REVIEW

REVISED Models for the retention of duplicate genes and their biological underpinnings [version 2; peer review: 2 approved]

Raquel Assis[1], Gavin Conant[2], Barbara Holland[3], David A. Liberles[4], Malgorzata M. O'Reilly[3], Amanda E. Wilson[4]

[1]Florida Atlantic University, Boca Raton, Florida, USA
[2]North Carolina State University, Raleigh, North Carolina, USA
[3]University of Tasmania, Hobart, Tasmania, Australia
[4]Temple University, Philadelphia, Pennsylvania, USA

v2 **First published:** 24 Oct 2023, **12**:1400
https://doi.org/10.12688/f1000research.141786.1
Latest published: 12 Feb 2024, **12**:1400
https://doi.org/10.12688/f1000research.141786.2

Abstract

Gene content in genomes changes through several different processes, with gene duplication being an important contributor to such changes. Gene duplication occurs over a range of scales from individual genes to whole genomes, and the dynamics of this process can be context dependent. Still, there are rules by which genes are retained or lost from genomes after duplication, and probabilistic modeling has enabled characterization of these rules, including their context-dependence. Here, we describe the biology and corresponding mathematical models that are used to understand duplicate gene retention and its contribution to the set of biochemical functions encoded in a genome.

Keywords
gene duplication, probabilistic modeling, theoretical biology, Markov model, synteny, phylogenetic analysis

This article is included in the Bioinformatics gateway.

This article is included in the Evolutionary Bioinformatics collection.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 12 Feb 2024		
version 1 24 Oct 2023	✓ view	✓ view

1. **Nadia El-Mabrouk**, Université de Montréal, Montréal, Canada
2. **Arturo Becerra**, Universidad Nacional Autónoma de México, Mexico City, Mexico

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: David A. Liberles (daliberles@temple.edu)

Author roles: Assis R: Writing – Original Draft Preparation, Writing – Review & Editing; **Conant G**: Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Holland B**: Writing – Original Draft Preparation, Writing – Review & Editing; **Liberles DA**: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **O'Reilly MM**: Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Wilson AE**: Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Assis R, Conant G, Holland B *et al.* **Models for the retention of duplicate genes and their biological underpinnings [version 2; peer review: 2 approved]** F1000Research 2024, **12**:1400 https://doi.org/10.12688/f1000research.141786.2

First published: 24 Oct 2023, **12**:1400 https://doi.org/10.12688/f1000research.141786.1

REVISED Amendments from Version 1

In response to reviewers, a few changes were made to the manuscript. Most importantly, a new table (Table 1) has been added to summarize the models described and their biological assumptions. A paragraph has been added to describe the mechanisms of small scale duplication and corresponding retention profiles, including two new references. A clarifying paragraph organizing the work has been added to the introduction. Lastly, a series of more minor presentation and grammatical changes has been made.

Any further responses from the reviewers can be found at the end of the article

Introduction

Richard Feynman left the quotation, "What I cannot create, I do not understand" on his blackboard at the time of his death. Creation in mathematical modeling is writing down a model that describes a process. The retention of duplicate genes over long evolutionary periods involves mechanisms and processes in population genetics, evolution, molecular biology, ecology, and biochemistry. Here we describe the current state of modeling in the field of duplicate gene retention.

The genome can experience duplications of its content across a range of sizes, from incomplete duplications of single genes to small-scale events (single duplicate genes created through tandem duplication or retrotransposition events) to large scale events that involve multiple genes or even whole genomes. These events are broadly classified as whole genome duplications (WGD) or small scale duplications (SSD): the two types have several key differences. WGD duplicates are duplicated together with their interacting partners, and the population (and cellular) genetic model for the initial fixation of WGD duplicates is less straightforward than for SSDs. WGD events can be either allopolyploidies or autopolyploidies (see below). SSD duplicates typically do not see their interaction partners duplicated and initially have a frequency of $1/(2N_e)$ in a diploid population. SSD duplicates are typically assumed to begin as identical copies, but this isn't always the case.[1–6]

In examining small-scale duplication, tandem duplicates may be identical at birth, but also may be born as chimeras or as partial duplicates.[2,7] Those that are non-identical can be viewed as partly along the way to functional divergence. Identical duplicates are more common at birth, but less common among older tandem duplicates.[2,7] Similarly, genes that emerge by retrotransposition are less common at birth, but because they are likely born in a different expression environment and chromosomal location, diverge faster and are, relative to their birth rate, more common at older ages.[8]

Once duplicated, the accumulation of nonfunctionalizing mutations leading to the eventual loss of one copy through neutral processes can be naturally modeled with an exponential decay distribution.[9–12] Several factors can lead to duplicate gene pairs losing their redundancy and falling under the action of natural selection; when this change occurs, that pair will deviate from the neutral expectation of exponential loss.[12–14] In general, it has been suggested by Wagner (2005)[15] that the expression cost of a duplicated gene leads to a fitness cost in its possessor, and a similar argument could be made for a replication cost, especially for organisms that have limitations on genome size. Such a limitation would occur when selection acts to minimize replication time in log phase growth.

More generally, duplicates can be retained as a mechanism to gain extra expression.[16] An example of this in humans and other mammals appears to be the convergent duplication and retention of copies of the amylase gene.[17] Trypanosomes seem to regulate gene expression more generally through gene duplication, with very few transcription factors in their genomes.[18,19] There is also a selective pressure to retain duplicates that physically interact in stoichiometric balance to prevent misassembly of imbalanced heteromultimers or pathological interactions among their exposed hydrophobic surfaces.[20–24] Mechanisms of duplicate preservation that involve changes of function through mutation accumulation include subfunctionalization,[25,26] the partitioning of functions among copies from the pre-duplication ancestral state, and neofunctionalization,[16] the acquisition of a new beneficial function.

This review is focused on the characterization of different types of models with distinct assumptions that characterize the duplicate gene retention process. The goal of all of these models is to probabilistically predict which genes are likely to be found in genomes that have WGD events of different ages and ongoing processes of SSD. Multiple models for all of the processes described above exist and are discussed below together with their biological assumptions. Table 1 summarizes the models that have been characterized here.

Table 1. A summary of the models described for duplicate gene retention together with their assumptions is presented.

Model Name or Description	What is Modeled and Model Features	What is Assumed about Biology	Key Reference(s)
Duplicate Copy Presence or Coding Sequence Models			
Population Moran Model	Compares duplicate frequency and age in a population against a neutral expectation	Relies upon dS to estimate age; assumes deviation from neutrality due to selection and not demographics	81
Exponential Loss Model	Models retention over time with a simple loss rate	The simple loss rate is consistent with a neutral process without any onset of negative selection	9
Generalized Weibull Hazard Model	Models retention over time with mechanism-specific hazards	Works on simplified assumptions about average processes acting on genes	12
POInT	A single rate describing presence and absence at a syntenic location across genomes	A single loss rate is consistent with neutral loss without the onset of negative selection	120
Ji et al.	Tests for asymmetry in evolutionary rate in the coding sequence to infer mechanisms	Assumes that asymmetry is due to direct selection when other processes can also generate asymmetry; does not consider gene expression changes	101
Duplicate Gene Expression Models			
Gene Expression Continuous Trait Nested Hierarchy	A hierarchy of continuous trait models is applied to look at the evolution of expression levels over a phylogeny	Works only on reconstructing a phenotypic trait without knowledge of the underlying genotypic changes	
CDROM	Using asymmetry of rates of gene expression divergence to infer mechanisms	Works on phenotype without considering underlying genotype; assumes that asymmetry is due to direct selection without other processes acting	50
CLOUD	Gene expression inference for duplicates using an Ornstein-Uhlenbeck Process with a neural network	Works on phenotype without considering the underlying genotype	69
Integrated Models			
Biophysical Subfunctionalization Plus Dosage Markov Model	Integrates a biophysical model for protein interaction/misinteraction with a previous Markov model for subfunctionalization	Assumes modularity of gene regulatory units and ultimate retention through complementary gene expression changes	76
jPRIME models	Exponential Loss model integrated with lateral transfer and speciation	Single rate exponential loss models are consistent with neutrality without negative selection	11
Maere et al. models	Exponential Loss model integrating across duplication and divergence processes	Paralog losses are due to drift, but with differing chances of being subject to such losses for different functional classes of genes	113
Phase Type models	Time to subfunctionalization or loss of one gene copy post duplication is modelled as Phase Type distributed.	Assumes time until a loss of each subfunction or the coding region are each exponential (with different rates). Assumes all functions are protected by selection	91
QBD models	Models the evolution of a gene family within a species. QBDs track 'level' the size of the gene family as well as information on the amount of redundancy (the 'phase')	Assumes time until a loss of each subfunction or the coding region are each exponential (with different rates). Assumes all functions are protected by selection	94,95

Gene Duplicability

Different retention mechanisms are differentially applicable to different genes. This has given rise to the notion of gene duplicability, that some genes are inherently more duplicable than other genes.[10,27,28] For a gene to be retained after duplication, it ultimately needs to be either subfunctionalizable or neofunctionalizable.[12] To be subfunctionalizable, a gene must have at least two modular functions (biochemical activities, including binding, or modular expression domains), such that there are mutations that can disable one subfunction without damaging others. The probability of eventual subfunctionalization for two identical duplicate gene copies scales as $1 - 0.5^{(f - 1)}$, where f is the number of functions. The probability of neofunctionalization is harder to quantify, but either new selectable biochemical functions or expression domains must be evolvable. At the coding sequence level, this is influenced by the fold, number of binding partners, including those that are obligate heteromultimerization partners (proteins that obligately form multimers with protein products from different genes),[22,23] and type of function encoded.[29–35] Network position and expression level also influence gene duplicability.[28,33,36–40] Empirically, there is a class of "duplication-resistant" genes where natural selection apparently acts against the maintenance of both copies.[35] Different genomes might have different proportions of genes that are duplicable, as well as that are subject to dosage constraints.[41]

