A two-type branching process model of gene family evolution

Arthur Zwaenepoel^{1,2*} Yves Van de Peer^{1,2,3,4}

1. Department of Plant Biotechnology and Bioinformatics, Ghent University, 9054 Ghent, Belgium

2. Center for Plant Systems Biology, VIB, 9054 Ghent, Belgium

3. Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

4. College of Horticulture, Nanjing Agricultural University, Nanjing, 210095, China

* Corresponding author: arzwa@psb.vib-ugent.be

Abstract

Phylogenetic models of gene family evolution based on birth-death processes (BDPs) provide an awkward fit to comparative genomic data sets. A central assumption of these models is the constant per-gene loss rate in any particular family. Because of the possibility of partial functional redundancy among gene family members, gene loss dynamics are however likely to be dependent on the number of genes in a family, and different variations of commonly employed BDP models indeed suggest this is the case. We propose a simple two-type branching process model to better approximate the stochastic evolution of gene families by gene duplication and loss and perform Bayesian statistical inference of model parameters in a phylogenetic context. We evaluate the statistical methods using simulated data sets and apply the model to gene family data for *Drosophila*, yeasts and primates, providing new quantitative insights in the long-term maintenance of duplicated genes.

Keywords: gene duplication and loss, phylogenetics, multi-type branching process, Bayesian inference

Introduction

The expansion and contraction of gene families by gene duplication and loss is one of the key features of genome evolution. Statistical phylogenetic analyses of the evolution of gene family sizes rely on probabilistic models of gene family evolution, which are typically based on birth-death processes (BDPs) (Hahn et al. 2005; Csűrös and Miklós 2009). Phylogenetic BDP models have been applied to analyze genome evolution using comparative genomic

data sets (e.g. Hahn, Han, and Han 2007; Csűrös and Miklós 2009; Thomas et al. 2020), infer reconciled gene trees (Åkerborg et al. 2009; Szöllősi et al. 2015), and infer ancient whole-genome duplications in a phylogeny (Rabier, Ta, and Ané 2014; Zwaenepoel and Van de Peer 2020). Importantly, such models seek to capture the long-term stochastic evolution of gene family sizes by duplication and loss, as opposed to the short-term population genetic processes leading to the establishment and loss of newly arisen copy number variants.

Standard phylogenetic BDP models provide however an awkward fit to comparative genomic data. Our study departs from the observation that estimated gene loss rates can vary dramatically depending on the data and assumed BDP model. For instance, when considering the full array of gene families across a set of eight *Drosophila* species, the standard linear BDP model yields an estimated loss rate of 0.24 expected loss events per gene per 100 My; whereas considering the set of gene families that are retained with at least one copy across all taxa, and assuming these gene families cannot go extinct (i.e. are essential), a loss rate of 3.94 expected loss events per *duplicated* gene per 100 My is obtained (see below for methods). The same discrepancy appears when comparing loss rate estimates from the linear BDP model with crude estimates based on age distributions of duplicated genes (Lynch and Conery 2003).

One of the key assumptions underlying the standard linear BDP model that is generally violated, and which may lead to the discrepancies noted above, is the assumption of a single and constant loss rate per gene in a family. Many gene duplicates, although stably established in the genome, exhibit some functional redundancy (examples abound in the molecular biological literature). When a duplicated gene is fully or partially functionally redundant with its parental gene, it seems likely that both copies are subject to higher gene loss rates until either one gene of the duplicate pair is lost or has adopted a distinct function (neofunctionalization), or both genes underwent ‘complete’ subfunctionalization (hereafter, neo- and subfunctionalization are referred to jointly as ‘x-functionalization’). In other words, when a set of genes is (partially) functionally redundant, we may expect an increased per-gene loss rate μ_r compared to a set of non-redundant genes with per-gene loss rate μ_{nr} , as in the former case there will be weaker purifying selection against pseudogenes (in the case of full redundancy for instance, a null-mutant will be effectively neutral (Walsh 2003)). This however induces non-independence among distinct genes in a family, because when one copy in a functionally redundant gene pair gets lost, the loss rate of the retained copy will drop back to μ_{nr} . This contrasts with the commonly employed BDP models which obey the branching property that distinct genes evolve independently. This non-independence represents a serious obstacle for developing more accurate stochastic models of gene family evolution.

In this study, we propose a two-type continuous-time branching process model of gene family evolution to better approximate the stochastic evolution of gene families, while retaining the independence assumption. In this model we treat a family as consisting of two types of genes: ‘base’ genes (type 1) and ‘excess’ genes (type 2). Any particular family at time t can be represented as a tuple $X(t) = (X_1(t), X_2(t))$ where $X_1(t)$ and $X_2(t)$ represent the number of type 1 and type 2 genes in the family at time t respectively. We can think of such a family as consisting of $Z(t) = X_1(t) + X_2(t)$ genes, coding for $X_1(t)$ functions (note that we define a

gene family as an *orthogroup*, irrespective of gene function, see methods). We allow for four types of events: gene duplication at rate λ per gene, loss of a type 1 gene at rate μ_1 per type 1 gene, loss of a type 2 gene at rate μ_2 per type 2 gene and a conversion of a type 2 to a type 1 gene (x-functionalization) at a rate ν per type 2 gene (see methods).

Importantly, because of the independence assumption, μ_2 cannot be directly interpreted as the rate of gene loss of a functionally redundant gene μ_r . To see this, consider a state, say, $X(t) = (1, 3)$, which we could interpret as a set of four genes coding for a single function. All four genes are functionally redundant in this case, and under the ideal model we would have a total loss rate $4\mu_r$, where μ_r is the loss rate for a functionally redundant gene. Instead under the two-type model, the total rate of gene loss will be $\mu_1 + 3\mu_2$, so that when $\mu_1 \ll \mu_2$ the rate of loss per redundant gene is $\approx 3\mu_2/4$. Under the two-type model the rate of gene loss per *excess* gene μ_2 is constant for different family sizes, while the rate of gene loss per *redundant* gene increases with increasing number of excess genes per base gene, approximately equaling $\mu_2(n - 1)/n$ for a group of n redundant genes. Another example illustrating the differences between the ideal model with non-independence and the two-type model is provided in fig. 1 (a).

In the remainder of this paper we describe the probabilistic model and an approach for Bayesian statistical inference of the relevant model parameters from comparative genomic data. We also present some innovations that apply more generally to phylogenetic BDP models, in particular the usage of a Beta-Geometric prior on the number of genes at the root and a BDP model for families that are assumed not to go extinct. We study the statistical properties of inference under the two-type branching process model using simulated data sets and study gene family evolution in *Drosophila*, yeasts and primates. Finally, we discuss the implications of our results for the long-term evolutionary dynamics of gene content and the evolutionary genetics of gene duplication.

Model and methods

Duplication-loss-functionalization model

We first describe an idealized duplication-loss-functionalization (DLF) model with non-independent evolution of gene copies. Throughout, we define a gene family as an orthogroup: a group of genes across a set of species derived from a common ancestral gene in the last common ancestor of the species under consideration. We assume a gene family of n genes is partitioned in $k \leq n$ functions, and we represent a gene family at time t as the length k array of gene counts in each functional class $X(t) = [X_1(t), X_2(t), \dots, X_k(t)]$, $X_i(t) > 0$, $i \leq k$. An extinct gene family is represented by the empty array. The total gene family size is denoted by $Z(t) = \sum_{i=1}^k X_i(t)$. An example gene family under this model with $Z = 5$ and $k = 3$ is shown in fig. 1 (a.iii). We assume birth-death dynamics, i.e. transitions increment or decrement $Z(t)$ by at most one unit. We assume all genes duplicate at rate λ and that genes in a functional class

i with $X_i > 1$ suffer loss at rate μ_r per gene (loss of a redundant copy). Genes in a single-copy functional class suffer loss at rate μ_{nr} (loss of a non-redundant copy), an event that leads to a decrease in the number of functional classes in the family. Lastly, we assume that for each class i for which $X_i > 1$, genes shift to a new, not yet existing functional class at rate ν per gene, decrementing X_i and increasing the number of functional classes in the family (neo- or subfunctionalization). Note that the latter event does not affect $Z(t)$. While this model does not admit efficient statistical inference (as far as we are aware), it is straightforward to simulate from using standard techniques.

Two-type branching process model

As an approximation to the idealized DLF model, we model the evolution of gene family content along a species tree as a two-type homogeneous continuous-time branching process $X(t) = (X_1(t), X_2(t)) \in \mathbb{N} \times \mathbb{N}$ (where we take $\mathbb{N}$ to include 0). Here $X_1(t)$ and $X_2(t)$ denote the number of type 1 and type 2 genes at time t respectively (fig. 1 a.iv). The stochastic evolution of the process is determined by the rate parameters $\theta = (\lambda, \mu_1, \mu_2, \nu)$ in the following way

$$p_{ij}(k, l, \Delta t) = \Pr\{X(t + \Delta t) = (k, l) | X(t) = (i, j)\}$$

$$= \begin{cases} (i + j)\lambda\Delta t + o(\Delta t) & k + l = i + j + 1 \\ i\mu_1\Delta t + o(\Delta t) & k = i - 1, l = j \\ j\mu_2\Delta t + o(\Delta t) & k = i, l = j - 1 \\ j\nu\Delta t + o(\Delta t) & k = i + 1, l = j - 1 \\ 1 - ((i + j)\lambda + i\mu_1 + j(\mu_2 + \nu))\Delta t + o(\Delta t) & k = i, l = j \\ o(\Delta t) & \text{else} \end{cases} \quad (1)$$

for Δt sufficiently small and $i, j, k, l \geq 0$. The process is a bivariate birth-death process and is Markovian (see fig. 1 (b) for a graph representation of the bivariate Markov chain). Throughout we assume $\{\mu_1, \lambda, \nu\} < \mu_2$. As a model of gene family evolution by gene duplication and loss, $X_1(t)$ denotes the number of ‘base’ genes in the family, while $X_2(t)$ denotes the number of ‘excess’ genes in the family for which the eventual fate (non, neo- or subfunctionalization) is yet to be determined at time t . The model defined above assumes that all genes duplicate at the same per-gene rate λ and that the fate of a type 2 gene is resolved after an exponentially distributed time with mean $\mu_2 + \nu$, with the probability of nonfunctionalization being $\mu_2/(\mu_2 + \nu)$ and the probability of sub- or neofunctionalization $\nu/(\mu_2 + \nu)$ (turning an excess gene into a base gene). We further assume type 1 genes are removed with rate μ_1 . We refer to this model as the ‘two-type duplication-loss (DL) model’.

