Effects of discordance between species and gene trees on phylogenetic diversity conservation

Kristina Wicke¹² · Mareike Fischer³ · Laura Kubatko⁴

Received: 20 December 2021 / Revised: 16 November 2022 / Accepted: 22 November 2022 /
Published online: 9 December 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Phylogenetic diversity indices such as the Fair Proportion (FP) index are frequently discussed as prioritization criteria in biodiversity conservation. They rank species according to their contribution to overall diversity by taking into account the unique and shared evolutionary history of each species as indicated by its placement in an underlying phylogenetic tree. Traditionally, phylogenetic trees were inferred from single genes and the resulting gene trees were assumed to be a valid estimate for the species tree, i.e., the “true” evolutionary history of the species under consideration.

However, nowadays it is common to sequence whole genomes of hundreds or thousands of genes, and it is often the case that conflicting genealogical histories exist in different genes throughout the genome, resulting in discordance between individual gene trees and the species tree. Here, we analyze the effects of gene and species tree discordance on prioritization decisions based on the FP index. In particular, we consider the ranking order of taxa induced by (i) The FP index on a species tree, and (ii) The expected FP index across all gene tree histories associated with the species tree.

On the one hand, we show that for particular tree shapes, the two rankings always coincide. On the other hand, we show that for all leaf numbers greater than or equal to five, there exist species trees for which the two rankings differ. Finally, we illustrate

Kristina Wicke
kristina.wicke@njit.edu

Mareike Fischer
email@mareikefischer.de

Laura Kubatko
lkubatko@stat.osu.edu

¹ Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, USA
² Department of Mathematics, The Ohio State University, Columbus, OH, USA
³ Institute of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany
⁴ Department of Statistics, Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH, USA
the variability in the rankings obtained from the FP index across different gene tree and species tree estimates for an empirical multilocus mammal data set.

**Keywords**  Biodiversity conservation · Fair Proportion index · Gene tree · Multispecies coalescent · Phylogenetic diversity · Species tree

**MSC Classification**  O5C05 · 92B05 · 92D15 · 60J90

## 1 Introduction

Phylogenetic diversity indices such as the Fair Proportion (FP) index (Isaac et al. 2007; Redding 2003), Equal Splits (ES) index (Redding and Mooers 2006), and Shapley value (Haake et al. 2008), have become popular tools for prioritizing species for conservation (e.g., Isaac et al. (2007); Redding and Mooers (2006); Redding et al. (2008, 2014); Vane-Wright et al. (1991); Vellend et al. (2011)). They quantify the importance of species for overall biodiversity based on their placement in an underlying phylogenetic tree and can thus, next to other criteria such as threat status, serve as prioritization criteria in conservation decisions. While these indices differ in their definitions and properties, they all depend on the underlying phylogenetic tree and its branch lengths. In particular, different phylogenetic trees for the same set of species or identical phylogenetic trees with different branch lengths, may result in different prioritization orders obtained by any of these phylogenetic diversity indices.

Until recently, phylogenetic trees were usually inferred from single genes and the resulting gene trees were assumed to be a valid estimate for the species tree, i.e., the “true” evolutionary history of the species under consideration. However, with the advent of rapid sequencing technologies in recent years, it is now common to sequence whole genomes composed of hundreds or thousands of genes, and it is well known that due to various biological phenomena, such as lateral gene transfer and incomplete lineage sorting, conflicting genealogical histories exist in different genes throughout the genome, resulting in discordance between individual gene trees and the species tree (e.g., Maddison (1997); Nichols (2001); Pamilo and Nei (1988)).

In this paper, we analyze the effects of gene and species tree discordance on prioritization decisions based on the FP index. We chose the FP index as it is employed (under the name ‘evolutionary distinctiveness’) in the so-called “EDGE of Existence” project established by the Zoological Society of London, a conservation initiative focusing specifically at threatened species that represent a high amount of unique evolutionary history (Isaac et al. (2007); see also https://www.edgeofexistence.org/). Moreover, while the FP index itself lacks a direct biological motivation, it coincides with the so-called Shapley value, a result that was first shown by Fuchs and Jin (2015). The Shapley value, originally a concept from cooperative game theory introduced by Lloyd Shapley (Shapley 1953), on the other hand, reflects the average contribution of a species to overall diversity and is characterized by certain desirable axioms (see, e.g., Haake et al. (2008)) that motivate its use (and hence that of the FP index) as a prioritization criterion in conservation planning.
Moreover, we focus on gene tree-species tree incongruence caused by incomplete lineage sorting by employing the multispecies coalescent model (Degnan and Rosenberg 2009). Under this model, given a species tree (an ultrametric phylogenetic tree with branch lengths), a distribution on possible gene tree histories evolving within the species tree is obtained, and the species tree as well as the distribution of gene tree histories are the basis for our analysis. More precisely, we compare the prioritization order of taxa induced by the FP index on the species tree with the prioritization order obtained when averaging the FP index over all gene tree histories evolving within the species tree.

The structure of this paper is as follows. We first introduce all relevant notation and concepts. We then analyze circumstances under which the FP index on a given species tree and the expected FP index across all gene tree histories evolving within the species tree induce the same ranking (Theorem 1). Afterwards, we show that for all \( n \geq 5 \), there exists a species tree on \( n \) leaves such that the two rankings (i.e., the ranking obtained from the species tree and the ranking obtained from averaging over all gene tree histories) differ (Theorem 2). Finally, we complement our theoretical results by analyzing the variability in the rankings obtained from the FP index across different gene trees and species trees for an empirical multilocus mammal data set. We end by discussing our results and indicating directions for future research.

# 2 Notation and preliminaries

To formally state our results, we need some terminology and notation. Throughout this manuscript, \( X \) denotes a non-empty finite set (of taxa or species). If not stated otherwise, \( X = \{x_1, \ldots, x_n\} \).

## 2.1 Species trees, gene tree histories, and probabilities under the multispecies coalescent model

A *species tree* \( S = (S, \tau) \) on \( X = \{x_1, \ldots, x_n\} \) is a pair consisting of a rooted binary phylogenetic tree \( S \) with leaf set \( X \) (i.e., a rooted binary tree whose leaves are identified with \( X \)) and a vector of interval lengths \( \tau = (\tau_0, \tau_1, \ldots, \tau_{n-2}) \), where \( \tau_i \) denotes the length of the interval between the \( i^{th} \) and \( (i+1)^{th} \) speciation event in \( S \) measured in *coalescent units*. Note that we assume here that no two speciation events in \( S \) occur at exactly the same time (this is motivated by the fact that under typical models for species trees, the probability for two speciation events to occur simultaneously is zero). Moreover, note that the length of the interval above the root of \( S \), denoted by \( \tau_{n-1} \), is infinite. Finally, note that the definition of \( \tau \) as a vector of interval lengths between speciation times implies that \( S = (S, \tau) \) is *equidistant* or *ultrametric*, meaning that the length of the path from the root of \( S \) to any leaf of the species tree is the same. This is also commonly referred to as the *molecular clock assumption*. Moreover, note that we sometimes refer to the length of the path from the root of \( S \) to any leaf as the *height* of \( S \).
Given a species tree $S = (S, \tau)$, we use the concept of a gene tree history $h$ associated with $S$ to indicate a sequence of coalescent events as well as the populations and time intervals in which they occur. Informally, a gene tree history $h$ can be thought of as a phylogenetic tree with leaf set $X$ evolving within $S$ (Fig. 1). Note that the set of gene tree histories associated with $S$ is finite, and we denote it by $H(S)$.

Moreover, we associate with each history $h$ a vector of coalescent times $t_h = (t_1, \ldots, t_{n-1})$, where, moving back in time, two distinct lineages coalesce or merge into a single lineage. For example, referring to Fig. 1, at time $t_1$, the lineages corresponding to $x_1$ and $x_2$ merge into a single lineage. Note that there are infinitely many choices for the values $t_1, \ldots, t_{n-1}$ satisfying the constraints of a history $h$ evolving within $S$, and in the following we will often integrate over all possible coalescent times consistent with $h$. Moreover, note that the probability that two coalescences occur at exactly the same time is zero, and thus we have $t_i \neq t_j$ for all $i, j \in \{1, \ldots, n-1\}$. Thus the coalescent times induce a ranking of the interior vertices of a gene tree history $h$ according to their distance from the root, where the interior vertex with the largest distance (i.e., the vertex corresponding to the most recent coalescence event) is assigned rank 1 and the root itself is assigned rank $n-1$. In particular, we consider two histories to be distinct if their coalescent times induce different rankings of the interior vertices (Fig. 2). Note that this notion is closely related to the concept of ranked gene tree histories (Degnan et al. 2012a, b) (similarly, ranked species trees are defined as species tree topologies together with an order on the speciation events). Moreover, note that when we refer to a history $h$, we usually mean $h$ and its associated vector of coalescent times $t_h$ as well as the information in which time interval each coalescent event occurred. Finally, note that we explicitly keep track of the time intervals and, next to the induced rankings of interior vertices, distinguish histories based on this information. For instance, considering the top part of Fig. 2, the scenario where the coalescent event $t_1$ occurs in time interval $\tau_1$ (rather than $\tau_2$) in the population ancestral to $x_1$ and $x_2$, is considered a different history, $h_3$ say, that is distinct from $h_1$.

For a given species tree $S = (S, \tau)$ and a particular (ranked gene tree) history $h \in H(S)$ with associated vector of coalescent times $t_h$, the probability of observing $h$ given $S$ under the multispecies coalescent model can be found by integrating over the joint density of coalescence times, which we denote by $f_h(h, t_h)$, where the limits of integration are determined by $t_h$ (i.e., we integrate over all possible coalescent times consistent with $h$). We refer the reader to Degnan et al. (2012a, b) for a careful derivation of this density. Essentially, it is a product of exponential terms arising from the occurrence of exponential distributions as waiting times to coalescent events. More precisely, for a sample of $k$ lineages from a single population, the waiting time to coalescence is exponentially distributed with rate $\frac{1}{k(k-1)}$. The probability that there is no coalescence for a sample of $k$ lineages in a time interval of length $\tau$ is then $e^{-\frac{\tau}{k(k-1)}}$. Based on this, the density $f_h(h, t_h)$ is obtained by multiplying exponential probabilities across the branches and time intervals of $S$. If there are $m_i \geq 0$ coalescences in an interval $\tau_i$ of $S$, the interval is subdivided into $m_i + 1$ subintervals and densities contributed by the subintervals are multiplied (again, see Degnan et al. (2012a, b) for a more rigorous definition; see also Fig. 2 for an example).
Fig. 1 Species tree $S = (S, \tau)$ on $X = \{x_1, \ldots, x_5\}$ and history $h$ evolving within $S$. Note that $S$ and $h$ depict different evolutionary relationships. Intervals of time between speciation events (horizontal lines) are denoted by the $\tau_i$ (left side of the figure), while times corresponding to coalescent events (black dots) are denoted by the $t_i$ (coalescent times are measured from the most recent speciation event). Circled numbers indicate a ranking on the speciation events with 1⃝ being the most recent and 3⃝ the most ancient speciation event, and $S$ together with this ordering defines a ranked species tree.

