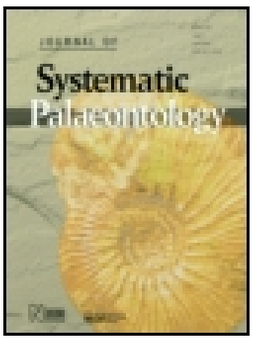

Journal of Systematic Palaeontology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjsp20

Comparative cladistics: identifying the sources for differing phylogenetic results between competing morphology-based datasets

Pablo A. Goloboff & Paul C. Sereno

To cite this article: Pablo A. Goloboff & Paul C. Sereno (2021): Comparative cladistics: identifying the sources for differing phylogenetic results between competing morphology-based datasets, Journal of Systematic Palaeontology, DOI: 10.1080/14772019.2021.1970038

To link to this article: https://doi.org/10.1080/14772019.2021.1970038

Journal of Systematic Palaeontology, 2021
http://dx.doi.org/10.1080/14772019.2021.1970038

Taylor & Francis
Taylor & Francis Group

Comparative cladistics: identifying the sources for differing phylogenetic results between competing morphology-based datasets

Pablo A. Goloboff[a,b]* and Paul C. Sereno[c]

[a]*Unidad Ejecutora Lillo, Consejo Nacional de Investigaciones Científicas y Técnicas, Fundación Miguel Lillo, Miguel Lillo 251, 4000 San Miguel de Tucumán, Argentina;* [b]*Research Associate, American Museum of Natural History, Central Park West at 79th St., New York, NY, 10024 USA;* [c]*Department of Organismal Biology and Anatomy and Committee on Evolutionary Biology, University of Chicago, Chicago, IL, 60637 USA*

(Received 3 August 2020; accepted 31 July 2021)

Competing morphology-based phylogenetic analyses are routinely compared and evaluated only by *a posteriori* phylogenetic topology and group support, with little or no analysis of *a priori* data sources responsible for differing results. Although discordant characters and character-state scores are usually key to differing results (more so than variation in terminal taxa), programs currently do not exist to facilitate even simple two-way comparisons of morphology-based datasets, despite the impracticality and imprecision of manual assessment for datasets involving hundreds of characters. This paper, a first step to remedy this circumstance, presents methods (within TNT) to identify, compile and evaluate differences in characters and character states between datasets that yield different trees and degrees of group support. These apparently simple and urgently needed computer-assisted routines involve conceptual and computational challenges, even when competing morphology-based datasets are grossly similar. Example two-way comparisons are presented using pairs of similar morphology-based datasets for hominin and basal dinosaur radiations.

Keywords: comparative cladistics; taxon; character; character state; scoring; index

Introduction

Morphology-based character data provides the sole basis for reconstructing the majority of the tree of life, which is known only from the fossil record. Although a framework for the tree of life may be traced with recent and ancient genomes, understanding most branchpoints as well as mapping morphological changes to any gene-based tree requires the parsing and analysis of morphology-based character data.

Challenges of morphology-based character data

Despite their continued centrality to the tree of life and a history of analytic ferment dating back more than a century, morphology-based data have proven far more complex than molecular data. Morphology-based datasets are typically compiled by an array of taxon experts steeped in direct observation of specimens, drawing on multifaceted data sources, often depending on disputed alternative interpretations of anatomy (Sereno 2007a; Vogt *et al.* 2010; Dahdul *et al.* 2010; Richter & Wirkner 2014). Characters are constructed as independent variables with mutually exclusive character states, some with weighted or ordered transformations. The input to phylogenetic analysis, thus, is quite particular and difficult to reproduce without the expertise of trained systematists. Efforts to automate phenotypic data gathering with ontological nomenclature and string-matching algorithms (Dahdul *et al.* 2012; Deans *et al.* 2015; Dececchi *et al.* 2015; Eliason *et al.* 2019) or to engage non-experts in assembly of morphological data (O'Leary *et al.* 2018) have yet to demonstrate that such methods are capable of rigorously comparing or building new morphology-based phylogenetic datasets of routine complexity.

A major challenge for comparison of morphology-based datasets is the characters themselves – how they are defined, coded and scored. Systematists have never adopted a standardized format for the structure, coding and expression of morphology-based characters, so the same character may be formulated and coded in a variety of ways (Hawkins 2000; Sereno 2007a). Locating comparable character data, thus, is a major challenge for either manual or computer-assisted dataset comparison (Sereno 2009; Dececchi *et al.* 2015). Another major challenge in data comparison is variation in terminal taxa, when terminal taxa are present in only one of the datasets or when they overlap but reference different taxonomic levels (specimen, species, genus). In these

*Corresponding author. Email: pablogolo@csnat.unt.edu.ar

Published online 18 Oct 2021

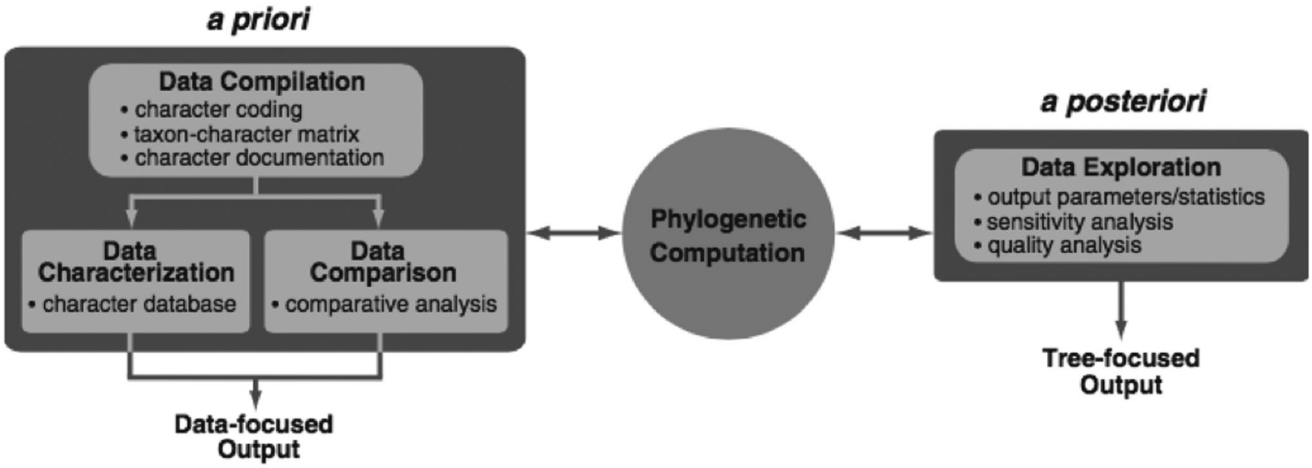

Figure 1. Scheme for morphology-based phylogenetic analysis that separates *a priori* data considerations and output from *a posteriori* tree considerations and output (from Sereno 2009).

cases, portions of competing datasets that do not overlap precisely still may play a significant role in generating different phylogenetic results.