Figure 1: (a) Schematic representation of a gene family of total size $Z = 5$, consisting of a set of genes coding for three functions, with its associated state X from the perspective of four models considered in this paper. The associated total (i.e. family-level) duplication, loss and x-functionalization rates are indicated below. In all models the per-gene duplication rate is λ . (i) Simple linear BDP model with per-gene loss rate μ ; (ii) Linear BDP without extinction, with loss rate per *excess* gene (in white) μ ; (iii) Duplication-loss-functionalization (DLF) model. The five gene copies are explicitly subdivided in three functional groups (depicted in different colors). Genes in groups consisting of multiple copies (gray and black genes) are subject to loss at rate μ_r per gene, while genes in single-copy groups are subject to loss at rate $\mu_{nr} < \mu_r$. Genes in multi-copy groups give rise to new groups (x-functionalize) at rate ν per gene; (iv) Two-type duplication-loss model with loss rate μ_1 per *base* (type 1, in black) gene, loss rate μ_2 per *excess* (type 2, in gray) gene and x-functionalization rate ν per type 2 gene. (b) Graph representation of the two-type continuous time branching process, where nodes depict states $X = (X_1, X_2)$ and edges transitions occurring with the associated rates. (c) Simplified probabilistic graphical model for the two-type DL model along a three-taxon phylogeny, following the directed factor graph notation of Dietz (2010). BG denotes the Beta-Geometric distribution and $2T(\theta, t_u)$ denotes the transient distribution for the two-type branching process with parameters θ along the branch leading to node u of length t_u . Priors for $\theta = (\lambda, \mu_1, \nu, \mu_2)$ and hyperparameters $\phi = (\eta, \zeta, r)$ are not shown.

Transition probabilities for the two-type model

Probabilistic inference of model parameters from comparative genomic data requires that we can compute transition probabilities under the model. In general, the transient distributions for multi-type branching processes like the above are analytically intractable, presenting a serious obstacle to their fruitful application in a statistical setting. Xu et al. (2015) however presented a numerical approach for computing transition probabilities for Markovian multi-type branching processes based on inversion of the associated probability generating functions (pgfs). Let $f_{ij}(s_1, s_2, t)$ denote the pgf for the two-type DL model in eq. 1 when starting at $X(0) = (i, j)$, i.e.

$$f_{ij}(s_1, s_2, t) = \sum_k \sum_l p_{ij}(k, l, t) s_1^k s_2^l$$

Importantly, the branching property implies the following relationship among the probability generating functions for different $X(0)$

$$f_{ij}(s_1, s_2, t) = f_{10}(s_1, s_2, t)^i f_{01}(s_1, s_2, t)^j$$

(e.g. Athreya and Ney (1972)), so we may work with f_{10} and f_{01} and recover the desired pgfs easily. As we derive in detail in Appendix 1, we can find the following system of ordinary differential equations (ODEs) for the pgfs

$$\begin{aligned} f'_{10} &= \mu_1 - (\lambda + \mu_1)f_{10} + \lambda f_{10}f_{01} \\ f'_{01} &= \mu_2 + \nu f_{10} - (\lambda + \nu + \mu_2)f_{01} + \lambda f_{01}^2 \end{aligned}$$

Where the arguments s_1, s_2 and t are omitted for clarity and differentiation is with respect to t .

No closed form solution for $f_{10}(s_1, s_2, t)$ and $f_{01}(s_1, s_2, t)$ can be obtained for this system, and we solve these ODEs numerically using the Tsit5 solver implemented in DifferentialEquations.jl (Rackauckas and Nie 2017; Tsitouras, Famelis, and Simos 2011). To obtain transition probabilities from the pgfs we use the numerical inversion method of Xu et al. (2015), which involves re-expressing the pgf as a Fourier series $f_{jk}(e^{2\pi i w_1}, e^{2\pi i w_2}, t)$ so that the coefficients corresponding to the transition probabilities are given by the inverse Fourier transform, which can be computed numerically using the fast Fourier transform (FFT) along an $N \times N$ grid. Choice of a higher N should lead to more accurate transition probabilities, and determines the maximum state from and to which we can compute transition probabilities. In practice the numerical error is dominated by the tolerance settings in the ODE solver, so that increasing N beyond relatively small values (e.g. $N = 16$ as in Xu et al. (2015)) does not lead to much gain in accuracy for transition probabilities among states with reasonable probability in our applications (fig. S1).

Count data likelihood along a phylogeny

With a reasonably efficient method for computing the transition probabilities available, likelihood-based statistical inference is possible. Our data set is an $n \times m$ matrix consisting of gene counts for n gene families for a set of m species related by a known phylogeny S . We refer to each row y_i , $i \in (1, \dots, n)$, of this matrix as a *phylogenetic profile*. In our notation, the profiles correspond to the m -tuples $y_i = (Z_i^{(u)})$ with $u \in \{1, \dots, m\}$, where $Z_i^{(u)}$ denotes the total gene count at leaf node u for family i , such that $Z_i^{(u)} = X_{i,1}^{(u)} + X_{i,2}^{(u)}$ (where $X_{i,1}^{(u)}$ denotes the number of type 1 genes at leaf node u for family i).

The model we consider assumes that the two-type branching process operates along the branches of species tree S , such that at each split in the tree, two independent copies of the ancestral process start with the same initial state. We assume phylogenetic profiles are independent and identically distributed (iid) under the two-type branching process model defined along S , and that we have no access to any information on the number of genes of each type at leaf nodes (i.e. we only observe Z , not X). The resulting probabilistic graphical model (PGM) is depicted for a hypothetical three-taxon tree ($m = 3$) in fig. 1. The likelihood of the observed data $p(y|\theta, \phi) = \prod_{i=1}^n p(y_i|\theta, \phi)$ conditional on the parameters of the branching process θ and the prior distribution for the root ϕ can be computed using variable elimination along the PGM (i.e. using Felsenstein’s pruning algorithm, and integrating the marginal likelihood values at the root over a suitable prior distribution on the root state).

Conditioning on non-extinction

To rule out *de novo* gain of genes in arbitrary subtrees of the phylogeny, we filter the data so that at least one gene is present in each clade stemming from the root of the species tree (as in e.g. Rabier, Ta, and Ané (2014) and Zwaenepoel and Van de Peer (2019)). We denote the event of extinction below a node u in the species tree by E_u and let a, b and c respectively label the root and its two daughter nodes. The likelihood of a profile y conditional on the event of non extinction in both clades is then

$$p(y|\theta, \bar{E}_b \cap \bar{E}_c) = \frac{p(y|\theta)}{\Pr(\bar{E}_b \cap \bar{E}_c)}$$

where $\bar{E}$ denotes the negation (or complement) of E . To obtain the relevant data likelihood we therefore need to compute the condition factor $\Pr(\bar{E}_b \cap \bar{E}_c)$. Assuming a suitable prior distribution on the number of lineages of each type at the root of the species tree (see below), we can rely on the conditional independence of subtrees of S and properties of the pgfs to obtain this probability. For ease of notation, denote by $X_{i,j}^{(a)}$ the event that $X^{(a)} = (i, j)$. The condition factor is then given by

$$\begin{aligned}\Pr(\bar{E}_b \cap \bar{E}_c) &= \sum_{k=1}^{\infty} \sum_{l=1}^k (1 - \Pr\{E_b | X_{l,k-l}^{(a)}\}) (1 - \Pr\{E_c | X_{l,k-l}^{(a)}\}) \Pr\{X_{l,k-l}^{(a)}\} \\ &= \sum_{k=1}^{\infty} \sum_{l=1}^k (1 - g_{l,k-l}^{(b)}(\mathbf{0})) (1 - g_{l,k-l}^{(c)}(\mathbf{0})) \Pr\{X_{l,k-l}^{(a)}\}\end{aligned}$$

Here $\Pr\{X_{l,k-l}^{(a)}\}$ is given by the prior distribution on the number of ancestral lineages in a gene family at the root of the species tree, and $g_{i,j}^{(u)}(\mathbf{s}) = g_{10}^{(u)}(\mathbf{s})^i g_{01}^{(u)}(\mathbf{s})^j$ is the joint pgf for the leaf observations in the subtree rooted in node u , conditional on the state at the parent of u , say u' , being (i, j) . Assuming there are m leaves below u and labeling the entries of $\mathbf{s}$ as $(s_{1,1}, s_{1,2}, s_{2,1}, s_{2,2}, \dots, s_{m,1}, s_{m,2})$, where the first index refers to the leaf node in the subtree below u and the second to the gene type, this pgf is written more explicitly as

$$g_{10}^{(u)}(\mathbf{s}) = \sum_{(k_1, l_1)} \sum_{(k_2, l_2)} \dots \sum_{(k_m, l_m)} \Pr\{X_{k_1, l_1}^{(1)}, X_{k_2, l_2}^{(2)} \dots X_{k_m, l_m}^{(m)} | X_{1,0}^{(u')}\} s_{1,1}^{k_1} s_{1,2}^{l_1} s_{2,1}^{k_2} s_{2,2}^{l_2} \dots s_{m,1}^{k_m} s_{m,2}^{l_m}$$

and of course analogously for $g_{01}^{(u)}$. By the branching property and conditional independence of disjoint subtrees, this pgf can be evaluated efficiently using a postorder traversal and the single branch pgf f_{10} . Specifically, consider a node u , a distance t_u from its parent, with (if it is not a leaf) child nodes v and w . We have the following recursive relation:

$$\begin{aligned}g_{10}^{(u)}(\mathbf{s}) &= f_{10}[h_{10}^{(u)}(\mathbf{s}), h_{01}^{(u)}(\mathbf{s}), t_u] \\ h_{10}^{(u)}(\mathbf{s}) &= \begin{cases} s_{u,1} & \text{if } u \text{ is a leaf} \\ g_{10}^{(v)}(\mathbf{s}) g_{10}^{(w)}(\mathbf{s}) & \text{else} \end{cases}\end{aligned}$$

with analogous recursions holding for $g_{01}^{(u)}$ and $h_{01}^{(u)}$.