As an example, consider the species tree $S = (S, \tau)$ and the gene tree history $h \in \mathcal{H}(S)$ depicted in Fig. 1. First consider the “left” subtree of $S$ containing leaves $x_1$ and $x_2$. Here, $x_1$ and $x_2$ fail to coalesce in an interval of length $\tau_1$, contributing a factor of $e^{-\tau_1}$ to the density $f_h(h, t_h)$. These lineages then coalesce at time $t_1$ within the next interval, contributing a factor of $e^{-t_1}$ to the density. Similarly, for the “right” subtree of $S$, lineages $x_4$ and $x_5$ fail to coalesce in an interval of length $\tau_2$ yielding a factor of $e^{-\tau_2}$. The interval $\tau_3$ then contains a coalescent event at time $t_2$, which contributes the term $e^{-3t_2}$ (since three lineages were available to coalesce). The remaining two lineages fail to coalesce for the remainder of this interval, contributing a term of $e^{-(\tau_3-t_2)}$. Finally, considering $\tau_4$, the most recent coalescent event occurs at time $t_3$, thus contributing a factor of $e^{-(t_4-t_3)}$. Thus, in this case, we have

$$f_h(h, t_h) = e^{-\tau_1} e^{-t_1} e^{-\tau_2} e^{-3t_2} e^{-(\tau_3-t_2)} e^{-3t_3} e^{-(t_4-t_3)} = e^{-\tau_1-\tau_2-\tau_3-t_1-2t_2-2t_3-t_4}.$$

Finally, the limits of integration are determined by the vector of coalescent times $t_h$ and the vector of interval lengths $\tau$. In the example discussed above, $t_1$ can range from 0 to $\tau_2$, $t_2$ can range from 0 to $\tau_3$, $t_3$ can range from 0 to $t_4$, and $t_4$ can range from 0 to infinity. This leads to

$$\mathbb{P}(h|S) = \int f_h(h, t_h) \, dt_h = \int_{t_4=0}^{\infty} \int_{t_3=0}^{t_4} \int_{t_2=0}^{t_3} \int_{t_1=0}^{t_2} e^{-\tau_1-\tau_2-\tau_3-t_1-2t_2-2t_3-t_4} \, dt_1 \, dt_2 \, dt_3 \, dt_4.$$
Fig. 2  Two distinct histories $h_1$ and $h_2$ evolving within the same species tree. Note that $h_1$ and $h_2$ have the same topology and their coalescent events occur in the same time intervals of the species tree, but the rankings of the interior vertices of $h_1$, respectively $h_2$, induced by vectors of coalescent times $t_{h_1}$ and $t_{h_2}$ differ. Moreover, both $h_1$ and $h_2$ are such that there are two coalescence events in interval $\tau_2$. By subdividing this interval into three subintervals, namely $[0, t_1]$, $[t_1, t_2]$, and $[t_2, \tau_2]$, and multiplying densities contributed by subintervals, we e.g. obtain

$$f_{h_1}(h_1, t_{h_1}) = e^{-\tau_1} \cdot e^{-t_1} \cdot e^{-(t_2 - t_1)} \cdot e^{-t_3} = e^{-\tau_1 - t_2 - t_3}$$

$$= \frac{1}{6} e^{-\tau_1 - 2\tau_2 - 3\tau_3} - \frac{1}{6} e^{-\tau_1 - \tau_2 - 3\tau_3} - \frac{1}{6} e^{-\tau_1 - 2\tau_2 - \tau_3} + \frac{1}{6} e^{-\tau_1 - \tau_2 - \tau_3}.$$ 

2.2 Lowest common ancestors and subtrees

Given a species tree $S$ (a history $h$) with root $\rho$ and two leaves $x_i, x_j \in X$, we refer to the last vertex in $S (h)$ that lies on both the path from $\rho$ to $x_i$ as well as the path from $\rho$ to $x_j$ as the lowest common ancestor of $x_i$ and $x_j$ and denote it as $lca(x_i, x_j)$. Moreover, given a vertex $v$ of $S (h)$, we refer to the subtree of $S (h)$ rooted at vertex $v$ as a pendant subtree of $S (h)$.

2.3 Exchangeability

First, recall that a pair of leaves, say $x_i$ and $x_j$, of a phylogenetic $X$-tree is called a cherry, denoted $[x_i, x_j]$, if $x_i$ and $x_j$ share a common parent. Considering the coalescent
process for some species tree \( S \) on \( X \) and a time interval \( \tau \), we call two leaves, \( x_i \) and \( x_j \) say, *exchangeable* (with respect to \( \tau \)) if they “enter” the same branch of the species tree \( S \) as single lineages in interval \( \tau \). In particular, from this point onward (back in time), \( x_i \) and \( x_j \) are “indistinguishable” and every history \( h \) with the property that neither \( x_i \) nor \( x_j \) coalesce before \( \tau \) either contains \([x_i, x_j]\) as a cherry, or if such a history \( h' \) does not contain the cherry \([x_i, x_j]\), there exists a history \( h'' \) which only differs from \( h' \) by a permutaiton of \( x_i \) and \( x_j \) and such that \( t_{h'} = t_{h''} \) and \( f_{h'}(h', t_{h'}) = f_{h''}(h'', t_{h''}) \).

As an example, for the species tree \( S \) shown in Fig. 1, leaves \( x_3, x_4 \), and \( x_5 \) are exchangeable with respect to interval \( \tau_3 \) if they enter the branch ancestral to \( x_3, x_4 \) and \( x_5 \) as single lineages in interval \( \tau_3 \), as is the case for the history depicted. The history \( h \) depicted contains \([x_3, x_4]\) as a cherry, but there also exist histories, \( h' \) and \( h'' \) say, where \( x_3 \) and \( x_4 \) do not form a cherry, but such that \( h' \) and \( h'' \) only differ by a permutation of \( x_3 \) and \( x_4 \), and such that \( t_{h'} = t_{h''} \) and \( f_{h'}(h', t_{h'}) = f_{h''}(h'', t_{h''}) \) (for instance, \( h' \) could be the history identical to \( h \) except that the coalescence at time \( \tau_2 \) involves \( x_3 \) and \( x_5 \), in which case the “matching” history \( h'' \) would be the history identical to \( h \) except that the coalescence at time \( \tau_2 \) involves \( x_4 \) and \( x_5 \)).

We will sometimes also consider exchangeability of subtrees (rather than single leaves) defined in the following way: Let \( S \) be a species tree. Then, we call subtrees \( T_{x_i} \) and \( T_{x_j} \) exchangeable if one of the following two conditions is met: Either (a) each history \( h \in \mathcal{H}(S) \) contains \( T_{x_i} \) and \( T_{x_j} \) as a “subtree cherry” (i.e., the roots of \( T_{x_i} \) and \( T_{x_j} \) have a common parent and the two trees form a pendant subtree \((T_{x_i}, T_{x_j})\) of \( h \)); or (b) if a history \( h' \) does not contain \((T_{x_i}, T_{x_j})\) as a pendant subtree, there exists a history \( h'' \) identical to \( h' \) except that the roles of \( T_{x_i} \) and \( T_{x_j} \) are exchanged and such that \( t_{h'} = t_{h''} \) and \( f_{h'}(h', t_{h'}) = f_{h''}(h'', t_{h''}) \).

This exchangeability property and the idea of pairing matching histories will be crucial for many proofs.

### 2.4 Caterpillar and pseudocaterpillar trees

Two particular tree topologies that will be relevant in the following are the *caterpillar tree* and the *pseudocaterpillar tree* (Fig. 3). Let \( n \geq 2 \). Then, a phylogenetic \( X \)-tree on \( X = \{x_1, \ldots, x_n\} \) is called a *caterpillar tree* if and only if it has precisely one cherry. Moreover, if \( n \geq 4 \), a phylogenetic \( X \)-tree on \( X = \{x_1, \ldots, x_n\} \) is called a *pseudocaterpillar tree* (adapted from Degnan et al. (2012a)\(^1\)) if it has precisely two cherries and those two cherries form a pendant subtree of size four.

### 2.5 The Fair proportion index

Let \( S \) be a rooted phylogenetic \( X \)-tree (not necessarily binary) with root \( \rho \) and leaf set \( X = \{x_1, \ldots, x_n\} \), where each edge \( e \) is assigned a non-negative length \( l(e) \in \mathbb{R}_{\geq 0} \). Then, the *Fair Proportion (FP) index* (Isaac et al. 2007; Redding 2003) for leaf \( x_i \in X \)

\(^1\) Note that Degnan et al. (2012a) defined pseudocaterpillar trees for \( n \geq 5 \).
is defined as

\[ FP_S(x_i) = \sum_{e \in P(S; \rho, x_i)} \frac{l(e)}{n(e)}, \]

where \( P(S; \rho, x_i) \) denotes the path in \( S \) from the root \( \rho \) to leaf \( x_i \) and \( n(e) \) is the number of leaves descended from \( e \). Note that \( \sum_{x_i \in X} FP_S(x_i) = \sum_{e \in S} l(e) \), i.e., the FP index apportions the total sum of branch lengths of \( S \) (also referred to as the ‘phylogenetic diversity’ of \( X \) (Faith 1992)) among the taxa in \( X \).

Note that we can calculate the FP index both for a given species tree \( S \) as well as for all histories \( h \in \mathcal{H}(S) \) (where in case of the latter, the FP index is expressed in terms of both speciation and coalescent times). We denote these variants as \( FP_S(x_i) \) and \( FP_h(x_i) \), respectively. As an example, for the species tree \( S = (S, \tau) \) and the history \( h \) with its vector \( t_h \) of coalescent times shown in Fig. 1, and the leaf \( x_3 \), we have...
The expected FP index for leaf $x_i$ on history $h$ given the species tree $S$ can be expressed as:

$$E[FP_h(x_i)|S] = \int FP_h(x_i) f_h(h, t_h) dt_h,$$



where $f_h(h, t_h)$ denotes the probability density of history $h$ and $t_h$ is the vector of coalescent times for $h$ and as before, the limits of integration are determined by the particular locations of the coalescent events of $h$ within the species tree. Based on this, we define the expected FP index of leaf $x_i$ on species tree $S$ as the expected FP index across all gene tree histories $h$ associated with $S$. More formally,

$$E[FP(x_i)|S] = \sum_{h \in \mathcal{H}(S)} E[FP_h(x_i)|S] = \sum_{h \in \mathcal{H}(S)} \int FP_h(x_i) f_h(h, t_h) dt_h. \quad (3)$$

Note that when two leaves, $x_i$ and $x_j$ say, are exchangeable with respect to some interval $\tau$, and we consider the subset of histories, $\mathcal{H}^\tau(S) \subseteq \mathcal{H}(S)$ say, containing all histories with the property that $x_i$ and $x_j$ enter the same branch of the species tree as single lineages in interval $\tau$, we have

$$E[FP(x_i) - FP(x_j)|S] = \sum_{h \in \mathcal{H}^\tau(S)} \int [FP_h(x_i) - FP_h(x_j)] f_h(h, t_h) dt_h = 0. \quad (4)$$

This is due to the following two observations: If a history $h \in \mathcal{H}^\tau(S)$ contains the cherry $[x_i, x_j]$, then by definition of the FP index, we have $FP_h(x_i) - FP_h(x_j) = l(e_i) - l(e_j)$, where $e_i$ denotes the pendant edge incident with $x_i$ and $e_j$ denotes the pendant edge incident with $x_j$. However, as all species trees and thus all histories considered in this manuscript are ultrametric, we necessarily have $l(e_i) = l(e_j)$, implying that $FP_h(x_i) - FP_h(x_j) = 0$. If, on the other hand, a history $h'$ does not contain the cherry $[x_i, x_j]$, there exists a history $h''$ which only differs from $h'$ by a permutation of $x_i$ and $x_j$ and such that $t_{h'} = t_{h''}$ and $f_{h'}(h', t_{h'}) = f_{h''}(h'', t_{h''})$. In particular, $FP_h'(x_i) = FP_{h''}(x_j)$ and $FP_h'(x_j) = FP_{h''}(x_i)$ and these terms cancel each other in the summation of Eq. (4). We will freely use this fact throughout the manuscript.
2.6 Rankings

The FP index induces a natural ranking of the elements in $X$, where taxa are ordered according to their FP index. Here, a ranking $r$ is an assignment of ranking numbers to the elements of $X$, where for any pair of taxa $x_i, x_j \in X$, $x_i$ either receives a higher or lower ranking number than $x_j$, or the ranking numbers are equal (we then call $x_i$ and $x_j$ tied). We say that a function $f : X \to \mathbb{R}$ induces a ranking $r_f(X)$ if the ranking number of $x_i$ in $r_f$ is smaller than the ranking number of $x_j$ precisely if $f(x_i) > f(x_j)$. If $f(x_i) = f(x_j)$ for some $x_i \neq x_j$, $x_i$ and $x_j$ receive the same ranking number. In this case, we use the so-called ‘standard competition ranking’ (also referred to as ‘1224’ ranking), where tied elements receive the same ranking number, say $\hat{r}$, and the subsequent ranking number is $\tilde{r} = \hat{r} + t - 1$, where $t$ denotes the number of tied elements with ranking number $\hat{r}$. In the following, we will consider the ranking of $X$ induced by the FP index on the species tree, denoted as $r(S, FP)(X)$, and the ranking of $X$ induced by the expected FP index, denoted as $r(S, E(FP))(X)$, and analyze the relationship between these two rankings.