In analyzing the retention of genes following two rounds of WGD, the Atlantic salmon genome paper[42] presented a conditional probability analysis suggesting that the gene duplicability hypothesis predicts that genes retained after one round of WGD might be more likely to be retained after the second round. However, prior analyses from plants tended to suggest that the factors that favor the retention of duplicates after a first polyploidy tend to be attenuated in subsequent polyploidies.[43] Support for the hypothesis from the test in Atlantic salmon was also lacking, but there is more complexity to the process, including changing gene duplicability, the time-dependence of the retention process, and other factors,[44,45] and probabilistic models that can be used as an expectation for different hypotheses are described below.[41]

The Biological Considerations as Building Blocks for Models

To model duplicate gene retention, one must describe what is mutable and selectable. Protein encoding genes must function as proteins after transcription and translation. They are expressed at a given concentration in specific places and at specific times. They then carry out various functions: binding, catalysis, or transport in interaction with other molecules in the cell. This is what we mean by function. The expression domains for a protein are a quantitative description of where and when expression occurs. There are cases where duplication is used as a mechanism for amplifying expression level, although this seems to be a temporary situation in most organisms, with trypanosomes being a possible exception.

Modeling Expression Evolution of Duplicate Genes as a Stand-Alone Process

Though classical models for the retention of duplicate genes often consider their levels of sequence divergence, gene expression data provide a promising source of underutilized information. In particular, gene expression data are now widely available for many species and often consist of measurements across multiple conditions, which can include tissues, sexes, and developmental stages. These measurements are an attribute of function, as knowledge of where and when a gene is expressed provides insight into its biological roles. Indeed, Ohno proposed that the first step of functional divergence between duplicate genes is their expression divergence.[16] Thus, gene expression is a trait that can be exploited to understand gene function and, in the case of duplicate genes, the divergence between their functions.

Many early studies compared expression levels between duplicate genes, finding that divergence between copies is often widespread, rapid, and asymmetric.[13,46–48] Yet expression divergence between duplicate genes does not provide information about the exact mechanisms of their retention. For instance, both neofunctionalization and subfunctionalization result in functional divergence between gene copies. Thus, it is important to compare the expression profiles of both copies to that of the ancestral single-copy gene, as this can elucidate how each copy has changed since duplication. Such an approach was developed about a decade ago[49] and later implemented as the software CDROM.[50] Applications of this approach have uncovered widespread neofunctionalization in *Drosophila,*[49] mammals,[51] honeybees,[52] and grasses.[53]

However, a key shortcoming of the approach of Assis and Bachtrog (2013)[49] is that it does not account for stochastic changes in gene expression arising from phenotypic drift.[54] This obstacle can be overcome by modeling gene expression evolution on the phylogenetic tree relating a pair of duplicates and their single-copy ancestor. There is a natural hierarchy of models for describing how gene expression evolves along the branches of a phylogenetic tree. The simplest is Brownian motion (BM), which models phenotypic drift without making any assumption of selection for a particular expression level.[55] The next level of the hierarchy is an Ornstein-Uhlenbeck (OU) process with stabilizing selection for a particular expression level.[56–60] Finally, one can utilize an OU process with a shift reflecting positive selection for an optimal expression level.[61–65] Another conceivable approach is to model the genotype driving expression evolution, but this requires an understanding of the combinatorial role of promoters and enhancers regulating expression,[66,67] which is currently beyond our modeling capabilities.[68]

With this in mind, researchers have recently begun to employ OU models for studying the expression evolution of duplicate genes.[24,69] Additionally, DeGiorgio and Assis (2021)[69] developed CLOUD, which predicts retention mechanisms of duplicate genes by overlaying their OU model with a neural network.[69] Though likelihood ratio tests (LRTs) have classically been used for similar tasks with single-copy genes,[24,58–61,70] machine learning approaches present several advantages, such as the optimization of model fit to training data, direct evaluation of performance on independent test data, and ability to make predictions from data with correlated or conflicting signals.[71] Further, such methods make predictions solely from data,[71] which can be advantageous when the underlying evolutionary model is unknown. Indeed, CLOUD demonstrates excellent predictive performance, outshining CDROM in classifying retention mechanisms while also being able to predict parameters corresponding to expression optima and strengths of selection and drift.[69]

Still, much remains to be done in this area. For one, many advanced machine learning algorithms have yet to be explored in this context. Moreover, though expression data can provide a lot of useful information about a gene, this does not necessarily mean that we should neglect other complementary sources, such as its sequence or protein structure. Another advantage of machine learning is that it is not weighed down by additional information, as correlated or conflicting signals can be reduced or even removed through regularization. However, the problem lies in extending the underlying OU model to accommodate diverse pieces of information. Last, one can argue that the most important extension of such work is to accommodate more species and gene copies. Currently, most researchers do not have access to expression data for multiple of the same conditions in many species, but this is soon to change as the cost of sequencing continues to decrease. And of course, it is critical to assay the fates of gene families with more than two members, as many such families are prominent across study systems and may be key to understanding adaptation.

Modeling the Evolutionary Cost of Gene Duplication

Analyses of duplicate genes often start from the premise that gene duplications are selectively neutral, creating redundant copies that can potentially degrade through degenerative mutations.[9,72] The patterns of which genes do and do not tend to survive in duplicate and the dosage-balance hypothesis (see above) already suggest that gene duplications are not all selectively neutral at birth. As mentioned above, Wagner has extended this argument by showing that, at least in microbial organisms, the gene expression costs associated with an extra gene copy are rarely if ever expected to be selectively neutral.[15,73] His model considers the per-time unit cost, in terms of high energy phosphate bonds, of expressing a duplicate gene for different ranges of mRNA and protein levels, finding that, for reasonably large values of the effective population size of microbial species, those energy costs are large enough to discount the hypothesis that the fate of a duplication is primarily driven by neutral evolution.[74] The principle that excess gene expression has measurable negative fitness effects due to the costs of transcription and translation has been elegantly experimentally explored in *E. coli.*[75]

Dosage Imbalance Cost

It is well established that maintaining stoichiometric balance with interacting partners is an important driving force to preserve duplicate genes in genomes while waiting for other preservation mechanisms to act. The mechanistic driving force behind this is thought to be the prevention of the accumulation of exposed hydrophobic residues that populate binding interfaces and can lead to misinteractions that might be deleterious to cells. An explicit model that relates fitness to the expected concentration of surface hydrophobic residues has been generated and used to explore how this model enables the transition to subfunctionalized states, with opposite trends observed after WGD and SSD.[76] This model is a mechanistic update over a previously described hazard function model that did not model this underlying biochemistry.[77]

Population Genetic Considerations

In eukaryotic organisms, the baseline state for most chromosomes is diploidy. Either across the whole genome or for individual loci, this diploid state is disrupted after gene duplication. Functional tetraploidy has meiotic implications that are not present for SSD events (see[42] for a discussion in Atlantic salmon). Over time, WGD events return to a state of functional diploidy and may start that way for alloduplication events with the chromosome sets already diverged, as may have been the case for Xenopus[78] and Brassica species.[79,80]

However, so far, this divergence characterization has been viewed without the underlying population-level dynamics. While more complex for WGD events, SSDs in diploid organisms begin with a frequency of $1/(2N_e)$ and must fix before they are lost if they are going to be retained. The neutral expectation for eventual fixation of such duplicates is that they will fix with a probability equal to their frequency. Classical population genetics then gives a time-dependent expectation for the frequency based upon the age of the duplicate. Stark *et al.* (2021)[81] have presented a population genetic model to evaluate if the age-dependent frequency is unexpected for a duplicate evolving neutrally. The power of this approach was evaluated using a Moran Model, with two selective parameters, one for selection on the duplicate itself, which can be positive when total gene dosage amplification is beneficial or negative due to factors like expression and replication cost, and the other for selection on the new function (for example neofunctionalization).

In using this model with actual segregating duplicates, the frequency of a duplicate in a population can be measured through population genomic sequencing. However, the age of the duplicate is the next question. For SNPs, population geneticists examine the length of tracts of identity by descent to estimate the age of an allele (e.g., duplicate locus) (see[82]), but a much simpler approach based upon pairwise pS values between copies may be possible. Application of these approaches to real data has not been performed yet to evaluate their performance.