As a consequence, a patchwork of overlapping morphology-based phylogenies has arisen over the last 40 years of the cladistic era, the underlying data for which are difficult to compare (Sereno 2009). These studies incorporate newly available character data and often modify, select from, or expand previous datasets, with minimal discussion of the differences between datasets. Typically, major new studies that involve first-hand observations will simply announce a different, or 'new', phylogenetic result and highlight select synapomorphies, the position of particular taxa, or particular clades. Comparison to previous studies usually focuses on '*a posteriori*' tree-based methods, such as decay indices or tree constraints, which highlight the significance of particular branchpoints (Fig. 1, right). Often little or no comment is given to assessing '*a priori*' character-based causes, i.e. character selection and character-state scoring, that are usually the primary factors generating different phylogenetic results (Fig. 1, left). Whereas quantitative measures for tree support or differences in tree configuration between analyses are routine, quantitative measures based on comparisons between the datasets themselves have only been more recently proposed (Sereno 2009), with little impact on actual practice.

A handful of studies take a more rigorous approach to character and character-state differences between datasets that consider similar phylogenetic problems. Harris *et al.* (2007) generated 'consensus' matrices to examine the subset of shared characters and taxa between two competing hypotheses or to isolate remaining characters used only by one hypothesis. Whitlock & Wilson (2013) compared datasets graphically (character distribution map) by dividing morphology-based characters into anatomical partitions with the percentage of character data those partitions comprise. This may be useful for a gross visual assessment of the differences between competing data matrices. These are rudimentary manual techniques that will never keep pace with the generation of new datasets of varying size and taxonomic scope.

Comparative cladistics

The analytical hurdles facing comparative cladistics were first outlined with examples compiled by hand (Sereno 2009; Sereno & Brusatte 2009). The present paper approaches comparative cladistics from within the phylogenetic program TNT (Goloboff *et al.* 2008; Goloboff & Catalano 2016). The implementation outlined below covers only the aspects most amenable to formalization, and assumes that equivalent characters in each dataset have similar, if not identical, names. When similar, or even identical, characters are expressed in a different form, it can be challenging to recognize their overlap without human (expert) intervention. The effects that different scorings have on character-state reconstructions on alternative trees or on group support with characters added or subtracted requires a more formal treatment. Such comparisons are very hard to perform manually or assess qualitatively, even when they appear relatively simple. We examine a number of such conceptual issues that arise in the course of dataset comparison.

We make comparisons between two pairs of grossly similar datasets. The first involves a revision of a dataset for hominin phylogeny by the same lead author

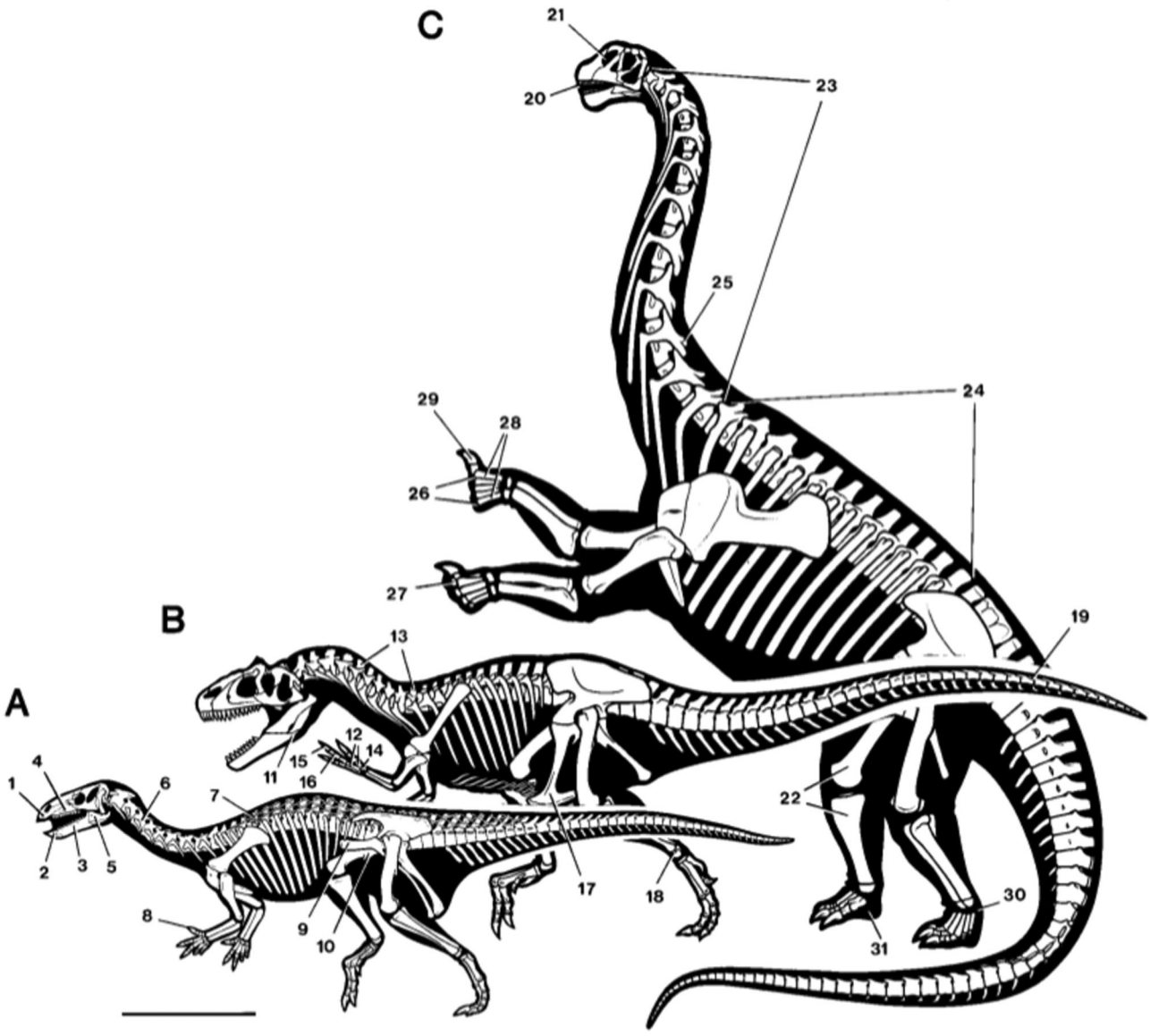

Figure 2. Three major clades of dinosaurs have long been recognized, their relationships remaining controversial: **A,** Ornithischia; **B,** Theropoda; and **C,** Sauropodomorpha; as represented by contemporaneous species (*Camptosaurus dispar*, *Allosaurus fragilis*, *Camarasaurus lentus*, respectively) from the Late Jurassic of North America. By the dawn of the Jurassic Period some ~200 Mya, these clades were already well defined morphologically. The phylogenetic dispute discussed in this paper is limited to their interrelationship and the allegiance of less derived, and often less complete, Late Triassic species. Numbers (1–31) highlight prominent clade-specific synapomorphies. Scale bar equals 1 m (modified from Sereno 1999a).

