BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 1 2012, pages 48–55
doi:10.1093/bioinformatics/btr592

Phylogenetics

Advance Access publication October 28, 2011

Determining the evolutionary history of gene families

Ryan M. Ames†, Daniel Money†, Vikramsinh P. Ghatge, Simon Whelan* and Simon C. Lovell*

Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Associate Editor: David Posada

ABSTRACT

Motivation: Recent large-scale studies of individuals within a population have demonstrated that there is widespread variation in copy number in many gene families. In addition, there is increasing evidence that the variation in gene copy number can give rise to substantial phenotypic effects. In some cases, these variations have been shown to be adaptive. These observations show that a full understanding of the evolution of biological function requires an understanding of gene gain and gene loss. Accurate, robust evolutionary models of gain and loss events are, therefore, required.

Results: We have developed weighted parsimony and maximum likelihood methods for inferring gain and loss events. To test these methods, we have used Markov models of gain and loss to simulate data with known properties. We examine three models: a simple birth–death model, a single rate model and a birth–death innovation model with parameters estimated from *Drosophila* genome data. We find that for all simulations maximum likelihood-based methods are very accurate for reconstructing the number of duplication events on the phylogenetic tree, and that maximum likelihood and weighted parsimony have similar accuracy for reconstructing the ancestral state. Our implementations are robust to different model parameters and provide accurate inferences of ancestral states and the number of gain and loss events. For ancestral reconstruction, we recommend weighted parsimony because it has similar accuracy to maximum likelihood, but is much faster. For inferring the number of individual gene loss or gain events, maximum likelihood is noticeably more accurate, albeit at greater computational cost.

Availability: www.bioinf.manchester.ac.uk/dupliphy

Contact: simon.lovell@manchester.ac.uk; simon.whelan@manchester.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 2, 2011; revised on October 14, 2011; accepted on October 20, 2011

1 INTRODUCTION

Recent large sequencing projects (Clark *et al.*, 2007; Liti *et al.*, 2009; Mills *et al.*, 2011; Sudmant *et al.*, 2010) and development of whole-genome tiling arrays have allowed comparative surveys of copy number variation (CNV) of genes. CNVs arise from gene duplication and loss, and play an important role in genome evolution (Ohno, 1970). Differences in copy number are often reflected by differences in gene family size between species, as the result of gene gain via duplication and gene loss. Indeed it has been argued that CNVs represent nascent gene families (Korbel *et al.*, 2008). Variation of copy- umber can have enormous functional consequences. CNVs have been shown to be related to a range of diseases (Lupski, 2007), including developmental defects (Turner *et al.*, 2007) and autism (Glessner *et al.*, 2009). In other cases, CNVs have shown to offer an adaptive advantage. In humans, an increase in the number of copies of the amylase gene is correlated with high-starch diet (Perry *et al.*, 2007), whereas in *Plasmodium falciparum* CNVs can lead to drug resistance (Nair *et al.*, 2008). CNVs have also been shown to be under selection in *Drosophila* (Emerson *et al.*, 2008). Large differences in copy number have been found within a number of species, including human (Redon *et al.*, 2006), fly (Dopman and Hartl, 2007), mouse (Egan *et al.*, 2007) and yeast (Liti *et al.*, 2009). CNVs can also be found between sets of closely related species (Clark *et al.*, 2007; Hahn *et al.*, 2007; Heger and Ponting, 2007) and within populations of the same species (Liti *et al.*, 2009).

Given the importance of the phenotypic effects of CNV, an understanding of the processes of gain and loss is key to understanding functional evolution. Specifically, both duplication and gene loss events must be mapped to the underlying phylogenetic tree if we are to correlate genotypic change with phenotypic change or understand the effects of selection. Moreover, mapping of duplications and losses to specific branches of a phylogeny allows us to identify lineage specific gain and loss, giving insight into the ongoing adaptation to particular environments (Ames *et al.*, 2010).

Advances in technology have only recently made large-scale resequencing projects and whole-genome tiling array studies cost effective, and hence it is only recently that the importance of CNVs has been recognized. The computational problem of mapping duplications and losses to a phylogeny has, therefore, not been tackled extensively. Many of the analysis tools available to determine duplication histories on phylogenetic trees use tree reconciliation techniques (Akerborg *et al.*, 2009; Chen *et al.*, 2000; Page, 1998; Tofigh *et al.*, 2010). These methods infer gene trees for each gene family, and then reconcile these trees with a known species tree to infer gain and loss events. This approach requires the generation of gene family trees, which may be time consuming and may be affected by bias in certain circumstances (Hahn, 2007).

More recently, maximum likelihood has been used to infer the ancestral copy number of gene families given a species tree and gene family sizes for each species (De Bie *et al.*, 2006; Hahn *et al.*, 2005), and for chromosome number in relation to polyploidy (Mayrose *et al.*, 2010). These methods differ in their models of gene

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

gain and loss, assuming either homogeneity (Hahn *et al.*, 2005) or heterogeneity (Iwasaki and Takagi, 2007). A third model uses three parameters whereby gene gain is split into two parameters based on the mechanism of gene gain (Csuros and Miklos, 2006). A common feature of likelihood models is that they achieve a high degree of accuracy at the expense of speed, and so may be slow when used to infer gene family evolution from whole genome data. Moreover, a highly parameterized model may have problems converging to a single global optimum (Hahn *et al.*, 2007).

The lack of a model that incorporates the biological complexity of duplication and loss may lead to reduced accuracy of a maximum likelihood methods; for this reason, other approaches should be considered. In cases where relatively few changes have occurred along a branch, parsimony is expected to be a reasonable approximation to maximum likelihood, as demonstrated by Csuros (2008). Weighted parsimony (Sankoff, 1975) can be used to infer the ancestral copy number of gene families and allow different costs to be set for different duplication and loss events. We hypothesize that if a model of gene family evolution accurately describes the biological process, it will outperform a parsimony method. However, if the model is miss-specified the parsimony method may provide a more accurate method for inferring gene family evolution.

