INTRODUCTION

Every organism needs to perform various biological functions to survive and reproduce; however, all such functions cannot be ideally and simultaneously performed because of the trade-offs that exist among them. For example, the bite force and jaw-closing speed of Darwin’s ground finches are negatively correlated because of the existence of a biomechanical trade-off between them (Herrel et al., 2009). Such trade-offs among different functions constrain phenotypes of organisms within limited areas in phenotype space, where the values of multiple traits are quantitatively represented as coordinates.

Received: 31 July 2020  Accepted: 21 December 2020
DOI: 10.1111/2041-210X.13553

RESEARCH ARTICLE

The flipping t-ratio test: Phylogenetically informed assessment of the Pareto theory for phenotypic evolution

Tomoyuki Mikami1  |  Wataru Iwasaki1,2,3,4,5

Abstract

1. An organism cannot be fully optimized for all tasks that are needed for its survival because of the existence of intrinsic trade-offs among those tasks. It was recently proposed that an economics-originated theory, the Pareto theory, is a general theory explaining the rules of phenotypic evolution under such trade-offs. To date, many phenotype datasets have been argued to fit the Pareto theory based on a statistical method named the t-ratio test.

2. Here, we show that this test suffers a serious defect in that general phenotype data can be wrongly concluded to fit the Pareto theory with a very high false-positive rate and that the claim that the Pareto theory is a general theory should definitely be considered with caution. This is because the t-ratio test assumes that all phenotypic traits are independent of one another, but this assumption does not hold true—different traits of organisms have usually been affected by the same phylogenetic history and are thus typically not independent.

3. We developed the flipping t-ratio test to accurately test the Pareto theory by considering phylogenetic background as well as artefacts that can be induced during dimensionality reduction. Using this improved method, we confirm that the phenotype data analysed in previous studies, including the well-known Darwin’s ground finch dataset, no longer support the Pareto theory.

4. We hope that the flipping t-ratio test will contribute to examining which phenotype datasets truly fit the Pareto theory and understanding how diverse phenotypes evolve in natural ecosystems.

KEYWORDS
ammonoid, Darwin’s ground finch, endothermic vertebrate, Pareto theory, phylogenetic constraint

1Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan
2Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan
3Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan
4Institute for Quantitative Biosciences, The University of Tokyo, Tokyo, Japan
5Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Tokyo, Japan

Correspondence
Tomoyuki Mikami
Email: mikami@bs.s.u-tokyo.ac.jp

Wataru Iwasaki
Email: iwasaki@bs.s.u-tokyo.ac.jp

Funding information
This study was supported by JSPS KAKENHI (grant numbers 18J21859, 19H05688, and 18H02493), the Masason Foundation and the Graduate Program for Leaders in Life Innovation of the University of Tokyo.

Handling Editor: Simone Blomberg
Recently, several attempts have been made to adopt an economics-originated theory, the Pareto theory, to explicitly formulate such constraints in phenotype space (Shoval et al., 2012). Assume the existence of an organism with a phenotype whose any function can be better performed without sacrificing other functions. In such a case, evolution is expected to improve that function to a point at which any function cannot be better performed without a sacrifice. The Pareto theory calls those resultant (i.e. realized) phenotypes Pareto optimal states (Sheftel et al., 2013; Shoval et al., 2012). Under some assumptions, the Pareto theory also predicts that (a) realized phenotypes are distributed within a polytope (e.g. triangle or tetrahedron) in phenotype space and (b) the vertices of those polytopes correspond to archetypes, i.e. phenotypes optimized to a single function. In one of the pioneering studies, Shoval et al. (2012) reported that a morphological phenotype distribution of 135 individuals of Darwin’s ground finch species (Geospiza spp.) fits a triangle in two-dimensional phenotype space and that its three vertices correspond to three archetypes specialized to different diets (Figure 1). As a potential general theory that is applicable to various biological phenomena, the Pareto theory has gained popularity in explaining many datasets in evolutionary biology (Kavanagh et al., 2013; Lewitus & Morlon, 2016; Shoval et al., 2012; Szekely et al., 2015; Tendler et al., 2015), molecular biology (Adler et al., 2019; Hart et al., 2015; Hausser & Alon, 2020; Hausser et al., 2019; Kocillari et al., 2018; Korem et al., 2015; Shoval et al., 2012; Thøgersen et al., 2013; Trink et al., 2018) and neurobiology (Cona et al., 2019; Forkosh et al., 2019; Gallagher et al., 2013; Karolis et al., 2019).