(Dembo *et al.* 2015; Dembo 2016). Revision, or modest alteration, of a dataset by subsequent authors is a common practice in the phylogenetic literature. The second pair of analyses focuses on the basal radiation of Dinosauria (Baron *et al.* 2017a; Langer *et al.* 2017; Fig. 2), which involves authors in opposing camps that reach different phylogenetic resolutions.

Both pairs of datasets capture differing or competing analyses that overlap to some degree, a very common occurrence in the literature. Datasets for the early radiation of dinosaurs, for example, date back to the mid-1980s and include some of the earliest quantitative cladistic studies incorporating fossils (Table 1). Many of these analyses have disparate character lists and terminal taxa. Stark differences in character selection and character-state scoring were noted among these relatively small datasets (Sereno 2007b). More recent datasets, which tend to include more taxa and characters, usually involve an amalgam of uniquely crafted characters and character-states. Disparate or conflicting character-state scores are buried within expansive datasets that are impossible to thoroughly compare manually. The oft-

Table 1. Sampling of quantitative morphology-based phylogenetic studies (some based on modifications of previous datasets) that have considered basal relations with Dinosauria.

No.	Analysis	No. ingroups	No. characters
1	Gauthier 1986	8	84
2	Novas 1992	5	17
3	Novas 1993	6	51
4	Sereno *et al.* 1993	12	132
5	Sereno 1999a	13	146
6	Rauhut 2003	58	224
7	Langer 2004	9	107
8	Langer & Benton 2006	10	98
9	Irmis *et. al.* 2007	26	127
10	Nesbitt *et al.* 2009	41	315
11	Nesbitt 2011	75	412
12	Sues *et al.* 2011	42	319
13	Excurra & Brusatte 2011	43	339
14	Cabreira *et al.* 2016	42	256
15	Baron *et al.* 2017a	74	457
16	Langer *et al.* 2017	83	457
17	Baron & Barrett 2017	76	457
18	Müller & Dias-da-Silva 2019	58	457
19	Marsh *et al.* 2019	45	352

repeated mantra that better results will be forthcoming with "more meticulous assessment of characters and homologies than those recently conducted" (Langer 2014, p. 1) rings hollow, when thorough comparison to the data in previous analyses is not presented. We delve into the competing dinosaur analyses in greater detail below, because they involve opposing camps that scored and re-scored the same characters, for reasons of rapid response and simultaneous publication. Use of the same characters, in turn, facilitated our automation of data comparison.

Future progress in comparative cladistics will require automation of a range of comparative procedures to handle differences in character construction and expression. Widespread adoption by systematists of a more uniform, logical character structure will greatly facilitate comparison, with character components and potentially terminal taxa drawn from ontological and taxonomic databases, respectively (Sereno 2009).

Material and methods

The implementation provided in TNT and outlined below assists the (1) comparison of two morphology-based taxon-character state matrices, (2) assessment of underlying causes for different phylogenetic results (trees), and (3) production of a combined dataset with annotated matrix cells that highlight unique and shared taxa, characters and character states that are responsible for differences in phylogenetic results. The Supplemental material for this paper (also available at http://www.lillo.org.ar/phylogeny/published/Goloboff_Sereno_2021.zip) includes comparisons among two pairs of datasets. The first comparison is between the datasets of Dembo *et al.* (2015) and Dembo (2016). The second is between the datasets of Baron *et al.* (2017a) and Langer *et al.* (2017).

Initiating dataset comparison

Subroutine in TNT. For ease of use, the method is implemented in a single TNT command, `dcomp`. In the Windows versions of TNT, this command is accessed through a single menu option (Fig. 3), with `File/MergeImportData/PairwiseComparisonCombination`. The command expects the name of an output file (where the combined dataset will be written), followed by the names of two input data files:

dcomp [options] combined-dataset newer-dataset older-dataset

The distinction between `newer` and `older` datasets facilitates taking default actions, when there are differences in the settings or scorings between two datasets. Users may want to preserve those of a particular dataset by default (e.g. accepting all `newer` scorings as correct). The `combined` dataset is automatically read with execution of the `dcomp` command. Thus, after comparing the two input matrices, TNT loads in memory the combined matrix and settings.

If the files with matrices to combine contain trees in parenthetical notation (with the `tread` command), then the `combined` dataset will also contain those trees, with the corresponding translation to new taxon numbers. These trees may have different taxon sets. For comparison the trees are (by default) pruned to their common taxon subset. Optionally, the taxa unique to each of the datasets can be placed at their most parsimonious location on trees from the other dataset. To place a taxon X that is present only in the `newer` dataset on a tree for the `older` dataset, the topology of the tree is forced (via constraints, with taxon X free of constraints), and taxon X is placed at the most parsimonious location on that tree, using character-state scores in the `newer` dataset. This can be done by checking the `Complete trees...` box (Fig. 3, in the `General Options` panel of the Windows version) or with the `++` option of the `dcomp` command.