Those considerations are used to evaluate selection on the copy itself. Selection on the sequence of the duplicate might be measured by a more rich data source and parametrization or in a simpler manner by examining ratios like pN/pS or using tests like the McDonald-Kreitman test applied to duplicates.[83]

Interspecific and Phylogenetic Models

Moving from intraspecific to interspecific analysis of gene duplicates, including models that run on a single genome, early work from Lynch and Conery (2000)[9] and from Lynch, Force, and coworkers[10,25,84] has been pioneering. The first model presented by Lynch and Conery (2000)[9] modeled duplicate gene retention expectations with a simple exponential distribution. This assumes that no matter how long a gene has been in a genome, the instantaneous probability of loss is constant, which is not consistent with retention mechanisms, but is useful as a nonfunctionalization null model. Konrad *et al*. (2011)[12] and Teufel *et al*. (2016)[77] described sets of hazard functions that did not have the property of time-independent hazard functions, and Zhao *et al*. (2015)[85] presented an age-dependent birth-death process inspired by this framework. Yohe *et al*. (2015)[86] presented a theoretical gene tree-species tree reconciliation framework using the Konrad *et al*. (2011)[12] model, but this was never implemented as software. Arvestad *et al*. (2009)[11] presented a formal probabilistic gene tree-species tree reconciliation framework using the exponential distribution model and generated software for this. Others have created similar software packages.[87,88] Additional innovations to this framework have included the treatment of synteny,[89] and species level processes.[90]

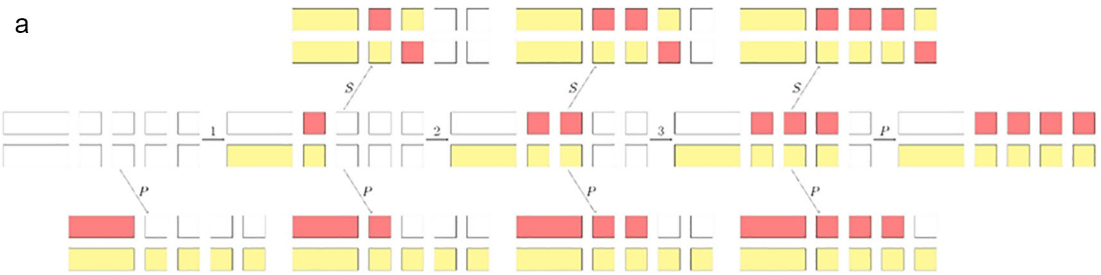
Contemporaneously with the Lynch and Conery (2000)[9] modeling, Force *et al*. (1999)[25] presented a more mechanistic framework for subfunctionalization and neofunctionalization as processes. This was formalized as a Markov Model by Stark *et al*. (2017)[91] and expanded upon by Wilson and Liberles (2023)[76] to enable consideration of dosage balance. Multi-scale Markov models reflect a further step in this trajectory.

Multi-Scale Markov Models

A wide range of multi-scale Markov models for the evolution of gene families has been studied in the literature. Models with states that record very detailed information about biology are suitable for simulation-based analysis. However, such models may not be useful for theoretical analysis due to the size of their state space. On the other hand, models with simplified state space are useful for in-depth theoretical and numerical analysis, which often leads to novel biological insights. Both types of models provide powerful tools for the analysis, and the choice of one over another may depend on the types of biological questions one might want to answer.

As an example, Stark (2017)[92] suggested a simulation model for the evolution of a family of genes, in which detailed information is recorded within a binary-matrix of 0's and 1s such that each row corresponds to a gene and each column corresponds to its function. Later, Diao *et al*. (2022)[93] applied the binary matrix model of Stark (2017)[92] in their simulation-based analysis, which led to some interesting biological insights. Their results suggested that when the rate of gene duplication dominates the rate of gene loss, then the distribution of tree shapes is close to following the uniform ranked tree shape (URT) distribution (i.e., the distribution for a constant birth-death process). However, when the opposite is true, then gene trees are more balanced than the URT distribution.

Next, Stark *et al*. (2017)[91] constructed a model with a simplified state space, for the theoretical analysis of the evolution of duplicate genes, see Figure 1. The manageable size of the state space allowed them to derive analytical expressions for the rates of subfunctionalization and pseudogenization. This led to the result, earlier predicted by classical models, that models with subfunctionalization provided a better fit to the age distribution of duplicate genes than models with a constant rate. Further, Diao *et al*. (2020)[94] developed a more advanced model for the theoretical analysis of the evolution of a family of duplicate genes that was based on the application of a level-dependent Quasi-Birth-and-Death (QBD) process. The state (n,m,k) of their QBD model consists of the variables n and m representing the number of genes and the number of the redundant genes respectively, and the variable k which, in a simplified manner, records the remaining information about the family. The authors took advantage of both types of models. They used the simulation-based binary matrix model of Stark (2017)[92] to obtain detailed outputs and then fitted the parameters of their QBD to data obtained from these outputs. Next, they derived biological insights by computing metrics based on the expressions from the theory of QBDs, such as the stationary distribution of observing various states within the model and the distribution of the time it takes for the family to lose a gene. Soewongsono *et al*. (2023)[95] then applied this QBD model to a more general



b

$$
\mathbf{Q} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ z-2 \\ z-1 \\ S \\ P \end{array}
\begin{array}{c}
\begin{array}{cccc|cc} 0 & 1 & \dots & z-1 & S & P \end{array} \\
\left[\begin{array}{cccc|cc}
-(2(u_e+zu_r)) & 2zu_r & 0 & \dots & 0 & 2u_e \\
0 & -(u_e+2(z-1)u_r)) & (z-1)u_r & \ddots & (z-1)u_r & u_e \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & -(u_e+2u_r) & 2u_r & u_e \\
0 & 0 & 0 & 0 & u_r & u_e+u_r \\
\hline
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{array}\right]
\end{array}.
$$

Figure 1. To model the evolution of gene duplicates, Stark *et al*. (2017)[91] constructed a Markov chain with state space {0,1, ...,z-1,S,P} and generator Q where z is the number of regulatory regions within the gene, and S and P is are the subfunctionalization and the pseudogenization (absorbing) state, respectively. In the above example of transitions with z=4, the regions hit by null mutations are in red, and the regions protected by selective pressure are in yellow. This figure is adapted from Stark *et al*. (2017),[91] which was published under an open access license.

problem of reconciliation, in which the task is to find a mapping of a gene tree to a species tree, to maximize the likelihood. The authors provided an algorithm to compute the likelihood of the reconciliation given the available incomplete data.

Modeling Asymmetric Divergence and Gene Conversion in Duplicate Genes

Another angle to evaluate duplicate gene retention mechanisms involves an examination of evolutionary symmetry. In fact asymmetric divergence between duplicate gene copies is relatively common.[96] This result is particularly striking when one considers that many tests have low power to detect asymmetry.[97] The test that was employed uses an outgroup gene from a relative lacking the duplication to polarize differences between the two paralogs, forming a triplet tree. Maximum likelihood was then used to compare a symmetric model, where dS (K_s), dN (K_a), or dN/dS (K_a/K_s or omega) was constrained to be equal for both paralogs, to an alternative model where the divergence statistic was allowed to differ between paralogs.[96] It is tempting to attribute to neofunctionalization cases where one paralog has accelerated evolution relative to the other, but in fact many different modes of evolution can induce asymmetry.[98] In fact, asymmetry in divergence is arguably more interesting for the differences seen between duplicate genes created by different mechanisms[3] or as a means for detecting recent gene conversion. We have used the ancient polyploidy in baker's yeast have experienced very recent gene conversion, such that ancient polyploidy-produced paralogs in one species are more similar to each other than those copies are to their orthologs in a closely related species, despite the fact that the divergence time between the ortholog pairs is probably ten-fold lower than the time since the paralogs were formed.[99]

Gene conversion refers to several mutational mechanisms that can allow part of the sequence of one member of a gene family to overwrite the corresponding region in another paralog, effectively erasing some of the divergence between the two paralogs.[100] Because such events violate the assumption of independent evolution between paralogs, they are difficult to treat with standard models. Ji *et al*. (2016),[101] have described a codon model of evolution that jointly considers the paired codons from two paralogous genes, incorporating a parameter t modeling the frequency with which conversion events alter the paralogs' sequences. This model confirms the surprisingly high rate of gene conversion among the yeast ribosomal proteins, which had previously and incorrectly been taken to represent the more general rate of gene conversion among yeast paralogs.[102]

Whole Genome Duplication: Duplicate Losses, Modeling and Synteny

While approaches such as standard time-independent birth and death models can be applied to duplicate genes produced by WGD, or polyploidies, there are complexities and opportunities introduced by WGD events that benefit from models that are specific to them. Polyploidy refers to a variety of events that result in eukaryotic cells with more than two copies of the genome.[103] Polyploid lineages are formed relatively often, but most quickly go extinct.[104] However, great trunks of the eukaryotic tree of life descend from surviving ancient polyploidy events, including all vertebrates and flowering plants, as well as specific lineages of yeasts, ciliates, and other plants.[105]