Here, we investigate the accuracy and robustness of parsimony and maximum likelihood approaches at inferring gene family evolution. The evolution of gene families is characterized in the most part by duplication events that increase family size and gene losses that lead to a decrease. We compared the accuracy of ancestral reconstruction and inferences of the number of events by our own implementations of parsimony and maximum likelihood with a previously published method, CAFE (De Bie *et al.*, 2006). Gene family evolutionary histories were generated under three separate models including a model based on the observed gene family sizes in nine species of *Drosophila*. We show that for estimation of the number of duplication and loss events, maximum likelihood gives very accurate results. For reconstruction of the ancestral state, weighted parsimony and maximum likelihood both perform well, with similar accuracy to previously published parsimony methods (Csuros, 2010). Interestingly, both our likelihood and parsimony tools show greater accuracy at inferring ancestral gene family sizes compared with CAFE, especially on trees with longer branches. We have also compared the performance of these methods on gene family data from nine *Drosophila* species and demonstrated that these methods show variation in inference of ancestral gene family sizes. Since weighted parsimony is much faster than maximum likelihood, we recommend it for reconstruction of the ancestral gene family, but suggest that maximum likelihood be used for inferring events on individual branches.

2 METHODS

2.1 Modelling gene gain and loss using birth–death models

We have implemented three models of gene gain and loss to describe the evolution of gene families. We have also developed two methods to infer the number of gain and loss events on a branch and reconstruct the ancestral gene family sizes at the internal nodes of a phylogenetic tree. We implement these methods in two programs: DupliPHY uses weighted parsimony to infer gain and loss events, whereas DupliPHY-ML implements maximum likelihood to infer these events.

2.1.1 Description of models In this study, we examine three Markov models of gain and loss, which treat the rate of change between the numbers of members in a gene family in a manner comparable to how substitution models in phylogenetics describe changes between, e.g. nucleotides in sequence evolution (Yang, 2006). All models only allow events that increase or decrease gene family size by one copy at a time, and for computational reasons we bound gene family size to a maximum of 75. Examining other maximum values suggest that our choice of bound does not affect our inference (data not shown). We also examine variants of our models that incorporate gene family rate variation, these models are denoted '$+\Gamma$' reflecting that family rates are drawn from a discrete Γ-distribution with four classes in a manner directly analogous to substitution models used in phylogenetics (Yang, 1993).

The first model examined, the birth–death-innovation (BDI) model (Karev *et al.*, 2002), is the most general model we consider, with its instantaneous rate matrix, **Q**, defined by Equation (1).

$$Q_{i,j}=\begin{cases} b \text{ if } j-i=1 \text{ and } i\neq 0 \text{ (birth)}\\ d \text{ if } i-j=1 \text{ (death)}\\ h \text{ if } i=0 \text{ and } j=1 \text{ (innovation)}\\ 0 \text{ if } |i-j|>1 \text{ (maximum one event)}\end{cases} \tag{1}$$

The birth and death parameters in this model represent natural gain and loss of genes in a family, whereas innovation represents the (re)gain of a gene family from other sources, such as lateral gene transfer or *de novo* gain. The birth parameter is constrained to be 1.0, whereas the death parameter must be positive. The innovation parameter has an upper bound of 10.0 to aid optimization in cases where there are short branch lengths and there is little information from which to infer its value. Our second model is a parsimony style model, termed the single rate model, which allows equiprobable gain and loss of single genes. This model is a special case of the BDI model where $b = d = h$

$$Q_{i,j}=\begin{cases} 1 \text{ if } |i-j|=1\\ 0 \text{ if } |i-j|>1\end{cases} \tag{2}$$

The final model, a birth–death model, which is also implemented by CAFE, differs from the BDI model in that the rate of birth or death is proportional to the current number of copies of a gene. Note this model has a sink state as state 0, meaning that once a gene family reaches zero copies, the family is extinct in that lineage.

$$Q_{i,j}=\begin{cases} i \text{ if } \quad |i-j|=1\\ 0 \text{ if } \quad |i-j|>1\end{cases} \tag{3}$$

For each model, the diagonal elements of **Q** are set so that each row of the matrix sums to zero. The matrices are then scaled so that the expected number of events (birth, death or innovation) per unit time is 1, allowing branch lengths to be interpreted as the number of events that have occurred on that specific branch. The stationary distribution, which describes the relative frequency of gene family sizes over long periods of time, is used to calculate the likelihood at the (pseudo-)root of the tree. For all three models, the α parameter of the Γ-distribution is constrained to be between 0.2 and 10.0, although no cases reach these bounds.

2.1.2 Implementation DupliPHY-ML uses maximum likelihood to infer branch lengths and parameters (Felsenstein, 2004), including accounting for unobservable states (Felsenstein, 1992). Standard numerical optimization techniques were used to sequentially optimize each parameter in turn until no improvement in likelihood is found. To infer ancestral states, we use the joint ancestral reconstruction method (Pupko *et al.*, 2000), where necessary using the branch-and-bound method (Pupko *et al.*, 2002). For likelihood computation, probability matrices for a branch length of t are calculated as $\mathbf{P}(t)=e^{\mathbf{Q}t}$. This exponentiation is usually performed via eigen decomposition, but the sparse nature of our matrices make this approach unstable. Instead we use the Taylor expansion for exponentiation. The stationary distribution of each Markov model is calculated by repeatedly applying a probability matrix to an arbitrary starting vector until a stable distribution is reached.