Completing the trees may be the only way to detect the source of differences in results for two datasets, particularly when the difference is due to the addition of taxa rather than changes in matrix cells for shared taxa and characters. Note that the influence of character and matrix differences in the results is evaluated only in the case that trees are included in both datasets. For each of the datasets, the trees included should be optimal tree(s)

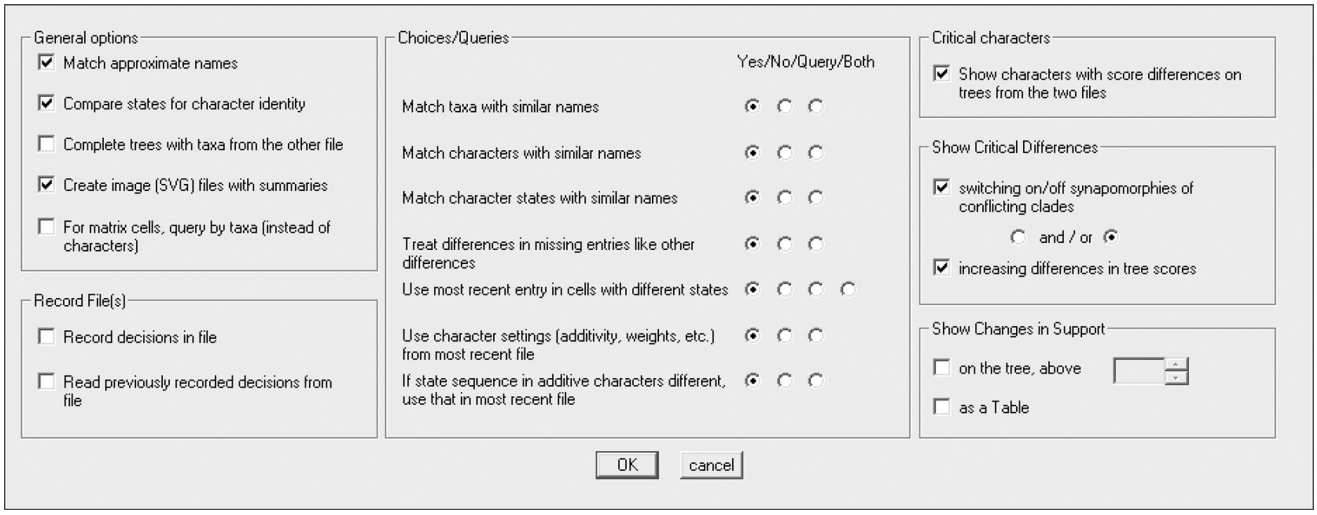

Figure 3. Windows dialog for comparing and combining datasets.

or their consensus. The user is responsible to ensure each dataset contains appropriate trees. When running the `dcomp` command, TNT does not automatically perform tree-searches for either dataset.

Via a series of options, the user decides how to establish matches between taxon and character names, as well as any decisions to resolve differences in character settings (i.e. character weights, additivities, ordering) or character-state scores (i.e. for particular matrix cells) between the two matrices, with default choices or queries for each specific decision.

Character structure. The implementation in TNT requires that all characters and character states in each dataset be named. The problem of deciding whether overlapping character conceptualizations truly describe the same character can be challenging. Variability in the expression of phenotypic characters and character states (collectively, 'character statements') cannot be 'solved' with a comparative computer program, but hopefully will emerge in future practice – as communities of systematists appreciate the need for speaking the same 'character language'. Only then will it be possible to advance beyond the current era of overlapping, yet in many ways non-comparable, morphology-based datasets.

Sereno (2007a, 2009) addresses this issue in detail, dividing character statements into two forms (neomorphic, transformational) with four components (locator, variables, variable qualifier, character state). Both character forms function as independent variables (Kluge & Farris 1969). In the future, one may envision a program for composing character statements by selecting these character components from an exhaustive or modifiable anatomical ontology. In this way, the current widespread variation in character form in morphology-based datasets could be minimized to facilitate comparison and consensus (e.g. Ramírez *et al.* 2007; Sereno 2009; Deans *et al.* 2015; Dececchi *et al.* 2015; Eliason *et al.* 2019).

Component comparison and combination

Taxon, character and character-state matching. All taxon and character names are first compared for exact matches. If each matrix contains two or more characters with identical names, TNT will first attempt to pair those which also have identical character states (state comparison can be turned off, with `dcomp-`). By default, the taxa and character names without an exact match are then compared using the Needleman–Wunsch (1970) string comparison algorithm (this can be changed by unchecking the `Match approximate names` box in the Windows dialog, Fig. 3, or with the ! option of the `dcomp` command). The Needleman–Wunsch algorithm (for alignment) counts the minimum number N of letters that must be changed or added/deleted to convert one name into another. If L is the length of the shortest name, then the similarity between names can be defined as $S = 1 - (N/L)$ (as in GB-to-TNT; Goloboff & Catalano 2012). For comparing matrices, the minimum similarity M for two names to be considered as possible synonyms is by default 0.75 (i.e. one letter difference every four), but this can be changed prior to running `dcomp` (with the command `help! M`). For every pair of non-identical names with similarity $S \geq M$, the action can be chosen to be either an automatic synonymization, or a query to the user for confirmation (see below, under Settings for matching and combination).

TNT reports the taxa, characters, and states that are not perfect matches (i.e. those which have been matched despite some minor differences in spelling, or despite

having different states, in the case of characters). The numbers of taxa and characters unique to one of the matrices are tallied and reported at the end of the comparison. TNT also reports the best possible match for unmatched taxa and characters, and this helps identifying many cases of synonymies. As the synonymy between taxa or characters can be pre-specified (see below, under Settings for matching and combination), this allows the comparison to be re-run.

As a test example, we used matrices for hominin phylogeny from Dembo *et al*. (2015) and Dembo (2016) (see Supplemental material). The first matrix (Dembo *et al*. 2015) emerged from doctoral research that was later modified (taxon names, characters, character-state scores) when formally submitted (Dembo 2016). Comparison of the two datasets by TNT indicates unmatched characters in each dataset: 7 of the 380 characters in Dembo *et al*. (2015) and 18 of the 391 characters in Dembo (2016). TNT provides a list of seven best possible matches (with `newer` name listed first, possible synonym in `older` dataset second): Temporal crest on supraorbital torus (45) *vs* Zygomatic arch relative to inferior orbital margin (97); Maxillary I2/C diastema (251) *vs* Presence of Maxillary I2/C diastema (247); Mid-trigonid crest in mandibular M1 and/or M2 (310) *vs* Presence of mid-trigonid crest in mandibular M1 and/or M2 (301); Cusp 5 in mandibular M1 and/or M2 (311) *vs* Absence of cusp 5 in mandibular M1 and/or M2 (302); Metaconid development on mandibular deciduous m1 (323) *vs* Presence of metaconid of mandibular deciduous m1 (313); Subalveolar fossa (333) *vs* Alveolar clivus shape (62); Anterior marginal tubercle (345) *vs* Anterior marginal tubercle position (334).

Human inspection of those seven TNT-identified possible pairings highlights two that seem spurious with no real correspondence between matrices (45/97 and 333/62). For these two, TNT reports the lowest similarity (0.380 and 0.429, respectively). The other five pairings (251/247, 310/301, 311/302, 323/313, 345/334) can be pre-specified in a file (Supplemental material, final results). The utility of the matching algorithms implemented in TNT is suggested by the higher similarities (0.604 to 0.789) shown between the scorings for these character matches.