3 Results

We compare the rankings induced by the FP index on a given species tree $S$ with the ranking induced by the expected FP index on $S$. We show that under certain circumstances the rankings always coincide, while there also exist cases for which the rankings differ.

3.1 Circumstances under which the FP index on the species tree and the expected FP index induce the same ranking

In this section, we consider certain tree topologies for which the rankings $r(S, FP)(X)$ and $r(S, E(FP))(X)$ are guaranteed to coincide. We begin by showing that if a species tree $S$ on $X = \{x_1, \ldots, x_n\}$ contains a pendant subtree $S'$ on a subset $X' \subseteq X$, $X' = \{x_1, \ldots, x_m\}$ with $3 \leq m \leq n$ say, and such that the two maximal pendant subtrees of $S'$ consist of $m - 1$ leaves labeled $x_1, \ldots, x_{m-1}$ and 1 leaf labeled $x_m$, respectively, then among these $m$ leaves, leaf $x_m$ has both the highest FP index on $S$ as well as the highest expected FP index. Note that we require $m \geq 3$, as for $m = 2$, the subtree $S'$ would be a cherry of $S$, in which case we would have $FP_S(x_1) = FP_S(x_2)$ and $E[FP(x_1)|S] = E[FP(x_2)|S]$ due to the exchangeability of $x_1$ and $x_2$.

**Proposition 1** Let $n \geq 3$ and let $S = (S, \tau)$ be a species tree on $X = \{x_1, \ldots, x_n\}$ containing a pendant subtree $S'$ on $3 \leq m \leq n$ leaves, which in turn consists of a maximal pendant subtree on $m - 1$ leaves, $x_1, \ldots, x_{m-1}$ say, and a maximal pendant subtree on one leaf, $x_m$ say, (Fig. 4). Then,

(i) $FP_S(x_m) > FP_S(x_i)$ for all $x_i \in \{x_1, \ldots, x_{m-1}\}$.

(ii) $E[FP(x_m)|S] > E[FP(x_i)|S]$ for all $x_i \in \{x_1, \ldots, x_{m-1}\}$.
Proof: (i) First, note that all edges of \( S \) above interval \( \tau^*_1 \) (see Fig. 4) contribute the same amount, \( \Delta_{FP} \) say, to \( FP_S(x_m) \) and \( FP_S(x_i) \) for \( x_i \in \{x_1, \ldots, x_{m-1}\} \). In particular, for taxon \( x_m \), we have \( FP_S(x_m) = \tau^*_0 + \tau^*_1 + \Delta_{FP} \), whereas for \( x_i \in \{x_1, \ldots, x_{m-1}\} \), we have \( FP_S(x_i) \leq \tau^*_0 + \frac{\tau^*_1}{m-1} + \Delta_{FP} < FP_S(x_m) \) (since the subtree of \( S' \) containing \( x_i \) has \( m - 1 \geq 2 \) leaves). This completes the first part of the proof.

(ii) Let \( x_i \in \{x_1, \ldots, x_{m-1}\} \). In order to show that \( \mathbb{E}[FP(x_m)|S] > \mathbb{E}[FP(x_i)|S] \), we show that

\[
\mathbb{E}[FP(x_m) - FP(x_i)|S] = \sum_{h \in \mathcal{H}(S)} \int \left[ FP_h(x_m) - FP_h(x_i) \right] f_h(h, t_h) dt_h > 0.
\]

Note that here (as everywhere in this manuscript) the limits of integration are determined by the particular locations of coalescent events of each history within the species tree. We now analyze the set of histories \( \mathcal{H}(S) \) more in-depth. In particular, we analyze them with respect to the interval \( \tau^*_2 \) in \( S \). Note that \( x_i \) enters interval \( \tau^*_2 \) as part of some subtree, \( T_{x_i} \), say, where the size of \( T_{x_i} \) can range from 1 to \( m - 1 \), whereas \( x_m \) enters interval \( \tau^*_2 \) as a subtree of size 1, \( T_{x_m} \) say.

However, subtrees \( T_{x_i} \) and \( T_{x_m} \) are exchangeable as defined on page 7 (see also Fig. 5). Let \( lca(x_i, x_m) \) denote the lowest common ancestor of \( x_i \) and \( x_m \) in the corresponding gene tree history, and consider the FP indices of \( x_i \) and \( x_m \).

(a) If \( h \) is such that \( T_{x_i} \) and \( T_{x_m} \) form a pendant subtree of \( h \) with root \( lca(x_i, x_m) \), all edges of \( h \) above \( lca(x_i, x_m) \) (i.e., all edges in the path from the root of \( h \) to \( lca(x_i, x_m) \)) contribute the same amount to \( FP_h(x_i) \) and \( FP_h(x_m) \). However,
considering the paths from \( lca(x_i, x_m) \) to \( x_i \) and \( x_m \) of length \( \tau_{0}^{*} + \tau_{1}^{*} + t_{lca} \), where \( t_{lca} \) denotes the time of the coalescence event that leads to \( lca(x_i, x_m) \), the contribution of edges in the path from \( lca(x_i, x_m) \) to \( x_m \) to \( FP_{h}(x_m) \) is precisely \( \tau_{0}^{*} + \tau_{1}^{*} + t_{lca} \), whereas the contribution of edges in the path from \( lca(x_i, x_m) \) to \( x_i \) to \( FP_{h}(x_i) \) is less or equal to \( \tau_{0}^{*} + \tau_{1}^{*} + t_{lca} \) (where the inequality is strict if \( T_{x_i} \) contains at least two leaves). In total, this implies that we have \( FP_{h}(x_m) \geq FP_{h}(x_i) \).

(b) Now, if \( h' \) and \( h'' \) are two histories that differ only by a permutation of the roles of \( T_{x_i} \) and \( T_{x_m} \), we first note that the lowest common ancestor of \( x_i \) and \( x_m \) corresponds to the same vertex \( lca(x_i, x_m) \) in \( h' \) and \( h'' \) (see Fig. 5b). Moreover, as in case (a), all edges above \( lca(x_i, x_m) \) contribute the same amount towards \( FP_{h'}(x_m), FP_{h'}(x_i), FP_{h''}(x_m), \) and \( FP_{h''}(x_i) \). Now, as \( T_{x_i} \) contains at least one leaf, whereas \( T_{x_m} \) contains precisely one leaf, referring to the notation in Fig. 5, we can conclude that the contribution of edges in the path from \( lca(x_i, x_m) \) to \( u \) in \( h'' \) towards \( FP_{h''}(x_m) \) is greater than or equal to the contribution of the corresponding edges in the path from \( lca(x_i, x_m) \) to \( u \) in \( h' \) towards \( FP_{h'}(x_i) \). Moreover, similar to case (a), the contribution of edges in the path from \( u \) to \( x_m \) in \( h'' \) towards \( FP_{h''}(x_m) \) is greater than or equal to the contribution of edges in the path from \( u \) to \( x_i \) in \( h' \) towards \( FP_{h'}(x_i) \). In summary, \( FP_{h''}(x_m) \geq FP_{h''}(x_i) \). An analogous argument shows that \( FP_{h'}(x_m) \geq FP_{h'}(x_i) \). In particular note that the inequalities are strict if \( T_{x_i} \) contains at least two leaves.

Thus, we can partition the set of histories \( \mathcal{H}(S) \) into the subset of histories of type (a), in which \( T_{x_i} \) and \( T_{x_m} \) form a pendant subtree and we have \( FP_{h}(x_m) \geq FP_{h}(x_i) \), and thus

\[
\int_{\geq 0} (FP_{h}(x_m) - FP_{h}(x_i)) f_{h}(h, t_{h}) dt_{h} \geq 0,
\]

and into the subset of histories of type (b), i.e., into pairs of “matching” histories \((h', h'')\), where, with the reasoning from above, in particular, recalling that \( t_{h'} = t_{h''} \) and \( f_{h'}(h', t_{h'}) = f_{h''}(h'', t_{h''}) \), we have for each such pair \((h', h'')\) that

\[
\int_{\geq 0} \left[ \left( FP_{h'}(x_m) - FP_{h'}(x_i) \right) + \left( FP_{h''}(x_m) - FP_{h''}(x_i) \right) \right] f_{h'}(h', t_{h'}) dt_{h'} \geq 0.
\]

Now, as in both cases the inequalities are strict for all histories \( h \) such that \( T_{x_i} \) contains at least two leaves (and such histories always exist), in summary, we have

\[
\mathbb{E}[FP(x_m) - FP(x_i)|S] = \sum_{h \in \mathcal{H}(S)} \int_{\geq 0} [FP_{h}(x_m) - FP_{h}(x_i)] f_{h}(h, t_{h}) dt_{h} > 0
\]
Fig. 5 Histories in $\mathcal{H}(S)$ of type a and b as described in the proof of Proposition 1

as claimed. This completes the proof.

We now show that the ranking induced by the FP index on a species tree $S$ coincides with the ranking induced by the expected FP index for $S$ if $S$ is a caterpillar or pseudocaterpillar tree.

**Theorem 1** Let $n \geq 2$ and let $S = (S, \tau)$ be a caterpillar species tree on $X = \{x_1, \ldots, x_n\}$. Then, if $FP_S(x_j) > FP_S(x_i)$ for $x_i, x_j \in X$, we also have $E[FP(x_j)|S] > E[FP(x_i)|S]$. Moreover, if $FP_S(x_j) = FP_S(x_i)$, then $E[FP(x_j)|S] = E[FP(x_i)|S]$. In particular, the rankings $r_{(S, FP)}(X)$ and $r_{(S, E(FP))}(X)$ coincide. The same statements hold if $n \geq 4$ and $S = (S, \tau)$ is a pseudocaterpillar species tree.

**Proof** First, assume that $S$ is a caterpillar species tree and without loss of generality let $S$ be as depicted in Fig. 3. Then, we clearly have $FP_S(x_n) > FP_S(x_{n-1}) > \ldots > FP_S(x_3) > FP_S(x_2) = FP_S(x_1)$. Let $x_i, x_j \in X$ with $FP_S(x_j) > FP_S(x_i)$. Then it follows immediately from Proposition 1 (with $x_j = x_m$ and $x_i \in \{x_1, \ldots, x_{m-1}\}$) that $E[FP(x_j)|S] > E[FP(x_i)|S]$.

