# Chapter 8
# Evolutionary Rate Change and the Transformation from Additive to Ultrametric: Modal Similarity of Orthologs in Fish and Flower Phylogenomics


**Daniella Santos Muñoz, Eric Lam and David Sankoff**

**Abstract** Branch lengths in a phylogeny may be in units of elapsed time, so that the nodes have dates associated with them, or in units of evolutionary change, such as the number of mutations that have accrued between the two endpoints of a branch. Methods to account for the mutational change in terms of an additive tree are generally incompatible with the ultrametric requirement of time-based tree representations because of changes in mutation rate. There are some principled ways of converting additive trees to ultrametric form, and these suggest which branches have seen increased or decreased rates. We spell these methods out and apply them to the ray-finned fishes and the plant families Solanaceae and Malvaceae. The methods based on nonparametric rate smoothing prove to be more revealing than the Farris transform methods.

**Keywords** Tree metrics · Peaks tree · Fish phylogeny · Solanaceae · Malvaceae

## 8.1 Introduction

Phylogenomics, like phylogenetics, attempts to reconstruct evolutionary history by converting data of various kinds on a set of genomes into a rooted tree whose nodes represent speciation events, historical points at which one existing lineage is split into two or more lineages, each of which continued to evolve independently. Accompanying the tree, which is basically a special combinatorial object consisting of vertices

D. Santos Muñoz · E. Lam · D. Sankoff (✉)
University of Ottawa, Ottawa, Canada
e-mail: sankoff@uottawa.ca

D. Santos Muñoz
e-mail: dsant041@uottawa.ca

E. Lam
e-mail: elam041@uottawa.ca



T. Warnow (ed.), *Bioinformatics and Phylogenetics*, Computational Biology 29,
https://doi.org/10.1007/978-3-030-10837-3_8

(nodes) and edges (branches), we usually associate positive branch lengths. These lengths may be in units of elapsed time, so that the nodes have dates associated with them, or in units of evolutionary change, such as the number of mutations that have accrued between the two endpoints of a branch.

Were evolution clock-like, so that the number of mutations during a period of time was strictly proportional to the duration of that period, the two types of tree would be congruent. They would both be consistent with the ultrametric inequalities, with the given genomes assigned the present time. The number of mutations along each lineage, from the root to each of the present-day genomes, would be the same. Evolution, however, proceeds at an uneven pace, so that the rate of mutation may be elevated in one branch and depressed in another. A phylogenetic representation faithful to the amount of evolution on each branch, such as an additive tree, satisfying only the weaker four-point metric, will not generally be consistent with the ultrametric inequalities; we cannot assign dates to the nodes in any direct way so that the mutational change from the root to each of the present-day genomes is the same. Methods such as neighbor joining (NJ) [29], which produce edge-weighted trees (and hence additive trees) from a matrix of distances between pairs of genomes, will not generally output an ultrametric tree. Methods like UPGMA [24], which produce an ultrametric tree from the same data, will not generally give a good fit to data that are consistent with a lopsided additive tree and are liable to output a wrong tree topology in forcing an ultrametric structure on the data.

Inherent in the non-ultrametric nature of additive data is the existence of branches in the "true" phylogeny that generated the data with very high or very low mutational rates. This suggests that we could compare the length of branches in an additive tree and an ultrametric tree for the same set of data. The problem with this is that methods designed to produce an ultrametric and those generating an additive tree would generally produce different topologies, so that some of the branches in one would not exist in the other, so no comparison would be possible. An alternative would be to adapt an additive tree by stretching or compressing branches in some principled way so that it assumed ultrametric shape. This is the approach we will take here, exploring three alternative published techniques for converting an additive tree to an ultrametric, in order to detect evolution speeding up or slowing down on specific branches.

Recently a new, uniquely phylogenomic, way of deriving evolutionary distances was introduced [33, 36, 37], something that has no counterpart in traditional gene-based phylogeny or even concatenation-of-many-genes extensions of classical methods. This is based on the distribution of *all* syntenically validated ortholog similarities between two genomes. Syntenic validation is assured by software such as SynMap on the CoGe platform [20, 21]. Whole Genome Doubling events and speciation events can be separated out by locating "peaks" (local modes) in this distribution. The most recent (highest similarity) peak is always due to speciation. Although there may be relatively few pairs of orthologous genes with peak similarity, its accuracy is buttressed by the hundreds, thousands, or tens of thousands of ortholog pairs distributed in both sides of it on the x-axis. (Our description is phrased in terms of similarity rather than $K_s$, but either is feasible.) The peak similarities $p$ can be trans-

formed to estimates of evolutionary divergence using a negative log transform, or simply by $1 - p$.

Since the peaks method depends on no one gene family, and does not require any single-copy constraint, it is not susceptible to the rapid expansion of gene families or paralog-based confusion; it is uniquely placed to measure whether entire genomes have adopted a faster or slower pace of evolution. This motivates the application of additive to ultrametric transformations to peaks phylogenies in situations where WGD may muddy the waters when using traditional kinds of distance. The goals of this paper are thus to apply three transformation methods to distance-based phylogenies based on peaks, and to use these techniques to detect increased or diminished rates of evolution in three phylogenetic domains.

## 8.2 Approaches to Transformation

### *8.2.1 Farris Transform*

Originally suggested in [10], the Farris transform is among the early methods for converting an additive tree metric into an ultrametric. Its original implementation, however, was to be used in conjunction with UPGMA [24]. The idea was to correct for the constant evolution rate across all lineages assumption, which may result in incorrect topologies given an additive tree, through the use of an outgroup [17].

Let $X = \{1, 2, \ldots, n\}$ represent the set of present-day taxa, and $T$ a rooted and dated phylogenetic tree on $X$ with vertex set $V_T \supset X$. Assume that for any two present-day taxa $i, j \in X$, the genetic distance between them is represented by $D(i, j)$. Now, consider any vertex $v \in V_T$, and any two distinct taxa $i, j$ of $v$, in the set $X(v) = \{k \in X : v \text{ is an ancestor of } k\}$. The set $X(v)$ contains all present-day descendants of $v$ in $X$. Finally, if we let $a \notin X(v)$ be a known outgroup and $D(i, j)$ be an additive distance matrix for the set of species $X$, then we have the following transformation:

$$D'(i, j) = \overline{D(a)} + \frac{1}{2}\left(D(i, j) - D(i, a) - D(j, a)\right) \tag{8.1}$$

where $\overline{D(a)} = \frac{1}{n}\sum_{i=1}^{n} D(i, a)$, which is the average distance between the outgroup and all ingroups [17]. An alternative constant that can be used is $D = \max_i D(i, a)$, the maximum distance value between the outgroup and any ingroup from the additive distance matrix [12].

