RESEARCH ARTICLE   **Open Access**

CrossMark

A generalized birth and death process for modeling the fates of gene duplication

Jing Zhao[1], Ashley I. Teufel[2,3], David A. Liberles[2,3*] and Liang Liu[1,4*]

Abstract

Background: Accurately estimating the timing and mode of gene duplications along the evolutionary history of species can provide invaluable information about underlying mechanisms by which the genomes of organisms evolved and the genes with novel functions arose. Mechanistic models have previously been introduced that allow for probabilistic inference of the evolutionary mechanism for duplicate gene retention based upon the average rate of loss over time of the duplicate. However, there is currently no probabilistic model embedded in a birth-death modeling framework that can take into account the effects of different evolutionary mechanisms of gene retention when analyzing gene family data.

Results: In this study, we describe a generalized birth-death process for modeling the fates of gene duplication. Use of mechanistic models in a phylogenetic framework requires an age-dependent birth-death process. Starting with a single population corresponding to the lineage of a phylogenetic tree and with an assumption of a clock that starts ticking for each duplicate at its birth, an age-dependent birth-death process is developed by extending the results from the time-dependent birth-death process. The implementation of such models in a full phylogenetic framework is expected to enable large scale probabilistic analysis of duplicates in comparative genomic studies.

Conclusions: We develop an age-dependent birth-death model for understanding the mechanisms of gene retention, which allows a gene loss rate dependent on each duplication event. Simulation results indicate that different mechanisms of gene retentions produce distinct likelihood functions, which can be used with genomic data to quantitatively distinguish those mechanisms.

Keywords: Gene duplication, Phylogenetic methods, Probabilistic models, Birth-death processes, Stochastic processes

Background

A gene family is a group of genes with similar sequences that show evidence of descent from a common ancestor [1–3]. This includes orthologs that originate through speciation as well as duplicates (modeled here) that can be found within a species or shared between species from an older duplication event that predated speciation. The large number of genes per family suggests that the newly arisen gene duplicates are potentially major contributors to evolutionary novelties [4–7]. Gene duplication can provide raw genetic material for evolutionary forces to act on. Although a majority of duplicate genes

may be silenced by degenerative mutations or lost due to population dynamics, some duplicated genes are able to evolve novel functions permanently preserved in the population [8, 9]. Accurately estimating the timing and mode of gene duplications along the evolutionary history of species can provide invaluable information about underlying mechanisms by which the genomes of organisms evolved and the genes with novel functions arose [10].

Several biological models have been proposed to depict the mechanisms that lead to different evolutionary fates of a gene duplicate [11–14]. Nonfunctionalization refers to the process in which mutations occur on one of the gene duplicates and produce a non-functional protein [11, 15]. The neofunctionalization model [16] assumes that duplication itself does not affect fitness. Although a duplicate is most likely to be pseudogenized by degenerative mutation (nonfunctionalization) or lost due to population dynamics [9], the redundant copy may

* Correspondence: daliberles@temple.edu; lliu@uga.edu
[2]Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA
[1]Department of Statistics, University of Georgia, 101 Cedar Street, Athens, GA 30602, USA
Full list of author information is available at the end of the article

Zhao *et al. BMC Evolutionary Biology* (2015) 15:275

Page 2 of 11

occasionally acquire a new beneficial function through mutation that will be preferentially preserved in the population. While this function may subsequently be optimized and accommodated within the genome structure (assuming a coding sequence change) by an evolutionary Stokes shift [17], the initial event leading to retention is a single beneficial change. The waiting time for this single change gives rise to a convexly decaying hazard function when modeled together with non-functionalizing changes and is referred to as the neofunctionalization model (see [15, 18, 19] for a review). The duplication-degeneration-complementation model [20] describes a so-called subfunctionalization mechanism in which two gene copies are partially damaged by degenerated mutations. Both copies must be maintained in order to perform the original function of the gene [21, 22]. This model, called subfunctionalization, involves a waiting time for multiple events to occur as deleterious substitutions accumulate in both copies before the retaining mutation can occur. This waiting time for multiple changes gives rise to a switch from a convex to a concave (sigmoidal) hazard function when modeled together with non-functionalizing mutations (again, see [15, 18, 19] for a review and engaged discussion). In addition to the processes acting on individual genes, large-scale gene duplication events (for example, whole genome duplication) may have occurred and produced multiple interacting genes together creating an additional retention mechanism. Dosage balance promotes the retention of duplicated interaction networks, as loss of interaction stoichiometry can lead to declines in fitness. This gives rise to very different retention dynamics compared to neofunctionalization or subfunctionalization (see [15, 18, 19] for a review). The mechanistic models described for nonfunctionalization, subfunctionalization, neofunctionalization, and dosage balance represent one of many conceivable modeling frameworks for duplicate gene retention (see [19] for an enhanced discussion). The models here are used within a single population, reflecting a lineage of a phylogenetic tree, but the ultimate aim is to extend their use into an interspecific phylogenetic framework with the population genetic assumptions that accompany this. Simpler models have already been incorporated into a fuller phylogenetic framework of this nature (see for example [23]).

Accurately reconstructing the evolution of gene families requires informative datasets, powerful mathematical models, and efficient computational algorithms. Advanced biotechnologies provide a vast amount of genetic data for understanding the evolution of gene families [24, 25]. Meanwhile, probabilistic models, describing the process of gene family evolution, significantly enhance our ability to extract useful information from genetic data [26–29]. The birth-death (BD) model [30], which has been broadly applied in analyzing species phylogenies [25, 29, 31, 32],

could also be adopted in phylogenetic analysis of gene families [33]. In 1975, Thompson [34] introduced a phylogenetic model based on the birth-death process to understand the evolution of human populations. Under the generalized birth-death model, Nee et al. [35] derived a reconstructed evolutionary process [36] to estimate birth and death rates in a interspecific phylogenetic framework. Rannala and Yang [37] developed a birth-death phylogenetic model for estimating phylogenetic trees from molecular sequence data. Aldous and Popovic [38] proposed a continuous-time critical branching process conditioned on the number of species in the present, with the assumption that the birth and death rates are identical in macroevolution, which was later relaxed by Gernhard [39, 40] to allow uncorrelated birth and death rates. With the assumption of constant birth and death rates, Stadler [41] derived the probability density function of a phylogenetic tree under the birth-death model. Recently, time-dependent BD processes have attracted more attention as a mode of performing hypothesis-driven research [42–45]. Rabosky [42] distinguished rate-variable models of diversification from rate-constant models by fitting BD models using likelihood methods. Hohna [44, 46] and Hallinan [45] studied the reconstructed process with time-dependent rates in a more general setting by relaxing the assumptions about the number of species and the time of the process. The BD model was first adopted in [47] and further extended by other researchers to reconcile gene and species trees (Arvestad et al. [48], Akerborg et al. [23], Rasmussen and Kellis [49] and Sjostrand et al. [50]). Recently, Boussau et al. [51] established a BD phylogenetic model for co-estimating gene and species trees without the need of estimation of divergence times in species trees and duplication and loss rates.

