


SOFTWARE

Open Access



WGDTree: a phylogenetic software tool to examine conditional probabilities of retention following whole genome duplication events

C. Nicholas Henry¹, Kathryn Piper^{1,3}, Amanda E. Wilson¹, John L. Miraszek^{1,4}, Claire S. Probst¹, Yuying Rong^{1,2,5} and David A. Liberles^{1*} 

*Correspondence:
daliberles@temple.edu

¹ Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA 19122, USA

² Department of Biology, Haverford College, Haverford, PA 19041, USA

³ Present Address: Department of Biological Sciences, University at Albany, Albany, NY 12222, USA

⁴ Present Address: Genetics Area Program, University of Missouri, Columbia, MO 65211, USA

⁵ Present Address: Groningen Institute for Evolutionary Life Sciences, University of Groningen, 9747 AG Groningen, The Netherlands

Abstract

Background: Multiple processes impact the probability of retention of individual genes following whole genome duplication (WGD) events. In analyzing two consecutive whole genome duplication events that occurred in the lineage leading to Atlantic salmon, a new phylogenetic statistical analysis was developed to examine the contingency of retention in one event based upon retention in a previous event. This analysis is intended to evaluate mechanisms of duplicate gene retention and to provide software to generate the test statistic for any genome with pairs of WGDs in its history.

Results: Here a software package written in Python, 'WGDTree' for the analysis of duplicate gene retention following whole genome duplication events is presented. Using gene tree-species tree reconciliation to label gene duplicate nodes and differentiate between WGD and SSD duplicates, the tool calculates a statistic based upon the conditional probability of a gene duplicate being retained after a second whole genome duplication dependent upon the retention status after the first event. The package also contains methods for the simulation of gene trees with WGD events. After running simulations, the accuracy of the placement of events has been determined to be high. The conditional probability statistic has been calculated for *Phalacrocorax auritus* on a monocot species tree with a pair of consecutive WGD events on its lineage, showing the applicability of the method.

Conclusions: A new software tool has been created for the analysis of duplicate genes in examination of retention mechanisms. The software tool has been made available on the Python package index and the source code can be found on GitHub here: <https://github.com/cnickh/wgdtree>.

Keywords: Whole genome duplication, Phylogenetic analysis, Gene duplicability, Mutational opportunity



Background

Gene duplication is an important driver of the evolution of genomes because without gene duplication, evolution is thought to act conservatively [1]. Gene duplication can relax selective constraints and enable faster evolution by creating redundant copies [2]. This redundancy can provide opportunity for favorable morphological innovative development in addition to other processes [3, 4]. Duplicate gene evolution comes in two broad types, smaller scale duplication (SSD) and whole genome duplication (WGD) which is rarer but in some cases can be a beneficial process [5]. These are differentiated by several functional features. Because all or a large piece of the genome is duplicated, WGD duplicates are duplicated together with their physical interacting partners [6–8].

Dosage balance theory states that selection favors gene products existing in stoichiometric balance which prevents the deleterious interaction of imbalanced partners [9–11]. Whole genome duplication (WGD) events preserve the dosage balance, so selection favors a slow initial duplicate gene loss rate [8, 12–17]. Alternatively, dosage constraints can favor removing gene duplicates quickly after small-scale duplication events when these events immediately throw off the stoichiometric balance of gene products [18]. Changes in gene expression can sometimes also aid in the initial retention of duplicate genes because it helps maintain the balance [14, 19]. Other processes that lead to the long-term retention of gene duplicates include subfunctionalization and neofunctionalization [1, 12, 20–22].

Duplicate genes that are retained over long evolutionary time periods do show patterns consistent with neofunctionalization and/or subfunctionalization [23]. Several factors affect the probability that an individual gene will neofunctionalize or subfunctionalize, including the number and specific functions of the gene, its length, and the complexity of its regulatory regions, among others [24–26]. The gene duplicability hypothesis states that some genes are more duplicable than other genes because of these gene characteristics [24, 27]. A naive expectation from this hypothesis is that when a genome undergoes consecutive WGD events, the genes retained after the first event are also more likely to be retained in following WGD events.

In a study performed on gene retention rates after consecutive whole genome duplication events in Atlantic salmon, a new statistic was developed [23]. This new statistic was developed for evaluating the conditional probability of duplicate gene retention from the second WGD based upon the retention status from the first WGD event for the analysis of the Atlantic salmon genome, which has had two relatively recent WGD events in its history [23]. The statistic is applied to a set of gene trees with consecutive WGD events on different species tree lineages. The probability ratio statistic is shown in Eq. 1.

$$Pratio = \frac{P(\text{retained after WGD \#2} | \text{retained after WGD \#1})}{P(\text{retained after WGD \#2} | \text{not retained after WGD \#1})} \quad (1)$$

The analysis from the Atlantic salmon genome unexpectedly gave a ratio of ~1, consistent with independence of duplicate retention in genes between events rather than the prior conceptualization of gene duplicability. Other lines of evidence might still support a more complex process involving non-independence [23, 28]. Because of this result there is interest in generalized software to characterize more genomes

to enable analysis of the ratio and mechanisms leading to retention generally and in different genomes. This type of analysis has the potential to spawn advancements in our understanding of duplicate gene retention together with additional future methodological refinement. Further, modeling with the gene duplicability hypothesis under different evolutionary scenarios and duplication times represents a parallel research track to explain the result from the Atlantic salmon genome. Other hypotheses beyond gene duplicability, including changes to mutational opportunity for functional change after duplicate gene retention can also be conceived and are being developed. It is important to note that the retention process is time dependent, meaning that the time between the WGD events and the time since the most recent event affects the probability that duplicate copies are retained. Evaluating this hypothesis requires modeling that is time-dependent and will be presented elsewhere.

