Chapter 3

Protocols for the Molecular Evolutionary Analysis of Membrane Protein Gene Duplicates

Laurel R. Yohe, Liang Liu, Liliana M. Dávalos, and David A. Liberles

Abstract

Gene duplication is an important process in the evolution of gene content in eukaryotic genomes. Understanding when gene duplicates contribute new molecular functions to genomes through molecular adaptation is one important goal in comparative genomics. In large gene families, however, characterizing adaptation and neofunctionalization across species is challenging, as models have traditionally quantified the timing of duplications without considering underlying gene trees. This protocol combines multiple approaches to detect adaptation in protein duplicates at a phylogenetic scale. We include a description of models for gene tree-species tree reconciliation that enable different types of inference, as well as a practical guide to their use. Although simulation-based approaches successfully detect shifts in the rate of duplication/retention, the conflation between the duplication and retention processes, the distinct trajectories of duplicates under non-, sub-, and neofunctionalization, as well as dosage effects offer hitherto unexplored analytical avenues. We introduce mathematical descriptions of these probabilities and offer a road map to computational implementation whose starting point is parsimony reconciliation. Sequence evolution information based on the ratio of nonsynonymous to synonymous nucleotide substitution rates (dN/dS) can be combined with duplicate survival probabilities to better predict the emergence of new molecular functions in retained duplicates. Together, these methods enable characterization of potentially adaptive candidate duplicates whose neofunctionalization may contribute to phenotypic divergence across species.

Key words Gene duplication, Gene tree, Birth-death models, Molecular evolution, dN/dS

1 Introduction

1.1 Gene Duplication and Membrane Proteins

The evolutionary mechanisms for generating novelty are key to understanding variation in phenotypic and taxonomic diversity across the Tree of Life. Identifying the genetic mechanisms behind the origin and maintenance of phenotypic diversity is therefore a fundamental objective of evolutionary genetics. While base pair substitutions provide a means for understanding the novel function of existing genes, the duplication of entire genes and genomes offers a source of new variation for functional diversification. Duplications are primary sources of innovation, from large-scale whole-

Tobias Sikosek (ed.), *Computational Methods in Protein Evolution*, Methods in Molecular Biology, vol. 1851,
https://doi.org/10.1007/978-1-4939-8736-8_3,

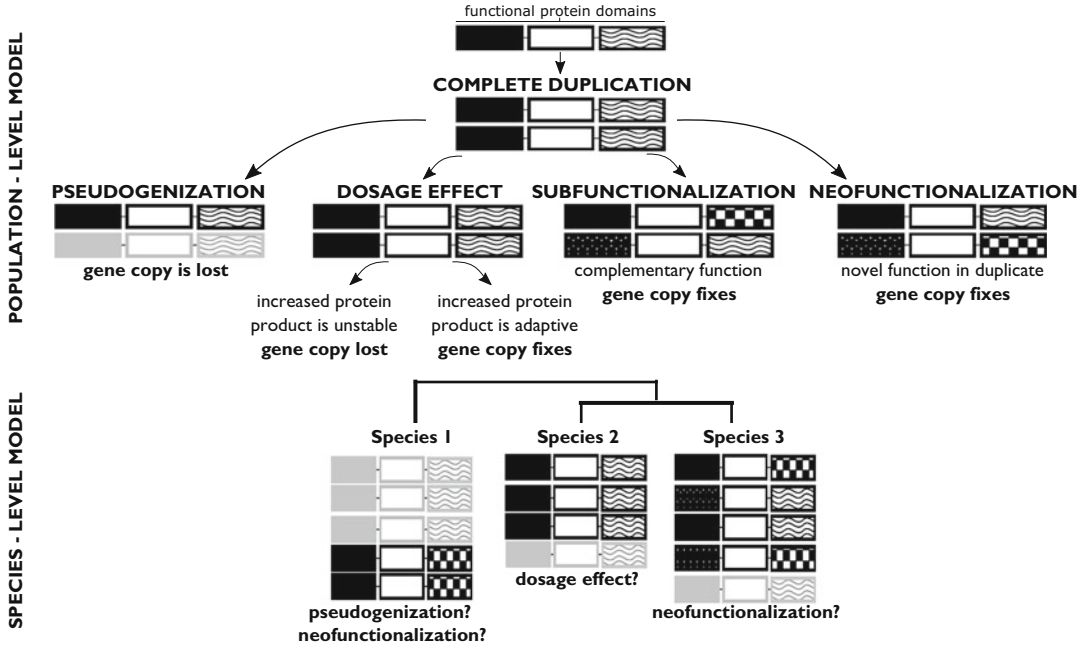

Fig. 1 Theoretical model of single-copy gene duplication and mechanisms for how a duplicate is fixed or lost in a population (top). Different patterns indicate different fixed amino acid differences. Grayed genes indicate loss of function. Note changes can also happen in regulatory regions, but are not shown here. The species-level model (bottom) is a cartoon of hypothetical scenarios that may be observed across species and their potential mechanisms. Figure adapted from [11]

genome duplications that may prompt speciation, seen in notable examples of teleost fish [1–3] or extraordinary polyploidy observed across plants [4–6], to duplications of a single gene, such as the expansion of multiple ion channels associated with the evolution of neural system complexity [7].