The current computational methods for analyzing gene family data (including gene duplication and loss) suffer a variety of weakness that need to be addressed. There is no probabilistic model embedded in a birth-death phylogenetic modeling framework that can take into account the effects of different evolutionary mechanisms of gene retention when analyzing gene family data. It is desirable to build a stochastic model as a good approximation to the real biological process of gene duplication and loss. Such probabilistic models can both add biological realism to improve the fit of the model to the data as well as enable mechanistic inference that is currently not possible. In this study, we integrate several evolutionary mechanisms of gene retention into the age-dependent BD model [42–45], in which the loss rate is a function of the ages of gene copies. Moreover, we derive the probability density function of gene duplication times for each mechanism. The conditional density

Zhao et al. BMC Evolutionary Biology (2015) 15:275

Page 3 of 11

function of a duplication time given the previous duplication time is derived from the reconstructed process under the generalized birth-death model [35, 52]. The conditional density function can be utilized to calculate the joint density of duplication times, and to efficiently simulate duplication times under the generalized BD model. The simulation results suggest that the maximum likelihood approach can accurately estimate the parameters in the generalized BD model for different mechanisms of gene retention, and the proposed gene-retention model can be used to detect the underlying mechanism that drives the evolutionary process of duplicates within a gene family.

Methods

Modeling the loss rate

For simplicity, we consider the process of gene duplication/loss in a single population. For a single population, we assume that a gene copy may duplicate or die at time t. The homogeneous birth-death model assumes that the rate of loss (hazard) of a duplicated gene is constant through time [11, 53]. This expectation is consistent with the nonfunctionalization process, but does not take into account any of the processes of neofunctionalization and subfunctionalization, which can affect the loss rate of gene duplicates. The birth-death model for the fates of gene retention (nonfunctionalization, subfunctionalization, neofunctionalization, and dosage balance) includes a time-dependent loss rate and a constant duplication rate λ. The time-dependent loss rates will be extended to age-dependent loss rates in the age-dependent birth-death model (see section 2.3). The process starts at time 0, and the number of gene copies at time 0 is 2. The process of gene duplication and loss occurs under the following postulates [54]: (1) the probability that a duplication will occur during an infinitesimal interval (t, t + Δt] is $n_t \lambda \Delta t$ + o(Δt), while the probability that no duplication will occur is 1- $n_t \lambda \Delta t$ + o(Δt), and (2) the probability that a gene duplicate will be lost during an infinitesimal interval (t, t + Δt] is $n_t \mu_t \Delta t$ + o(Δt), while the probability that no loss will occur is 1- $n_t \mu_t$ Δt + o(Δt), in which the loss rate μ_t is a function of time t.

We introduce three formulas for the loss rate μ_t based on the processes of nonfunctionalization, neofunctionalization, and subfunctionalization, with assumptions about these processes made in the introduction and also described in [45]. For nonfunctionalization, the loss rate μ_t is constant over time t, i.e., $\mu_t = \mu$. The neofunctionalization hazard rate (instantaneous rate of duplicate copy loss) declines with time [55]. Averaging across the probability of hitting a neofunctionalizing substitution, the nonfunctionalization probability for duplicate genes declines, leading to the overall decline of duplicate loss over long evolutionary time periods [19]. This convexly

declining loss rate has been described with a Weibull hazard function to characterize the average process (the process for a single gene with a known neofunctionalization event would be a discrete jump in the hazard rate) [18]. We use an exponential function to model the loss rate of neofunctionlization, i.e., $\mu_t = \alpha e^{-t\alpha}$ for $0 < \alpha < 1$. Further, the subfunctionalization loss rate behavior has been characterized to be concavely (sigmoidally) declining based upon theoretical expectations of a waiting time for complementary mutations [18, 20]. Konrad [15] introduced an extended exponential hazard function to describe the instantaneous rate of loss. We adopt a generalized logistic function for the loss rate μ_t of subfunctionalization, i.e., $\mu_t = \frac{\alpha e^{\gamma - t}}{1 + e^{\gamma - t}}$, in which the scale parameter $0 < \alpha < 1$ and known location parameter $\gamma > 0$.

The time-dependent birth-death model

We are interested in the probability distribution of duplication times of the reconstructed lineages (the lineages that have survived to the present time), because the phylogeny reconstructed from the sequences of contemporary species does not include the extinct lineages [35]. The pure birth process of the reconstructed lineages can be derived from a generalized birth-death process [34, 36]. We use the following notations which are defined closely to Nee et al. [35] throughout this paper. Let $t_2 = 0$ be the first duplication time at the root of the tree, and T be the present time (we are looking forward in time, i.e., $T > 0$). Let n_T be the number of lineages at the present time T. Let n_i be the number of reconstructed lineages alive at t_i that survive to the present. We use $\{t_i \mid i = 2, \ldots, n_T\}$ to denote the duplication times of n_T lineages at the tips of a phylogenetic tree, and $t_2 < t_3 < t_4 < \ldots < T$. Let P($\tau$, T) be the probability that one lineage at time τ leaves multiple descendants at the present time T, i.e., P(τ, T) = $P(n_T > 0 \mid n_\tau = 1)$ [34–36, 44],

$$P(\tau, T) = \left[1 + \int_\tau^T \mu_t e^{\rho(\tau, t)} dt \right]^{-1}. \qquad (1)$$

