementary Material online), but with more than nine parameters (up to differences of ±4 in the asymmetric model, ±7 in the symmetric model) the model often failed to find the global maximum likelihood. Even when the estimated error matrix for the best model had nonzero error rates for differences in family size greater than ±4, the estimated error rates for differences greater than ±4 were smaller than 0.01 in all the genomes analyzed (supplementary table S4, Supplementary Material online). On the basis of the errors observed from real genomes, we conclude that our error models are reasonable tradeoffs between realistic descriptions of the errors and ease of optimization.

### Estimating the Error Distribution without External Data

In our second approach to estimating error distributions, we did not use external data; instead we compare the likelihood of CAFE runs with varying error parameters. Because the likelihood of any particular run using the error model is calculated with fixed error parameters, we anticipated that comparison of likelihood scores among runs with varying error distributions would lead to accurate estimates of the error distributions themselves. Although the resulting estimates are not MLEs across the whole parameter space—and are therefore not guaranteed to be global maxima—as discussed in the Materials and Methods section, we can find the pseudo-MLE (PMLE) within the resolution of the grid we are searching.

To assess our ability to estimate the error distribution, we focused on estimating the value of ε, the proportion of gene families with error. We limit our search to estimate only a one-parameter error model (equal probability of ±1). To do this, we again simulated data sets with a λ value of 0.0012, adding either ε = 0.1 or 0.4. We then ran CAFE on both of these data sets with error models having a value of ε ranging from 0.0 to 0.9, in fixed increments. Figure 2 shows that the likelihood of the data is maximized (−ln L is minimized) when

<table>
<thead>
<tr>
<th>Error Added (ε)</th>
<th>l (No Error Model)</th>
<th>l (Correct Error Model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 (1A)</td>
<td>λ₁ = 0.00206⁵</td>
<td>λ₁ = 0.00097</td>
</tr>
<tr>
<td></td>
<td>λ₂ = 0.05784⁴</td>
<td>λ₂ = 0.02513</td>
</tr>
<tr>
<td>0.4 (1B)</td>
<td>λ₁ = 0.00616</td>
<td>λ₁ = 0.00091</td>
</tr>
<tr>
<td></td>
<td>λ₂ = 0.18173</td>
<td>λ₂ = 0.02690</td>
</tr>
</tbody>
</table>

⁵Error distributions correspond to the given panel in figure 1.
⁶The simulated values without error in all cases are λ₁ = 0.0009 and λ₂ = 0.025.
⁷Species with λ₂ are highlighted in supplementary figure S1, Supplementary Material online.
the model having the correct value of $\varepsilon$ is used: Models with $\varepsilon = 0.1$ and $\varepsilon = 0.4$ represent the best fit to the data when simulations with $\varepsilon = 0.1$ and $\varepsilon = 0.4$ are used, respectively. We have provided a program that automates the process of running CAFE with varying error models while maximizing the likelihood score by performing a simple optimization (caferor.py); this program will report the error model used to obtain the best score. These analyses show that we are able to use CAFE to estimate simple error models from the input data set by finding the value that maximizes the likelihood.

When the error distribution varies across the genomes considered, the simple one-dimensional search implemented above will only tell us about the average error across the tree. We therefore simulated a data set with varying levels of error among genomes and followed the one-dimensional search for an average error parameter with a species-by-species search. This species-by-species search is achieved by repeatedly adding or subtracting 10% of the average error predicted in the first step to each species separately, until the likelihood score has been maximized. Again, more complex search strategies that simultaneously try to maximize the average and species-specific error distributions can be considered in the future, but here we show that a simple approach works well on the simulated data set, as illustrated in table 3. The error in the data set resulted in an average error across all species estimated as $\varepsilon = 0.26$. We then continued to predict error on each individual species: D. melanogaster was simulated with $\varepsilon = 0.1$ and was predicted to have $\varepsilon = 0.101$. Drosophila simulans, D. sechellia, and D. persimilis had $\varepsilon = 0.4$ applied to their genomes and were all predicted to have error values of 0.406 (table 3). All other species had $\varepsilon = 0.2$ and were predicted to have error values from a range of 0.203 to 0.228 (table 3). It is notable that the largest deviation we saw was in D. pseudoobscura (simulated with $\varepsilon = 0.2$), as it is sister to D. persimilis (simulated with $\varepsilon = 0.4$); this suggests that genomes with high error rates can make it seem as though genomes of closely related species also have higher than expected error rates. Again, this pattern has been seen previously in the Drosophila data (Hahn, Han, et al. 2007). The species-by-species search has been implemented as an option that can be run after an average error parameter has been estimated using the caferor.py script.