**Theorem 1** ([3]) *Let $D$ be an additive distance matrix. If $D'$ is the Farris transform of $D$, then $D'$ is ultrametric.*

For more details on the mathematical context of the Farris transform, see [9].

### 8.2.2 Nonparametric Rate Smoothing (NPRS)

NPRS [31] is a method that estimates divergence times without assuming evolutionary rates are constant across lineages. The method relaxes the assumption of the molecular clock by using a least squares smoothing of local estimates of substitution rates. NPRS is dependent on the minimization of ancestor-descendent local rate changes and is motivated by the likelihood that evolutionary rates are autocorrelated in time. It also estimates divergence times for all unfixed nodes. Sanderson's paper [31] demonstrates that NPRS-produced divergence time estimates are more consistent with paleobotanical evidence than tests using clock-based estimates.

The main idea of Sanderson's method was to construct an estimate of the evolutionary local rate of each branch in a tree while minimizing the difference between that estimate and the local rate estimates of their immediate descendants.

A simple local estimate of rate is

$$\hat{r} = \frac{b}{t} \tag{8.2}$$

where $b$ is the length of branch and $t$ is its temporal duration.

All branches on the path between the root and a terminal node are directed away from the root. We denote by $\hat{r}_k$ the rate on the unique branch directed to node $k$ and $\mathscr{D}(k)$ to the set of nodes immediately descended from node $k$. Now, for each internal node $k$ of the tree, we define

$$w_k = \sum_{j \in \mathscr{D}(k)} \left| \hat{r}_k - \hat{r}_j \right|^2 \tag{8.3}$$

Thus, an overall function to be minimized is then attained by summing the terms over all internal nodes as defined by

$$W = \sum_{k \in \text{internal nodes}} w_k \tag{8.4}$$

A reasonable estimator of local rate will take into account the unknown time endpoints of the branches and some function of the distance matrix, **D**. Hence, we can rewrite the function $W$ as the function $W(t_1, t_2, \ldots, t_m|\mathbf{D})$, where $m$ is the number of internal nodes and $\{t_i\}$ are the unknown times of internal nodes. The minimization of $W$ over these unknown times can then give estimates of those divergence times (comparable to nonparametric regression techniques). The objective is to smooth the local transformations in rate as the rates change over the tree. Equation (8.3) is a minimization of changes in rate from an ancestral lineage to its descendant lineages.

Sanderson [31] estimates the root branch local rate as the mean of all the estimated descendant rates throughout the tree and then constructs an expression similar to Eq. (8.3) for the root node. The mean estimated rate is defined by

$$\hat{r}_{\text{root}} = \frac{1}{n} \sum_{k} \hat{r}_k \tag{8.5}$$

where $n$ is the number of branches. Thus the objective function to be minimized, $W$, should also include the term

$$w_{\text{root}} = \sum_{j \in \mathscr{D}(root)} \left| \hat{r}_{root} - \hat{r}_j \right|^2 . \tag{8.6}$$

### 8.2.3 Penalized Likelihood

Penalized likelihood [32] is a semiparametric smoothing method. It features a trade-off between a parametric model having a different substitution rate on every branch with a nonparametric model which penalizes the model more for rapid rate changes on a tree. This is controlled by an optimality criterion, namely the log likelihood minus $\lambda$ times a roughness penalty, where $\lambda$ is the "smoothing parameter." As smoothing values increase, the variation in rates are smoother and the model is reasonably clock-like; whereas for small values of smoothing, the parametric component dominates and rates heavily vary among branches. Optimal values of the smoothing parameter can be determined by performing cross-validation, a resampling procedure that successively removes each terminal branch and estimates the remaining parameters of the reduced tree using penalized likelihood [11].

Consider a rooted phylogenetic tree with $n$ taxa and $m$ internal nodes, where the root node is labeled by $a$, the remaining internal nodes are labeled by $\{1, \ldots, m-1\}$ and the leaf nodes are labeled by $\{m, \ldots, m+n-1\}$. Branches are labeled by the node they are directed away from, as in the NPRS method. Let node $k$ have an age $t_k$ and its ancestral node, called $anc(k)$, have an age $t_{anc(k)}$. The branch defined by these two nodes has a duration in time given by $t_{anc(k)} - t_k$.

In a clock-like (CL) model, the rate parameters are the same for every branch, $\hat{r}_k = \hat{r}$. In a saturated model (SAT), each branch can have a unique rate, $\hat{r}_k$. The known parameters of each model can be written respectively as $\theta_{\text{CL}} = \left\{t_a, \ldots, t_{m-1}; \hat{r}\right\}$ and $\theta_{\text{SAT}} = \left\{t_a, \ldots, t_{m-1}; \hat{r}_1, \ldots, \hat{r}_{m+n-1}\right\}$, for $m+1$ or $2m+n-1$ free parameters. Next, let $P(x|\xi) = \frac{\xi^x \exp(-\xi)}{x!}$ be the usual probability that an observation $x$ is taken from a Poisson distribution with parameter $\xi$. Then, the log likelihood of $\theta$ for the saturated model is

$$\log L\left(\theta_{\text{SAT}} | x_1, \ldots, x_{m+n-1}\right) = \sum_{k=1}^{m+n-1} \log P\left(x_k | \hat{r}_k \left[t_{anc(k)} - t_k\right]\right) \tag{8.7}$$

