Divergent subgenome evolution after allopolyploidization in African clawed frogs (*Xenopus*)

Benjamin L. S. Furman[1], Utkarsh J. Dang[2], Ben J. Evans[1], and G. Brian Golding[1]

[1]Department of Biology, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada
[2]Department of Health Outcomes and Administrative Sciences, School of Pharmacy and Pharmaceutical Sciences, Binghamton University, State University of New York, Binghamton, USA

Abstract

Whole genome duplication (WGD), the doubling of the nuclear DNA of a species, contributes to biological innovation by creating genetic redundancy. One mode of WGD is allopolyploidization, wherein each genome from two ancestral species becomes a 'subgenome' of a polyploid descendant species. The evolutionary trajectory of a duplicated gene that arises from WGD is influenced both by natural selection, that may favour redundant, new, or partitioned functions, and by gene silencing (pseudogenization). Here, we explored how these two phenomena varied over time and within allopolyploid genomes in several allotetraploid clawed frog species (*Xenopus*). Our analysis demonstrates that, across these polyploid genomes, purifying selection was greatly relaxed compared to a diploid outgroup, was asymmetric between each subgenome, and that coding regions are shorter in the subgenome with more relaxed purifying selection. As well, we found that the rate of gene loss was higher in the subgenome under weaker purifying selection and that this rate has remained relatively consistent over time after WGD. Our findings provide perspective from recently evolved vertebrates on the evolutionary forces that likely shape allopolyploid genomes on other branches of the tree of life.

keywords

Whole genome duplication, pseudogenization, allopolyploidization, *Xenopus*, purifying selection

Divergent subgenome evolution after allopolyploidization in African clawed frogs (*Xenopus*)

Abstract

Whole genome duplication (WGD), the doubling of the nuclear DNA of a species, contributes to biological innovation by creating genetic redundancy. One mode of WGD is allopolyploidization, wherein each genome from two ancestral species becomes a 'subgenome' of a polyploid descendant species. The evolutionary trajectory of a duplicated gene that arises from WGD is influenced both by natural selection, that may favour redundant, new, or partitioned functions, and by gene silencing (pseudogenization). Here, we explored how these two phenomena varied over time and within allopolyploid genomes in several allotetraploid clawed frog species (*Xenopus*). Our analysis demonstrates that, across these polyploid genomes, purifying selection was greatly relaxed compared to a diploid outgroup, was asymmetric between each subgenome, and that coding regions are shorter in the subgenome with more relaxed purifying selection. As well, we found that the rate of gene loss was higher in the subgenome under weaker purifying selection and that this rate has remained relatively consistent over time after WGD. Our findings provide perspective from recently evolved vertebrates on the evolutionary forces that likely shape allopolyploid genomes on other branches of the tree of life.

keywords

Whole genome duplication, pseudogenization, allopolyploidization, *Xenopus*, purifying selection

Introduction

Whole genome duplication (WGD) creates redundancy in genetic pathways and can lead to biological innovation (Ohno et al. 1970). For instance, WGD is thought to have contributed to phenotypic diversity in jawed vertebrates, which experienced at least two rounds of WGD (2R hypothesis; Dehal and Boore 2005), and the success of angiosperms, which experienced numerous WGD events (Fawcett et al. 2009; Jiao et al. 2011). However, despite the potential evolutionary advantages of WGD, the most common evolutionary outcome of a gene pair generated by WGD (homeologs) is that one becomes non-functional ("pseudogenization") (Lynch and Conery 2000).

One possible route for both gene copies to be retained is "neofunctionalization", where one homeolog acquires a novel function (Ohno et al. 1970). Another route involves a partitioning of ancestral function among duplicated genes, thus making both copies necessary ("subfunctionalization") (Force et al. 1999; Stoltzfus 1999). Additionally, because WGD doubles entire genetic pathways, natural selection may favor the functional persistence of gene duplicates in order to maintain the stoichiometric balance of epistatic interactions among the protein products of duplicated genes (Papp et al. 2003; Gout et al. 2010; Qian et al. 2010).

Post-WGD, genes involved in dosage sensitive functions, such as protein complexes and transcription factors, were preferentially retained in several species (Blanc and Wolfe 2004; Makino and McLysaght 2010; McGrath et al. 2014). However, analysis of ancient WGD events indicate that many of these genes will also eventually be lost, with only a small proportion of duplicates surviving over the long haul: 8% for yeast (Scannell et al. 2006); 18% for teleost WGD (Inoue et al. 2015); likely $<$ 10% for 2R vertebrate WGD (Dehal and Boore 2005); 20-30% for *Brassicaceae* WGD (Liu et al. 2014). One exception appears in *Paramecium*, in which 40-50% of its homeologous pairs have remained functional after 350 my (Aury et al. 2006; McGrath et al. 2014).

There is evidence that some genes are rapidly lost after WGD (Scannell et al. 2006; Inoue et al. 2015). This makes sense if gene copies are initially functionally redundant, and if natural selection to retain both homeologs is correspondingly weak. However, if gene dosage is important after WGD, it may take a long time for gene loss to be selectively neutral (Gout and Lynch 2015). Thus, the rate of pseudogenization could be constant or potentially increase over time (Gout et al. 2010; Gout and Lynch 2015). As well, if WGD is the result of allopolyploidization, duplicates may not be fully redundant due to divergence in lower-ploidy progenitors (Adams 2007). This divergence could also introduce biases in rates of pseudogenization between each half of an allopolyploid genome, i.e., each subgenome (Evans 2007; Comai 2000).

Rates of pseudogenization and the extent of purifying selection on homeologs are interrelated and potentially dynamic through time. Thus, to best understand the interactions of each one, a comparative approach – that uses data from multiple species in a time-calibrated phylogenetic context – is paramount (Inoue et al. 2015). To better understand these phenomena, we examined several tetraploid African clawed frog (*Xenopus*) species that are derived from a shared allotetraploid ancestor (4x=2n=36 chromosomes, where x is the number of chromosomes in a gamete from an ancestral diploid species (here, 9), and n is the number of chromosomes in an allotetraploid gamete (here, 18) (Tymowska 1991; Evans et al. 2015). Each of these allotetraploid species have two subgenomes of 9 homologous chromosome pairs each (the 'L' and the 'S' subgenome; Matsuda et al. 2015), that were inherited from different diploid ancestral species that generated the shared allotetraploid ancestor. Extensive multigene sequencing, genome analyses, and cytogenetic work has demonstrated closer intraspecific relationships than interspecific for the two subgenomes, supporting the hypothesis that one allotetraploidization event between two diploid progenitor lineages gave rise to the extant *Xenopus* allotetraploids (Tymowska 1991; Evans et al. 2004; Evans et al. 2015; Furman and Evans 2016; Session et al. 2016). Other scenarios are possible, e.g., supplementary information of Bewick et al. 2011, but one allotetraploidization event is the most parsimonious scenario. Estimates for the time for the initial allopolyploid WGD range from as recently as 17 my ago (Session et al. 2016) to between 25–65 my ago (Chain and Evans 2006; Hellsten et al. 2007; Bewick et al. 2011; Furman and Evans 2016), depending on the calibration point and analytical methods used. In one of these species, *X. laevis*, about 60% of the homeologous pairs are functional, and there exists substantial asymmetry in subgenome evolution (Session et al. 2016). For instance, the S-subgenome of *X. laevis* experienced more genomic rearrangements compared to the L, and has fewer intact and functional genes (Session et al. 2016).