Now, if $FP_S(x_j) = FP_S(x_i)$, we must have $\{x_i, x_j\} = \{x_1, x_2\}$. In particular, $x_i$ and $x_j$ form the unique cherry in $S$. This immediately implies that $x_i$ and $x_j$ are exchangeable and thus $E[FP(x_j)|S] = E[FP(x_i)|S]$ as claimed.

As $FP_S(x_n) > FP_S(x_{n-1}) > \ldots > FP_S(x_3) > FP_S(x_2) = FP_S(x_1)$ and all inequality and equality signs are “preserved” when considering the expected FP index, we clearly have $r_{(S, FP)}(X) = r_{(S, E(FP))}(X)$. This completes the first part of the proof.
Now, assume that $S$ is a pseudocaterpillar species tree with $n \geq 4$ leaves as depicted in Fig. 3. In this case, we have $FP_S(x_n) > FP_S(x_{n-1}) > \ldots > FP_S(x_5) > FP_S(x_4) = FP_S(x_3) > FP_S(x_2) = FP_S(x_1)$. In particular, if $FP_S(x_i) = FP_S(x_j)$, we have $\{x_i, x_j\} = \{x_1, x_2\}$ or $\{x_i, x_j\} = \{x_3, x_4\}$. In both cases, $x_i$ and $x_j$ form a cherry and are clearly exchangeable. We thus have $\mathbb{E}[FP(x_j) | S] = \mathbb{E}[FP(x_i) | S]$ as claimed. Now, consider $x_i, x_j \in X$ with $FP_S(x_j) > FP_S(x_i)$. If $x_j \in \{x_5, \ldots, x_n\}$ (and $x_i \in \{x_1, \ldots, x_n\}$ with $i < j$) it follows from Proposition 1 that $\mathbb{E}[FP(x_j) | S] > \mathbb{E}[FP(x_i) | S]$ as claimed. It thus only remains to consider the case that $x_j \in \{x_3, x_4\}$ and $x_i \in \{x_1, x_2\}$. Without loss of generality, assume $x_j = x_3$ and $x_i = x_1$. Considering

$$\mathbb{E}[FP(x_3) - FP(x_1) | S] = \sum_{h \in \mathcal{H}(S)} \int [FP_h(x_3) - FP_h(x_1)] f_h(h, t_h) \, dt_h,$$

we now note that $\mathcal{H}(S)$ can be partitioned into five disjoint subsets:

(i) The subset of histories, $\mathcal{H}_1(S)$ say, for which both the coalescence involving $x_1$ and the coalescence involving $x_3$ occur in interval $\tau_3$ or later. In this case, $x_1$ and $x_3$ are exchangeable and we can conclude that

$$\sum_{h \in \mathcal{H}_1(S)} \int [FP_h(x_3) - FP_h(x_1)] f_h(h, t_h) \, dt_h = 0.$$

(ii) The subset of histories, $\mathcal{H}_2(S)$ say, for which both the coalescence involving $x_1$ and the coalescence involving $x_3$ occur in interval $\tau_2$. By the symmetry of the 4-leaf subtree of $S$ containing leaves $\{x_1, x_2, x_3, x_4\}$, leaves $x_1$ and $x_3$ are also exchangeable (for each history $h'$ where the cherry $\{x_1, x_2\}$ is formed prior to the cherry $\{x_3, x_4\}$, there exists a history $h''$ identical to $h'$ except that the cherry $\{x_3, x_4\}$ is formed prior to the cherry $\{x_1, x_2\}$; in particular, the roles of $x_1$ and $x_3$ are interchanged between $h'$ and $h''$) and we can conclude that

$$\sum_{h \in \mathcal{H}_2(S)} \int [FP_h(x_3) - FP_h(x_1)] f_h(h, t_h) \, dt_h = 0.$$

(iii) The subset of histories, say $\mathcal{H}_3(S)$, for which the coalescence involving $x_1$ occurs in interval $\tau_1$ and the coalescence involving $x_3$ occurs in interval $\tau_2$ or later. In this case, $x_1$ enters the interval $\tau_3$ as part of a subtree, $T_{x_1}$ say, containing precisely two leaves, whereas $x_3$ enters the interval $\tau_3$ as part of a subtree, $T_{x_3}$ say, containing at most two leaves.

As in the proof of Proposition 1, $T_{x_1}$ and $T_{x_3}$ are exchangeable as defined on page 7. Moreover, as $x_1$ coalesces strictly before $x_3$, for each history in $\mathcal{H}_3(S)$ the length of the pendant edge incident to $x_3$ is strictly larger than the length of the pendant edge incident to $x_1$. Using this and the exchangeability of $T_{x_1}$ and $T_{x_3}$, together with the fact that $T_{x_3}$ contains at most two leaves, whereas $T_{x_1}$ contains precisely two leaves, we can thus conclude that
\[ \sum_{h \in \mathcal{H}_3(S)} \int [FP_h(x_3) - FP_h(x_1)] f_h(h, t_h) \, dt_h > 0. \]

(iv) The last two remaining subsets are the subsets, \( \mathcal{H}_4(S) \) and \( \mathcal{H}_5(S) \) say, where \( \mathcal{H}_4(S) \) is the subset of histories for which the coalescence involving \( x_1 \) occurs in \( \tau_2 \) and the coalescence involving \( x_3 \) occurs in \( \tau_3 \) or later, and \( \mathcal{H}_5(S) \) is the subset of histories for which the coalescence involving \( x_1 \) occurs in \( \tau_3 \) or later and the coalescence involving \( x_3 \) occurs in \( \tau_2 \). By the symmetry of the 4-leaf subtree of \( S \) containing leaves \( \{x_1, x_2, x_3, x_4\} \), the sets \( \mathcal{H}_4(S) \) and \( \mathcal{H}_5(S) \) have the same cardinality. Now, for each history \( h' \in \mathcal{H}_4(S) \), there exists precisely one history \( h'' \in \mathcal{H}_5(S) \) identical to \( h' \), except that in \( h'' \) the most recent coalescence at time \( t_1 \) corresponds to the formation of the cherry \( \{x_1, x_2\} \), whereas in \( h'' \), the most recent coalescence at time \( t_1 \) corresponds to the formation of the cherry \( \{x_3, x_4\} \); in particular, the roles of \( x_1 \) and \( x_3 \) are interchanged between \( h' \) and \( h'' \) (similar to Case (ii) discussed above, except that not necessarily both cherries \( \{x_1, x_2\} \) and \( \{x_3, x_4\} \) are formed in \( h' \), respectively \( h'' \)). Moreover, \( t_{h'} = t_{h''} \) and \( f_{h'}(h', t_{h'}) = f_{h''}(h'', t_{h''}) \). Furthermore, \( FP_h(x_1) = FP_{h''}(x_3) \) and \( FP_{h''}(x_1) = FP_{h'}(x_3) \), and thus we can cancel the contribution of \( h' \) towards \( \mathbb{E}[FP(x_3) - FP(x_1)|S] \) with the corresponding contribution of \( h'' \).

Taking all cases together, this implies that \( \mathbb{E}[FP(x_3) - FP(x_1)|S > 0 \) and thus \( \mathbb{E}[FP(x_3)|S] > \mathbb{E}[FP(x_1)|S] \) as claimed. Now, as \( FP_S(S_n) > FP_S(S_{n-1}) > \ldots > FP_S(S_4) = FP_S(S_3) > FP_S(S_2) = FP_{S_1} \) and all inequality and equality signs are preserved when considering the expected FP index, we clearly have \( r_{(S,FP)}(X) = r_{(S,FP)}(X) \). This completes the proof.

As all species trees with at most four leaves are either caterpillar or pseudocaterpillar trees, Theorem 1 immediately leads to the following corollary:

**Corollary 1** If \( S = (S, \tau) \) is a species tree on \( X = \{x_1, \ldots, x_n\} \) with \( n \leq 4 \), the rankings \( r_{(S,FP)}(X) \) and \( r_{(S,FP)}(X) \) coincide.

### 3.2 On the existence of species trees with different rankings induced by the FP index and the expected FP index for all \( n \geq 5 \)

After having analyzed some circumstances under which the FP index calculated for a given species tree and the expected FP index across all gene tree histories associated with this species tree induce the same ranking, in this section we show that for all leaf numbers greater or equal to 5, there exist species trees for which the two rankings differ.

**Theorem 2** For all \( n \geq 5 \), there exists a species tree \( S = (S, \tau) \) on \( X = \{x_1, \ldots, x_n\} \) such that \( S \) contains two leaves, \( x_i \) and \( x_j \) say, with \( FP_S(x_i) > FP_S(x_j) \) but \( \mathbb{E}[FP(x_i)|S] < \mathbb{E}[FP(x_j)|S] \). In particular, \( S \) is such that the rankings \( r_{(S,FP)}(X) \) and \( r_{(S,FP)}(X) \) are different.

The strategy to prove this theorem is to show that there exists a species tree \( S = (S, \tau) \) on precisely 5 leaves for which the rankings \( r_{(S,FP)}(X) \) and \( r_{(S,FP)}(X) \) differ. We
then construct a species tree $S'$ on $n > 5$ leaves containing $S$ as a subtree and show that the rank swap in $S$ is preserved in $S'$. For this purpose, we will decompose all histories evolving within $S$, respectively $S'$, by “cutting” them at a certain height determined by the species tree. We now explain this procedure in more detail.

Let $S = (S, \tau)$ be a species tree of height $\eta$ and let $h \in \mathcal{H}(S)$ be a history evolving within $S$. Let $0 < \eta' \leq \eta$. By cutting history $h$ at height $\eta'$, we mean subdividing all edges of $h$ present at time $\eta'$ into two new edges, thereby disconnecting $h$. In case there is an interior vertex $v$ of $h$ present at time $\eta'$, we also subdivide it into two copies, one of which forming the root of a subtree of $h$ with the outgoing edges, and the other being a new leaf for the incoming edge. This procedure will subdivide $h$ into a forest of several trees, one of which, say $h^{top}$, will contain the root of $h$. We call $h^{top}$ the top part of $h$ concerning cut height $\eta'$ and we refer to the other trees, say $h^1, \ldots, h^k$, in the forest as the chopped parts of $h$ concerning cut height $\eta'$. Moreover, if we attach to each leaf of $h^{top}$ the number of descending leaves it had in $h$ by edges of length 0 (note that the resulting tree need not be binary and may contain degree-2 vertices), we derive the extended top part of $h$ concerning cut height $\eta'$, which we denote by $h^{ext}$. Whenever there is no ambiguity concerning the cut height, we drop the explicit mention of $\eta'$. As an example, consider the species tree $S$ and the history $h$ evolving within $S$ depicted in Fig. 1 and let $\eta' = \tau_0 + \tau_1 + \tau_2$. Then cutting $h$ at height $\eta'$ leads to the decomposition of $h$ into its (extended) top and chopped parts as depicted in Fig. 6.

We now show that given a cut height $0 < \eta' \leq \eta$, the FP index of a taxon $x_i$ on history $h$ can be calculated by considering the decomposition of $h$ into its extended top part and chopped parts concerning cut height $\eta'$.

**Lemma 1** Let $S = (S, \tau)$ be a species tree on taxon set $X = \{x_1, \ldots, x_n\}$ and let $\eta = \tau_0 + \ldots + \tau_{n-2}$ denote its height. Let $h \in \mathcal{H}(S)$ be a history evolving within $S$. Moreover, let $0 < \eta' \leq \eta$, and let $h^{ext}$ be the extended top part of $h$ concerning cut height $\eta'$ and let $h^1, \ldots, h^k$ be the chopped parts of $h$. Then, for each $x_i$ in $X$, we have

$$FP_h(x_i) = FP_{h^i}(x_i) + FP_{h^{ext}}(x_i),$$

where $h^i$ is the chopped part of $h$ containing $x_i$ (note that the root of $h^i$ might have out-degree 1) and $v_{x_i}$ is one of the newly attached leaves of $h^{ext}$ whose parent is an ancestor of $x_i$ in $h$.