Just as new species evolve from ancestral lineages, new genes can evolve from those already present in the genome, and gene duplication is a primary molecular mechanism for the evolution of novel genes [8]. However, testing whether gene duplication is adaptive remains an unresolved challenge in evolutionary biology. In this chapter, we present an overview of the current methods used for studying gene duplication across species, and we describe a theoretical approach that integrates across several methodologies. We focus specifically on detecting adaptation in small-scale duplications from a single gene. Our primary emphasis is on membrane proteins, as many of these proteins are encoded by genes that evolve through a birth-death process that is a central mechanism to the model we propose.

A new gene may follow one of several trajectories after duplication (Fig. 1). Most probably, the duplicate is deleterious or neutral, does not fix in the population, and is lost [9–12]. It may also be

retained, either because it is adaptive or because of drift. The adaptive scenario may occur by either taking on a novel coding sequence or expression function or maintaining identical coding and expression domain functions as its ancestor, but increasing the expression of the gene product from redundant gene copies—a phenomenon known as dosage effect [8–10, 13]. In a nonadaptive scenario, the copy may fix but will likely pseudogenize after many generations unless subfunctionalization occurs. In each of these outcomes, the probability of gene retention and loss can be modeled as a function of time. The rates of amino acid-changing and silent substitutions that occur in each of these outcomes differs and can be informative in determining the fate of a gene duplicate. The overarching objective of this chapter is to quantify these distinct processes, present methods of simulation for different models, and synthesize the outcomes into biologically relevant interpretations of adaptation and loss.

The domain of a protein is the coding sequence that encodes the amino acid residues, and proteins can be composed of a single domain or several. These domains are the evolutionary unit of a protein, as part or all of the domain may undergo duplication or recombination or accumulate mutations that may affect protein function [14]. Membrane proteins are critical to several indispensable cellular functions including signal recognition, signal transduction, and transportation of materials into and out of the cell. In addition to these functions, membrane proteins are constrained to maintaining domains that enable the insertion, and that preserve the orientation, of the protein in the lipid bilayer of the cell membrane [15]. Membrane proteins also show a preference for positively charged residues that interact with the cytoplasmic side of the membrane [16]. With these constraints in mind, membrane proteins that respond to extracellular signals from the environment must also have binding sites for their respective ligands. Chemosensory receptors and immune-related membrane proteins involved in pathogen recognition encounter natural selection to detect ever-changing environmental cues. Many genes that encode these proteins evolve in a concerted birth-death fashion, in which genes duplicate, and duplicates may evolve a new function or pseudogenize [17]. This mechanism leads to a pattern of many closely related genes with similar and divergent function that can be classified as a multigene family.

2 Methods

2.1 Approaches and Limitations to Studying Gene Duplication

There are two major approaches to investigating the evolutionary process of gene duplication among species: birth-death models fit to a species tree and gene tree-species tree reconciliation. Several methodologies have been published using gene tree-species tree reconciliation [18–22]. This approach allows detection of branches in

which duplications and losses of particular gene copies occur, modeling the history of gene copies as a function of speciation events. However, currently available methods are either parsimony-based or do not estimate rates of gene retention [18, 23]. Importantly, any computed rate of loss is a homogeneous function of time along branches of the species tree, instead of a function relating loss to the age of the duplicate. This is a problem because the loss rate should not be constant through time. Instead, the probability of gene retention decays with duplicate age, making the loss rate a function of the time since duplication. Current interspecific models also conflate mutation and fixation, overlooking the time between these events. Future work could include the development of mutation-selection style models for gene duplication.

The second approach estimates rates of birth (duplication) and death (pseudogenization/loss) and tests if there are increased rates of either in different parts of the tree [24–26]. These methods calculate the likelihood of gene family data based on a birth and death rate while also considering branch lengths of species divergence times [24, 25]. This framework allows for explicit hypothesis testing of different birth-death rates in different parts of the tree but is subject to several assumptions, discussed below.

We provide an overview of these methods used for studying adaptation of gene duplication. Our examples provide a conceptual framework on how to define biologically meaningful questions in gene duplication analyses in a way that enables quantitative tests. Our examples also demonstrate strong caveats and ever-present assumptions in gene duplication analyses at the phylogenetic scale. First, we demonstrate a gene tree-species tree reconciliation method using parsimony. Second, we show how to test if the number of inferred duplications and losses is significantly higher or lower than expected under a null birth-death process through simulations. Third, we present the theory for developing a more integrated approach to characterize the different fates of gene duplicates.