Using a phylogenetic framework and sets of expressed gene sequences from Furman and Evans (2016), we explore duplicate gene evolution and pseuodgenization post-WGD in several allotetraploid *Xenopus* species. Our findings indicate that selection is substantially relaxed post-WGD, and has not returned to pre-duplicate levels. We found that the extent of this relaxation differs between the two subgenomes, with the S-subgenome having a more relaxed level of purifying selection and shorter coding sequences than the L-subgenome. Using a probabilistic model in a maximum likelihood framework (an extension of Dang et al. 2016), we also found that the rate of pseudogenization is higher in the S-subgenome across the *Xenopus* clade. Furthermore, we find that these rates have remained relatively constant over time. Our results are consistent with those of the *X. laevis* genome sequencing project (Session et al. 2016), but extend them by demonstrating that subgenome differences are a phenomena prevalent across several species in the *Xenopus* subgenus. We conclude that genome restructuring post-WGD is an ongoing feature of the *Xenopus* subgenus, drawn out over millions of years.

Methods

Homeolog Identification

Sequence data analyzed in this study was obtained from a recent phylogenetic analysis of *Xenopus* (Furman and Evans 2016). The dataset was generated from a total of six species (1 diploid, 5 allotetraploids), including previously published RNASeq data from four *Xenopus* (*X. borealis*, *X. clivii*, *X. largeni*, and *X. allofraseri*) and downloaded Unigene libraries from *X. laevis* and *Xenopus Silurana tropicalis* (Unigene database, last modified March 2013). RNASeq libraries consisted of 17–19 million reads per sample after quality trimming and transcriptomes were assembled using Trinity (Grabherr et al. 2011), with 72,000 – 97,000 unique transcripts per assembly and an average depth per transcript of 37–42x across species. A full description of the treatment of RNASeq data and assembly is available in Furman and Evans (2016).

As described in Furman and Evans (2016), we identified homeologous and orthologous sequences using a reciprocal `BLAST` approach (Altschul et al. 1997) with each species transcriptome assembly and the Unigene database of *X. laevis* to the *X. S. tropicalis* database. Multiple rounds of tree building and parsing allowed us to distinguish orthologous from homeologous sequences. We used a bioinformatic filter that required closer interspecific than intraspecific relationships among orthologous sequences to match an expectation associated with WGD preceding speciation of the allopolyploids (see Furman and Evans 2016 for full details). We analyzed only those gene

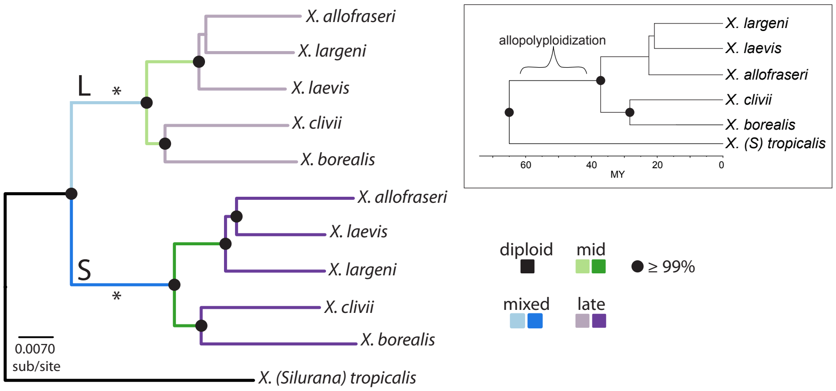

Figure 1: Phylogram recovered from a `RAxML` analysis of concatenated sequence data, with inset of species relationships from Furman and Evans (2016). L and S refer to the two subgenomes of *Xenopus* species (Matsuda et al. 2015), and the colors and corresponding labels (diploid, mixed, mid, late) refer to how branches were grouped for `Codeml` analysis and modeling of pseudogenization rates. Somewhere along the "mixed" branch an allopolyploidization event took place (indicated by an asterisk), generating a tetraploid ancestor of all extant *Xenopus*. The different resolution of sister relationships among the *X. laevis*, *X. allofraseri*, and *X. largeni* clade between the L and S-subgenomes reflect the poor resolution obtained by any analysis (Furman and Evans 2016).

alignments that had data for both homeologs for at least one species. Orthologous sequences for each species included only the longest coding region for each gene and allelic and splice variants were not analyzed. We considered only those alignments with at least 300 bp of ungapped sequence (an arbitrary cutoff to ensure enough data was present for phylogenetic analyses performed in Furman and Evans 2016). These data included a total of 1,585 genes. From these data, we realigned each using the codon aligner `MACSE` (Ranwez et al. 2011). We then removed alignments with no sequence data for *X. laevis* (a constraint of our pseudogenization model, supplemental methods section S1.3), and determined whether sequences belonged to the L or S subgenome (supplemental methods section S1), leaving a total of 1,417 genes spanning 2,235,636 base pairs (bp) (711,340 bp ungapped characters total across alignments), with a wide range of L- or S-subgenome copies missing across species (10%:*X. laevis* – 64%:*X. clivii*).

Quantifying Selective Constraint Over Time

To assess selective constraints on DNA sequences over time, we obtained estimates of ω (dN/dS) that were specific to a lineage or a group of lineages (described below), using `Codeml` (part of the `PAML` package v4.8; Yang 1997; Yang 2007). We first used a Perl script to remove any stop codons in each alignment (a requirement of `Codeml`). We then concatenated sequences across all genes for each subgenome for each species such that each allotetraploid species was represented by two concatenated sequences – one from the L- and one from the S-subgenome. If only one copy or no copies were present for a species, we inserted gaps equal to the alignment length of the missing gene. To generate a starting tree topology, we used `RAxML` v.8.2.4 (Stamatakis 2014) and set a `GTRGAMMA` model, followed by 500 bootstrap replicates to assess support. Strongly supported nodes are consistent with the phylogenetic analyses of Furman and Evans (2016), but include different relationships among *X. laevis*, *X. allofraseri*, and *X. largeni* between the L-lineage compared to the S-lineage. This is due to poorly supported, short internal branch lengths and a lower mutation rate of the L-subgenome (Fig. 1).

We used six evolutionary models to evaluate the impact of different subgenomes and time on purifying selection. The models are distinguished by the number of ω values and by the grouping schemes of branches in the phylogeny for which ω values were jointly estimated (see Fig. 2 for a visualization of all models). The simplest "ploidy" model has three separate ω (dN/dS) values that are defined by the ploidy of the branches. One ω value was estimated for the diploid branch, one was estimated for the pair of branches where the whole genome duplication took place across both subgenomes (the "mixed" lineages that have part diploid and part tetraploid histories), and one ω value was estimated for the other branches that are entirely allotetraploid.

From this simplest model accounting for just differences in ploidy, models took one of two forms. Either they tested for the effect of different subgenomes on purifying selection, or they tested the effect of time since duplication on purifying selection. A final model evaluated both of these factors together. Fig. 2 provides a visual description of these models, and the supplemental methods (section S1.2) has an in-depth description of each. Essentially, the models estimated ω values for different sections of the phylogeny separating L- and S-subgenome estimates ('subgenome' model), or by separating epochs of time along the phylogeny ('time' model), then extending each previous model by separating species clades (subgenome- or time-species models), and finally by estimating independent ω values for all divisions ('time-subgenome-species' model).

For model selection, we used the Bayesian information criterion (BIC) (Schwarz 1978):

$$\mathrm{BIC} = 2l(\hat{\boldsymbol{\Theta}}) - \log n \times m,$$

where n is the number of codons in the alignment (745,212), and $\hat{\boldsymbol{\Theta}}$ is the log-likelihood estimate of the model. m is the number of parameters estimated for each model and included nine parameters for the estimated codon frequencies (`CodonFreq = 2`), 19 parameters estimated for the branch lengths (`clock = 0`, and $2 * n - 3$, where n is the number of tips, as outlined in the PAML manual), one for the estimated value of κ (the transition/transversion ratio, `fix_kappa = 0`), and the estimated number of ω values for each model (either three, four, five, six, seven, or 11; see outlined models above and Fig. 2). Thus, m was 32, 33, 34, 35, 36, 40 for the "ploidy", "time", "subgenome", "time-species" ,"subgenome-species", and "time-subgenome-species" models, respectively (Fig. 2).

