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]

[1]National Evolutionary Synthesis Center, Durham, North Carolina

[2]School of Informatics and Computing, Indiana University

[3]Department of Biology, Indiana University

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

Associate editor: Sudhir Kumar

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

Article

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:

$$\theta_{w|x} = P(W = w \mid X = x),$$

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 θ_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 θ denotes the discrete probability distribution θ_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 λ and μ 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):

$$P(X(t) = c \mid X(0) = s) = \sum_{j=0}^{\min(s,c)} \binom{s}{j} \binom{s+c-j-1}{s-1} \alpha^{s-j} \beta^{c-j} (1-\alpha-\beta)^j$$

$$\alpha = \frac{\mu(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}, \quad \beta = \frac{\lambda(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}$$

where λ is the rate of gene gain and μ 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 $X_n = (X_{n1}, X_{n2}, \ldots X_{ns})$, in which each item X_{ni} describes the size of the nth gene family in each genome of species i ($i \in S$). Similarly, $Z_n = (Z_{n1}, Z_{n2}, \ldots Z_{nu})$ is the vector in which each item Z_{nj} 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:

$$\begin{aligned} \lambda, \mu &= \text{argmax}_{\lambda,\mu} \left(\prod_{n=1}^{N} P(X_n \mid \lambda, \mu, T) \right) \\ &= \text{argmax}_{\lambda,\mu} \left(\prod_{n=1}^{N} P(X_n \mid Z_n, \lambda, \mu, T) P(Z_n \mid \lambda, \mu, T) \right) \\ &= \text{argmax}_{\lambda,\mu} \left(\prod_{n=1}^{N} \left\{ \begin{array}{l} \sum_{z_{n1}=0}^{M} \sum_{z_{n2}=0}^{M} \cdots \sum_{z_{nu}=0}^{M} P(X_n \mid Z_n \\ = (z_{n1}, z_{n2} \ldots z_{nu}), \lambda, \mu, T) \\ \times P(Z_n = (z_{n1}, z_{n2} \ldots z_{nu}) \mid \lambda, \mu, T) \end{array} \right\} \right). \end{aligned}$$

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

$$(\lambda, \mu)_W = \text{argmax}_{\lambda,\mu} \left(\prod_{n=1}^{N} P(W_n \mid \lambda, \mu, T) \right),$$

where we define the vector $W_n = (W_{n1}, W_{n2}, \ldots W_{ns})$. Similar to X_n, W_{ni} 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:

$$\lambda,\mu = \text{argmax}_{\lambda,\mu}\left(\prod_{n=1}^{N}\left\{\begin{array}{l}\sum_{x_{n1}=0}^{M}\sum_{x_{n2}=0}^{M}\cdots\sum_{x_{ns}=0}^{M}P(W_n\,|\,X_n\\ \quad=(x_{n1},x_{n2},\ldots x_{ns}))\\ \quad\times P(X_n,\lambda,\mu,T)\end{array}\right\}\right)$$

$$=\text{argmax}_{\lambda,\mu}\left(\prod_{n=1}^{N}\left\{\begin{array}{l}\sum_{x_{n1}=0}^{M}\sum_{x_{n2}=0}^{M}\cdots\sum_{x_{ns}=0}^{M}P(W_{n1}\,|\,X_{n1}\\ \quad=x_{n1})P(W_{n2}\,|\,X_{n2}=x_{n2})\\ \quad\times\ldots P(W_{ns}\,|\,X_{ns}=x_{ns})\\ \quad\times P(X_n,\lambda,\mu,T)\end{array}\right\}\right).$$

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

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

$$(\lambda,\mu,\Theta)$$

$$=\text{argmax}_{\lambda,\mu,\Theta}\left(\prod_{n=1}^{N}\left\{\begin{array}{l}\sum_{X}^{M}\Theta_{W_{n1}\,|\,X_{n1}}\Theta_{W_{n2}\,|\,X_{n2}}\\ \quad\times\ldots\Theta_{W_{ns}\,|\,X_{ns}}P(X_n,\lambda,\mu,T)\end{array}\right\}\right)$$