Prior distribution on the number of lineages at the root in a gene family

Plausible inferences for the proposed model do not only depend on the stochastic process describing the transient evolution of gene family counts, but also on the assumptions of the ancestral gene content for a family. Since neither the single-type linear BDP nor the two-type DL model have a stationary distribution on $\mathbb{N}$, the model itself does not directly suggest a natural prior for the number of ancestral lineages. Conditional on non-extinction however, the linear birth-death process has a Geometric(η) stationary distribution $\Pr\{X = k\} = \eta(1 - \eta)^{k-1}$, and a Geometric prior has been considered previously in this context (Rabier, Ta, and Ané 2014; Zwaenepoel and Van de Peer 2019).

Figure 2: Top row: Geometric and Beta-Geometric stationary distributions fitted to observed gene family size distributions for three data sets. The black dots show the observed size distribution (across all taxa), whereas the lines show the mean frequencies and 95% posterior predictive intervals based on 1000 simulations from the posterior predictive distribution for both models. Bottom row: Phylogenies for the relevant data sets, refer to table S1 for complete taxon names.

The size distribution of gene families is however universally overdispersed with respect to the geometric distribution, showing an approximate power-law tail (Huynen and Van Nimwegen 1998). As we plan to study in detail elsewhere, the major cause of the apparent power-law frequency distribution is likely heterogeneity in gene duplication and loss rates across families, a hypothesis also investigated by Hughes and Liberles (2008). A simple model showing this behavior, corresponding to a special case of the two-type process with $\mu_1 = \nu = 0$, is derived in Appendix 2. The resulting stationary distribution is a Beta-Geometric BG(α, β) distribution, which has probability mass function (pmf)

$$\Pr\{Z = k\} = \frac{B(\alpha + 1, \beta + k - 1)}{B(\alpha, \beta)}, \quad k = 1, 2, \dots$$

where $B(\alpha, \beta)$ denotes the Beta function $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$. This pmf shows approximate power-law behavior in its tail when α and β are fairly small. The BG distribution provides an excellent fit to the observed size distribution for non-zero gene counts (fig. 2). Note furthermore that the mean of the underlying beta distribution (see Appendix 2) may provide a crude idea of the ratio λ/μ_2 in the two-type model.

The BG distribution is a reasonable prior distribution for the *total* number of genes $Z^{(a)}$ at the root node a of the species tree, however for inference purposes a prior for $X^{(a)}$ is required. Here

we assume that there is at least one type 1 gene in each family, and that among the $Z^{(a)} - 1$ remaining genes, each gene is of type 1 with probability r and type 2 with probability $1 - r$, so that the number of type 2 genes is a $\text{Binomial}(Z^{(a)} - 1, 1 - r)$ random variable. We use the resulting distribution for $X^{(a)}$ in our analyses for the two-type DL model (fig. 1) and refer to it as the BG-Binomial prior. Importantly, while we motivate the usage of the BG distribution using the fact that the observed size distribution is overdispersed with respect to the geometric distribution (and that this is likely caused by rate heterogeneity across families, see Appendix 2), we do ignore this rate heterogeneity in the transient regime for reasons of computational efficiency (but see fig. 5 and associated discussion).

Bayesian inference and posterior predictive simulation

With an algorithm for the likelihood $p(y|\theta, \phi)$ available we can perform Bayesian inference for the two-type DL model using Markov Chain Monte Carlo (MCMC) given suitable prior distributions. Unless stated otherwise, we adopt the following priors in our analyses:

$$\begin{aligned}\mu_2 &\sim \text{Exponential}(5) \\ a_1, a_2, a_3 &\sim \text{Beta}(1, 1) \quad (\text{iid}) \\ (\lambda, \mu_1, \nu) &= (a_1\mu_2, a_2\mu_2, a_3\mu_2) \\ r &\sim \text{Beta}(1, 1)\end{aligned}$$

Assuming $\{\lambda, \mu_1, \nu\} < \mu_2$. Unless stated otherwise, we fix η and ζ to the posterior mean values obtained from the fit of the stationary distribution of the single-type model without extinction to the relevant data (see fig. 2). We implemented a simple adaptive Metropolis-within-Gibbs (MWG) algorithm (Roberts and Rosenthal 2009) to sample from the posterior distribution. We make extensive use of posterior predictive simulations to assess model fit. Throughout, posterior predictive simulations are conducted by sampling 1000 random parameter vectors from the joint posterior distribution. For each sample we simulate a data set of equal size as the data set for which inference was performed, and for each such simulation replicate we compute the size distribution for each lineage and across all lineages.

Single-type BDP models

We compare analyses under the two-type DL model with analyses for three distinct single-type BDP models. First, we consider a critical linear BDP model where genes duplicate and get lost at a single ‘turnover’ rate λ . As a second model, we consider the default linear BDP with duplication and loss rates λ and μ per gene respectively. Lastly, we consider the single-type linear BDP model without extinction (see Appendix 2), where genes duplicate at rate λ , and the loss rate per *excess* gene is μ . The latter model is equivalent to a linear birth-death and

immigration process (or duplication, loss and gain model in Csűrös and Miklós (2009)) on the number of excess genes (i.e. $Z - 1$ instead of the family size Z) with immigration (gain) rate $\kappa = \lambda$. We note that a discrete-time deterministic analog of this model was used by Lynch and Conery (2003) to model age distributions of duplicated genes. An illustration of how these various models differ in terms of family-level duplication and loss rates from each other, the DLF model and the two-type DL model is provided in fig. 1 (a). The single-type model without extinction can further be seen as a special case of the two-type DL model where $\mu_1 = \nu = 0$. Inference for single-type models is performed using the DeadBird.jl library (which builds on Zwaenepoel and Van de Peer 2020), which employs the algorithm for computing the phylogenetic BDP likelihood devised by Csűrös and Miklós (2009). We further make use of the Turing.jl library (Ge, Xu, and Ghahramani 2018) using the NUTS algorithm (Hoffman and Gelman 2014) for Bayesian inference. We use the same BG prior as for the two-type model as a distribution on the number of lineages at the root (see Appendix 3) and use non-informative (flat) priors for the rate parameters unless stated otherwise. Posterior predictive simulations were performed as for the two-type DL model.

Data sets, implementation and availability

We consider three data sets, consisting of (1) eight *Drosophila* species (a subset from the widely studied 12 *Drosophila* data set (Hahn, Han, and Han 2007; Drosophila 12 Genomes Consortium 2007)), (2) eight yeast species (the set of species included in the yeast gene order browser (YGOB, v7-Aug2012; Byrne and Wolfe (2005)) that did not undergo the *S. cerevisiae* genome duplication) and (3) 11 primate species. Gene families for each data set were obtained using OrthoFinder (v2.4) (Emms and Kelly 2019), using the sequence data listed in table S1 (when the data included isoforms, we retained the longest isoform as a representative for each gene). For the sake of computational feasibility, we focus for the primates phylogeny on a subset of the data by identifying those families that contain a human homolog annotated with the gene ontology term GO:0002376 (immune system process) or any of its descendant terms. The data was filtered such that there was at least one representative for each family in both clades stemming from the root. For computational feasibility, we further excluded families with more than 10 copies in any of the included taxa, so that in the end we obtain a data set containing 12,163, 4855 and 1921 gene families for the *Drosophila*, yeast and primates phylogenies respectively. For the *Drosophila* and primates data sets, a dated phylogeny was downloaded from the TimeTree database (Kumar et al. 2017) (fig. 2). For the yeast data set, we inferred a dated time tree using r8s (Sanderson 2003) under the molecular clock assumption, using the phylogeny inferred by OrthoFinder and a root calibration of 112 My in line with Beimforde et al. (2014). All methods are implemented in the Julia programming language (Bezanson et al. 2017) and the associated packages are freely available at <https://github.com/arzwa/TwoTypeDLModel> and <https://github.com/arzwa/DeadBird.jl>. The data sets and scripts for conducting the analyses performed in this paper are also available from the former repository.

Results

Parameter estimation for simulated data under the two-type DL model

We first assess our ability to recover true parameter values of the two-type DL model for simulated data from the same model. In particular, the fact that in real empirical data sets we can only observe the total gene count $Z^{(u)}$ for each family at each leaf node u (and not the counts for each type $X^{(u)} = (X_1^{(u)}, X_2^{(u)})$) is a potential source of identifiability issues. Indeed, it is clear that for a single branch, it would be impossible to identify model parameters of the two-type DL model based on nothing more than a single observation of Z for each family. By considering gene family counts along a phylogeny, a single family provides multiple correlated observations of the evolutionary process, providing information about gene content in ancestral branches of the tree. Observations of Z along the leaves of a phylogeny should therefore also provide information about ancestral gene content at the type level (X_1, X_2) .