By analyzing these phylogenetic divisions with unique ω parameters, we assessed how selective constraints on duplicate copies changed over time. As well, by assigning subgenome of origin, we evaluated whether each genome that contributes to an allopolyploidization event experiences different selective constraints after allopolyploidization. For the best fit model, we estimated 95% confidence intervals for the ω parameters by analyzing 100 bootstrap replicates where codons were re-sampled with replacement. Here, missing data are expected to widen the confidence intervals on the branch specific estimates of dN/dS rates, which would limit our power to detect differences between branches.

Coding Sequence Length

To evaluate changes in selective constraints as evidenced by the evolution of premature stop codons, we tested if homeolog coding sequences were of different lengths between the subgenomes. To accomplish this, we first measured the ungapped sequence length of each of the ingroup sequences

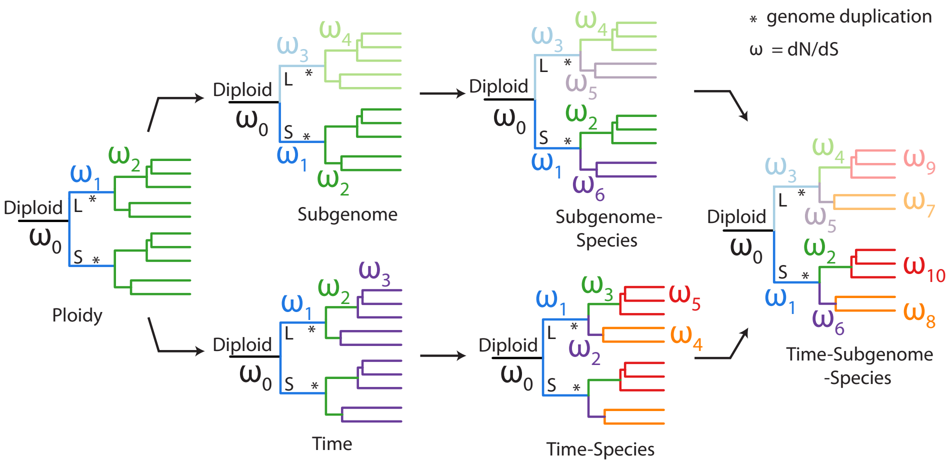

Figure 2: Model schemes for the `codeml` analysis of purifying selection. The * marks the branches where the whole genome duplication took place, thus these branches were diploid for some length of time and tetraploid for the rest. The diploid branch is the lineage that extends to *X. S. tropicalis*. Analyses were performed on unrooted trees. Colors reflect branches that have a distinct dN/dS value estimated for them.

for each alignment (using a Perl script), retaining only those that had copies from both subgenomes. We then analyzed these data using a Markov chain Monte Carlo generalized linear mixed model (MCMCglmm) using the R package `MCMCglmm` (Hadfield 2010). We set as a fixed effect the subgenome of origin (L or S), and used random effects of gene and phylogeny. The tree used in this analysis is based of the topology obtained by the `*Beast` (Heled and Drummond 2010) analysis performed by Furman and Evans (2016), and we obtained a chronogram using `mcmctree` (Yang 2007) (see supplemental methods section 1.1 for details). Using this phylogeny and the measured sequence lengths, we ran the mcmcGLMM Markov chain for 1000000 generations, with a 10000 generation burn-in and a thinning interval of 500. We set an inverse-Gamma prior for both random effects (phylogeny: $V = 1, \nu = 2$; gene: $V = 1, \nu = 0.002$) and the residual effect ($V = 1, \nu = 0.002$, and a Gaussian family link; following similar example analyses in the package documentation and in Garamszegi (2014). We modified this method to ensure adequate exploration of the posterior distribution of parameter values).

Table 1: Frequencies of the different observed states in the data. (0,1) indicates the presence of an L-subgenome homeolog and not an S homeolog; (1,0) denotes that the S homeolog is present, but not L. Less S homeologs were recovered for each species.

Species	*Dataset A* ($N = 1417$)			
	(0,0)	(0,1)	(1,0)	(1,1)
X. laevis		151	121	1145
X. largeni	370	525	372	150
X. allofraseri	314	475	397	231
X. clivii	480	470	371	96
X. borealis	346	530	404	137

Modeling Variation in the Rate of Duplicate Gene Loss Over Time

Whether polyploid genomes are rapidly reshaped after the duplication event, or if this process occurs over a protracted period of time will influence what happens to duplicate copies (e.g., whether they are preserved by rare events, such as beneficial mutations that lead to neofunctionalization). As well, the rate of pseudogenization allows for evaluation of the extent that global mechanisms of duplicate preservation, such as dosage balance on interacting gene networks, influence genome structure post-duplication. Here, to infer the rate at which genes are lost after a duplication event, we used a model-based approach. We used this model to estimate the rate of pseudogenization in different time intervals demarcated by speciation events, and subgenome specific rates.

Using a continuous time Markov chain model, we simultaneously estimate the rate of pseudogenization and the fraction of missing (or mis-recorded) data in the fashion of Dang et al. (2016). The model constructed here also accounts for substantial sampling bias in the data due to constraints imposed by Furman and Evans (2016) during dataset construction (see details below). Our model uses the number and type of gene copies present in each species in each gene alignment. An observation of (1,1) denotes the presence of two gene homeologs (L and S), (0,1) and (1,0) indicate presence of one homeolog and not the other (only L, or only S, respectively), and finally (0,0) would indicate that neither gene copy was observed (i.e., no data for a given gene in a species). Excluding the last category of no data, the observations of either one or both homeologs form a phyletic pattern of presence/absence of homeologous sequence for the species in the phylogenetic tree (Table 1). The model assumes that the tetraploid descendants of the diploid progenitors inherited both copies of all genes in the analysis (i.e., the root of the phylogeny had a (1,1) state for all genes). With this three state Markov chain (Table 1) and employing the pruning algorithm (Felsenstein 1973; Felsenstein 1981) to calculate the likelihood of these phyletic patterns, we can model the rates of pseudogenization along the *Xenopus* species phylogeny (see supplemental methods section S1.3 for full details).

For this model, it is assumed that pseudogenization occurs independently for each gene in each species and at constant rates. The substitution rate matrix for the Markov chain is

$$\boldsymbol{Q} = \begin{array}{c} \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{ccc} (0,1) & (1,0) & (1,1) \end{array} \\ \left(\begin{array}{ccc} - & 0 & 0 \\ 0 & - & 0 \\ \theta_S & \theta_L & - \end{array} \right) \end{array}, \qquad (1)$$

where the rows (columns) represent the current (future) state, respectively. The substitution rate matrix $\boldsymbol{Q}$ only allows moves from the (1,1) state to a (0,1) or (1,0) state, i.e., from a two gene copy state to having either the L or S-subgenome homeolog.

Because our data are from transcriptomes, pseudogenization of duplicate gene copies here represents genes that are either no longer expressed, or expressed in a low enough amount to not be detected. The latter category of genes (low expression) are quite possibly on their way to becoming pseudogenized (Gout and Lynch 2015), and we explore this further in the discussion section. Additionally, there may be genes not expressed in the tissue we analyzed (liver) at the developmental stage we surveyed (adult), and these would be picked up as “pseudogenized” genes in our analysis. If it is common for genes to be missing for these reasons, our estimate of the rate of pseudogenization will be upwardly biased. But, because our interest is in comparing subgenomes to one another and time periods to one another (and subgenomes across time periods), we focus on comparison of rates rather than the magnitudes of these rates. Our model does attempt to estimate the proportion of gene copies that were not detected due to missed data, separately from pseudogenization rates. Whole genome assembly of these species and analysis using the model developed here could help further account for some of these issues, though assemblies are not without error and their analysis could benefit from a model that accommodates missing data (such as this one).