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 $2 + M^2$; 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 λ and Θ, we estimate λ 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 (ε) was added to each data set by randomly adding or subtracting genes from ε 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. 1*A*, *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 λ 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 λ 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 λ, 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 λ when errors involving larger changes in family size (e.g., ±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

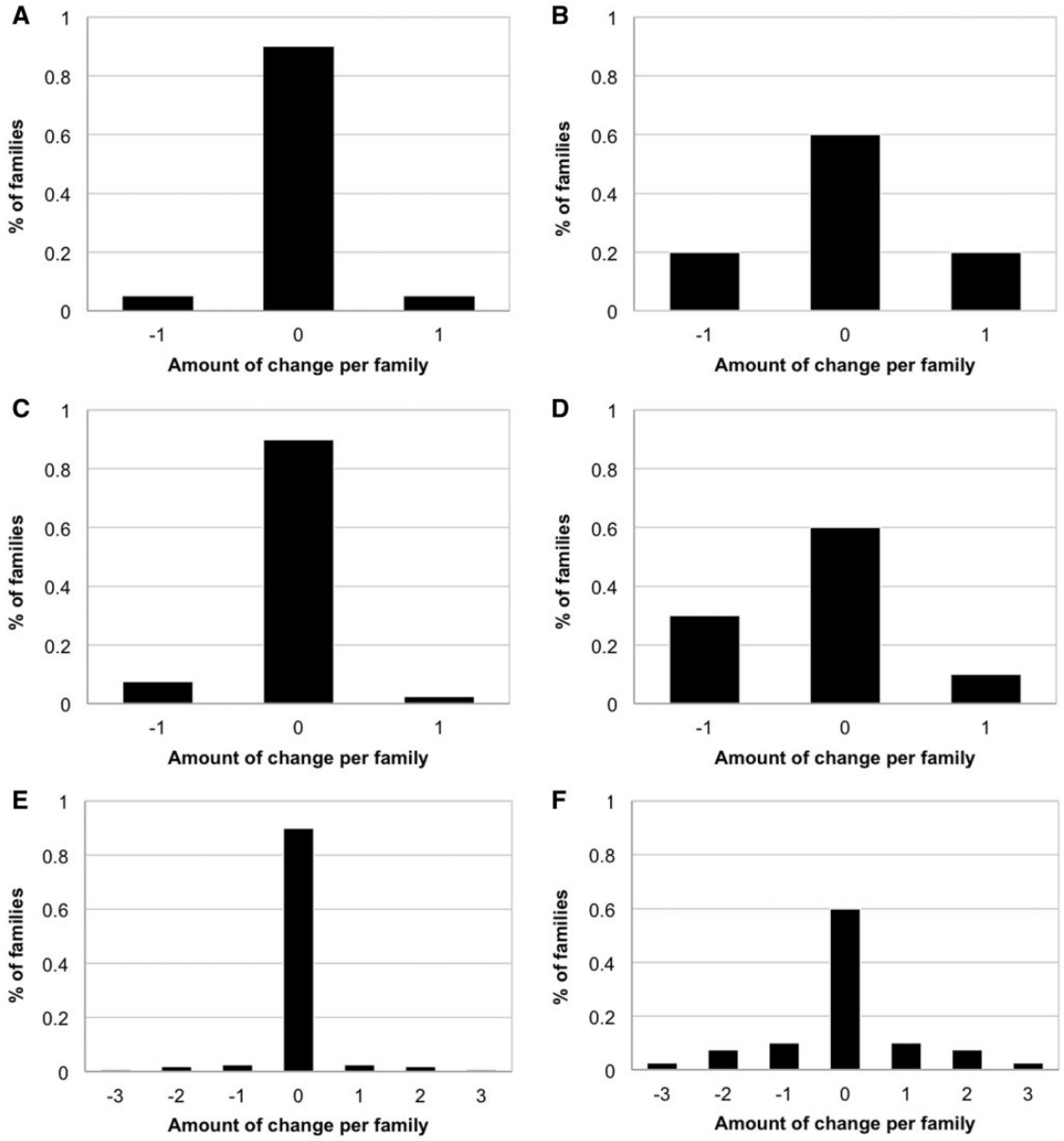

FIG. 1. Error distributions used in simulations. These distributions include total errors of $\varepsilon = 0.1$ (*A*, *C*, *E*) and $\varepsilon = 0.4$ (*B*, *D*, *F*) but vary in the manner in which errors are spread across the error spectrum. In panels (*A*–*D*), only errors of $+1$ or -1 gene per 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.

described earlier could be used to account for this error to correctly infer the true value of λ. 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, ε, 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 λ 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 λ 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

Table 1. Performance of the Error Model.

	Error Added (ε)	λ (No Error Model)	λ (Correct Error Model)	λ (Incorrect Error Model)[a]
Symmetric error distributions	0.1 (1A)[b]	0.00280	0.00122[c]	0.00043
	0.4 (1B)	0.00897	0.00124	0.00447
Asymmetric error distributions	0.1 (1C)	0.00283	0.00124	0.00047
	0.1 (1C)[d]	0.00276	0.00120	0.00050
	0.4 (1D)	0.00960	0.00139	0.00502
	0.4 (1D)[d]	0.00765	0.00111	0.00366
Varying error across species	0.4 for low-quality species[e] (1B) 0.1 for all other branches (1A)	0.00417	0.00121	0.00090
Symmetric error with more error classes	0.1 (1E)	0.00324	0.00130	0.00050
	0.4 (1F)	0.00903	0.00154	0.00406

[a]The incorrect error model for simulations with $\varepsilon = 0.1$ is $\varepsilon = 0.4$ and vice versa.
[b]Error distributions correspond to the given panel in figure 1.
[c]The simulated value without error in all cases is $\lambda = 0.0012$.
[d]These distributions follow the same pattern as in figure 1*C* and *D* but with the asymmetry in the opposite direction.
[e]Low-quality species are highlighted in supplementary figure S1, Supplementary Material online.