To statistically assess whether given phenotype datasets fit the Pareto theory, the t-ratio test was developed (Hart et al., 2015; Shoval et al., 2012). Briefly, the t-ratio test estimates null distributions of phenotypes by means of the randomization (shuffling) of trait values among organisms (or individuals) in each dimension and compares the goodness-of-fit of the original and randomized phenotype distributions to a polytope. Here, the goodness-of-fit is quantified using a t-ratio, which is a volume ratio of the polytope to the convex hull of the phenotype distribution. As visualized in Figure 1, the closer the t-ratio is to 1, the better the phenotype distribution fits a polytope. The p-value is then calculated as the fraction of randomized datasets whose t-ratios are closer to 1 than or equal to that of the original dataset (note that most previous studies ignored cases that t-ratios are exactly the same; in this study, we included those cases by following Edelaar (2013)). Using the t-ratio test, many biological datasets have been reported to fit the Pareto theory: those related to animal morphology (Shoval et al., 2012; Szekely et al., 2015; Tendler et al., 2015), gene expression profiles of cells (Adler et al., 2019; Hart et al., 2015; Hausser et al., 2019; Kocillari et al., 2018; Shoval et al., 2012), brain functional lateralization (Karolis et al., 2019), animal phylogeny (Lewitus & Morlon, 2016), animal behaviour (Forkosh et al., 2019; Gallagher et al., 2013) and human cognition (Cona et al., 2019).

However, despite its broad acceptance, it has been argued that the t-ratio test may fail in controlling type-1 error rates and provide too many false-positive results (Edelaar, 2013). The principal problem is that the different traits of an organism are usually not independent of one another, although the randomization step of the t-ratio test assumes independence. Thus, the t-ratio test generates null distributions that contain cases that are too unrealistic and accordingly overestimates the goodness-of-fit of the original dataset to a polytope, providing false-positive conclusions. The most intrinsic cause of trait dependency is phylogenetic signals (constraints): phylogenetically related organisms necessarily have similar traits because they share a common ancestor (Felsenstein, 1985; Harvey & Pagel, 1991). Another important problem underlying the t-ratio test is the arch effect (also known as the horseshoe effect), which is the ‘distortion’ of data distributions characterized by the curved line-like shapes often observed in principal component analysis (PCA) plots (Digby & Kempton, 1987; Gauch, 1982; Legendre & Legendre, 2012; Minchin, 1987). Such distributions have been reported to cause the t-ratio test to provide false-positive results (Hart et al., 2015); however, many studies using the t-ratio test adopted PCA to reduce dimensionalities without identifying this risk. These frequent adoptions of PCA occur because the dimensionalities of the convex hull of the phenotype distribution and a polytope must be equalized to calculate t-ratio. Without equalizing dimensionalities, the t-ratio closer to

**Figure 1** Calculation of the t-ratio. This example is from a morphological trait dataset of 135 individuals of Darwin’s ground finch (Grant et al., 1985; Shoval et al., 2012). The trait data were fed into PCA, and the first and second principal components are shown. Each colour represents each Geospiza species following Lamichhaney et al. (2015). The red and blue lines represent an SDVMM-estimated polytope (triangle) and the convex hull respectively. Each of the three vertices of the polytope corresponds to an archetype: The top (G. scandens), bottom left (G. fuliginosa) and bottom right (G. magnirostris) are specialists in terms of eating insects/nectar, small seeds and large seeds respectively.
one might not mean the better fit of the phenotype distribution to a polytope. For this reason, dimensionality reduction is a prerequisite to the t-ratio test in many cases in which the archetype number (i.e. the number of polytope vertices) is equal to or smaller than that of the evaluated traits.

In this study, we developed the flipping t-ratio test to ameliorate these two problems associated with the ‘naïve’ t-ratio test. First, the flipping randomization method was developed to randomize phenotype data by considering phylogenetic signals and retaining phenotype distances between sibling groups. Second, the time at which PCA is conducted was changed to prevent false positives due to the arch effect. The flipping t-ratio test was confirmed to be effective using simulated datasets. Our results based on the proposed method suggest that previous claims that datasets containing information on life-history traits of endothermic vertebrates and morphological traits of ammonoids and Darwin’s ground finches support the Pareto theory appear to be artefacts of phylogenetic signals.

2 | MATERIALS AND METHODs

2.1 | ‘Naïve’ t-ratio test and polytope estimation

The naïve t-ratio test is a method that tests the Pareto theory given multitrait phenotype data of multiple organisms/individuals and an archetype number. If the archetype number is equal to or smaller than that of the traits, dimensionality reduction is first conducted using PCA so that the dimensionality is the archetype number minus one. Then, a fitted polytope and the convex hull of the phenotype distribution are calculated, and their area (or volume) ratio, or a t-ratio, is calculated as the goodness-of-fit of the phenotype distribution to the polytope (Figure 1). Here, the polytope is fitted with an unmixing algorithm that was originally developed to decompose hyperspectral images into endmembers and abundances of endmembers (see the next paragraph). Next, a large number of randomized datasets (e.g. 1,000) are generated by shuffling trait values among organisms in each dimension by assuming that all traits (and dimensions) are independent of one another. The convex hulls, fitted polytopes and t-ratios of the randomized datasets are calculated in the same manner. Finally, the p-value is calculated as the fraction of the randomized datasets whose t-ratios are closer to 1 than or equal to that of the original dataset.

