TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution

CHAO ZHANG¹, YIMING ZHAO², EDWARD L BRAUN³, AND SIAVASH MIRARAB^{2,*}

¹ *Department, University of California, San Diego, 92093, USA*

² *Electrical and Computer Engineering Department, University of California, San Diego, 92093, USA*

³ *Department of Biology and Genetics Institute, University of Florida, Gainesville, FL 32607, USA*

**Siavash Mirarab, smirarab@ucsd.edu*

ABSTRACT

Erroneous data can creep into sequence datasets for reasons ranging from contamination to annotation and alignment mistakes. These errors *can* reduce the accuracy of downstream analyses such as tree inference and *will* diminish the confidence of the community in the results even when they do not impact the analysis. As datasets keep getting larger, it has become difficult to visually check for errors, and thus, automatic error detection methods are needed more than ever before. Alignment masking methods, which are widely used, completely remove entire aligned sites. Therefore, they *may* reduce signal as much as or more than they reduce the noise. An alternative is designing targeted methods that look for errors in small species-specific stretches of the alignment by detecting outliers. Crucially, such a method should attempt to distinguish the real heterogeneity, which includes signal, from errors. This type of error filtering is surprisingly under-explored. In this paper, we introduce TAPER, an automatic algorithm that looks for small stretches of error in sequence alignments. Our results show that TAPER removes very little data yet finds much of the error and cleans up the alignments.

Key words: Error removal, alignment filtering, homology errors, phylogenomics

Multiple sequence alignments used in phylogenetics and other evolutionary analyses

are susceptible to errors. The input to phylogenetic inference is often prepared through a long pipeline of several error-prone steps (Fig. 1). Together, these steps can leave datasets riddled with many types of errors, called *data pipeline errors* here. For example, contaminating DNA (Simion *et al.*, 2018; Laurin-Lemay *et al.*, 2012; Olson and Hassanin, 2003) and sequencing errors can persist even after the assembly of sequences (Francois *et al.*, 2020; Breitwieser *et al.*, 2019). The process for establishing homology using genome annotations, whole genome alignment, or sequence matching involves complex computational problems (Lunter *et al.*, 2008), and thus, errors in homology are not just possible but rampant (Springer and Gatesy, 2018). Most commonly used methods further assume *orthology*, and errors in orthology detection are common (Laurin-Lemay *et al.*, 2012; Salichos and Rokas, 2011). Alignment errors are also ubiquitous and can impact tree accuracy (Li-San Wang *et al.*, 2011; Liu *et al.*, 2009; Ogdenw and Rosenberg, 2006; Fletcher and Yang, 2010; Smirnov and Warnow, 2020). The prevalence of these errors in phylogenomic datasets has been appreciated (Springer and Gatesy, 2018, 2016; Hosner *et al.*, 2016; Sayyari *et al.*, 2017; Philippe *et al.*, 2017; Laurin-Lemay *et al.*, 2012), and several phylogenomics studies have now been criticized (Springer and Gatesy, 2018, 2016; Gatesy and Springer, 2014; Jeffroy *et al.*, 2006; Salichos and Rokas, 2013; Shen *et al.*, 2017). Data pipeline errors represent a major source of that criticism.

Despite the widespread recognition of the challenges associated with data pipeline errors, there are no truly satisfactory ways to address those errors. Researches can not afford to visually curate their phylogenomic datasets. Many authors have mentioned the need for better methods for detecting errors in data (e.g., Philippe *et al.*, 2017). However, detecting and eliminating error comes with its own caveats and issues, and caution is warranted. Aggressive filtering of data upon any suspicion of error can add bias (e.g., by eliminating highly-divergent but genuinely homologous site patterns that should be considered in analyses). Thus, excessive filtering has the potential to eliminate signal. Consistent with this explanation, studies have found that commonly used alignment

TAPER: PINPOINTING ERRORS IN SEQUENCE ALIGNMENTS

3

Fig. 1. **Data pipeline errors.** Top: Many error-prone steps are needed to produce gene multiple sequence alignments (MSAs) used as input to phylogenomic reconstruction methods. Bottom: an example data pipeline error in the avian dataset of Jarvis *et al.* (2014), identified by Springer and Gatesy (2018), where in gene CEPT1, for three species (Phoenicopterus, Mesitornis, Leptosomus) Intron 3 is aligned with exon 4 of other taxa (left). TAPER is able to detect most but not all of the mis-aligned parts for these sequences (X positions).

masking algorithms often have limited impact on accuracy (Portik and Wiens, 2020) and can even reduce the accuracy of phylogeny inference (Tan *et al.*, 2015). Regardless of whether small errors *actually* impact the tree inference, a question on which there is disagreement, the existence of data pipeline errors has the *potential* to impact the results, which can diminish confidence in analyses. Thus, even when errors do not impact the analyses, researchers benefit from detecting and removing them as long as removing the errors does not remove signal. Achieving this objective requires error detection methods that are targeted and find *minimal* portions of the data with putative errors.

The existing methods for data filtering mostly focus on finding entire genes or entire species that should be eliminated (e.g., Hosner *et al.*, 2016; Molloy and Warnow, 2018; Huang *et al.*, 2016). Somewhat more targeted are alignment masking methods that eliminate entire sites from a sequence alignment in order to avoid mis-alignment (e.g., Castresana, 2000; Dress *et al.*, 2008; Capella-Gutiérrez *et al.*, 2009; Rajan, 2012). Another form of filtering is to keep all genes and all species but to remove specific species from

specific genes because of fragmentation (Sayyari *et al.*, 2017), evidence of unexpected patterns of tree topology and branch length (de Vienne *et al.*, 2012; Wickett *et al.*, 2014; Mai and Mirarab, 2018), or detection of rogue taxa (Westover *et al.*, 2013).

Existing trimming methods operate at coarse levels. Many forms of pipeline errors can be limited to a small stretch of a sequence in a particular species (not all species) and in a particular set of positions (not an entire gene). For example, the dominant form of error that Springer and Gatesy (2018) found in the avian data of Jarvis *et al.* (2014) relates to small pieces of introns being annotated as exons (e.g., Fig.1b). Such errors are limited to a handful of sequence in a small stretch of sites. Eliminating entire genes, entire sites, or entire species because a small stretch includes errors wastes data. We need methods to find specific stretches of a specific species in a specific gene that appear erroneous.

Finding such small stretches of errors can be formulated as outlier positions in a multiple sequence alignment (MSA). Imagine a MSA that is almost fully conserved for all species across all sites except that a small stretch from a single species is close to random with respect to other sequences aligned to it. Such outliers *can* be detected. One would hope alignment methods would leave such sequences unaligned, but most commonly used alignment methods are known to over-align (Löytynoja and Goldman, 2008; Löytynoja and Goldman, 2005). Thus, these stretches are suspect and likely to be erroneous.

Alignment trimming using the outlier detection paradigm needs to contend with two related issues. On the one hand, sequence divergence among species is a function of their phylogenetic relationships and evolutionary rates, and simply being divergent from other sequences cannot be viewed as evidence of error. If one naively looks for species that look unusually divergent compared to the remaining sequences, an outgroup or an ingroup with a highly accelerated rate of evolution would be mistakenly taken as erroneous (a false positive error detection). On the other hand, rates of evolution change across sites, and thus, how much divergence is “normal” depends on the sequence context.

We advocate for two-dimensional (2D) error detection: finding stretches in a

sequence that are unusually divergent compared to other sequences, calibrating the *normal* level of divergence based on both genomic positions (columns) *and* the species (rows). An outlier should be detected only if a sequence is unusually divergent along both axes. For tree-based trimming, de Vienne *et al.* (2012) has pioneered such a two-dimensional approach in their mathematically elegant method Phylo-MCOA, and TreeShrink (Mai and Mirarab, 2018) follows a similar philosophy. However, we are not aware of two-dimensional outlier methods for sequence data. The most relevant method is DivA (Zepeda Mendoza *et al.*, 2014), which does look for stretches to remove from individual sequences using sliding windows, but is only applicable to amino acid sequences.