To generate data across the genome from consecutive duplication events, a phylogenetic approach was developed [23, 28]. This approach relied upon the construction of gene trees for all genes in the genome and a reference species tree, with the need to differentiate between smaller scale events and the whole genome events of interest. The original script [28], which was based upon algorithms for gene tree/species tree reconciliation, was hard coded for properties of salmonids, including syntenic information in the Atlantic salmon [23] and rainbow trout [29] genomes.

Other phylogenetic methods of identifying and differentiating SSDs and WGDs rely on the number of gene trees that show the event in question [30, 31]. Additionally, some methods to strengthen the identification of WGD events can use syntenic information because one expects conserved synteny after a WGD event but not with SSD events [32]. This information provides support for the identified events and complements a purely phylogenetic approach.

Here, using the Python scripting language, generalized software to calculate the probability ratio for any pair of WGD events in a tree specified by a user has been created. The software takes a collection of gene trees and a species tree as input in doing so and is available at <https://github.com/cnickh/wgdtree>.

Implementation

Algorithm for inference

Software has been created to enable evaluation of conditional retention probability ratios (the test statistic) for any genome that has a pair of WGD events in its history. Written in Python, the conditional probability ratio statistic is calculated from a collection of gene trees derived from systematic comparative genomic analysis and a reference species tree with whole genome duplication events labeled. The statistic can be calculated for every pair of WGD events that occurs serially in the evolutionary history of a species. This refers to two whole genome duplication events that occurred on different species tree lineages that are both on the same phylogenetic trajectory from the species tree root to the extant species tip. At the heart of the package, is a gene tree-species tree algorithm that labels specific nodes in each gene tree, as described below.

The input to the software for inference are a reference rooted species tree and a set of unrooted gene trees from the genomes of the species involved. Gene trees do not need to contain genes from all species in the species tree but are assumed to contain all

members that existed descended from the root node of the species tree. Once rooted using the species tree, gene trees containing more than 1 species with root nodes that are duplication events are split iteratively until the root node is a speciation event.

In order to calculate the conditional probability ratio, it is needed to map WGD events onto nodes of a gene tree. In principle, a WGD event could correspond to any duplication event on a gene tree where all leaves under that node correspond to a species under the event on the species tree. Here it will be assumed the most parsimonious solution, where the placement that results in the smallest number of small scale duplications is utilized. When equally parsimonious solutions are possible, the algorithm will place WGD at the earlier node. There may be scenarios where this choice can lead to a bias, but this has not been detected here. The mapping has a linear time complexity with respect to the number of nodes on the species tree.

Notation

G, denotes an input node. This should be the root node of a rooted and labeled gene tree.

S, denotes an input node. This should be the root of the corresponding species tree with labeled WGD events.

g, is used to represent specified nodes on the gene tree.

s, is used to represent specified nodes on the species tree.

r(n), denotes the right child of node **n**.

l(n), denotes the left child of node **n**.

L(n), denotes all leaves under **n**.

add_event(n,*), labels a gene tree node **n** as a WGD duplication, where the duplicate was either retained so the event is “Present” or the duplicate was lost so the event is “Missing”.

lca(s,g), denotes a method that returns the mapping of node **s** onto **g** by getting the last [least] common ancestor of all leaves present under **s** on **g**. For example, let **g** define a gene tree node: **((a1, b1), (a2, b2))g**. Let **s** be a parent node on the species tree with only **a,b** as children. Here the left child and the right child include all species labels. Then **lca(s,g)** will simply return **g**.

Implementation

The placement algorithm takes as input binary gene trees that are rooted and reconciled such that duplication and loss events are present and labeled and one species tree where the branches containing the WGD events are labeled. This is generated by the software as described below, given a rooted species tree and a set of unrooted gene trees. Given the root node of a gene tree **G** and the root node of the species tree **S** the algorithm maps the WGD events on to the gene tree to give as output a new gene tree with nodes labeled as WGD.

First it is necessary to root and reconcile the gene trees. The software package presented here implements a known reconciliation algorithm [33], which differs from an algorithm previously implemented in the group [34]. The software also roots the

tree by iterating over all branches and selecting the root that minimizes the number of duplication and loss events on the reconciled tree. From the reconciliation a rooted gene tree with both duplication and loss events labeled is obtained. Although it is possible and computationally more efficient to map WGD events onto the gene tree as part of the reconciliation it was decided to keep these methods separate to allow other reconciliation algorithms to be used with the WGD placement software. As long as the resulting gene tree is labeled with duplication and loss events the placement algorithm will work. The loss events are treated as leaves for the purposes of the mapping. Let \mathbf{G} be the root node of a gene tree and \mathbf{S} the root of a species tree. Both should be labeled rooted binary trees. The set of leaves of a gene tree $\mathbf{L}(\mathbf{G})$ and a species tree $\mathbf{L}(\mathbf{S})$ are taken from the same set of species. The first step is to select the most recent possible node \mathbf{n} under \mathbf{g} an arbitrary gene tree node such that $\mathbf{L}(\mathbf{s}) \subseteq \mathbf{L}(\mathbf{n})$, where \mathbf{s} represents a node on the species tree directly after the WGD event and $\mathbf{L}(\mathbf{s})$ denotes all leaves under \mathbf{s} . Let $\mathbf{lca}(\mathbf{s}, \mathbf{g})$ be a function for this mapping. This mapping could return an \mathbf{n} such that $\mathbf{L}(\mathbf{n}) \not\subseteq \mathbf{L}(\mathbf{s})$, for example if \mathbf{g} was a node $(((\mathbf{a1}, \mathbf{b1}), \mathbf{c1}), ((\mathbf{a2}, \mathbf{b2}), \mathbf{c2}))$ and \mathbf{s} was only a parent to \mathbf{a}, \mathbf{b} . If this is the case \mathbf{n} is a duplication occurring before the whole genome duplication event. This means it is possible the event corresponds to two locations on the gene tree. So we use $\mathbf{r}(\mathbf{n})$, the right child of \mathbf{n} and $\mathbf{l}(\mathbf{n})$, the left child of \mathbf{n} and get two new mappings. This process continues until we have a likely candidate, a node or nodes \mathbf{n} such that $\mathbf{L}(\mathbf{s}) \subseteq \mathbf{L}(\mathbf{n})$ and $\mathbf{L}(\mathbf{n}) \subseteq \mathbf{L}(\mathbf{s})$.