We assume that observations of no gene copies, i.e., (0,0), are not possible and are instead due to these data being generated from gene expression data. Thus, our model allows for the possibility that an observation of a (0,0) was truly either (0,1), (1,0), or (1,1). As well, this model allows for the possibility that in actuality a species has both homologs (L and S), but was recorded as only having one copy in this dataset (i.e., a (0,1) or a (1,0) is truly a (1,1)). Here, it is assumed that each copy of each gene has an equal probability of not being recorded as present. Briefly, this missing data is accommodated using the pruning algorithm where the conditional probabilities (of the same observation conditional on different events) of the tips states do not have to sum to one so, for example, an observation of (0,1) with some model estimated probability could be (1,1) and so on. For more details, please see Table S2, and supplemental methods section S1.3).

This parameter can be estimated in the likelihood calculation representing the fraction of missing (or mis-recorded) gene copies in the data for each species; details on how this is done are provided in the supplement methods section S1.3. In addition to missing data, Furman and Evans (2016) set a variety of data constraints in order to reconstruct the *Xenopus* phylogeny. The result of these constraints is that certain phyletic patterns of genes are not possible in the dataset (e.g., one constraint involved a minimum of 3 in-group taxa, thus no more than two (0,0) tips are possible). We correct the likelihood calculation for these impossible patterns, with details in the supplemental methods section S1.3.

In the context of accounting for duplication and loss events using the principle of parsimony, Eulenstein et al. (2010) note that accounting for loss events can be problematic, because “it is impossible to differentiate between gene loss and missing data”. Here, the probabilistic methodology utilized accounts for sampling bias and missing data, based on the framework used in Dang et al. (2016), while evaluating variation in the evolutionary rates of duplicate gene loss following WGD. Markov models have been successfully used to estimate evolutionary rates (e.g., insertion and deletion rates) of gene families in closely related sequences (Hao and Golding 2006; Marri et al. 2006; Cohen and Pupko 2010). These likelihood-based analyses typically require that the sequences being investigated have complete genome sequences available to ensure that phenomena such as genome rearrangement does not affect detection of homeologs (Hao and Golding 2006). As mentioned above, because we are working with transcriptome data, gene absence may not be due to pseudogenization only, and may additionally reflect failure to detect lowly expressed transcripts.

Similar models have also been used by Han et al. (2013) to correct for missing data, but similar to the models in Dang et al. (2016), our modeling needs were different. The models by Han et al. (2013) were implemented for gene family size-type data, used a birth death process, and there was no way to constrain the transitions between the different states (which we needed to do). We also simultaneously estimate what they call the error-model matrix (and provide standard errors) along with the pseudogenization rates. In contrast, Han et al. (2013) first estimate the errormodel matrix (without knowledge of this already from an external source) and then the gene family size expansion or reduction rates. Estimation of rates of discrete trait evolution has been done previously by Pagel (1994). That model was developed to assess correlation among two traits and the rate of character change across a phylogeny. While our model does estimate a rate of change for discrete characters, it does not assess correlated evolution among the characters in question (copy numbers of a gene), and also attempts to estimate the amount of missing data (along with requiring other customization outlined in the supplemental methods section S1.3).

Overall and with these assumptions and limitations, our model simultaneously corrects for incomplete gene copy membership data and provides probabilistic rate estimates of pseudogenization of duplicate gene copies, starting from a two-copy state. The model is implemented in R (R Core Team 2017) and C++, with parameter estimates obtained from numerical optimization using PORT routines (Gay 1990) as implemented in the `nlminb` function in R. Code was adapted and extended from the `indelmiss` (Dang et al. 2016) and `markophylo` (Dang and Golding 2016) packages for R. Time expensive computations were written in C++ using `Rcpp` (Eddelbuettel et al. 2011). We performed simulations to ensure reliable parameter recovery for the rates, and to show that the model can differentiate between missing data and pseudogenization; see Supplemental methods section S2 for details.

Our objective with this model was to investigate differences in the rate of pseudogenization in different time periods (soon after compared to long after WGD) and differences between the subgenomes. Thus, we fit four versions of this model to these gene presence/absence data, correcting for sampling bias in each, estimating a proportion of missing data for each of the non-*X. laevis* taxa (see supplemental methods section S1.3), and generated

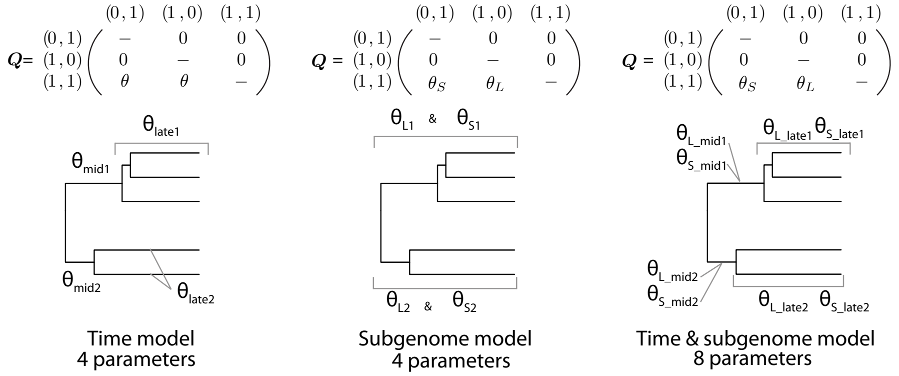

Figure 3: Model schemes for estimating the rate of pseudogenization (θ) along with the corresponding instantaneous rate matrix, indicating when one or two θs were inferred (either independently for loss of S or loss of L, or joint indicating a transition from two copies to one ignoring which subgenome copy was lost). Not shown here is a fourth model that was also assessed, wherein a single θ for each clade was estimated, with no partitioning of time or subgenome.

confidence intervals for all parameters with 1000 bootstrap replicates. As mentioned, these models were all fit using the species tree, and as with the `Codeml` analyses, we estimate rates separately for the 'borealis/clivii clade' and the 'laevis/largeni/allofraseri clade'. The first version of this model estimated a single rate of pseudogenization for the borealis/clivii clade, and another for the laevis/largeni/allofraseri clade (i.e., homogeneous rates with no partitioning of subgenome or time). For the other three, we followed a similar partitioning scheme as the `Codeml` analyses above. The second model assessed the effect of time on the rate of pseudogenization, allowing for unique rates for the mid and late-tetraploid linages, within each clade, but estimated a single pseudogenization rate (i.e., not distinguishing whether an S or L copy was lost) for each time period. The second model assessed the effects of the subgenome on the pseudogenization rate, ignoring time, by estimating a rate for the loss of the L copy separately from the loss of the S copy, in each clade. Finally, the time-subgenome model combined the previous two models, now estimating the rate of pseudogenization for each subgenome within each time period (mid and late). See Fig. 3 for a visual depiction of the models and their corresponding rate matrices, and Fig. 1 for a visual depiction of branch groups. Comparisons of these models permit statistical evaluation of the hypothesis that there were differences in pseudogenization rates between subgenomes and between time periods. For model selection, again, we assessed model fit using BIC (BIC $= 2l(\hat{\boldsymbol{\Theta}}) - \log n \times m$,) where $l(\hat{\boldsymbol{\Theta}})$ is the log-likelihood at the maximum likelihood estimates, n is the number of gene families, and m are the number of parameters estimated for the model. As above, higher BIC values are better.

Results

Subgenome-specific relaxation of selective constraints

Model comparisons by BIC indicated that the 'subgenome-species' model fit best and was 11.2 times more likely (BIC weight of 0.92 probability of being the best model; following Wagenmakers and Farrell 2004) than the second best model of 'time-subgenome-species' model, which was the most complex model (BIC weight of 0.082; Table 2). The statistically preferred model indicated that the diploid lineage experienced the strongest purifying selection (i.e., the lowest $\omega = 0.1245$), and the all S-subgenome lineages experience the weakest compared to their corresponding L-subgenome lineages (Fig. 4). For the mixed lineages, along which the whole genome duplication occurred, the S-subgenome had the weakest purifying selection detected ($\omega = 0.21$). However, this was not true for the L-subgenome lineage, where the laevis-clade had weaker purifying selection than the mixed lineage of the L-subgenome (Fig. 4). As well, for the laevis-clade, the two subgenomes had similar levels of purifying selection, with overlapping 95% confidence intervals. The more complicated time-subgenome-species model had a BIC that was only 5 units less, but included four more parameters, indicating that the time variable did not explain a large portion of the variance in these data.