In this paper, we introduce the Two-dimensional Algorithm for Pinpointing ERrors (TAPER) that takes a multiple sequence alignment as input and outputs outlier sequence positions. Using both simulated and real data, we show that TAPER is able to pinpoint errors in multiple sequence alignments without removing large parts of the alignment.

DESCRIPTION

2D outlier detection algorithm

We first describe our general-purpose 2D outlier detection algorithm (Alg. 1). The input is a set of n aligned sequences with length L on any arbitrary alphabet Γ (e.g., nucleotide) plus missing data (e.g., gaps). The output is a delineation of each sequence into alternating normal and outlier regions. We use *letter* to refer to a position in a sequence, not counting gaps (or ambiguous letters like X for amino acid) as letters.

Step 1. Compute a divergence score for each letter c in each column i of the alignment. Small scores should indicate agreement with a strong consensus in that site and the largest values should indicate disagreement from an otherwise strong consensus. High scoring letters are candidate outliers. Thus, the drop from large values to small values should be not be gradual; instead, it should have a fast drop as deviation from the consensus weaken.

Algorithm 1 TAPER algorithm. *A*: Input alignment on n sequences of length L on

alphabet Γ and gap letter $-$. k, p, q, T : user-provided parameters. $[X]$ denotes $1 \dots X$.

```

procedure TAPER( $A, k, p, q, T$ )
   $M \leftarrow \text{DIVERGENCESCORES}(A)$ 
  for each row  $m_i \in M$  do
     $\mathbf{v}_i \leftarrow \text{WINDOWSCORES}(m_i, k)$ 
     $c_i \leftarrow \text{JENKSBREAKPOINT}(\mathbf{v}_i)$ 
     $(t_1, \dots, t_n) \leftarrow \text{ADJ}((\mathbf{v}_1, \dots, \mathbf{v}_n), (c_1, \dots, c_n), p, q, T)$ 
    for each row  $a_i \in A$  do
       $s_i \leftarrow \text{REMOVEOUTLIER}(a_i, \mathbf{v}_i, t_i, k)$ 
  return  $(s_1, \dots, s_n)$ 

procedure DIVERGENCESCORES( $A$ ) ▷ Step 1
  for  $i \in [L]$  do
    for  $c \in \Gamma$  do
       $p_c \leftarrow$  relative frequency of  $c$  in the  $i$ -th column
       $u_i \leftarrow$  number of letters  $c$  in  $\Gamma$  with  $p_c \neq 0$ .
      for  $j \in [n]$  do
         $M_{j,i} \leftarrow \begin{cases} \emptyset, & \text{if } A_{j,i} = - \\ 1/(u_i \times p_{A_{j,i}}), & \text{elsewise} \end{cases}$ 
  return  $M$ 

procedure WINDOWSCORES( $m, k$ ) ▷ Step 2
   $m' \leftarrow$  remove  $\emptyset$  from  $m$ 
   $l \leftarrow$  length of  $m'$ 
  for  $i \in [l - k + 1]$  do
     $\mathbf{v}_i \leftarrow \text{Median}(m'_i, \dots, m'_{i+k-1})$ 
  return  $\mathbf{v}$ 

procedure JENKSBREAKPOINT( $\mathbf{v}$ ) ▷ Step 3
   $l \leftarrow$  length of  $\mathbf{v}$ 
   $\mathbf{v}' \leftarrow$  sorted  $\mathbf{v}$ 
   $s_0 \leftarrow 0$ 
  for  $i \in [l]$  do
     $\mathbf{s}_i \leftarrow \mathbf{s}_{i-1} + \mathbf{v}'_i$ 
   $j \leftarrow \text{argmax}_{i \in [l-1]} \frac{\mathbf{s}_i^2}{i} + \frac{(\mathbf{s}_l - \mathbf{s}_i)^2}{l-i}$ 
  return  $\mathbf{v}'_j$  ▷ Step 4

procedure ADJ( $(\mathbf{v}_1, \dots, \mathbf{v}_n), (c_1, \dots, c_n), p, q, T$ )
   $a \leftarrow [p \times n]$ -th largest value in  $c_1, \dots, c_n$ 
  for  $i \in [n]$  do
     $l \leftarrow$  length of  $\mathbf{v}_i$ 
     $b_i \leftarrow [q \times l]$ -th largest value in  $\mathbf{v}_i$ 
     $t_i \leftarrow \max\{a, b_i, c_i, T\}$ 
  return  $(t_1, \dots, t_n)$ 

procedure REMOVEOUTLIER( $a, \mathbf{v}, t, k$ ) ▷ Step 5
   $l \leftarrow$  length of  $\mathbf{v}$ 
   $s \leftarrow$  remove gaps from  $a$ 
   $N_0 \leftarrow 0$ 
   $O_0 \leftarrow 0$ 
  for  $i \in [k-1]$  do
     $N_i \leftarrow N_{i-1} + \begin{cases} 1, & \text{if } \mathbf{v}_i \leq t \\ 0, & \text{elsewise} \end{cases}$ 
     $O_i \leftarrow O_{i-1} + \begin{cases} 0, & \text{if } \mathbf{v}_i \leq t \\ 1, & \text{elsewise} \end{cases}$ 
     $B_i^N \leftarrow (N, i-1)$ 
     $B_i^O \leftarrow (O, i-1)$ 
  for  $i \in \{k, \dots, l\}$  do
    if  $N_{i-1} > O_{i-k}$  then
       $N_i \leftarrow N_{i-1} + \begin{cases} 1, & \text{if } \mathbf{v}_i \leq t \\ 0, & \text{elsewise} \end{cases}, B_i^N \leftarrow (N, i-1)$ 
    else
       $N_i \leftarrow O_{i-k} + \begin{cases} 1, & \text{if } \mathbf{v}_i \leq t \\ 0, & \text{elsewise} \end{cases}, B_i^N \leftarrow (O, i-k)$ 
    if  $O_{i-1} > N_{i-k}$  then
       $O_i \leftarrow O_{i-1} + \begin{cases} 0, & \text{if } \mathbf{v}_i \leq t \\ 1, & \text{elsewise} \end{cases}, B_i^O \leftarrow (O, i-1)$ 
    else
       $O_i \leftarrow N_{i-k} + \begin{cases} 0, & \text{if } \mathbf{v}_i \leq t \\ 1, & \text{elsewise} \end{cases}, B_i^O \leftarrow (N, i-k)$ 
    if  $N_i > O_l$  then
       $(S, i) \leftarrow (O, l)$ 
    else
       $(S, i) \leftarrow (N, l)$ 
  while  $i \neq 0$  do
    if  $S = O$  then
      for  $j \in \{i, \dots, i+k-1\}$  do
         $s_j \leftarrow \emptyset$  ▷ Mark position  $j$  as outlier
       $(S, i) \leftarrow B_i^S$ 
     $s' \leftarrow$  add gaps of  $a$  back to  $s$ 
  return  $s'$ 

```

111 While several such functions can be imagined, here, we use a scoring method that Henikoff
 112 and Henikoff (1992) used for sequence weighting. We score a letter c in column i as $\frac{1}{u_i \times p_{i,c}}$
 113 where u_i is the number of unique letters in the column, and $p_{i,c}$ is the fraction of the letters
 114 in the column that are c . It can be checked that this score satisfies our criteria (Fig. S1).