Simulations of data sets along the *Drosophila* phylogeny consisting of 1000 gene families across a range of randomly drawn parameter values indicate that overall, parameters can be estimated accurately even for relatively small data sets (fig. S2). As expected, using the completely observed two-type data does lead to a considerably lower variance in the marginal posterior distribution for ν compared to the collapsed data. It may be that the marginal posterior mean for ν slightly overestimates the true value when using the incompletely observed data, however all uncertainty intervals contain the true value, so this risk seems to be minor. Also noticeable, but less dramatic is the difference in posterior variance for μ_2 and r (the probability of excess genes at the root to be of type 1). In addition, we simulated a data set of 10,000 gene families along the eight-taxon *Drosophila* phylogeny using model parameters that seemed reasonable based on exploratory analyses of subsets of the actual *Drosophila* data set ($\lambda = 0.2, \mu_1 = 0.1, \nu = 0.2, \mu_2 = 5, \eta = 0.95, \zeta = 4$ and $r = 0.5$). We obtain similar results as for our smaller simulations (fig. 3), in particular we find that the posterior variance for ν is much higher when using the incompletely observed data. Nevertheless, the associated posterior mean values seem to align rather well with each other and the true value, suggesting that inference for the two-type model in a phylogenetic context is indeed possible without observing a type-specific census. Very similar observations can be made for independent replicate simulations (fig. S3, fig. S4). When we perform parameter inference for the default single-type BDP model and the single-type model without extinction for the same simulated data set, we find that the duplication rate tends to be underestimated in both models (table S2). The loss rate of the default single-type model only very slightly overestimates μ_1 while the loss rate in the single-type model without extinction strongly underestimates μ_2 .

Parameter estimation for simulated data under the DLF model

We next evaluate to what extent the two-type DL model can approximate the dynamics of the idealized DLF model, which entails non-independent evolution of gene copies within a family

Figure 3: Marginal posterior distributions for a simulated data set of 10,000 gene families, simulated along the eight-taxon *Drosophila* phylogeny from the two-type DL process with parameters $\lambda = 0.2, \mu_1 = 0.1, \nu = 0.2, \mu_2 = 5$ and a bounded Beta-geometric prior on the number of lineages at the root with $\eta = 0.95$, dispersion $\zeta = 4$ and bound at $Z = 10$. When there is more than one lineage at the root, each additional lineage has a $r = 0.5$ probability of being a type 2 gene. In black the posteriors conditional on the fully observed data (i.e. $X^{(u)} = (X_1^{(u)}, X_2^{(u)})$ tuples for each leaf node u) are shown, while in red the posteriors conditional on the total gene count $Z^{(u)}$ are shown.

(see methods). We again simulated multiple data sets of 1000 gene families as well as a large data set of 10,000 gene families for the eight-taxon *Drosophila* phylogeny. Note that the total loss rate for a family consisting of one redundant pair will be $2\mu_r$ in the DLF model, whereas the total loss rate for such a family under the two-type DL model (corresponding to the state $X = (1, 1)$) will be $\mu_1 + \mu_2 \approx \mu_2$. Because such small families dominate the data, we expect that $\mu_2 \approx 2\mu_r$.

In line with this expectation, we find that the estimated value of μ_2 under the two-type DL model corresponds to twice the simulated loss rate per redundant gene (μ_r) under the DLF model (fig. S5). Similar observations hold for ν , although the large posterior variance blurs the expected relationship of ν in the DLF model to twice ν in the two-type DL model. Additionally, we note that the posterior mean μ_1 and λ values for the two-type DL model are very accurate estimators of the corresponding parameters under the DLF model. For the large data set simulated under the DLF model with $\lambda = 0.2, \mu_{nr} = 0.1, \mu_r = 3$ and $\nu = 0.2$, we obtain an estimate for the loss rate per excess gene μ_2 of 5.99 (5.35, 6.65), again coinciding with twice the true loss rate per redundant gene (fig. S6). Similarly, the posterior mean estimate for ν under the two-type DL model is obtained as 0.35 (0.10, 0.62), also approximately corresponding to twice the true value of the underlying DLF model. The much higher posterior variance for this parameter makes the correspondence again however less clear. These simulations therefore suggest it is reasonable to assume that twice the loss rate per *excess* gene, estimated under

the two-type DL model, may serve as an approximation to the more intuitive loss rate per *redundant* gene in the DLF model. Posterior predictive simulations for the fitted two-type DL model further show that the posterior predictive distribution is fully compatible with the data simulated under the DLF model (fig. S7), indicating that the two-type DL model indeed can serve as an approximation to the DLF model.

Analysis of *Drosophila*, yeast and primate data

We performed Bayesian inference for the two-type DL model using gene count data from eight *Drosophila* species, eight yeast species and 11 primates (see methods) and compare the inferred model parameters to parameter estimates for different single-type BDP models (table 1). We note that for the yeast and primates data sets, initial tests indicated that numerical inaccuracies in the likelihood for the two-type model could lead to a failure of the MCMC algorithm to converge in some runs. Increasing the FFT length from $N = 16$ (as employed in our simulations) to $N = 32$ and decreasing tolerance settings in the ODE solver alleviated these issues and resulted in proper convergence. Furthermore, we found that the posterior distribution for the primates data set was bimodal, with one of the modes at $\mu_2 > 20$, $\lambda > 1.5$ and $r \approx 0.2$. We deem this unrealistic in light of the results for the single type models and other data sets, signaling the need for a more informative prior. We therefore further restrict $\mu_2 < 10$ for this data set.

Interestingly, the estimated rates for slow processes in the *Drosophila* and primates data sets are roughly on the same scale, corresponding to about 0.1 to 0.4 events per gene per 100 My, whereas the yeast data set yields parameter estimates that are considerably lower. Of course, this is not wholly unexpected, given the vastly different genomic organization of yeast species, with small genomes typically consisting of 5000 to 7000 predicted genes. The overall behavior of parameter estimates across models is similar for all three data sets. We find that the duplication rate estimated for the two-type DL model is higher than for the single-type models, which is in line with our simulations above. The loss rate for an excess (type 2) gene μ_2 is an order of magnitude higher than other rate parameters, with an implied half-life ($t_{1/2}$) of a type 2 gene of about 11, 39 and 16 My for the three data sets respectively (table 1). If we assume $\mu_2 \approx 2\mu_r$ (see above), these can be interpreted as the half-lives of duplicate *pairs* under the DLF model. Note that the ratios of λ/μ_2 of about 0.04, 0.02 and 0.08 are in complete agreement with the Beta-Geometric stationary distribution fit of the single-type model (fig. 2, Appendix 1). The marginal posterior mean estimate of the probability that a new duplicate gets established eventually (i.e. becomes a type 1 gene, rather than suffering loss, p_x) is a mere 3% for the *Drosophila* data, about 2% for the yeast data and 2% for the primates data. Again, under the DLF model, this value can be interpreted as the probability that a duplicate *pair* undergoes successful sub- or neofunctionalization so that it is stably established in the genome.

Clearly, a failure to model the different loss rates within a family caused by functional redundancy leads to a loss rate estimate μ that is dominated by the large number of non-redundant genes that only rarely get lost. The loss rate in the default single-type BDP model therefore

more closely resembles μ_1 than μ_2 in the two-type model. The opposite holds for the single-type model for the number of *excess* genes in non-extinct families. However the loss rates for this model (estimated at 3.94, 1.61 and 3.46 expected loss events per excess gene per 100 My for the three data sets respectively) are considerably lower than the estimates for μ_2 for the two-type model. This is consistent with the dynamics the two-type process is supposed to model, as duplicated genes that sub- or neofunctionalize should pull back the loss rate towards the loss rate of non-redundant genes in the corresponding single-type model. In other words, when assuming all duplicated genes have the same loss rate (as in the single-type non-extinct model), the presence of duplicate genes that have become essential leads to a downwardly biased loss rate when interpreted as the rate of pseudogenization of duplicate genes. An additional explanatory factor may be that the subset of families which do not go extinct has a lower average loss rate per excess gene than the full data set. Lastly, all single-type models seem to underestimate the duplication rate with respect to the two-type model, which may be a result of the biases in the loss rates caused by a failure to separate the dynamics of redundant versus non-redundant duplicated genes.

Table 1: Marginal posterior parameter estimates. For all analyses, the prior on the number of lineages at the root was a Beta-Geometric distribution with η and ζ parameters fixed to the marginal posterior mean values obtained from the stationary distribution fit (fig. 2, *Drosophila*: $\eta = 0.96, \zeta = 4.01$, yeasts: $\eta = 0.98, \zeta = 4.06$, primates: $\eta = 0.92, \zeta = 3.23$). For the two-type model we use an exponential prior for μ_2 with mean equal to the posterior mean value of μ under the single-type model with no extinction. For the primates data we further restrict $\mu_2 < 10$. Other priors are as described in the methods section. $t_{1/2}$ and p_x denote the half-life (in My) and probability of x-functionalization (in %) of a type 2 gene respectively.

Parameter	<i>Drosophila</i>	Yeasts	Primates (GO:0002376)
<i>Single-type (critical)</i>			
λ	0.19 (0.19, 0.20)	0.040 (0.038, 0.041)	0.18 (0.18, 0.19)
<i>Single-type</i>			
λ	0.17 (0.16, 0.17)	0.022 (0.020, 0.023)	0.20 (0.19, 0.21)
μ	0.24 (0.23, 0.25)	0.059 (0.056, 0.062)	0.17 (0.16, 0.18)
<i>Single-type (non-extinct)</i>			
λ	0.18 (0.17, 0.19)	0.016 (0.014, 0.019)	0.24 (0.22, 0.26)
μ	3.94 (3.71, 4.18)	1.606 (1.385, 1.813)	3.46 (3.24, 3.67)
<i>Two-type</i>			
λ	0.26 (0.24, 0.28)	0.027 (0.024, 0.030)	0.34 (0.32, 0.38)
μ_1	0.19 (0.18, 0.20)	0.055 (0.053, 0.058)	0.13 (0.12, 0.14)
μ_2	6.37 (5.98, 6.93)	1.794 (1.638, 1.972)	4.32 (3.95, 4.71)
ν	0.20 (0.08, 0.35)	0.045 (0.007, 0.101)	0.08 (0.03, 0.14)
r	0.12 (0.10, 0.15)	0.013 (0.002, 0.031)	0.04 (0.02, 0.07)
$t_{1/2}$ (My)	11 (10, 12)	39 (35, 42)	16 (15, 18)
p_x (%)	3.1 (1.3, 5.0)	2.3 (0.4, 5.2)	1.8 (0.7, 3.1)

Figure 4: Posterior predictive densities of the gene family size distributions for each of the eight leaves of the *Drosophila* phylogeny. The distribution of the differences of the observed gene family size frequencies (p_k) with the frequencies observed in 1000 posterior predictive simulations ($\tilde{p}_k$) is plotted. In red and gray posterior predictive distributions for the two-type and single-type model are shown respectively.