Next, one checks if \mathbf{n} is labeled as a duplication by the rooting/reconciliation method. If the node is not a duplication, then the node is labeled as a missing WGD event. If the node was a duplication both children are checked for duplication events. If both children are duplication nodes, then the event is placed on both children. If not, the event is placed on the current node and labeled as present. The reasoning behind checking the children for duplication events works as follows. If a duplication, where all leaves under that node correspond to a species under the event on the species tree, is in fact not a result of WGD it be would expected to see both copies of the gene duplicate as a result of WGD. This method greatly reduces the number of duplication events attributed to small scale duplication compared to placement methods that do not check children node for duplications. Here, **event_num** is an integer that tracks which event is being placed. Upon successfully placing an event, it is incremented for the next consecutive event. For example, when **place_event()** is called if **event_num** is 0 the node will be labeled “event 0”.

Pseudo-code

Pseudo-code for the algorithm is provided as Fig. 1.

Now with WGD events mapped onto the gene tree, computing the conditional retention is straight forward. First one iterates over each node until we find a node labeled as event 0 then we iterate over all descendant nodes until we find a node labeled as event 1. For every event 1 node found the number of possible duplicate copies increases by one. All the leaves descending from the event 1 node are checked and the total number of retained duplicate copies is counted. The number of duplicate copies along with the number of possible duplicate copies is added to a counter that tracks the total copies and possible copies across all gene families being analyzed. If event 0 was

Algorithm 1 *Place*($S, G, event_num$)

```

if the parent branch of  $S$  has WGD then
   $g = lca(S, G)$ 
  if  $L(g) == L(S)$  then
    if  $g$  is duplication then
      if  $r(g)$  and  $l(g)$  are duplications then
         $Place(r(g), S, event\_num)$ 
         $Place(l(g), S, event\_num)$ 
      else
         $add\_event(g, Present)$ 
         $Place(r(g), r(S), event\_num + 1)$ 
         $Place(l(g), r(S), event\_num + 1)$ 
         $Place(r(g), l(S), event\_num + 1)$ 
         $Place(l(g), l(S), event\_num + 1)$ 
      end if
    else
       $add\_event(g, Missing)$ 
       $Place(g, r(S), event\_num + 1)$ 
       $Place(g, l(S), event\_num + 1)$ 
    end if
  else
     $Place(r(g), S, event\_num)$ 
     $Place(l(g), S, event\_num)$ 
  end if
else if  $S$  is not a leaf then
   $Place(g, r(S), event\_num)$ 
   $Place(g, l(S), event\_num)$ 
end if

```

Fig. 1 Pseudo-code for the algorithm is given in this figure

present, then the copies and possible copies get added to the RR count and if event 0 was missing, the copies get added to the LR count. This process repeats for all events 1 descending from the event 0. This is done for every event 0 node on the tree.

For example, In the case where a gene "A" was retained after the first WGD, there will be two copies of "A", "A1" and "A2", before the second WGD happens. If, after the second WGD, the new copies of "A1" are retained but not those "A2". The software will find one event 0 with two descended nodes labeled event 1 so 2 possible duplicate copies, and since the "A1" duplicate was retained and not "A2" we would see 1 duplicate copy of the gene still present on the tree. Thus for just this tree the conditional probability of RR would be 0.5 and since the first event was retained there would be no data for LR.

Comparative genomic bioinformatic pipeline

The following seven species of plants with 5 pairs of WGD events between them [35–38] were selected to be analyzed: *Ananas comosus*, *Elaeis guineensis*, *Nelumbo nucifera*, *Oryza brachyantha*, *Panicum hallii*, *Phalaenopsis equestris*, and *Phoenix dactylifera*. All plant species selected are autopolyploids with available high-quality genomes, allopolyploids were eliminated. A species tree was generated using NCBI [39] with the timing of species divergences generated using Timetree [40]. Protein sequences for all seven species were gathered from NCBI as FASTA files. A total of 255,312 protein sequences were gathered, 35,775 from *Ananas comosus* at 400x, 41,887 from *Elaeis*