Step 2. Since per-column scores are noisy, we combine them along small windows *per each sequence*. We first remove gaps from each sequence. Then, selecting an odd constant value k (e.g., 11), for each overlapping window of size k of each sequence, we assign the median score of the letters in that window as the score of that window. This step produces a distribution of scores for each sequence.

Step 3. For each sequence, we seek to find which of the windows have abnormally high scores; these are considered candidate outliers. To do so, we divide windows of *each sequence* into a low scoring and a high scoring group. We find a cutoff point, t , such that the squared deviations from the mean of the scores below t plus the squared deviations from the mean of points above t is minimized. This approach is the Jenks (1967) natural breaks optimization and is equivalent to the 2-means clustering.

Step 4. The initial cutoffs from Step 3 include high-scoring windows for all sequences, regardless of whether any outlier exists. To allow sequences with no outliers, we use three parameters $0 < p, q < 1$ and $T > 1$ to adjust cutoffs. We set the final cutoff of a sequence to the maximum of four quantities: the highest p -quantile of all cutoff values across *all* sequences, the highest q -quantile of all window scores for *that* sequence, the threshold value T , and the initial cutoff t . Thus, the adjusted cutoff will include no high-scoring windows for sequences where the windows have homogeneous scores (controlled by q) or all scores are within the normal range compared to other sequences (controlled by p). User can adjust p , q , and T to control the aggressiveness of the method; p controls how many species can have error while q controls error length; T controls overall aggressiveness. Windows with scores greater than the final cutoff are called red and the rest are called green.

Step 5. We divide the original sequence without gaps into alternating *normal* and *outlier* sections. Note that a window can span both sections. The sections boundaries are set using dynamic programming, seeking to maximize the number of red windows fully contained in outlier sections and green windows fully contained in normal sections. The point of this

step is smoothing of red and green assignments. Since windows that fall on the section boundaries do not count towards the optimization score, frequent switches between normal and outlier regions are eliminated in this step.

The 5-step procedure we described above is called two-dimensional because scores are computed along the columns (step 1), but outliers are detected (Step 2 and 3) and smoothed (Step 4) along the rows. Thus, for a letter to be marked as an outlier, it must be in several windows, all of which have abnormally high scores compared to the rest of their respective columns, when compared to other windows of the same sequence.

TAPER Algorithm

2D outlier detection is expected to be more effective at catching smaller errors with smaller values of k and longer errors with larger values of k (our results confirm this notion; see Fig. S2). To be able to catch a wider range of error lengths, we have devised a strategy to combine several k values, each with a different p, q setting. We run the 2D outlier algorithm on multiple k values and report their union. However, to ensure that a specific k value only catches errors of its intended length, we define an upper limit to the length of the detected error, such that only detected errors with lengths less than the upper limit are flagged. In our preliminary analyses, we confirmed that using two or more values of k dramatically improves recall for short errors of length 22 (Fig. S3–S4). In our default setting, we use k values 5, 9, and 17, with p set respectively to 0.25, 0.25, 0.1, and q set to 0.1, 0.25, 0.5; we only keep errors of length up to $6 \times k$ for $k = 5$ or 9. These settings are motivated by preliminary analyses but are kept fixed as we study various datasets.

Table 1. *Datasets used in simulations.*

Dataset	Type	# sequences	Seq. Length	Alignment
16S.B (371 sub-clades)	RNA	23–1958 (mean: 544)	854–2940 (mean: 1362)	structure-based+curated
Early-bird (19 genes)	DNA	72–171 (mean: 158)	461–2335 (1168)	Mafft + curated
Small-AA (RV100-BBA0039)	Protein	807	375	Gold-standard

BENCHMARK

Evaluation setup

Datasets.— To benchmark TAPER, we inserted random errors into MSAs of three real datasets with different properties (Table1) and studied whether TAPER can detect them. The 16S.B dataset is an RNA dataset of 16S, with gold-standard alignments built by Cannone *et al.* (2002) and used for benchmarking alignment methods (e.g., Liu *et al.*, 2009; Mirarab *et al.*, 2015; Nguyen *et al.*, 2015). Because this dataset includes 27643 sequences, it enables us to sub-sample the alignment to create subsets with controlled divergence levels. We selected subtrees with a range of diameters (i.e., maximum distance between species) from a phylogeny built from the original MSA. We first find *all* clades in small diameter ranges in increments of 0.025 ($[0, 0.025]$, $[0.025, 0.05]$, $\dots$, $[0.975, 1]$) and then select up to 10 largest clades in each diameter range, requiring at least 20 species. This procedure gave us 371 sub-datasets of the 16S.B dataset, ranging in tree diameter from 0.043 to 0.990. The avian early-bird dataset consists of DNA sequences from 19 genes, aligned automatically but also curated manually by Hackett *et al.* (2008). The RV100-BBA0039 is one of the largest AA alignments available as part of the BALiBASE datasets of gold-standard curated alignments used for benchmarking (Thompson *et al.*, 2005).

Simulating errors.— Errors are added to a predefined number of sequences in the alignment (m) and for a predefined length (l). Sequences with errors are selected uniformly at random. For each of the m erroneous sequences, a position to start the error is selected uniformly at random, and l of the original non-gap letters are replaced with a randomly chosen letter. For DNA, we choose randomly among the four possible nucleotides. For

proteins, we draw the replaced letter from the set of all amino acids such that the chance of selecting each amino acid is proportional to the number of codons that encode to it (e.g., the chance of flipping to Leucine is six times higher than that of Methionine).

Our experiments explore several error profiles (m and l). First, we fix m to be 5% of the total number of sequences, and vary l between $(2, 3, 5, 8, 16, 32, 64) \times 11$, including 64×11 only for 16S.B as the length is long enough to accommodate such long errors, and excluding 32×11 from early-bird genes with mean sequence length below 704 as the error would be more than half of the length. We then fix l to 8×11 and set m to 1 or to 2%, 5%, 10%, or 20% of the number of sequences. To ensure these discrete choices do not impact results, on 16S.B, we also draw l and m from a normal distribution. We determine m by first drawing a value x from a normal distribution centered at 20 and a standard deviation of 2 and set $m = \frac{N}{x} + 1$, where N is the total number of sequences in the alignment (so around 5% of sequences are erroneous). We set the length of the error by drawing from another normal distribution centered at 50 with a standard deviation of 10.

Evaluation criteria. — We define a False Positive (FP) as any letter that is not an error, but is marked erroneous; a False Negative (FN) as any letter that is erroneous, but is not marked; a True Positive (TP) as any letter that is erroneous, and is marked erroneous; and, a True Negative (TN) as any a letter that is not erroneous, and is not marked. Each letter in the MSA, excluding gap letters, is categorized into one of the four groups, allowing us to compute recall ($\frac{TP}{TP+FN}$) and FPR ($\frac{FP}{TN+FP}$). We also report the percentage of the alignment made of errors before ($\frac{FN+TP}{FN+FP+TP+TN}$) and after ($\frac{FN}{FN+TN}$) filtering, as well as the percentage of alignment retained ($\frac{FN+TN}{FN+FP+TP+TN}$).

Results

16S.B and impact of parameters. — On the 16S.B dataset, TAPER is effective in finding erroneous sequences (Fig. 2). The FPR is low, never exceeding 0.16% on average

Fig. 2. (a) ROC curve (Recall versus FPR) for the subsamples from 16S.B dataset with varying error length, error frequency, grouped into five categories (colours) based on their subtree diameter. Left: Fixing percent of erroneous sequences to 5%, we change the error length from 2×11 nt to 64×11 nt. We also draw both length and frequency from Normal distributions, as describe in text (cross). Right: Fixing error length to 8×11 nt, we vary the number of erroneous sequence between 1 and 2%, 5%, 10%, or 20% of the size of the subtree. Vertical lines on top show the FPR before adding error. See also Figure S6. (b) Reduction in the portion of the alignment nucleotides that are in error. Arrows show percent error before and after filtering. Note the log-scale (see S5b for normal scale).

across model conditions, and is even lower before adding errors. Note that we assume that the starting alignment is fully correct, but some FPs may in fact be undetected errors. TAPER retains more than 99% in most cases and never less than 97% of alignment letters (Fig. S5); thus, the method does not overzealously remove data. Depending on the model condition, 70% to 98% of erroneous letters are detected (Fig. 2a). The remaining error is never more than 0.47% of the alignment after filtering, compared to 2.5% before (Fig. 2b).