2.1.1 Parsimony-Based Reconciliation

One early and common approach to gene tree-species tree reconciliation is to use the principle of parsimony to minimize either the duplication or the loss cost associated with mapping lineages of gene trees to branches of the species tree. This approach provides a valuable preliminary analysis for identifying discordance between the gene tree topology and the species tree (when the species tree relationship is not recovered within the gene family). Early approaches required the gene and species trees to be fully resolved with binary nodes, but subsequent approaches relaxed this assumption (*see* [27] for a review). As in parsimony-based tree reconstruction, the insensitivity of parsimony to duplication rates on branches with different lengths is a potential problem. A previous study has evaluated the relationship of different costs of accounting for gene

tree discordance to each other in a parsimony context, which represents a starting point for comparing these with model-based reconciliations under different models [28].

Here we provide an example of the amino acid transport protein gene family known as the amino acid-polyamine-organocation (APC) transporters in the sap-feeding insect suborder Sternorrhyncha. These insects have evolved a tight symbiotic relationship with gut bacteria that provides essential amino acids to supplement a nutrient-poor diet of phloem. Amino acid transport proteins facilitate the exchange of amino acids between the symbiont and its host across the bacteriocytes. It was known that some species of sap-feeding insects had multiple gene copies of APC transporters [29], but whether these duplications occurred prior to the radiation of sap-feeding insects was unclear. If an expansion of the number of APC transport proteins had occurred within this clade, it might be related to the increased reliance on nutrient supplements from gut symbionts. To answer this question, a published study implemented several reconciliation and birth-death methods to model the evolutionary history of the gene tree [30]. We first present the reconciliation of the APC gene tree with the Hemiptera species tree to demonstrate parsimony inference of duplications and losses (Fig. 2). Parsimony reconciliation was inferred using Notung [18]. Reconciliation can also be performed using a likelihood-based method (in this case, DupliPHY-ML [31]) that yields similar results (Fig. 3a).

Figure 2a shows that several lineages within Sternorrhyncha have experienced an expansion in the number of copies of APC transporters, as well as an expansion at the base of the group. However, in addition to the statistical inconsistency of parsimony inference when many changes accumulate, there is no hypothesis testing involved in describing whether any of these duplications or losses differ than from what is expected under a null evolutionary model of birth-death.

2.1.2 Birth-Death Models of Gene Duplication

Early models for gene duplication were traditional birth and death models. In these, the number of duplicate copies evolves through a stochastic birth-death process in which retention and loss are modeled with an exponential distribution [32]. Key parameters estimated in birth-death models are the birth and death rates of the genes, as well as the number of gene copies at each internal node. These models set up a statistical framework that describes how rates of gene duplication and loss may vary in different parts of the tree.

In the context of our example with the APC transporters in hemipteran insects, the parsimony inference suggests there may be an increased rate of gene duplication in Sternorrhyncha compared to other insects in the order. Likelihood-based birth-death models