**Proof** Considering the path $P(h; \rho, x_i)$ from the root $\rho$ in $h$ to leaf $i$ with respect to cut height $\eta'$, we can decompose this path into two subpaths, $P_i$ and $P^{top}$ say, where $P_i$ is subpath belonging to $h^i$, except possibly for a subdivided edge $e_{cut}$, and $P^{top}$ is the subpath belonging to the top part $h^{top}$ of history $h$, again except for edge $e_{cut}$ (if it exists). In case such a subdivided edge $e_{cut}$ exists, let us denote by $e_i$ the part of $e_{cut}$ that belongs to the chopped part $h^i$ of history $h$ and by $e_{top}$ the part of $e_{cut}$ which belongs to the top part $h^{top}$ of history $h$. Setting $l(e_{cut}) = 0$ (implying $l(e_i) = l(e_{top}) = 0$) in case no edge in $P(h; \rho, x_i)$ was subdivided and noting that $n(e_{cut}) = n(e_i) = n(e_{top})$, we get
Fig. 6 Decomposition of gene tree history \( h \) evolving within species tree \( S \) as depicted in Fig. 1 when cutting at height \( \eta' = \tau_0 + \tau_1 + \tau_2 \) into its a top part, b extended top part, and c chopped parts

\[
FP_h(x_i) = \sum_{e \in P(h; \rho, x_i)} \frac{l(e)}{n(e)} = \sum_{e \in P} \frac{l(e)}{n(e)} + \sum_{e \in P_{top}} \frac{l(e)}{n(e)} + \frac{l(e_{top})}{n(e_{top})}
\]

\[
= \left( \sum_{e \in P} \frac{l(e)}{n(e)} + \frac{l(e_i)}{n(e_i)} \right) + \left( \sum_{e \in P_{top}} \frac{l(e)}{n(e)} + \frac{l(e_{top})}{n(e_{top})} \right)
\]

\[
= FP_{h^1}(x_i) + FP_{h_{ext}(v_{x_i})}
\]

This completes the proof.

As an example consider history \( h \) depicted in Fig. 1 and its decomposition concerning cut height \( \eta' = \tau_0 + \tau_1 + \tau_2 \) depicted in Fig. 6. Here, we have \( FP_h(x_1) = \tau_0 + \tau_1 + \frac{\tau_2 + \tau_3 + \tau_1 + \tau_4}{2} \), \( FP_{h^1}(x_1) = \tau_0 + \tau_1 + \frac{\tau_2 + \tau_1}{2} \), and \( FP_{h_{ext}(v_{x_1})} = \frac{\tau_3 + \tau_4}{2} \). In particular, \( FP_h(x_1) = FP_{h^1}(x_1) + FP_{h_{ext}(v_{x_1})} \) as stated by Lemma 1.

Now, considering the expected FP index across all gene tree histories associated with a species tree \( S \) and given a cut height \( 0 < \eta' \leq \eta \), as a consequence of Lemma 1 and
using linearity of the expectation, we obtain the following decomposition of Eq. (3):

\[ \mathbb{E}[FP(x_i) | S] = \sum_{h \in \mathcal{H}(S)} \mathbb{E}[FP_h(x_i) | S] \]

\[ = \sum_{h \in \mathcal{H}(S)} \mathbb{E}[FP_h^\prime(x_i) + FP_{h_{\text{top}}}^{\prime}(v_{x_i}) | S] \]

\[ = \sum_{h \in \mathcal{H}(S)} \left( \mathbb{E}[FP_h^\prime(x_i) | S] + \mathbb{E}[FP_{h_{\text{top}}}^{\prime}(v_{x_i}) | S] \right) \]

\[ = \sum_{h \in \mathcal{H}(S)} \mathbb{E}[FP_h^\prime(x_i) | S] + \sum_{h \in \mathcal{H}(S)} \mathbb{E}[FP_{h_{\text{top}}}^{\prime}(v_{x_i}) | S]. \quad (6) \]

In the following, we will refer to the first summand in Eq. (6) as the ‘expected chopped FP index’ of \( x_i \), denoted as \( \mathbb{E}[FP_{\text{chop}}^\eta(x_i) | S] \), and to the second summand as the ‘expected top FP index’ of \( x_i \), denoted as \( \mathbb{E}[FP_{\text{top}}^\eta(x_i) | S] \). With this we are now finally in the position to prove Theorem 2.

**Proof of Theorem 2** We first show that the statement holds for \( n = 5 \). Let \( S = (S, \tau) \) be as depicted in Fig. 1 and set \( \tau = (\tau_0, \tau_1, \tau_2, \tau_3) = (2.114, 0.1269, 0.00293, 0.37684) \). Then, it can easily be seen that \( FP_0(x_3) \approx 2.36944 > 2.367335 = FP_0(x_1) \). Moreover, we used the computer algebra system Mathematica (Wolfram Research I 2020) to explicitly enumerate all gene tree histories in \( \mathcal{H}(S) \) and calculate the expected FP indices of all taxa according to Eq. (3) (the full expressions in terms of \( (\tau_0, \tau_1, \tau_2, \tau_3, \tau_4) \) are given in Appendix A). In particular, \( \mathbb{E}[FP(x_3) | S] \approx 3.23846178 < 3.2940675 \approx \mathbb{E}[FP(x_1) | S] \). This completes the proof for \( n = 5 \).

In the following, we will construct a species tree \( S' \) on \( n > 5 \) leaves containing \( S \) as a pendant subtree such that the rank swap of taxa \( x_1 \) and \( x_3 \) is preserved in \( S' \). However, before we describe this construction in more detail, we remark that if we cut all histories in \( \mathcal{H}(S) \) at the height of \( S \), i.e., at height \( \eta = \tau_0 + \tau_1 + \tau_2 + \tau_3 \), using Mathematica (Wolfram Research I 2020) we obtained the following expected chopped FP indices for taxa \( x_1 \) and \( x_3 \) and the choice of \( \tau \) given above:

\[ \mathbb{E}[FP_{\text{chop}}^\eta(x_1) | S] \approx 2.5661 \quad \text{and} \quad \mathbb{E}[FP_{\text{chop}}^\eta(x_3) | S] \approx 2.5641. \]

In particular, the order of \( x_1 \) and \( x_3 \) induced by the expected chopped FP index coincides with the order induced by the overall expected FP index and contradicts the order induced by the FP index on the species tree. Moreover, considering the extended top parts of all histories \( h \in \mathcal{H}(S) \) with respect to cut height \( \eta \), using Mathematica (Wolfram Research I 2020) we calculated the distribution of the number of new leaves that are attached to the last ancestor of \( x_1 \), respectively \( x_3 \), in the top part \( h_{\text{top}} \) of each history \( h \) (again for the choice of \( \tau \) given above). Let \( n_1 \) be the number of leaves attached to the last ancestor of \( x_1 \) and let \( n_3 \) denote the number of leaves attached to the last ancestor of \( x_3 \). Then, \( n_1 \in \{1, 2\} \) (if a history \( h \) is such that \( x_1 \) and \( x_2 \) coalesced
before time $\eta$, $n_1$ will be two; otherwise $n_1$ will be one) and $n_3 \in \{1, 2, 3\}$ (where $n_3 = 3$ if $h$ is such that all of $x_3, x_4, x_5$ coalesced before time $\eta$, $n_3 = 2$ if $x_3$ coalesced with $x_4$ or $x_5$ but all three taxa did not coalesce, and $n_3 = 1$ if $x_3$ did not coalesce) and we have the following distribution (Table 1):

| $m$ | 1   | 2   | 3   |
|-----|-----|-----|-----|
| $P(n_1 = m)$ | 0.602499 | 0.397501 | 0   |
| $P(n_3 = m)$ | 0.504977 | 0.362098 | 0.132925 |

Table 1   Distribution of $n_1$ and $n_3$ as described in the proof of Theorem 2

Keeping these findings in mind, we now construct a species tree $S' = (S', \tau')$ on $n > 5$ leaves as follows. We take any tree topology $S'$ on $n - 4$ leaves and replace one of the leaves by $S$. For $\tau'$, we keep the interval lengths as induced by $\tau$ and choose all remaining interval lengths such that $S' = (S', \tau')$ is an ultrametric species tree. We now argue that $S'$ has the properties that $FP_{S'}(x_3) > FP_{S'}(x_1)$, but $E[FP(x_3)|S'] < E[FP(x_1)|S']$.

First, as $S$ replaced a leaf of $S'$, all edges on the path from the root of $S'$ to the root of $S$ contribute the same amount, say $c$, to the FP index of any leaf in $\{x_1, \ldots, x_5\}$. Moreover, no edge lengths within the subtree $S$ of $S'$ were changed. Hence, we have $FP_{S'}(x_1) = FP_{S}(x_i) + c$ for all $x_i \in \{x_1, \ldots, x_5\}$ and thus in particular $FP_{S'}(x_3) > FP_{S'}(x_1)$.

In order to show the second assertion, we now consider the decomposition of the expected FP index into the expected chopped FP index and expected top FP index as given in Eq. (6) with respect to $S'$ and with respect to cut height $\eta = \tau_0 + \tau_1 + \tau_2 + \tau_3$ (i.e., with respect to the same cut height as discussed for $S$). If we cut all histories $h \in H(S')$ at height $\eta$, the expected chopped FP indices for taxa $x_1, \ldots, x_5$ on $S'$ are identical to the ones observed when cutting all histories evolving within $S$ at height $\eta$. In particular, we have

$$E[FP_{\text{chop}}(x_1)|S'] \approx 2.5661 > 2.5641 \approx E[FP_{\text{chop}}(x_3)|S'].$$

It remains to consider $E[FP_{\text{top}}(x_1)|S']$ and $E[FP_{\text{top}}(x_3)|S']$. Instead of calculating these expected values explicitly, we will consider $E[FP_{\text{top}}(x_1)|S'] - E[FP_{\text{top}}(x_3)|S'] = E[FP_{\text{top}}(x_1) - FP_{\text{top}}(x_3)|S']$ and show that this expectation is strictly positive. Together with the fact that $E[FP_{\text{chop}}(x_1)|S'] > E[FP_{\text{chop}}(x_3)|S']$ this will complete the proof.

For each history $h \in H(S')$ with top part $h^{top}$, let $l_1$ denote the leaf of $h^{top}$ corresponding to the last ancestor of $x_1$ and let $l_3$ denote the leaf of $h^{top}$ corresponding to the last ancestor of $x_3$. Then, $l_1$ and $l_3$ are exchangeable in the sense that they either (a) form a cherry of $h^{top}$; or (b) if a top part history $\tilde{h}^{top}$ does not contain the cherry $[l_1, l_3]$, there exists a top part history $\tilde{h}^{top}$ identical to $\tilde{h}^{top}$ except that the roles of $l_1$ and $l_3$ are interchanged and such that for the corresponding complete histories $\tilde{h}$ and $\tilde{\eta}$, we have $t_{\tilde{\eta}}(\tilde{h}, t_{\tilde{\eta}}) = f_{\tilde{h}}(\tilde{h}, t_{\tilde{\eta}})$. Now, recall the extended top part $h_{\text{ext}}^{top}$ of a history $h$ is obtained by attaching to each leaf of $h^{top}$
the number of descending leaves it had in \( h \) by edges of length 0. Moreover, recall that 
\[
\mathbb{E}[FP_\text{top}^h(x_i)|S'] = \sum_{h \in \mathcal{H}(S')} \mathbb{E}[FP_\text{top}^h(v_{x_i})|S'],
\]
where \( v_{x_i} \) is one of the newly attached leaves of \( h_\text{ext} \) whose parent is an ancestor of \( x_i \) in \( h \). We thus need to analyze 
\[
\sum_{h \in \mathcal{H}(S')} \mathbb{E}[FP_\text{top}^h(v_{x_1}) - FP_\text{top}^h(v_{x_3})|S'].
\]
For this purpose, we partition the set \( \mathcal{H}(S') \) of histories associated with \( S' \) into two disjoint subsets: (a) the set of histories whose induced top parts contain the cherry \( [l_1, l_3] \), and (b) the set of histories whose induced top parts do not contain the cherry \( [l_1, l_3] \). In either case, this will lead to six subcases for the extended top part histories depending on the number of leaves attached to \( l_1 \) and to \( l_3 \), respectively.