For taxon disparities between matrices, only one plausible pairing was made by TNT on the basis of taxon names alone (i.e. ‘H sapiens’ a suggested synonym of ‘Homo sapiens’). ‘Asian H erectus’ is suggested as a synonym of ‘H erectus’, which seems to correspond, although this could not be discovered from taxon names alone given the presence in the 2016 dataset of ‘African H erectus’ and ‘Georgian H erectus’. Taxon synonymies in this case are better established on the basis of character scorings (see below, under Tallies).

A Character Similarity Index (CSI) measures the proportion of shared relevant character data between two hypotheses (Sereno 2009). In the final matching, 2 of 380 characters in the matrix of Dembo *et al.* (2015) and 13 of 391 characters in the matrix of Dembo (2016) are unique to each matrix. The total number of characters between the matrices sums to 396, of which 378 are shared (possible to match). The CSI is $378/393 = 0.962$, indicating an overlap (similarity) of approximately 96% of total character data. A CSI this high (approaching identity, 1.0) usually occurs only when comparing versions of studies by the same or similar authors. Comparison between independent phylogenetic studies of the same group results in considerably lower CSI values, sometimes as low as 0.10, or 10% character overlap (Sereno 2007b; Sereno & Brusatte 2009).

Character settings. The differences in settings for the characters are tallied as well, checking differences in additivity and weight. Step-matrix characters are not handled by the present implementation (this would also require comparing the transformation cost between equivalent states in both datasets).

TNT expects states in additive characters to be at similar distances from each other in both matrices. For example, transforming between states `absent/present` in a matrix with only those two states will cost a single step, but it will cost two steps in a matrix with states `absent/rudimentary/present` (this, again, will use the automatic choice of either the `newer` or `older` setting, or elicit a query from TNT, and will be recorded for final tallies as a difference in state-ordering).

Cell scorings. After establishing matches between taxon, character, and character-state names, TNT makes comparisons between cells in the two matrices. In the case of discrete characters, the program automatically accommodates differences in state numbering. For example, states 0 and 1 may represent `absent` and `present`, respectively, in one dataset with an inverted assignment in the other matrix. In that case, TNT will expect a taxon scored as 0 for that character in the first matrix to be scored as 1 in the second. Otherwise, a decision on how to resolve such scoring conflicts must be made by the user (e.g. automatically retaining by default the `newer` scoring, or else the `older`, the combination of both, or querying user for a choice).

A Character-State Similarity Index (CSSI) measures the similarity between corresponding character states of the same characters in identical or comparable taxa (Sereno 2009). This index most easily applies to discrete

character data, where character-state conflict (e.g. 0 *vs* 1) and disparity (e.g. 0 *vs* ?) are more easily quantified. After matching attempts, a penalty of 1.0 and 0.5 are summed for each character-state conflict and disparity, respectively. That sum is subtracted from the total number of character states and then divided by the total number of character states to indicate overall character-state similarity. These hominin datasets, which are composed exclusively of discrete character data, share 378 characters including a total of 9450 character states. Their comparison reveals 510 character-state conflicts and 432 character-state disparities. The CSSI is 0.916, indicating an overlap (similarity) of approximately 91% of character-state assignments for the same characters. As with CSI, comparison of character-state assignments between independent phylogenetic studies typically results in lower CSI values (Sereno & Brusatte 2009).

Missing entries. By default, missing entries are treated as any other scoring difference. Optionally, if a cell is scored as missing in one dataset and as an observed entry in the other, it is possible to make TNT automatically use the observed entry (see below, under Settings for matching and combination, for details on how to set these options). TNT has routines for inapplicable characters, but these use a recoding method (Goloboff *et al.* 2021) based on the character hierarchy determined from the character names (see `help smatrix`). Thus, TNT does not distinguish among inapplicable and missing entries on input.

Continuous characters. In the case of continuous characters, the first check involves comparing the ranges (= scaling) of the character in both datasets, because the same observations may appear different just by virtue of a different rescaling in both datasets. For this, the two taxa present in both datasets that are most distant from each other in the first matrix are chosen, and the character is rescaled in both datasets so that those two taxa have the same distance in both datasets. The entries for all shared taxa are then compared; if the number of observed entries that become identical after rescaling outnumbers the number of entries that are different by more than 5:1, then the character is considered to be the 'same', and the user is queried for retaining the scores in either the character from the first dataset, or from the second. Otherwise, the differences in both characters are considered to be irreconcilable by means of rescaling, and the user is just warned that the entries differing in the two datasets cannot be meaningfully combined.

Landmark characters. For landmark-based characters, differences in scoring for the same taxon in two datasets may arise from differences in alignment (or sizing). The distance between corresponding pairs of taxa is calculated first as the sum of lineal distances between each pair of landmarks (Catalano *et al.* 2010; Palci & Lee 2018). If this distance is below 10^{-4}, the configuration is considered to be the same. Otherwise, the two configurations (i.e. for the same taxon in each datasets) are aligned (again, minimizing the sum of lineal distances between landmarks). If the distance is now below 10^{-4}, a warning message advising realignment of the landmark configuration is issued; if the distance after realignment is greater than 10^{-4}, landmark differences are considered irreconcilable and added to the tally of matrix scoring differences.

Combination of datasets. After all taxa, characters, and character states have been matched (with possible user input), a report specifies what is unique to each matrix. The `combined` matrix will contain all the taxa, with the taxa from the first matrix listed first in the same order as they appear in the `newer` dataset (regardless of whether they are shared with the `older` matrix), followed by the taxa unique to the `older` matrix. The same is done with characters when no continuous characters are present in the matrices. In TNT continuous characters cannot be preceded by any discrete or landmark character in the matrix, and thus (when continuous characters are present), the `combined` matrix will contain first the continuous characters from the `newer` matrix, then the continuous characters unique to the `older`, then the discrete or landmark characters from the `newer` matrix, and finally the discrete or landmark characters unique to the `older` matrix. All the character and state names are exported (with the numbering correspondence) to the `combined` dataset.

Graphic representation. To facilitate visualization, the differences between the datasets are plotted in the form of coloured diagrams in image files (SVG, scalable vector graphics). The file `matrix_differences.svg` contains a graphic depiction of the matrices, with red taxa, characters, or character states derived from the `newer` matrix, blue characters or character states derived from the `older` matrix, and black for entities that are shared. In a comparison of two matrices (Fig. 4), Taxon E and character 3 in the `combined` matrix (in red) are exclusive to the `newer` dataset, while taxon B and character 4 (in blue) are exclusive to the `older`. For the first character ('feet'), taxon A is scored as 1 (red) in the `newer` dataset ('big'), and as 0 (blue) in the `older` ('small'); this conflict is indicated by the colors of the corresponding cell.