Table 1. Duplicate prediction for nine *Drosophila* species' genomes

Species	Annotated Seqs	Duplicate genes	Duplicate genes (%)
Drosophila ananassae	11257	2794	24.82
Drosophila erecta	13348	3404	25.5
Drosophila grimshawi	9261	1945	21
Drosophila melanogaster	14058	3730	26.53
Drosophila mojavenis	9245	1992	21.54
Drosophila pseudoobscura	10658	2459	23.07
Drosophila simulans	13183	3174	24.07
Drosophila virilis	9473	2034	21.47
Drosophila yakuba	13445	3516	26.15

DupliPHY implements Sankoff's dynamic programming procedure (Sankoff and Rousseau, 1975), to assign duplication and loss events on a phylogenetic tree. This algorithm uses a post-order tree-traversal to assign each internal node a cost for each potential character at that node given the characters at the descendants of the node, followed by a pre-order tree-traversal to assign ancestral states. When calculating weighted parsimony with DupliPHY, it is possible that multiple gene family sizes have the same parsimony score at the root. In cases of multiple family sizes having the same parsimony score at the root, we arbitrarily choose the family with the fewest members. To ensure this choice does not affect the accuracy of DuliPHY, we compared the accuracy of choosing the family with the fewest members to choosing a random family; we find there is little difference (Supplementary Material). The program uses a user-defined matrix of weights or costs for each gain and loss event. For this analysis, we use a single weights matrix where we assign the cost of a gain or loss of one or more genes equal to the number of events. Here gain and loss are equally likely, as has been considered in previous studies of gene family evolution (Hahn *et al.*, 2005).

2.2 Data

2.2.1 Gene families in Drosophila To test the performance of our methods on data with real biological properties, we identified gene families in nine *Drosophila* species. *Drosophila melanogaster* sequence data was taken from Adams *et al.* (2000) (release 5.12), *D.pseudoobscura* from Richards *et al.* (2005) and the remaining species (*D.simulans*, *D.yakuba*, *D.erecta*, *D.ananassae*, *D.mojavenis*, *D.virilis* and *D.grimshawi*) from the *Drosophila* comparative genomics project (Clark *et al.*, 2007). All data were downloaded from flybase (http://flybase.org/). One coding sequence was selected from each *D.melanogaster* gene at random to avoid multiple transcripts from the same gene being identified as duplicates (Hakes *et al.*, 2007). The total number of *D.melanogaster* coding sequences used in this investigation was 14 058, which excludes RNA genes and pseudogenes. BLAST (Altschul *et al.*, 1990), with the cutoff 10^{-8} was used to annotate the genes from the other species, again selecting only one coding sequence from each gene.

Duplicates were identified using GenomeHistory (Conant and Wagner, 2002) with the following parameters; BLAST threshold 10^{-8}, minimum open reading frame (ORF) translation length 100, minimum aligned residues 100 and percent identity threshold 40%. An identity threshold of 40% was used to decrease the occurrence of potential false positive paralogy assignments (Hakes *et al.*, 2007).

The number of annotated genes and identified duplicates for each species are shown in Table 1. We focus on duplicate genes because gene duplication is the main mechanism by which gene families increase in size. On average, 23.79% of genes in each genome are identified as duplicates. The number of annotated sequences are similar to the number of predicted genes identified for these species in a previous study (Heger and Ponting, 2007).

Duplicate pairs were organized into gene families using agglomerative hierarchical clustering, where duplicate pairs were clustered by common members of a pair until all clusters had no overlap. This clustering yielded a total of 1481 gene families. The size distribution of the identified gene families shows that the majority of gene families are small, with few large gene families present in the data; the average family size in the *Drosophila* data is only 2.9 genes, with only 1 gene family having >75 members. (Fig. 1). This distribution is similar to that identified in other studies looking at gain and loss of protein domains (Karev *et al.*, 2002) and gene duplication ages (Lynch and Conery, 2000, 2003). We remove three families from our dataset as these had unusual patterns of variation, leaving 1478 families. We find that many of the genes in these high variation families have no functional annotation, and so may represent erroneously annotated families. These families show very different properties when compared with the majority of families identified and as such are likely to be under different selective constraint. Therefore, removing these families allows us to remove those families which are unlikely to be adequately described by simple models of gain and loss.

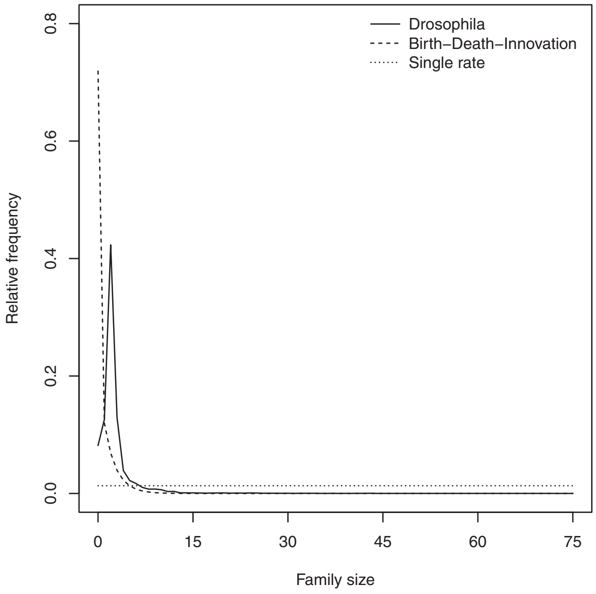

Fig. 1. The distribution of average gene family sizes across nine species of *Drosophila*, compared with the stationary distribution of our single rate model and birth–death–innovation model with parameters estimated from the *Drosophila* data. Note that the models used assume a homogeneous distribution throughout the tree.