Several unmixing algorithms have been used to calculate the polytope that fits a phenotype distribution. In this study, we adopted the successive decoupled volume max-min (SDVMM) method (Chan et al., 2013). The SDVMM method is a pure pixel algorithm; such algorithms identify the polytope that best matches a data distribution under the assumption that the archetypes occur within the convex hull of the distribution. It should be noted that four other unmixing algorithms have been proposed for use in the t-ratio test (Hart et al., 2015), and some previous studies (Adler et al., 2019; Cona et al., 2019; Forkosh et al., 2019; Hart et al., 2015; Koçillari et al., 2018; Lewitus & Morlon, 2016; Szekely et al., 2015; Tendler et al., 2015) have used one of the following four methods: principal convex hull analysis (PCHA; Mørup & Hansen, 2012), the minimum volume enclosing simplex (MVES) method (Chan et al., 2012), minimum volume simplex analysis (MVSA; Li & Bioucas-Dias, 2008) and the simplex identification via split augmented Lagrangian (SISAL) method (Bioucas-Dias, 2009).

Although this is not a key message of this paper, here we claim that the SISAL method is inappropriate for calculating fitted polytopes in the t-ratio test. The SISAL method is one of the unmixing algorithms most frequently used in studies of the Pareto theory and is a minimum volume algorithm; such algorithms can estimate archetype coordinates outside of the convex hull. The problem is that this method allows some outliers to be located outside of the polytope; thus, the t-ratio can become 1 even if the convex hull does not perfectly fit the polytope. In addition, we found that the MVES, MVSA and PCHA methods are sensitive to noise and often estimate archetypes to have forbidden or unrealistic coordinates (unpublished results).

2.2 | Two problems with the naïve t-ratio test

As described in the Introduction, the naïve t-ratio test makes the unrealistic assumption that different traits are independent of one another. Among many sources of independence between traits, we focused on the phylogenetic signals as a general and unignorable factor. Figure 2 provides a schematic example showing that a phenotype distribution fits a polytope (a triangle in this case) just because of phylogenetic signals, not because of the Pareto theory. In such a case, the naïve t-ratio test can present a false-positive conclusion that the dataset fits the Pareto theory by ignoring phylogenetic signals and the trait dependency resulting from them. Note that a dataset containing phenotype data for multiple individuals in each species (e.g. Darwin’s ground finch dataset, Shoval et al., 2012) also has the same problem, where phenotypes are apparently more similar among individuals of the same species than among those of different species.

Another fundamental problem with the naïve t-ratio test is the arch effect, or the curved line-like distributions observed when PCA is...
conducted for dimensionality reduction (Legendre & Legendre, 2012). It has been suggested that if the naive t-ratio test is applied to phenotype data distributions that have curved line-like shapes, positive results can be falsely obtained (Hart et al., 2015). The arch effect becomes particularly prominent if data points gradually vary or have a gradation structure (Legendre & Legendre, 2012). A typical case is seen when phenotypes evolve on a pectinate phylogenetic tree, such as has been observed in virus evolution (Poon et al., 2013). As a simulation example in which phenotypes follow neutral evolution and as has been observed in virus evolution (Poon et al., 2013). As a simulation example in which phenotypes follow neutral evolution and population example in which phenotypes follow neutral evolution and •

2.3 Flipping randomization

We developed a novel randomization method of phenotype data by considering phylogenetic signals: The flipping randomization method. The flipping randomization method generates randomized datasets under the null hypothesis that traits evolve independent of each other along each internal branch of the given phylogenetic tree. Comparison between the input dataset and the generated randomized datasets enables us to evaluate structures behind trait values by subtracting effects of correlations solely due to the phylogenetic signal.

The flipping randomization method uses maximum-likelihood estimation of ancestral trait values for trait-value randomization that considers phylogenetic signals and retains phenotype distances between sibling groups. With this aim, the flipping randomization method additionally requires a full-binary phylogenetic tree with branch lengths as input. Given phenotypic data and a phylogenetic tree, the method recursively visits the ancestral nodes along the leaf-to-root paths using an algorithm developed by Felsenstein (1985) and estimates each ancestral trait value using the branch length-weighted means of those of the child nodes (i.e. under the assumption of Brownian-motion evolution; Figure 4a). Let 1, 2, ..., n be indices of the external nodes of the given phylogenetic tree of n organisms and n + 1, n + 2, ... 2n − 1 be internal nodes labelled in the leaf-to-root order (i.e. the root is 2n − 1). Let \( v_k \) be the branch length between node k and its parent node (1 ≤ k ≤ 2n − 2). Let

\[
\begin{align*}
\nu_k' &= \begin{cases} 
\nu_k & \text{(if node } k \text{ is an external node or } \nu_k' = \nu_j' = 0) \\
\nu_k' + \frac{\nu_i' + \nu_j'}{2} & \text{(otherwise)}
\end{cases}
\end{align*}
\]

where nodes i and j are two child nodes of k. Note that \( \nu_k' \) refers to an ‘adjusted’ branch length that considers uncertainty in the estimation of internal node positions (Felsenstein, 1985). For each trait, assume that a trait value of organism k is given by \( X_k \) (1 ≤ k ≤ n). Then, the values of that trait for the ancestral organisms \( X_{n+1}, X_{n+2}, ..., X_{2n-1} \) are estimated as

\[
X_k = \begin{cases} 
\nu_k' \frac{X_i + X_j}{\nu_i' + \nu_j'} & \text{(if } \nu_i' \neq 0 \text{ or } \nu_j' \neq 0) \\
\nu_k' X_i + X_j & \text{(otherwise)}
\end{cases}
\]