The following penalized likelihood to be maximized is given by:

$$\Psi\left(\theta_{\mathrm{SAT}}|x_1,\ldots,x_{m+n-1}\right)=\log L\left(\theta_{\mathrm{SAT}}|x_1,\ldots,x_{m+n-1}\right)-\lambda\Phi\left(\hat{r}_1,\ldots,\hat{r}_{m+n-1}\right) \tag{8.8}$$

where $\Phi$ is a roughness penalty and $\lambda$ is the smoothing parameter previously discussed above. The roughness penalty $\Phi$ should be chosen to reflect change in rate between neighboring branches of the tree. Following the NPRS method above, this penalty was chosen to penalize squared difference in rates between ancestral and descendant branches and the variance in rate between the branches descended from the root node:

$$\Phi\left(\hat{r}_1,\ldots,\hat{r}_{m+n-1}\right)=\sum_{k\notin a\bigcup\mathscr{D}(a)}\left(\hat{r}_k-\hat{r}_{anc(k)}\right)^2+\mathrm{Var}\left(\hat{r}_k:k\in\mathscr{D}(a)\right) \tag{8.9}$$

where $\mathscr{D}(k)$ is the set consisting of the descendants of node $k$. The summation extends to all internal nodes except the root node and the descendants of the root. The second term compares the branches descended from the root node and minimizes the variances of their rates [32].

The choice of $\lambda$ will affect the estimated rates and times. We use cross-validation [11] to assist in choosing the value for this parameter.

We drop each leaf successively, leaving its ancestral node in place and repeat the analysis with the remaining tree, using the **ape** package [26] in R [27]. For the $i$th leaf, the following is calculated:

$$\sum_{j=1}^{m-1}\frac{\left(t_j-t_j^{-i}\right)^2}{t_j} \tag{8.10}$$

where $t_j$ is the estimated date for the $j$th node with the full phylogeny, $t_j^{-i}$ is the estimated date for the $j$th node after removing leaf $i$ from the tree, and lastly, $m$ is the number of internal nodes.

## 8.3 Pipeline

The data for this study were collected from the genomes stored on the CoGe platform [20] for the fish genomes and for some of the others, and from the NCBI genome website for many of the flower genomes; the latter were stored in a private account in CoGe. The SynMap [22] tool in CoGe was used to compare the CDS versions of each of the pairs of the genomes included in the fish, Solanaceae and Malvaceae data sets. Histograms of measures of dissimilarity $(1-p)$, $K_s$ (also written as $D_s$, the rate of synonymous substitutions in the coding sequence) and $\log_{10} K_s$ were compiled for the syntenically validated orthologous gene pairs detected by SynMap, and examined to obtain the location of "peak" frequencies of these measures. These locations were arrayed as a genome distance matrix for each data set, and was used as input to

UPGMA to derive an ultrametric tree and to the NJ algorithm to create an additive tree.

To apply the transformations, we first derived the branch matrix from the NJ results, containing the sum of the branch lengths on the path through the tree for each pair of species. We applied each of the transformations: Farris, NPRS and penalized likelihood, to the branch matrix. Then, NJ was repeated on the transformed data to produce an ultrametric tree for display purposes.

## 8.4 Applications

### *8.4.1 Fish*

Fish phylogeny has long been controversial, particularly in the classification of the perciforme genomes due to the numerous speciation events that have occurred within a relatively short time span. Previously, some groups of perciformes were thought to be monophyletic, but are now considered polyphyletic [5, 13]. To help resolve these issues, we set out to construct a peaks-based phylogeny of 15 fish species, distributed among 10 orders of ray-finned fishes including five perciforme orders, as well as a shark (a cartilaginous fish) and a coelacanth (a lobe-finned fish) (Table 8.1).

Figure 8.1 illustrates the use of histograms derived from SYNMAP analyses. The use of a log scale for the $K_s$ value helps separate the various peaks. In the salmon-

**Table 8.1** Species included in the study

| Order | Species | Common name |
|---|---|---|
| Chimaeriformes | *Callorhinchus milii* [35] | Elephant shark |
| Coelacanthiformes | *Latimeria chalumnae* [1] | African coelacanth |
| Semionotiformes | *Lepisosteus oculatus* [6] | Spotted gar |
| Characiformes | *Astyanax mexicanus* [23] | Mexican tetra |
| Cypriniformes | *Danio rerio* [14] | Zebrafish |
| Esociformes | *Esox lucius* [28] | Northern pike |
| Salmoniformes | *Salmo salar* [19] | Atlantic salmon |
| Pleuronectiformes | *Cynoglossus semilaevis* [8] | Tongue sole |
| Perciformes | *Oreochromis niloticus* [7] | Nile tilapia |
| Gasterosteiformes | *Gasterosteus aculeatus* [16] | Three-spined stickleback |
| Tetraodontiformes | *Takifugu rubripes* [2] | Torafugu |
| | *Tetraodon nigroviridis* [15] | Spotted green pufferfish |
| | *Mola mola* [25] | Ocean sunfish |
| Cyprinodontiformes | *Poecilia reticulata* [18] | Guppy |
| | *Xiphophorus maculatus* [34] | Southern platyfish |