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 λ estimate when different amounts of error are applied to individual genomes, simulations were carried out as above with one difference: An error distribution of $\varepsilon = 0.1$ was applied to all species except *D. simulans*, *D. sechellia*, and *D. persimilis*, which were all given $\varepsilon = 0.4$ (these species were observed to have lower quality genome assemblies; Begun et al. 2007; *Drosophila* 12 Genomes Consortium 2007). When we specify these same error distributions in our error model—correctly assigning each distribution to each species—we again recover the true single value of λ across the tree (third row in table 1). In addition to searching for a single λ value across the tree, CAFE has the ability to search for separate λ values on individual branches or clades. Without accounting for error, models with separate λ parameters for terminal branches leading to each of the three species with higher error rates fit significantly better (data not shown); this is because the error-prone assemblies appear to have faster rates of gene family evolution (cf., Hahn, Han, et al. 2007). Conversely, if the data are simulated with two λ parameters corresponding to these two parts of the tree (supplementary fig. S1, Supplementary Material online), we are able to correctly infer these parameter values even in the presence of error (table 2). We note that there is a potential identifiability problem when we have separate parameters for gain and loss on a specific branch along with a separate error model for the same branch. We may not always be able to distinguish higher gain and loss rates on the branch with the higher error rates.

We also tested the effect of using an incorrect error model when inferring rates of change. We anticipated that using an error distribution larger than the error distribution that is present in the data—that is, a larger value of ε—would lead to an overcorrection and that a lower λ would be observed with respect to the value of λ with which the data were simulated. Similarly, using an error distribution that corrects for less error than is present in the data might lead to an undercorrection and a higher λ than the true value. Our simulations confirmed our expectations: When an error model with $\varepsilon = 0.4$ was applied to a data set simulated with $\varepsilon = 0.1$, the λ 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 λ 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.