where nodes i and j are two child nodes of k. Then, for the randomization of trait values by considering phylogenetic signals, \( X_{2n-1}^{\text{rand}}, X_{2n-2}^{\text{rand}}, ..., X_1^{\text{rand}} \) are recursively calculated along the root-to-leaf paths by random ‘flipping’ (Figure 4b). Let \( X_{2n-1}^{\text{rand}} = X_{2n-1} \). Given the ancestral node k and its two child nodes i and j,

\[
X_k^{\text{rand}} = X_k^{\text{rand}} + \delta_k (X_i - X_k) \]

\[
X_j^{\text{rand}} = X_j^{\text{rand}} + \delta_k (X_i - X_k),
\]

where \( \delta_k \) is a random variable that takes a value of 1 or −1 with equal probability. Because \( \delta_k \) is independently determined for each internal node k, the flipping randomization method randomizes directions of

**FIGURE 3** Simulation-based example showing that the naive t-ratio test provides a false-positive conclusion because of the arch effect. (a) Pectinate phylogenetic tree topology used for the evolutionary simulation of 10 traits. (b) After PCA for the reduction to two-dimensional phenotype space, a curved line-like distribution appeared because of the arch effect, making the naive t-ratio test provide false-positive conclusions. The red and blue lines represent an SDVMM-estimated polytope (triangle) and the convex hull respectively. The numbers correspond to the node numbers in (a).
Figure 4  Schematic example of the flipping randomization method. (a) Input trait values ($X$) and a phylogenetic tree. Numbers beside the branches indicate branch lengths ($v$) and adjusted branch lengths ($v^{'}$). The ancestral trait values were estimated according to the maximum-likelihood method. (b) A 'flipped' phylogenetic tree and randomized trait values at the external nodes. Note that trait-value distances between sibling nodes are maintained after the randomization.

Finally, this randomization process is repeated for all traits as in the naïve t-ratio test. It may be noted that the same trait distributions can appear among generated datasets by chance, especially when $n$ is small. The use of phylogenetic trees with sufficiently large $n$ (e.g. >30) is thus recommended.

While the above procedure requires a full-binary tree, the flipping randomization method can be easily extended to accept a tree containing polytomies, i.e. internal nodes that have more than two child nodes. In this case, each polytomy is replaced with a binary subtree following Pagel (1992) by applying the unweighted pair group method with arithmetic mean (UPGMA; Sokal & Michener, 1958) to the trait values of its child nodes. The branch lengths of those subtrees are set to zero by following Felsenstein (1985) and Pagel (1992).

### 2.4 Improvement of PCA timing

The other improvement provided by the flipping t-ratio test is related to the time at which PCA is conducted. In contrast to the naïve t-ratio test, in which PCA is conducted before randomization (the original PCA timing; principal components are randomized), PCA is performed immediately after randomization in the flipping t-ratio test (the improved PCA timing; raw trait values are randomized). Because this change causes the original and randomized datasets to be affected by the arch effect in the same manner, the flipping t-ratio test prevents the overestimation of goodness-of-fit due to the arch effect. It may be noted that use of phylogenetic PCA (Revell, 2009) instead of PCA may also prevent false positives caused by the arch effect; however, phylogenetic PCA might attenuate trait changes due to phylogeny-dependent adaptive evolution (Polly et al., 2013).

### 2.5 Evaluation of false-positive rates using simulated datasets

To evaluate the false-positive rates of the naïve and flipping t-ratio tests in relation to the phylogenetic signal and the arch effect, we conducted two evolutionary simulation analyses.

To evaluate the false positives caused by phylogenetic signals, we randomly generated 1,000 different phylogenetic trees with 100 organisms using the rtree function of the package ape (Paradis & Schliep, 2019) in r. Then, the evolution of 10 traits along each tree was simulated using the rTraitCont function with the default evolutionary model (Brownian motion at $\sigma = 0.1$ with initial trait values of zero).

To evaluate the false positives caused by the arch effect, the evolution of 20 organisms with 10 traits along a pectinate phylogenetic tree with all internal and external branch lengths of one (Figure 3a) was simulated 1,000 times using the rTraitCont function with the default evolutionary model (Brownian motion at $\sigma = 0.1$ with initial trait values of zero).
The phenotype data of the simulated extant organisms were then fed into the naïve $t$-ratio test, $t$-ratio test with the flipping randomization method, $t$-ratio test with the improved PCA timing and flipping $t$-ratio test (i.e. with full improvement). The number of archetypes was set to three, and the dimensionality of the phenotype data was reduced to two using PCA. In each test, 1,000 randomizations were conducted to calculate p-values. Any positive conclusions were regarded as false positives because these datasets were generated under the assumption of neutral evolution and did not conform to the Pareto theory.