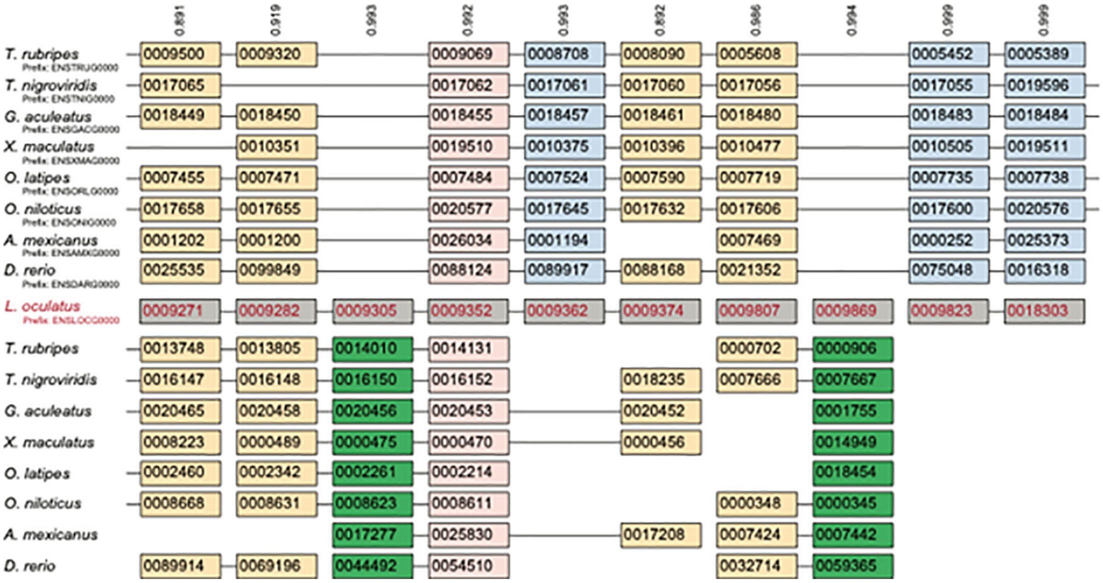
Polyploid individuals can form through the merger of genomes from the same species, known as autopolyploidy, or of distinct species, referred to as allopolyploidy.[106] The relative frequency of formation of these two types of polyploids may be approximately equal,[107] but because allopolyploidy confers the potential benefits of both polyploidy and hybridization, there is reason to suspect that most surviving ancient polyploidy events were allopolyploidies.[107–109]

The term WGD is potentially slightly misleading because it suggests that all genes in the genome are duplicated. Initially they are. However, for any reasonably old polyploidy event, many or even most of the duplicate genes will have been lost.[109,110] Probably most of these losses occur through the fixation of loss-of-function mutations in one copy by genetic drift, a process common to duplicates of all types.[9] As described in the earlier characterization of duplicate retention mechanisms, selection from various sources can also play a role.

Another Class of Models for Evolution After Polyploidization Events

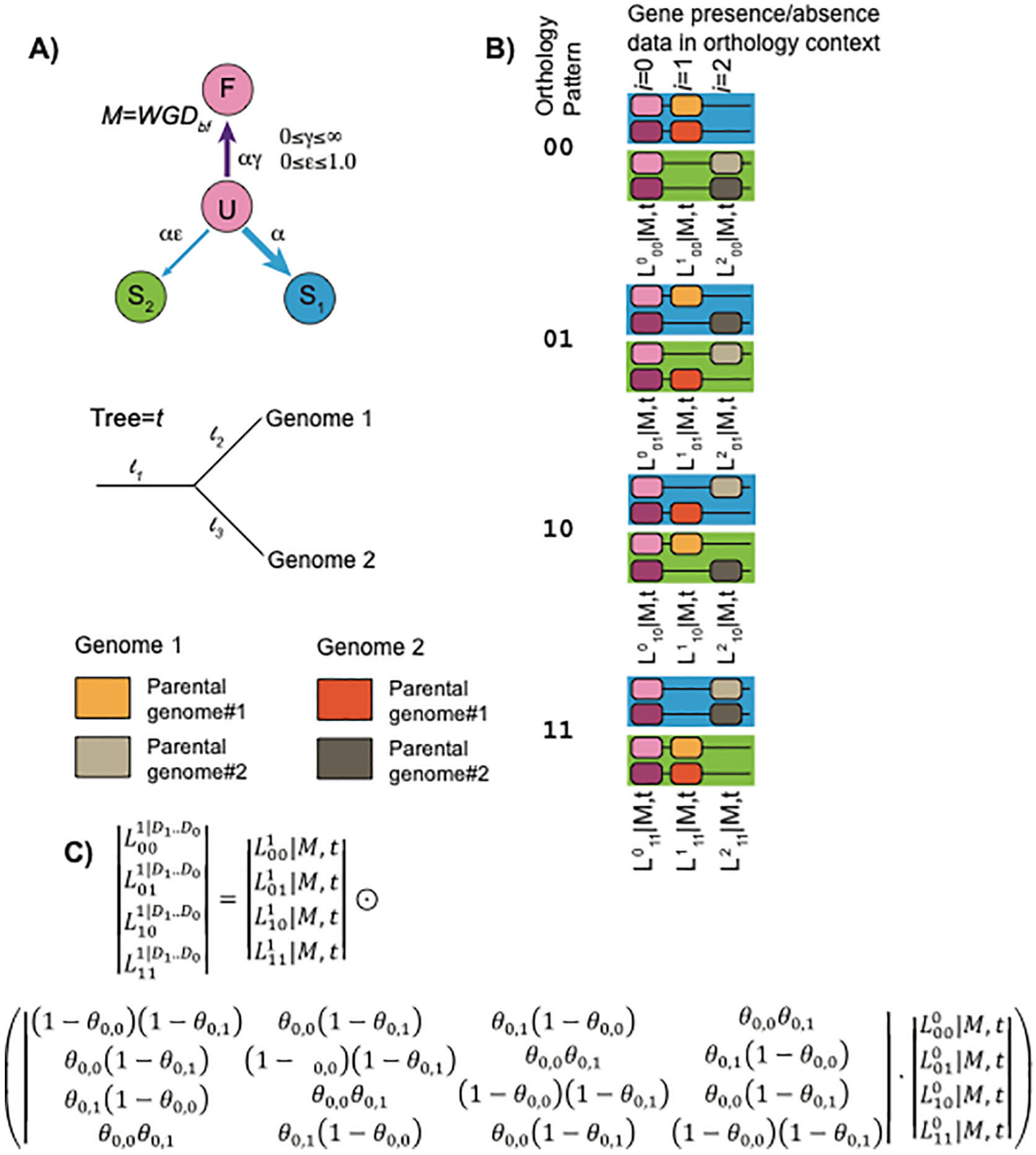
The question of demonstrating that a particular genome has an ancient polyploidy in its history is a complex one[111] and somewhat distinct from our concerns here. However, one obvious consequence of a polyploidy is the production of a group of duplicate genes that were all formed “at the same moment.” In principle, a neutral measure of paralog divergence, such as the number of synonymous substitutions per synonymous site (K_s) should be able to detect a polyploidy through the excess of duplicates with similar K_S values.[9,47] While the actual practice of detecting polyploidy events in this way requires care,[112] it has been an extremely illuminating approach. For instance, in a pioneering study, Maere *et al*. (2005),[113] were able to fit a mixture of age models to the *Arabidopsis thaliana* genome and detect three ancient polyploidies in its history. They further showed convergent retention of genes of similar function in duplicate after these events.[113] To do so, they modeled three distinct processes: 1) a basal continuous rate of single gene duplication, 2) a set of between one and three ancient polyploidy events and 3) continuous losses of duplicates created by processes 1 and 2. They evolved these three processes in simulated discrete K_s time intervals and fit the simulations to the observed set of K_s values from duplicate genes found in the *A. thaliana* genome.

Maere *et al*.’s (2005)[113] approach is elegant but challenging to implement: other analyses of a similar form have instead fit mixtures of models to the observed duplicate divergences, combining a basal steady-state duplicate birth-death model with one or more discrete events duplicating the entire genome (with the potential for the immediate removal of some of these duplicates).[112,114] Such approaches allow for testing hypotheses regarding the number of polyploidy events in the



	0.891	0.919	0.993	0.992	0.993	0.892	0.986	0.994	0.999	0.999
T. rubripes Prefix: ENSTRUG0000	0009500	0009320		0009069	0008708	0008090	0005608		0005452	0005389
T. nigroviridis Prefix: ENSTNIG0000	0017065			0017062	0017061	0017060	0017056		0017055	0019596
G. aculeatus Prefix: ENSGACG0000	0018449	0018450		0018455	0018457	0018461	0018480		0018483	0018484
X. maculatus Prefix: ENSXMAG0000		0010351		0019510	0010375	0010396	0010477		0010505	0019511
O. latipes Prefix: ENSORLG0000	0007455	0007471		0007484	0007524	0007590	0007719		0007735	0007738
O. niloticus Prefix: ENSONIG0000	0017658	0017655		0020577	0017645	0017632	0017606		0017600	0020576
A. mexicanus Prefix: ENSAMXG0000	0001202	0001200		0026034	0001194		0007469		0000252	0025373
D. rerio Prefix: ENSDARG0000	0025535	0099849		0088124	0089917	0088168	0021352		0075048	0016318
L. oculatus Prefix: ENSLOCG0000	0009271	0009282	0009305	0009352	0009362	0009374	0009807	0009869	0009823	0018303
T. rubripes	0013748	0013805	0014010	0014131			0000702	0000906		
T. nigroviridis	0016147	0016148	0016150	0016152		0018235	0007666	0007667		
G. aculeatus	0020465	0020458	0020456	0020453		0020452		0001755		
X. maculatus	0008223	0000489	0000475	0000470		0000456		0014949		
O. latipes	0002460	0002342	0002261	0002214				0018454		
O. niloticus	0008668	0008631	0008623	0008611			0000348	0000345		
A. mexicanus			0017277	0025830		0017208	0007424	0007442		
D. rerio	0089914	0069196	0044492	0054510			0032714	0059365		