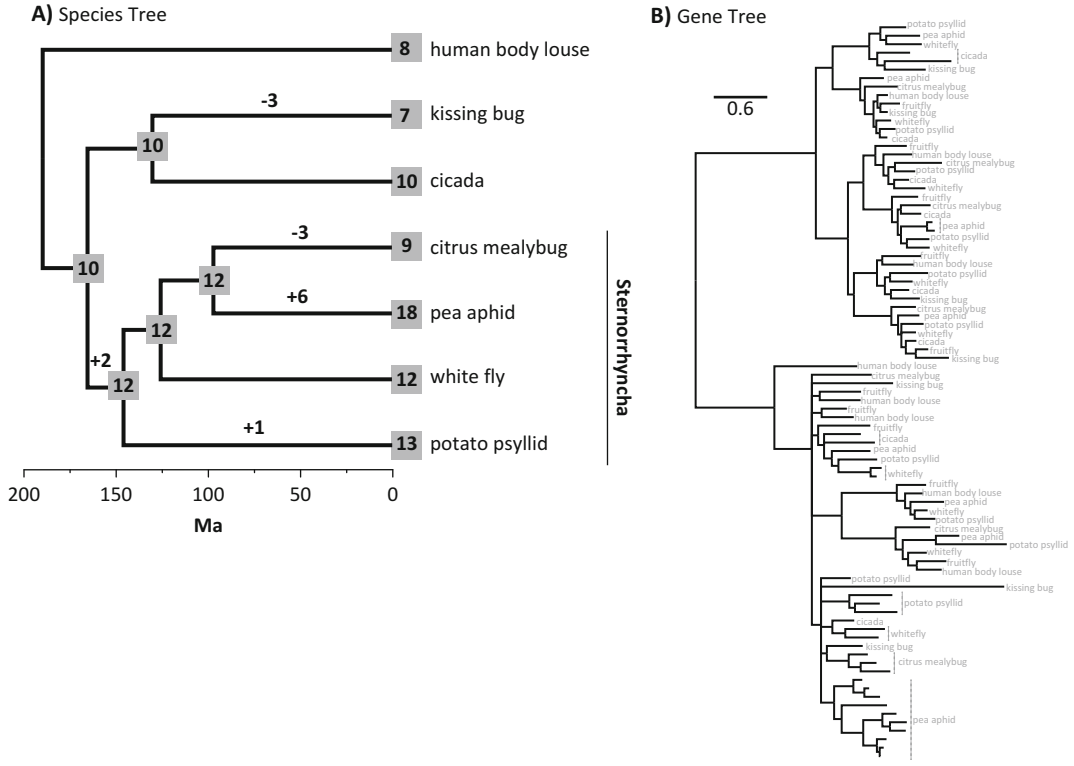

Fig. 2 (**a**) Species tree for Hemiptera insect order, denoted with the Sternorrhyncha sap-feeding insect suborder. The human body louse is an outgroup. The fruit fly (*Drosophila melanogaster*) was omitted from the species tree for clarity. Gray boxes indicate the number of gene copies inferred for each species and at each ancestral node. Branch labels indicate the number of duplications (+) or losses (−) inferred to have occurred at each respective branch as inferred using parsimony. (**b**) Gene tree of the APC amino acid transporter family. Each tip is a unique gene copy belonging to the species labeled at the tip

explicitly test whether multiple birth rates in different parts of the tree (in this case Sternorrhyncha v. background branches) better fit the data than a single birth rate for the entire phylogeny. A previous study estimated the birth rate (b) for different parts of the tree and found that a model with a single b for the entire phylogeny was a better fit than a model with a separate estimate of b for the Sternorrhyncha suborder (Table 1) [30]. Thus, from this approach, evidence does *not* support increased rates of duplication in sap-feeding insects.

While this approach can identify the species tree branches in which increased rates of duplication events occurred, it ignores the gene tree. Unlike reconciliation approaches, phyletic birth-death models simply fit parameters to numbers of gene copies, instead of actually considering if particular orthologs or paralogs are observed across species. Simulations of gene trees under similar birth and death rates estimated from one's data can provide a more thorough understanding of a null model of birth and death rate estimates

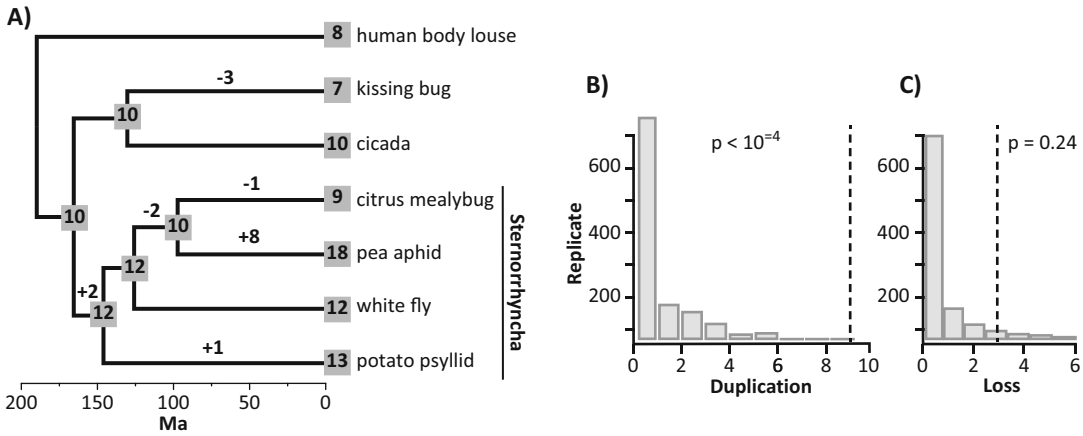

Fig. 3 Likelihood-based reconciliation of the APC transport proteins in Hemiptera. (**a**) Duplications and losses labeled on branches were inferred from reconciliation analyses in DupliPHY-ML v. 1.2 [31]. Gray boxes are number of APC transporter gene copies in each species or inferred at the ancestral node. (**b**) Simulation of expected number of duplications for Sternorrhyncha under a null birth-death process. The dotted line is the cumulative number of duplications observed from the DupliPHY-ML results. (**c**) Simulation of expected number of losses for Sternorrhyncha under a null birth-death process. The dotted line is the cumulative number of losses observed from the DupliPHY-ML results. P-values test whether the observed value is significantly different than the null distribution. Simulations were performed using GenPhyloData within the JPrIME v. 0.3.6 software [21]. Code for simulations is available in the supplementary material of [30]