Posterior predictive simulations indicate that the two-type model provides a better fit to the size distribution than the single-type model for all data sets (fig. 4, fig. S8, fig. S10, fig. S11). For all three data sets we find that the Kullback-Leibler (KL) divergence $D_{\text{KL}}(p, \tilde{p})$ from the posterior predictive frequency distribution $\tilde{p}$ to the observed frequency distribution p for the two-type model is less than half of the same KL divergence obtained under the default single-type model (fig. S8, fig. S9). We note that the different degrees of correspondence of the posterior predictive distributions to the observed size frequency distributions for different taxa suggests rate heterogeneity across *branches* of the species tree, a complication we ignore in the present paper. As expected, the posterior predictive size distribution is underdispersed with respect to the true distribution further in the tail (fig. S12), which is a consequence of ignoring rate heterogeneity across *families*. This is less so for the single-type model, where the tail of the posterior predictive distribution is closer to the observed distribution. This is simply a result of the distribution on the number of lineages at the root being better preserved under the process, which has a lower overall event rate. We note that the posterior predictive distribution for the non-zero counts under the two-type model is nearly indistinguishable from the posterior predictive distribution for the single-type model for non-extinct families.

While rate heterogeneity across families could be accounted for by using a mixture modeling approach, we refrain from doing so for reasons of computational feasibility. We may however apply an *ad hoc* procedure to apply rate heterogeneity across families in our posterior predictive simulations by considering the fit of the BG distribution to the data. As we discuss in the

Figure 5: Posterior predictive simulations assuming λ/μ_2 is Beta distributed across gene families with dispersion $\zeta = 4.01$. On the left the joint distribution of family-specific λ and posterior mean μ_2 values is shown. The eight panels on the right show the posterior predictive family size distributions with (red) and without (black) the *ad hoc* procedure to account for rate heterogeneity across families based on the Beta-Geometric assumption.

methods section and Appendix 2, the BG distribution is the stationary distribution of a special case of the two-type DL model where $\mu_1 = \nu = 0$ and where the ratio λ/μ_2 is iid distributed according to a Beta distribution across families. The parameters of this stationary distribution can be easily estimated from the data, and we estimated the dispersion parameter for the *Drosophila* data for instance at $\zeta = 4.01$ (fig. 2). If we assume λ/μ_2 to be Beta distributed across families, and take for each replicate simulation of the posterior distribution the value of λ/μ_2 as the mean of this Beta distribution, we may simulate from the posterior distribution under this assumption as follows: For each of the $i \in (1, \dots, 1000)$ replicate simulations we sample a $(\lambda/\mu_2, \mu_2)_i$ pair from the posterior and compute $\alpha_i = \zeta(\lambda_i/\mu_{2,i})$ and $\beta_i = \zeta(1 - \lambda_i/\mu_{2,i})$. For each of the $j \in (1, \dots, n)$ families in the i th simulation replicate, we sample a random value for $\xi_{ij} = \lambda_{ij}/\mu_{2,ij}$ by sampling from $\text{Beta}(\alpha_i, \beta_i)$. We then obtain λ_{ij} and $\mu_{2,ij}$ by assuming $\mu_{2,ij} = \mu_{2,i}$ and $\lambda_{ij} = \xi_{ij}\mu_{2,i}$. Note that this procedure translates the assumed variation in λ/μ_2 across families to heterogeneity in λ alone across families. The joint posterior distribution of the family-specific rates obtained following this procedure is shown in fig. 5. The resulting posterior predictive distribution still fits the small family sizes well (as in fig. 4), but now no longer underestimates the proportions of larger families, and yields predictions compatible with the observed power law tail (fig. 5). This again clearly shows that the power-law tail of the gene family size distribution can be explained by rate heterogeneity across families.

Discussion

We describe a two-type continuous-time branching process model of gene family evolution, show the feasibility of estimating its parameters from incomplete data in a phylogenetic context using simulations, and perform Bayesian inference of model parameters for comparative genomic data sets. Comparisons with closely related single-type phylogenetic BDP models

highlight the shortcomings of these models and indicate how a two-type process may provide a first step towards more realistic stochastic models of gene family evolution, providing more detailed and biologically meaningful quantitative insights in the associated evolutionary processes.

The main virtue of the two-type DL model is the discrimination between non-redundant functional genes that are only rarely lost, and duplicates which may be largely redundant and therefore more prone to loss by pseudogenization. We noted that the model proposed here can be viewed as an approximation to the gene family evolution dynamics where duplicated genes are (partially) functionally redundant and evolve in a non-independent manner (as in the DLF model). Compared to the DLF model, the loss rate μ_2 in the two-type process does *not* correspond to the loss rate per redundant gene, but rather the loss rate per *excess* gene in a family. We explicitly examined the connection between the two models using simulated data and find that parameter estimates under the two-type DL model can be interpreted in the context of the more intuitive DLF model.

An alternative approach to account for differences in loss rates within a family which may be natural to adopt is to consider an age-dependent process (more precisely a ‘budding’ age-dependent branching process (Greenman and Chou 2016)), where the loss rate of a duplicated gene decreases through time. Such a model has been considered by Zhao et al. (2015), albeit not in the context of gene content evolution along a phylogeny. Inference of model parameters relies however on knowledge of the gene family trees and ages of duplication events (dated reconciled gene trees in the comparative genomic setting), the estimation of which is of course an extremely challenging statistical problem in its own right. Without access to such high-quality gene trees, inference from gene counts alone seems highly challenging. Indeed, statistical inference of age-dependent branching processes more generally is an active research topic of considerable mathematical sophistication (e.g. Fok and Chou 2013; Greenman and Chou 2016).

It is important to stress what is *not* modelled by the phylogenetic two-type DL process. Most importantly, we do not explicitly model the population genetics of fixation of copy number variants (CNVs). Clearly, most duplication events will either be lost due to drift if neutral, or lost by purifying selection when deleterious, and as a result leave no trace in extant genomes. The duplication rate λ in the model should therefore be interpreted as a proxy for the rate at which gene duplications that rise to high frequency occur. We have however no guarantee that the observed copy numbers are monomorphic in their respective populations, so the presence of low-frequency CNVs in the data may cause λ to overestimate the rate at which duplications that fix in the population occur. Similar considerations hold for the loss rate parameters μ_1 and μ_2 , which should be interpreted as rates of gene deletion events, not the rates of loss of unfixed duplicate genes by genetic drift or purifying selection. Of course, exactly the same issues hold for single-type models as well. We note that, if gene duplications were predominantly neutral (which is of course extremely unlikely), the estimated λ should roughly correspond to the per-gene duplicative mutation rate. For *Drosophila*, the duplicative mutation rate has been estimated at 1.25×10^{-7} duplications per gene per generation (Schrider et al. 2013).

Taking this as a crude estimate of the order of magnitude of the duplicative mutation rate, and considering a long-term average generation time between 7 to 20 days, the expected duplication rate λ under neutrality would be *at least* a 100 times larger than the rates we estimated under the two-type DL model, suggesting that the vast majority of duplications is deleterious. A more detailed picture of the distribution of fitness effects of new duplications remains however elusive.

The two-type DL model provides new quantitative insights in the long-term evolutionary dynamics of duplicated genes. In our analyses of comparative genomic data sets, we find that the loss rate per excess gene is much higher than the loss rate per base gene, suggesting that even when gene duplicates establish in a population, the selective pressures ensuring their maintenance differ strongly from those maintaining typical single-copy genes. For instance, for the *Drosophila* genus, we estimated that on average half of the duplicated genes are maintained over a period of approximately 11 My and that about 3% of the duplicated genes may eventually become stably established (complete x-functionalization) so that their loss rates reflect those of base genes. The half-life of a typical base gene on the other hand is estimated at 371 My. Concomitantly, our results indicate that estimates of long-term gene loss rates based on simple BDP models (as in e.g. Hahn, Han, and Han (2007) for *Drosophila*) likely underestimate the actual loss rates of duplicated genes significantly, while estimates of duplication rates tend to be in rough agreement with those presented here.

What then causes gene duplicates that establish in a population to remain prone to higher loss rates? Or conversely, how are gene duplicates that are non-essential and prone to loss at relatively high rates established in the first place? Of course, completely redundant duplicates may drift to fixation, and it is unsurprising that such genes would be prone to higher loss rates than their essential counterparts (Walsh 2003). Many gene duplicates that rise to fixation may however be adaptive (Han et al. 2009; Innan and Kondrashov 2010; Kondrashov 2012), demanding an answer to the first question. As Kondrashov (2012) stressed, it may be misleading to think of gene duplicates as either completely redundant or non-redundant, as many adaptive duplications may establish as a consequence of positive selection for increased dosage in a stressful environment, despite being *qualitatively* redundant. In such cases, a changing environment or the emergence of other genetic variants may alter selection pressures over time, and while some duplicated gene may promote adaptation during some environmental challenge, it may return to a state of complete redundancy or even come at a fitness cost later (Kondrashov 2012). Our phylogenetic analyses may go some way supporting this view of long-term gene family evolution, where many of the duplicated genes in multi-copy gene families reside for some time in the genome, but eventually suffer loss before undergoing complete sub- or neofunctionalization.