2.2.2 Data simulation schemes To simulate gene family evolution, we use a standard Monte Carlo simulation that draws from the stationary distribution at the root of the tree and uses transition matrices to model changes in gene number along branches of a tree, an approach common in phylogenetic applications (Yang, 2006). To ground our simulations in biological reality, we use the *Drosophila* data as inspiration for our simulation scheme. Each simulation uses one of our three Markov model to create the evolutionary history of 1481 gene families over a tree, while ensuring that no family has an unobservable pattern. The parameters for our simulations are based on those estimated for the BDI model applied to the *Drosophila* data. The α parameter for the Γ-distribution was also estimated from the *Drosophila* data using the BDI model. The same value of α was used for simulation under all models. The relative estimated parameters are 1.0 (birth), 1.741 (death), 0.289 (innovation) and 0.432 (α). These estimates mean for every birth event that occurs there are 0.74 innovation events and 1.74 death events.

We examine trees with 4 or 8 taxa under 10 different tree lengths (sum of all branch lengths), which is intended to represent a range of biologically

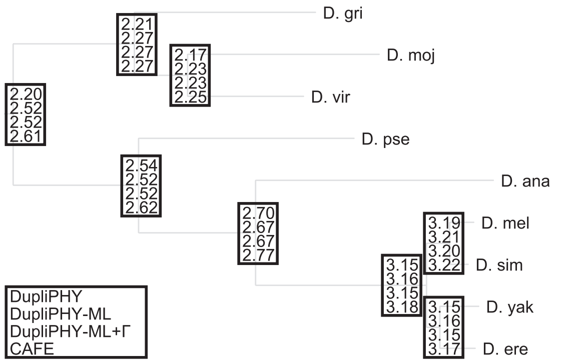

Fig. 2. The average ancestral gene family sizes of 1478 *Drosophila* gene families inferred by DupliPHY, DupliPHY-ML, DupliPHY-ML+Γ and CAFE. The values at each internal node shows the average ancestral family size predicted by DupliPHY (top), DupliPHY-ML (second from top), DupliPHY-ML+Γ (third from top) and CAFE (bottom). We can see that these values are more variable nearer the root of the tree and on longer branches.

plausible scenarios. We use a range of different lengths as we expect this to be representative of real trees. We expect parsimony to have worse performance on longer trees (Felsenstein, 1978). We use two different numbers of taxa to see how well the different methods perform when long branches are split and more information is provided. To ensure our tree topologies and relative branch lengths are representative of those that occur in real data we randomly sample trees with an appropriate number of taxa from TreeBASE (Sanderson *et al.*, 1994). For each tree selected, the branch lengths are rescaled to the tree length we wish to examine. Note that the BDI and birth–death model are non-reversible, so roots are chosen using the midpoint rooting. After scaling, any trees with very short branch lengths ($<5\times10^{-6}$) were removed. These trees were removed as CAFE requires that branch lengths are non-zero integer numbers and the short branch lengths fall below the limit of our conversion factor to produce integer branch lengths. For each set of conditions examined, we perform 50 simulations.

3 RESULTS

3.1 Real data analysis

3.1.1 Drosophila data We inferred the ancestral reconstruction of gene family size in 1478 *Drosophila* gene families on the nine species phylogeny. This phylogeny is based on that of Pollard *et al.* (2006), and was provided, including branch lengths by the author (D. Pollard, personal communication). CAFE, DupliPHY, DupliPHY-ML and DupliPHY-ML+Γ were used to infer the ancestral family sizes and the estimates were averaged over all families. We can see that the methods produce very similar estimates of ancestral gene family size toward the tips of the tree. However, as we move toward the root and on longer branches, there is more variation in the estimates (Fig. 2).

3.1.2 Model fit In order to assess how accurately our models fit the *Drosophila* data, we first compare the maximum likelihood of real data under both the single rate model and the BDI innovation model. A likelihood ratio test shows that the BDI+Γ model (3 df; log-likelihood: −10514.8) provides a significantly better fit than the single rate model +Γ (1 df; log-likelihood: −14626.2; $P \ll 0.001$). We next compare the stationary distributions of the single rate and BDI models with the real distribution of family size from the *Drosophila* data (Fig. 1). The stationary distribution of the single rate model is one where each state is equally likely and is significantly different from the *Drosophila* data ($P \ll 0.001$; Pearson's χ^2 test). The stationary distribution of the BDI model is also significantly different to the *Drosophila* data ($P \ll 0.001$; Pearson's χ^2 test), although its shape is much closer to that of the *Drosophila* data. These differences suggest that neither the single rate or BDI model are adequate descriptions of the *Drosophila* data. Note that the sink-state in the birth–death model mean its stationary distribution is a point mass on zero, which is not useful to compare with the real data. For the following analyses, we only perform maximum likelihood inference under the BDI model because from the models we examine it appears to provide the best description of real data.

3.2 Simulation

Here, we assess the performance and robustness of these methods on a variety of trees with different lengths and number of taxa. We aim to identify the type of data upon which specific models perform well or otherwise. We tested the accuracy of inferring the number of events on a branch and ancestral reconstruction of DupliPHY, CAFE and DupliPHY-ML for simulated data under all three models of gene family evolution.

Note that CAFE is only used for benchmarking ancestral reconstruction because under its birth–death model one cannot compute the number of events on a branch as the scaling factor requires a non-zero stationary distribution. No other programs are available for benchmarking. The probabilistic model implemented in COUNT (Csuros, 2010) only annotates ancestral species as containing 0, 1 or more members of a family, whereas the parsimony method produces indistinguishable results from DupliPHY (Supplementary Material). The method of Iwasaki and Takagi (2007) allows a maximum gene family size of three.