Table 1
Hemiptera APC transporter gene family parameter estimates of likelihood-based birth-death model and likelihood ratio test results of model comparisons between a null model of a single birth rate (*b*) for the entire tree or two rates of *b*, one for the background branches and one for sternorrhynchans

Model	$b_{background}$	$b_{Sternorrhyncha}$	ML	np	LR	*p*-value
Single *b*	1.22×10^{3}	–	34.3	1	–	–
Multiple *b*	0.73×10^{3}	2.50×10^{3}	33.5	2	1.52	0.20

ML is the log-likelihood. np is the number of parameters. LR is the likelihood ratio. Inferences were made using CAFE v. 3.1 [40]. This model assumed the rate of birth to be equal to the rate of death. Analysis derived from [30].

under a neutral process. If the number of observed fixed duplicates or losses differs significantly from what is estimated from simulated data, the probability of fixation might be higher or lower than is expected by the null birth-death process. In our example with the APC transporter genes, the study used 1000 birth-death simulations based upon a birth rate estimated from the single *b* model in Table 1 [30]. From these gene trees, the expected number of duplications and losses could be estimated for each node of the tree. The study compared the observed values from the likelihood-based reconciliation (Fig. 3a) and found that sternorrhynchan insects did indeed have a significantly higher number of

duplications (but no difference in losses) compared to what was expected under a null birth-death scenario (Fig. 3b, c). While this approach is still subject to assumptions made by birth-death models, simulation experiments can provide a useful insight into null expectations for the underlying evolutionary process.

We argue, however, that these methods may be testing the wrong question. All models discussed so far conflate an increased rate of birth, which is a Poisson process similar to mutation events, with an increased rate of gene retention. In other words, instead of testing for an increased "birth rate," which should be intrinsically stochastic and homogeneous throughout long time scales, it would be ideal to measure an increased rate of gene retention. In the case of increased rates of gene retention, duplicates may be subject to selection and may indicate adaptation. Different processes lead to gene retention (Fig. 1), and these processes can be modeled. We propose an integrated framework to quantitatively differentiate among different gene retention scenarios that may lead to more biologically meaningful interpretations of adaptation that result from gene duplication.

2.2 Modeling Different Fates of Gene Duplicates: Integrating Reconciliation and Birth-Death

Several biological models have been proposed to depict the mechanisms that lead to different evolutionary fates for a gene duplicate (Fig. 1), including pseudogenization, neofunctionalization, subfunctionalization, or dosage effect. These mechanisms give rise to quite different retention dynamics that can lead to a time-dependent loss rate of gene duplicates, expressed as a function $\lambda(t)$. For nonfunctionalization, the loss rate is constant over time. In contrast, the loss rates of neofunctionalization and subfunctionalization decline over time and have been described with a Weibull hazard function [8]. For dosage effect, the rate of loss increases over time unless dosage effects are combined with subsequent neofunctionalization or subfunctionalization [33]. Alternative formulations with very similar dynamics have also been proposed [13]. Figure 4 depicts the shapes of these hazard functions under different scenarios.

From Reconciliation Probabilities to Birth-Death Models

In most birth-death model frameworks, the time-dependent loss rates have been incorporated in a generalized birth-death process to model the fate of gene duplicates. This means the evolution of the gene copies in a gene family is modeled as a pure birth process with a time-dependent birth rate, which is a function of the loss and birth rates in the original birth-death process. Since the loss rate characterizes the underlying retention mechanisms, the inference of the loss rates can identify either nonfunctionalization, subfunctionalization, dosage, or neofunctionalization as responsible for the observed site patterns of gene family data. However, an important caveat of all time-dependent models is that any rate of loss that is computed is a function of time along branches of the

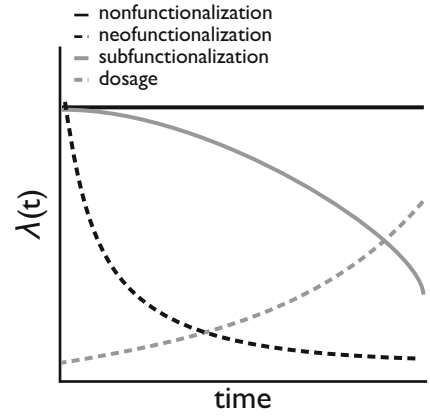

Fig. 4 Shape of the hazard function through time representing the rate of gene loss under the four different gene retention scenarios. Figure modified from [39]

species tree, instead of relating to the age of the gene duplicate. This is a problem because the loss rate should not be constant through time but instead be a function of the time since duplication, as the probability of gene retention decays with duplicate age.