Table 2. Performance of the Error Model on a 2-λ Search.

Error Added (ε)	λ (No Error Model)	λ (Correct Error Model)
0.1 (1A)[a]	$\lambda_1 = 0.00206$[b] $\lambda_2 = 0.05784$[c]	$\lambda_1 = 0.00097$ $\lambda_2 = 0.02513$
0.4 (1B)	$\lambda_1 = 0.00616$ $\lambda_2 = 0.18173$	$\lambda_1 = 0.00091$ $\lambda_2 = 0.02690$

[a]Error distributions correspond to the given panel in figure 1.
[b]The simulated values without error in all cases are $\lambda_1 = 0.0009$ and $\lambda_2 = 0.025$.
[c]Species with λ_2 are highlighted in supplementary figure S1, Supplementary Material online.

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 that comprise a whole genome. We can either compare these two error-prone data sets with each other to estimate the error matrix or we can compare one to the data set without error. The former comparison is equivalent to having two equally error-prone annotations of the same genome, whereas the latter comparison is equivalent to having one accurate annotation and one error-prone assembly. In both cases, we can find the parameters in the error matrix, Θ, that maximize the log likelihood of the data (see Materials and Methods). We simulated data sets with varying numbers of error parameters, from two parameters (error probabilities for differences of 0 and ±1) to seven parameters (error probabilities for differences of $+3$, $+2$, $+1$, 0, -1, -2, and -3) and varying amounts of error (with ε ranging from 0.1 to 0.6). The results of these simulations suggest that we are able to accurately estimate up to four error parameters, across all values of ε (supplementary table S1, Supplementary Material online). With more than four parameters, we did not get convergence of the log likelihood scores. In addition, when the simulated error model is simpler (i.e., has fewer parameters), the more complex error models are not better fitting, suggesting that we are able to find both estimates of the error parameters and the model complexity (supplementary fig. S2 and table S1, Supplementary Material online). Finally, we found that estimating the error distribution by comparing two error-prone data sets was only very slightly less accurate than estimation via comparison between one error-prone data set and one high-quality data set (supplementary table S1, Supplementary Material online). This result is encouraging for systems in which no reference-quality genome is available, although we are making the assumption that the two available error-prone data sets share the same error distribution.

To evaluate the error distributions found in real genomes, we compared low- and high-quality assemblies and annotations for eight species (supplementary table S2, Supplementary Material online). These comparisons range from genomes of highly disparate quality (1.92X coverage vs. 6.79X coverage for the two versions of the guinea pig genome) to genomes in the final stages of finishing (e.g., *D. melanogaster* assembly version 4 vs. version 5.4). As a result, we find a wide range of error distributions, with ε going from 0.687 (in *Apis*) to 0.104 (in *Drosophila*). The distribution of differences in gene family sizes in these pairs of genomes is shown graphically in supplementary figure S3, Supplementary Material online. As can be seen, these distributions match our simulated distributions quite well, with most errors involving $+1$ or -1 differences and very few 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 genome data, 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 $\varepsilon = 0.1$ 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