3.2.1 Inferring the number of gain and loss events Inferring the number of gain and loss events on branches allows the identification of lineages with a high turnover of genes, which may be the result of factors such as relaxation of natural selection, adaptation or changes in the effective population size. Despite this, few available methods explicitly provide this information. We compared the inference of the number of events along the tree by weighted parsimony as implemented in DupliPHY and maximum likelihood as implemented in DupliPHY-ML(+Γ), on a range of simulated data (Fig. 3).

Under simulated data produced from the birth-death and single rate model, we observe a decrease in accuracy of the number of events inferred by DupliPHY as tree length increases (Fig. 3), although the inclusion of additional taxa reduces the degree of error. A similar effect of tree length on accuracy occurs under DupliPHY-ML. In contrast, DupliPHY+Γ produces the best estimates of duplication tree lengths. Under simulations from BDI, the performance of DupliPHY, and to a lesser extent DupliPHY-ML, appear to worsen, although DupliPHY-ML+Γ still recovers accurate tree estimates. The unusual performance of DupliPHY, where parsimony overestimates the amount of evolution, appears to be caused by the lack of a correction for sites removed from the analysis. The reason this problem affects the BDI simulations and not the others may be because of an interaction between innovation and the frequency of missing data, whereby allowing innovation in a

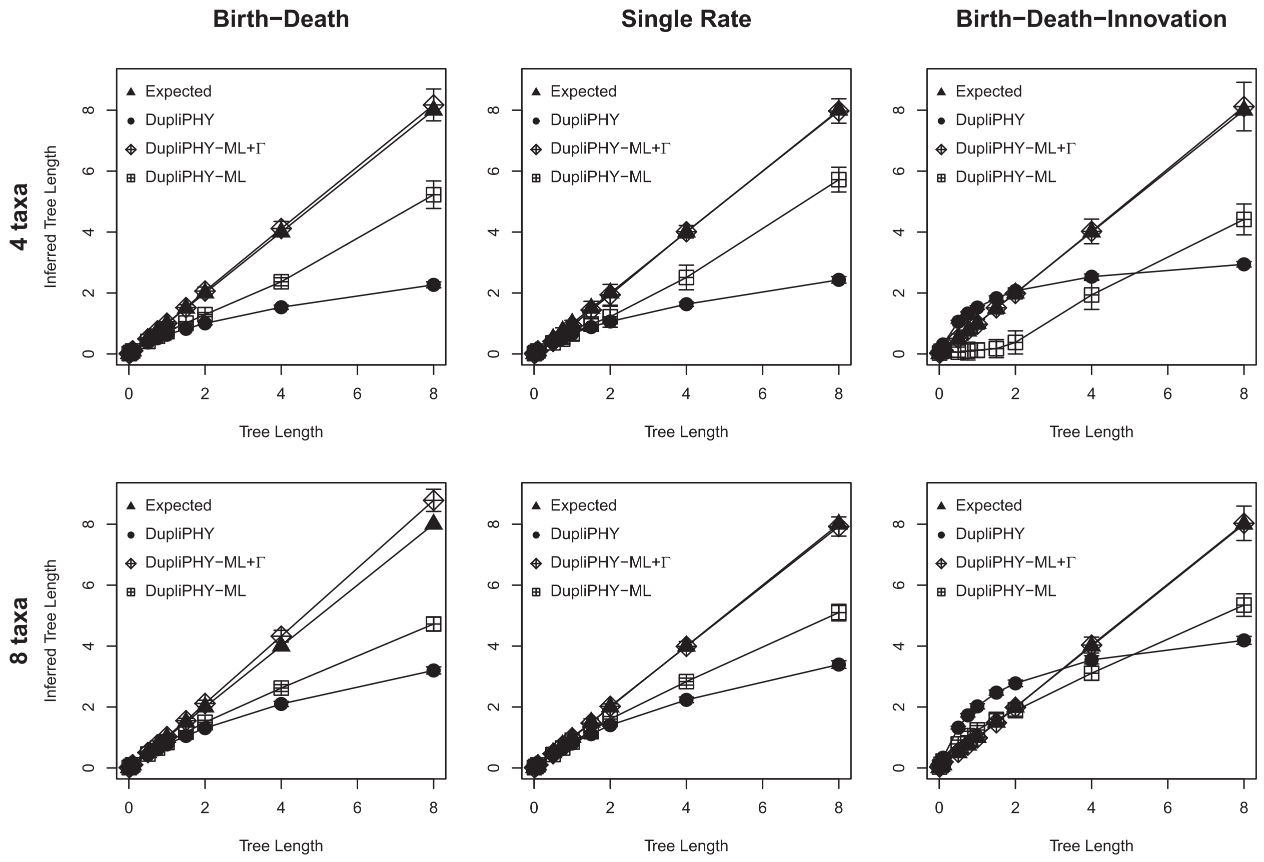

Fig. 3. Accuracy of duplication tree length inference by weighted parsimony and maximum likelihood. Inferences were made over 10 tree lengths each with 50 repetitions containing 1481 gene families. Data were simulated under our birth–death, single rate and birth–death–innovation models. The triangles, closed circles, crossed diamonds and crossed squares show the inferred duplication branch lengths for expected length, DupliPHY, DupliPHY-ML+Γ and DupliPHY-ML, respectively. Error bars are SDs.

model also results in a higher death parameter, which in turn means more genes are expected to be missing from one or more genomes. Extra leaf nodes appear to alleviate this problem.

Examining the number of events inferred across an entire tree may miss important differences in the number of events inferred on single branches. To ensure we are not missing any branch-specific bias, we calculated the root mean square deviation (RMSD) between simulation and inferred branch lengths values (Supplementary Fig. S1). These data follow similar patterns to those in Figure 3 and do not suggest any obvious form of bias.