Tree diameter, which is an indicator of the divergence level, has a substantial impact on the effectiveness of TAPER (Fig. 2 and S6). Over all conditions and using a simple linear model, diameter explains a statistically significant 12% of the variance in the recall (p-value according to an ANOVA test: $\ll 10^{-5}$; see Table S1). As the diameter increases, recall reduces gradually, but the largest reductions happen when errors are relatively short (Fig. S6). Increasing the diameter also increase FPR, especially when diameter is >0.5 . The number of sequences in each alignment has a small impact on the recall (0.5% of total variance) and FPR values (Fig. S7).

FPR is not substantially impacted by the error length but the recall is. Error length explains 32% of variance in the recall (Table S1), which is the lowest with small errors of length 22nt, quickly increases as errors become longer, peaks somewhere between 88 and 352nt, and degrades slightly after that (Fig. 2a). The low recall for short errors should be compared with the amount of error left in the alignment after filtering (Fig. 2b), which is less than 0.03% on average even for high diameters. The amount of remaining error is the highest (0.47% on average) when inserted errors are long, but even then, error has reduced dramatically (2.7% before filtering). When we vary error frequency, we observe small and inconsistent changes in the recall, decreasing for low diameter and increasing for high diameters. FPR increases substantially as the error frequency goes up but remains below 0.16% even when 20% of sequences are erroneous. Finally, when error length and frequency are drawn from a normal distribution, we see consistent results (Fig. 2).

Early-bird dataset. — On the early-bird dataset, which includes 19 genes, TAPER in most cases has FPR below 0.1% and recall above 60% (Fig. 3a) and reduces the error to less than 1% of the alignment in all cases (Fig. 3b). However, the effectiveness of TAPER varies across genes (Fig. S8a). At one end of the spectrum, on EEF and HMG genes, which have high diameter (1.1 and 0.97), TAPER has FPR close to 0.1% and relatively low recall (as low as 33% when errors are long and no more than 78% under other conditions). In addition to high diameter, the HMG gene includes 72 out of the 171 species. The other

Fig. 3. (a,c) ROC results with various error profiles for all genes of the early bird (a) and the AA (c) datasets. (b,d) Reduction in percentage of the nucleotide positions in the alignment that consists of errors for early bird (b) and AA datasets. Arrows show percent error before and after filtering. Note the log-scale. For each gene, below it, we indicate the diameter, mean sequence length, and the number of species present in that gene. See also Fig. S8.

extreme is the BDNF gene (diameter: 0.40) where the mean recall is between 74% and 94% and FPR is below 0.2% across all conditions and below 0.05% in most. While BDNF is known to have patterns of branch length variation that differ from the other genes in the early-bird dataset (cf., Fig. 8 of Braun *et al.*, 2019), several other genes such as NT3 and IRF2 also have high accuracy (Fig. S8). Overall, diameter has modest negative impact on the recall of TAPER and the impact is most noticeable for longer errors (Fig. S8b). According to a linear model only 9% of variation in recall is explained by diameter

(p-value: $\ll 10^{-5}$; see Table S2). For example, the IRF2 gene with a moderately high diameter (0.8) has very high recall (ranging from 71% to 93%). The impact of sequence length and number of sequences on the recall was significant (p-values: 0. $< 10^{-5}$, 0.003) but modest (1.3% and 1% of total variance).

The error profile matters. The error frequency has small impacts on the recall (Table S2), but error length has a large impact (20% of variance; p-value: $\ll 10^{-5}$). In particular, short errors are difficult for all genes, while long errors are difficult for many but not all genes (e.g., EGR1). Overall, the five factors examined and their interactions explain only 42% of the total variance (Table S2).

Small-AA dataset. — On the AA dataset, where the original alignment includes a set of sequences that are substantially divergent from the rest (Fig. S10), TAPER has higher FPR (mean: 1.34%) than the previous datasets . We attribute the high FPR to the uncertainty in the alignment, consistent with the observation that even before adding errors, TAPER removes 1.30% of the alignment. This uncertainty makes it harder for TAPER to find inserted errors. Depending on the model condition, the median recall ranges between 54% and 90%. Increasing the frequency of error and the length of the error both improve the recall. A relatively small portion of alignments, typically less than 3%, is removed (Fig. S11) and the remaining error after filtering is never more than 0.5%.

BIOLOGICAL EXAMPLES

To test the effectiveness of TAPER on real data, we revisited 56 genes from the avian dataset of Jarvis *et al.* (2014) analyzed manually by Springer and Gatesy (2018). Springer and Gatesy (2018) found errors in all 56 of these genes (some small and others relatively large). TAPER marked between 0% and 1.2% (mean: 0.2%) of nucleotides in these genes as erroneous (Fig. 4). Since the ground truth of where *all* errors lie is not known, to evaluate results of TAPER, we rely on a likelihood-based metric. If the

Fig. 4. Increase in the log likelihood after filtering computed both for the ML gene tree (GT) and the species tree (ST) versus the portion of the nucleotides in an alignment that are removed. Grey: TAPER filtering. Red: the same amount of data is removed from each gene as TAPER removes but selected from random stretches with the same length. y-axis: change in log-likelihood normalized by likelihood before filtering.

nucleotides removed are in fact erroneous, we expect the likelihood of the maximum likelihood gene tree to increase, more than it would if we remove the same number of nucleotides from random positions. Further, since gene trees for this dataset are notoriously difficult to estimate accurately (Mirarab *et al.*, 2014), we also compute the likelihood on the best-estimate of the species tree from Zhang *et al.* (2018) (we present this tree as Fig. S12a) after estimating branch lengths for each specific gene; despite a potential for true gene tree discordance, we expect that removing errors would increase the likelihood of the species tree. For both gene trees and the species tree, we see dramatic improvements in the log likelihood (measured using RAxML (Stamatakis, 2014)), far exceeding small increases in the likelihood we get by random removal of the same amount of data (Fig. 4). The more data is removed, the more the likelihood increases. Removing as little as 1% of the data can result in 10% improvements in the log likelihood. Reassuringly, the two outgroups were not removed more often than other species (Fig S12b).

Visual inspection of genes shows that out of 68 cases of error identified by Springer and Gatesy (2018), 24 of them are fully or mostly found by TAPER, and in eight cases, a minority of erroneous positions are found (Table S3). Among the remaining 36 cases (Table S4), 21 are errors that are too short (≤ 10 bp) or too frequent (≥ 10 out of 48

species) for TAPER to be effective, and in 7 cases, they are somewhat frequent (≥ 5). In another three cases, only a handful of species are present in those sites, making the errors to be a high *portion* of non-gapped species, which TAPER cannot detect (Table S4). Overall, TAPER marked 21 out of 29 cases that were not too short or too frequent.