### Applying the Error Model to Real Data Sets

Comparative data on the size of gene families has been analyzed in multiple different groups of organisms using earlier versions of CAFE (e.g., Sackton et al. 2007; Martin et al. 2008; Sharpton et al. 2009; Brown et al. 2010; Ohm et al. 2012; Qiu et al. 2012). Here, we use the new error model feature in CAFE 3 to revisit data from three clades that we have previously analyzed: fungi, mammals, and Drosophila. We analyzed new gene family data from 16 fungi (Butler et al. 2009; Rasmussen and Kellis 2011) and 10 mammals (Worley K et al., in preparation), as well as previously analyzed data from Drosophila (Hahn, Han, et al. 2007). For the Drosophila data set, we used the gene families and tree as described in Hahn, Han, et al. (2007); the gene families and tree from the expanded fungal data set are described in Rasmussen and Kellis (2011) and for the mammals is described in Worley K et al. (in preparation). Each of these groups of species is certain to have heterogeneous levels of genome assembly and annotation quality among them. Each group contains one or two focal species—Saccharomyces cerevisiae among the fungi, Homo sapiens and Mus musculus among the mammals, and D. melanogaster among the flies—that has a very high-quality assembly and annotation and likely several species with below-average genomes (e.g., the Drosophila species mentioned in the previous section). Because of the presence of genomes with lower quality, it seems likely that the error model introduced here will provide a more accurate estimate of the rate of gene family evolution in each group.

Analyzing the data without applying an error model, we found $\lambda = 0.0008$, 0.0023, and 0.0012, for the fungi, mammals, and Drosophila, respectively (table 4). For each data set, we then estimated a one-parameter error model without using

---

**FIG. 2.** Estimating the error distribution. Each panel shows the $-\ln$ likelihood of individual CAFE runs on a simulated data set with error. The points represent the score of the run with the corresponding error model ($\varepsilon$ value) on the x axis; the dashed vertical line indicates the true amount of error added to the simulated data. Panel (A) shows a simulation with $\varepsilon = 0.1$ and panel (B) shows one with $\varepsilon = 0.4$. In both cases, the maximum likelihood score occurs when the correct error model is used.
In the fungi and mammals, the new estimates suggest rates of evolution approximately 75% of the uncorrected estimates, whereas in Drosophila, the new estimate is 50% of the original one.

In addition to fitting a global rate parameter, we examined the fit of a model having three \( \lambda \) parameters on the mammalian tree. Previous research found higher rates of gene gain and loss in the great apes, with an intermediate rate in other primates, and the lowest rate on the rest of the tree (Hahn, Demuth, et al. 2007). We wished to know whether this pattern would still be observed after errors were accounted for, so we estimated the likelihood of the three-parameter model while setting \( \varepsilon = 0.0732 \). As expected if this pattern is due to a true rate acceleration and not error in assemblies or annotation, the three-parameter model with error fit the data significantly better than a one-parameter model with error (\( P < 0.0001 \); likelihood ratio test). Indeed, as previously observed, the value of the \( \lambda \) parameter on the human–chimp shared lineage was more than three times higher than the average value (0.0062 vs. 0.0019), with the orangutan–macaque shared lineage having an intermediate rate (\( \lambda = 0.0044 \)). These results are not wholly unexpected, as other previous studies have found the same rate acceleration in a set of analyses that is free from biases due to heterogeneous quality in genome assemblies (Marques-Bonet et al. 2009). Our results also confirm earlier conclusions about the amount of genomic copy number divergence between humans and chimpanzees (Demuth et al. 2006) and demonstrate that these results were not due to genome assembly or annotation error.

### Table 3: Estimating the Error Distribution (\( \varepsilon \)) with Heterogeneous Error across Species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Error Added</th>
<th>Estimated Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila willistoni</td>
<td>0.2</td>
<td>0.20283</td>
</tr>
<tr>
<td>D. virilis</td>
<td>0.2</td>
<td>0.20283</td>
</tr>
<tr>
<td>D. persimilis</td>
<td>0.4</td>
<td>0.40566</td>
</tr>
<tr>
<td>D. mojavensis</td>
<td>0.2</td>
<td>0.20283</td>
</tr>
<tr>
<td>D. sechellia</td>
<td>0.4</td>
<td>0.40566</td>
</tr>
<tr>
<td>D. pseudoobscura</td>
<td>0.2</td>
<td>0.22819</td>
</tr>
<tr>
<td>D. yakuba</td>
<td>0.2</td>
<td>0.20283</td>
</tr>
<tr>
<td>D. grimshawi</td>
<td>0.2</td>
<td>0.20283</td>
</tr>
<tr>
<td>D. erecta</td>
<td>0.2</td>
<td>0.20283</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>0.1</td>
<td>0.10141</td>
</tr>
<tr>
<td>D. ananassae</td>
<td>0.2</td>
<td>0.20283</td>
</tr>
<tr>
<td>D. simulans</td>
<td>0.4</td>
<td>0.40566</td>
</tr>
</tbody>
</table>

*The estimated error distribution was symmetric with only \( \pm 1 \) allowed.