Hence, it is more realistic to treat the loss rate as a function of the ages of gene copies. We propose a theoretical solution. Let $\lambda(t^*)$ be the loss rate of a gene copy at age t^*. The age-dependent model assumes 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 as follows [12]:
the probability of a gene duplication

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

the probability of a gene loss

$$P(n_{t+\Delta t} = n_t - 1) = \sum_{i=1}^{n_t} \lambda\left(t_i^*\right)\Delta t + o(\Delta t),$$

and the probability that the number of copies stays the same

$$P(n_{t+\Delta t} = n_t) = 1 - \left(n_t b + \sum_{i=1}^{n_t} \lambda\left(t_i^*\right)\right)\Delta t + o(\Delta t).$$

The parameter b is the birth rate; n is the number of gene copies at the present time; $\lambda\left(t_i^*\right)$ is the loss rate of gene copy i at age t_i^*. The three equations lead to a stochastic differential equation characterizing the age-dependent birth-death process. When the loss rate is constant (nonfunctionalization), the age-dependent birth-death model is identical to the time-dependent birth-death model derived from the reconstructed process (*see* [34] for derivation). For neofunctionalization and subfunctionalization, it has been demonstrated by simulation that the likelihood function of the

age-dependent model differs from that of the time-dependent model [12], and presumably for dosage as well. However, at the present time, there is no analytic solution to the stochastic differential equation when subfunctionalization, neofunctionalization, and dosage are the underlying mechanisms governing the age-dependent birth and loss rates. Research on the age-dependent model will provide indispensable insights on the evolution of gene duplicates.

The model we propose differs from existing approaches, as it constrains the inference of duplication events with speciation events while also calculating an age-dependent survival probability of gene copies. If a speciation event occurs at t_i, the probability of gene copy retention is a survival probability E_j calculated from the hazard function $\lambda(t)$, which represents the instantaneous loss at time t. Instead of modeling the time associated with retention or loss as constant through time, it will actually be calculated from the moment the duplication occurred, which can be denoted as t^*, reflecting the age-related duplicate notation described in the equations above [8, 9, 13].

We present a simple example to demonstrate how these probabilities may be calculated and how these probabilities can then be integrated with a gene tree-species tree reconciliation framework. Figure 5 shows an example gene tree with one specific reconciliation solution that may have occurred throughout the history of the gene family and species phylogeny. The solution is shown based on parsimony. In this scenario, two duplications and one loss have occurred in the phylogeny. The probability of retention is the product of all survival probabilities of different events in Table 2. The hazard function $\lambda(t)$ and its corresponding survival function are different for each outcome in Fig. 5 [8], including nonfunctionalization, neofunctionalization, subfunctionalization, and dosage effect. The product of all survival probabilities occurring for each event (e.g., Table 2) will reflect the survival probability of all

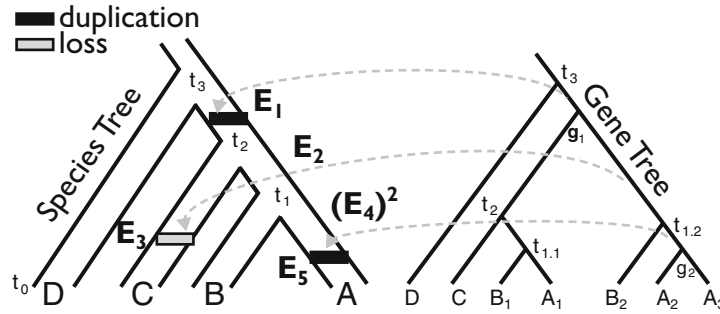

Fig. 5 Cartoon of species tree-gene tree reconciliation. Speciation times (t_i) and gene divergence times (g_i) are noted on nodes. E_4 is squared because it is counting both branches from time t_1. Event probabilities are listed in Table 2

Table 2
Events and probabilities of Fig. 5

Event	Description	Probability
E_1	Duplication and retention	$e^{-\int_0^{g_1-t_2}\lambda(t)dt}$
E_2	Retain duplicate	$e^{-\int_{g_1-t_2}^{g_1-t_1}\lambda(t)dt}$
E_3	Lose duplicate	$1-e^{-\int_{g_1-t_2}^{g_1}\lambda(t)dt}$
E_4	Retain duplicate	$e^{-\int_{g_1-t_1}^{g_1}\lambda(t)dt}$
E_5	Duplication and retention	$e^{-\int_0^{g_2}\lambda(t)dt}$

The probability of the reconciled tree in Fig. 5 is the product of all event probabilities. Gray arrows indicate probabilities that do not include a speciation event. The branch length-dependent birth rate can also be incorporated, when relevant.

duplicates in the gene tree. The best-fit hazard function model can be determined by model selection using the Akaike or Bayesian Information Criterion.

It should be emphasized that this example only accounts for a single set of events for one proposed reconciliation solution, as opposed to multiple hidden events that may have also occurred. Integrating over all possible reconciliation histories is, in theory, the only way to account for all possible hidden events. However, this is not a feasible solution given the possible number of hidden events that may have occurred. A more tractable solution is to begin with a parsimonious reconciliation and iteratively consider hidden events and alternative reconciliations according to a branch and bound-style approach. In this regard, a finite set of events (such as those shown in Table 2) for each reconciliation history can be compared with one another, and the most probable solution among this finite set of specific histories can be calculated.