Acknowledgements & Funding

Arthur Zwaenepoel acknowledges funding from the Flemish fund for scientific research (FWO). Yves Van de Peer acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01).

References

- Athreya, Krishna B., and Peter E. Ney. 1972. *Branching Processes*. Grundlehren Der Mathematischen Wissenschaften. Berlin Heidelberg: Springer-Verlag. <https://doi.org/10.1007/978-3-642-65371-1>.
- Åkerborg, Örjan, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. 2009. "Simultaneous Bayesian Gene Tree Reconstruction and Reconciliation Analysis." *Proceedings of the National Academy of Sciences* 106 (14): 5714–9.
- Bailey, Norman T. J. 1990. *The Elements of Stochastic Processes with Applications to the Natural Sciences*. John Wiley & Sons.
- Beimforde, Christina, Kathrin Feldberg, Stephan Nylander, Jouko Rikkinen, Hanna Tuovila, Heinrich Dörfelt, Matthias Gube, et al. 2014. "Estimating the Phanerozoic History of the Ascomycota Lineages: Combining Fossil and Molecular Data." *Molecular Phylogenetics and Evolution* 78: 386–98.
- Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B Shah. 2017. "Julia: A Fresh Approach to Numerical Computing." *SIAM Review* 59 (1): 65–98.
- Byrne, Kevin P, and Kenneth H Wolfe. 2005. "The Yeast Gene Order Browser: Combining Curated Homology and Syntenic Context Reveals Gene Fate in Polyploid Species." *Genome Research* 15 (10): 1456–61.
- Csűrös, Miklós, and István Miklós. 2009. "Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model." *Molecular Biology and Evolution* 26 (9): 2087–95.
- Dietz, Laura. 2010. "Directed Factor Graph Notation for Generative Models." *Technical Report*.
- Drosophila 12 Genomes Consortium. 2007. "Evolution of Genes and Genomes on the Drosophila Phylogeny." *Nature* 450 (7167): 203.
- Emms, David M, and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 1–14.

- Fok, Pak-Wing, and Tom Chou. 2013. “Identifiability of Age-Dependent Branching Processes from Extinction Probabilities and Number Distributions.” *Journal of Statistical Physics* 152 (4): 769–86.
- Ge, Hong, Kai Xu, and Zoubin Ghahramani. 2018. “Turing: A Language for Flexible Probabilistic Inference.”
- Greenman, Chris D, and Tom Chou. 2016. “Kinetic Theory of Age-Structured Stochastic Birth-Death Processes.” *Physical Review E* 93 (1): 012112.
- Hahn, Matthew W, Tijn De Bie, Jason E Stajich, Chi Nguyen, and Nello Cristianini. 2005. “Estimating the Tempo and Mode of Gene Family Evolution from Comparative Genomic Data.” *Genome Research* 15 (8): 1153–60.
- Hahn, Matthew W, Mira V Han, and Sang-Gook Han. 2007. “Gene Family Evolution Across 12 Drosophila Genomes.” *PLoS Genet* 3 (11): e197.
- Han, Mira V, Jeffery P Demuth, Casey L McGrath, Claudio Casola, and Matthew W Hahn. 2009. “Adaptive Evolution of Young Gene Duplicates in Mammals.” *Genome Research* 19 (5): 859–67.
- Hoffman, Matthew D, and Andrew Gelman. 2014. “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.” *J. Mach. Learn. Res.* 15 (1): 1593–1623.
- Hughes, Timothy, and David A Liberles. 2008. “The Power-Law Distribution of Gene Family Size Is Driven by the Pseudogenisation Rate’s Heterogeneity Between Gene Families.” *Gene* 414 (1-2): 85–94.
- Huynen, Martijn A, and Erik Van Nimwegen. 1998. “The Frequency Distribution of Gene Family Sizes in Complete Genomes.” *Molecular Biology and Evolution* 15 (5): 583–89.
- Innan, Hideki, and Fyodor Kondrashov. 2010. “The Evolution of Gene Duplications: Classifying and Distinguishing Between Models.” *Nature Reviews Genetics* 11 (2): 97–108.
- Kondrashov, Fyodor A. 2012. “Gene Duplication as a Mechanism of Genomic Adaptation to a Changing Environment.” *Proceedings of the Royal Society B: Biological Sciences* 279 (1749): 5048–57.
- Kumar, Sudhir, Glen Stecher, Michael Suleski, and S Blair Hedges. 2017. “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.” *Molecular Biology and Evolution* 34 (7): 1812–9.
- Lynch, Michael, and John S Conery. 2003. “The Evolutionary Demography of Duplicate Genes.” *Genome Evolution*, 35–44.
- Rabier, Charles-Elie, Tram Ta, and Cécile Ané. 2014. “Detecting and Locating Whole Genome Duplications on a Phylogeny: A Probabilistic Approach.” *Molecular Biology and Evolution* 31 (3): 750–62.

- Rackauckas, Christopher, and Qing Nie. 2017. “DifferentialEquations.Jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia.” *Journal of Open Research Software* 5 (1): 15. <https://doi.org/10.5334/jors.151>.
- Roberts, Gareth O, and Jeffrey S Rosenthal. 2009. “Examples of Adaptive Mcmc.” *Journal of Computational and Graphical Statistics* 18 (2): 349–67.
- Sanderson, Michael J. 2003. “R8s: Inferring Absolute Rates of Molecular Evolution and Divergence Times in the Absence of a Molecular Clock.” *Bioinformatics* 19 (2): 301–2.
- Schrider, Daniel R, David Houle, Michael Lynch, and Matthew W Hahn. 2013. “Rates and Genomic Consequences of Spontaneous Mutational Events in *Drosophila Melanogaster*.” *Genetics* 194 (4): 937–54.
- Szöllősi, Gergely J, Eric Tannier, Vincent Daubin, and Bastien Boussau. 2015. “The Inference of Gene Trees with Species Trees.” *Systematic Biology* 64 (1): e42–e62.
- Thomas, Gregg WC, Elias Dohmen, Daniel ST Hughes, Shwetha C Murali, Monica Poelchau, Karl Glastad, Clare A Anstead, et al. 2020. “Gene Content Evolution in the Arthropods.” *Genome Biology* 21 (1): 1–14.
- Tsitouras, Ch., I. Th. Famelis, and T. E. Simos. 2011. “On Modified Runge–Kutta Trees and Methods.” *Computers & Mathematics with Applications* 62 (4): 2101–11. <https://doi.org/10.1016/j.camwa.2011.06.058>.
- Walsh, Bruce. 2003. “Population-Genetic Models of the Fates of Duplicate Genes.” In *Origin and Evolution of New Gene Functions*, 279–94. Springer.
- Xu, Jason, Peter Gutter, Midori Kato-Maeda, and Vladimir N. Minin. 2015. “Likelihood-Based Inference for Discretely Observed Birth–Death–Shift Processes, with Applications to Evolution of Mobile Genetic Elements.” *Biometrics* 71 (4): 1009–21. <https://doi.org/10.1111/biom.12352>.
- Zhao, Jing, Ashley I Teufel, David A Liberles, and Liang Liu. 2015. “A Generalized Birth and Death Process for Modeling the Fates of Gene Duplication.” *BMC Evolutionary Biology* 15 (1): 1–11.
- Zwaenepoel, Arthur, and Yves Van de Peer. 2019. “Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates.” *Molecular Biology and Evolution* 36 (7): 1384–1404.
- . 2020. “Model-Based Detection of Whole-Genome Duplications in a Phylogeny.” *Molecular Biology and Evolution*.

Supplementary material

Appendix 1: Transition probabilities for the two-type DL model

Here we derive the probability generating function (pgf) for the two-type DL process defined in the main text using techniques from Bailey (1990) and Xu et al. (2015). Let $f_{ij}(s_1, s_2, t)$ denote the pgf for the two-type branching process model in eq. 1 of the main text when starting at $X(0) = (i, j)$, i.e.

$$f_{ij}(s_1, s_2, t) = \sum_k \sum_l p_{ij}(k, l, t) s_1^k s_2^l$$

Importantly, the branching property implies the following relationship among the pgfs for different $X(0)$

$$f_{ij}(s_1, s_2, t) = f_{10}(s_1, s_2, t)^i f_{01}(s_1, s_2, t)^j$$

(e.g. Athreya and Ney (1972)), so we may work with f_{10} and f_{01} and recover the desired pgfs easily. Let $p_{ij}(k, l, dt) = r_{ij}(k, l)dt + o(dt)$, we also define the auxilliary generating functions

$$u_{10}(s_1, s_2) = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} r_{10}(k, l) s_1^k s_2^l$$

and $u_{01}(s_1, s_2)$ analogously in terms of $r_{01}(k, l)$. Note that the r_{10} and r_{01} functions are determined by the definition of the process in terms of its infinitesimal rates in eq. 1. Filling in the relevant parameters we obtain

$$\begin{aligned} u_{10}(s_1, s_2) &= \mu_1 - (\lambda + \mu_1)s_1 + \lambda s_1 s_2 \\ u_{01}(s_1, s_2) &= \mu_2 + \nu s_1 - (\lambda + \nu + \mu_2)s_2 + \lambda s_2^2 \end{aligned}$$

The pgfs for the process are related to the auxilliary generating functions

$$f_{10}(s_1, s_2, t) = \sum_k \sum_l \left(\mathbb{1}_{k=1, l=0} + r_{10}(k, l)t + o(t) \right) s_1^k s_2^l$$

and similarly for f_{01} , so that we get

$$\begin{aligned}f_{10}(s_1, s_2, t) &= s_1 + u_{10}(s_1, s_2)t + o(t) \\f_{01}(s_1, s_2, t) &= s_2 + u_{01}(s_1, s_2)t + o(t)\end{aligned}$$