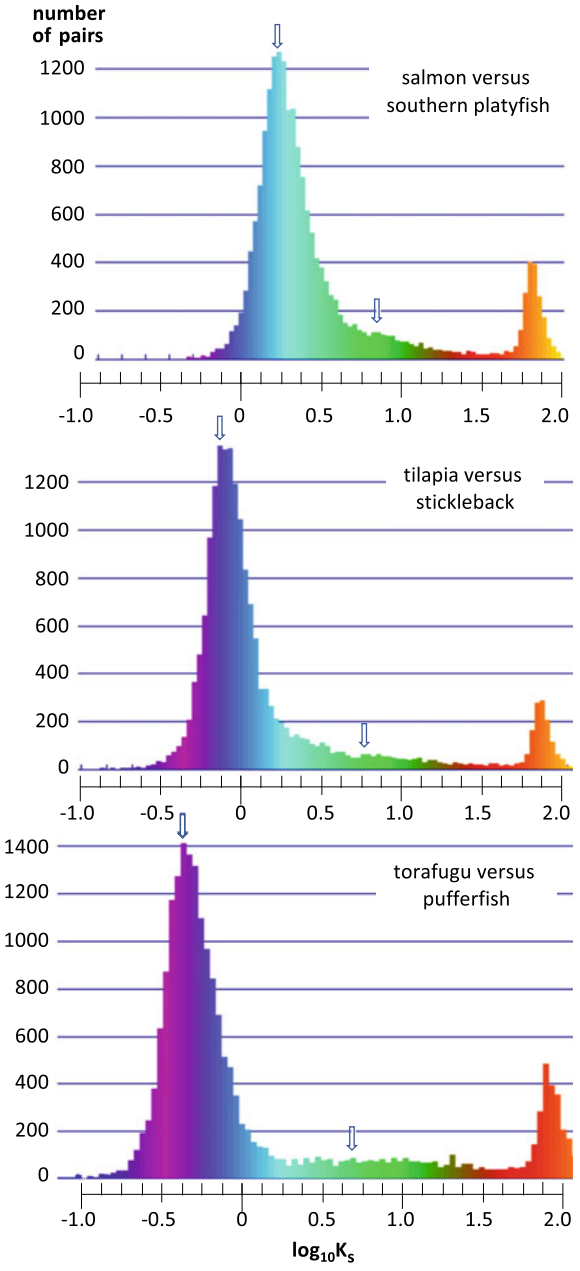



**Fig. 8.1** Distribution of log $K_s$ of duplicate gene pairs in salmon-platyfish, tilapia-stickleback and torafugu-pufferfish comparisons, showing a common WGD before speciation. Arrows indicate speciation peaks and WGD peaks. Peaks at far right represent noise (gene fragments, common domains in unrelated genes, etc.) accumulating in exponentially larger intervals, as well as earlier vertebrate WGDs

platyfish comparison, we can identify the speciation peak at $\log_{10} K_s = 0.25$ and a WGD that, being pre-speciation, is shared by both genomes, with $\log_{10} K_s$ in the range of 0.76–0.81. The salmonid WGD, occurs later than this speciation, and so does not show up as a peak. The visible peak is the "teleost WGD" at the root of the ray-finned fish radiation, and is confirmed in the other two comparisons, with tilapia versus stickleback ($\log_{10} K_s$ range 0.75–81) and torafugu versus pufferfish (diffuse peak for $\log_{10} K_s$ between 0.68 and 1.0), respectively. More important for our purposes are the more precisely defined speciation peaks: salmon-platyfish, tilapia-stickleback, and torafugu-pufferfish at $\log_{10} K_s = 0.25, -0.15$, and $-0.4$, respectively.

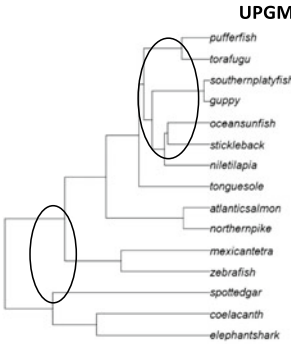
We use the matrix of speciation peak locations (in $K_s$ terms) in all $\binom{15}{2} = 105$ comparisons as the input to UPGMA and a Java-implemented NJ algorithm. We filled in the one missing data point, for the northern pike-spotted gar comparison (due to very sparse $K_s$ information) with the salmon-spotted gar value, since salmon and northern pike paralleled each other in all the other comparisons. The ultrametric tree produced by UPGMA and the additive tree output by NJ, as well as two of its transforms, are depicted in Fig. 8.2. The Farris transformation analysis failed, for reasons explained in Sect. 8.5 below.

The tree produced by UPGMA is a good illustration of why we avoid dating evolutionary events by directly producing an ultrametric from a distance matrix. The topology of the tree should first be determined by methods less sensitive to violations of the constant rate condition. Adjusting the branch length while retaining the topology produces a more meaningful timing of event history. The fish UPGMA produces a biologically thoroughly implausible sister group to the ray-finned fishes consisting of the elephant shark and the coelacanth, while it also scrambles aspects of perciform evolution.

The NJ tree and its transforms, on the other hand, are in complete accord with the current understanding of fish phylogeny. This is true of every branching in the tree, although the tongue sole should probably be grouped with tilapia, guppy and platyfish, rather than as an outgroup to these and the tetradontiform/stickleback clade [4, 5, 30].

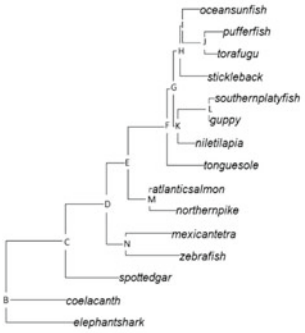
The ratios of transformed branch lengths suggest conservative evolutionary tendencies in the earliest branching fish species (elephant shark, coelacanth and spotted gar), since these branches need to be multiplied by a relatively large factor in the transition to ultrametric, although the lineage represented by the internal branches leading through the ray-finned fish (including spotted gar) to the teleosts (including the remaining ten species) underwent rapid evolution, since these branches need to be multiplied by a very small factor in the transformation to an ultrametric representation. Salmon and guppy both seem much more conservative than their sister species, northern pike and platyfish, respectively, although the common ancestor of salmon and pike also appears to have been conservative.

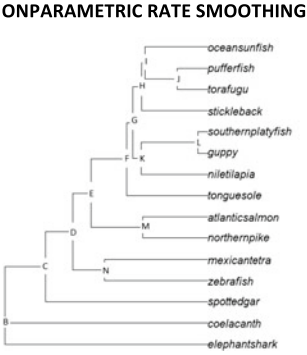
Compared to nonparametric rate smoothing, the roughness penalty in the penalized likelihood method attenuates somewhat the discordance between salmon and pike and between guppy and platyfish.