2.6 | Reevaluation of previous results obtained based on the naïve $t$-ratio test

Finally, using the flipping $t$-ratio test, we reevaluated multispecies phenotype datasets that were previously reported to fit the Pareto theory. These datasets included information on life-history traits of endothermic vertebrates and morphological traits of ammonoids and Darwin’s ground finches (Shoval et al., 2012; Szekely et al., 2015; Tendler et al., 2015). Another multispecies phenotype dataset from microbats (Shoval et al., 2012) was also examined but not reevaluated in this study because we found that this dataset has another fundamental problem associated with outlier data points (see Section 4). The naïve and flipping $t$-ratio tests were conducted with 10,000 randomizations. The two major indices of phylogenetic signals, Blomberg’s $K$ (Blomberg et al., 2003) and Pagel’s $\lambda$ (Pagel, 1999), were calculated with the phylosig function of the package phytools (Revell, 2012) in R for all cases except for the finch dataset, which contained data on intraspecific individuals. Large values of both indices correspond to strong phylogenetic signals. We tested for the presence of phylogenetic signals with randomization tests using $K$ (Blomberg et al., 2003) and likelihood ratio tests using $\lambda$ (Pagel, 1999). We used the contMap function of the package phytools (Revell, 2012) in R to visualize the ancestral traits estimated by the fastAnc function, which implements the method developed by Felsenstein (1985). To run phytools without errors, branch lengths of zero were replaced with a value of $1.0 \times 10^{-6}$.

The dataset pertaining to life-history traits of endothermic vertebrates (Magalhães & Costa, 2009) was retrieved from Table S1 in Szekely et al. (2015) and log-transformed. As in Szekely et al. (2015), we examined the two-trait (log-adult weight and log-longevity) and four-trait (log-adult weight, log-longevity, log-female maturity and log-birth weight) datasets. The exclusion of marsupials in the four-trait dataset in accordance with Szekely et al. (2015) resulted in a two-trait dataset of 2,105 species and a four-trait dataset of 566 species. For the flipping $t$-ratio test, a phylogenetic tree of mammalian animals reconstructed with a supertree approach (Bininda-Emonds et al., 2007) was used, with the intersection containing 966 and 537 species for the two-trait and four-trait datasets respectively. PCA was conducted for only the four-trait dataset to reduce its dimensionality to three. It may be noted that Szekely et al. (2015) used the SISAL method (Bioucas-Dias, 2009) for the polytope estimation, which is not appropriate for the $t$-ratio test, as described previously.

The dataset of morphological traits of ammonoids (Tendler et al., 2015) was retrieved from the Appendix of Saunders et al. (2004). As a convention in ammonoid studies, the ammonoid traits were described by the three geometrical parameters of $W$, $D$ and $S$ (Raup, 1966). Tendler et al. (2015) previously argued that this dataset fits the Pareto theory in terms of both the three-trait ($W$, $D$ and $S$) and two-trait ($W$ and $D$) phenotype spaces, but we focused on the two-trait phenotype space in this study because archetypes of randomized datasets in the $W$-$D$-$S$ space have often been estimated to be at unrealistic outlier coordinates (PCHA, Mørup & Hansen, 2012). It may also be noted that Tendler et al. (2015) used the SISAL method (Bioucas-Dias, 2009) for the estimation of archetypes in the $W$-$D$ space, which is not appropriate for the $t$-ratio test, as described previously. The exclusion of genera occurring before the Frasnian–Famennian extinction in accordance with Tendler et al. (2015) resulted in the retrieval of a two-trait dataset of 504 genera. Because $W$ and $D$ are constrained to $0 < D < 1$ and $1 < W$ by definition, the straightforward application of the flipping randomization method could break this constraint. Thus, the trait values were transformed to $\log(D)$ and $\log(1/W)$ and inverse transformed afterwards. Dimensionality reduction was not conducted for this dataset because the number of archetypes (three) was already larger than that of traits (two). For the flipping $t$-ratio test, the conventional classification system of ammonoids (Korn, 2006) was used as the phylogenetic tree by setting every phylogenetic branch to be of equal length, with the intersection containing 466 genera.

For the dataset of Darwin’s ground finches (Shoval et al., 2012), data on five morphological traits of 135 individuals were retrieved from Grant et al. (1985) and log-transformed as previously described. Here, a maximum-likelihood phylogenetic tree inferred from autosomal genome sequences (Lamichhaney et al., 2015) was used after digitalization using TreeSnatcher plus (Laubach et al., 2012). The taxonomic names proposed by Lamichhaney et al. (2015) were used. Dimensionality was reduced to two dimensions with PCA.

3 | RESULTS

3.1 | Evaluation of false-positive rates using simulated datasets

To assess to what extent the naïve $t$-ratio test gives false-positive conclusions for datasets influenced by phylogenetic signals and/or the arch effect, we generated simulated phenotype datasets and applied the naïve and flipping $t$-ratio tests as well as $t$-ratio tests with either the flipping randomization method or improved PCA timing.

For the datasets with phylogenetic signals, the use of the flipping randomization method drastically decreased the false-positive
rate from approximately 0.68 to 0.06 (Figure 5a). The very large false-positive rate of 0.68 strongly suggests that the naïve t-ratio test is very erroneous if phenotype datasets show a phylogenetic signal. The additional adoption of the improved PCA timing (i.e. the flipping t-ratio test) slightly reduced the false-positive rate (from 0.06 to 0.04, Figure 5a).