Note furthermore that $\partial f_{ij}(s_1, s_2, t)/\partial t = u_{ij}(s_1, s_2)$. We can expand f_{10} as a Taylor series around t

$$f_{10}(s_1, s_2, t + h) = f_{10}(s_1, s_2, t) + h \frac{\partial f_{10}(s_1, s_2, t + h)}{\partial h} \Big|_{h=0} + o(h)$$

Now exploiting the property of the branching process that

$$f_{10}(s_1, s_2, t + h) = f_{10}\{f_{10}(s_1, s_2, t), f_{01}(s_1, s_2, t), h\}$$

we can rewrite the Taylor expansion as

$$\begin{aligned}f_{10}(s_1, s_2, t + h) &= f_{10}(s_1, s_2, t) + h \frac{\partial f_{10}\{f_{10}(s_1, s_2, t), f_{01}(s_1, s_2, t), h\}}{\partial h} \Big|_{h=0} + o(h) \\&= f_{10}(s_1, s_2, t) + hu_{10}\{f_{10}(s_1, s_2, t), f_{01}(s_1, s_2, t)\} + o(h)\end{aligned}$$

Which shows that

$$\frac{\partial f_{10}(s_1, s_2, t)}{\partial t} = u_{10}\{f_{10}(s_1, s_2, t), f_{01}(s_1, s_2, t)\}$$

and analogously for $\partial f_{01}/\partial t$. Combining this result with the generating functions u_{10} and u_{01} , we arrive at the following system of non-linear ordinary differential equations (ODEs)

$$f'_{10} = \mu_1 - (\lambda + \mu_1)f_{10} + \lambda f_{10}f_{01} \quad (2)$$

$$f'_{01} = \mu_2 + \nu f_{10} - (\lambda + \nu + \mu_2)f_{01} + \lambda f_{01}^2 \quad (3)$$

Where the arguments s_1, s_2 and t are omitted for notational convenience and differentiation is with respect to t .

Appendix 2: The stationary distribution of a birth-death process with no extinction and rate heterogeneity across families

We propose the Beta-geometric distribution as a reasonable prior for the number of ancestral lineages. This choice is motivated by two main factors: (1) the observed size distribution of gene families shows an approximate power-law decay and (2) the stationary distribution of a special case of the two-type process with duplication and loss rate heterogeneity across families can be shown to lead to a Beta-geometric probability mass function (pmf).

Consider the two-type branching process model proposed in the main text, where we set $\mu_1 = \nu = 0$ and $\mu_2 = \mu$, and where we assume the process starts with a single type 1 gene. In this model, the number of type 1 genes stays 1 perpetually, while the number of type 2 genes evolves as a stochastic birth-death-immigration process (BDIP) with immigration rate λ . The system of differential-difference equations for the transient distribution of the number of type 2 genes is

$$\begin{aligned}\frac{dp_0(t)}{dt} &= \mu p_1(t) - \lambda p_0(t) \\ \frac{dp_k(t)}{dt} &= k\lambda + (k+1)\mu p_{k+1}(t) - [k(\lambda + \mu) + \lambda]p_k(t)\end{aligned}$$

Setting $dp_k(t)/dt = 0$, the stationary pmf π can be obtained as

$$\pi_k = \left(\frac{\lambda}{\mu}\right)^k \pi_0$$

With the constraint $\sum_{k=0}^{\infty} \pi_k = 1$ and under the assumption that $\lambda < \mu$, we can find that $\pi_0 = 1 - \lambda/\mu$ so that the stationary pmf under this model becomes

$$\pi_k = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^k$$

Being a Geometric distribution on $k \geq 0$ with parameter $1 - \lambda/\mu$. Note that the distribution for the *total* number of genes is this Geometric distribution shifted by one, $\pi_k = (1 - \lambda/\mu)(\lambda/\mu)^{k-1}$, $k > 0$.

Now (considering point (1) above), empirical size distributions of non-extinct gene families are overdispersed with respect to the Geometric stationary prediction under this model. If however duplication and loss rates vary across families (which is of course expected to be the rule rather than the exception given the many functional and structural features of genes that should affect their propensity to duplicate or pseudogenize), the model above with Geometric stationary distribution gives rise to power-law like distributions for some models of rate heterogeneity.

Consider now λ and μ for each gene family to be a random variable iid from some distribution. A particularly straightforward and flexible model is to assume $1 - \lambda/\mu$ to be distributed according to a $\text{Beta}(\alpha, \beta)$ distribution across families (or equivalently, $\lambda/\mu \sim \text{Beta}(\beta, \alpha)$). The resulting compound stationary distribution becomes a so-called Beta-Geometric $\text{BG}(\alpha, \beta)$ distribution with pmf

$$\begin{aligned}\pi_k &= \int_0^1 \eta(1-\eta)^{k-1} f(\eta; \alpha, \beta) d\eta \\ &= \frac{1}{\text{B}(\alpha, \beta)} \int_0^1 \eta^\alpha (1-\eta)^{\beta+k-2} d\eta \\ &= \frac{\text{B}(\alpha+1, \beta+k-1)}{\text{B}(\alpha, \beta)}\end{aligned}$$

for $k > 0$ and where $f(\eta; \alpha, \beta) = \text{B}(\alpha, \beta)^{-1} \eta^{\alpha-1} (1-\eta)^{\beta-1}$ is the density of the Beta distribution. The mean of this distribution is $(\alpha + \beta)/\alpha$, and the expected value of λ/μ across families under this model is $(\alpha + \beta)/\beta$. In our work, we mostly employ the alternative parameterization using the mean $\eta = \alpha/(\alpha + \beta)$ and dispersion parameters $\zeta = \alpha + \beta$ of the underlying Beta distribution.

Having a meaningful interpretation as the stationary distribution of a simple birth-death process with rate heterogeneity, the BG distribution provides a flexible two-parameter alternative to the Geometric distribution employed by Rabier, Ta, and Ané (2014), Zwaenepoel and Van de Peer (2019) and Zwaenepoel and Van de Peer (2020) which provides an excellent fit to observed gene family size distributions. In addition to serving as an apt prior for the number of ancestral genes in a family, a fit of the BG distribution to observed data provides information on the ratio of λ/μ (λ/μ_2 in the context of the two-type process) and the variation of this ratio across families. We infer parameters of the BG distribution using a uniform prior on η and flat improper prior on ζ , using Turing.jl (Ge, Xu, and Ghahramani 2018) with the NUTS sampler (Hoffman and Gelman 2014).

Appendix 3: Inference for the single-type model with the Beta-Geometric prior distribution

For the single-type linear birth-death process model we use the same Beta-Geometric prior distribution for the number of genes at the root as for the two-type process. We implemented this in the *DeadBird* package for statistical inference of linear phylogenetic birth-death-immigration processes (which builds on work in Zwaenepoel and Van de Peer (2020)), which uses the algorithm of Csűrös and Miklós (2009) to compute the data log-likelihood. The algorithm of Csűrös and Miklós (2009) makes use of *conditional survival likelihoods*, and is exact, in the sense that it does not require a truncation of the state space. The conditional survival likelihood is defined as

$$\ell_k = \Pr(y|\theta, k \text{ lineages at the root leave observed descendants})$$

Note that the maximum k is bounded by the observed data (Csűrös and Miklós 2009), so that $1 \leq k \leq b$ where b is the relevant bound. Csűrös and Miklós (2009) describe an efficient algorithm to compute ℓ_k . Given a prior on the number of extant genes at the root $Z^{(a)}$, the complete data likelihood can be obtained as

$$\ell = \sum_{k=1}^b \ell_k \left(\sum_{i=0}^{\infty} \binom{k+i}{i} \epsilon^i (1-\epsilon)^k \Pr\{Z^{(a)} = k+i\} \right)$$

Where ϵ is the extinction probability at the root, i.e. the probability that a lineage extant at the root leaves no observed descendants.

For a Geometric(η) prior $\Pr\{Z = k\} = \eta(1-\eta)^{k-1}$, it is easy to show that

$$\ell = \sum_{k=1}^b \ell_k \frac{(1-\epsilon)^k (1-\eta)^{k-1} \eta}{(1-\epsilon(1-\eta))^{k+1}}$$

For the Beta-Geometric prior distribution we did not obtain a closed form solution. Using the property of the Beta function that $B(\alpha, \beta+1) = B(\alpha, \beta) \frac{\beta}{\alpha+\beta}$ we can find

$$\ell = \sum_{k=1}^b \ell_k \sum_{i=0}^{\infty} \binom{k+i}{i} \epsilon^i (1-\epsilon)^k \frac{B(\alpha+1, \beta+k+i-1)}{B(\alpha, \beta)} = \sum_{k=1}^b \ell_k (1-\epsilon)^k \sum_{i=0}^{\infty} A_{k,i} \quad (4)$$

With the recursion relations

$$\begin{aligned} A_{1,0} &= \frac{\alpha}{\alpha+\beta} \\ A_{k,0} &= A_{k-1,0} \frac{\beta+k-2}{\alpha+\beta+k-1} \\ A_{k,i} &= A_{k,i-1} \frac{\epsilon(k+i)(\beta+k+i-2)}{i(\alpha+\beta+k+i-1)} \end{aligned}$$

Which allows to efficiently approximate the infinite series in (4) above by some partial sum sufficiently far in the series. The sum converges very rapidly, so that in practice no more than 10 terms are needed. Similar formulas hold in the case where the domain of the Beta-Geometric distribution is $k \geq 0$. Note that the sum in (4) reduces to the result for the Geometric distribution with $\eta = \alpha/(\alpha+\beta)$ as $(\alpha+\beta) \rightarrow \infty$.