**RATIOS**

| | NPRS | PL |
|---|---|---|
| Elephant shark | 1.02 | 1.02 |
| Coelacanth | 1.14 | 1.14 |
| Spotted gar | 1.03 | 1.02 |
| Zebrafish | 0.65 | 0.64 |
| Mexican tetra | 0.76 | 0.75 |
| Salmon | 7.61 | 5.33 |
| Northern pike | 0.81 | 0.57 |
| Tongue sole | 0.74 | 0.70 |
| Tilapia | 1.28 | 1.19 |
| Guppy | 3.40 | 2.45 |
| Platyfish | 0.61 | 0.44 |
| Stickleback | 0.81 | 0.77 |
| Torafugu | 0.80 | 0.76 |
| Pufferfish | 0.55 | 0.53 |
| Sunfish | 1.19 | 1.13 |
| B-C | 0.23 | 0.24 |
| C-D | 0.23 | 0.24 |
| D-N | 0.54 | 0.52 |
| D-E | 0.28 | 0.30 |
| E-M | 0.81 | 1.05 |
| E-F | 0.30 | 0.30 |
| F-G | 0.34 | 0.34 |
| G-K | 0.61 | 0.65 |
| K-L | 0.61 | 0.59 |
| G-H | 0.44 | 0.42 |
| H-I | 0.49 | 0.48 |
| I-J | 0.49 | 0.45 |



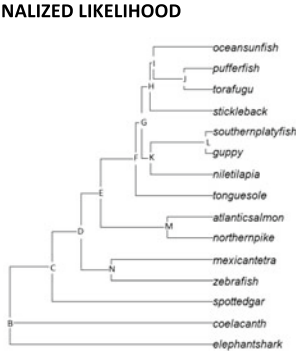




**Fig. 8.2** UPGMA tree, NJ tree and its transformations by nonparametric smoothing and penalized likelihood methods, for the fish data. "Ratios" summarize transformed branch lengths versus original NJ values. Ovals surround regions of UPGMA tree that contradict accepted biological knowledge

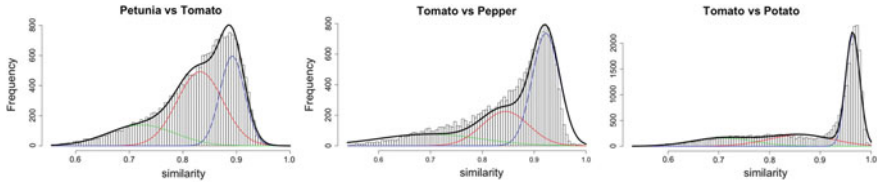



**Fig. 8.3** Solanaceae Whole Genome Doubling (WGD) and speciation peaks

### *8.4.2 Solanaceae*

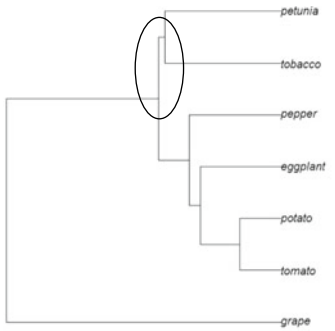
The Solanaceae family, here represented by tomato, potato, eggplant, pepper, tobacco, and petunia, descends from two whole genome triplications events, the first one, 120 million years ago affecting all core eudicots, and the second one, specific to this family [37], dating from around $40 \pm 10$ million years ago. The distributions of gene pair similarities thus have three peaks, two dating from the ancient polyploidization events, and one from speciation, as seen in Fig. 8.3. As seen in the previous section, more recent peaks tend to be less dispersed. This holds not only within each comparison, but also for the speciation comparisons, where potato is the most closely related to tomato, and petunia is the most distantly related.

As with the fish data, Fig. 8.4 shows an error in the UPGMA tree, this time for our sample of plants from the Solanaceae family; petunia and tobacco are grouped together separately from pepper and the *Solanum* species: tomato, potato, and eggplant. The NJ tree and its transforms correctly place tobacco as branching from the other species *after* petunia does.

The branch length ratios indicate that evolutionary change has accelerated in the *Solanum*, though not dramatically. This trend emerges most clearly in the nonparametric rate smoothing and penalized likelihood methods. All three methods show that after the speciation event leading to petunia, the ancestor of the other genomes underwent rapid evolutionary change.

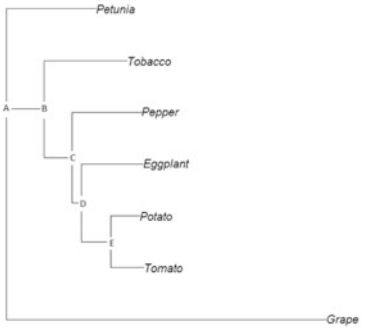
### *8.4.3 Malvaceae*

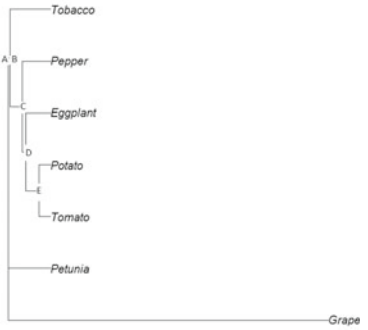
Genomes that have been sequenced in the Malvaceae include cacao, monkey cacao, jute (two species), durian, hibiscus and cotton (two species). Once again, there is a discrepancy between the UPGMA tree and the NJ tree. Figure 8.5 shows that durian branches after jute in the lineage toward cocoa in the UPGMA tree, and before jute in that lineage in the NJ tree. In this case, however, it may very well be that the UPGMA tree is correct, and that the NJ tree results from statistical fluctuations involving the cotton genome [36]. In any case, there is as yet no consensus in the literature about the placement of durian.




**RATIOS**

| | Farris | NPRS | PL |
|---|---|---|---|
| Grape | 1.15 | 0.57 | 0.57 |
| Tomato | 0.40 | 1.13 | 1.33 |
| Potato | 0.46 | 1.28 | 1.51 |
| Eggplant | 0.46 | 1.23 | 1.45 |
| Pepper | 0.47 | 1.56 | 1.54 |
| Tobacco | 0.57 | 1.81 | 1.99 |
| Petunia | 0.54 | 2.04 | 2.04 |
| A-B | 0.05 | 0.90 | 0.52 |
| D-E | 0.50 | 1.32 | 1.54 |
| C-D | 0.50 | 3.55 | 1.86 |
| B-C | 0.50 | 1.49 | 2.05 |



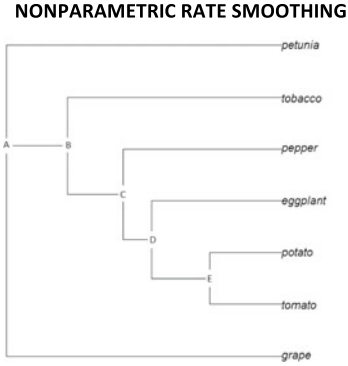
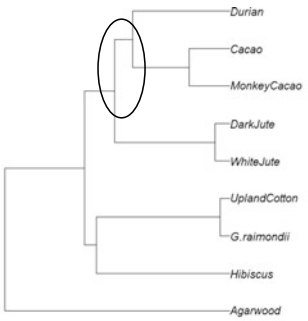