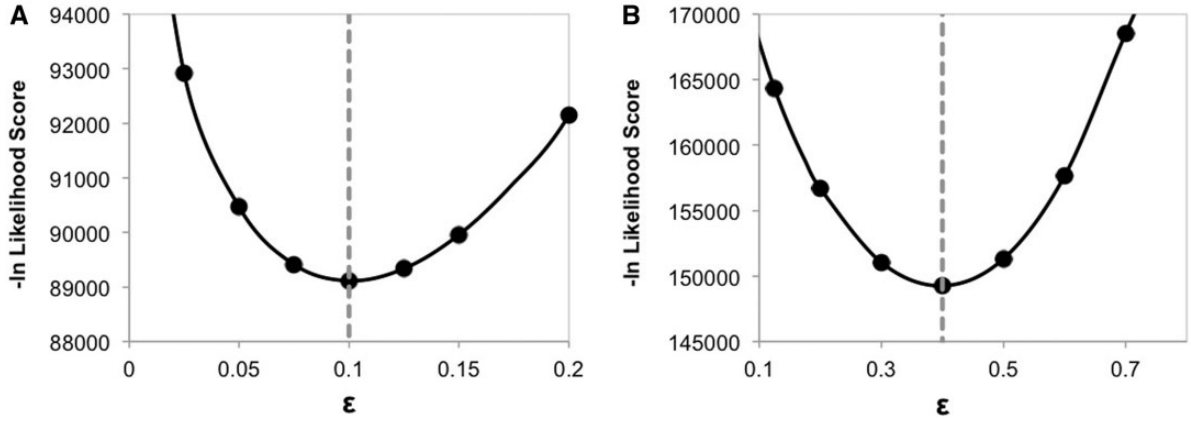

Fig. 2. Estimating the error distribution. Each panel shows the –ln likelihood of individual CAFE runs on a simulated data set with error. The points represent the score of the run with the corresponding error model (ε value) on the *x* axis; the dashed vertical line indicates the true amount of error 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.

the model having the correct value of ε 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 (caferror.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 caferror.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

Table 3. Estimating the Error Distribution (ε) with Heterogeneous Error across Species.

Species	Error Added	Estimated Error
Drosophila willistoni	0.2	0.20283
D. virilis	0.2	0.20283
D. persimilis	0.4	0.40566
D. mojavensis	0.2	0.20283
D. sechellia	0.4	0.40566
D. pseudoobscura	0.2	0.22819
D. yakuba	0.2	0.20283
D. grimshawi	0.2	0.20283
D. erecta	0.2	0.20283
D. melanogaster	0.1	0.10141
D. ananassae	0.2	0.20283
D. simulans	0.4	0.40566

Table 4. Analysis of Previously Published Data Sets.

Clade	λ (No Error Model)	ε (Estimated)[a]	λ (Error Model = ε)
Fungi	0.00080	0.02771	0.00061
Mammals	0.00238	0.07324	0.00186
Drosophila	0.00121	0.04102	0.00059

[a]The estimated error distribution was symmetric with only ±1 allowed.

external data, limiting ε to symmetric errors of ±1. Although this is a highly simplified error distribution, as shown earlier it captures most of the effect of error on rates of gene family evolution. Assuming an average error distribution shared across all genomes in each analysis, we searched for the PMLE of the error parameter, ε, finding $\varepsilon = 0.0277$, 0.0732, and 0.041 for the fungi, mammals, and *Drosophila*, respectively (table 4). These values imply that on average 2.8%, 7.3%, and 4.1% of gene families in each data set have observed sizes that are in error—that is, the observed size of the families are not equal to the true size. It is interesting to note that the average error in each clade roughly coincides with the number of repetitive elements, number of total gene duplicates, and total size of the analyzed genomes (mammals > *Drosophila* > fungi), all of which are expected to coincide with errors in assembly and annotation. These conclusions of course rest on the assumption that our error model is an accurate representation of the true error distributions. As we have shown earlier, an analysis of error-prone assemblies supports our modeling assumptions. Nevertheless, this is an association among three clades, arising from genome assemblies constructed in very different ways, and should be considered to be a preliminary indication of factors that can affect assembly and annotation quality.

Analyzing the data with the best-fit error models, we found new rate parameters of $\lambda = 0.0006$, 0.0019, and 0.0006, for the fungi, mammals, and *Drosophila*, respectively (table 4). In all three data sets, models with a single error-parameter fit the data significantly better than models without error parameters (all $P \ll 0.0001$; likelihood ratio test), indicating that the corrected λ values reported here are more accurate reflections of the rate of gene family evolution in these three clades. 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 λ 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 ($P \ll 0.0001$; likelihood ratio test). Indeed, as previously observed, the value of the λ 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 genic 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.