3.2.2 Ancestral reconstruction of gene family sizes Ancestral reconstruction of gene family sizes is the focus of the majority of methods that analyze gene family evolution. We measured the accuracy of ancestral reconstruction by taking the average of the absolute value of the difference between the inferred and simulated family size, averaged across all the ancestral nodes. We can therefore determine how far the inferences of each method are from the simulated value.

COUNT (Csuros, 2010) provides a parsimony reconstruction of the ancestral size of gene families. We find that there is very little difference between DupliPHY and COUNT for simulations under any of the three models (Supplementary Fig. S2). Both methods use Wagner parsimony and so the small advantages in accuracy for DupliPHY are probably due to the differences in the handling of tied parsimony scores at the root. Since the differences are so small we include only DupliPHY as a parsimony methodology for subsequent analysis.

On trees with short branch lengths, we find that there is very little difference between the accuracy of our weighted parsimony and the two maximum likelihood approaches (Fig. 4). As branch length increases, DupliPHY-ML+Γ consistently produces the most accurate inference of ancestral gene family sizes. We conclude from these results that on trees with short average branch lengths, weighted parsimony is a viable method to infer ancestral gene family sizes. As branch lengths increase, maximum likelihood methods are needed to get the most accurate estimates. Under BDI, both DupliPHY and DupliPHY-ML produce reasonable ancestral reconstructions despite the problems they have inferring tree lengths.

Interestingly, the maximum likelihood methodology implemented by CAFE shows reduced accuracy when compared to the three DupliPHY implementations (Fig. 4). This result is in line with the differences seen for the ancestral reconstruction of *Drosophila* gene families above. Our birth–death model was developed to recreate the characteristics of CAFE's model. On data simulated under this

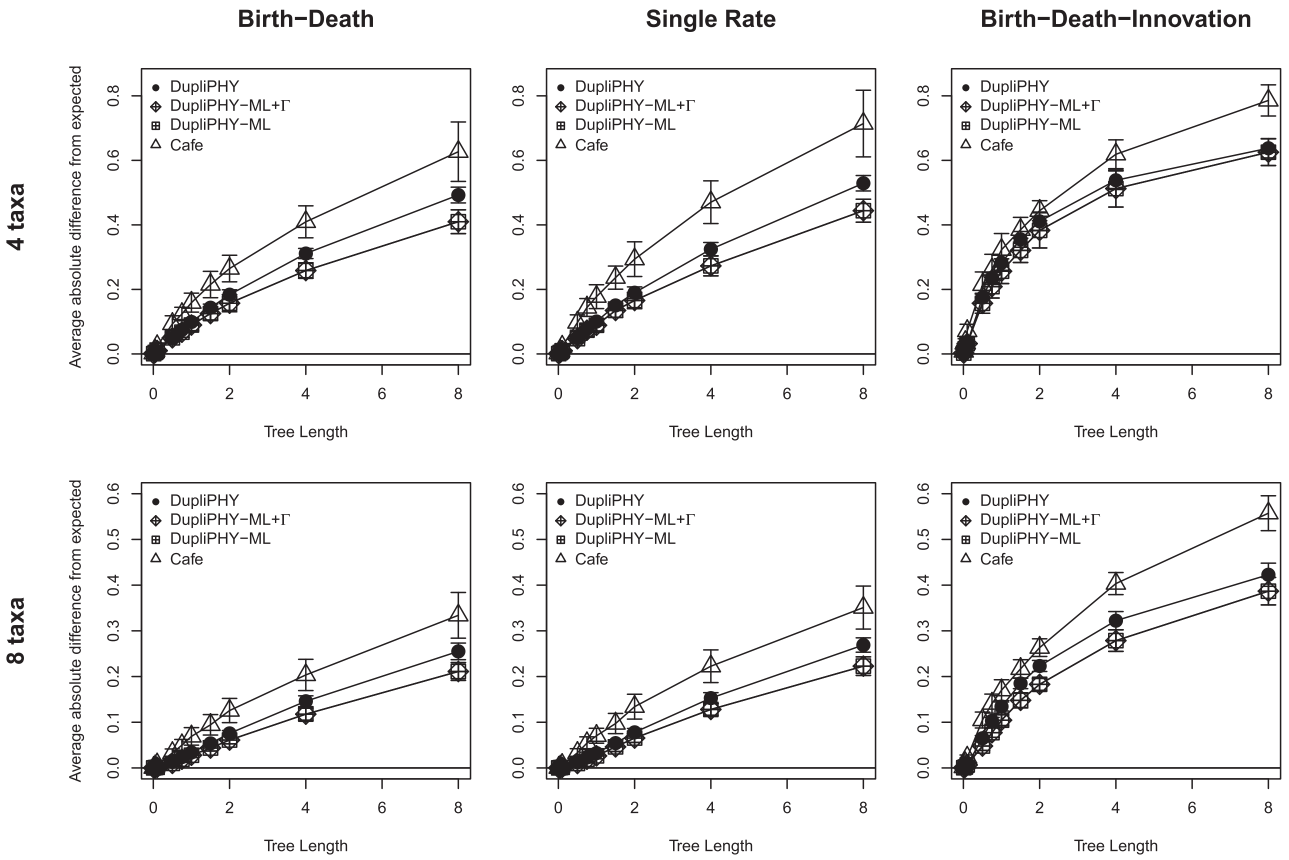

Fig. 4. The accuracy of several methods on determining ancestral gene family sizes. Inferences were made over 10 tree lengths each with 50 repetitions containing 1481 gene families. Data were simulated under our birth–death, single rate and birth–death–innovation models. The closed circles, crossed diamonds, crossed squares and open triangles show the performance of DupliPHY, DupliPHY-ML+Γ, DupliPHY-ML and CAFE, respectively. Error bars are SDs. The line at 0 represents the simulated value.

model CAFE's performance is close to that of the other methods on trees with short branch lengths, and becomes less accurate as branch length increases. However, on data simulated under the single rate and BDI models CAFE shows reduced accuracy on trees with short branch lengths. The two maximum likelihood methods may show substantially different results because of the differences in the methods implementation or the underlying models of gene family evolution.