Figure 2. A region of ten ancestral genes duplicated through the teleost-specific genome duplication (TGD).[125] Shown in the center in gray are the ten genes as they are found in the genome of the spotted gar (*L. oculatus*), which lacks the TGD. The paralogous regions created by the TGD in the eight genomes possessing it are then shown above and below the gar genes. The lines joining pairs of genes indicate that these genes are neighbors in the genome (i.e., they are in synteny). After the TGD, some duplicates survive in all (pink) or some (tan) genomes, while others have been returned to single copy, either from the subgenome with more surviving genes (blue) or than with fewer (green). Numbers at the top of each column/pillar are the orthology confidence estimates from POInT. In other words, this figure gives the confidence for placing the genes in this orthology state relative to the other 2^8-1=255 orthology configurations. Genes are shown with their Ensemble identifiers[126] for reference. This figure is an original figure produced by the authors for this review article.



$$\begin{vmatrix} L_{00}^{1|D_1..D_0} \\ L_{01}^{1|D_1..D_0} \\ L_{10}^{1|D_1..D_0} \\ L_{11}^{1|D_1..D_0} \end{vmatrix} = \begin{vmatrix} L_{00}^{1}|M,t \\ L_{01}^{1}|M,t \\ L_{10}^{1}|M,t \\ L_{11}^{1}|M,t \end{vmatrix} \odot$$

$$\left(\begin{vmatrix} (1-\theta_{0,0})(1-\theta_{0,1}) & \theta_{0,0}(1-\theta_{0,1}) & \theta_{0,1}(1-\theta_{0,0}) & \theta_{0,0}\theta_{0,1} \\ \theta_{0,0}(1-\theta_{0,1}) & (1-\ _{0,0})(1-\theta_{0,1}) & \theta_{0,0}\theta_{0,1} & \theta_{0,1}(1-\theta_{0,0}) \\ \theta_{0,1}(1-\theta_{0,0}) & \theta_{0,0}\theta_{0,1} & (1-\theta_{0,0})(1-\theta_{0,1}) & \theta_{0,0}(1-\theta_{0,1}) \\ \theta_{0,0}\theta_{0,1} & \theta_{0,1}(1-\theta_{0,0}) & \theta_{0,0}(1-\theta_{0,1}) & (1-\theta_{0,0})(1-\theta_{0,1}) \end{vmatrix} \cdot \begin{vmatrix} L_{00}^{0}|M,t \\ L_{01}^{0}|M,t \\ L_{10}^{0}|M,t \\ L_{11}^{0}|M,t \end{vmatrix} \right)$$

Figure 3. Modeling duplicate gene loss after polyploidy. A) Following Lewis (2001),[122] a discrete state model *M* allows an ancestral position to be duplicated (**D**), single copy (**S$_1$** or **S$_2$**) or a fixed duplicate (**D$_f$**). Transitions between these states occur at rates proportional to model parameters α, ε, and γ. Losses occur along an assumed phylogenetic tree *t* with branch lengths $l_1..l_t$. The extant genomes are phased into a series of homologous columns or *pillars*: each genome may have one or two homologs present at a pillar (a state for complete homolog absence will be added to future versions of POInT). Different parental subgenomes within an extant genome can be distinguished (orange verses tan) but subgenome identities between the genomes are unknown. **B)** For *N = 2* polyploid genomes, there are 2^N possible orthology relations. At each pillar *i*, we can compute the likelihood of the observed gene presence and absence data for a given orthology pattern XX using the model *M* and the tree *t:* $L^i_{xx}|M,t$. **C)** Using the synteny relationships, the values $L^i_{00}|M,t$ .. $L^i_{11}|M,t$ can be conditioned on $L^{i-1}_{00}|M,t$ .. $L^{i-1}_{11}|M,t$ with a transition probability matrix Θ. The elements of Θ depend on $\Theta_{i,g}$, where *i* is the pillar number and *g* is the genome. If synteny in maintained between pillars *i* and *i+1* for genome *g,* $\Theta_{i,g} = \Theta^M$, a global constant estimated by maximum likelihood ($0 \le \Theta^M \le 1$). Otherwise $\Theta_{i,g}$=0.5, meaning the orthology pattern at *i* is independent of that at *i-1.* This equation can be applied recursively to compute the likelihood of the entire dataset with standard hidden Markov model approaches[123]: the ⊙ operator represents an element-wise vector product. The tree branch lengths and model parameters are estimated from the data by maximum likelihood using standard numerical techniques.[127] This figure is an original figure produced by the authors for this review article.

lineage of a genome, but the results require some caution in their interpretation due to the relatively modest information provided by K_s values.[112]

Polyploidy and Gene Synteny

Our discussion so far has considered signals such as gene tree topologies and divergence measures which are applicable to all types of duplicate genes. However, for the specific case of a polyploidy, another type of highly informative data is present: the gene order patterns among the duplicated and non-duplicated genes. These patterns are commonly referred to as gene *synteny.* They were critical in identifying the first ancient polyploidy found in a eukaryotic genome[115] and have been used in many subsequent analyses of polyploid genomes.[116–118] Figure 2 illustrates the principle that a WGD, in contrast to SSDs, produces duplicate genes that preserve the gene order present in the unduplicated ancestor. Indeed, these patterns can identify ancient polyploidies even in the limiting case where all the duplicate genes were subsequently lost, provided that an outgroup genome lacking the polyploidy is available and the degree of chromosomal rearrangement is not too large. Hence, synteny is often considered the best evidence of the presence of an ancient polyploidy, even if formal tests using it are hard to develop.[112,119]

One example of the power of combining syntenic information with models of duplicate gene gain and loss can be seen with POInT (the Polyploid Orthology Inference Tool)[120,121] Assuming that the duplicate products of a WGD are known through syntenic information (as in Figure 2), one can use the generic discrete character evolutionary model of Paul Lewis (2001)[122] to model the preservation or loss of duplicate copies in different genomes that share this WGD (Figure 3A). Briefly, the loss model presumes that all loci (or *pillars*) start in a duplicated state **D** and then can undergo fates such as loss (resulting in states $\mathbf{S_1}$ or $\mathbf{S_2}$) or fixation ($\mathbf{D_f}$). Such a model can be applied to the duplicate presence and absence data for a group of genomes sharing the polyploidy. However, the difficulty arises that the orthology assignments between those genomes are unknown. POInT hence computes the likelihood of the observed gene presence/absence data at each pillar for all possible orthology relationship under a duplicate loss model. It then uses a hidden Markov model to condition that set of likelihoods at the current pillar *i* on those from pillars 0..*i-1-* using a transition matrix Q.[123] The elements of Q are determined by whether or not synteny is preserved between *i-1* and *i* in each genome.[109]

Once such a framework is in place, standard likelihood ratio tests[124] can be used to test hypotheses about the evolution of polyploid genomes such as what fraction of the duplicates appear to have been fixed[120] or whether one of the two parental genomes from an allopolyploid is favored when duplicates are lost.[121] This second pattern, termed *biased fractionation*, is likely indicative of an allopolyploidy[108] and raises questions as to whether the subgenomes of allopolyploid hybrids are functionally compatible.[79]

Concluding Thoughts

A number of models have been generated that describe different levels of duplicate gene retention with different levels of mechanistic detail and as standalone models for individual problems, or as models that are integrated with other problems. These models are summarized in Table 1. Two key elements of duplicate gene retention are coding sequence function modeled using summary statistics (like dN/dS) or Markov models describing increasing layers of complexity and expression evolution. These can also include models for syntenic position. While some of the Markov models attempt to integrate the two layers of evolution reflecting coding sequence function and expression in a sophisticated manner, other approaches either use a simpler unifying factor, like [P] in a biophysical model or treat them independently or without differential specification. Work in these directions is making substantial progress in capturing biological realism. Modeling of duplicate gene retention can converge with the broader modeling frameworks being developed for the genotype-phenotype map. While much of this modeling is in the realm of additive statistical association, the field of computational systems biology includes modeling frameworks that add another layer to the genotype-phenotype map that have not been touched much at the boundaries of the duplicate gene retention modeling field. Mechanistic models for gene expression evolution will also be fruitful in this field. There is a lot of room to keep expanding these modeling frameworks as genomic and other omic data accumulate for species and underlying populations and as biological domain-specific modeling improves that can improve mechanistic duplicate gene retention models.

Data availability

No data are associated with this article.