Tallies. TNT reports how similar compared matrices are regarding the proportion of shared (matched) characters (CSI) and the proportion of shared (matched) character states (CSSI, for discrete character data) among shared

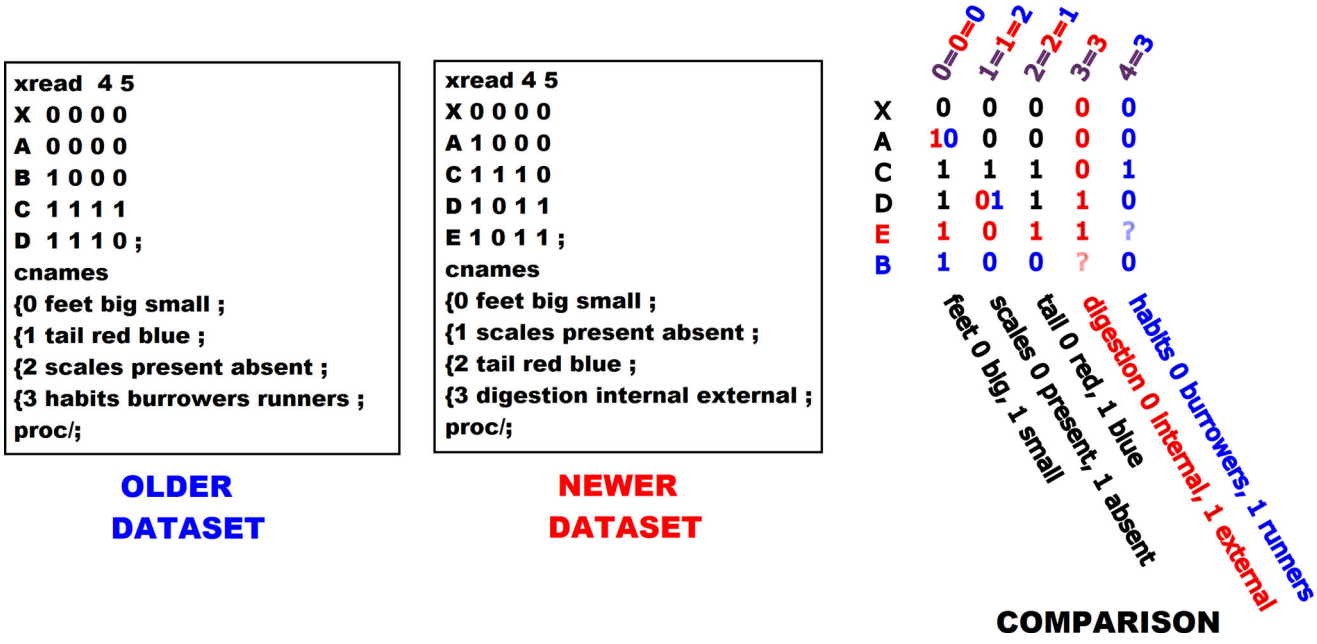

Figure 4. Two datasets showing the colour-coded comparison made by the program (saved to file `matrix_differences.svg`). Characters, taxa and character-state scores unique to one of the matrices are colour-coded in the combined dataset (`older` in blue, `newer` in red). The correspondence between character numbers in the original matrices and the comparative matrix is shown above the latter, with overlapping or new characters and character states shown below.

(matched) characters. Also reported are the number of characters with different settings (or character-state ordering, in the case of additive characters). All these numbers can also be accessed and manipulated via scripting expressions (e.g. Goloboff & Morales 2020).

Importantly, TNT also reports the number of cells with 'critical' character-state differences, i.e. individual character states that are most likely causing differences in results between datasets (see below, Comparison of results). Note that the number of cells with different scorings may well depend on decisions of how to synonymize taxa, characters or character states.

In addition to global tallies, it is also possible (using > as option for `dcomp`) to output number of differences in scorings on a per character basis, and a per taxon basis. This can help pinpoint characters or taxa which are harder to score (e.g. poorly preserved or rare materials). In addition to summaries for shared characters and taxa, this option also reports, in the case of unshared taxa, the number of different scorings (N.B. this count includes cases of a positive score *vs* a missing entry) for every possible pairing of unmatched taxa. Obviously, those taxa with closest scorings are candidates for synonymy. In the case of the Dembo *et al.* (2015) and Dembo (2016) datasets, this criterion points to the synonymy of four taxon names: 'African_H_erectus' of Dembo (2016) and 'H_ergaster' of Dembo *et al.* (2015) are most similar (note that *H. ergaster* is sometimes considered as a synonym of *H. erectus*); 'Asian_H_erectus' of Dembo (2016) and 'H_erectus' of Dembo *et al.* (2015) are most similar; and 'Georgian_H_erectus' of Dembo (2016) and 'small_bodied_Dmanisi' of Dembo *et al.* (2015) are most similar (note that Dmanisi is an archaeological site in Georgia). The fourth synonymy suggested by this comparison is between 'Homo_sapiens' in Dembo (2016) and 'H_sapiens' of Dembo *et al.* (2015), which had been suggested on the basis of names alone (see above). These four synonymies were pre-specified for the final comparison (see Supplemental material).

Settings for matching and combination. In the Windows GUI version, these options are set in the central panel of the dialog for comparing and combining datasets (Fig. 3, under `Choices/Queries`). In the command-driven versions, the name of the `combined` dataset can be optionally preceded by the default choice to be made, in specific comparisons, within parentheses. The default setting is to preserve all the options of the `newer` dataset, without queries. The series of optional actions to take is indicated by a string of letters, within parentheses. The letters *t* (taxa), *c* (characters), *s* (states), *d* (data, scoring of individual cells), *m* (missing entries), *o* (options, character settings of additivity and weights), *a* (ordering of states in additive characters), *q* (quantitative or continuous characters), and *l* (landmark characters) indicate the comparison to which the subsequent action is to be applied. The string of letters is followed

by the action to take, with *y* (yes, match, or use condition from `newer` dataset), *n* (the opposite), or ? (query). The action *i* is used only for *m*, missing entries (ignore, treat like the rest of entries), with automatic replacement of missing by observed entries indicated with *y*. The action *b* is used only for *d*, cell scoring differences, to merge the differing observed states into a superset.