2.3 Combining Survival Probabilities with dN/dS

For each outcome in Fig. 1, there is an expected behavior of the ratio rates of nonsynonymous (dN) to synonymous (dS) substitutions (dN/dS or ω) for the gene copy (Fig. 6). The behaviors of this ratio can reveal biologically meaningful interpretations relevant to molecular adaptation. For example, analyses of mammalian olfactory receptors, a hyperdiverse gene family that encodes G-protein-coupled chemosensory receptors, have shown that some particular orthologous gene groups have undergone rapid expansions and have high dN/dS relative to the median, suggesting functional diversification of these receptor types [35]. However, dN/dS is not currently modeled in any methodology used to study gene duplication, despite predictable functions under different gene retention scenarios. When genes are initially redundant following

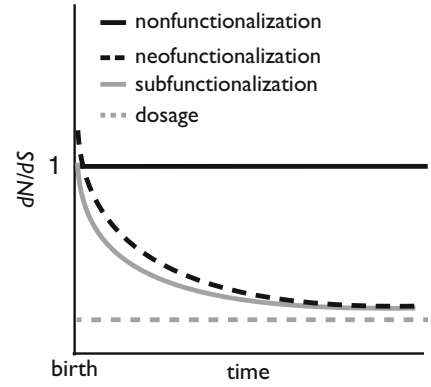

Fig. 6 Expected *dN*/*dS* of duplicated copy after gene duplication

duplication, they are expected to show neutral evolution or at least relaxation from purifying selection. Genes that nonfunctionalize should continue to evolve with $dN/dS = 1$, whereas duplicates that are retained through either the neofunctionalization or subfunctionalization process should see dN/dS decay toward a rate consistent with non-duplicated genes as an asymptote (Fig. 6). Indeed, it has been empirically shown that accelerated rates of dN/dS occur after duplication and then subsequently decline [36]. There may be little information to differentiate between neofunctionalization and subfunctionalization with these data, although it might be anticipated that neofunctionalizing genes at some early point have $dN/dS > 1$ (depending upon several factors, including the starting value of purifying selection and the strength of positive selection), something not expected for subfunctionalization. For subfunctionalization, dN/dS may not initially be as high as with neofunctionalization, as part of the gene is still under strong purifying selection to maintain ancestral function. In the case of selection for increased dosage, strong purifying selection is expected from the moment of duplication, as duplicates are functionally the same ($dN/dS << 1$ and constant).

One previously used approach is to approximate the age of the duplication event by building a histogram of pairwise dN/dS values of duplicates related to dS values [9]. Across collections of genes from a genome, each empirical frequency distribution is a sample of an underlying duplication process. When a gene family is known, an alternative is to examine branch-specific changes in dN/dS in lineages downstream from a duplication event. In this scenario, the onset of selection post-duplication in individual lineages can be evaluated.

The dN/dS statistic is one of the most commonly used approaches to measure the strength of selection among species, but it can be susceptible to false positives if there is purifying selection on synonymous mutations [37]. Meaningful dN/dS estimates may also be problematic for recent duplicates in closely

related lineages [38]. Mutation-selection models can offer a complementary set of tools to estimate the strength of selection and should also be considered in this framework [37].

3 Concluding Thoughts

Gene duplication is a fundamental mechanism underlying novel protein function. However, the fate of a gene duplicate is complex, and it can be challenging to determine whether or not gene duplication events are adaptive at phylogenetic time scales. Reconciling the evolutionary history of the gene family with the species tree and estimating rates of duplication and loss are the two most common approaches to analyzing gene duplication, but current methods are prone to assumptions that hinder a meaningful biological interpretation of parameter estimates. We proposed an approach that integrates both reconciliation and birth-death models to estimate the probabilities of different gene retention scenarios. Future research on the implementation of such an approach will bridge theory to practical application for a more comprehensive understanding of adaptive gene duplication, a key process in protein evolution.

Acknowledgements

This research was supported in part by DEB-1442142 to L.M.D., DEB-1701414 to L.M.D., D.A.L., and L.R.Y., and DBI-1222940 to D.A.L. and L.L.