For the datasets under the arch effect (generated by evolutionary simulation with a pectinate phylogenetic tree), both the flipping randomization and improved PCA timing substantially reduced the false-positive rates (Figure 5b). The flipping randomization decreased the false-positive rate from 0.31 to 0.12, and the additional adoption of the improved PCA timing drastically reduced the false-positive rate to 0.05 (Figure 5b). This result suggests that the improved PCA timing is also effective for the improvement of the naïve t-ratio test in preventing the arch effect.

### 3.2 Reevaluation of previous results obtained based on the naïve t-ratio test

Finally, using the flipping t-ratio test, we reevaluated multispecies phenotype datasets that were previously reported to fit the Pareto theory (Shoval et al., 2012; Szekely et al., 2015; Tendler et al., 2015) because the naïve t-ratio test has been proven to give false-positive results. Specifically, the flipping t-ratio test was applied to three datasets containing information on life-history traits of endothermic vertebrates and morphological traits of ammonoids and Darwin’s ground finches. Overall, none of the previous results that we assessed remained significant after the application of the flipping t-ratio test, as described below (Table 1; Figure 6).

For the dataset containing information on life-history traits of endothermic vertebrates, it may be noted that Szekely et al. (2015) insisted that phylogeny had limited effects when analysing these data; however, we observed significant phylogenetic signals based on both Pagel’s $\lambda$ and Blomberg’s $K$, strongly supporting the presence of phylogenetic signals for all traits ($p < 0.01$ in all cases; Table 2). The visual inspection of the ancestral states reconstructed according to the maximum-likelihood method also showed that related taxa have similar phenotypes (Figure S1a–d).

For the dataset of D-W morphological traits of ammonoids (Tendler et al., 2015), Pagel’s $\lambda$ significantly indicated the presence of phylogenetic signals for both $D$ ($p < 0.01$) and $W$ ($p < 0.01$), whereas Blomberg’s $K$ did not ($p = 0.23$ and $p = 0.25$, respectively), possibly because of its lower sensitivity (Münkemüller et al., 2012; Table 2). The visual inspection of the reconstructed ancestral states also showed that related taxa have similar phenotypes (Figure S1e–f).

Finally, in the dataset of the five morphological traits of Darwin’s ground finches, dependencies among different traits are particularly evident because the phenotypes of the individuals from the same species clustered together in phenotype space (Figure 1a). It should

---

**FIGURE 5** Comparison of false-positive rates using simulated datasets. (a) Results for datasets with phylogenetic signals alone. (b) Results for datasets showing the arch effect. Error bars indicate 95% confidence intervals.

**TABLE 1** Summary of the reevaluation of the results of previous studies

| Subject                          | Previous study  | $p$-value in flipping t-ratio test | $p$-value in t-ratio test with flipping randomization | $p$-value in t-ratio test with improved PCA timing | $p$-value in naïve t-ratio test |
|----------------------------------|----------------|-----------------------------------|----------------------------------------------------|--------------------------------------------------|--------------------------------|
| Endothermic vertebrates          | Szekely et al. (2015) | 0.123 | NA                                  | NA                                               | 0.037                                      |
| (2 traits)                       |                |                                   |                                                    |                                                  |                                |
| Endothermic vertebrates          | Szekely et al. (2015) | 0.257 | 0.252                              | <0.001                                           | <0.001                                      |
| (4 traits)                       |                |                                   |                                                    |                                                  |                                |
| Ammonoids                        | Tendler et al. (2015) | 0.319 | NA                                  | NA                                               | <0.001                                      |
| Darwin’s ground finches          | Shoval et al. (2012) | 0.120 | 0.077                              | <0.001                                           | <0.001                                      |

Note: NA indicates that PCA was not necessary for dimensionality reduction. Bold font indicates statistical significance.
be noted that while the randomization step of the naïve method apparently destroyed those cluster structures (Figure S2a), that of the flipping t-ratio test did not (Figure S2b).

### 4 DISCUSSION

In this study, we developed the flipping t-ratio test, which involves the flipping randomization method and improved PCA timing, and applied it to multispecies phenotype datasets that were previously reported to fit the Pareto theory. As anticipated, the consideration of phylogenetic signals (Edelaar, 2013) and artefacts in dimensionality reduction resulted in none of the previous conclusions remaining significant. Therefore, we conclude that claims that the Pareto theory can explain many biological phenomena should definitely be judged with caution unless those factors are appropriately considered. Our report is consistent with a recent report by Sun & Zhang (2021), which also showed that the naïve t-ratio test has large false-positive rates.

We analysed Darwin’s ground finch dataset and showed that the p-values provided by the t-ratio test with the flipping randomization method were large when phylogenetic signals were strong (Supporting Information, Figure S3). We also confirmed that the p-value provided by the t-ratio test with the flipping randomization method is relatively large when a binary tree reconstructed by applying UPGMA to the trait values was used (Supporting Information, Figure S3). This result indicates that the use of UPGMA in resolving polytomies would enable conservative tests for preventing

---

**FIGURE 6** Histograms of null distributions of t-ratios in the reevaluation of previous results. The different panels show the results for (a) endothermic vertebrates, two traits; (b) endothermic vertebrates, four traits; (c) ammonoids, and (d) Darwin’s ground finches. The naïve and flipping t-ratio tests are represented by light and dark grey colours respectively. The t-ratios of the original dataset are shown by arrows.