In Eq. (1), $\rho(\tau, T) = \int_\tau^T (\mu_s - \lambda) ds$. Since the integral $\int_\tau^T \mu_t e^{\rho(\tau, t)}$ in Eq. (1) is analytically intractable, it is approximated by a Monte Carlo method. We define u_{ij} as the probability $P(n_j > 1 \mid n_i = 1)$ that one lineage at time t_i leaves multiple descendant reconstructed lineages at a later time t_j. This probability has been derived under the time-dependent BD model, i.e., $u_{ij} = P(n_j > 1 | n_i = 1) = 1 - P(t_i, t_j) e^{\rho(t_i, t_j)}$ (see Eq. (8) in [45]). Given the number n_T of lineages at the present time T and the number n_0 of lineages at time 0, the probability

Zhao *et al. BMC Evolutionary Biology* (2015) 15:275

Page 4 of 11

density function of the duplication times $t = \{t_i \mid i = n_0 + 1, ..., n_T\}$ is given by [45]

$$f(t|n_T, n_0, T) = \frac{\prod_{i=n_0+1}^{n_T} (i-1)\lambda P(t_i, T)\left(1-\eta_{t_{i-1}, t_i}\right)^{i-1}}{\binom{n_T-1}{n_0-1}\left(1-\eta_{0,T}\right)^{n_0}\eta_{0,T}^{n_T-n_0}}.$$

(2)

In (2), $\eta_{ij} = 1 - \frac{1-u_{iT}}{1-u_{jT}}$. The conditional probability distribution of duplication time t_i $(i > 2)$, given its previous duplication time t_{i-1}, T and n_T, is given by [45]

$$f(t_i|t_{i-1}, n_T, T) = \frac{f(t_i|t_{i-1})P(n_T|n_{t_i}, T)}{P(n_T|n_{t_{i-1}}, T)}.$$

(3)

In Eq. (3), $f(t_i|t_{i-1}) = (i-1)\lambda P(t_i, T)\left(1-\eta_{t_{i-1}, t_i}\right)^{i-1}$ (see Eq. (19) and (23) in [45]). With the conditional densities $f(t_i|t_{i-1}, n_T, T)$ of duplication times, the duplication events between times 0 and T can be simulated recursively in forward direction. The conditional density in (3) differs from the density of duplication times derived by Hohna [44], in which the duplication events are treated as a random sample from a common probability distribution.

The age-dependent birth-death model
The time-dependent birth-death model described in the previous section starts with a single population corresponding to the lineage of a phylogenetic tree and assumes a molecular clock that starts ticking for all duplicates at the root. Thus, in the time-dependent birth-death model, the loss rate μ_t of a gene copy is a function of time t. However, the loss rate μ_t should be a function of the ages of gene copies. In this section, the time-dependent birth-death process is extended to the age-dependent process, where the clock for each duplicate starts ticking at its birth. When the loss rate is constant (i.e., nonfuncitonalization), the age-dependent model is identical with the time-dependent model. Thus, we only describe the age-dependent model for neofunctionalizaiton and subfunctionalization. In the age-dependent model, the expressions for the loss rates of neofunctionalization and subfunctionalization remain unchanged (see section Modeling the loss rate), except that time t is replaced with the age t' of the gene copy, i.e., $\mu_{t'} = \alpha e^{-t'\alpha}$ for neofunctionalization and $\mu_{t'} = \frac{\alpha e^{\gamma - t'}}{1 + e^{\gamma - t'}}$ for subfunctionalization. Moreover, it is assumed that the number of gene copies increases or decreases by 1 or

remains the same during an infinitesimal interval $(t, t + \Delta t]$ with probabilities described in (4a-c)

$$P(n_{t+\Delta t} = n_t + 1) = n_t\lambda\Delta t + o(\Delta t)$$

(4a)

$$P(n_{t+\Delta t} = n_t - 1) = \sum_{i=1}^{n_t} \mu_{t_i'}\Delta t + o(\Delta t)$$

(4b)

$$P(n_{t+\Delta t} = n_t) = 1 - \left(n_t\lambda + \sum_{i=1}^{n_t}\mu_{t_i'}\right)\Delta t + o(\Delta t)$$

(4c)

In (4b), $\mu_{t_i'}$ is the loss rate of gene copy i at the age of t_i' for $i = 1, 2, ..., n_t$. Let t_i^0 be the duplication time of gene copy i. The age t_i' of gene copy i is a random variable, because it is a function of the random duplication time t_i^0, i.e., $t_i' = t - t_i^0$. Therefore, (4b) and (4c) are integrated over all possible values of $\mu_{t_i'}$ with respect to the probability density function $f(t')$ of the age t' of a gene copy. The age-dependent loss rate $\mu_{t_i'}$ in (4b) and (4c) is replaced with its expectation $E\left(\mu_{t_i'}\right)$. Since all t_i' s have the same probability distribution, the loss rates of n_t gene copies have the same expected values. Let t^0 be the most recent duplication time of a gene copy that survives to time t. Since t^0 is the most recent duplication time, it indicates that no duplication or loss events have occurred between t^0 and t on the gene copy. It has been shown that the number of duplication or loss events follows the Poisson distribution with mean $\int_0^t (\lambda + \mu_x)dx$. The probability of no duplication or loss events occurring within the time interval $[0, t]$ is equal to $e^{-\int_0^t (\lambda + \mu_x)dx}$. Thus, the probability density of duplication time t^0 is proportional to $D_{t^0}e^{-\int_0^t (\lambda + \mu_x)dx}$ for $0 < t^0 < t$, in which D_{t^0} is the duplication rate at time t^0 and $e^{-\int_0^t (\lambda + \mu_x)dx}$ is the probability that t^0 is the most recent duplication time of the gene copy. Given that duplication occurs on a specific lineage, D_{t^0} is equal to the duplication rate λ. Thus, the probably density of the most recent duplication time t^0 is

$$f(t^0) = \frac{e^{-\int_{t^0}^t (\lambda + \mu_x)dx}}{\int_0^t \left(e^{-\int_{t^0}^t (\lambda + \mu_x)dx}\right)dt^0}$$

(5)

Zhao *et al. BMC Evolutionary Biology* (2015) 15:275

Page 5 of 11

Because the gene age t' is equal to $t - t^0$, the probability density of age t' for $0 < t^0 < t$ is given by

$$f(t') = \frac{e^{-\int_{t-t'}^{t}(\lambda + \mu_x)dx}}{\int_0^t \left(e^{-\int_{t-t'}^{t}(\lambda + \mu_x)dx} \right) dt'} \qquad (6)$$