4 DISCUSSION

In order to be able to further our understanding of functional evolution, we must understand the processes of gene gain and loss. Here, we have developed methods for inferring these events and the ancestral gene family sizes on a tree. We have compared the inferences of these methods with CAFE (De Bie *et al.*, 2006), on gene families identified from *Drosophila* data and on simulated data. We see that on *Drosophila* data all methods perform similarly for internal nodes near the tips of the tree, but vary more on longer branches toward to the root. Over all gene families we see that the methods produce inferences of gene family size that are more similar to each other than to CAFE. Even where the average variation in the inferences made by these methods is small it may be important in specific cases, particularly for those families with lots of variation between species.

We use simulated data to compare the accuracy of our methods on data with a known evolutionary history and to compare the robustness of these methods on data produced under known conditions. The accuracy of three methods for inferring the number of birth and death events across the whole tree (weighted parsimony, as implemented in DupliPHY, and maximum likelihood, as implemented in DupliPHY-ML and DupliPHY-ML+Γ) was compared across all three models (Fig. 3). The maximum likelihood methods provide an accurate estimate of the number of gain–loss events, provided rate variation between genes is incorporated in the model. The accurate inference for all three simulation schemes suggest that the model may be reasonably robust to minor misspecification when describing the process of gene gain and loss, although failure to incorporate events that change gene family number by greater than one (Spencer *et al.*, 2006) or affect multiple genes may still cause inaccurate inference. The failure of parsimony to infer correctly the number of events along a branch is a well-known shortcoming, with the problem being analogous to long branch attraction (Felsenstein, 1978). We conclude that maximum likelihood, with an appropriate probabilistic model, is well suited for inferring the number of gene gain and loss events along a branch, which may reveal interesting evolutionary factors in a particular region of the tree.

Finally, we examined the accuracy of maximum likelihood methods implementing probabilistic models and weighted parsimony on ancestral reconstruction of gene family sizes on data simulated under three models (Fig. 4). Under all models of data simulation, DupliPHY-ML+Γ is the most accurate method for inferring ancestral gene family sizes. This difference is most pronounced on longer trees, while on trees with shorter branches DupliPHY, DupliPHY-ML and DupliPHY-ML+Γ perform equally well, confirming the findings of previous studies (Iwasaki and Takagi, 2007). CAFE is the least accurate of the three methods over all models and shows reduced accuracy on short branches under the single rate and BDI models. This reduction in accuracy is probably because of the implementation of birth–death model that CAFE uses for inference rather than a property of maximum likelihood inference. The birth–death model causes problems for likelihood computations because of the presence of a sink-state, which precludes simple likelihood computation and the approximation required appears to affect the accuracy of inference. We observed similar results when analyzing the *Drosophila* data, where CAFE produced the most divergent estimates of ancestral gene family size when compared to all three versions of DupliPHY. We conclude that for those trees with shorter branch lengths or where a reliable probabilistic model is unavailable, weighted parsimony produces similar results to maximum likelihood. Both likelihood and parsimony have additional benefits that may mean these methods are more suited to specific situations. Likelihood can generate confidence intervals to demonstrate the reliability of the inference and parsimony is much faster, running in ∼3 min on the *Drosophila* data compared with ∼6 h for DupliPHY-ML and 14 h for DupliPHY-ML+Γ . This difference in runtime may be useful when many runs on large datasets are required (Felsenstein, 1978; Hahn *et al.*, 2007; Iwasaki and Takagi, 2007).

The differences between the stationary distribution of the BDI model and the empirical distribution of the *Drosophila* data suggests that the BDI model is not an adequate description of gene gain and loss (Fig. 1). BDI is a simple model and does not describe many known biological mechanisms for gene duplication and loss, for example large-scale duplication events, such as segmental or whole genome duplication, and large-scale gene loss. These models also assume that innovation is a frequent mechanism, which seems unlikely in eukaryotes (Cai *et al.*, 2008; Knowles and McLysaght, 2009; Zhou *et al.*, 2008). Another potential issue is the assumption that the process describing gene family evolution is stationary, and that the stationary distribution can be derived from this model. This, and other assumptions may lead to the discrepancy between the stationary distribution and the observed size of gene families shown in Figure 1. Although the tools described here seem adequate for inferring ancestral states and the number of changes in gene number, more biologically sophisticated models may tell us more about the specific mechanisms of gene family evolution, allowing us to address fundamentally important questions about copy number variation.

5 CONCLUSION

In order to understand the evolution of gene families through gene duplication and loss, we must be able to map gain and loss events on a phylogenetic tree. The two methods we have developed allow us to map these events to a tree. Using gene family data from nine *Drosophila* species, we found that the methods tended to vary more in their inferences of ancestral gene family size on longer branches near the root of the tree. On data simulated under a variety of models, maximum likelihood provides the most accurate and robust method of determining ancestral gene family sizes and identifying the individual events along a branch. However, we also see that weighted parsimony performs equally well as maximum likelihood at ancestral reconstruction on trees with shorter branch lengths. Overall, we find that the accuracy of maximum likelihood is dependent on the underlying probabilistic model used to infer gain and loss and that more work is required to accurately describe the processes of gene family evolution.