References

1. Katju V, Lynch M: **The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome.** *Genetics.* 2003 Dec; **165**(4): 1793–1803.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Katju V, Lynch M: **On the formation of novel genes by duplication in the Caenorhabditis elegans genome.** *Mol. Biol. Evol.* 2006 May; **23**(5): 1056–1067.
PubMed Abstract | Publisher Full Text

3. Cusack BP, Wolfe KH: **Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates.** *Mol. Biol. Evol.* 2007 Mar; **24**(3): 679–686.
PubMed Abstract | Publisher Full Text

4. Hakes L, Pinney JW, Lovell SC, *et al.*: **All duplicates are not equal: the difference between small-scale and genome duplication.** *Genome Biol.* 2007; **8**(10): R209.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Cardoso-Moreira M, Arguello JR, Gottipati S, *et al.*: **Evidence for the fixation of gene duplications by positive selection in Drosophila.** *Genome Res.* 2016 Jun; **26**(6): 787–798.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Rogers RL, Shao L, Thornton KR: **Tandem duplications lead to novel expression patterns through exon shuffling in Drosophila yakuba.** *PLoS Genet.* 2017 May; **13**(5): e1006795.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Konrad A, Flibotte S, Taylor J, *et al.*: **Mutational and transcriptional landscape of spontaneous gene duplications and deletions in *Caenorhabditis elegans*.** *Proc. Natl. Acad. Sci. U S A.* 2018 Jul 10; **115**(28): 7386–7391.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Wang Z, Dong X, Ding G, *et al.*: **Comparing the retention mechanisms of tandem duplicates and retrogenes in human and mouse genomes.** *Genet. Sel. Evol*. 2010; **42**: 24.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science.* 2000 Nov 10; **290**(5494): 1151–1155.
PubMed Abstract | Publisher Full Text

10. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics.* 2000 Jan; **154**(1): 459–473.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Arvestad L, Lagergren J, Sennblad B: **The gene evolution model and computing its associated probabilities.** *J. ACM.* 2009 Apr 17; **56**(2): 1–44.
Publisher Full Text

12. Konrad A, Teufel AI, Grahnen JA, *et al.*: **Toward a general model for the evolutionary dynamics of gene duplicates.** *Genome Biol. Evol.* 2011 Sep 13; **3**: 1197–209.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Zhang P, Min W, Li WH: **Different age distribution patterns of human, nematode, and Arabidopsis duplicate genes.** *Gene.* 2004 Nov 24; **342**(2): 263–268.
Publisher Full Text

14. Hughes T, Liberles DA: **The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation.** *J. Mol. Evol.* 2007 Nov; **65**(5): 574–588.
PubMed Abstract | Publisher Full Text

15. Wagner A: **Energy constraints on the evolution of gene expression.** *Mol. Biol. Evol.* 2005 Jun; **22**(6): 1365–1374.
Publisher Full Text

16. Ohno S: *Evolution by Gene Duplication.* Berlin Heidelberg: Springer-Verlag; 1970; 160.

17. Perry GH, Dominy NJ, Claw KG, *et al.*: **Diet and the evolution of human amylase gene copy number variation.** *Nat. Genet.* 2007 Oct; **39**(10): 1256–1260.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Reis-Cunha JL, Valdivia HO, Bartholomeu DC: **Gene and Chromosomal Copy Number Variations as an Adaptive Mechanism Towards a Parasitic Lifestyle in Trypanosomatids.** *Curr. Genomics.* 2018 Feb; **19**(2): 87–97.
PubMed Abstract | Publisher Full Text

19. Jan-Peter D, Keith G, Bill W: **Cell Biology of the Trypanosome Genome.** *Microbiol. Mol. Biol. Rev.* 2010 Dec 1; **74**(4): 552–569.
Publisher Full Text

20. Freeling M, Thomas BC: **Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity.** *Genome Res.* 2006 Jul; **16**(7): 805–814.
Publisher Full Text

21. Hughes T, Ekman D, Ardawatia H, *et al.*: **Evaluating dosage compensation as a cause of duplicate gene retention in Paramecium tetraurelia.** *Genome Biol.* 2007; **8**(5): 213.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Freeling M: **Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition.** *Annu. Rev. Plant Biol.* 2009; **60**: 433–453.
Publisher Full Text

23. Birchler JA, Veitia RA: **Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines.** *Proc. Natl. Acad. Sci. U. S. A.* 2012 Sep 11; **109**(37): 14746–14753.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Gillard GB, Grønvold L, Røsæg LL, *et al.*: **Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication.** *Genome Biol.* 2021 Apr 13; **22**(1): 103.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Force A, Lynch M, Pickett FB, *et al.*: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics.* 1999 Apr; **151**(4): 1531–1545.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Stoltzfus A: **On the possibility of constructive neutral evolution.** *J. Mol. Evol.* 1999 Aug; **49**(2): 169–181.
Publisher Full Text

27. Davis JC, Petrov DA: **Preferential duplication of conserved proteins in eukaryotic genomes.** *PLoS Biol.* 2004 Mar; **2**(3): E55.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Yang J, Lusk R, Li WH: **Organismal complexity, protein complexity, and gene duplicability.** *Proc. Natl. Acad. Sci. U. S. A.* 2003 Dec 23; **100**(26): 15661–15665.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Roux J, Liu J, Robinson-Rechavi M: **Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates.** *Mol. Biol. Evol.* 2017 Nov 1; **34**(11): 2773–2791.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Wang L, Ma H, Lin J: **Angiosperm-Wide and Family-Level Analyses of AP2/ERF Genes Reveal Differential Retention and Sequence Divergence After Whole-Genome Duplication.** *Front. Plant Sci.* 2019 Feb 26; **10**: 196.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Woods S, Coghlan A, Rivers D, *et al.*: **Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses.** *PLoS Genet.* 2013 May; **9**(5): e1003330.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Geiser C, Mandáková T, Arrigo N, *et al.*: **Repeated Whole-Genome Duplication, Karyotype Reshuffling, and Biased Retention of Stress-Responding Genes in Buckler Mustard.** *Plant Cell.* 2016 Jan; **28**(1): 17–27.
PubMed Abstract | Publisher Full Text | Free Full Text

33. McGrath CL, Gout JF, Johri P, *et al.*: **Differential retention and divergent resolution of duplicate genes following whole-genome duplication.** *Genome Res.* 2014 Oct; **24**(10): 1665–1675.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Li Z, Defoort J, Tasdighian S, *et al.*: **Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms.** *Plant Cell.* 2016 Feb; **28**(2): 326–344.
PubMed Abstract | Publisher Full Text | Free Full Text

35. De Smet R, Adams KL, Vandepoele K, *et al.*: **Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants.** *Proc. Natl. Acad. Sci. U. S. A.* 2013 Feb 19; **110**(8): 2898–2903.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Ascencio D, Diss G, Gagnon-Arsenault I, *et al.*: **Expression attenuation as a mechanism of robustness against gene duplication.** *Proc. Natl. Acad. Sci. U. S. A.* 2021 Feb 9; **118**(6): e2014345118.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Aury JM, Jaillon O, Duret L, *et al.*: **Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia.** *Nature.* 2006 Nov 9; **444**(7116): 171–178.
PubMed Abstract | Publisher Full Text

38. Qian W, Zhang J: **Gene dosage and gene duplicability.** *Genetics.* 2008 Aug; **179**(4): 2319–2324.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Kuzmin E, Taylor JS, Boone C: **Retention of duplicated genes in evolution.** *Trends Genet.* 2022 Jan; **38**(1): 59–72.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Zhang Z, Luo ZW, Kishino H, *et al.*: **Divergence pattern of duplicate genes in protein-protein interactions follows the power law.** *Mol. Biol. Evol.* 2005 Mar; **22**(3): 501–505.
PubMed Abstract | Publisher Full Text

41. Wilson AE, Liberles DA: **Expectations of duplicate gene retention under the gene duplicability hypothesis.** *BMC Ecol Evol.* 2023 Dec 14; **23**(1): 76.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

42. Lien S, Koop BF, Sandve SR, *et al.*: **The Atlantic salmon genome provides insights into rediploidization.** *Nature.* 2016 May 12; **533**(7602): 200–205.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

43. Schnable JC, Wang X, Pires JC, *et al.*: **Escape from preferential retention following repeated whole genome duplications in plants.** *Front. Plant Sci.* 2012 May 15; **3**: 94.
    Publisher Full Text

44. Hermansen RA, Hvidsten TR, Sandve SR, *et al.*: **Extracting functional trends from whole genome duplication events using comparative genomics.** *Biol Proced Online.* 2016 May 10; **18**: 11.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

45. Henry CN, Piper K, Wilson AE, *et al.*: **WGDTree: a phylogenetic software tool to examine conditional probabilities of retention following whole genome duplication events.** *BMC Bioinformatics.* 2022 Nov 24; **23**(1): 505.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

46. Gu Z, Nicolae D, Lu HHS, *et al.*: **Rapid divergence in expression between duplicate genes inferred from microarray data.** *Trends Genet.* 2002 Dec; **18**(12): 609–613.
    PubMed Abstract | **Publisher Full Text**

47. Blanc G, Wolfe KH: **Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes.** *Plant Cell.* 2004 Jul; **16**(7): 1667–1678.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

48. Wagner A: **Asymmetric functional divergence of duplicate genes in yeast.** *Mol. Biol. Evol.* 2002 Oct; **19**(10): 1760–1768.
    PubMed Abstract | **Publisher Full Text**

49. Assis R, Bachtrog D: **Neofunctionalization of young duplicate genes in Drosophila.** *Proc. Natl. Acad. Sci. U. S. A.* 2013 Oct 22; **110**(43): 17409–17414.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

50. Perry BR, Assis R: **CDROM: Classification of Duplicate gene RetentiOn Mechanisms.** *BMC Evol. Biol.* 2016 Apr 14; **16**: 82.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

51. Assis R, Bachtrog D: **Rapid divergence and diversification of mammalian duplicate gene functions.** *BMC Evol. Biol.* 2015 Jul 15; **15**: 138.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

52. Chau LM, Goodisman MAD: **Gene duplication and the evolution of phenotypic diversity in insect societies.** *Evolution.* 2017 Dec; **71**(12): 2871–2884.
    PubMed Abstract | **Publisher Full Text**

53. Jiang X, Assis R: **Rapid functional divergence after small-scale gene duplication in grasses.** *BMC Evol. Biol.* 2019 May 2; **19**(1): 97.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

54. Yanai I, Graur D, Ophir R: **Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control.** *OMICS.* 2004 Spring; **8**(1): 15–24.
    PubMed Abstract | **Publisher Full Text**