WGD duplicates differ between subgenomes in coding sequence length

For the 1417 gene alignments in our analysis, both the L- and the S-subgenome homeologs were recovered from an average of 1.24 species per gene. The vast majority of these genes (1132) had only one species with both copies present and the other allotetraploids with one copy. 896 alignments had both copies only in *X. laevis* data.

Table 2: Model Support ordered by BIC (note that higher BIC is better, see Methods). BIC weights (w_i(BIC)) calculated following Wagenmakers and Farrell (2004). The ω column is the number of estimated dN/dS values in the model.

model	ω	lnL	BIC	w_i(BIC)
Subgenome-species	7	-4831284	-9663055	0.92
Time-Subgenome-species	11	-4831260	-9663060	0.082
Time-species	6	-4831316	-9663106	$9.0e^{-12}$
Subgenome	5	-4831349	-9663157	$6.9e^{-23}$
Ploidy	3	-4831386	-9663205	$2.4e^{-33}$
Time	4	-4831384	-9663214	$3.9e^{-35}$

On average, the S homeologs are shorter than corresponding L homeologs (-40.62 bp, 95% credible interval = -68.02 – -13.77). For the MCMCglmm (see Methods) fit, effective samples sizes for parameter estimates were all above 1800 (and similar between parameters) and the amount of autocorrelation among samples was low (range: -0.007 – 0.04), indicating the chain had reached convergence.

The rates of pseudogenization differ between subgenomes of several allotetraploid *Xenopus* species

Model comparisons revealed that, similar to the `Codeml` analysis, the model estimating unique rates of pseudogenization for the two subgenomes was preferred over the model with homogeneous rates for each clade, the model separating time periods, or the most complex model that estimated separate parameters for both time and subgenome (BIC: subgenome = -15,933, single rate = -15,963, time = -15,957, time-subgenome = -15,935). The results of the subgenome model indicate that the S-subgenome has a higher pseudogenization rate than the L-subgenome (borealis/clivii clade: $\hat{\theta}_S = 0.358$, $\hat{\theta}_L = 0.270$; laevis/largeni/allofraseri clade: $\hat{\theta}_S = 0.144$, $\hat{\theta}_L = 0.096$), with non-overlapping 95% confidence intervals (Fig. 5). The subgenome model also indicated that the borealis/clivii clade has a higher pseudogenization rate than the laevis/largeni/allofraseri clade (Fig. 5). As with the `Codeml` results, the more complex model incorporating time and subgenome had a similar fit as the subgenome model (BIC only 2 units less, compared to the other models that were over 20 BIC units less), indicating that time does not

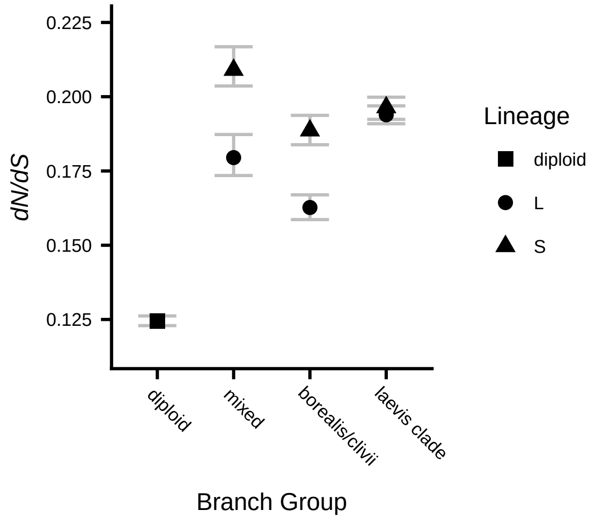

Figure 4: `Codeml` results from the favored 'subgenome-species' model. *dN*/*dS* estimates by `Codeml` 95% CIs from bootstrap replicates of the concatenated alignment.

have a large effect on these estimated pseudogenization rates. But, the results of this more complex model indicated that the laevis/largeni/allofraseri clade may have an increasing rate of pseudogenization over time, with non-overlapping 95% confidence intervals within each subgenome (Fig. S3). In the BIC-favored subgenome model, the estimated missing data proportions for each taxa ranged from 0.34 to 0.55 (all values with confidence intervals for all models are presented in Table S1).

Analyses on posterior distribution from `*Beast`

As described above, the consensus tree used in these models (and simulations) was constructed from the posterior distribution of trees recovered from the original `*Beast` analysis (Furman and Evans 2016). However, that analysis (as well as other analyses they performed and our maximum likelihood analysis above; Fig. 1) failed to confidently resolve the relationships among the *X. largeni*, *X. allofraseri*, and *X. laevis* clade (at least for the L-subgenome). To test the effect that alternate resolutions of these relationships have on our estimation of pseudogenization rates, we fitted the

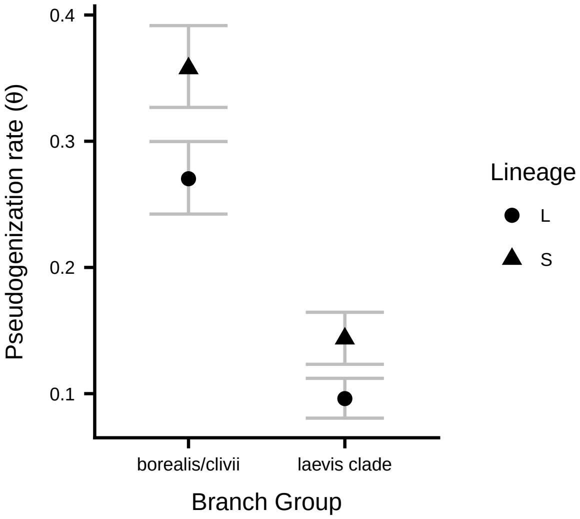

Figure 5: Estimated rates of pseudogenization from the BIC favored "subgenome model", with 95% confidence intervals based on 1000 bootstrap replicates. Missing data for non-*X. laevis* taxa varied from 0.34 – 0.55. See Table S1 for pseudogenization rate estimates of all models and estimated missing data proportions.

BIC preferred "subgenome model" to 1000 trees that were randomly sampled from the post-burn-in posterior distribution of the `*Beast` analysis and transformed to chronograms using mcmctree, as outlined above. These trees had differences in branch lengths and tree topology (i.e., one of three possible resolutions of relationships among *X. largeni*, *X. allofraseri*, and *X. laevis*). For ease of comparison while taking into account the effect of the varying branch lengths on the estimated rates, we calculated the ratio of the estimated pseudogenization rates for the clade-specific S- to L-subgenomes. For the borealis/clivii clade, we found the median (interquartile range) ratio for S- to L-subgenome estimated rates to be 1.323 $(1.323, 1.323)$. Similarly, for the laevis/largeni/allofraseri clade estimated rates, we found the median (interquartile range) ratio to be 1.495 $(1.488, 1.495)$. Thus, similar to the results from the consensus tree that are discussed above, this analysis which incorporates phylogenetic uncertainty also found that the S-subgenome had a higher pseudogenization rate than the L-subgenome. We also obtained ratios for the estimated pseudogenization rates for the borealis/clivii clade versus the laevis/largeni/allofraseri clade. Median (interquartile range) ratios for the S-subgenome and L-subgenome between the two clades are 2.487 (2.442, 2.490) and 2.812 (2.746, 2.814), respectively, i.e., the borealis/clivii clade had higher pseudogenization rates than the laevis/largeni/allofraseri clade. These ratios are nearly identical to those from the maximum likelihood estimates on the consensus tree, indicating that alternate resolutions of the *X. largeni*, *X. allofraseri*, and *X. laevis* clade did not affect the conclusions. This sort of sensitivity analysis shows that the phylogenetic comparative model was fairly robust to the provided phylogenetic tree.