guineensis at $16\times$ coverage, 38,191 from *Nelumbo nucifera* at $100\times$ coverage, 26,803 from *Oryza brachyantha* at $104\times$, 44,192 from *Panicum hallii* at $202\times$ coverage, 29,894 from *Phalaenopsis equestris* at $99.5\times$ coverage, and 38,570 from *Phoenix dactylifera* at $139\times$ coverage. High coverage genomes were used to reduce any potential for bias due to missing gene duplicates. It should be noted that the Pratio statistic is expected to be more robust than other retention statistics because missing duplicates might be expected to occur in the statistic numerator or denominator without bias. BLAST all-against-all was run for each pair of species, including against themselves, at an e-value threshold of 10^{-10} to identify homologous protein pairs. These pairs of homologs were run through a script to ensure that a pair of homologs had both percent identity and percent ungapped were both $\geq 60\%$. 255,187 gene families were formed by single linkage clustering of all the gene pairs. Protein alignments of the gene families were generated using MAFFT [41]. Maximum likelihood trees were created for gene families of size 4 or greater by PhyML [42] using SMS model selection [43] and Neighbor-Joining. Due to the size of some gene families (size > 100) PhyML was not an efficient method to use, Neighbor-Joining was used for these families. The gene trees were rooted using a Python script based off [7, 34]. As described above, gene trees with more than 1 species and a root node as a duplication event were iteratively split until the root node was a speciation event. In total 12,852 trees with size > 4 were generated, 81 of which used neighbor joining due to size.

Generation of simulated data

For testing and as a companion to the inference tool, a simulation tool was built that is capable of producing a statistically probable gene trees for any given species tree [44–46]. Using a Poisson process to dictate the arrival of events (duplication/loss), a set of gene trees is produced. The evolutionary history of a gene is simulated over each branch of a given species tree.

Duplication affects the gene tree by duplicating the corresponding branch and subtree and losses affect the tree by dropping one branch and subtree. The simulation also includes the ability to add WGD duplication events where every copy of the gene present for the event is duplicated. The location for these events is determined via comments with a specific T value on the input species tree. The T value represents time and affects the probability that the simulation will place a SSD before or after the event on the same branch. This is a comment used with NHX format. The package uses ETE 3 [47] to read and manipulate trees. The resulting tree is a rooted simulated gene tree corresponding to the inputted species tree with WGD events labeled. Using these simulated trees, the inference method used to place WGD events on the tree can be tested for accuracy under different sets of conditions.

The simulation method was run over two tree types, balanced and caterpillar (Fig. 2) to produce gene trees with correctly labeled WGD nodes. 1,000 trees were produced for every combination of loss and SSD rates = {0.01, 0.009, 0.002, 0.0002, 0.00002} (units of ‘event per million years’) (Additional file 1: Table S1). The values were chosen based on the current estimate of the rate of evolutionary events given in [48] and simulation run time limitations.

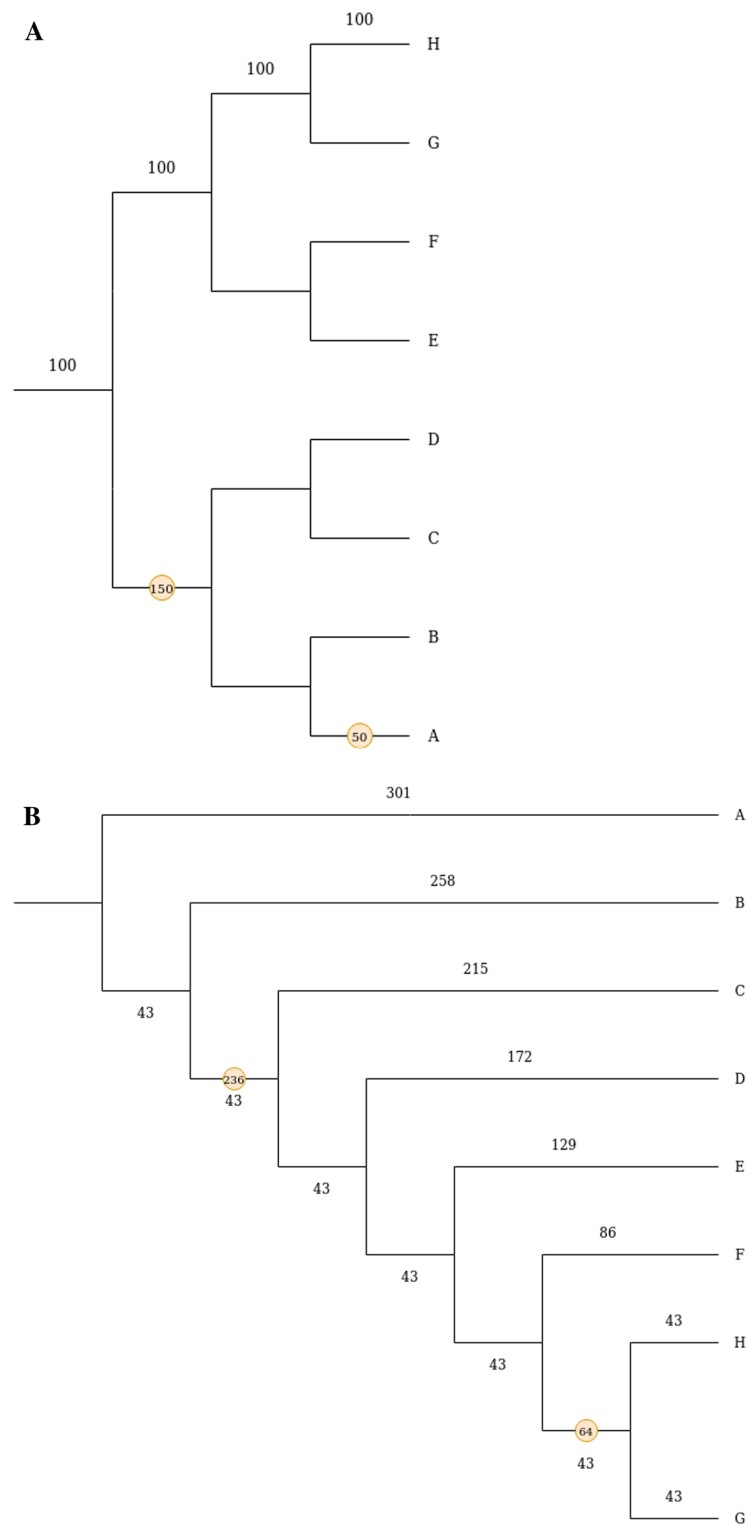


Fig. 2 The trees in this figure represent the two tree types used to generate simulated phylogenetic trees in order to test the accuracy of the inference method. Branch lengths are in millions of year. The dots represent WGD events. The balanced (**A**) and caterpillar (**B**) trees shown here have events one and three speciation events apart respectively