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):

$$\lambda, \mu = \text{argmax}_{\lambda,\mu}\left(\prod_{n=1}^{N} P(W_n \mid X_n, \lambda, \mu, T)\right)$$

$$P(W_n \mid X_n, \lambda, \mu, T)$$
$$= \sum_{x_{n1}=0}^{M}\sum_{x_{n2}=0}^{M}\cdots\sum_{x_{ns}=0}^{M} P(W_n \mid X_n = (x_{n1}, x_{n2}, \ldots x_{ns}))P(X_n, \lambda, \mu, T)$$
$$= \sum_{x_{n1}=0}^{M}\sum_{x_{n2}=0}^{M}\cdots\sum_{x_{ns}=0}^{M}\sum_{z_{n1}=0}^{M}\sum_{z_{n2}=0}^{M}\cdots$$
$$\times \sum_{z_{nu}=0}^{M}\left\{\begin{array}{l} P(W_n \mid X_n = (x_{n1}, x_{n2}, \ldots x_{ns}))P(X_n \mid Z_n \\ = (z_{n1}, z_{n2}\ldots z_{nu}), \lambda, \mu, T)P(Z_n \mid \lambda, \mu, T)\end{array}\right\},$$

where $P(W = w \mid X = x) = \hat{\theta}_{w \mid x}$ and is estimated through external data.

Simplifying the Error Matrix

Because we are dealing with count data that can go up to M, if we allow for error from any true count to any observed count the number of parameters specifying the error matrix for this full model is M^2, which is unnecessarily complex. To simplify the parameters and to make the model useful, we can constrain three aspects of the model.

First, we can assume that the error rate depends on the difference between the observed count and the true count but does not depend on the true count itself. This is equivalent to having a single homogeneous parameter along the diagonals of the error matrix. Although this assumption may not be biologically realistic, because most of our gene families are sizes of three or smaller and large families are rare, only three rates would normally have to be used. Our modeling framework also allows this assumption to be relaxed when error is estimated from external data, and any error structure can be entered into CAFE 3 if specified by the user. Second, we can restrict the errors that are allowed to be at most D differences from the true count. This forces the corner parameters that are $D + 1$ or more steps away from the diagonal to a probability ω that is constrained to be smaller than all other parameters. Again, D is a user-specified parameter that can be quite large in practice. Third, we can assume symmetry on the rates of errors that increase the counts and the rates of errors that decrease the counts, reducing the number of parameters to half. This last assumption is again optional, and we have explored models with and without it. In our simulated results presented above, we explore a range of models that are combinations of the value of D ($\leq$3) and the state of symmetry. 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 measure 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 $W_1 = i$ and $W_2 = j$. The probability of each pairwise observation is defined based on the true count probabilities θ_x and the error rates $\theta_{w \mid x}$. Assuming that the probability of observing pairwise observations is p_{ij}, with

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

then the log likelihood of the data matrix R given the probability distribution of θ_x and the error matrix model Θ can be calculated using the multinomial likelihood. Ignoring the coefficient, the log likelihood ln L is:

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

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 ln L_c is:

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

We do not have data for the true θ, 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 θ. For each model of error matrix Θ, 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 θ_x and the error rates $\theta_{w|x}$, but the probability of observing pairwise observations p_{ij} is simpler:

$$p_{ij} = \theta_{i|j} \cdot \theta_j.$$

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

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

The distribution θ can be estimated using the assumed true count data X. MLEs of parameters that determine the error matrix Θ 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, ± 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 predoctoral fellowship to J.L.-M., and by National Science Foundation grant DBI-0845494 to M.W.H.

References

Ames RM, Money D, Ghatge VP, Whelan S, Lovell SC. 2012. Determining the evolutionary history of gene families. *Bioinformatics* 28:48–55.

Bailey NTJ. 1964. The elements of stochastic processes. New York: John Wiley & Sons, Inc.

Begun DJ, Holloway AK, Stephens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* 5:e310.

Brawand D, Soumillon M, Necsulea A, et al. (18 co-authors). 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478:343–348.

Brown CA, Murray AW, Verstrepen KJ. 2010. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr Biol.* 20:895–903.