Since the denominator in (6) is intractable, it is approximated by Monte Carlo simulation. It follows that the mean loss rate at time t is $\phi_t = E\left(\mu_{t'_i} \right) = \int_0^t \mu_{t'} f(t')dt'$. Thus, the postulates in (4b) and (4c) become $P(n_{t+\Delta t} = n_t - 1) = n\phi_t \Delta t + o(\Delta t)$ and $P(n_{t+\Delta t} = n_t) = 1 - n_t(\lambda + \phi_t)\Delta t + o(\Delta t)$. The loss rate in Eq. (1) is replaced by the mean loss rate ϕ_t accordingly and $P(\tau, T)$ is modified as

$$P(\tau, T) = \left[1 + \int_\tau^T \phi_t e^{\rho(\tau, t)} dt \right]^{-1} \qquad (7)$$

Finally, the joint and conditional probability density of duplication times (in Eq. 2–3) for the age dependent model remain unchanged, except that the loss rate μ_t in Eq. (2–3) is replaced with the mean loss rate ϕ_t.

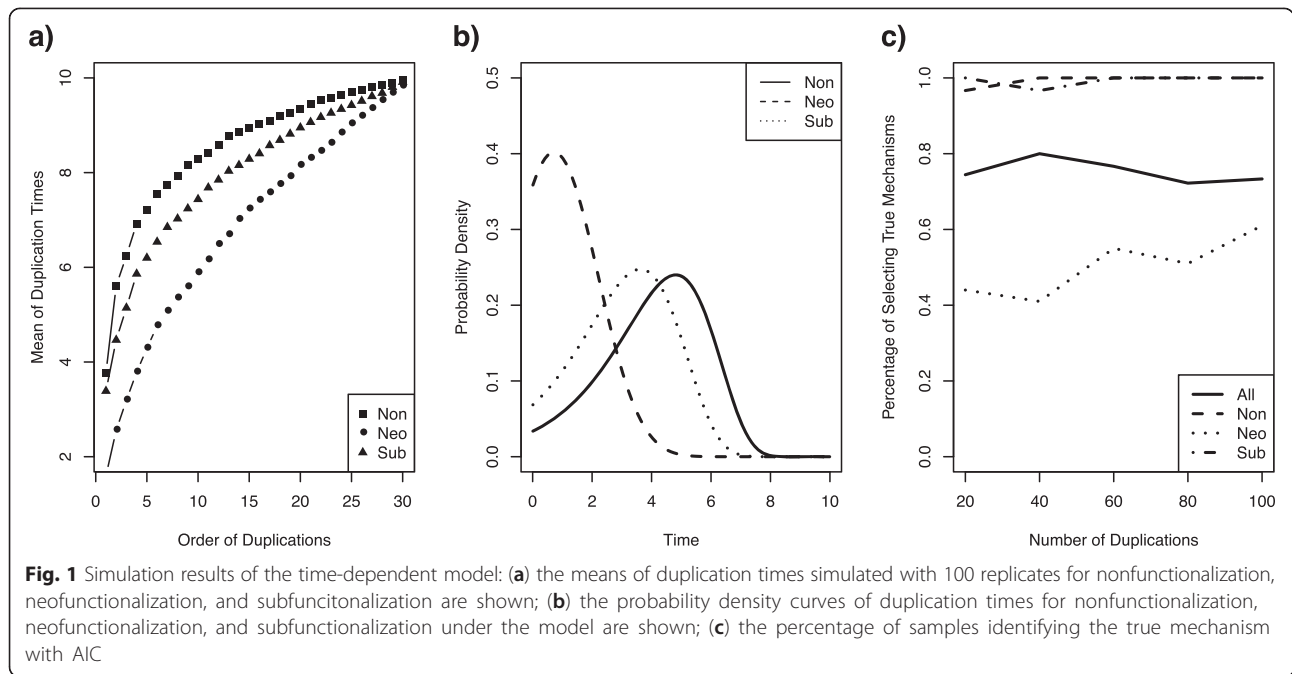## Results

Simulation for the time-dependent model

To evaluate the performance of the time-dependent birth-death model on simulated data where the true values of parameters are known, we generated duplication times of gene copies using the rejection-sampling algorithm with the conditional probability density function of duplication times in Eq. (3). We found the maximum likelihood score for the conditional probability distribution using an optimization function *optim* in R. The maximum score was used as the upper bound in the rejection-sampling algorithm. Specifically, duplication times were simulated from Eq. (3) with a fixed current time $T = 10$ and a fixed number of gene copies $n_T = 32$ at time T. The first duplication time is set to 0, i.e., $t_2 = 0$; the second one is simulated conditional on the first one and so on so that additional 30 duplication times are generated sequentially. Duplication events were generated under each of 3 duplication mechanisms (nonfunctionalization, neofunctionalization, and subfunctionalization) with different parameterizations specified in Table 1. We set a constant duplication rate $\lambda = 0.2$ for all simulations (Table 1). The loss rates were determined by the equations described previously for nonfunctionalization, neofunctionalization, and subfunctionalization models

Table 1 The values of parameters used in simulating duplication times under nonfunctionalization, neofunctionalization, and subfunctionalization are shown

	λ	μ	α
Nonfunctionalization	0.2	0.8	
Neofunctionalization	0.2		0.8
Subfunctionalization	0.2		0.8

with parameters shown in Table 1. The values of parameters were selected such that three mechanisms have the same initial deletion rate.