Discussion

Genomic dynamics of relaxed purifying selection post-allopolyploidization

Allotetraploid species inherit each half of their genome (each subgenome) from different ancestral species. Our analyzes indicate that, after the *Xenopus* allotetraploid genome was generated, the strength of purifying selection on each subgenome differed significantly in the ancestors of some (but not all) of the species we studied. Specifically, in the lineages prior to speciation of the tetraploid ancestor and in the borealis/clivii clade, the S-subgenome experienced a greater relaxation of purifying selection than the L-subgenome. However, in the laevis/largeni/allofraseri clade, the strength of purifying selection was similar between the subgenomes. We also found that time since the speciation of the tetraploid ancestor of extant *Xenopus* did not have a large effect on the rate of purifying selection, and that gene coding regions are shorter in the S-subgenome than the homeologous coding regions in the L-subgenome. This latter result may be a consequence of more pronounced relaxation of purifying selection in this subgenome in the mixed lineage and in the borealis/clivii clade. Larger-scale subgenome-specific effects also have been observed: for instance, more rearrangements occurred since divergence of the diploid ancestors of *Xenopus* allotetraploids in the S subgenome than the L subgenome (Session et al. 2016).

That the level of purifiying selection did not vary substantially over time contrasts with the expectation that duplicates are most similar soon after allotetraploidization and thus maximally redundant, and that this redundancy would wane as homeologs diverge over time. Session et al. (2016) estimated divergence of the diploid ancestors of extant subgenus *Xenopus* species occurred about 34 million years ago (mya) and that allotetraploidization between these diploid species then occurred roughly 15–17 mya. Using our dataset and alternative methods and date calibrations, we recovered a similar estimate of 32 mya (31.6–35.4 95% CI) for the divergence of the diploid ancestors (supplemental methods section S3). Thus, the relaxation of purifying selection has persisted and not returned to pre-duplication like levels for millions of years (Fig. 4; Session et al. (2016)).

These findings are consistent with those of previous studies on *Xenopus* duplicate genes (Chain and Evans 2006; Hellsten et al. 2007), and of WGD events in other taxa (Lynch and Conery 2000; Brunet et al. 2006; Scannell and Wolfe 2008), which find relaxed purifying selection post-WGD. Using a much smaller dataset, Chain et al. (2008) also explored duplicate gene evolution over time in *Xenopus*, and did not detect a difference in the level of purifying selection between the two categories in their comparison (equivalent to our "mixed" and a combined "mid" and "late" categories, similar to our subgenome model). In this analysis, we were able to assign duplicate copies to the L and S subgenomes and test for differences between the gene copies that were inherited from separate diploid ancestors, something Chain et al. (2008) were not able to do as the *X. laevis* genome sequence was not available.

Genomic dynamics of pseudogenization post-allopolyploidization

When an allotetraploid genome first forms, it is expected to have similar gene content in each subgenome because each one is derived from a different ancestral diploid species that carried the full complement of genes required for survival. Numerous genes can be simultaneously pseudogenized immediately after WGD, possibly due to large scale changes in the regulation of gene expression (Buggs et al. 2012; Lovell et al. 2014; Inoue et al. 2015). Here, we did not explore pseudogenization in the period of evolution that (i) immediately followed WGD but that (ii) preceded diversification of the extant allotetraploids. This is because the presence of at least one species with two homeologous sequences was required to establish orthology (Furman and Evans 2016). However, 60% of homeologous pairs are still both functional in *X. laevis* (Session et al. 2016), so presumably the most recent common allotetraploid ancestor of the allotetraploids we studied retained at least this proportion (and probably an even higher proportion).

Thus, our analysis of rates of pseudogenization focused specifically on gene pairs that (a) survived an initial period following WGD before speciation of allotetraploids, and also that (b) continue to be maintained as functional duplicates in at least one of the allotetraploid species, which describes the majority of duplicate genes in the *Xenopus* genome.

Though it is possible that the extant tetraploids were the result of independent allopolyploidization events of these ancestral diploid lineages, our comparisons focus on the evolution of L and S subgenome sequences in descendant species regardless of the number of times these two lineages underwent allopolyploidization. So long as S and L sequences are each reciprocally monophyletic, our analysis remains unchanged because our model focuses on what happens to each of these subgenomes in the species that carry both. Our results indicate that (i) each progenitor lineage contributed unequally to the functional gene content in extant allotetraploids, and that (ii) the conditions at the time of allopolyploidization (e.g., divergence between diploid ancestral lineages) and after allopolyploidization (e.g., species-specific population dynamics and mutation) strongly influence allopolyploid genome evolution.

After allotetraploidization in *Xenopus*, we found that the S-subgenome had a faster rate of pseudogenization than the L in several *Xenopus* allotetraploids (by about 30–50%; Fig. 5). The complete genome sequence of *X. laevis* reveals that the S-subgenome lost 31.5% of gene copies, where as the L had only lost 8% (Session et al. 2016). Our findings thus extend these *X. laevis* genomic results to several other allotetraploid *Xenopus* species with 2n=4x=36 chromosomes, including *X. borealis*, *X. clivii*, *X. largeni*, and *X. allofraseri* (Fig. 5). Asymmetry in subgenome pseudogenization was observed in a smaller scale in two genes across a diversity of species in subgenus *Xenopus* (*RAG1*: Evans 2007; *DMRT1* loci: Bewick et al. 2011). Using `BLAST`, we assigned subgenome of origin for the results of Evans (2007) and Bewick et al. (2011), which indicated that the homeologs with the higher rates of pseudogenization are in the S-subgenome (data not shown), a finding that is consistent with the higher rate of pseudogenization in the S subgenome. This suggests the rate of pseudogenization is also higher in the S subgenomes of allo-octoploid and allo-dodecaploid *Xenopus* as well. Overall, in terms of gene copies recovered from the transcriptome data, single copy S genes were 15–30% less common than single copy L genes across the species, but this figure reflects a combination of pseudogenization and missing data (Table 1).

Our analyses did not detect evidence for a substantial change in the rate of pseudogenization over time since allotetraploidization in *Xenopus*. In teleosts, the rate of pseudogenization was highest soon after WGD (Inoue et al. 2015), but the period over which there was most rapid gene loss was about 60 my – greater than the time since the WGD event in *Xenopus*. Furthermore, a slowdown in teleost pseudogenization occurred only after about 80% of duplicates were lost (Inoue et al. 2015), whereas in *X. laevis* <40% of duplicates have been lost so far (Session et al. 2016). Yeast also exhibits a tempo of pseudogenization similar to teleosts, with more rapid gene loss earlier on (Scannell et al. 2006), but this pattern also played out over a longer period of time (and many more generations) than the *Xenopus* WGD analyzed here (> 60 my, Marcet-Houben and Gabaldón 2015). These results from *Xenopus*, which is a comparatively recent WGD event, indicate that in the early stages of genome restructuring post-WGD the rate of gene loss may be relatively constant until most gene copies are lost. Millions of frog generations in the future, it is certainly possible that the rate of gene loss also will slow down in *Xenopus*.

We note that the borealis/clivii clade had a higher rate of pseudogenization in our analysis than the laevis/largeni/fraseri clade, which suggests that the rate of gene loss can be species-specific post-WGD. Similar to the species-specific relaxation of purifying selection discussed above, this could be a consequence of differences in life history, natural selection, demography, or other factors. Species-specific rates of pseudogenization after WGD also have been reported in yeast (Scannell et al. 2006). Purifying selection was stronger in the borealis/clivii clade for both subgenomes than in the laevis/largeni/allofraseri clade (Fig. 4), which could be because there are more singleton genes in the former clade. Indeed, in the expression data we analyzed here, the fewest number of genes in duplicate copy were recovered for the borealis/clivii clade (Table 1). The lack of substantial variation in the rate of pseudogenization over time coupled with pronounced variation among species in this rate, argues that aspects of the evolutionary fates of allopolyploid genomes are influenced to a great degree by species-specific phenomena (e.g., mutations, effective population size, demography). A high quality complete genome sequence for *X. borealis* and the other species will make possible further exploration of these interpretations.