To compute the condition factor required for correcting the likelihood in accord with the data filtering strategy, we require the probability of a gene family going extinct given the extinction probability of a single lineage ϵ . This probability can be obtained as

$$\Pr\{E\} = \sum_{k=1}^{\infty} \epsilon^k \Pr\{Z = k\}$$

Which can be easily shown to correspond to $\eta\epsilon/(1 - \epsilon(1 - \eta))$ for the Geometric distribution on $Z^{(a)}$. For the Beta-Geometric distribution, a little more work leads to

$$\Pr\{E\} = \sum_{k=1}^{\infty} \epsilon^k \frac{B(\alpha + 1, \beta + k - 1)}{B(\alpha + \beta)} = \sum_{k=1}^{\infty} A_k$$

with the recursion relation

$$A_1 = \epsilon \left(\frac{\alpha}{\alpha + \beta} \right)$$

$$A_k = A_{k-1} \frac{\epsilon(\beta + k - 2)}{\alpha + \beta + k - 1}$$

So that we can efficiently approximate the infinite sum by some partial sum sufficiently far in the series ($k > 100$, say). Together these results allow the application of a Beta-Geometric prior distribution on the number of lineages at the root in the context of the algorithm of Csrs and Mikls (2009).

Supplementary figures & tables

Figure S1: Comparison of log transition probabilities computed using the pgf method (blue lines) and estimated using Monte Carlo simulations (black dots) for different algorithmic settings (tolerance settings in the ODE solver (tol) and FFT length N). Parameters were $\lambda = 0.2$, $\mu_1 = 0.1$, $\nu = 0.2$ and $\mu_2 = 5$. Transition probabilities are computed from the state $X(0) = (2, 3)$ to states $X(t) = (i, j)$ where $0 \leq i < 8$, $0 \leq j < 10$ and $t = 1$. Target states along the x -axes are ordered $[(0, 0), (0, 1), \dots, (0, 9), (1, 0), (1, 1) \dots]$. Monte Carlo estimates are based on 10 million independent simulations from the model.

Table S1: Taxon names, data sources and ‘accession’ strings for sequence data used in this study.

id	taxon	source	accession
dere	<i>Drosophila erecta</i>	NCBI	GCF_003286155.1_DereRS2
dana	<i>Drosophila ananassae</i>	NCBI	GCF_003285975.2_DanaRS2.1
dsim	<i>Drosophila simulans</i>	NCBI	GCF_000754195.2_ASM75419v2
dper	<i>Drosophila persimilis</i>	NCBI	GCF_003286085.1_DperRS2
dpse	<i>Drosophila pseudoobscura</i>	NCBI	GCF_009870125.1_UCI_Dpse_MV25
dmel	<i>Drosophila melanogaster</i>	NCBI	GCF_000001215.4_Release_6_plus_ISO1_MT
dyak	<i>Drosophila yakuba</i>	NCBI	GCF_000005975.2_dyak_caf1
dsec	<i>Drosophila sechellia</i>	NCBI	GCF_004382195.1_ASM438219v1
Ecymbalariae	<i>Eremothecium cymbalariae</i>	YGOB, v7-Aug2012	Ecymbalariae
Egossypii	<i>Eremothecium gossypii</i>	YGOB, v7-Aug2012	Egossypii
Klactis	<i>Kluyveromyces lactis</i>	YGOB, v7-Aug2012	Klactis
Lkluyveri	<i>Lachancea kluyveri</i>	YGOB, v7-Aug2012	Lkluyveri
Lthermotolerans	<i>Lachancea thermotolerans</i>	YGOB, v7-Aug2012	Lthermotolerans
Lwaltii	<i>Lachancea waltii</i>	YGOB, v7-Aug2012	Lwaltii
Tdelbrueckii	<i>Torulaspora delbrueckii</i>	YGOB, v7-Aug2012	Tdelbrueckii
Zrouxii	<i>Zygosaccharomyces rouxii</i>	YGOB, v7-Aug2012	Zrouxii
Hsapi	<i>Homo sapiens</i>	Ensembl, release 102	Homo_sapiens.GRCh38
Ptrog	<i>Pan troglodytes</i>	Ensembl, release 102	Pan_troglodytes.Pan_tro_3.0
Pabel	<i>Pongo abelii</i>	Ensembl, release 102	Pongo_abelii.PPYG2
Nleuc	<i>Nomascus leucogenys</i>	Ensembl, release 102	Nomascus_leucogenys.Nleu_3.0
Mmula	<i>Macaca mulatta</i>	Ensembl, release 102	Macaca_mulatta.Mmul_10
Panub	<i>Papio anubis</i>	Ensembl, release 102	Papio_anubis.Panu_3.0
Csaba	<i>Chlorocebus sabaeus</i>	Ensembl, release 102	Chlorocebus_sabaeus.ChlSab1.1
Cjacc	<i>Callithrix jacchus</i>	Ensembl, release 102	Callithrix_jacchus.ASM275486v1
Csyri	<i>Carlito syrichta</i>	Ensembl, release 102	Carlito_syrichta.Tarsius_syrichta-2.0.1
Mmuri	<i>Microcebus murinus</i>	Ensembl, release 102	Microcebus_murinus.Mmur_3.0
Pcoqu	<i>Propithecus coquereli</i>	Ensembl, release 102	Propithecus_coquereli.Pcoq_1.0
Ogarn	<i>Otolemur garnettii</i>	Ensembl, release 102	Otolemur_garnettii.OtoGar3

Figure S2: Posterior mean parameter estimates and 95% uncertainty intervals (y-axis) as a function of the true simulated value (x-axis) for simulations of data sets of 1000 gene families. μ_2 values were drawn uniformly from the interval (1, 10), while λ/μ_2 , μ_1/μ_2 and ν/μ_2 were drawn from a Beta(9, 1) distribution. r was drawn from a Beta(3, 1) distribution and ζ from a log normal distribution with mean 3 and variance 0.2. Plots are shown in pairs, with on the left the results conditioning on the incompletely observed data $Z = X_1 + X_2$ while on the right results conditioning on the fully observed data $X = (X_1, X_2)$ are shown.

Figure S3: An independent replicate of the simulation shown in fig. 3

Figure S4: An independent replicate of the simulation shown in fig. 3

Table S2: Posterior distribution summary showing marginal posterior means and 95% uncertainty intervals (lower, upper) for various models for the simulation replicate associated with fig. 3.

model	parameter	mean	lower	upper
Single-type	λ	0.14	0.14	0.15
	μ	0.12	0.11	0.13
	ζ	6.77	4.62	9.96
Single-type (fixed ζ)	λ	0.14	0.14	0.15
	μ	0.12	0.11	0.12
Single-type (non extinct)	λ	0.14	0.13	0.15
	μ	1.25	1.11	1.42
	ζ	3.69	2.65	5.11
Single-type (non extinct fixed ζ)	λ	0.14	0.14	0.15
	μ	1.27	1.13	1.41
Two-type	λ	0.19	0.18	0.21
	μ_1	0.10	0.09	0.11
	ν	0.25	0.03	0.54
	μ_2	4.94	4.44	5.51
	ζ	4.20	3.15	5.60
	r	0.48	0.41	0.54

Figure S5: Posterior mean parameter estimates and 95% uncertainty intervals (y-axis) as a function of the true simulated value (x-axis) for simulations of data sets of 1000 gene families. Data sets were simulated under the DLF model, while inference was performed under the two-type DL model. μ_2 values were drawn uniformly from the interval (1, 5), while λ/μ_2 , μ_1/μ_2 and ν/μ_2 were drawn from a Beta(9, 1) distribution. r was drawn from a Beta(3, 1) distribution and ζ from a log normal distribution with mean 3 and variance 0.2. Inference is based on the total gene count (Z). The red lines for μ_2 and ν mark the expected approximate relationship between the simulated parameter value in the DLF model and the corresponding parameter in the two-type DL model (i.e. $\mu_2 \approx 2\mu_r$ and $\nu \approx 2\nu_{\text{DLF}}$).

Figure S6: Marginal posterior distributions for the two-type DL model applied to data simulated under the DLF model. Vertical solid lines mark the true simulated values, while dashed lines mark the expected approximate correspondences $\mu_2 \approx 2\mu_r$ and $\nu \approx 2\nu_{\text{DLF}}$ (see main text) .

Table S3: Summary of the posterior distributions showing marginal posterior means and 95% uncertainty intervals (lower, upper) for various single-type models fitted to the data simulated from the DLF model (fig. S6).

Model	Parameter	Posterior mean	Lower	Upper
Single-type	λ	0.13	0.13	0.14
Single-type	μ	0.11	0.10	0.12
Single-type	ζ	5.44	3.81	7.51
Single-type (fixed ζ)	λ	0.13	0.13	0.14
Single-type (fixed ζ)	μ	0.11	0.10	0.12
Single-type (non extinct)	λ	0.14	0.13	0.14
Single-type (non extinct)	μ	1.00	0.87	1.15
Single-type (non extinct)	ζ	3.50	2.54	4.59
Single-type (non extinct fixed ζ)	λ	0.14	0.13	0.14
Single-type (non extinct fixed ζ)	μ	1.01	0.88	1.15

Figure S7: Posterior predictive simulations for the two-type DL model and the default single-type linear BDP model applied to data simulated under the DLF model.

Figure S8: Kullback-Leibler (KL) divergence from posterior predictive simulations of the gene family size distribution to the observed family size distribution (across all species) for the *Drosophila*, yeast and primates data. In red the KL divergences for the simulations under the two-type model are shown, whereas in gray the distribution of divergences for the single-type model is shown.

Figure S9: Kullback-Leibler (KL) divergence (along the x -axis) from posterior predictive simulations of the gene family size distribution to the observed family size distribution as in fig. S8, but now for each species separately. Results for the two-type model are displayed in red, for the single-type model in gray. See table S1 for species name abbreviations.

Figure S10: As fig. 4 but for the yeast data set.

Figure S11: As fig. 4 but for the primates data set.

Figure S12: Posterior predictive family size distribution (across all lineages) for the *Drosophila* (top row), yeast (middle row) and primates data (bottom row) for the two-type, default single-type and single-type for non-extinct ('single-type-ne') families models. The dots show the observed size distribution.