**TABLE 2** Indices of phylogenetic signals in the endothermic vertebrate and ammonoid datasets

| Subject                  | Trait       | \(\lambda\) | Significance of \(\lambda\) | \(K\) | Significance of \(K\) |
|--------------------------|-------------|--------------|------------------------------|------|------------------------|
| Endothermic vertebrates  | Adult weight| 1.00         | \textless 0.01              | 1.56 | \textless 0.01         |
|                          | Maximum longevity | 0.98 | \textless 0.01              | 0.66 | \textless 0.01         |
| (2 traits)               | Adult weight | 1.00         | \textless 0.01              | 2.25 | \textless 0.01         |
|                          | Maximum longevity | 0.96 | \textless 0.01              | 1.02 | \textless 0.01         |
|                          | Female maturity | 0.93 | \textless 0.01              | 0.63 | \textless 0.01         |
|                          | Birth weight  | 1.00         | \textless 0.01              | 2.70 | \textless 0.01         |
| Endothermic vertebrates  | Adult weight | 1.00         | \textless 0.01              | 1.56 | \textless 0.01         |
| (4 traits)               | Maximum longevity | 0.96 | \textless 0.01              | 1.02 | \textless 0.01         |
|                          | Female maturity | 0.93 | \textless 0.01              | 0.63 | \textless 0.01         |
|                          | Birth weight  | 1.00         | \textless 0.01              | 2.70 | \textless 0.01         |
| Ammonoids                | \(D\)        | 0.92         | \textless 0.01              | 0.00 | 0.29                   |
|                          | \(W\)        | 0.72         | \textless 0.01              | 0.01 | 0.25                   |

Note: Pagel’s \(\lambda\) and Blomberg’s \(K\) were calculated for each trait. Bold font indicates statistical significance.
false-positive conclusions. Similarly, the flipping randomization method can also be extended to a case in which a phylogenetic tree is totally unavailable. In this case, a binary tree with branch lengths is reconstructed by applying UPGMA to the trait values of all organisms for a conservative test.

Although we focused on multispecies phenotype datasets in this study, we would like to note that many other biological datasets also have hierarchical structures that resemble phylogenetic signals. For example, in comparative transcriptome datasets (Adler et al., 2019; Hart et al., 2015; Haussler et al., 2019; Korem et al., 2015; Shoval et al., 2012; Thégersen et al., 2013; Trink et al., 2018), the gene expression profiles of cells are likely to be similar if the cells are from the same lineage. Such hierarchical structures will make different ‘traits’ dependent on one another, and the naïve t-ratio test will provide false-positive conclusions, as we proved in this study. Because the flipping randomization method uses binary trees but without assumptions specific to evolutionary processes, the method may also be adopted to reevaluate such datasets.

We would like to point out that three underrated problems in the assessment of the Pareto theory remain. One problem lies in the functional assessment of archetypes. The Pareto theory assumes that the vertices of polytopes correspond to archetypes that have specialized phenotypes, but most previous studies focused on the goodness-of-fit alone. While Hart et al. (2015) examined biological tasks at the polytope vertices, objective and quantitative assessment with that purpose needs to be more widely conducted with a general method. The second problem is related to how to deal with outliers, which often drastically changes the shapes of the convex hull of a data distribution and the fitted polytope (Hart et al., 2015). In an extreme case, three far outliers in two-dimensional phenotype space make a convex hull perfectly fit a triangle irrespective of other phenotype data and incorrectly make the t-ratio test provide positive conclusions. For this reason, we did not include the interspecies phenotype dataset of microbes analysed by Shoval et al. (2012) in this study because we confirmed that the removal of one outlier from the dataset substantially changed the t-ratios regardless of the method used and that the results were almost irreproducible. In addition, while we analysed Darwin’s ground finch dataset by regarding individuals as data points, mean trait values of higher-level units (e.g. species groups) might also be regarded as data points to alleviate effects of outliers (e.g. non-adaptive phenotypes) that strongly affect the calculation of the t-ratio. The third problem lies in how archetype numbers are determined, which directly affects the t-ratio test. While most studies determine numbers of archetypes arbitrarily, Hart et al. (2015) proposed a method to determine the number of archetypes using explained variance curves. It should also be noted that multiple t-ratio tests with different numbers of archetypes will increase family-wise error rate due to the multiple testing problem. Similar problems that would inflate false-positive rates are also discussed by Sun and Zhang (2021).

We envision that the flipping t-ratio test will contribute to the examination of which datasets fit the Pareto theory most reliably and systematically. Beyond the popular view that the Pareto theory is a general theory, the improved method will delineate the true conditions under which the Pareto theory holds true and deepen our understanding of the trade-offs in phenotypic evolution and how diverse phenotypes evolve in natural ecosystems. Furthermore, the flipping randomization method can also be used to extend phylogenetic comparative methods (PCMs), which test evolutionary hypotheses by considering phylogenetic signals (Garamszegi, 2014). Recently, some nonparametric statistical methods were proposed for PCMs using multivariate datasets (Adams, 2014; Adams & Collyer, 2015; Klingenberg & Marugán-Lobón, 2013), but these methods are applicable to simple phylogenetic regressions only. Because flipping randomization is a general method that can be used to generate null models under phylogenetic signals, it can be used to improve such nonparametric statistical methods irrespective of the considered hypotheses.