Summary

TAPER was able to reduce error dramatically under varied conditions, but it also had its limitations. TAPER is not very effective in finding very short errors (e.g., below a length 20). On the other extreme, since our method is looking for outlier regions, it cannot detect very large errors (e.g., those that are more than half the sequence length) or those that are too frequent. In between these extremes of short and long is the sweet spot for the error detection by TAPER where there is enough signal to detect oddity of the pattern but the error is not so large that it looks like a real phylogenetic divergence. For shorter errors, changing settings of TAPER (e.g., reducing k) could perhaps make the method more sensitive, but that sensitivity would come at the expense of more FP filtering, which we tried hard to avoid. For longer errors, methods like TreeShrink that look for unexpected patterns in branch lengths provide viable alternatives. Besides error length, other factors also mattered. Errors in alignments with very few sequences or very short sequences were harder to detect, as were errors in alignments that were very divergent.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

AVAILABILITY

TAPER is implemented fully in Julia and is made available open-source at <https://github.com/chaoszhang/AlignmentErrorRemoval>. It takes as input a fasta file and outputs a masked alignment.

REFERENCES

- Braun, E. L., Cracraft, J., and Houde, P. 2019. Resolving the Avian Tree of Life from Top to Bottom: The Promise and Potential Boundaries of the Phylogenomic Era. In *Avian Genomics in Ecology and Evolution*, pages 151–210. Springer International Publishing, Cham.
- Breitwieser, F. P., Perte, M., Zimin, A. V., and Salzberg, S. L. 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome research*, 29(6): 954–960.
- Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D’Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Müller, K. M., Pande, N., Shang, Z., Yu, N., and Gutell, R. R. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics*, 3: 2.
- Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15): 1972–1973.
- Castresana, J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, 17(4): 540–552.
- de Vienne, D. M., Ollier, S., and Aguilera, G. 2012. Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis. *Molecular Biology and Evolution*, 29(6): 1587–1598.
- Dress, A. W., Flamm, C., Fritsch, G., Grünwald, S., Kruspe, M., Prohaska, S. J., and Stadler, P. F. 2008. Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular Biology*, 3(1): 7.
- Fletcher, W. and Yang, Z. 2010. The Effect of Insertions, Deletions, and Alignment Errors

on the Branch-Site Test of Positive Selection. *Molecular Biology and Evolution*, 27(10):
2257–2267.

Francois, C. M., Durand, F., Figuet, E., and Galtier, N. 2020. Prevalence and Implications
of Contamination in Public Genomic Resources: A Case Study of 43 Reference
Arthropod Assemblies. *G3*, 10(2): 721–730.

Gatesy, J. and Springer, M. S. 2014. Phylogenetic Analysis at Deep Timescales: Unreliable
Gene Trees, Bypassed Hidden Support, and the Coalescence/Concatallescence
Conundrum. *Molecular phylogenetics and evolution*, 80: 231–266.

Hackett, S. J., Kimball, R. T., Reddy, S., Bowie, R. C. K., Braun, E. L., Braun, M. J.,
Chojnowski, J. L., Cox, W. A., Han, K.-L., Harshman, J., Huddleston, C. J., Marks,
B. D., Miglia, K. J., Moore, W. S., Sheldon, F. H., Steadman, D. W., Witt, C. C., and
Yuri, T. 2008. A phylogenomic study of birds reveals their evolutionary history. *Science*,
320(5884): 1763–1768.

Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein
blocks. *Proceedings of the National Academy of Sciences*, 89(22): 10915–10919.

Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L., and Kimball, R. T. 2016.
Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the
Landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, 33(4): 1110–1125.

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A.,
Al-Shehbaz, I., and Edger, P. P. 2016. Resolution of Brassicaceae phylogeny using
nuclear genes uncovers nested radiations and supports convergent morphological
evolution. *Molecular biology and evolution*, 33(2): 394–412.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth,
B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J.,
Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid,

REFERENCES

19

- 365 M. S., Zavidovych, V., Subramanian, S., Gabaldon, T., Capella-Gutierrez, S.,
366 Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C.,
367 Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y.,
368 Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V.,
369 Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdociimi, F., Samaniego, J. A.,
370 Velazquez, A. M. V., Alfaro-Nunez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten,
371 T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P.,
372 Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao,
373 J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E.,
374 Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jonsson, K. A., Johnson, W., Koepfli,
375 K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R.,
376 Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alstrom, P., Edwards, S. V.,
377 Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert,
378 M. T. P., and Zhang, G. 2014. Whole-genome analyses resolve early branches in the tree
379 of life of modern birds. *Science*, 346(6215): 1320–1331.
- 380 Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. 2006. Phylogenomics: the
381 beginning of incongruence? *Trends in Genetics*, 22(4): 225–231.
- 382 Jenks, G. F. 1967. The data model concept in statistical mapping. *International yearbook*
383 *of cartography*, 7: 186–190.
- 384 Laurin-Lemay, S., Brinkmann, H., and Philippe, H. 2012. Origin of land plants revisited in
385 the light of sequence contamination and missing data. *Current Biology*.
- 386 Li-San Wang, Leebens-Mack, J., Wall, P. K., Beckmann, K., de Pamphilis, C. W., and
387 Warnow, T. 2011. The Impact of Multiple Protein Sequence Alignment on Phylogenetic
388 Estimation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*,
389 8(4): 1108–1119.
- 390 Liu, K., Raghavan, S., Nelesen, S. M., Linder, C. R., and Warnow, T. 2009. Rapid and

- Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees.
Science, 324(5934): 1561–1564.
- Loytynoja, A. and Goldman, N. 2005. An algorithm for progressive multiple alignment of
sequences with insertions. *Proceedings of the National Academy of Sciences*, 102(30):
10557–10562.
- Löytynoja, A. and Goldman, N. 2008. Phylogeny-aware gap placement prevents errors in
sequence alignment and evolutionary analysis. *Science*, 320(5883): 1632–1635.
- Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., and Hein, J. 2008.
Uncertainty in homology inferences: assessing and improving genomic sequence
alignment. *Genome research*, 18(2): 298–309.
- Mai, U. and Mirarab, S. 2018. TreeShrink: fast and accurate detection of outlier long
branches in collections of phylogenetic trees. *BMC Genomics*, 19(S5): 272.
- Mirarab, S., Bayzid, M. S., Boussau, B., and Warnow, T. 2014. Statistical binning enables
an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215):
1250463–1250463.
- Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. 2015. PASTA:
Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences.
Journal of computational biology, 22(5): 377–86.
- Molloy, E. K. and Warnow, T. 2018. To Include or Not to Include: The Impact of Gene
Filtering on Species Tree Estimation Methods. *Systematic Biology*, 67(2): 285–303.
- Nguyen, N.-p. D., Mirarab, S., Kumar, K., and Warnow, T. 2015. Ultra-large alignments
using phylogeny-aware profiles. *Genome Biology*, 16(1): 124.
- Ogdenw, T. H. and Rosenberg, M. S. 2006. Multiple sequence alignment accuracy and
phylogenetic inference. *Systematic biology*, 55(2): 314–328.

- 415 Olson, L. E. and Hassanin, A. 2003. Contamination and chimerism are perpetuating the
416 legend of the snake-eating cow with twisted horns (*Pseudonovibos spiralis*). A case study
417 of the pitfalls of ancient DNA. *Molecular Phylogenetics and Evolution*, 27(3): 545–548.
- 418 Philippe, H., Vienne, D. M. d., Ranwez, V., Roure, B., Baurain, D., and Delsuc, F. 2017.
419 Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*.
- 420 Portik, D. M. and Wiens, J. J. 2020. Do Alignment and Trimming Methods Matter for
421 Phylogenomic (UCE) Analyses? *Systematic Biology*.
- 422 Rajan, V. 2012. A Method of Alignment Masking for Refining the Phylogenetic Signal of
423 Multiple Sequence Alignments A Method of Alignment Masking for Refining the
424 Phylogenetic Signal of. *Molecular biology and evolution*, 30(3): 689–712.
- 425 Salichos, L. and Rokas, A. 2011. Evaluating ortholog prediction algorithms in a Yeast
426 Model Clade. *PLoS ONE*.
- 427 Salichos, L. and Rokas, A. 2013. Inferring ancient divergences requires genes with strong
428 phylogenetic signals. *Nature*, 497(7449): 327–31.
- 429 Sayyari, E., Whitfield, J. B., and Mirarab, S. 2017. Fragmentary Gene Sequences
430 Negatively Impact Gene Tree and Species Tree Reconstruction. *Molecular Biology and*
431 *Evolution*, 34(12): 3279–3291.
- 432 Shen, X.-x., Hittinger, C. T., and Rokas, A. 2017. Contentious relationships in
433 phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*,
434 1(5): 0126.
- 435 Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J. C., Manuel, M., Philippe, H.,
436 and Telford, M. J. 2018. A software tool 'CroCo' detects pervasive cross-species
437 contamination in next generation sequencing data. *BMC Biology*.
- 438 Smirnov, V. and Warnow, T. 2020. Phylogeny Estimation Given Sequence Length
439 Heterogeneity. *Systematic Biology*, (0): 1–47.