ACKNOWLEDGEMENTS

We thank John Archer for providing helpful discussion on the implementation of DupliPHY.

Funding: Biotechnology and Biological Sciences Research Council, UK.

Conflict of Interest: none declared.

REFERENCES

Adams,M. *et al.* (2000) The Genome Sequence of Drosophila melanogaster. *Science,* **287**, 2185–2195.

Akerborg,O. *et al.* (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl Acad. Sci. USA,* **106**, 5714–5719.

Altschul,S. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.,* **215**, 403–410.

Ames,R.M. *et al.* (2010) Gene duplication and environmental adaptation within yeast populations. *Genome Biol. Evol.,* **2**, 591–601.

Cai,J. *et al.* (2008) De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics,* **179**, 487–496.

Chen,K. *et al.* (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.,* **7**, 429–447.

Clark,A. *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature,* **450**, 203–218.

Conant,G. and Wagner,A. (2002) GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.,* **30**, 3378–3386.

Csuros,M. (2008) Ancestral reconstruction by asymmetric wagner parsimony over continuous characters and squared parsimony over distributions. In *Comparative Genomics: International Workshop, RECOMB-CG 2008, Paris, France, October 13-15, 2008, Proceedings.* vol. 5267. Springer, New York, pp. 72–78.

Csuros,M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics,* **26**, 1910–1912.

Csuros,M. and Miklos,I. (2006) A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *Lect. Notes Comput. Sci.,* **3909**, 206–220.

De Bie,T. *et al.* (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics,* **22**, 1269–1271.

Dopman,E. and Hartl,D. (2007) A portrait of copy-number polymorphism in Drosophila melanogaster. *Proc. Natl Acad. Sci. USA,* **104**, 19920–19925.

Egan,C. *et al.* (2007) Recurrent DNA copy number variation in the laboratory mouse. *Nat. Genet.,* **39**, 1384–1389.

Emerson,J. *et al.* (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science,* **320**, 1629–1631.

Felsenstein,J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.,* **27**, 401–410.

Felsenstein,J. (1992) Phylogenies from restriction sites: a maximum-likelihood approach. *Evolution,* **46**, 159–173.

Felsenstein,J. (2004) *Inferring Phytogenies.* Sinauer Associates, Sunderland, Massachusetts.

Glessner,J. *et al.* (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature,* **459**, 569–573.

Hahn,M. (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.,* **8**, R141–R149.

Hahn,M. *et al.* (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.,* **15**, 1153–1160.

Hahn,M. *et al.* (2007) Gene family evolution across 12 Drosophila genomes. *PLoS Genet.,* **3**, e197–e209.

Hakes,L. *et al.* (2007) All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.,* **8**, R209–R222.

Heger,A. and Ponting,C. (2007) Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res.,* **17**, 1837.

Iwasaki,W. and Takagi,T. (2007) Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. *Bioinformatics,* **23**, i230–i239.

Karev,G. *et al.* (2002) Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol. Biol.,* **2**, 18–34.

Knowles,D. and McLysaght,A. (2009) Recent de novo origin of human protein-coding genes. *Genome Res.,* **19**, 1752–1759.

Korbel,J. *et al.* (2008) The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr. Opin. Struct. Biol.,* **18**, 366–374.

Liti,G. *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature,* **458**, 337–341.

Lupski,J. (2007) Genomic rearrangements and sporadic disease. *Nat. Genet.,* **39**, S43–S47.

Lynch,M. and Conery,J. (2000) The evolutionary fate and consequences of duplicate genes. *Science,* **290**, 1151–1155.

Lynch,M. and Conery,J. (2003) The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics,* **3**, 35–44.

Mayrose,I. *et al.* (2010) Probabilistic models of chromosome number evolution and the inference of polyploidy. *Syst. Biol.,* **59**, 132–144.

Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature,* **470**, 59–65.

Nair,S. *et al.* (2008) Adaptive copy number evolution in malaria parasites. *PLoS Genet.,* **4**, e1000243–e1000253.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer, New York.

Page,R. (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics,* **14**, 819–820.

Perry,G. *et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.,* **39**, 1256–1260.

Pollard,D. *et al.* (2006) Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS Genet.,* **2**, 1634–1647.

Pupko,T. *et al.* (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.,* **17**, 890–896.

Pupko,T. *et al.* (2002) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: application to the evolution of five gene families. *Bioinformatics,* **18**, 1116–1123.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature,* **444**, 444–454.

Richards,S. *et al.* (2005) Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res.,* **15**, 1–18.

Sanderson,M. *et al.* (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am. J. Botany,* **81**, 183–187.

Sankoff, D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math.,* **28**, 35–42.

Sankoff,D. and Rousseau,P. (1975) Locating the vertices of a steiner tree in an arbitrary metric space. *Math. Program.,* **9**, 240–246.

Spencer,M. *et al.* (2006) Modelling prokaryote gene content. *Evol. Bioinform Online,* **2**, 157–78.

Sudmant,P.H. *et al.* (2010) Diversity of human copy number variation and multicopy genes. *Science,* **330**, 641–646.

Tofigh,A. *et al.* (2010) Simultaneous identification of duplications and lateral gene transfers. *IEEE IEEE/ACM Trans. Comput. Biol. Bioinformatics,* **8**, 517–535.

Turner,D. *et al.* (2007) Germline rates of de novo meiotic deletions and duplications causing s everal genomic disorders. , 90–95.

Yang,Z. (1993) Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Mol. Biol. Evol.,* **10**, 1396–1401.

Yang,Z. (2006) *Computational Molecular Evolution*. Oxford University Press, USA.

Zhou,Q. *et al.* (2008) On the origin of new genes in Drosophila. *Genome Res.,* **18**, 1446–1455.