ACKNOWLEDGEMENTS
We thank Mengyi Sun and Jianzhi Zhang for sharing their manuscript before publication and two reviewers for their helpful comments. We appreciate Naoki Irie, Daichi Funamoto and Saori Mikami for their insightful suggestions and thank Motomu Matsui, Masaki Hoso, Masaya Mukai and Ken Kuroki for critically reading the manuscript.

AUTHORS’ CONTRIBUTIONS
T.M. and W.I. designed research and wrote the paper; T.M. performed research and analysed the data. All authors contributed critically to the drafts and gave final approval for publication.

PEER REVIEW
The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13553.

DATA AVAILABILITY STATEMENT
All empirical data used in this study have already been published. The flipping t-ratio test including the flipping randomization method implemented in MATLAB is available at GitHub https://github.com/TomoyukiMikami/flippping_t-ratio_test and deposited at Zenodo https://doi.org/10.5281/zenodo.4430178 (Mikami, 2021).

ORCID
Wataru Iwasaki https://orcid.org/0000-0002-9169-9245

REFERENCES
Adams, D. C. (2014). A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. Evolution, 68(9), 2675–2688. https://doi.org/10.1111/evo.12463
Adams, D. C., & Collyer, M. L. (2015). Permutation tests for phylogenetic comparative analyses of high-dimensional shape data: What you shuffle matters. Evolution, 69(3), 823–829. https://doi.org/10.1111/evo.12596
Adler, M., Korem, K. Y., Tendler, A., Mayo, A., & Alon, U. (2019). Continuum of gene-expression profiles provides spatial division of labor within a differentiated cell type. Cell Systems, 8(1), 43–52. https://doi.org/10.1016/j.cels.2018.12.008
Bininda-Emonds, O. R. P., Cardillo, M., Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., Price, S. A., Vos, R. A., Gittleman, J. L., &
Polly, P. D., Michelle, L. A., Fabre, A. C., & Goswami, A. (2013). Phylogenetic principal components analysis and geometric morphometrics. *Hystrix, 24*(1), 33–41. https://doi.org/10.4404/hystrix-24.1-6383

Poon, A. F. Y., Walker, L. W., Murray, H., McCloskey, R. M., Harrigan, P. R., & Liang, R. H. (2013). Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS ONE, 8*(11), e78122. https://doi.org/10.1371/journal.pone.0078122

Raup, D. M. (1966). Geometric analysis of shell coiling: General problems. *Journal of Paleontology, 40*(5), 1178–1190.

Revell, L. J. (2009). Size-correction and principal components for interspecific comparative studies. *Evolution, 63*(12), 3258–3268. https://doi.org/10.1111/j.1558-5646.2009.00804.x

Revell, L. J. (2012). *phytools: An R package for phylogenetic comparative biology* (and other things). *Methods in Ecology and Evolution, 3*(2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Saunders, W. B., Work, D. M., & Nikolaeva, S. V. (2004). The evolutionary history of shell geometry in Paleozoic ammonoids. *Paleobiology, 30*(1), 19–43. https://doi.org/10.1666/0094-8373(2004)030<0019:TEHOSG>2.0.CO;2

Sheftel, H., Shoval, O., Mayo, A., & Alon, U. (2013). The geometry of the Pareto front in biological phenotype space. *Ecology and Evolution, 3*(6), 1471–1483. https://doi.org/10.1002/ece3.528

Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., Dekel, E., Kavanagh, K., & Alon, U. (2012). Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science, 336*(6085), 1157–1160. https://doi.org/10.1126/science.1217405

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38*, 1409–1438.

Sun, M., & Zhang, J. (2021). Rampant false detection of adaptive phenotypic optimization by ParTI-based Pareto front inference. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msaa330

Szekely, P., Korem, Y., Moran, U., Mayo, A., & Alon, U. (2015). The mass-longevity triangle: Pareto optimality and the geometry of life-history trait space. *PLoS Computational Biology, 11*(10), e1004524. https://doi.org/10.1371/journal.pcbi.1004524

Tendler, A., Mayo, A., & Alon, U. (2015). Evolutionary tradeoffs, Pareto optimality and the morphology of ammonite shells. *BMC Systems Biology, 9*, 12. https://doi.org/10.1186/s12918-015-0149-z

Thegersen, J. C., Mørup, M., Damkær, S., Molin, S., & Jelsbak, L. (2013). Archetypal analysis of diverse Pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC Bioinformatics, 14*, 279. https://doi.org/10.1186/1471-2105-14-279

Trink, A., Kanter, I., Pode-Shakked, N., Urbach, A., Dekel, B., & Kalisky, T. (2018). Geometry of gene expression space of Wilms’ tumors from human patients. *Neoplasia, 20*(8), 871–881. https://doi.org/10.1016/j.neo.2018.06.006

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Mikami T, Iwasaki W. The flipping t-ratio test: Phylogenetically informed assessment of the Pareto theory for phenotypic evolution. *Methods Ecol Evol*. 2021;12:696–706. https://doi.org/10.1111/2041-210X.13553