References

1. Hoegg S, Brinkmann H, Taylor JS et al (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J Mol Evol 59:190–203
2. Jaillon O, Aury J-M, Brunet F et al (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–957
3. Lien S, Koop BF, Sandve SR et al (2016) The Atlantic salmon genome provides insights into rediploidization. Nature 533:200–205
4. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815
5. De Bodt S, Maere S, Van De Peer Y (2005) Genome duplication and the origin of angiosperms. Trends Ecol Evol 20:591–597
6. Hollister JD (2015) Polyploidy: adaptation to the genomic environment. New Phytol 205:1034–1039
7. Liebeskind BJ, Hillis DM, Zakon HH (2015) Convergence of ion channel genome content in early animal evolution. Proc Natl Acad Sci U S A 112:E846–E851
8. Konrad A, Teufel AI, Grahnen JA et al (2011) Toward a general model for the evolutionary dynamics of gene duplicates. Genome Biol Evol 3:1197–1209
9. Hughes T, Liberles DA (2007) The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. J Mol Evol 65:574–588
10. Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered 100:605–617

11. Sikosek T, Bornberg-Bauer E (2010) Evolution after and before gene duplication? In: Dittmar K, Liberles D (eds) Evolution after gene duplication. Wiley-Blackwell, Hoboken, NJ, pp 105–131
12. Zhao J, Teufel AI, Liberles DA et al (2015) A generalized birth and death process for modeling the fates of gene duplication. BMC Evol Biol 15:275
13. Teufel A, Zhao J, O'Reilly M et al (2014) On mechanistic modeling of gene content evolution: Birth-death models and mechanisms of gene birth and gene retention. Computation 2:112–130
14. Chothia C, Gough J, Vogel C et al (2003) Evolution of the protein repertoire. Science 300:1701–1703
15. von Heijne G (2006) Membrane-protein topology. Nat Rev Mol Cell Biol 7:909–918
16. Poolman B, Geertsma ER, Slotboom D-J (2007) A missing link in membrane protein evolution. Science 315:1229–1231
17. Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152
18. Chen K, Durand D, Farach-colton M (2000) NOTUNG: a program for dating gene duplications. J Comput Biol 7:429–447
19. Berglund-Sonnhammer AC, Steffansson P, Betts MJ et al (2006) Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. J Mol Evol 63:240–250
20. Doyon JP, Ranwez V, Daubin V et al (2011) Models, algorithms and programs for phylogeny reconciliation. Brief Bioinform 12:392–400
21. Sjöstrand J, Sennblad B, Arvestad L et al (2012) DLRS: gene tree evolution in light of a species tree. Bioinformatics 28:2994–2995
22. Hermansen RA, Hvidsten TR, Sandve SR et al (2016) Extracting functional trends from whole genome duplication events using comparative genomics. Biol Proced Online 18:11
23. Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genom 3:201–212
24. Hahn MW, De Bie T, Stajich JE et al (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res 15:1153–1160
25. Liu L, Yu L, Kalavacharla V et al (2011) A Bayesian model for gene family evolution. BMC Bioinformatics 12:426
26. Han MV, Thomas GWC, Lugo-Martinez J et al (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol 30:1987–1997
27. Eulenstein O, Huzurbazar S, Liberles DA (2010) Reconciling phylogenetic trees. In: Dittmar K, Liberles D (eds) Evolution after gene duplication. Wiley-Blackwell, Hoboken, NJ, pp 185–206
28. Górecki P, Eulenstein O (2014) Refining discordant gene trees. BMC Bioinformatics 15:S3
29. Duncan RP, Husnik F, Van LJT et al (2014) Dynamic recruitment of amino acid transporters to the insect/symbiont interface. Mol Ecol 23:1608–1623
30. Dahan RA, Duncan RP, Wilson AC et al (2015) Amino acid transporter expansions associated with the evolution of obligate endosymbiosis in sap-feeding insects (Hemiptera: Sternorrhyncha). BMC Evol Biol 15:52
31. Ames RM, Money D, Ghatge VP et al (2012) Determining the evolutionary history of gene families. Bioinformatics 28:48–55
32. Arvestad L, Lagergren J, Sennblad B (2009) The gene evolution model and computing its associated probabilities. J ACM 56(7):44
33. Teufel AI, Liu L, Liberles DA (2016) Models for gene duplication when dosage balance works as a transition state to subsequent neo-or sub-functionalization. BMC Evol Biol 16:45
34. Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. Philos Trans R Soc Lond Ser B Biol Sci 344:305–311
35. Niimura Y, Matsui A, Touhara K (2014) Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res 24:1485–1496
36. Pegueroles C, Laurie S, Albà MM (2013) Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. Mol Biol Evol 30:1830–1842
37. Spielman SJ, Wilke CO (2015) The relationship between dN/dS and scaled selection coefficients. Mol Biol Evol 32:1097–1108
38. Mugal CF, Wolf JBW, Kaj I (2014) Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol Biol Evol 31:212–231
39. Liberles DA, Teufel AI, Liu L et al (2013) On the need for mechanistic models in computational genomics and metagenomics. Genome Biol Evol 5:2008–2018
40. De Bie T, Cristianini N, Demuth JP et al (2006) CAFE: A computational tool for the study of gene family evolution. Bioinformatics 22:1269–1271