(a) We first consider the subset of histories in \( \mathcal{H}(S') \) whose induced top parts contain the cherry \( [l_1, l_3] \). Consider such a top part and denote the time of coalescence of \( l_1 \) and \( l_3 \) by \( t_{l_3} \). Now, note that each such fixed top part gives rise to a distribution of extended top parts obtained by attaching \( E \) to the cherry in \( h_\text{ext} \). Moreover, note that \( n_1 \) and \( n_3 \) are clearly independent. Referring to Fig. 7, among the histories in \( \mathcal{H}(S') \) whose induced top parts contain the cherry \( [l_1, l_3] \) and the time of coalescence of \( l_1 \) and \( l_3 \) is \( t_{l_3} \), a proportion of \( P(n_1 = 1) \cdot P(n_3 = 1) \) of extended top parts is of type (i), a proportion of \( P(n_1 = 1) \cdot P(n_3 = 2) \) is of type (ii) and so forth. Now, considering these types of extended top parts more in-depth, we observe the following:

- If the extended top part of a history \( h \) is of type (i) or (v), we have 
  \[
  FP_\text{top}^h(v_{x_1}) - FP_\text{top}^h(v_{x_3}) = 0.
  \]
- If the extended top part of a history \( h \) is of type (iii), we have 
  \[
  FP_\text{ext}^h(v_{x_1}) - FP_\text{ext}^h(v_{x_3}) = \frac{2n_3}{3} > 0.
  \]
  Similarly, if the extended top part of a history \( h \) is of type (vi), we have 
  \[
  FP_\text{top}^h(v_{x_1}) - FP_\text{top}^h(v_{x_3}) = \frac{t_{l_3}}{6} > 0.
  \]
- If the extended top part of a history \( h \) is of type (ii), we have 
  \[
  FP_\text{ext}^h(v_{x_1}) - FP_\text{ext}^h(v_{x_3}) = \frac{t_{l_3}}{2} > 0,
  \]
  and if it is of type (iv), we have 
  \[
  FP_\text{top}^h(v_{x_1}) - FP_\text{top}^h(v_{x_3}) = \frac{-t_{l_3}}{2} < 0.
  \]
  However, as type (ii) has probability \( P(n_1 = 1) \cdot P(n_3 = 2) \approx 0.218 \), whereas type (iv) has probability \( P(n_1 = 2) \cdot P(n_3 = 1) \approx 0.201 \), the proportion of histories with extended top parts contributing \( \frac{t_{l_3}}{2} \) is larger than the proportion of histories with extended top parts contributing \( -\frac{t_{l_3}}{2} \).

In summary, we can thus conclude that

\[
\sum_{h \in \mathcal{H}(S'): h_\text{top} \text{ contains cherry } [l_1, l_3]} \mathbb{E}[FP_\text{top}^h(v_{x_1}) - FP_\text{top}^h(v_{x_3})|S'] > 0. \tag{7}
\]

(b) Now, consider the subset of histories in \( \mathcal{H}(S') \) whose induced top parts do not contain the cherry \( [l_1, l_3] \). As explained above, for each such history \( \tilde{h} \) with top
part $\tilde{h}_{\text{top}}$, there exists a history $\tilde{h}$ with top part $\tilde{h}_{\text{top}}$ such that $\tilde{h}_{\text{top}}$ and $\hat{h}_{\text{top}}$ are identical up to a permutation of the leaves $l_1$ and $l_3$, and such that $t_{\tilde{h}} = t_{\hat{h}}$ and $f_{\tilde{h}}(\tilde{h}, t_{\tilde{h}}) = f_{\hat{h}}(h, t_{\hat{h}})$. Similar to case (a), each fixed pair of matching top parts gives rise to a distribution of matching pairs of extended top parts by attaching $n_1 \in \{1, 2\}$ leaves to $l_1$ and $n_3 \in \{1, 2, 3\}$ leaves to $l_3$, where $n_1$ and $n_3$ have the distribution given in Table 1. Using the same enumeration as in case (a), we thus have the following subcases:

(i) $n_1 = n_3 = 1$;
(ii) $n_1 = 1$ and $n_3 = 2$;
(iii) $n_1 = 1$ and $n_3 = 3$;
(iv) $n_1 = 2$ and $n_3 = 1$;
(v) $n_1 = 2$ and $n_3 = 2$;
(vi) $n_1 = 2$ and $n_3 = 3$.

Referring to this enumeration scheme and Fig. 8 and fixing the time of coalescence leading to $lca(x_1, x_3)$, among the matching pairs of histories in $\mathcal{H}(S)$ with induced top parts $\tilde{h}_{\text{top}}$ and $\hat{h}_{\text{top}}$, respectively, a proportion of $P(n_1 = 1) \cdot P(n_3 = 1)$ of pairs of extended top parts is of type (i), a proportion of $P(n_1 = 1) \cdot P(n_3 = 2)$ is of type (ii) and so forth. Similar to case (a), we now analyze these types more in-depth, and observe the following:

- For pairs of matching histories $(\tilde{h}, \hat{h})$ with extended top parts of type (i) or (v), we have
  \[
  FP_{\tilde{h}_{\text{ext}}}^{\tilde{h}_{\text{top}}}(v_{x_1}) - FP_{\hat{h}_{\text{ext}}}^{\hat{h}_{\text{top}}}(v_{x_3}) + FP_{\hat{h}_{\text{ext}}}^{\tilde{h}_{\text{top}}}(v_{x_1}) - FP_{\tilde{h}_{\text{ext}}}^{\hat{h}_{\text{top}}}(v_{x_3}) = 0.
  \]

- For pairs of matching histories $(\tilde{h}, \hat{h})$ with extended top parts of type (iii) or (vi), we have
  \[
  FP_{\tilde{h}_{\text{ext}}}^{\tilde{h}_{\text{top}}}(v_{x_1}) - FP_{\hat{h}_{\text{ext}}}^{\hat{h}_{\text{top}}}(v_{x_3}) + FP_{\hat{h}_{\text{ext}}}^{\tilde{h}_{\text{top}}}(v_{x_1}) - FP_{\tilde{h}_{\text{ext}}}^{\hat{h}_{\text{top}}}(v_{x_3}) > 0.
  \]

- For pairs of matching histories $(\tilde{h}, \hat{h})$ with extended top parts of type (ii), we have
  \[
  FP_{\tilde{h}_{\text{ext}}}^{\tilde{h}_{\text{top}}}(v_{x_1}) - FP_{\hat{h}_{\text{ext}}}^{\hat{h}_{\text{top}}}(v_{x_3}) + FP_{\hat{h}_{\text{ext}}}^{\tilde{h}_{\text{top}}}(v_{x_1}) - FP_{\tilde{h}_{\text{ext}}}^{\hat{h}_{\text{top}}}(v_{x_3}) = \delta > 0,
  \]

whereas for pairs of matching histories $(\tilde{h}, \hat{h})$ with extended top parts of type (iv), we have

\[
FP_{\tilde{h}_{\text{ext}}}^{\tilde{h}_{\text{top}}}(v_{x_1}) - FP_{\hat{h}_{\text{ext}}}^{\hat{h}_{\text{top}}}(v_{x_3}) + FP_{\hat{h}_{\text{ext}}}^{\tilde{h}_{\text{top}}}(v_{x_1}) - FP_{\tilde{h}_{\text{ext}}}^{\hat{h}_{\text{top}}}(v_{x_3}) = -\delta < 0.
\]
Fig. 7 Types of extended top parts induced by top parts of histories containing the cherry \([l_1, l_3]\) as described in the proof of Theorem 2

However, as in Case (a), pairs of type (ii) have a higher probability than pairs of type (iv), and thus the proportion of matching pairs of histories with extended top parts contributing a positive value is larger than the proportion of histories with extended top parts contributing the same negative value.

As we can partition the subset of histories in \(\mathcal{H}(S')\) whose induced top parts do not contain the cherry \([l_1, l_3]\) in pairs of matching histories \((\tilde{h}, \hat{h})\) as described above, we can thus also conclude that

\[
\sum_{h \in \mathcal{H}(S') \text{ such that } h_{top} \text{ does not contain cherry } [l_1, l_3]} \mathbb{E}[FP_{\eta_{\text{top}}}(v_{x_1}) - FP_{\eta_{\text{top}}}(v_{x_3})|S'] > 0. \tag{8}
\]

Taking Eqs. (7) and (8) together, we obtain

\[
\mathbb{E}[FP_{\eta_{\text{top}}}(x_1) - FP_{\eta_{\text{top}}}(x_3)|S'] = \sum_{h \in \mathcal{H}(S')} \mathbb{E}[FP_{\eta_{\text{top}}}(v_{x_1}) - FP_{\eta_{\text{top}}}(v_{x_3})|S'] > 0
\]

as claimed. This completes the proof. \(\square\)

Remarks.
We conclude this section with a few remarks. First, note that the proof of Theorem 2 establishes a slightly stronger result than the statement of Theorem 2. More precisely, we not only prove the existence of one species tree \(S\) on \(n \geq 5\) leaves with the desired property, but in fact establish this result for a family of such trees for \(n\) large enough,
as we did not fix the tree on \( n - 4 \) leaves used in the construction, such that any tree on \( n - 4 \) leaves can be used.

Second, note that in the proof of Theorem 2, we used a very particular choice of \( \tau \) leading to \( FP_S(x_3) > FP_S(x_1) \) and \( \mathbb{E}[FP(x_3)|S] < \mathbb{E}[FP(x_1)|S] \) for the species tree \( S \) on \( n = 5 \) leaves depicted in Fig. 1. However, there are infinitely many numerical values for \( \tau \) with this property and we depict slices of the corresponding regions of parameter (branch length) space in Fig. 9. Note, however, that these regions were solely plotted by considering \( FP_S(x_3) > FP_S(x_1) \) and \( \mathbb{E}[FP(x_3)|S] < \mathbb{E}[FP(x_1)|S] \) for the species tree \( S \) on \( n = 5 \) leaves depicted in Fig. 1.²

Third, we note that there exist two additional species trees on 5 leaves, \( S' \) and \( S'' \) say, with the same topology as the species tree \( S \) used in our construction, but with a different order of speciation events (see Fig. 10). Expressions for the FP indices on the species trees \( S' \) and \( S'' \) and the respective expected FP indices are given in Appendix A. It is an interesting question whether a rank swap can also occur for these two species trees. In case of the species tree \( S' \), we can answer this affirmatively. For instance, for \( \tau = (1.0000, 0.0050, 0.0075, 0.0100) \), we have \( FP_{S'}(x_3) \approx 1.01583 > 1.01375 = FP_S'(x_1) \), but \( \mathbb{E}[FP(x_3)|S'] \approx 1.8510951 < 1.8518378 \approx \mathbb{E}[FP(x_1)|S'] \). For the species tree \( S'' \), on the other hand, we conjecture that a rank swap is not possible, but leave a proof or counterexample to this conjecture to future research.