To test if the number of speciation events between WGD events affected the accuracy of the placement, the simulation was run for variations of the balanced and caterpillar tree where WGD events were placed a different number of speciation events apart. 1000 trees for each placement, loss rate and SSD rate were produced (Additional file 1: Tables S2 and S3). When the inference tool was run over the simulated trees to identify the branches inferred to have duplication events, the accuracy was inferred. Accuracy is the total number of correctly placed events divided by the total number of events called on the tree.

Bootstrap analysis

Bootstrap analysis was performed to generate p-values and intervals containing 95% of the data. For the comparative genomic analysis, 1000 bootstrap samples of the 8013 trees containing *P. equestris* were generated. From the simulated data, 1000 bootstrap samples of the 1000 trees for each of the 150 data points was generated.

Software user information

The tool developed in this paper is available on the Python package index under the name WGDTree from <https://github.com/cnickh/wgdtree>. The software provides functions for simulating likely gene tree phylogeny for a given species tree with WGD events, placing WGD events onto gene trees given a labeled species tree, and determining the conditional retention rate of duplicates resulting from WGD events. There is a user guide available with the source code on GitHub. In addition to the user guide there is also example code displaying the expected usage of the functions provided by this package.

Results

Here, WGDTree is evaluated with simulated data to characterize its performance before being run on a plant genome taken from a larger comparative genomic study. The results of these analyses are shown below in presenting the new software.

Simulated data

Using the set of simulated data, the accuracy for each of the trees of different tree topologies, SSD rates and loss rates was determined and is shown in Fig. 3 and is generally high. The tool's accuracy on both caterpillar and balanced trees was higher with small loss rates, regardless of SSD rates, but did even better when the SSD rate was also small. For caterpillar trees, the accuracy increased as the distance (in branch length, also corresponding to the number of intervening speciation events) between the WGD events increased. The balanced tree type allowed for less distance between events making the overall accuracy lower. When distance between the events was the same, the tool performed better on the balanced tree type than on the caterpillar (equivalent data points had non-overlapping 95% confidence intervals in comparisons). With events spaced one speciation apart and high SSD and loss rates (0.01–0.002) accuracy was significantly higher on the balanced tree type (Additional file 1: Tables) (again, equivalent data points had non-overlapping 95% confidence intervals in comparisons).