55. Felsenstein J: **Phylogenies and the Comparative Method.** *Am. Nat.* 1985 Jan 1; **125**(1): 1–15.
    Publisher Full Text

56. Hansen TF: **Stabilizing Selection and the Comparative Analysis of Adaptation.** *Evolution.* 1997 Oct; **51**(5): 1341–1351.
    Publisher Full Text

57. Butler MA, King AA: **Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution.** *Am. Nat.* 2004 Dec; **164**(6): 683–695.
    Publisher Full Text

58. Bedford T, Hartl DL: **Optimization of gene expression by natural selection.** *Proc. Natl. Acad. Sci. U. S. A.* 2009 Jan 27; **106**(4): 1133–1138.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

59. Schraiber JG, Mostovoy Y, Hsu TY, *et al.*: **Inferring evolutionary histories of pathway regulation from transcriptional profiling data.** *PLoS Comput. Biol.* 2013 Oct 10; **9**(10): e1003255.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

60. Kalinka AT, Varga KM, Gerrard DT, *et al.*: **Gene expression divergence recapitulates the developmental hourglass model.** *Nature.* 2010 Dec 9; **468**(7325): 811–814.
    PubMed Abstract | **Publisher Full Text**

61. Brawand D, Soumillon M, Necsulea A, *et al.*: **The evolution of gene expression levels in mammalian organs.** *Nature.* 2011 Oct 19; **478**(7369): 343–348.
    Publisher Full Text

62. Rohlfs RV, Harrigan P, Nielsen R: **Modeling gene expression evolution with an extended Ornstein--Uhlenbeck process accounting for within-species variation.** *Mol. Biol. Evol.* 2014; **31**(1): 201–211.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

63. Uyeda JC, Harmon LJ: **A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data.** *Syst. Biol.* 2014 Nov; **63**(6): 902–918.
    PubMed Abstract | **Publisher Full Text**

64. Rohlfs RV, Nielsen R: **Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution.** *Syst. Biol.* 2015 Sep; **64**(5): 695–708.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

65. Khabbazian M, Kriebel R, Rohe K, *et al.*: **Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models.** *Methods Ecol. Evol.* 2016 Jul; **7**(7): 811–824.
    Publisher Full Text

66. Panigrahi A, O'Malley BW: **Mechanisms of enhancer action: the known and the unknown.** *Genome Biol.* 2021 Apr 15; **22**(1): 108.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

67. Wittkopp PJ, Haerum BK, Clark AG: **Evolutionary changes in cis and trans gene regulation.** *Nature.* 2004 Jul 1; **430**(6995): 85–88.
    Publisher Full Text

68. Karollus A, Mauermeier T, Gagneur J: **Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers.** *Genome Biol.* 2023 Mar 27; **24**(1): 56.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

69. DeGiorgio M, Assis R: **Learning Retention Mechanisms and Evolutionary Parameters of Duplicate Genes from Their Expression Data.** *Mol. Biol. Evol.* 2021 Mar 9; **38**(3): 1209–1224.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

70. Perry GH, Melsted P, Marioni JC, *et al.*: **Comparative RNA sequencing reveals substantial genetic variation in endangered primates.** *Genome Res.* 2012 Apr; **22**(4): 602–610.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

71. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second ed. New York: Springer; 2009; 757.

72. Li WH: **Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes.** *Genetics.* 1980 May; **95**(1): 237–258.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

73. Wagner A: **Energy costs constrain the evolution of gene expression.** *J. Exp. Zool. B Mol. Dev. Evol.* 2007 May 15; **308**(3): 322–324.
    PubMed Abstract | **Publisher Full Text**

74. Charlesworth B: **Effective population size and patterns of molecular evolution and variation.** *Nat. Rev. Genet.* 2009 Mar; **10**(3): 195–205.
    Publisher Full Text

75. Stoebel DM, Dean AM, Dykhuizen DE: **The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products.** *Genetics.* 2008 Mar; **178**(3): 1653–1660.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

76. Wilson AE, Liberles DA: **Dosage balance acts as a time-dependent selective barrier to subfunctionalization.** *BMC Ecol Evol.* 2023 May 3; **23**(1): 14.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

77. Teufel AI, Liu L, Liberles DA: **Models for gene duplication when dosage balance works as a transition state to subsequent neo- or sub-functionalization.** *BMC Evol. Biol.* 2016 Feb 20; **16**: 45.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

78. Session AM, Uno Y, Kwon T, *et al.*: **Genome evolution in the allotetraploid frog Xenopus laevis.** *Nature.* 2016 Oct 20; **538**(7625): 336–343.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

79. Hao Y, Mabry ME, Edger PP, *et al.*: **The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible.** *Genome Res.* 2021 May; **31**(5): 799–810.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

80. Cai X, Chang L, Zhang T, *et al.*: **Impacts of allopolyploidization and structural variation on intraspecific diversification in Brassica rapa.** *Genome Biol.* 2021 May 31; **22**(1): 166.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

81. Stark TL, Kaufman RS, Maltepes MA, *et al.*: **Detecting Selection on Segregating Gene Duplicates in a Population.** *J. Mol. Evol.* 2021 Oct; **89**(8): 554–564.
    PubMed Abstract | **Publisher Full Text**

82. Platt A, Pivirotto A, Knoblauch J, *et al.*: **An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations.**

PLoS Genet. 2019 Aug; **15**(8): e1008340.
PubMed Abstract | Publisher Full Text | Free Full Text

83. McDonald JH, Kreitman M: **Adaptive protein evolution at the Adh locus in Drosophila.** *Nature.* 1991 Jun 20; **351**(6328): 652–654.
PubMed Abstract | Publisher Full Text

84. Lynch M, O'Hely M, Walsh B, *et al.*: **The probability of preservation of a newly arisen gene duplicate.** *Genetics.* 2001 Dec; **159**(4): 1789–1804.
PubMed Abstract | Publisher Full Text | Free Full Text

85. Zhao J, Teufel AI, Liberles DA, *et al.*: **A generalized birth and death process for modeling the fates of gene duplication.** *BMC Evol. Biol.* 2015 Dec 8; **15**: 275.
PubMed Abstract | Publisher Full Text | Free Full Text

86. Yohe LR, Liu L, Dávalos LM, *et al.*: **Protocols for the Molecular Evolutionary Analysis of Membrane Protein Gene Duplicates.** *Methods Mol. Biol.* 2019; **1851**: 49–62.
Publisher Full Text

87. Rasmussen MD, Kellis M: **A Bayesian approach for fast and accurate gene tree reconstruction.** *Mol. Biol. Evol.* 2011 Jan; **28**(1): 273–290.
Publisher Full Text

88. Górecki P, Eulenstein O: **DrML: probabilistic modeling of gene duplications.** *J. Comput. Biol.* 2014 Jan; **21**(1): 89–98.
PubMed Abstract | Publisher Full Text | Free Full Text

89. Parey E, Louis A, Montfort J, *et al.*: **An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication.** *Genome Res.* 2022; **32**: 1685–1697.
PubMed Abstract | Publisher Full Text | Free Full Text

90. Nakhleh L: **Computational approaches to species phylogeny inference and gene tree reconciliation.** *Trends Ecol. Evol.* 2013 Dec; **28**(12): 719–728.
PubMed Abstract | Publisher Full Text | Free Full Text

91. Stark TL, Liberles DA, Holland BR, *et al.*: **Analysis of a mechanistic Markov model for gene duplicates evolving under subfunctionalization.** *BMC Evol. Biol.* 2017 Jan 31; **17**(1): 38.
PubMed Abstract | Publisher Full Text | Free Full Text

92. Stark T: **Markov models for the evolution of duplicate genes, and microsatellites [PhD Thesis]. University of Tasmania.** 2017.

93. Diao J, O'Reilly M, M, Holland B.: **A subfunctionalisation model of gene family evolution predicts balanced tree shapes.** *Mol. Phylogenet. Evol.* 2022 Nov; **176**: 107566.
PubMed Abstract | Publisher Full Text

94. Diao J, Stark TL, Liberles DA, *et al.*: **Level-dependent QBD models for the evolution of a family of gene duplicates.** *Stoch. Model.* 2020 Apr 2; **36**(2): 285–311.
Publisher Full Text

95. Soewongsono AC, Diao J, Tristan ST, *et al.*: **Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation.** Completed Manuscript. arXiv preprint arXiv: 2309.0644. 2023.

96. Conant GC, Wagner A: **Asymmetric sequence divergence of duplicate genes.** *Genome Res.* 2003 Sep; **13**(9): 2052–2058.
PubMed Abstract | Publisher Full Text | Free Full Text

97. Seoighe C, Scheffler K: **Very Low Power to Detect Asymmetric Divergence of Duplicated Genes.** *Comparative Genomics.* Berlin Heidelberg: Springer; 2005; p. 142–52.