**Fig. 8.4** Ultrametric, NJ tree and its transformations by the Farris, nonparametric rate smoothing and penalized likelihood methods, for the family Solanaceae and outgroup grape. "Ratios" summarize transformed branch lengths versus original NJ values

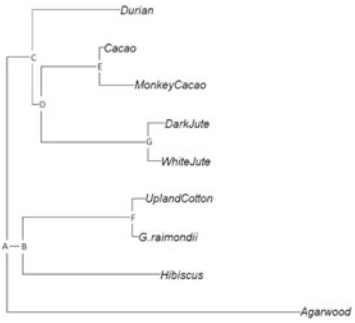
**UPGMA**



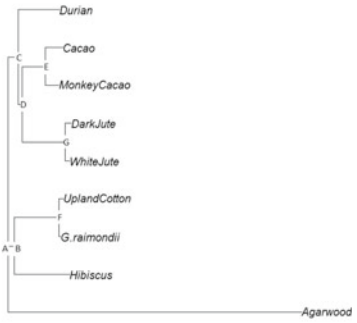

**RATIOS**

| | Farris | NPRS | PL |
|---|---|---|---|
| Agarwood | 1.24 | 0.67 | 0.67 |
| Hibiscus | 0.49 | 1.13 | 1.25 |
| G. raimondii | 0.35 | 5.68 | 1.96 |
| Upland Cotton | 0.43 | 2.79 | 0.96 |
| White Jute | 0.42 | 2.94 | 1.31 |
| Dark Jute | 0.44 | 2.20 | 0.98 |
| Monkey Cacao | 0.47 | 2.03 | 1.46 |
| Cacao | 0.48 | 1.58 | 1.13 |
| Durian | 0.58 | 1.82 | 1.65 |
| A-B | 0.50 | 2.66 | 1.61 |
| B-F | 0.50 | 1.08 | 1.46 |
| A-C | 0.50 | 1.44 | 2.05 |
| C-D | 0.50 | 1.86 | 1.34 |
| D-G | 0.50 | 1.00 | 1.10 |
| D-E | 0.50 | 1.26 | 1.43 |

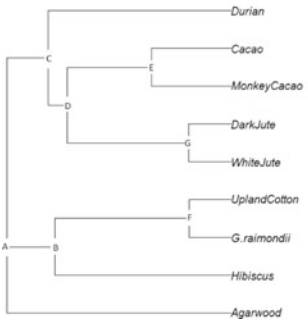
**NEIGHBOUR JOINING**




**FARRIS**




**NONPARAMETRIC RATE SMOOTHING**




**PENALIZED LIKELIHOOD**

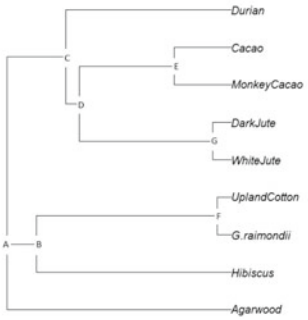



**Fig. 8.5** Ultrametric, NJ tree and its transformations by the Farris, nonparametric smoothing and penalized likelihood methods, for the family Malvaceae and outgroup agarwood. "Ratios" summarize transformed branch lengths versus original NJ values

In examining the branch length ratios, there is little that emerges that is consistent across all three methods, except that *G. raimondii* is somewhat conservative. Nonparametric rate smoothing and penalized likelihood both have differential rates of evolution at the origin of the two main clades, but they differ on which side is more conservative, with nonparametric rate smoothing pointing to the cacao side, and with penalized likelihood suggesting the cotton/hibiscus clade. In addition, the ancestor of the jute species appears to have undergone accelerated evolution.

## 8.5 Discussion

Preliminary to the evaluation of the transformation protocols, we demonstrated in all three evolutionary domains that the NJ with the peaks approach recovers the correct topology for the phylogenetic tree, and that the UPGMA was not able to. (Although the correct topology is not known for the Malvaceae, we do know that the UPGMA result was not consistent with the additive results.)

In applying the three transformations, the two that incorporate smoothing between coincident branches were clearly preferable to the Farris method. In addition, penalized likelihood stood out as being able to achieve ultrametric status with the least stretching or contracting of branches.

The Farris transform displaces much of the rate variation to the branch leading to the outgroup. In the case of the fish phylogeny in Fig. 8.2, this could not be achieved because the outgroup was too conservative, and other branches could not be shortened enough by the transformation without becoming negative.

Finally, we have demonstrated that the non-ultrametric status of additive trees, at least on these data sets, may be resolved by identifying one or two branches on which evolutionary rate has clearly slowed or accelerated. It would be useful to be able to confirm these cases with data on individual gene trees.

**Acknowledgements** This work was supported in part by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics.

## References

1. Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H., MacCallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., Organ, C., Chalopin, D., Smith, J.J., Robinson, M., Dorrington, R.A., Gerdol, M., Aken, B., Biscotti, M.A., Barucca, M., Baurain, D., Berlin, A.M., Blatch, G.L., Buonocore, F., Burmester, T., Campbell, M.S., Canapa, A., Cannon, J.P., Christoffels, A., De Moro, G., Edkins, A.L., Fan, L., Fausto, A.M., Feiner, N., Forconi, M., Gamieldien, J., Gnerre, S., Gnirke, A., Goldstone, J.V., Haerty, W., Hahn, M.E., Hesse, U., Hoffmann, S., Johnson, J., Karchner, S.I., Kuraku, S., Lara, M., Levin, J.Z., Litman, G.W., Mauceli, E., Miyake, T., Mueller, M.G., Nelson, D.R., Nitsche, A., Olmo, E., Ota, T., Pallavicini, A., Panji, S., Picone, B., Ponting, C.P., Prohaska, S.J., Przybylski, D., Saha, N.R., Ravi, V., Ribeiro,