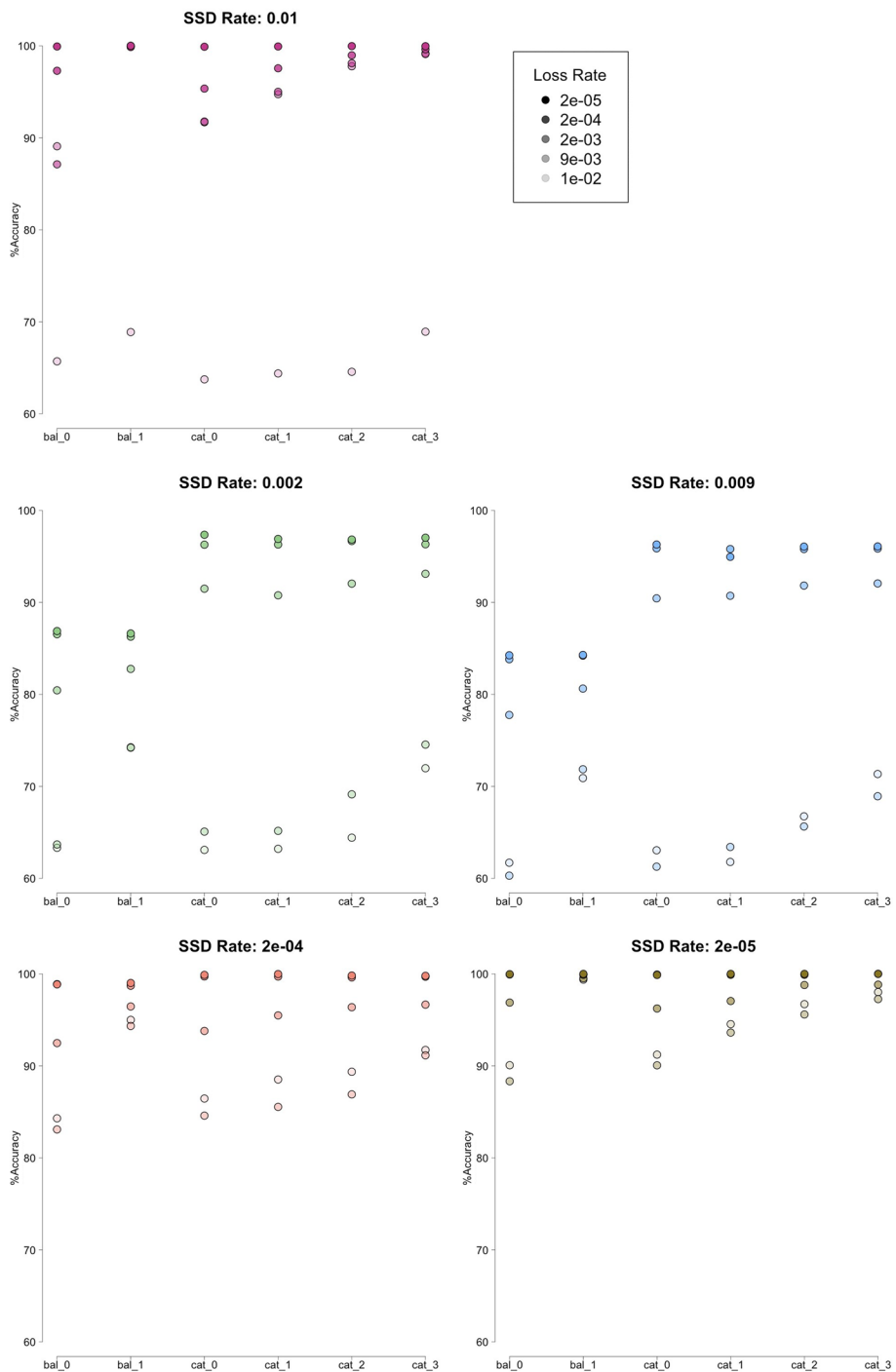


Fig. 3 Each point shows the total accuracy (z) of the inference method's placement of events across all simulated trees for a given species tree, loss (x) and SSD (y) rate in units per million years. The simulation conditions for each data point as described are where "bal_x" and "cat_x" represent a balanced and caterpillar tree type respectively, with WGD events placed "x + 1" speciation events apart on the tree

The four distinct clusters in Fig. 3 are due to using noncontinuous loss and SSD rate parameters for the simulation. The four clusters are correlated with trees generated with; high (0.009, 0.01) loss and ssd rates, low (0–0.002) loss and SSD rates, high loss and low SSD rates, low loss and high SSD rates. The color of the data points indicates tree type.

Application of the inference method to comparative genomic data

To demonstrate the utility of the software tool presented here, one pair of recent whole genome duplication events on the lineage of *P. equestris* was analyzed (Fig. 4). The data was taken from a larger comparative genomic study that will be published elsewhere. Gene trees were created using the described methods. The software tool rooted the trees and calculated the conditional probability of retention of duplicates resulting from WGD. The conditional probability ratio (Pratio) was found to be 0.94, indicating that the calculated statistic is significantly smaller than the Pratio 1 ($p < 0.001$, the limits of the bootstrapping analysis that was performed). Results indicated that genes retained after the first event were not more likely to be retained again after the second event. Values for other genomes in the comparative study will be reported elsewhere as part of a paper focused on the underlying biology.

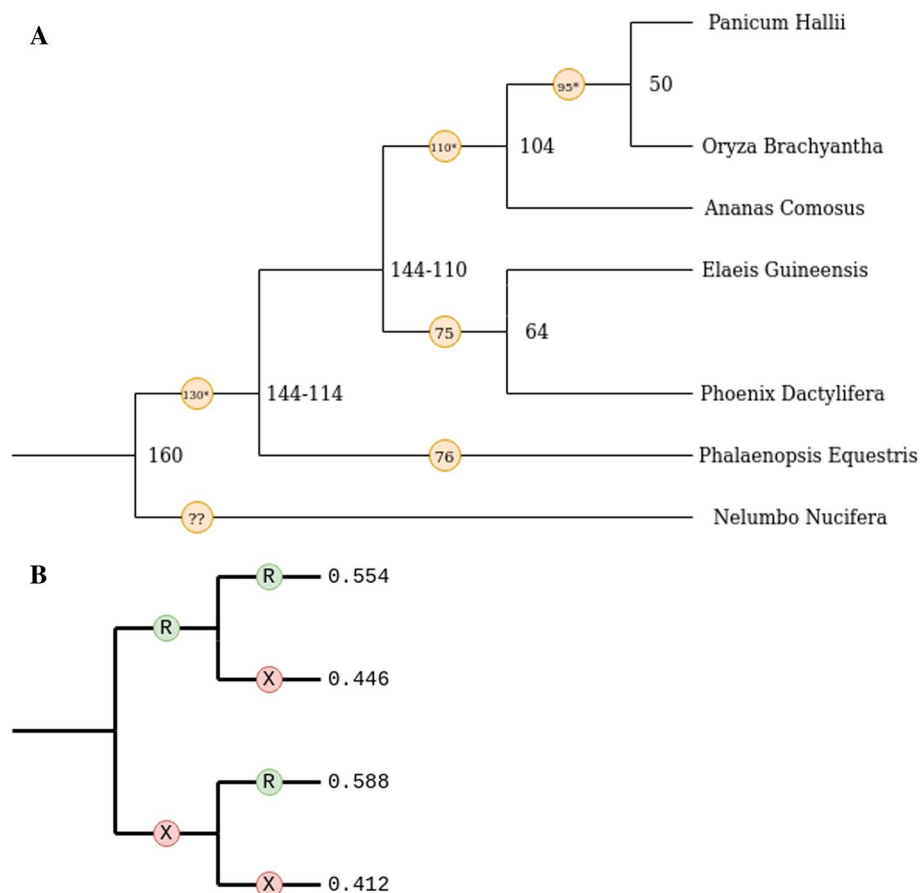


Fig. 4 *P. equestris* Retention Rates. **A** This is the monocot species tree, where all units are in MYA. The events being analyzed are the events at ~130 MYA and 76 MYA. **B** From data analysis using WGDTree, *P. equestris* was found to have a Pratio of 0.94 (0.93–0.96 95% confidence interval)