An alternative way to force synonymization choices for taxa or characters is by using a file with a list of the taxa or characters. For this, the letter *f* inside the parentheses, followed by the name of the file, can be used. The file must contain the string '`*taxa`' followed by all the pairs of taxa (two numbers, corresponding to the numbering in the `newer` dataset, and the numbering in the `older` dataset; separate the numbers with ! to force non-synonymization instead of synonymization), and the same for '`*characters`'.

Saving (and reusing) choices. The comparison of two datasets with numerous differences may involve many decisions on the part of the user if the automatic choice is not used for all differences. For example, a pair of taxon names with similar but not identical spelling can be synonymized, while another pair is not. To avoid having to go again through the list of possible decisions, it is possible to save a list of the decisions made to a file, subsequently using the decisions stored in file for repeating automatically the same choices. In the Windows GUI version, this is set in the lower left panel (under `Record File(s)`, Fig. 3). In command-driven versions, the symbol > inside the parentheses used to delimit the settings for queries instructs TNT to store decisions in a file, and the < symbol tells TNT to read decisions from a file previously created with the > option. Both > and < can be used simultaneously within the parentheses (obviously, reading from one file and writing to another).

If the user presses the `Escape` key when faced with a given query, the parsing is interrupted. If the decisions are being saved to a file, subsequent reading of decisions from that file will start asking questions (e.g. about taxon or character identities) from the same point at which the comparisons were last saved. Thus matrices may be compared over time in several sessions.

To read decisions from a previously created file, it is necessary that the settings (specified inside the parentheses, or from the `Choices/Queries` panel of the Dialog, Fig. 3) be identical. Identical settings are required because whether two matrix cells have a difference in scoring may depend on whether the corresponding taxa and characters have been considered synonymous. If a file (with the *f* option, see above, under Settings for matching and combination) forces taxon or character correspondences, TNT does not check for identical settings. The user must make sure that the same file for predefined lists of correspondences is used, when reading decisions from a previously created file.

Requirements and limitations

Requirements of the data files. The files with the matrices must be in the `xread` format (i.e. the native data-input format for TNT). The two matrices must share the same first taxon. The taxa other than the first in each dataset can be in any sequence.

The comparison of results for each dataset requires that trees be included in parenthetical notation, after the matrix, in each dataset. The trees must be rooted on the (shared) first taxon, and contain all of the taxa included in the corresponding dataset. Either a single tree (perhaps a consensus tree, with polytomies), or multiple trees can be included in each dataset; in the latter case, the comparisons automatically calculate the consensus of the trees in each dataset (as well as character fits or synapomorphies in each of the trees, as in the common mapping options of TNT, see Goloboff *et al.* 2008, p. 781).

All characters and their states must be named. Because taxon and character names can be long strings, it is advisable to make sure that the tree labels can be long enough (which, in turn, also requires that enough RAM is allocated to TNT before comparing and combining the data). By default, prior to executing the comparison, TNT automatically sets the tree labels to have a minimum length of up to 2048 characters, and allocates at least 400 MB of RAM. To make TNT allocate longer labels or more RAM prior to comparing and combining the datasets, use the corresponding commands, `taxname + L` and `mxram MB`, where `L` is the label length and `MB` RAM in MegaBytes.

One-to-one taxon and character correspondence. The current implementation allows only a one-to-one correspondence between taxa and characters from each dataset. For correspondence between taxa, Sereno (2009) discussed the circumstance where a specimen, species, genus or higher taxon in one dataset corresponds with several species or terminal taxa in another dataset. One solution is to substitute multiple terminals with the reconstructed ancestral states (HTUs or hypothetical taxonomic units) of the corresponding group (ideally, the down-pass states), and name the equivalent replacement terminal taxon accordingly (Sereno 2009). TNT will not do this automatically. The user can issue commands (e.g. with the `xread:` command, see online help for TNT) to save the down-pass states corresponding to some internal node, but then the matrix will have

to be manually edited so that the names of the HTUs correspond to those of the taxon with which it is to be matched. The same approach could be handled with scripts, although requiring some programming effort.

An alternative strategy for datasets with different taxon sets is to prune the trees to their common taxon subsets (Sereno 2009). This is the default setting of TNT. Pruning the trees to their common taxon subsets, however, may be insufficient to identify all sources for differing results. Different results may arise due to the addition of taxa with new character combinations, rather than only the addition of new characters or altered character-state scores (see examples below). Thus, TNT also allows the option to complete trees from each of the datasets by adding the taxa present only in the other dataset.

In the case of characters, a single character in one dataset may be represented by more than one in another dataset. The simplest occasion for this kind of character mismatch is the consideration of 'absence' of a condition. Some authors include this as an additional state of a multistate character, whereas others prefer scoring as a separate neomorphic (presence-absence) character (Sereno 2007a). The current implementation has no way to deal with that situation; a character in one of the datasets cannot be made to correspond to more than a single character in another. The current means of handling such situations is to manually edit the matrices, fusing or splitting characters to obtain one-to-one correspondence, and compare the adjusted datasets.

Tree searches, optimality criteria. The addition of taxa missing from each dataset, and all subsequent comparisons of character fit on the different trees, are computed with either prior weights or standard implied weights (if implied weighting was turned *on* when starting the dataset comparison). Note that the same criterion is used for placing taxa missing from a dataset into tree(s). Those two criteria are most commonly employed for analysis of morphology-based datasets. Neither tree completion nor scoring can be computed with user-defined weighting functions (which must be redefined after reading a new dataset), auto-weighted optimization (Goloboff 1997), or extended implied weights (Goloboff 2013). If results from one or both datasets have been obtained under any of these optimality criteria or with different weighting strategies (e.g. implied weights for one dataset but not the other, or values of k for both datasets), the results obtained by comparing tree scores are only an approximation.

Comparison of results

Besides summary index comparisons between matrices (character and character-state similarity indices, CSI, CSSI), results may be compared by their effect on the trees they support in two fundamental ways: first is a comparison of synapomorphies and character fits on the optimal topologies of each dataset; second is a comparison of the degree of support for clades when one matrix is transformed into the other. The first comparison (comparisons on optimal topologies for each dataset) is shown in graphic form in the files `newer-not-older.svg` and `older-not-newer.svg`, as well as the file (already mentioned above) `matrix_differences-svg`. The second aspect is saved to a file called `bremer_differences.svg` (also, optionally, the normal text output of the program).