Springer, M. S. and Gatesy, J. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94(Part A): 1–33.

Springer, M. S. and Gatesy, J. 2018. On the importance of homology in the age of phylogenomics. *Systematics and Biodiversity*, 16(3): 210–228.

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

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., and Dessimoz, C. 2015. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology*, 64(5): 778–791.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. 2005. BALiBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins: Structure, Function, and Bioinformatics*, 61(1): 127–136.

Westover, K. M., Rusinko, J. P., Hoin, J., and Neal, M. 2013. Rogue taxa phenomenon: A biological companion to simulation analysis. *Molecular Phylogenetics and Evolution*, 69(1): 1–3.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E. J., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., Rothfels, C. J., Pokorný, L., Shaw, A. J., DeGironimo, L., Stevenson, D. W., Surek, B., Villarreal, J. C., Roure, B., Philippe, H., DePamphilis, C. W., Chen, T., Deyholos, M. K., Baucom, R. S., Kutchan, T. M., Augustin, M. M., Wang, J. J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G. K.-S., and Leebens-Mack, J. J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45): 4859–4868.

REFERENCES

23

- 465 Zepeda Mendoza, M., Nygaard, S., and da Fonseca, R. R. 2014. DivA: detection of
466 non-homologous and very divergent regions in protein sequence alignments. *BMC*
467 *Research Notes*, 7(1): 806.

468 Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. 2018. ASTRAL-III: polynomial time
469 species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*,
470 19(S6): 153.

24

REFERENCES

471

APPENDIX

472

Supplementary Materials

473

SUPPLEMENTARY FIGURES

Fig. S1. **Score function.** Left: DNA, Right: Protein.

REFERENCES

25

Fig. S3. A comparison of various strategies for selecting k as the length of error changes (boxes). We either use a single value of $k = 11$ or the union of 2, 3, or 4 values of k . When using unions of multiple k s, we take results for each k only at a certain range of error lengths; for 2k setting: $k = 5$ for error length $[0, 30]$, and $k = 9$ for other lengths; For 3k setting: $k = 5$ for error length $[0, 30]$ $k = 9$ for error length $[0, 54]$ $k = 19$ for other lengths; For 4k setting: $k = 5$ for error length $[0, 20]$ $k = 7$ for error length $[0, 35]$ $k = 11$ for error length $[0, 66]$ $k = 17$ for other lengths. Using one k has very low recall in the case with short error length. The other three settings do not universally dominate each other (there is tradeoff between FPR and recall). Overall, the 2k setting seems to have substantially less recall than 4k, with small advantages in FPR. 3k setting generally has better FPR than the other two methods, but also slightly lower recall. Overall, to protect against FP error filtering, we chose the 3k setting that provides a balance between high recall and low FPR. See Figure S4 for more details.

REFERENCES

27

Fig. S4. A comparison of various strategies for selecting k as the length of error changes (boxes). We either use a single value of $k = 11$ or the union of 2, 3, or 4 values of k . When using unions of multiple k s, we take results for each k only at a certain range of error lengths; for 2k setting: $k = 5$ for error length $[0, 30]$, and $k = 9$ for other lengths; For 3k setting: $k = 5$ for error length $[0, 30]$ $k = 9$ for error length $[0, 54]$ $k = 17$ for other lengths; For 4k setting: $k = 5$ for error length $[0, 20]$ $k = 7$ for error length $[0, 35]$ $k = 11$ for error length $[0, 66]$ $k = 17$ for other lengths. Using one k has very low recall in the case with short error length. The other three settings do not universally dominate each other (there is tradeoff between FPR and recall). Overall, the 2k setting seems to have substantially less recall than 4k, with small advantages in FPR. 3k setting generally has better FPR than the other two methods, but also slightly lower recall. Overall, to protect against FP error filtering, we chose the 3k setting that provides a balance between high recall and low FPR.

Fig. S5. (a) Percentage of the alignment remaining after filtering (the total number of non-gap nucleotide positions in the alignment after divided by before filtering) across model conditions as the error length and frequency changes on the 16S.B dataset. (b) Similar to Figure 2b, we show change in percent error but without log scale.

REFERENCES

29

Fig. S6. Impact of diameter on Recall and FPR on the 16S dataset.

Fig. S7. Impact of sequence count on the Recall and FPR on the 16S dataset.

REFERENCES

31

Fig. S8. a) Recall and FPR of all genes as we change sequence error or error length. Genes are sorted by their average recall. We show diameter, mean sequence length, and the number of species for each gene under its name. b) Impact of Diameter, sequence count (top left), sequence length (top right), and sequence count (bottom) on the recall. For Sequence count, we show results with and without the outlier gene HMG.

Fig. S9. Results on AA dataset.

REFERENCES

33

Fig. S10. The AA alignment RV100_BBA0039 from the BALIBASE benchmarking dataset. The alignment includes a minority of sites that look very different from the rest.

Fig. S11. The AA alignment RV100_BBA0039 from the BALIBASE benchmarking dataset. The alignment includes a minority of sites that look very different from the rest.

SUPPLEMENTARY TABLES

Table S1. ANOVA test on the 16S dataset, showing impact of four factors and their interactions: Error Length (ErrLen), Error Frequency (n), Diameter, and Sequence Count (N). X:Y corresponds to interactions of variables X and Y.

	Df	Sum Sq	Mean Sq	F value	Pvalue	PctExp
ErrLen	6	21.17	3.53	991.32	0.000000	31.9
n	4	0.01	0.00	0.99	0.410483	0.0
Diameter	1	7.73	7.73	2171.08	0.000000	11.6
N	1	0.32	0.32	90.54	0.000000	0.5
ErrLen:Diameter	6	4.94	0.82	231.22	0.000000	7.4
n:Diameter	4	0.13	0.03	9.04	0.000000	0.2
ErrLen:N	6	0.13	0.02	6.15	0.000002	0.2
n:N	4	0.11	0.03	7.75	0.000003	0.2
Diameter:N	1	0.03	0.03	9.44	0.002130	0.1
ErrLen:Diameter:N	6	0.24	0.04	11.10	0.000000	0.4
n:Diameter:N	4	0.00	0.00	0.17	0.951870	0.0
Residuals	8860	31.53	0.00			47.5

REFERENCES

37

Table S2. ANOVA test on the early-bird dataset, showing impact of five factors and their interactions: Error Length (ErrLen), Error Frequency (n), Diameter, Sequence Length (SL), and Sequence Count (N). X:Y corresponds to interactions of variables X and Y.