Discussion

The Pratio for *P. equestris* from two recent WGD events ~130 million years ago (MYA) and 76 MYA was found to be 0.94. This finding is superficially similar to the Pratio found for Atlantic Salmon in previous work, which was ~1. Again, these results support that genes retained after the first event were not more likely to be retained again after the second event. Because the time since the most recent events are similar and the time between the WGD events was significantly shorter than the pair of WGD events explored in Atlantic Salmon, this result is particularly interesting because it potentially could have supported the hypothesis that the probability of retention is independent after consecutive events. In *P. equestris*, the ratio was statistically less than 1, which is not consistent with expectations of independence of retention between the two events. Future work needs to be done to identify more Pratio data points in other species with other WGD events in their lineages to determine if these results are consistent for WGD events of different ages.

The inference tool performed well for a range of tree topologies and SSD rates particularly when loss and small-scale duplication rates were small and when event pairs were placed further apart. Therefore, this software can be used to reliably calculate Pratio values in other lineages. Future modeling studies can use the data generated by this tool to identify the dominant process(es) involved in the retention of duplicate genes. Studying consecutive WGD events provide a unique opportunity to explore the dominant processes involved in the retention of duplicates because they can illuminate the probability each gene will remain as a duplicate after different lengths of time, and after its already been duplicated or lost. The results found in Atlantic Salmon and now *P. equestris* potentially call into question the gene duplicability hypothesis because they do not initially appear to align with the idea that some genes are more duplicable than other genes. Additional modeling is necessary to fully explore the expectations of the gene duplicability hypothesis. If the gene duplicability hypothesis is not supported through model testing, then this could challenge the extent to which the function and complexity of a gene's interactions affect retention of the gene. However, the gene duplicability hypothesis could be supported if the hypothesis also interacts with the law of diminishing returns (either mutational opportunity for different events or reduced selection for events that are mutationally accessible), or even dosage constraints. Dosage balance may play a role in duplicate gene retention, especially in WGD events that are closer together and more recent. Alternatively, the hypothesis that the landscape of mutational opportunity could affect the likelihood of being subfunctionalized or neofunctionalized, affecting the probability of being retained as a duplicate, is a novel hypothesis in relation to gene duplication. Future work will model these hypotheses and conduct model testing on the data generated by the tool presented in this paper and identify what process(es) and to what extent do they affect the probability a gene will be retained as a duplicate. All of this analysis is supported by the software package described here, that produces processed Pratio data for analysis.

One caveat to this method is that homologs that are massively diverged are difficult to identify. Syntenic information would be helpful to incorporate in the algorithm because it would help identify orthologs and paralogs from WGD events [49]. A potential extension to this method could include analysis of syntenic regions in different genomes from genome alignments as a generalized feature, as was performed in the analysis of the Atlantic salmon genome [23].

Conclusions

Here, we presented a useful software tool that is capable of rooting gene trees and reconciling them with species trees and then accurately identifying and differentiating between speciation events, WGD events, SSD events, and loss events. From there, it calculates this statistic developed for evaluating the conditional probability of duplicate gene retention from the second WGD based upon the retention status from the first WGD event. With this tool, the conditional probability ratio for *P. equestris* was determined to be 0.94, which like the conditional probability ratio of Atlantic Salmon does not result in a ratio greater than 1. More species that have undergone two recent duplication events can be identified to provide a large enough dataset for model testing.

Availability and requirements

Project name: WGDTree

Project home page: <https://github.com/cnickh/wgdtree>. This package includes both the analysis tool and the tool for simulating the data that was used to test the analysis software.

Operating system(s): Platform independent

Programming language: Python

Other requirements: 3.9.2 Python version

License: GNU General Public License (GNU GPL)

Any restrictions to use by non-academics: none

Abbreviations

WGD	Whole genome duplication
SSD	Small-scale duplication
Pratio	Conditional probability ratio defined in Eq. 1
NHX	New Hampshire extended file format
MYA	Million years ago

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-022-05042-w>.

Additional file 1. This file contains 6 supplemental tables that present the analysis of simulated data under different sets of parameterizations and different types of trees.

Acknowledgements

We thank other members of the Liberles Research Group for contributions to the bioinformatic pipeline.

Author contributions

CNH wrote software and performed simulations and inference analysis with the software. KP, CSP, and JLM wrote the phylogenetic comparative genomic pipeline, which was run by KP and CSP. AEW and DAL conceived the study and wrote the manuscript, together with KP, YR, and CNH. All authors have read and approved the final manuscript.

Funding

The first author is an undergraduate researcher and funding for this project came from Temple University. Temple University beyond the contribution of authors affiliated with Temple University did not contribute to the design of the study, collection, analysis, and interpretation of data, or writing of the manuscript. No external funding support for this work was received.

Availability of data and materials

Software produced in this manuscript is freely available as described from <https://github.com/cnickh/wgdtree>. This package includes the software for both analysis of data and simulation of data.

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

Competing interests

David Liberles is a Section Editor for another BMC series journal, *BMC Ecology and Evolution*. The other authors declare that they have no competing interests.