For each mechanism, simulation was repeated 100 times. The mean of simulated duplication times for each of three mechanisms are shown in Fig. 1a. Duplication times simulated under different mechanisms show distinct patterns. Given the present time T and the number of gene copies n_T, the overall duplication times for nonfunctionalization tend to be larger than those for neofunctionalization and subfunctionalization, and duplication times for neofunctionalization appear to be smaller than subfunctionalization. The curves of duplication times for nonfunctionalization, neofunctionalization, and subfunctionalization are well separated (Fig. 1a), even though three mechanisms have the same duplication rate and the same starting deletion rate. These results indicate that duplication times can be used to distinguish different mechanisms of gene retention, and to make inference about the underlying mechanism that generated the observed duplication times given the assumptions of the duplication models and their relationship to the underlying biology. These results are consistent with the caveat that the time-dependent process uses a tree-dependent clock rather than the more biological situation of a duplication-event specific process. The extension to the age-dependent birth-death model will be discussed below. The joint probability density function in Eq. (2) can be used to obtain the maximum likelihood estimates (MLE) of parameters in the time-dependent model, when duplication times are given as input data. To visualize the divergence of the probability density functions of three mechanisms, we plotted the density curves of the first duplication time for nonfunctionalization, neofunctionalization, and subfunctionalization (Fig. 1b) with the values of parameters in Table 1. Since each mechanism has a unique density curve, this result indicates that it is possible to distinguish three mechanisms using the time-dependent birth-death model. Moreover, we employed the Akaike Information Criterion (AIC) [56] to evaluate the relative quality of the time-dependent models for nonfunctionalization, neofunctionalization, and subfunctionalization. The data sets simulated from the time-dependent model were used as input data to calculate AIC for nonfunctionalization,
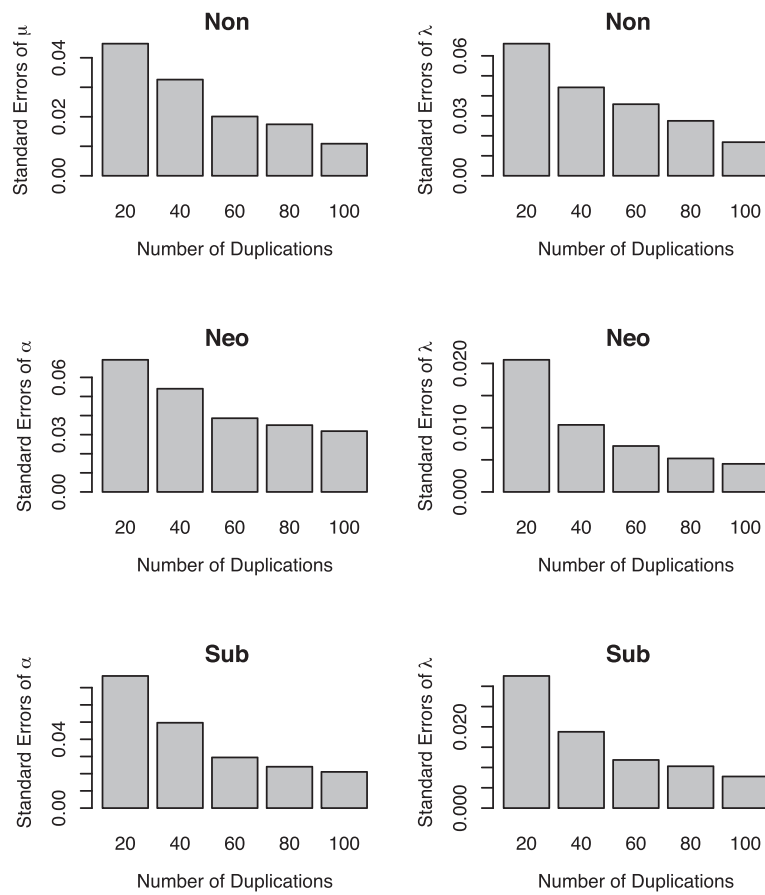
Zhao *et al. BMC Evolutionary Biology* (2015) 15:275

Page 6 of 11

Fig. 1 Simulation results of the time-dependent model: (**a**) the means of duplication times simulated with 100 replicates for nonfunctionalization, neofunctionalization, and subfuncitonalization are shown; (**b**) the probability density curves of duplication times for nonfunctionalization, neofunctionalization, and subfunctionalization under the model are shown; (**c**) the percentage of samples identifying the true mechanism with AIC

neofunctionalization, and subfunctionalization. For each simulated data set, the mechanism with the lowest AIC score was selected and compared with the true mechanism from which the data sets were generated. We reported the percentage of the simulated data sets successfully identifying the true mechanism (Fig. 1c). The overall average of the percentages of samples recovering the true mechanism is about 80 % (Fig. 1c). In addition, subfunctionalization appears to be more difficult than neofunctionalization to distinguish from nonfunctionalization in this modeling framework (Fig. 1c).

To examine the performance of maximum likelihood estimation, we use the simulated duplication times as data to estimate model parameters. The sample size (the number of duplication times) ranges from 20 to 100. The maximum likelihood estimates of parameters were obtained using Metropolis-Hastings Markov Chain Monte Carlo algorithm. The standard errors of the maximum likelihood estimates are displayed in Fig. 2. For nonfunctionalization, the standard errors of the estimates of μ and λ decrease as the number of duplication times increases from 20 to 100. Similarly, the standard errors of the estimates of parameters for subfunctionalization and neofunctionalization decrease as the number of duplications grows. However, the estimation of parameter α for neofunctionalization does not improve well with the increased number of gene copies (Fig. 2), because duplication times in the simulated data are highly correlated and the autocorrelation between two adjacent duplication times increases as the number of duplication times increases. As a result, when the number of highly correlated duplication times reaches a certain number, adding even more

duplication times does not contribute more information for accurately estimating model parameters, especially for neofunctionalization where the loss rate quickly declines to a very low level. Similar results about biases and parameter estimates under constant and time-dependent birth-death processes have been obtained in [57]. Nevertheless, these results suggest that maximum likelihood methods can accurately estimate parameters in the time-dependent birth-death model when the sample size is large.