	Df	Sum Sq	Mean Sq	F value	<i>p</i> -value	% Var Exp
ErrLen	5	4.65	0.93	66.21	0.000000	20.3
n	4	0.23	0.06	4.11	0.002645	1.0
Diameter	1	2.13	2.13	151.73	0.000000	9.3
SL	1	0.29	0.29	20.62	0.000006	1.3
N	1	0.29	0.29	20.77	0.000006	1.3
ErrLen:Diameter	5	0.28	0.06	3.96	0.001478	1.2
n:Diameter	4	0.08	0.02	1.49	0.202735	0.4
ErrLen:SL	5	0.16	0.03	2.23	0.049866	0.7
n:SL	4	0.04	0.01	0.78	0.538640	0.2
Diameter:SL	1	0.17	0.17	12.38	0.000455	0.8
ErrLen:N	5	0.27	0.05	3.81	0.002017	1.2
n:N	4	0.03	0.01	0.51	0.729552	0.1
Diameter:N	1	0.04	0.04	3.14	0.076594	0.2
SL:N	1	0.18	0.18	12.99	0.000330	0.8
ErrLen:Diameter:SL	5	0.08	0.02	1.16	0.327340	0.4
n:Diameter:SL	4	0.02	0.00	0.31	0.868535	0.1
ErrLen:Diameter:N	5	0.09	0.02	1.30	0.260338	0.4
n:Diameter:N	4	0.06	0.01	1.07	0.371634	0.3
ErrLen:SL:N	5	0.15	0.03	2.19	0.052842	0.7
n:SL:N	4	0.14	0.04	2.50	0.041228	0.6
Diameter:SL:N	1	0.03	0.03	2.31	0.128471	0.1
ErrLen:Diameter:SL:N	5	0.14	0.03	2.06	0.067832	0.6
n:Diameter:SL:N	4	0.08	0.02	1.40	0.230659	0.3
Residuals	940	13.20	0.01			57.8

Table S3. Errors identified by Springer and Gatesy (2018) that TAPER is able to detect fully (Found), mostly (Majority), or to a lesser degree (Minority). **Red: Error is either too short (length ≤ 10) or involves too many sequences (≥ 10). Orange: Error involves somewhat high numbers of sequences (between 5 and 10).** †: erroneous homology is restricted to a subset of the region identified by Springer and Gatesy (2018). See the supplementary file **supplementary-error-pictures.xlsx** for pictures of errors found.

Gene	Positions	L	n	Description of error by Springer and Gatesy (2018)	Found
1066	859-1212	353	2	Partial intron 8 (Pterocles, Podiceps) is aligned with different region of intron 8 in Balearica (no homology with other sequences) and exons 9-11 in remaining taxa	Full
82	413-498	85	2	<i>Manacus</i> and <i>Acanthisitta</i> exon 6 is combination of intron 5 (5') and exon 6 (3')	Full
1087	673-753	80	2	Intron 8 (Columba, Anas) aligned against exon 8 in other taxa	Full
1077	1150-1206	56	2	Intron 11 (Melopsittacus, Colius) aligned with exon 10 in other taxa	Full
1028	334-386	53	2	<i>Gavia</i> and <i>Struthio</i> (intron 3) aligned with exon 3 in other taxa	Full
1054	577-707	30	2	Part of intron 5 (Merops, Nestor) is aligned against exons 6 and 7 in others	Full
1079	184-201	17	2	3' end of intron 2 (Eurypyga, Haliaeetus leucocephalus) aligned against exon 2 in other taxa	Full
82	136-146	11	2	Intron 2 (Cathartes, Tauraco) is aligned with exon 2 in other taxa	Full
16	2212-2213	2	2	5 end of intron 21 in Geospiza and Acanthisitta is aligned against 5 end of exon 22 in other taxa	Full
1039	1784-1999	210	3	Exon 14 sequences in Gavia, Phalacrocorax, and Opisthocomus are poorly aligned with other sequences and show different splice site boundaries	Full
1016	82-126	44	3	Part of intron 1 and exon 2 (Struthio, Tinamus, Pelecanus) aligned with exon 1 in others	Full
44	634-737	103	4	Exon 9 in Colius, Acanthisitta, Cariama, and Pterocles is a combination of the 5' region of intron 8 and the 3' region of intron 9	Full
1042	289-310	21	5	Exon 1 (Gallus, Haliaeetus leucocephalus, Melopsittacus, Tauraco, Columba) aligned against intron 1 in other taxa	Full
1098	161-309	140	2	Intron 1 (Opisthocomus, Phalacrocorax) aligned with exon 1 in other taxa	Majority
90	733-915	80	2	Intron 8 (Charadrius, Tauraco) aligned with exon 9 in other taxa	Majority
1014	705-768	64	2	Part of intron 7 (Eurypyga and Columba) aligned against exon 8 in others	Majority
1013	250-273	24	2	<i>Gallus</i> and <i>Tinamus</i> not homologous with other sequences that are present (which are intron 1); problems with <i>Meleagris</i> exon 5	Majority
93	124-139	15	2	3' end of intron 1 (Tyto, Pelecanus, Melopsittacus, Meleagris) is aligned with 3' end of exon 1 in others	Majority
30	691-762	71	3	Intron 3 (Phoenicopterus, Mesitornis, Leptosomus) aligned with exon 4 of other taxa	Majority
1044	342-441	99	4	Intron 5 (Opisthocomus, Eurypyga, Balearica, Cathartes) aligned against exon 5 in other taxa	Majority
1039	1307-1368	61	4	Intron 12 (Corvus, Columba, Pelecanus, Pygoscelis) aligned against exon 12 in others	Majority
89	740-796	54	4	Four taxa (Gallus, Eurypyga, Fulmarus, Manacus) have partial intron 6 sequence instead of partial exon 6	Majority
1087	1006-1019	13	5	5' end of intron 10 (Cuculus, Apaloderma, Columba, Colius, Anas) aligned against 5' end of exon 10 in others	Majority
1039	1360-1524	164	5?	Exon 11 sequences are poorly aligned across avian tree and have different intron-exon boundaries. The numbers of exons in different taxa also varies.	Majority
1077	448-583	120	2	Regions of intron 5 (?) and 6 (Melopsittacus, Colius) are aligned with exon 7 in others	Minority
1028	1158-1232	80	2	Sequences are not homologous in Merops and Nipponia versus other taxa	Minority
1089	367-421	50	2	Two taxa (Eurypyga, Pygoscelis) have intron 4 aligned against exon 4 in other taxa	Minority
99	401-506	100	7	Intron 3 (Columba, Pycoides, Chlamydotis, Tauraco, Pygoscelis, Cathartes, Tinamus) aligned against exon 3 in other taxa	Minority
1083	127-144	17	7	Seven taxa (Chlamydotis, Anas, Tinamus, Cariama, Tauraco, Acanthisitta, Opisthocomus) with intron 2 aligned against exon 2 in others	Minority
1097	912-1120	200	10	Intron 6 in some taxa (e.g., Tinamus) aligned with exon 6 in other taxa (e.g., Struthio)	Minority
28	1-81† (48-80)	32	13	Intron 2 in 13 taxa (e.g., Haliaeetus albicilla) is aligned with exon 1 and exon 2 in others (e.g., H. leucocephalus)	Minority
20	754-842	88	14	5' region of intron 1 (Gallus and others) aligned against 3' region of intron 1 (e.g., Falco and others); middle portion of intron 1 (Gallus group) aligned against 5' end of exon 2 (Falco group). Also some taxa in Gallus group (palaeognaths, Cathartes, Podiceps, Charadrius) are misaligned with others in group	Minority

REFERENCES

39

Table S4. Errors identified by Springer and Gatesy (2018) and missed by TAPER. Notations same as Table S3.

†: real error boundary. Error is short (≤ 10 bp) or frequent (≥ 10); Error is somewhat frequent ($\geq 5, \leq 10$).