Finally, we remark that there is a curious connection between some of the results presented in this manuscript and results obtained by Degnan et al. (2012b, a) in relation to the existence of so-called anomalous ranked gene trees (ARGTs), whereby a ranked gene tree history that does not match the ranked species tree can have a greater probability than the matching ranked gene tree. More precisely, Degnan et al. (2012b) showed that ARGTs do not exist for 3-taxon or 4-taxon species trees, and Degnan et al. (2012a) showed that more generally ARGTs do not exist for caterpillar and pseudocaterpillar species trees. This directly parallels Corollary 1 and Theorem 1 in the present manuscript, which state that the ranking induced by the FP index on the

² In particular, we did not re-compute the probability distribution corresponding to Table 1 and it might be the case that not every point in the plotted region has the property that \( P(n_1 = 1) \cdot P(n_3 = 2) > P(n_1 = 2) \cdot P(n_3 = 1) \), which was used in the proof of Theorem 2 for \( n > 5 \).
species tree and the ranking induced by the expected FP index cannot differ in these cases. Moreover, the species tree $S$ depicted in Fig. 1, i.e., the species tree we used in the proof of Theorem 2 to show that there are species trees for which the rankings differ, is exactly the ranked species tree used by Degnan et al. (2012b, a) to prove the existence of ARGTs for 5-taxon trees. As indicated above, rank swaps for the FP index exist for at least one other ranking of this species tree topology (namely for the ranked species tree $S'$ shown in Fig. 10) and it would be interesting to see if this again parallels or maybe contrasts the case of ARGTs. Based on the characterization of species trees producing ARGTs given by Degnan et al. (2012a) it is unfortunately not possible to decide if the two species trees $S'$ and $S''$ induce ARGTs.

Fig. 9 Slices of parameter regions (shaded) for the species tree $S$ depicted in Fig. 1 such that $FP_S(x_3) > FP_S(x_1)$, whereas $E[FP(x_3)|S] < E[FP(x_1)|S]$. In each subfigure, a different parameter is fixed and four different choices are shown, respectively.
Fig. 10 Species trees $S'$ and $S''$ that have the same unranked topology as the species tree $S$ depicted in Fig. 1, but a different ranked topology. In particular, they depict a different order of speciation events (compare the circled numbers to see this)

4 Empirical data

In order to illustrate the variability of prioritization orders obtained from the FP index for empirical data, we considered the multilocus mammal data set of Song et al. (2012). As the original collection of gene trees for the data was criticized in the literature for methodological errors, we used those from a re-analysis of Springer and Gatesy (2016). These data consist of 447 gene trees for 37 species (33 of which are mammals and four of which are outgroup species; a list of species names and their abbreviations can be found in Table 2 Appendix B). As the gene trees provided by Springer and Gatesy (2016) were unrooted, we rooted them using *Gallus gallus* as the outgroup. We then calculated the FP index for all 37 species on all 447 gene trees. In addition, we calculated the average FP index of each species across the 447 gene trees. We also calculated the FP index for all species using two methods: the species tree estimation method SVDQuartets (Chifman and Kubatko 2014) as implemented in the PAUP* package (Swofford 2021) and the maximum likelihood method for the concatenated

3 Note that we use this data set as a mere example for the variability of the FP rankings across gene trees and do not attempt to make any statement about the conservation implications for any of those species.
data using RAxML version 8.2.10 (Stamatakis 2014). Maximum likelihood branch lengths were computed for both trees under the GTR+Γ model in PAUP*, additionally enforcing a molecular clock. For reproducibility, we provide the exact commands used in Appendix C. Finally, the FP indices on the 447 gene trees and the two species trees as well as the average FP indices were transformed into rankings and a boxplot summarizing the results was generated (Fig. 11).

As evident from the boxplot, there is a huge variability of ranking positions across the 447 gene trees for most species. In particular, there are species such as *Sus scrofa* (Sus), whose ranking positions span the entire range of ranking positions possible or are close to it (as in the case of *Callithrix jacchus* (Cal), *Equus caballus* (Equ), or *Macaca mulatta* (New)). On the other hand, there are some species such as *Gorilla gorilla* (Gor), *Homo sapiens* (Hom), *Pan troglodytes* (Pan), and *Pongo pygmaeus* (Pon) that show comparatively little variability in their ranking positions and consistently receive low ranks. These four species also show the interesting property that the ranks obtained from the average FP index across the 447 gene trees and the ranks obtained from the two species tree estimates all coincide except for one case (for Pon, the rank from the mean over gene trees differs slightly from the rank on the two species tree estimates). In general, it is interesting to note that the average FP index and the FP index on the two species tree estimates yield similar ranks in many cases. In particular, the rank of the SVDQuartets tree almost always coincides with either the rank based on the average FP index or the rank based on the concatenated data. However, there are nevertheless some interesting rank swaps. For instance, considering species *Procavia capensis* (Pro) and *Tupaia belangeri* (Tup), based on the average FP index, *Tupaia belangeri* receives are higher rank than *Procavia capensis*, but on both species tree estimates, *Procavia capensis* receives a higher rank than *Tupaia belangeri*. This last observation highlights that our theoretical results on possible rank swaps between the FP index on a species tree and the expected FP index across all gene tree histories associated with it (Theorem 2) might occur in practice and need to be considered in the process of conservation planning.

5 Concluding remarks

The FP index is a popular phylogenetic diversity index used to prioritize species for conservation. It assesses the importance of species for overall biodiversity based on their placement in an underlying phylogenetic tree. As such, the prioritization order obtained from the FP index heavily depends on the topology and branch lengths of the phylogenetic tree used. In particular, different phylogenetic trees for the same set of species might lead to different prioritization orders.

In this note, we analyzed the variability of prioritization orders obtained from the FP index for different phylogenetic trees on the same set of species in the context of gene tree species tree discordance. More precisely, we compared the ranking obtained from the FP index on a given species tree with the ranking obtained from the expected FP index across all gene tree histories evolving within the species tree under the multispecies coalescent model. On the one hand, we showed that the two prioritization orders are identical for all trees with at most 4 leaves, and more generally for all
caterpillar and pseudocaterpillar trees. On the other hand, we showed that for all \( n \geq 5 \), there exists a species tree on \( n \) leaves such that the two rankings are different. We ended by illustrating the variability in ranking orders obtained from the FP index for an empirical multilocus mammal data set. Here, we observed a high variability in ranking positions for most species, indicating that if conservation decisions are based on single genes, the prioritization order may be very different depending on which gene is used. We also observed rank swaps between the average FP index across gene trees and the FP index on two species tree estimates, empirically underpinning our theoretical results stated in Theorem 2.

Our work suggests several broad questions that might be interesting to explore in the future. From a theoretical point of view, it would be interesting to analyze if there are classes of tree topologies other than the caterpillar and pseudocaterpillar trees for which the FP index on the species tree and the expected FP index across all gene tree histories necessarily induce the same prioritization order. Moreover, in cases for which there are differences between the two prioritization orders, it would be interesting to assess how different the two rankings can be in the most extreme cases. In particular, we remark that our construction for the existence of a species tree on \( n \geq 5 \) leaves leading to two different ranking orders relies on a switch in ranking positions among the leaves of a certain 5-taxon subtree, and if \( n \) is large and this is the only difference between the two rankings, this might be considered a minor difference. It would thus be interesting to see if there are other constructions that lead to more severe differences between the two rankings. From an empirical point of view, it would also be interesting to study the variability of the prioritization orders obtained from the FP index for different gene trees and species tree estimates for a broader class of data sets, in particular for data sets containing conservation-relevant species, i.e., species at high risk of extinction. More generally, as both our theoretical results as well the analysis of the empirical data set show that different prioritization orders may be obtained from considering the FP index on a given species tree, on a particular gene tree, or even on all possible gene trees associated with the species tree, an immediate question that arises is the question of which evolutionary information should be used as the basis for conservation decisions. Should these decisions be based on the species tree, on a collection of gene trees, or taking into account a combination of both? We leave all of these questions to future research, but remark that particularly answering the last question will likely require joint efforts from theoreticians and empiricists.

Acknowledgements MF was supported by the joint research project DIG-IT! funded by the European Social Fund (ESF), reference: ESF/14-BM-A55-0017/19, and the Ministry of Education, Science and Culture of Mecklenburg-Vorpommern, Germany. KW was supported by The Ohio State University’s President’s Postdoctoral Scholars Program. All authors thank three anonymous reviewers for detailed comments on an
earlier version of this manuscript. Moreover, all authors thank James Degnan for a helpful discussion on the existence of anomalous ranked gene trees.

**Funding** MF was supported by the joint research project DIG-IT! funded by the European Social Fund (ESF), reference: ESF/14-BM-A55- 0017/19, and the Ministry of Education, Science and Culture of Mecklenburg-Vorpommern, Germany. KW was supported by The Ohio State University’s President’s Post-doctoral Scholars Program.

**Data Availability** Not applicable.

**Declarations**

**Conflicts of interest** The authors declare that there is no conflict of interest.

**Code availability** Not applicable.

**Appendix A FP indices for the species trees depicted in Figs. 1 and 10**

In the following, we give expression for the FP indices on the species tree and the expected FP indices for the species trees \(S\), \(S'\), and \(S''\) on 5 leaves considered in the main part of this manuscript (Figs. 1 and 10). The expected FP indices were obtained using Mathematica (Wolfram Research I 2020) to enumerate all gene tree histories in \(H(S)\), respectively \(H(S')\) and \(H(S'')\).

**Species tree \(S\) (Fig. 1).**

- FP indices on the species tree:

\[
FP_S(x_1) = \tau_0 + \frac{\tau_1}{2} + \frac{\tau_2}{2} + \frac{\tau_3}{2} = FP_S(x_2)
\]
\[
FP_S(x_3) = \tau_0 + \tau_1 + \tau_2 + \frac{\tau_3}{3}
\]
\[
FP_S(x_4) = \tau_0 + \tau_1 + \frac{\tau_2}{2} + \frac{\tau_3}{3} = FP_S(x_5).
\]

- Expected FP indices:

\[
\mathbb{E}[FP(x_1)|S] = \tau_0 + \frac{\tau_1}{2} + \frac{\tau_2}{2} + \frac{\tau_3}{3} + 1 - 12 - \frac{1}{24} - \frac{1}{24} - \frac{1}{24} = \mathbb{E}[FP(x_2)|S]
\]
\[
\mathbb{E}[FP(x_3)|S] = \tau_0 + \tau_1 + \tau_2 + \frac{\tau_3}{3} + 1 - \frac{1}{9} - \frac{1}{12} - \frac{1}{12} = \mathbb{E}[FP(x_4)|S]
\]
\[
\mathbb{E}[FP(x_4)|S] = \tau_0 + \tau_1 + \frac{\tau_2}{2} + \frac{\tau_3}{3} + 1 - \frac{1}{9} - \frac{1}{24} - \frac{1}{24} - \frac{1}{24} = \mathbb{E}[FP(x_5)|S].
\]
Species tree $S'$ (Fig. 10, top)

- FP indices on the species tree:

$$FP_{S'}(x_1) = \tau_0 + \tau_1 + \frac{\tau_2}{2} + \frac{\tau_3}{2} = FP_{S'}(x_2)$$
$$FP_{S'}(x_3) = \tau_0 + \tau_1 + \tau_2 + \frac{\tau_3}{3}$$
$$FP_{S'}(x_4) = \tau_0 + \frac{\tau_1}{2} + \frac{\tau_2}{2} + \frac{\tau_3}{3} = FP_{S'}(x_5).$$