Simulation for the age-dependent birth-death model

The simulation for the age-dependent model was conducted with the same parameterization and simulation procedure used for the time-dependent model. We generated duplication times from the age-dependent models for nonfunctionalization, neofunctionalization, and subfunctionalization. The mean duplication times given the current time T and gene copy number n_T for the age-dependent models (Fig. 3a) appear to be less dispersed among nonfunctionalization, neofunctionalization, and subfunctionalization than those for the time-dependent models (Fig. 1a). In addition, the density curve for subfunctionalization becomes closer to the nonfunctionalization curve under the age-dependent model (Fig. 3b), compared to the curves for the time-dependent model (Fig. 1b). This is consistent with our expectation, because the age of a gene copy is less than the absolute time t and the beginning portion of the concavely declining loss rate of subfunctionalization is similar to the constant rate of nonfunctionalization. In Fig. 3b, the density curve for neofunctionalization is well separated from the density curves for nonfunctionalization

Zhao *et al. BMC Evolutionary Biology* (2015) 15:275

Page 7 of 11

Fig. 2 The standard errors of the maximum likelihood estimates of parameters in the age-dependent models for nonfunctionalization, neofunctionliazation, and subfunctionalization
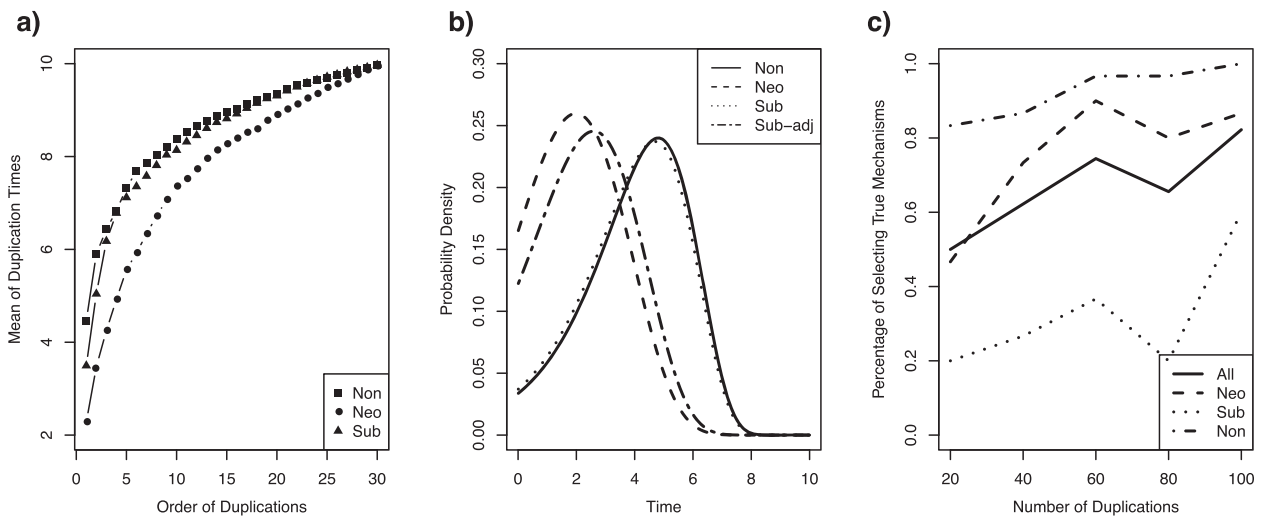
Fig. 3 Simulation results of the age-dependent model: (**a**) the means of duplication times simulated with 30 replicates for nonfunctionalization, neofunctionalization, and subfuncitonalization are shown; (**b**) the probability density curves of duplication times for nonfunctionalization, neofunctionalization, and subfunctionalization under the model are shown; (**c**) the percentage of samples identifying the true mechanism with AIC

Zhao *et al. BMC Evolutionary Biology* (2015) 15:275

Page 8 of 11

and subfunctionalization. In contrast, the loss rate of sub-functionalization is assumed to be a backwards-S-shaped logistic function of time, which is an intermediate state between the loss rates of nonfunctionalization and neo-functionalization. If the loss curve of subfunctionalization moves to the right, it becomes closer to nonfunctionaliza-tion (Fig. 3b). Conversely, when the loss rate curve moves to the left, it gets closer to neofunctionalization (Fig. 3d). Although subfunctionalization is an intermediate state be-tween nonfunctionalization and neofunctionalization, it is expected to be more similar to neofunctionalization, which can be tested in real data analysis. The ultimate similarity comes with increasing time, as both neofunctio-nalization and subfunctionalization culminate in reduced hazard rates, unlike nonfunctionalization. With a fixed duplication rate, these processes are expected to result in an increased number of copies. Conditional on the num-ber of copies, subfunctionalization and neofunctionaliza-tion would be consistent with a reduced duplication rate and older duplication times. The overall percentage of samples identifying the true mechanism increases as the number of gene copies grows (Fig. 3c). The percent-ages of nonfunctionalization and neofunctionalization are significantly higher than the overall percentage. Although the performance of subfunctionalization is below average, the percentage of samples successfully identifying the true subfunctionalization increases to 60 % when the number of gene copies reaches 100. Moreover, the standard errors of the estimates of parameters in the age-dependent model appear to de-crease as the number of gene copies grows, suggest-ing that maximum likelihood methods can accurately estimate parameters in the age-dependent model, when the sample size is large (Fig. 4).

Discussion
Summary of the gene family evolution model
We have derived the probability density function for the age-dependent birth-death model, in which the loss rate is a function of the ages of gene copies. In addition, the conditional density function and a joint density function of duplication times with age-dependent loss rate have been developed in above age-dependent model, given the current time T and the number of gene copies at the time T. The conditional density function is used to efficiently simulate duplication times, and the simulation results suggest that maximum likelihood methods can accurately estimate model parameters in both time-dependent and age-dependent models. In addition, the relative quality of various birth-death models was assessed with AIC. Both time-dependent and age-dependent models can distinguish the three mechanisms (nonfunctionalization, neofunctionaliza-tion, and subfunctionalization) with high probabilities

when the sample size is large. These results indicate that the probabilistic models derived from the birth-death process with a time-dependent and age-dependent loss rates are useful for understanding the duplication and loss mechanisms of gene families that evolve over time in a single population with caveats discussed.

Limitations and future study
As duplication times are often not observable, it is desir-able to generalize the current model to DNA sequence data. We are currently working along this line to build a generalized model that includes two stochastic processes. The birth and death process is used to de-rive the probability distribution of a gene family tree, while the mutation process is used to derive the probability distribution of DNA sequence data given the gene family tree. With this generalized model, we can estimate model parameters (duplication and loss rates) from DNA sequence data.

One of the limits of the current model is that it considers gene family evolution in a single population. This model cannot be applied as currently imple-mented to understand the evolutionary process of gene families from multiple species. To overcome this limit, the current model will be extended in the context of species trees, in which duplication process occurs along the lineages of species trees. This generalization will cer-tainly involve intensive computation, but such a model is quite useful for understanding gene family evolution in the context of the evolution of species. Another limitation of the current age-dependent model is that the likelihood is conditioned on observed extant duplicate copies and does not consider the full generative process including du-plicates that were lost before the present. Future work will examine this in the context of Approximate Bayesian Computation [58]. Further, the current model exists in the classes of interspecific models that treat all observations from a single individual from a species as fixed relative to observations from single individuals from other species. Recently, a correction for the effects of population dynamics has been introduced and can be considered in modeling efforts [9]. Missing data and genome as-sembly error are also not specifically addressed in the modeling framework and their impact on inference also needs to be addressed.

