Determining the evolutionary history of gene families
Ryan M. Ames 1*, Daniel Money 1†, Vikramsinh P Ghatge 1, Simon Whelan 1‡, and Simon C. Lovell 1†

1Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
Received on XXXXX; revised on XXXXX; accepted on XXXXX

ABSTRACT

Motivation: Recent large-scale studies of individuals within a population have demonstrated that there is wide-spread 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

1 INTRODUCTION

Recent large sequencing projects (Clark et al., 2007; Liti et al., 2009; Sudmant et al., 2010; Mills et al., 2011) 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-number 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 & 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 & 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 so it is only recently that the importance of CNVs has been recognised. 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 (Page, 1998; Chen et al., 2000; Akerborg et al., 2009; 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 gain and loss, assuming either homogeneity (Hahn et al., 2005) or...
heterogeneity (Iwasaki & Takagi, 2007). A third model uses three parameters whereby gene gain is split into two parameters based on the mechanism of gene gain (Csuros & Miklov, 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 parameterised 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 hypothesise 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. 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 3 separate models including a model based on the observed gene family sizes in 9 species of Drosophila. We showed 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 to CAFE, especially on trees with longer branches. We have also compared the performance of these methods on gene family data from 9 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 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 “+Γ” 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.

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

\[
Q_{i,j} = \begin{cases} 
 1 & \text{if } |i - j| = 1 \\
 0 & \text{if } |i - j| > 1 
\end{cases}
\] (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 one a gene family reaches zero copies the family is extinct in that lineage.

\[
Q_{i,j} = \begin{cases} 
 i & \text{if } |i - j| = 1 \\
 0 & \text{if } |i - j| > 1 
\end{cases}
\] (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 \(\alpha\) parameter of the \(\Gamma\)-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 optimisation techniques were used to sequentially optimise 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 $P(t) = Q_t$. 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.

DupliPHY implements Wagner parsimony (Farriss, 1970), a special case of weighted parsimony (Sankoff, 1975), to assign duplication and loss events on a phylogenetic tree. This algorithm uses a post-order tree-traversal and 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. In cases of ambiguity when assigning ancestral states we arbitrarily choose that with lowest gene count, although other choices are not expected to affect our results (see 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 9 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. mojavensis, 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 & Wagner, 2002) with the following parameters; BLAST threshold 10^{-8}, minimum open reading frame (ORF) translation length 100, minimum aligned residues 100, 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. 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 & Ponting, 2007).

Duplicate pairs were organised 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 more than 75 members. (Figure 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 & 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 to 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.

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 $\alpha$ parameter for the $\Gamma$-distribution was also estimated from the Drosophila data using the BDI model. The same value of $\alpha$ was used for simulation under all models. The relative estimated parameters are 1.0 (birth), 1.741 (death), 0.289 (innovation) and 0.432 ($\alpha$). 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

<table>
<thead>
<tr>
<th>Species</th>
<th>Annotated Seqs</th>
<th>Duplicate Genes</th>
<th>Duplicate Genes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. ananassae</td>
<td>11257</td>
<td>2794</td>
<td>24.82</td>
</tr>
<tr>
<td>D. erecta</td>
<td>13348</td>
<td>3404</td>
<td>25.5</td>
</tr>
<tr>
<td>D. grimshawi</td>
<td>9261</td>
<td>1945</td>
<td>21</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>14058</td>
<td>3735</td>
<td>26.53</td>
</tr>
<tr>
<td>D. mojavensis</td>
<td>9245</td>
<td>1992</td>
<td>21.54</td>
</tr>
<tr>
<td>D. pseudoobscura</td>
<td>10658</td>
<td>2459</td>
<td>23.07</td>
</tr>
<tr>
<td>D. simulans</td>
<td>13183</td>
<td>3174</td>
<td>24.07</td>
</tr>
<tr>
<td>D. virilis</td>
<td>9473</td>
<td>2034</td>
<td>21.47</td>
</tr>
<tr>
<td>D. yakuba</td>
<td>13445</td>
<td>3516</td>
<td>26.15</td>
</tr>
</tbody>
</table>

Table 1. Duplicate prediction for 9 Drosophila species’ genomes.
of biologically 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 mid-point rooting. After scaling, any trees with very short branch lengths (< 5 x 10⁻⁶) were removed. These trees were removed as CAFE requires that branch lengths are non zero integer numbers and the short branch lengths fell 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 9 species phylogeny (Pollard et al., 2006). 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 towards the tips of the tree. However, as we move towards the root and on longer branches there is more variation in the estimates. (Figure 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-value < 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 (Figure 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 < 0.001; Pearson’s χ² test). The stationary distribution of the BDI model is also significantly different to the Drosophila data (P < 0.001; Pearson’s χ² test), although its shape is much closer to that of the Drosophila data. These differences suggest 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 provides 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 because 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 (see Supplementary Material). The method of Iwasaki & 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(+) on a range of simulated data (Figure 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 (Figure 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(+) 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(+) still recovers accurate tree estimates. The unusual performance of DupliPHY, where parsimony over-estimates 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 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 Figure 1). 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 analyse 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 Figure 2). 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 (Figure 4). As branch length increases DupliPHY-ML(+) 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 (Figure 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 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 towards 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(+) was compared across all three models (Figure 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 mis-specification 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 (Figure 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 & Takagi, 2007). CAFE is the least accurate of the three methods under 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 analysing 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 about 3 minutes on the Drosophila data compared to about 6 hours for DupliPHY-ML and 14 hours for DupliPHY-ML+Γ. This difference in runtime may be useful when many runs on large data sets are required (Felsenstein, 1978; Iwasaki & Takagi, 2007; Hahn et al., 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...
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+, DupliPHY-ML and CAFE respectively. Error bars are standard deviations. The line at 0 represents the simulated value.

and loss (Figure 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, large-scale gene loss. These models also assume that innovation is a frequent mechanism, which seems unlikely in eukaryotes (Cai et al., 2008; Zhou et al., 2008; Knowles & McLysaght, 2009). 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 9 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: This work was supported by the Biotechnology and Biological Sciences Research Council, UK.
REFERENCES

Ames et al

BMC Evolutionary Biology, 2 (1), 18–34.