F.J., Sauka-Spengler, T., Scapigliati, G., Searle, S.M.J., Sharpe, T., Simakov, O., Stadler, P.F., Stegeman, J.J., Sumiyama, K., Tabbaa, D., Tafer, H., Turner-Maier, J., van Heusden, P., White, S., Williams, L., Yandell, M., Brinkmann, H., Volff, J.N., Tabin, C.J., Shubin, N., Schartl, M., Jaffe, D.B., Postlethwait, J.H., Venkatesh, B., Di Palma, F., Lander, E.S., Meyer, A., Lindblad-Toh, K.: The African coelacanth genome provides insights into tetrapod evolution. Nature **496**, 311 (2013)

2. Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.m., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D.S., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J.K., Doggett, N., Zharkikh, A., Tavtigian, S.V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., Brenner, S.: Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science **297**(5585), 1301–1310 (2002). https://doi.org/10.1126/science.1072104
3. Bandelt, H.J.: Recognition of tree metrics. SIAM J. Discrete Math. **3**(1), 1–6 (1990)
4. Betancur-Rodriguez, R., Ortí, G., Pyron, R.A., Morlon, H.: Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes. Ecol. Lett. **18**(5), 441–450 (2015). https://doi.org/10.1111/ele.12423
5. Betancur-Rodriguez, R., Wiley, E.O., Arratia, G., Acero, A., Bailly, N., Miya, M., Lecointre, G., Ortí, G.: Phylogenetic classification of bony fishes. BMC Evol. Biol. **17**(1), 162 (2017). https://doi.org/10.1186/s12862-017-0958-3
6. Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A.M., Campbell, M.S., Barrell, D., Martin, K.J., Mulley, J.F., Ravi, V., Lee, A.P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro, C., Sydes, J., Beaudry, F.E.G., Sun, Y., Hertel, J., Beam, M.J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J.H., Litman, G.W., Litman, R.T., Mikami, M., Ota, T., Saha, N.R., Williams, L., Stadler, P.F., Wang, H., Taylor, J.S., Fontenot, Q., Ferrara, A., Searle, S.M.J., Aken, B., Yandell, M., Schneider, I., Yoder, J.A., Volff, J.N., Meyer, A., Amemiya, C.T., Venkatesh, B., Holland, P.W.H., Guiguen, Y., Bobe, J., Shubin, N.H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., Postlethwait, J.H.: The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. **48**, 427 (2016)
7. Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H.J., Russell, P., Aken, B., Alföldi, J., Amemiya, C., Azzouzi, N., Baroiller, J.F., Barloy-Hubler, F., Berlin, A., Bloomquist, R., Carleton, K.L., Conte, M.A., D'Cotta, H., Eshel, O., Gaffney, L., Galibert, F., Gante, H.F., Gnerre, S., Greuter, L., Guyon, R., Haddad, N.S., Haerty, W., Harris, R.M., Hofmann, H.A., Hourlier, T., Hulata, G., Jaffe, D.B., Lara, M., Lee, A.P., MacCallum, I., Mwaiko, S., Nikaido, M., Nishihara, H., Ozouf-Costaz, C., Penman, D.J., Przybylski, D., Rakotomanga, M., Renn, S.C.P., Ribeiro, F.J., Ron, M., Salzburger, W., Sanchez-Pulido, L., Santos, M.E., Searle, S., Sharpe, T., Swofford, R., Tan, F.J., Williams, L., Young, S., Yin, S., Okada, N., Kocher, T.D., Miska, E.A., Lander, E.S., Venkatesh, B., Fernald, R.D., Meyer, A., Ponting, C.P., Streelman, J.T., Lindblad-Toh, K., Seehausen, O., Di Palma, F.: The genomic substrate for adaptive radiation in African cichlid fish. Nature **513**, 375 (2014)
8. Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., Song, W., An, N., Chalopin, D., Volff, J.N., Hong, Y., Li, Q., Sha, Z., Zhou, H., Xie, M., Yu, Q., Liu, Y., Xiang, H., Wang, N., Wu, K., Yang, C., Zhou, Q., Liao, X., Yang, L., Hu, Q., Zhang, J., Meng, L., Jin, L., Tian, Y., Lian, J., Yang, J., Miao, G., Liu, S., Liang, Z., Yan, F., Li, Y., Sun, B., Zhang, H., Zhang, J., Zhu, Y., Du, M., Zhao, Y., Schartl, M., Tang, Q., Wang, J.: Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat. Genet. **46**, 253 (2014)
9. Dress, A., Huber, K.T., Moulton, V.: Some uses of the Farris transform in mathematics and phylogenetics—a review. Ann. Comb. **11**, 1–37 (2007)
10. Farris, J.: On the phenetic approach to vertebrate classification. In: Major Patterns in Vertebrate Evolution, pp. 823–850 (1997)

11. Green, P., Silverman, B.: Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman and Hall (1994)
12. Gusfield, D.: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press (1997)
13. Harrington, R.C., Faircloth, B.C., Eytan, R.I., Smith, W.L., Near, T.J., Alfaro, M.E., Friedman, M.: Phylogenomic analysis of crangimorph fishes reveals flatfish asymmetry arose in a blink of the evolutionary eye. BMC Evol. Biol. **16**(1), 224 (2016). https://doi.org/10.1186/s12862-016-0786-x
14. Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.J., White, S., Chow, W., Kilian, B., Quintais, L.T., Guerra-Assunção, J., Zhou, Y., Gu, Y., Yen, J., Vogel, J.H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S.F., Laird, G.K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Elliott, D., Threadgold, G., Harden, G., Ware, D., Begum, S., Mortimore, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Lloyd, C., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J.D., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C.M., Ersan-Ürün, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberländer, M., Rudolph-Geiger, S., Teucke, M., Lanz, C., Raddatz, G., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Schuster, S.C., Carter, N.P., Harrow, J., Ning, Z., Herrero, J., Searle, S.M.J., Enright, A., Geisler, R., Plasterk, R.H.A., Lee, C., Westerfield, M., de Jong, P.J., Zon, L.I., Postlethwait, J.H., Nüsslein-Volhard, C., Hubbard, T.J.P., Crollius, H.R., Rogers, J., Stemple, D.L.: The zebrafish reference genome sequence and its relationship to the human genome. Nature **496**, 498 (2013)
15. Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., de Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K.J., McEwan, P., Bosak, S., Kellis, M., Volff, J.N., Guigó, R., Zody, M.C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E.S., Weissenbach, J., Roest Crollius, H.: Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature **431**, 946 (2004)
16. Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Team, B.I.G.S.P..W.G.A., Lander, E.S., Di Palma, F., Lindblad-Toh, K., Kingsley, D.M.: The genomic basis of adaptive evolution in threespine sticklebacks. Nature **484**, 55 (2012)
17. Krane, D., Raymer, M.: Fundamental Concepts of Bioinformatics. The Genetics Place Series. Benjamin Cummings (2003)