The gene loss models and their interpretations (the relationship between the best fit curve shape and the underlying biology) make assumptions about the relationship between the accumulation of synonymous changes and of non-synonymous changes whereas there is information in the evolution of dN/dS vs. dS that can be taken advantage of in alternative formulations of the likelihood (see [18]). Lastly, the models can be used to make predictions about functional evolution in the
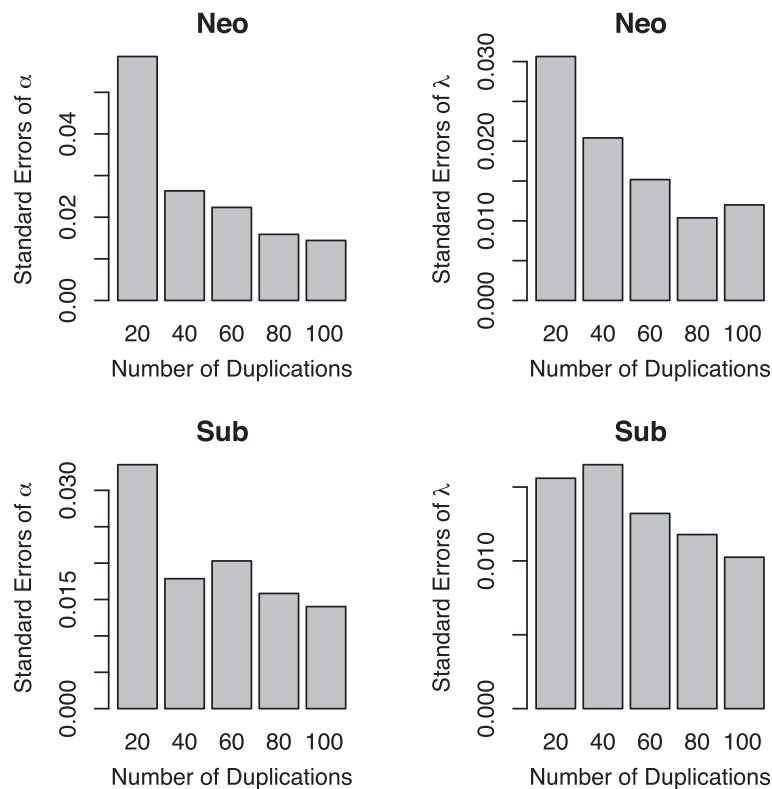
Fig. 4 The standard errors of maximum likelihood estimates of parameters in the age-dependent models for neofunctionliazation and subfunctionalization

absence of actual functional data. While such data does not currently exist in large scale, the future may bring data on the expression levels of protein duplicates compared to an ancestral state as well as binding and enzyme specificities (and enzyme kinetics), all of which can be integrated into a phylogenetic framework. However, even with future comparative proteomic data, one still needs models that treat signals associated with selective pressures (like the models presented here), as neutral changes in expression and functional properties would not lead to changes in retention profiles (the gene loss hazard/survival model) and meaningful lineage-specific biology (see [59] for a discussion of the interplay between molecular phenotypes and biological function in an evolutionary context).

The model as currently developed also assumes that all duplicates in a gene family evolve under the same process. A future opportunity is in examination of large gene family databases like Ensembl [60], HOGENOM [61], or TAED [62], a mixture model of duplicate processes can be applied across all gene families and duplication events to enable a posteriori probabilistic identification of duplication retention mechanisms for individual gene duplication events. The work presented in this manuscript, with a birth-

death model in a phylogenetic context, brings this scale of modeling one step closer.

Conclusions

We develop a generalized birth-death model for probabilistic inference of the evolutionary mechanism for duplicate gene retention based upon the average rate of loss over time of the duplicate. The time-dependent birth-death model assumes a molecular clock that starts ticking for all duplicates at the root. The time-dependent model is then extended to the age-dependent model, which allows the gene loss rate dependent of duplication events. Simulation results indicate that the mechanisms of gene retentions (nonfuncitonalization, neofunctionalization, and subfunctionalization) produce distinct likelihood functions, which can be used with comparative genomic data to quantitatively distinguish those mechanisms.

Availability of supporting data

This study of a theoretical nature has not generated any novel supporting data.

Zhao *et al. BMC Evolutionary Biology* (2015) 15:275

Page 10 of 11

Authors' contributions
The study was conceived by DAL and LL. Mathematical derivations were performed by JZ and LL. Simulations and programming were performed by JZ and AIT. The manuscript was written by JZ, AIT, DAL, and LL. All authors have read and approved the final version of the manuscript.

Author details
[1]Department of Statistics, University of Georgia, 101 Cedar Street, Athens, GA 30602, USA. [2]Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA. [3]Center for Computational Genetics and Genomics and Department of Biology, Temple University, Philadelphia, PA 19122, USA. [4]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA.