### Table 4: Analysis of Previously Published Data Sets.

<table>
<thead>
<tr>
<th>Clade</th>
<th>( \lambda ) (No Error Model)</th>
<th>( \varepsilon ) (Estimated)</th>
<th>( \lambda ) (Error Model = ( \varepsilon ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fungi</td>
<td>0.00080</td>
<td>0.02771</td>
<td>0.00061</td>
</tr>
<tr>
<td>Mammals</td>
<td>0.00238</td>
<td>0.07324</td>
<td>0.00186</td>
</tr>
<tr>
<td>Drosophila</td>
<td>0.00121</td>
<td>0.04102</td>
<td>0.00059</td>
</tr>
</tbody>
</table>

Conclusions

Here, we have provided a new software package that enables the accurate estimation of rates of gene family evolution when there are errors in the observed gene family sizes. By allowing users to marginalize over the uncertainty in the observed gene family sizes, CAFE 3 provides a platform for expanding comparative genomic analyses into clades consisting solely of draft genome sequences. Our software is freely available (http://sourceforge.net/projects/cafehahnlab/) and can be compiled on multiple operating systems. Although it is likely that there are typical error distributions associated with different sequencing technologies used to assemble genomes (e.g., Illumina vs. 454), our program does not require that such distributions are known ahead of time. If prior information about either the sequencing technology or the depth of coverage is known, more accurate results may be obtained. We have demonstrated three alternative approaches to estimating error distributions, each of which requires slightly different types of data; regardless of how error distributions are estimated, CAFE 3 allows any arbitrary distribution to be specified. Finally, although we have applied this approach to correcting for error in gene family sizes, similar methods may be applicable to errors in nucleotide data (e.g., Heid et al. 2008; Hubisz et al. 2011) or any trait for which a realistic error model can be constructed (e.g., RNA-seq; Brawand et al. 2011).
Materials and Methods

Application of the Error Model to Infer the Gain and Loss Parameters

The general approach we take uses PMLEs because we do not include the error parameters in the full likelihood formula. Instead, we estimate the measurement error parameters from external data as described later, and then the likelihood is calculated using the observed data, with the error parameters fixed at their estimates (Buonaccorsi 2010):

\[
\lambda, \mu = \arg\max_{\lambda, \mu} \sum_{n=1}^{N} P(W_n \mid X_n, \lambda, \mu, T) \]

The numbers of parameters for the error matrix are then \(2D + 1\) for asymmetric models and \(D + 1\) for symmetric models.

Estimation of the Error Matrix from External Data

If we know that one measurement (i.e., one set of gene families from a well-annotated genome) is more accurate than another, we can estimate the error matrix by assuming the better measure as the true value and comparing the lesser measurement to the truth. Although having a well-annotated genome might seem to obviate the need for using the error model, the estimated error matrix from such a comparison could be usefully applied to poorly annotated genomes. If we have two sets of measurements with unknown the relative accuracy, we can estimate the error matrix based on the observed agreement between the two measures. We describe these two cases in reverse order.

First, when the two measures are similarly error prone, we find the triangular matrix \(R = r_{ij}\) (where \(i = 0 \ldots M, j \geq i\)) of pairwise observations, with \(r_{ij}\) being the number of observations with \(W_j = i\) and \(W_k = j\). The probability of each pairwise observation is defined based on the true count probabilities \(\theta_x\) and the error rates \(\theta_{w \mid x}\). Assuming that the probability of observing pairwise observations is \(p_{ij}\) with

\[
p_{ij} = \begin{cases} 
2 \cdot \theta_{i \mid k} \cdot \theta_{j \mid k} \cdot \theta_k & \text{if } i \neq j \\
\sum_{k=0}^{M} \theta_{i \mid k} \cdot \theta_{j \mid k} \cdot \theta_k & \text{if } i = j 
\end{cases}
\]

then the log likelihood of the data matrix \(R\) given the probability distribution of \(\theta_x\) and the error matrix model \(\Theta\) can be calculated using the multinomial likelihood. Ignoring the coefficient, the log likelihood \(L\) is:

\[
\ln L(R \mid \Theta, \theta) = \sum_{i=0}^{M} \sum_{j=0}^{M} r_{ij} \ln (p_{ij}).
\]