98. Hahn MW: **Distinguishing among evolutionary models for the maintenance of gene duplicates.** *J. Hered.* 2009 Jul 13; **100**(5): 605–617.
PubMed Abstract | Publisher Full Text

99. Evangelisti AM, Conant GC: **Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling.** *Genome Biol. Evol.* 2010 Oct 21; **2**: 826–34.
PubMed Abstract | Publisher Full Text | Free Full Text

100. Li WH: *Molecular Evolution.* Sunderland, MA: Sinauer Associates; 1997; 432.

101. Ji X, Griffing A, Thorne JL: **A Phylogenetic Approach Finds Abundant Interlocus Gene Conversion in Yeast.** *Mol. Biol. Evol.* 2016 Sep; **33**(9): 2469–2476.
PubMed Abstract | Publisher Full Text | Free Full Text

102. Gao LZ, Innan H: **Very low gene duplication rate in the yeast genome.** *Science.* 2004 Nov 19; **306**(5700): 1367–1370.
PubMed Abstract | Publisher Full Text

103. Mayfield-Jones D, Washburn JD, Arias T, *et al.*: **Watching the grin fade: tracing the effects of polyploidy on different evolutionary time scales.** *Semin. Cell Dev. Biol.* 2013 Apr; **24**(4): 320–331.
PubMed Abstract | Publisher Full Text

104. Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, *et al.*: **Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose *et al.* ().** *New Phytol.* 2014 Jun; **202**(4): 1105–1117.
PubMed Abstract | Publisher Full Text

105. Van de Peer Y, Mizrachi E, Marchal K: **The evolutionary significance of polyploidy.** *Nat. Rev. Genet.* 2017 Jul; **18**(7): 411–424.
Publisher Full Text

106. Otto SP: **The evolutionary consequences of polyploidy.** *Cell.* 2007 Nov 2; **131**(3): 452–462.
Publisher Full Text

107. Barker MS, Arrigo N, Baniaga AE, *et al.*: **On the relative abundance of autopolyploids and allopolyploids.** *New Phytol.* 2016 Apr; **210**(2): 391–398.
PubMed Abstract | Publisher Full Text

108. Garsmeur O, Schnable JC, Almeida A, *et al.*: **Two evolutionarily distinct classes of paleopolyploidy.** *Mol. Biol. Evol.* 2014 Feb; **31**(2): 448–454.
PubMed Abstract | Publisher Full Text

109. Hao Y, Fleming J, Petterson J, *et al.*: **Convergent evolution of polyploid genomes from across the eukaryotic tree of life.** *G3.* 2022 May 30; **12**(6)
PubMed Abstract | Publisher Full Text | Free Full Text

110. Scannell DR, Frank AC, Conant GC, *et al.*: **Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication.** *Proc. Natl. Acad. Sci. U. S. A.* 2007 May 15; **104**(20): 8397–8402.
PubMed Abstract | Publisher Full Text | Free Full Text

111. Zwaenepoel A, Li Z, Lohaus R, *et al.*: **Finding Evidence for Whole Genome Duplications: A Reappraisal.** *Mol. Plant.* 2019 Feb 4; **12**(2): 133–136.
PubMed Abstract | Publisher Full Text

112. Tiley GP, Barker MS, Burleigh JG: **Assessing the Performance of Ks Plots for Detecting Ancient Whole Genome Duplications.** *Genome Biol. Evol.* 2018 Nov 1; **10**(11): 2882–2898.
PubMed Abstract | Publisher Full Text

113. Maere S, De Bodt S, Raes J, *et al.*: **Modeling gene and genome duplications in eukaryotes.** *Proc. Natl. Acad. Sci. U. S. A.* 2005 Apr 12; **102**(15): 5454–5459.
PubMed Abstract | Publisher Full Text | Free Full Text

114. Rabier CE, Ta T, Ané C: **Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach.** *Mol. Biol. Evol.* 2014 Mar; **31**(3): 750–762.
PubMed Abstract | Publisher Full Text | Free Full Text

115. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature.* 1997 Jun 12; **387**(6634): 708–713.
PubMed Abstract | Publisher Full Text

116. Byrne KP, Wolfe KH: **The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species.** *Genome Res.* 2005 Oct; **15**(10): 1456–1461.
PubMed Abstract | Publisher Full Text | Free Full Text

117. Lyons E, Freeling M: **How to usefully compare homologous plant genes and chromosomes as DNA sequences.** *Plant J.* 2008 Feb; **53**(4): 661–673.
PubMed Abstract | Publisher Full Text

118. Nakatani Y, Shingate P, Ravi V, *et al.*: **Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution.** *Nat. Commun.* 2021 Jul 23; **12**(1): 4489.
PubMed Abstract | Publisher Full Text | Free Full Text

119. Li J, Van de Peer Y, Li Z: **Inference of Ancient Polyploidy Using Transcriptome Data.** *Methods Mol. Biol.* 2023; **2545**: 47–76.
PubMed Abstract | Publisher Full Text

120. Conant GC, Wolfe KH: **Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast.** *Genetics.* 2008 Jul; **179**(3): 1681–1692.
PubMed Abstract | Publisher Full Text | Free Full Text

121. Emery M, Willis MMS, Hao Y, *et al.*: **Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization.** *PLoS Genet.* 2018 Mar; **14**(3): e1007267.
Publisher Full Text

122. Lewis PO: **A likelihood approach to estimating phylogeny from discrete morphological character data.** *Syst. Biol.* 2001 Nov-Dec; **50**(6): 913–925.
PubMed Abstract | Publisher Full Text

123. Felsenstein J, Churchill GA: **A Hidden Markov Model approach to variation among sites in rate of evolution.** *Mol. Biol. Evol.*

1996 Jan; **13**(1): 93–104.
PubMed Abstract | Publisher Full Text

124. Sokal RR, Rohlf FJ: *Biometry.* 3rd ed. New York: W. H. Freeman and Company; 1995.

125. Conant GC: **The lasting after-effects of an ancient polyploidy on the genomes of teleosts.** *PLoS One.* 2020 Apr 16; **15**(4): e0231356.
PubMed Abstract | Publisher Full Text | Free Full Text

126. Aken BL, Achuthan P, Akanni W, *et al.*: **Ensembl 2017.** *Nucleic Acids Res.* 2017 Jan 4; **45**(D1): D635–D642.
PubMed Abstract | Publisher Full Text | Free Full Text

127. Press WH, Teukolsky SA, Vetterling WA, *et al.*: *Numerical Recipes in C.* New York: Cambridge University Press; 1992.

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 02 January 2024

https://doi.org/10.5256/f1000research.155263.r228194

✓ **Arturo Becerra**
[1] Universidad Nacional Autónoma de México, Mexico City, Mexico
[2] Universidad Nacional Autónoma de México, Mexico City, Mexico

The authors have provided a succinct yet comprehensive overview of the primary models governing the retention of duplicate genes—a critical biological process. The manuscript is well-crafted and explores a subject of significant originality, mainly as there are few or no existing text reviews of this nature. However, I have a minor concern and a suggestion:

The text appears densely packed given the format constraints, resulting in brief coverage of specific models, such as the Dosage Imbalance Cost. While I understand the limitations imposed by the format and the challenge of expanding on each model, I propose including a table or images that summarize the characteristics of these models, both mathematically and biologically. This addition would greatly aid in providing readers with a more precise and concise understanding of the models.

Minor point.
Page 7, third paragraph, last line:
"including all vertebrates and flowering plants, as well as lineages of yeasts, ciliates and other plants."

Change by:
... including all vertebrates, flowering plants, lineages of yeasts, and ciliates.

Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Is the review written in accessible language?

Yes

Are the conclusions drawn appropriate in the context of the current research literature?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Early evolution of life

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 22 November 2023

https://doi.org/10.5256/f1000research.155263.r218598

✓ **Nadia El-Mabrouk**

[1] Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Quebec, Canada

[2] Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Quebec, Canada

This is an interesting review on the dynamic of gene retention and loss, depending on gene structure, gene function, the mechanism creating the duplicated copies, and other biological constraints. It presents various mathematical models characterizing such gene duplicates retention/loss dynamics in a context-dependent way. In particular, the difference between gene duplicates evolution depends on whether gene copies result from small scale duplication or whole genome duplication (WGD). The last part of the paper also explains the probabilistic and algorithmic ways of predicting WGD events.

My main criticism is that the review intermingles a variety of biological concepts and mathematical models, with no clear subdivision. The introduction should clearly explain the objective of the review and how it is organized. The content should be subdivided into coherent subsections. I would suggest adding a table with all the various constraints for gene duplicability and retention listed in the different sections, summarizing which of them are considered by which model/paper.

In addition, it would be interesting to discuss gene retention as a function not only of small-scale duplication versus whole-genome duplication, but also as a function of the type of small-scale events: duplication of a single gene versus duplication of larger segments, tandem duplication versus retrotransposition.

Few other comments:

- Mathematical formulae should be better written. This may be a problem linked to F1000Research's writing constraints? Writing “1-(0.5)^(f-1)” or event 1/2N is unclear. Is it 1/(2N) or (1/2)N?
- Although the target audience should be familiar with most of the biological and mathematical terms used, some are less obvious and should be explained, such as “heteromultimerization partners”.
- On page 7, correct the sentence: “The relative frequency of formation of these two types of polyploids many be approximately equal”. I guess “many” should be replaced by “may”.
- On the same page, 6th line before end: “As described, selection from various sources can also play a role.” As described where?
- First part of Figure 2: I guess the segments, for each genome, above and below that for L. oculatus correspond to the two paralogous syntenies? That should be specified.
- Caption of Figure 3: “Transitions between these states occur at rates proportional to model parameters ?, ε, and γ”. “?” should be corrected.

Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Is the review written in accessible language?

Yes

Are the conclusions drawn appropriate in the context of the current research literature?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology, phylogeny, gene/species tree reconciliation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.