Gene	Positions	L	n	Description of error by Springer and Gatesy (2018)
1040	106-138 [†] (-129)	24	2	Intron 1 (Opisthocomus, Apaloderma) aligned against exon sequences in others
1085 ¹	73-97	24	2	Two taxa (Anas, Nipponia) have intron 1 aligned against exon 1
62	93-129 [†] (-105)	12	2	3' end of intron 1 in Cathartes and Cariama is aligned with 3' end of exon 1 in other taxa
51	478-489 [†] (-487)	10	2	3' end of intron 2 in Balearica and Eurypyga is aligned with 5' end of exon 3 in other taxa
1045	415-425	10	2	Intron 3 in Colius and Pelecanus aligned with exon 3 in other taxa
1050 ²	1372-1395	23	3	Intron 12 (Chaetura, Melopsittacus, Acanthisitta) versus exon 13 (Cuculus, Tinamus, Gallus, Nestor, Taeniopygia, Geospiza)
12	1-324 [†] (300-)	20	3	Intron 1 (Eurypyga, Fulmarus, Columba) aligned against exon 1 in other taxa
1078	98-128 [†] (-117)	10	4	3' end of exon 1 in some taxa (Tauraco, Taeniopygia, Eurypyga, Apaloderma) in some taxa aligned against 5' end of exon 2 in various others
1038 ³	230-252	22	4	Part of intron 2 (Fulmarus, Melopsittacus, Nestor, Tyto) is aligned with exon 2 in other taxa
1001 ⁴	1596-1617	21	4	Intron 14 (Merops, Tauraco, Charadrius, Mesitornis) aligned with exon 14
64 ⁵	1-147 [†] (133-)	14	4	Exon 1 (Pycoides, Melopsittacus, Gallus, Manacus) is aligned with part of intron 1 (Nestor, Chlamydotis, Mesitornis, Buceros, Merops, Falco, Corvus)
88	1114-1122	8	4	Four taxa (Taeniopygia, Caprimulgus, Eurypyga, Chlamydotis) have intron sequence
1036	838-921	84	5	Sequences in Aptenodytes, Gavia, Phoenicopterus, Chlamydotis, and Phaethon are not homologous to other sequences (exon versus intron sequences are unclear)
1044	25-52	27	5	Intron 1 (Fulmarus, Gavia, Opisthocomus, Colius, Eurypyga) aligned against exon 1 in others
1094	1069-1095	26	5	Five taxa (Gavia, Apaloderma, Chaetura, Pterocles, Pycoides) with intron 7 and five others with exon 7 (Gallus, Meleagris, Pelecanus, Podiceps, Merops)
1062	58-70	12	5	Exon 1 (Cariama, Meleagris, Manacus, Gallus, Colius) versus intron 1 in other taxa
56	106-115	10	5	3' end of exon 1 in Taeniopygia, Columba, Corvus, Geospiza, and Gallus is aligned with 3' end of intron 1 in other taxa
1056	157-180	23	6	Exon 1 (Charadrius, Melopsittacus, Gallus, Cuculus, Taeniopygia, Columba) versus intron 1 in other taxa
1085	187-191	4	6	Six taxa with intron 2 (Charadrius, Buceros, Opisthocomus, Merops, Haliaeetus albicilla, Egretta) aligned against exon 2
1038	76-115	41	7	Part of intron 1 (Egretta, Apaloderma, Podiceps, Balearica, Eurypyga, Phalacrocorax, Caprimulgus) aligned against exon 1 in other taxa (region also missing for many taxa)
84	124-129	5	7	Seven taxa with different intron 1 boundaries (Fulmarus, Anas, Phoenicopterus, Balearica, Gavia, Charadrius, Struthio) versus other taxa
1020	84-174	90	8	Part of intron 2 (Phaethon, Nipponia, Aptenodytes, Phoenicopterus, Cariama, Columba, Mesitornis, Falco) aligned with exon 2 (Pycoides, Calypete, Meleagris, Charadrius, Egretta, Gallus, Haliaeetus leucocephalus, Cuculus)
37	262-312 [†] (-273)	9	8	3' region of exon 1 (or 5' region of intron 1?) in eight taxa (Gallus, Falco, Columba, Balearica, Buceros, Taeniopygia, Geospiza, Cuculus) aligned with 5' region of exon 2 (or 3' region of exon 1?) in other taxa
1083	43-45	2	9	Nine taxa (Leptosomus, Chaetura, Corvus, Taeniopygia, Cuculus, Phaethon, Eurypyga, Balearica, Egretta) share 3 bp from 3' end of intron 1, which is aligned with last 3 bp of exon 1 in many other taxa (also some with missing data)
1089	25-69	44	10	Seventeen taxa with intron 1 sequences aligned against 10 taxa with exon 1 sequences
5	147-161	14	10	Exon 1 (Meleagris, Gallus, Fulmarus, Cariama, Pygoscelsis, Taeniopygia, Aptenodytes, Haliaeetus leucocephalus, Colius) aligned against intron 1 (other taxa with sequence)
1029	58-85	28	11	Intron 1 (Opisthocomus, Anas, Podiceps, Eurypyga, Cariama, Aptenodytes, Tauraco, Caprimulgus, Leptosomus, Tinamus, Merops) aligned against exon 1 (Struthio, Falco, Haliaeetus leucocephalus, Gavia, Phaethon, Pygoscelsis, Gallus, Meleagris)
74	241-252	3	11	5' end of intron 2 in Eurypyga, Corvus, and Taeniopygia is aligned with 5' end of exon 3 in other taxa
1005	121-124	3	11	Exon 1 (Fulmarus, Podiceps, Melopsittacus, Cathartes, Meleagris, Falco, Tinamus, Gallus, Haliaeetus leucocephalus, Cariama) aligned with intron 1 in other taxa
1014	787-849	75	14	Part of intron 7 (Acanthisitta, Balearica, Colius, Meleagris, Pterocles, Apaloderma, Eurypyga, Columba, Pygoscelsis, Falco, Opisthocomus, Caprimulgus, Nipponia, Leptosomus) aligned against exon 8 in others.
1027	1055-1153	99	16	Part of intron 8 in 16 taxa (e.g., Struthio, Pygascelsis, Galga) aligned against non-homologous region in other taxa (e.g., Tinamus, Aptenodytes, Meleagris)
1025	412-518	106	16	Part of intron 4 (Calypete, Leptosomus, Phalacrocorax, Balearica, Chlamydotis, Pelecanus, Fulmarus, Cariama, Tyto, Anas, Aptenodytes, Columba, Apaloderma, Tauraco, Corvus, Acanthisitta) aligned with part of exon 5(?) in Haliaeetus leucocephalus, Nipponia, Gallus, Geospiza, Taeniopygia, and Tinamus (and to a lesser extent with other taxa that are largely missing in this region)
1035	58-64	6	17	Intron 1 (Tauraco, Phaethon, Caprimulgus, Fulmarus, Tyto, Pelecanus, Gavia, Apaloderma, Haliaeetus albicilla, Leptosomus, Balearica, Podiceps, Charadrius, Pygoscelsis, Struthio, Phoenicopterus, Egretta) aligned against exon 1 in other taxa
1067	529-532	4	20	First four bp of intron 5 in 20 taxa (e.g., Gallus, Pelecanus) aligned against first four bp of exon 6 in 28 taxa (e.g., Meleagris, Egretta)
1037	361-408	45	20?	5' and 3' ends of intron 2 are included for some taxa are are misaligned across avian alignment, including misalignment with last 3 bp of exon 2
1049	2725-2762	37	20?	Mix-up of intron 8 and exon 9 sequences

¹ 2 out of 6 aligned have the error; thus, high frequency.

² 3 out of 9 aligned have the error; thus, high frequency.

³ Starts at 247 for 3 of the four species; thus, short for most sequences.

⁴ 6 columns in the middle are all gaps, others are not visually clear errors.

⁵ 4 out of 11 aligned have the error; thus, high frequency.