Received: 24 March 2022 Accepted: 8 November 2022

Published online: 24 November 2022

References

- Ohno S. Evolution by gene duplication. Berlin Heidelberg: Springer-Verlag; 1970.
- Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000;290:1151–5.
- Freeling M, Thomas BC. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res*. 2006;16:805–14.
- Jin G, Ma P-F, Wu X, Gu L, Long M, Zhang C, et al. New genes interacted with recent whole-genome duplicates in the fast stem growth of bamboos. *Mol Biol Evol*. 2021;38:5752–68.
- Marsit S, Hénault M, Charron G, Fijarczyk A, Landry CR. The neutral rate of whole-genome duplication varies among yeast species and their hybrids. *Nat Commun*. 2021;12:3126.
- Veitia RA. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics*. 2004;168:569–74.
- Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, et al. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol*. 2007;308:58–73.
- Liang H, Plazonic KR, Chen J, Li W-H, Fernández A. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet*. 2008;4: e11.
- Veitia RA. Exploring the etiology of haploinsufficiency. *BioEssays*. 2002;24:175–84.
- Birchler JA, Veitia RA. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell*. 2007;19:395–402.
- 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;16:45.
- Konrad A, Teufel AI, Grahnen JA, Liberles DA. Toward a general model for the evolutionary dynamics of gene duplicates. *Genome Biol Evol*. 2011;3:1197–209.
- Teufel AI, Zhao J, O'Reilly M, Liu L, Liberles DA. On mechanistic modeling of gene content evolution: birth-death models and mechanisms of gene birth and gene retention. *Computation*. 2014;2:112–30.
- Li J-T, Hou G-Y, Kong X-F, Li C-Y, Zeng J-M, Li H-D, et al. The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (*Cyprinus carpio*). *Sci Rep*. 2015;5:8199.
- Geiser C, Mandáková T, Arrigo N, Lysak MA, Parisod C. Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in Buckler mustard. *Plant Cell*. 2016;28:17–27.
- 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;34:2773–91.
- Gillard GB, Grønvald L, Røsaeg LL, Holen MM, Monsen Ø, Koop BF, et al. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol*. 2021;22:103.
- Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. *Nature*. 2003;424:194–7.
- Huang KM, Chain FJJ. Copy number variations and young duplicate genes have high methylation levels in sticklebacks. *Evolution*. 2021;75:706–18.
- Hughes AL. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci*. 1994;256:119–24.
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. 1999;151:1531–45.
- Stoltzfus A. On the possibility of constructive neutral evolution. *J Mol Evol*. 1999;49:169–81.

23. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. *Nature*. 2016;533:200–5.
24. Lynch M, Force AG. The origin of interspecific genomic incompatibility via gene duplication. *Am Nat*. 2000;156:590–605.
25. Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends Genet*. 2006;22:597–602.
26. Stark TL, Liberles DA, Holland BR, O'Reilly MM. Analysis of a mechanistic Markov model for gene duplicates evolving under subfunctionalization. *BMC Evol Biol*. 2017;17:38.
27. Davis JC, Petrov DA. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol*. 2004;2:E55.
28. Hermansen RA, Hvidsten TR, Sandve SR, Liberles DA. Extracting functional trends from whole genome duplication events using comparative genomics. *Biol Proced Online*. 2016;18:11.
29. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*. 2014;5:3657.
30. Meyer A, Van de Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays*. 2005;27:937–45.
31. Eulenstein O, Huzarbazar S, Liberles DA. Reconciling Phylogenetic Trees. In: Dittmar K, Liberles DA, editors. *Evolution after gene duplication*. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2011. p. 185–206.
32. Delabre M, El-Mabrouk N, Huber KT, Lafond M, Moulton V, Noutahi E, et al. Evolution through segmental duplications and losses: a Super-Reconciliation approach. *Algorithms Mol Biol*. 2020;15:12.
33. Bonizzoni P, Della Vedova G, Dondi R. Reconciling gene trees to a species tree. In: Petreschi R, Persiano G, Silvestri R, editors. *Algorithms and Complexity. CIAC 2003. Lecture Notes in Computer Science*, vol 2653. Berlin, Heidelberg: Springer. 2003. https://doi.org/10.1007/3-540-44849-7_18.
34. Berglund-Sonnhammer AC, Steffansson P, Betts MJ, Liberles DA. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J Mol Evol*. 2006;63:240–50. <https://doi.org/10.1007/s00239-005-0096-1>.
35. Wang X, Shi X, Hao B, Ge S, Luo J. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol*. 2005;165:937–46.
36. Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li L-T, et al. Genome of the long-living sacred lotus (*Nelumbo nucifera Gaertn.*). *Genome Biol*. 2013;14:41.
37. Cai J, Liu X, Vanneste K, Proost S, Tsai W-C, Liu K-W, et al. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat Genet*. 2015;47:65–72.
38. Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet*. 2015;47:1435–42.
39. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20–6.
40. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 2017;34:1812–9.
41. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772–80.
42. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59:307–21.
43. Lefort V, Longueville J-E, Gascuel O. SMS: smart model selection in PhyML. *Mol Biol Evol*. 2017;34:2422–4.
44. Arvestad L, Lagergren J, Sennblad B. The gene evolution model and computing its associated probabilities. *J ACM*. 2009;56:1–44.
45. Górecki P, Burleigh GJ, Eulenstein O. Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinform*. 2011;12(Suppl 1):S15.
46. Górecki P, Eulenstein O. DrML: probabilistic modeling of gene duplications. *J Comput Biol*. 2014;21:89–98.
47. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016;33:1635–8. <https://doi.org/10.1093/molbev/msw046>.
48. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*. 2006;22:1269–71.
49. Pary E, Louis A, Cabau C, Guiguen Y, Roest Crollius H, Berthelot C. Synteny-guided resolution of gene trees clarifies the functional impact of whole-genome duplications. *Mol Biol Evol*. 2020;37:3324–37.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.