References
1. Ohta T. Simulating evolution by gene duplication. Genetics. 1987;115(1):207–13.
2. Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, et al. Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. 2004;2(7):E207.
3. Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 2005;39:121–52.
4. Lynch M, O'Hely M, Walsh B, Force A. The probability of preservation of a newly arisen gene duplicate. Genetics. 2001;159(4):1789–804.
5. Hurles M. Gene duplication: the genomic trade in spare parts. PLoS Biol. 2004;2(7):E206.
6. Ohta T. Role of gene duplication in evolution. Genome. 1989;31(1):304–10.
7. Zhang JZ. Evolution by gene duplication: an update. Trends Ecol Evol. 2003;18(6):292–8.
8. Lynch M. Genomics. Gene duplication and evolution. Science. 2002;297(5583):945–7.
9. Teufel AI, Masel J, Liberles DA. What fraction of duplicates observed in recently sequenced genomes is segregating and destined to fail to fix? Genome Biol Evol. 2015;7(8):2258–64. doi:10.1093/gbe/evv139.
10. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 2005;15(8):1153–60.
11. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.
12. Hughes AL, Friedman R. Gene duplication and the properties of biological networks. J Mol Evol. 2005;61(6):758–64.
13. Liberles DA, Kolesov G, Dittmar K. Understanding gene duplication through biochemistry and population genetics. In: Dittmar K, Liberles DA, Editors. Evolution After Gene Duplication. Hoboken (NJ): Wiley-Blackwell, 2010.
14. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 2010;11(2):97–108.
15. Konrad A, Teufel AI, Grahnen JA, Liberles DA. Toward a general model for the evolutionary dynamics of gene duplicates. Genome Biol Evol. 2011;3:1197–209.
16. Ohno S. Evolution by gene duplication. New York: Springer; 1970.
17. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci U S A. 2012;109(21):E1352–9.
18. Hughes T, Liberles DA. The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. J Mol Evol. 2007;65(5):574–88.
19. Teufel AI, Zhao J, O'Reilly M, Liu L, Liberles DA. On mechanistic modeling of gene content evolution: birth-death models and mechanisms of gene birth and gene retention. Computation. 2014;2:3.
20. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151(4):1531–45.
21. Rastogi S, Liberles DA. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol. 2005;5:28.
22. Khan AA, Janke A, Shimokawa T, Zhang H. Phylogenetic analysis of kindlins suggests subfunctionalization of an ancestral unduplicated kindlin into three paralogs in vertebrates. Evol Bioinform Online. 2011;7:7–19.
23. Akerborg O, Sennblad B, Arvestad L, Lagergren J. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci U S A. 2009;106(14):5714–9.
24. Basten CJ, Ohta T. Simulation study of a multigene family, with special reference to the evolution of compensatory advantageous mutations. Genetics. 1992;132(1):247–52.
25. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. Genetics. 2007;177(3):1941–9.
26. Ohta T. An extension of a model for the evolution of multigene families by unequal crossing over. Genetics. 1979;91(3):591–607.
27. Thornton JW, DeSalle R. Gene family evolution and homology: genomics meets phylogenetics. Annu Rev Genomics Hum Genet. 2000;1:41–73.
28. Yanai I, Camacho CJ, DeLisi C. Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys Rev Lett. 2000;85(12):2641–4.
29. Karev GP, Wolf YI, Berezovskaya FS, Koonin EV. Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. BMC Evol Biol. 2004;4:32.
30. Bailey N. The elements of stochastic processes. New York: Wiley; 1964.
31. Huynen MA, van Nimwegen E. The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol. 1998;15(5):583–9.
32. Csuros M, Miklos I. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol. 2009;26(9):2087–95.
33. Szollosi GJ, Tannier E, Daubin V, Boussau B. The inference of gene trees with species trees. Syst Biol. 2015;64(1):e42–62.
34. Thompson. The likelihood approach. In: Human evolutionary trees. 1975.
35. Nee S, May RM, Harvey PH. The reconstructed evolutionary process. Philos Trans R Soc Lond B Biol Sci. 1994;344(1309):305–11.
36. Kendall DG. On the generalized birth-and-death process. Ann Math Stat. 1948;19(1):1–15.
37. Rannala B, Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol. 1996;43(3):304–11.
38. Aldous D, Popovic L. A critical branching process model for biodiversity. Adv Appl Probab. 2005;37(4):1094–115.
39. Gernhard T. The conditioned reconstructed process. J Theor Biol. 2008;253(4):769–78.
40. Gernhard T. New analytic results for speciation times in neutral models. Bull Math Biol. 2008;70(4):1082–97.
41. Stadler T. Sampling-through-time in birth-death trees. J Theor Biol. 2010;267(3):396–404.
42. Rabosky DL. Likelihood methods for detecting temporal shifts in diversification rates. Evolution. 2006;60(6):1152–64.
43. Morlon H, Parsons TL, Plotkin JB. Reconciling molecular phylogenies with the fossil record. Proc Natl Acad Sci U S A. 2011;108(39):16327–32.
44. Hohna S. Fast simulation of reconstructed phylogenies under global time-dependent birth-death processes. Bioinformatics. 2013;29(11):1367–74.
45. Hallinan N. The generalized time variable reconstructed birth-death process. J Theor Biol. 2012;300:265–76.
46. Hohna S. The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. J Theor Biol. 2015;380:321–31.
47. Arvestad L, Berglund AC, Lagergren J, Sennblad B. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics. 2003;19 Suppl 1:i7–15.
48. Arvestad L, Lagergren J, Sennblad B. The gene evolution model and computing its associated probabilities. J ACM. 2009;56(2):1–44.
49. Rasmussen MD, Kellis M. A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol. 2011;28(1):273–90.
50. Sjostrand J, Sennblad B, Arvestad L, Lagergren J. DLRS: gene tree evolution in light of a species tree. Bioinformatics. 2012;28(22):2994–5.
51. Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. Genome-scale coestimation of species and gene trees. Genome Res. 2013;23(2):323–30.
52. Liu L, Yu L, Kalavacharla V, Liu Z. A Bayesian model for gene family evolution. BMC Bioinformatics. 2011;12:426.
53. Cotton JA, Page RD. Rates and patterns of gene duplication and loss in the human genome. Proc Biol Sci. 2005;272(1560):277–83.
54. Feller W. An introduction to probability theory and its applications. New York: Wiley; 1954.

Zhao *et al. BMC Evolutionary Biology* (2015) 15:275

Page 11 of 11

55. Zhang P, Min W, Li WH. Different age distribution patterns of human, nematode, and Arabidopsis duplicate genes. Gene. 2004;342(2):263–8.

56. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F, editors. 2nd international symposium on information theory. Budapest: Akadémiai Kiadó; 1973. p. 267–81.

57. Hohna S. Likelihood inference of non-constant diversification rates with incomplete taxon sampling. PLoS One. 2014;9(1):e84184.

58. Janzen T, Höhna S, Etienne RS. Approximate Bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT. Methods Ecol Evol. 2015;6:5.

59. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5(3):578–90.

60. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res. 2012;40(Database issue):D84–90.

61. Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, et al. Databases of homologous gene families for comparative genomics. BMC Bioinformatics. 2009;10 Suppl 6:S3.

62. Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA. The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. Nucleic Acids Res. 2005;33(Database issue):D495–7.