However, for our data set, we have a limitation that we can never observe gene families that are size zero in both measures. To account for this missing data, we find the likelihood that is conditional on the event (\(E\)) that we observed at least one gene in either one of the measurements. Because

\[
P(R \mid \Theta, \theta, E) = \frac{P(R, E \mid \Theta, \theta)}{P(E)} = \frac{P(R, E \mid \Theta, \theta)}{1 - p(0, 0)},
\]

the conditional log likelihood \(L_c\) is:

\[
\ln L_c(R \mid \Theta, \theta, E) = \sum_{i=0}^{M} \sum_{j=0}^{M} r_{ij} \ln (p_{ij}) - \ln (1 - p_{00}).
\]

We do not have data for the true \(\theta\), but assuming the distribution is similar to the counts found in \(W_1\) and \(W_2\), we can substitute the count distributions observed in \(W_1 + W_2\) (\(\hat{\theta}\)) as the approximation of \(\theta\). For each model of error matrix \(\Theta\), we find the parameters that maximize the
conditional log likelihood of the data \( R \) using the Nelder–Mead method.

For the case when one set of annotations is treated as the true measure, the observation matrix \( R \) is a full matrix with columns that correspond to the true measure and rows to the incorrect measure. \( r_{ij} \) is now the number of observations with \( W = i \) and \( X = j \). The probability of each pairwise observation is again defined based on the true count probabilities \( \theta_i \) and the error rates \( \theta_{ij-x} \), but the probability of observing pairwise observations \( p_{ij} \) is simpler:

\[
p_{ij} = \theta_{ij} \cdot \theta_j.
\]

The conditional log likelihood follows the same approach as above but is summed across the whole discordance matrix, \( R \):

\[
\ln L(R | \Theta, \theta, E) = \sum_{i=0}^{M} \sum_{j=0}^{M} r_{ij} \ln (p_{ij}) - \ln (1 - p_{00}).
\]

The distribution \( \theta \) can be estimated using the assumed true count data \( X \). MLEs of parameters that determine the error matrix \( \Theta \) are then found by comparison of the true and error-prone data.

**Estimation of the Error Matrix from Eight Genomes**

To collect data on realistic error matrices, we compared annotations for two versions of each of eight published genomes. We used the gene models for honey bee (Apis mellifera), cow (Bos taurus), guinea pig (Cavia porcellus), zebrafish (Danio rerio), fruit fly (Drosophila melanogaster), human (H. sapiens), mouse (M. musculus), and fugu (Takifugu rubripes) from Ensembl (Flicek et al. 2012). For each species, we compared an earlier, lower quality assembly and annotation to a later, higher quality assembly and annotation (supplementary table S2, Supplementary Material online). For each pair of genome annotations (i.e., for each species), we assigned genes to gene families using an all-against-all BLASTP sequence similarity search, followed by clustering using the MCL algorithm (Enright et al. 2002). Because genes from both annotation versions were clustered simultaneously, we can simply compare the size of each resulting family to estimate the error matrix. We applied our method of estimating the error matrix from external data to the pairs of genomes, assuming that the updated annotation is the true measure. We compared symmetric and asymmetric error models with a range of parameters (supplementary table S3, Supplementary Material online).

The number of parameters in the asymmetric models ranged from three (difference of \(-1, 0, 1\)) to up to nine (difference of \(-4, -3, \ldots 3, 4\)). The number of parameters on the symmetric models ranged from two (difference of 0, \pm 1) to eight (difference of 0, \pm 1, \ldots \pm 7).

**Estimation of the Error Matrix via Search without External Data**

We also demonstrate how to estimate the error matrix when there is no external validation data available. In this case, we find the birth-and-death parameters that maximize the pseudolikelihood using a fixed error model but repeat the procedure across a grid of error-parameter values. The grid consists of error parameters in fixed increments across a fixed region, and the error parameter that yields the maximum pseudolikelihood across the grid space is determined as the estimate. This process has the limitation that it cannot search the whole continuous parameter space, and an \( n \)-dimensional grid search becomes impractical for complicated error matrices as the number of parameters in the model increases. Nevertheless, it performs fairly well in practice, as is shown in the simulations presented in the Results and Discussion sections.

**Supplementary Material**

Supplementary figures S1–S3 and tables S1–S4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors thank Matt Rasmussen for providing the fungal data and David Swofford for helpful discussions. This work was supported by a National Evolutionary Synthesis Center postdoctoral fellowship to M.V.H., by a Ford Foundation postdoctoral fellowship to J.L.-M., and by National Science Foundation grant DBI-0845494 to M.W.H.

**References**