18. Künstner, A., Hoffmann, M., Fraser, B.A., Kottler, V.A., Sharma, E., Weigel, D., Dreyer, C.: The genome of the Trinidadian guppy, Poecilia reticulata, and variation in the Guanapo population. PLoS ONE **11**(12), 1–25 (2016). https://doi.org/10.1371/journal.pone.0169087
19. Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., von Schalburg, K., Rondeau, E.B., Di Genova, A., Samy, J.K.A., Olav Vik, J., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Inge Våge, D., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A.J., Tooming-Klunderud, A., Jakobsen, K.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J.M., Jonassen, I., Maass, A., Omholt, S.W., Davidson, W.S.: The Atlantic salmon genome provides insights into rediploidization. Nature **533**, 200 (2016)
20. Lyons, E., Freeling, M.: How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. **53**(4), 661–673 (2008). https://doi.org/10.1111/j.1365-313X.2007.03326.x
21. Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D.M., Freeling, M.: Finding and comparing syntenic regions among *arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids. Plant Physiol. **148**, 1772–1781 (2008)
22. Lyons, E., Pedersen, B., Kane, J., Freeling, M.: The value of nonmodel genomes and an example using synmap within coge to dissect the hexaploidy that predates the rosids. Trop. Plant Biol. **1**(3), 181–190 (2008). https://doi.org/10.1007/s12042-008-9017-y
23. McGaugh, S.E., Gross, J.B., Aken, B., Blin, M., Borowsky, R., Chalopin, D., Hinaux, H., Jeffery, W.R., Keene, A., Ma, L., Minx, P., Murphy, D., O'Quin, K.E., Rétaux, S., Rohner, N., Searle, S.M.J., Stahl, B.A., Tabin, C., Volff, J.N., Yoshizawa, M., Warren, W.C.: The cavefish genome reveals candidate genes for eye loss. Nat. Commun. **5**, 5307 (2014)
24. Michener, C.D., Sokal, R.R.: A quantitative approach to a problem in classification. Evolution **11**(2), 130–162 (1957). https://doi.org/10.1111/j.1558-5646.1957.tb02884.x
25. Pan, H., Yu, H., Ravi, V., Li, C., Lee, A.P., Lian, M.M., Tay, B.H., Brenner, S., Wang, J., Yang, H., Zhang, G., Venkatesh, B.: The genome of the largest bony fish, ocean sunfish (Mola mola), provides insights into its fast growth rate. GigaScience **5**(1), 36 (2016). https://doi.org/10.1186/s13742-016-0144-3
26. Paradis, E., Claude, J., Strimmer, K.: APE: analyses of phylogenetics and evolution in R language. Bioinformatics **20**, 289–290 (2004)
27. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). http://www.R-project.org/
28. Rondeau, E.B., Minkley, D.R., Leong, J.S., Messmer, A.M., Jantzen, J.R., von Schalburg, K.R., Lemon, C., Bird, N.H., Koop, B.F.: The genome and linkage map of the northern pike (Esox lucius): conserved synteny revealed between the salmonid sister group and the neoteleostei. PLoS ONE **9**(7), 1–18 (2014). https://doi.org/10.1371/journal.pone.0102089
29. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**(4), 406–425 (1987)
30. Sanciangco, M.D., Carpenter, K.E., Betancur-Rodriguez, R.: Phylogenetic placement of enigmatic Percomorph families (teleostei: Percomorphaceae). Mol. Phylogenet. Evol. **94**, 565–576 (2016). https://doi.org/10.1016/j.ympev.2015.10.006
31. Sanderson, M.J.: A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. **14**(12), 1218–1231 (1997)
32. Sanderson, M.J.: Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. **19**(1), 101–109 (2002). https://doi.org/10.1093/oxfordjournals.molbev.a003974
33. Sankoff, D., Zheng, C., Zhang, Y., Meidanis, J., Lyons, E., Tang, H.: Models for similarity distributions of syntenic homologs and applications to phylogenomics. IEEE/ACM Trans. Comput. Biol. Bioinform. (2018). https://doi.org/10.1109/TCBB.2018.2849377
34. Schartl, M., Walter, R.B., Shen, Y., Garcia, T., Catchen, J., Amores, A., Braasch, I., Chalopin, D., Volff, J.N., Lesch, K.P., Bisazza, A., Minx, P., Hillier, L., Wilson, R.K., Fuerstenberg, S.,

Boore, J., Searle, S., Postlethwait, J.H., Warren, W.C.: The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. Nat. Genet. **45**, 567 (2013)
35. Venkatesh, B., Lee, A.P., Ravi, V., Maurya, A.K., Lian, M.M., Swann, J.B., Ohta, Y., Flajnik, M.F., Sutoh, Y., Kasahara, M., Hoon, S., Gangu, V., Roy, S.W., Irimia, M., Korzh, V., Kondrychyn, I., Lim, Z.W., Tay, B.H., Tohari, S., Kong, K.W., Ho, S., Lorente-Galdos, B., Quilez, J., Marques-Bonet, T., Raney, B.J., Ingham, P.W., Tay, A., Hillier, L.W., Minx, P., Boehm, T., Wilson, R.K., Brenner, S., Warren, W.C.: Elephant shark genome provides unique insights into Gnathostome evolution. Nature **505**, 174 (2014)
36. Zhang, Y., Zheng, C., Islam, S., Kim, Y.M., Sankoff, D.: Branching out to speciation with a birth-and-death model of fractionation: the Malvaceae. IEEE/ACM Trans. Comput. Biol. Bioinform. (2019)
37. Zhang, Y., Zheng, C., Sankoff, D.: Speciation and rate variation in a birth-and-death account of WGD and fractionation; the case of Solanaceae. In: Proceedings of the 16th RECOMB Comparative Genomics Satellite Workshop. Lecture Notes in Computer Science 11183 (2018)