Asymmetric subgenome evolution

A unique implication of speciation by allopolyploidization is the merging of diverged genomes into a single species. While there may be beneficial consequences, such as higher dosages of beneficial alleles, there are also potentially negative consequences, such as poorly coordinated epistatic interactions from diverged members of a genetic network (Otto and Whitton 2000; Riddle and Birchler 2003; Comai 2005). Establishment of disomic inheritance (i.e., the formation of bivalents rather than multivalents at meiosis; Wolfe 2001) may confer greater genomic stability to allopolyploid genomes (Comai et al. 2003), and also allow for a preservation of subgenome differences that otherwise would be homogenized by recombination.

It has been well demonstrated in both old and young allopolyploids plants that the subgenome from one of the progenitors is often expressed less, and more frequently the target of pseudogenizing mutations, referred to as 'biased fractionation' (Flagel and Wendel 2010; Cheng et al. 2012; Garsmeur et al. 2013; Renny-Byfield et al. 2015). If each subgenome has a distinctive repertoire of transposible elements (TEs), a possible mechanism for differential subgenome evolution is RNA-mediated silencing of TEs that also represses adjacent genes (Woodhouse et al. 2014; Steige and Slotte 2016). Reduced gene expression is linked to weaker purifying selection and a higher rate of mutation and pseudogenization (Rocha 2006; Steige and Slotte 2016). Analysis of the extant *X. laevis* is consistent with this possibility, in that the subgenomes have different TE classes and abundances, with the S-subgenome carrying subgenome-specific TE classes at high abundance (Session et al. 2016). Exploration of gene expression in this species found that L-subgenome homeologs tended to have higher expression than S-subgenome copies (mean difference of 10–25%, but a more modest median difference of 1.8% or less), along with 760 homeologous gene pairs where the homeolog with little to no expression had more relaxed purifying selection than the other (Session et al. 2016). These differences in TEs between the subgenomes were probably inherited from the diploid ancestors (Session et al. 2016), and this could have set the stage for higher pseudogenization in the S subgenome. Related to this, when the expression level of one homeolog is or evolves to be sufficient for survival, the fitness cost if the other becomes a pseudogene becomes tolerable (Freeling et al. 2012; Gout and Lynch 2015). The higher L-subgenome expression in *X. laevis* (Session et al. 2016) is thus consistent with these homeologs being less amenable to loss (Fig. 5). But, measuring only an extant species makes it unclear whether the differences between the L and the S in these species existed before WGD (i.e., in the diploid species) or arose after WGD.

Though the (presumably) extinct diploid progenitors of tetraploid *Xenopus* cannot be directly assayed for differences in expression or natural selection, leveraging of multiple species and the branch specific dN/dS models provide some insight into differences at the time of WGD. We detected weaker purifying selection along the 'mixed lineages' branches compared to the purely diploid lineage, and also significantly weaker purifying selection on the S mixed lineage than the L mixed lineage (Fig. 1, 4). The dN/dS estimates of the mixed lineages probably underestimate dN/dS immediately after WGD because a portion of the mixed lineages was diploid. Overall, however, these dN/dS estimates indicate that either S and L subgenome differences were rapidly established after WGD, or possibly a result of divergence between the diploid progenitors, as was suggested by Session et al. (2016) in the analysis of just *X. laevis*. Allopolyploid green toads indicate that the formation of polyploids can occur from progenitors that are more diverged than diploids that hybridize without polyploidy, indicating that divergence between progenitors is not necessarily a hurdle for allopolyploidization (Betto-Colliard et al. 2018). If expression intensity is negatively correlated with purifying selection (Drummond et al. 2005; Rocha 2006), the expression differences between the L and S subgenomes may have been present soon after WGD in *Xenopus*, and persisted for many millions of years thereafter, as seen for other allopolyploids (Cheng et al. 2012; Renny-Byfield et al. 2015). Our analysis supports that S-subgenome loss has been higher than the L-subgenome, and remained consistently so over time (Fig. 5), and suggests that the differences between the subgenome was a persistent feature of *Xenopus* allotetraploids.

Asymmetry in subgenome evolution has several interesting consequences for genome evolution. For instance, allopolyploid cotton species, also show asymmetry in subgenome evolution with a bias in gene conversion rates (one subgenome is more frequently converted by the other subgenome), and there exists subgenome biases in gene involvement in certain phenotypes (Paterson et al. 2012). Wheat species subgenomes carry different levels of genetic diversity, with the more diverse subgenome involved in local adaptation phenotypes and the other preserving more core function genes (Feldman et al. 2012). In *Xenopus*, cyto-nuclear incompatibilities appear to be limited to one subgenome or the other (Gibeaux et al. 2018), indicating that subgenome specific evolution may additionally contribute to the origin of reproductive incompatibilities among species. Interestingly, biases in subgenome evolution may exist even though genetic exchange between subgenomes does occasionally occur. Genetic exchange between subgenomes of allotetraploid *Xenopus* is illustrated, for example, by the sex determining gene *DMW* which resides on the L-subgenome but was formed from a partial gene duplicate of the S-subgenome copy of a gene called *DMRT1* (Bewick et al. 2011).

Overall, this study paints a dynamic portrait of allopolyploid genome evolution by highlighting among several closely related allotetraploids evidence for persistent relaxed purifying selection with species-specific subgenome patterns, and ongoing pseudogenization with asymmetric rates in each subgenome. Many functional duplicate genes still remain in these and other allotetraploids, and do so for millions of years (Renny-Byfield et al. 2015, this study). As such, events that are thought to depend on rare mutational events that promote the retention of duplicate genes, such as neofunctionalization, may in fact have a protracted time-frame within which to occur.

Competing interests

The authors declare that they have no competing interests.

Data Availability

Transcriptome sequence data are available in the NCBI short read archive, as submitted by Furman and Evans (2016) (accessions PRJNA318484, PRJNA318394, PRJNA318474, PRJNA318404).

Additional Files

Additional file 1 — Supplemental Text and Figures

Additional figures, explanation of divergence estimates, and details of simulations of pseudogenization rates can be found in the supplemental file.

References

Adams, K. L. (2007). Evolution of duplicate gene expression in polyploid and hybrid plants. *Journal of Heredity* 98(2), 136–141.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17), 3389–3402.

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444(7116), 171–178.

Betto-Colliard, C., Hofmann, S., Sermier, R., Perrin, N., and Stöck, M. (2018). Profound genetic divergence and asymmetric parental genome contributions as hallmarks of hybrid speciation in polyploid toads. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 285(20172667), 1–8.

Bewick, A. J., Anderson, D. W., and Evans, B. J. (2011). Evolution of the closely related, sex-related genes DM-W and DMRT1 in African clawed frogs (*Xenopus*). *Evolution* 65(3), 698–712.

Blanc, G. and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *The Plant Cell* 16(7), 1679–1691.

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution* 23(9), 1808–1816.

Buggs, R. J., Chamala, S., Wu, W., Tate, J. A., Schnable, P. S., Soltis, D. E., Soltis, P. S., and Barbazuk, W. B. (2012). Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Current Biology* 22(3), 248–252.

Chain, F. J. J. and Evans, B. J. (2006). Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. *PLoS Genetics* 2(4), e56.

Chain, F. J. J., Ilieva, D., and Evans, B. J. (2008). Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evolutionary Biology* 8(1), 43.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PloS one* 7(5), e36442.

Cohen, O. and Pupko, T. (2010). Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution* 27(3), 703–713.