- Expected FP indices:

$$\mathbb{E}[FP(x_1)|S'] = \tau_0 + \tau_1 + \frac{\tau_2}{2} + \frac{\tau_3}{2} + 1 - \frac{e^{-\tau_1 - 2\tau_2 - 2\tau_3}}{24} - \frac{e^{-\tau_2 - 2\tau_3}}{12} - \frac{e^{-\tau_2 - \tau_3}}{72} + \frac{e^{-\tau_3}}{36} = \mathbb{E}[FP(x_2)|S']$$
$$\mathbb{E}[FP(x_3)|S'] = \tau_0 + \tau_1 + \tau_2 + \frac{\tau_3}{3} + 1 - \frac{e^{-\tau_1 - \tau_2}}{9} - \frac{e^{-\tau_2 - 2\tau_3}}{12} + \frac{e^{-\tau_2 - \tau_3}}{18}$$
$$\mathbb{E}[FP(x_4)|S'] = \tau_0 + \frac{\tau_1}{2} + \frac{\tau_2}{2} + \frac{\tau_3}{3} + 1 - \frac{e^{-\tau_1 - \tau_2}}{9} - \frac{e^{-\tau_2 - 2\tau_3}}{24} + \frac{e^{-\tau_2 - \tau_3}}{18} + \frac{e^{-\tau_3}}{36} = \mathbb{E}[FP(x_5)|S'] .$$

Species tree $S''$ (Fig. 10, bottom)

- FP indices on the species tree:

$$FP_{S''}(x_1) = \tau_0 + \tau_1 + \tau_2 + \frac{\tau_3}{2} = FP_{S''}(x_2)$$
$$FP_{S''}(x_3) = \tau_0 + \tau_1 + \frac{\tau_2}{3} + \frac{\tau_3}{3}$$
$$FP_{S''}(x_4) = \tau_0 + \frac{\tau_1}{2} + \frac{\tau_2}{3} + \frac{\tau_3}{3} = FP_{S''}(x_5).$$

- Expected FP indices:

$$\mathbb{E}[FP(x_1)|S''] = \tau_0 + \tau_1 + \tau_2 + \frac{\tau_3}{2} + 1 - \frac{e^{-\tau_2 - 2\tau_3}}{12} - \frac{e^{-\tau_1 - 2\tau_3}}{24} + \frac{e^{-\tau_2 - \tau_3}}{36} + \frac{e^{-\tau_3}}{72} = \mathbb{E}[FP(x_2)|S'']$$
$$\mathbb{E}[FP(x_3)|S''] = \tau_0 + \tau_1 + \frac{\tau_2}{3} + \frac{\tau_3}{3} + 1 - \frac{e^{-\tau_1}}{9} - \frac{e^{-\tau_2 - 2\tau_3}}{12} - \frac{e^{-\tau_2 - \tau_3}}{18} + \frac{e^{-\tau_3}}{12}$$

Springer
\[
\mathbb{E}\left[ FP(x_4)|S''\right] = \tau_0 + \frac{\tau_1}{2} + \frac{\tau_2}{3} + \frac{\tau_3}{3} + 1 - \frac{e^{-\tau_1}}{9} - \frac{e^{-\tau_2 - 2\tau_3}}{24}
\]

\[
- \frac{e^{-\tau_1 - \tau_2 - 2\tau_3}}{24} + \frac{e^{-\tau_2 - \tau_3}}{36} - \frac{e^{-\tau_1 - \tau_2 - \tau_3}}{18} + \frac{e^{-\tau_2}}{18}
\]

\[
= \mathbb{E}\left[ FP(x_5)|S''\right].
\]

Appendix B Species name abbreviations used in Fig. 11

| Abbreviation | Scientific name             | Common name                      |
|--------------|-----------------------------|----------------------------------|
| Bos          | Bos taurus                  | Cattle                           |
| Cal          | Callithrix jacchus          | White-tufted-ear marmoset        |
| Can          | Canis familiaris            | Dog                              |
| Cav          | Cavia porcellus             | Domestic guinea pig              |
| Cho          | Choloepus hoffmann          | Hoffmann’s two-fingered sloth    |
| Das          | Dasypus novemcinctus        | Nine-banded armadillo            |
| Dip          | Dipodomys ordii             | Ord’s kangaroo rat               |
| Ech          | Echinops telfairi           | Lesser hedgehog tenrec           |
| Equ          | Equus caballus              | Domestic horse                    |
| Eri          | Erinaceus europaeus         | European hedgehog                |
| Fel          | Felis catus                 | Cat                              |
| Gal          | Gallus gallus              | Chicken                          |
| Gor          | Gorilla gorilla            | Gorilla                          |
| Hom          | Homo sapiens                | Human                            |
| Lox          | Loxodonta africana          | African bush elephant            |
| Mac          | Macropus eugenii           | Tammar wallaby                   |
| Mic          | Microcebus murinus          | Gray mouse lemur                  |
| Mon          | Monodelphis domestica      | Gray short-tailed opossum         |
| Mus          | Mus musculus               | House mouse                      |
| Myo          | Myotis lucifugus            | Little brown bat                 |
| New          | Macaca mulatta             | Rhesus macaque                   |
| Och          | Ochotona princeps           | American pika                    |
| Orn          | Ornithorhynchus anatinus    | Platypus                         |
| Ory          | Oryctolagus cuniculus      | European rabbit                  |
| Oto          | Otolemur garnettii         | Northern greater galago          |
| Pan          | Pan troglodytes             | Cimpanzee                        |
| Pon          | Pongo pygmaeus              | Bornean orangutan                |
| Pro          | Procavia capensis           | Rock hyrax                       |
| Pte          | Pteropus vampyrus           | Large flying fox                 |
| Rat          | Rattus                      | Rat                              |
Table 2 (continued)

| Abbreviation | Scientific name | Common name       |
|--------------|----------------|-------------------|
| Sor          | *Sorex araneus* | Common shrew      |
| Spe          | *Spermophilus tridecemlineatus* | Thirteen-lined ground squirrel |
| Sus          | *Sus scrofa*    | Wild boar         |
| Tar          | *Tarsius syrichta* | Philippine tarsier |
| Tup          | *Tupaia belangeri* | Northern treeshrew |
| Tur          | *Tursiops truncatus* | Bottlenose dolphin |
| Vic          | *Vicugna pacos* | Alpaca            |

Appendix C Details on species tree estimation for Sect. 4

In order to obtain species tree estimates for the multilocus mammal data set of Song et al. (2012), we downloaded the full sequences of the 447 genes from [https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.3629v](https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.3629v) (file gene447_final.nexus).

After converting to PHYLIP format, a maximum likelihood tree based on the concatenated data was obtained using RAxML version 8.2.10 (Stamatakis 2014) via the following command:

```
raxMLHPC-AVX -s gene447_final.phy -p 12345 -m GTRGAMMA -o Gal -n raxml_GTR.tre
```

Afterwards, maximum likelihood branch lengths under the GTR+Γ model and enforcing a molecular clock were computed in PAUP* (Swofford 2021) via the following commands:

```
exe gene447_final.nexus
gettrees file=raxml_GTR.tre
set criterion=likelihood
lset nst=6 basefreq=empirical rates=gamma rmatrix=estimate shape=estimate clock=yes
lscores
savetrees RAxML_GTR_MLbranchlengths.tre brlens
```

Moreover, a species tree using the species tree estimation method SVDQuartets (Chifman and Kubatko 2014) and maximum likelihood branch lengths for this species tree were obtained in PAUP* via the following commands:

```
exe gene447_final.nex
svdq
outgroup 35 (where species 35 corresponds to Gal)
roottrees rootmethod=outgroup
set criterion=likelihood
lset nst=6 basefreq=empirical rates=gamma rmatrix=estimate shape=estimate clock=yes
lscores
savetrees svdq_GTR_MLbranchlengths.tre brlens
```
References

Chifman J, Kubatko L (2014) Quartet inference from SNP data under the coalescent model. Bioinformatics 30(23):3317–3324. https://doi.org/10.1093/bioinformatics/btu530

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24(6):332–340. https://doi.org/10.1016/j.tree.2009.01.009

Degnan JH, Rosenberg NA, Stadler T (2012) A characterization of the set of species trees that produce anomalous ranked gene trees. IEEE/ACM Trans Comput Biol Bioinf 9(6):1558–1568. https://doi.org/10.1109/tcbb.2012.110

Degnan JH, Rosenberg NA, Stadler T (2012) The probability distribution of ranked gene trees on a species tree. Math Biosci 235(1):45–55. https://doi.org/10.1016/j.mbs.2011.10.006

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Cons 61(1):1–10. https://doi.org/10.1016/0006-3207(92)91201-3

Fuchs M, Jin EY (2015) Equality of shapley value and fair proportion index in phylogenetic trees. J Math Biol 71(5):1133–1147. https://doi.org/10.1007/s00285-014-0853-0

Haake CJ, Kashiwada A, Su FE (2008) The Shapley value of phylogenetic trees. J Math Biol 56(4):479–497. https://doi.org/10.1007/s00285-007-0126-2

Isaac NJ, Turvey ST, Collen B et al (2007) Mammals on the EDGE: conservation priorities based on threat and phylogeny. PLoS ONE 2(3):e296

Maddison WP (1997) Gene trees in species trees. Syst Biol 46(3):523–536. https://doi.org/10.1093/sysbio/46.3.523

Nichols R (2001) Gene trees and species trees are not the same. Trends Ecol Evol 16(7):358–364. https://doi.org/10.1010/10.1016/S0169-5347(01)02203-0

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5(5):568–583. https://doi.org/10.1093/100006-32079219201-3

Redding DW (2003) Incorporating genetic distinctness and reserve occupancy into a conservation prioritization approach. University Of East Anglia, UK

Redding DW, Mooers AØ (2006) Incorporating evolutionary measures into conservation prioritization. Conserv Biol 20(6):1670–1678. https://doi.org/10.1111/j.1523-1739.2006.00555.x

Redding DW, Hartmann K, Mimoto A et al (2008) Evolutionarily distinctive species often capture more phylogenetic diversity than expected. J Theor Biol 251(4):606–615. https://doi.org/10.1016/j.jtbi.2007.12.006

Redding DW, Mazel F, Mooers AØ (2014) Measuring evolutionary isolation for conservation. PLoS ONE 9(12):e113,490. https://doi.org/10.1371/journal.pone.0113490

Shapley LS (1953) A value for n-person games. In: Contributions to the theory of games (AM-28), Volume II. Princeton University Press, Princeton, pp 307–317. https://doi.org/10.1515/9781400881970-018

Song S, Liu L, Edwards SV et al (2012) Resolving conflict in eutherian mammal phylogeny using phylodynamics and the multispecies coalescent model. Proc Nat Acad Sci 109(37):14942–14947. https://doi.org/10.1073/pnas.1211733109

Springer MS, Gatesy J (2016) The gene tree delusion. Mol Phylogenet Evol 94:1–33. https://doi.org/10.1016/j.mpev.2015.07.018

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

Swofford D (2021) Paup*: Phylogenetic analysis using parsimony (*and other methods). wwwpaupphylosolutionscom, version 4a168

Vane-Wright R, Humphries C, Williams P (1991) What to protect?-Systematics and the agony of choice. Biol Cons 55(3):235–254. https://doi.org/10.1016/0006-3207(91)90030-d

Vellend MK, Cornwell W, Magnuson-Ford K, et al (2011) Measuring phylogenetic biodiversity. In: Magurran AE, McGill BJ (Eds.) Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press, Oxford, chap 14, pp 194–207

Wolfram Research I (2020) Mathematica, Version 12.2. https://www.wolfram.com/mathematica, champaign, IL, 2020

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.