Buonaccorsi JP. 2010. Measurement error: models, methods and applications. Boca Raton (FL): Chapman and Hall/CRC Press.

Butler G, Rasmussen MD, Lin MF, et al. (51 co-authors). 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459:657–662.

Colbourne JK, Pfrender ME, Gilbert D, et al. (68 co-authors). 2011. The ecoresponsive genome of *Daphnia pulex*. *Science* 331:555–561.

Costello JC, Han MV, Hahn MW. 2008. Limitations of pseudogenes in identifying gene losses. In: Nelson C, Vialette S, editors. Proceedings of the Sixth Annual RECOMB Satellite Workshop on Comparative Genomics; 2008 Oct 13–15; Paris, France. Heidelberg (Germany): Springer Berlin. p. 14–25.

De Bie T, Demuth JP, Cristianini N, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22:1269–1271.

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. 2006. The evolution of mammalian gene families. *PLoS One* 1:e85.

Demuth JP, Hahn MW. 2009. The life and death of gene families. *BioEssays* 31:29–39.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–218.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* 320:1629–1631.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30: 1575–1584.
Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst Biol.* 22:240–249.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17:368–376.
Flicek P, Amode MR, Barrell D, et al. (57 co-authors). 2012. Ensembl 2012. *Nucleic Acids Res.* 40:D84–D90.
Floudas D, Binder M, Riley R, et al. (71 co-authors). 2012. The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* 336:1715–1719.
Gibbs RA, Rogers J, Katze M, et al. (176 co-authors). 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.
Gibbs RA, Weinstock GM, Metzker ML, et al. (241 co-authors). 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521.
Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* 15:1153–1160.
Hahn MW, Demuth JP, Han S-G. 2007. Accelerated rate of gene gain and loss in primates. *Genetics* 177:1941–1949.
Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet.* 3:e197.
Heid IM, Lamina C, Küchenhoff H, et al. (18 co-authors). 2008. Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *Am J Epidemiol.* 168:878–889.
Holt RA, Subramanian GM, Halpern A, et al. (123 co-authors). 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298:129–149.
Hubisz MJ, Lin MF, Kellis M, Siepel A. 2011. Error and error mitigation in low-coverage genome assemblies. *PLoS One* 6:e17034.
Kidd JM, Cooper GM, Donahue WF, et al. (46 co-authors). 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* 453:56–64.
Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics* 28:279–281.
Li R, Fan W, Tian G, et al. (123 co-authors). 2009. The sequence and de novo assembly of the giant panda genome. *Nature* 463:311–317.
Liu L, Yu L, Kalavacharla V, Liu Z. 2011. A Bayesian model for gene family evolution. *BMC Bioinformatics* 12:426.
Marques-Bonet T, Kidd JM, Ventura M, et al. (20 co-authors). 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* 457:877–881.
Martin F, Aerts A, Ahren D, et al. (68 co-authors). 2008. The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452:88–92.
Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet.* 39:121–152.
Ohm RA, Feau N, Henrissat B, et al. (28 co-authors). 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen *Dothideomycetes* fungi. *PLoS Pathog.* 8: e1003037.
Qiu Q, Zhang GJ, Ma T, et al. (47 co-authors). 2012. The yak genome and adaptation to life at high altitude. *Nat Genet.* 44:946–949.
Rasmussen MD, Kellis M. 2011. A Bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol.* 28:273–290.
Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in *Drosophila*. *Nat Genet.* 39:1461–1468.
Schrider DR, Stevens KA, Cardeño CM, Langley CH, Hahn MW. 2011. Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*. *Genome Res.* 21:2087–2095.
Sebat J, Lakshmi B, Troge J, et al. (21 co-authors). 2004. Large-scale copy number polymorphism in the human genome. *Science* 305: 525–528.
Sharpton TJ, Stajich JE, Rounsley SD, et al. (24 co-authors). 2009. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Res.* 19:1722–1731.
Stark A, Lin MF, Kheradpour P, et al. (46 co-authors). 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450:219–232.