Comai, L. (2000). Genetic and epigenetic interactions in allopolyploid plants. *Plant Molecular Biology* 43(2-3), 387–399.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6(11), 836.

Comai, L., Madlung, A., Josefsson, C., and Tyagi, A. (2003). Do the different parental 'heteromes' cause genomic shock in newly formed allopolyploids? *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358(1434), 1149–1155.

Dang, U. J., Devault, A. M., Mortimer, T. D., Pepperell, C. S., Poinar, H. N., and Golding, G. B. (2016). Estimation of Gene Insertion/Deletion Rates with Missing Data. *Genetics* 204(2), 513–529.

Dang, U. J. and Golding, G. B. (2016). markophylo: Markov Chain Analysis on Phylogenetic Trees. *Bioinformatics* 32(1), 130–132.

Dehal, P. and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* 3(10), e314.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences* 102(40), 14338–14343.

Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D., and Ushey, K. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* 40(8), 1–18.

Eulenstein, O., Huzurbazar, S., and Liberles, D. A. (2010). Reconciling Phylogenetic Trees. In: *Evolution after Gene Duplication*. Ed. by K. Dittmar and D. Liberles. Hoboken, NJ, USA: John Wiley & Sons, Inc. Chap. 10, 185–206.

Evans, B. J. (2007). Ancestry influences the fate of duplicated genes millions of years after polyploidization of clawed frogs (*Xenopus*). *Genetics* 176(2), 1119–1130.

Evans, B. J., Kelley, D. B., Tinsley, R. C., Melnick, D. J., and Cannatella, D. C. (2004). A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. *Molecular Phylogenetics and Evolution* 33(1), 197–213.

Evans, B. J., Carter, T. F., Greenbaum, E., Gvoždík, V., Kelley, D. B., McLaughlin, P. J., Pauwels, O. S., Portik, D. M., Stanley, E. L., Tinsley, R. C., et al. (2015). Genetics, Morphology, Advertisement Calls, and Historical Records Distinguish Six New Polyploid Species of African Clawed Frog (*Xenopus*, Pipidae) from West and Central Africa. *PLoS One* 10(12), e0142823.

Fawcett, J. A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceedings of the National Academy of Sciences* 106(14), 5737–5742.

Feldman, M., Levy, A. A., Fahima, T., and Korol, A. (2012). Genomic asymmetry in allopolyploid plants: wheat as a model. *Journal of Experimental Botany* 63 (14), 5045–5059.

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Biology* 22(3), 240–249.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. English. *Journal of Molecular Evolution* 17(6), 368–376. ISSN: 0022-2844.

Flagel, L. E. and Wendel, J. F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist* 186(1), 184–193.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4), 1531–1545.

Freeling, M., Woodhouse, M. R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J. C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology* 15(2), 131–139.

Furman, B. L. S. and Evans, B. J. (2016). Sequential turnovers of sex chromosomes in African clawed frogs (*Xenopus*) suggest some genomic regions are good at sex determination. *G3: Genes|Genomes|Genetics* 6(11), 3625–3633.

Garamszegi, L. Z. (2014). *Modern phylogenetic comparative methods and their application in evolutionary biology. Concepts and Practice.* London, UK: Springer.

Garsmeur, O., Schnable, J. C., Almeida, A., Jourda, C., D'Hont, A., and Freeling, M. (2013). Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution* 31(2), 448–454.

Gay, D. M. (1990). *Usage summary for selected optimization routines.* AT&T Bell Laboratories. Murray Hill, New Jersey.

Gibeaux, R., Acker, R., Kitaoka, M., Georgiou, G., Kruijsbergen, I. van, Ford, B., Marcotte, E. M., Nomura, D. K., Kwon, T., Veenstra, G. J. C., et al. (2018). Paternal chromosome loss and metabolic crisis contribute to hybrid inviability in *Xenopus. Nature*.

Gout, J.-F., Kahn, D., Duret, L., and Consortium, P. P.-G. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genetics* 6(5), e1000944.

Gout, J.-F. and Lynch, M. (2015). Maintenance and loss of duplicated genes by dosage subfunctionalization. *Molecular Biology and Evolution* 32 (8), 2141–2148.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology* 29(7), 644.

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* 33(2), 1–22.

Han, M. V., Thomas, G. W., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Molecular Biology and Evolution* 30(8), 1987–1997.

Hao, W. and Golding, G. B. (2006). The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research* 16(5), 636–643.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27(3), 570–80.

Hellsten, U., Khokha, M. K., Grammer, T. C., Harland, R. M., Richardson, P., and Rokhsar, D. S. (2007). Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis. BMC Biology* 5, 31.

Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K., and Nishida, M. (2015). Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proceedings of the National Academy of Sciences* 112(48), 14918–14923.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., and Soltis, D. E. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345), 97–100.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A., Zhao, M., Ma, J., Yu, J., Huang, S., et al. (2014). The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications* 5, 3930.

Lovell, P. V., Wirthlin, M., Wilhelm, L., Minx, P., Lazar, N. H., Carbone, L., Warren, W. C., and Mello, C. V. (2014). Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biology* 15(12), 565.

Lynch, M. and Conery, J. S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science* 290(5494), 1151–1155.

Makino, T. and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences* 107(20), 9270–9274.

Marcet-Houben, M. and Gabaldón, T. (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology* 13(8), e1002220.

Marri, P. R., Hao, W., and Golding, G. B. (2006). Gene Gain and Gene Loss in *Streptococcus*: Is It Driven by Habitat? *Molecular Biology and Evolution* 23(12), 2379–2391.

Matsuda, Y., Uno, Y., Kondo, M., Gilchrist, M. J., Zorn, A. M., Rokhsar, D. S., Schmid, M., and Taira, M. (2015). A new nomenclature of *Xenopus laevis* chromosomes based on the phylogenetic relationship to *Silurana/Xenopus tropicalis*. *Cytogenetic and Genome Research* 145(3-4), 187–191.

McGrath, C. L., Gout, J.-F., Johri, P., Doak, T. G., and Lynch, M. (2014). Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research* 24(10), 1665–1675.

Ohno, S. et al. (1970). *Evolution by Gene Duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Otto, S. P. and Whitton, J. (2000). Polyploid incidence and evolution. *Annual review of genetics* 34(1), 401–437.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* 255(1342), 37–45.

Papp, B., Pal, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424(6945), 194–197.

Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K. C., Shu, S., Udall, J., et al. (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* 492(7429), 423–427.

Qian, W., Liao, B.-Y., Chang, A. Y.-F., and Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics* 26(10), 425–430.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One* 6(9), e22594.

Renny-Byfield, S., Gong, L., Gallagher, J. P., and Wendel, J. F. (2015). Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Molecular Biology and Evolution* 32(4), 1063–1071.

Riddle, N. C. and Birchler, J. A. (2003). Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. *Trends in Genetics* 19(11), 597–600.

Rocha, E. P. (2006). The quest for the universals of protein evolution. *Trends in Genetics* 22(8), 412–416.

Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S., and Wolfe, K. H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440(7082), 341–345.

Scannell, D. R. and Wolfe, K. H. (2008). A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research* 18(1), 137–147.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., et al. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* 538(7625), 336–343.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9), 1312–1313.

Steige, K. A. and Slotte, T. (2016). Genomic legacies of the progenitors and the evolutionary consequences of allopolyploidy. *Current Opinion in Plant Biology* 30, 88–93.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* 49(2), 169–181.

Tymowska, J. (1991). Polyploidy and cytogenetic variation in frogs of the genus *Xenopus*. *Amphibian Cytogenetics and Evolution*, 259–297.

Wagenmakers, E.-J. and Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* 11(1), 192–196.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2(5), 333–341.

Woodhouse, M. R., Cheng, F., Pires, J. C., Lisch, D., Freeling, M., and Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proceedings of the National Academy of Sciences* 111(14), 5283–5288.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in Bioscience* 13(5), 555–556.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24(8), 1586–91.