Synapomorphies. The differences in results for both datasets are calculated by TNT, finding the groups that are present in (all) the tree(s) for the `newer` dataset but not present in (all) the tree(s) for the `older`, and vice versa. The synapomorphies for each of those groups are saved to two files with tree-diagrams, `newer-not-older.svg`, and `older-not-newer.svg` (the conflicting groups cannot be simultaneously displayed on a single tree). We present an arbitrarily modified part of a published matrix as an example (Goloboff 1993; Fig. 5). The bottom group with a red legend (character 14, black; character 22, red) is a group found only for the `older` dataset, with character 14 as a synapomorphy in both `older` and `newer` scoring (hence black), and character 22 as a synapomorphy only as the character is scored in the `newer` matrix (hence red). By so colour-coding synapomorphies for the conflicting clades between analyses, visual inspection allows a general perspective on the characters supporting alternative results. Note that some of the characters in one dataset (say, `older`) may still be favourable to the results of the other dataset (say, `newer`), with the colour-coding method facilitating their identification. As an example, the clade (in `newer` results) of BARYCHELIDAE plus its sister group (Fig. 5) has character 67 as a synapomorphy only as scored in the `older` matrix (blue).

Characters. The synapomorphies of the conflicting groups provide clues to the causes underlying differing results. Strictly speaking, a character favours a `newer` tree over an `older` tree only when the character has fewer steps in the `newer` than in the `older` tree (and vice versa). Length differences for characters on trees narrows the set of characters responsible for differing results. If multiple trees are present for the datasets, TNT will calculate the maximum and minimum length of the character in each set of trees.

In another example dataset (Fig. 6), two matrices for the same sets of taxa and characters have five character-

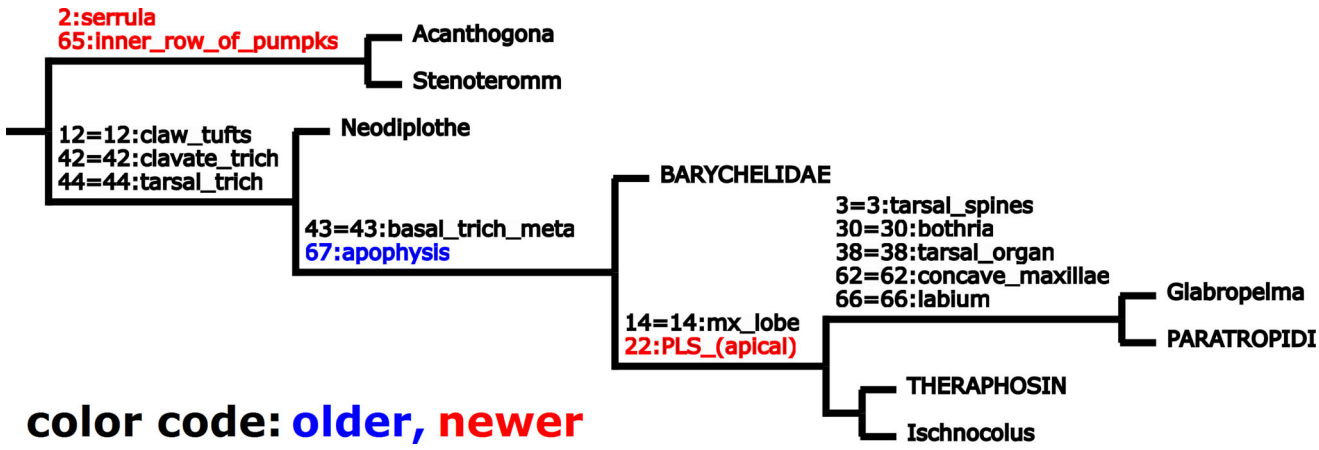

Figure 5. Diagram showing synapomorphies for groups present in the trees in only one of the datasets. Colour codes indicate whether the character is synapomorphic in one or the other character-state scores.

state score differences (in characters 1, 2 and 5). A `solid-line frame` surrounds the character with shorter length in the tree(s) for `newer`, whereas a `dashed-line frame` surrounds the character with shorter length in the tree(s) for `older`. Frame colour indicates whether the character favours one set of trees when scored as in `older` (blue) or in `newer` (red) datasets. It is possible, of course, that the changes in scoring made to a character in the `newer` dataset favour the `older` results rather than the `newer` results; this would be indicated by a blue solid-line frame. Likewise, an `older` scoring may favour `newer` trees when `newer` scoring does not; this would be indicated by a red dashed-line frame. When both `older` and `newer` scoring favour a given result, this is indicated with a black frame.

In Figure 6 (top), thus, the new scoring of characters 'one' and 'two' favours the `newer` results, whereas only the `older` scoring of character 'two' favoured the `older` result (the `older` scoring of character 'one' was indifferent to the differences in results). Character 'three' favours the `newer` results in both datasets, whereas character 'four' favours the `older` results in both datasets. Characters 'zero' and 'five' favour neither `newer` nor `older` results.

In the example, a similar picture of the influences of each character would have been obtained by examining the synapomorphies supporting the groups in conflict (Fig. 6, bottom). The group CD, supported by the `older` dataset, has character 'two' as a synapomorphy only in the `older` scoring, with character 'four' supporting the group in both datasets; the group BC, supported by the `newer` dataset, has a synapomorphy in characters 'one' and 'two' only in the `newer` scoring, with character 'three' supporting the group in both datasets. Note, however, that it is possible for a character that maps as a synapomorphy for a group present only in one tree to still favour the other tree (if the character has fewer steps on that other tree). Therefore, the information provided by the frames is not always identical and sometimes more precise than that provided by synapomorphy schemes.

Critical cells. The identification of 'critical' character-state cells provides an even more precise identification of the data supporting differing results than status of individual characters. Character 'two' in the hypothetical matrix comparison shows three changed character-state cells in the `newer` scoring (Fig. 6). It is entirely possible that some of those changes are more important than others in determining the differing results.

TNT uses two methods to identify the individual cells that are responsible for differing results, both based on evaluating the effect of changing different combinations of individual cells for each character. The first method identifies 'critical' cells as those that support the character as synapomorphy only with the `newer` scoring and lose that status with the `older` scoring. The second method identifies 'critical' cells in `newer` scoring as those that increase length difference ($\texttt{Length}_{\texttt{oldtree}} - \texttt{Length}_{\texttt{newtree}}$) when changed back to the `older` scoring (i.e. character states that favour `newer` results more strongly). The cells are changed in both methods individually and in combination. Given that the number of possible cell combinations may become too large for a given character, selecting cells for reassignment according to specific rules reduces the computation time.

Method 1 checks vanishing synapomorphies for a given conflicting group, reducing the number of cell changes to combine by examining the states assigned around the group. The character is first mapped for both

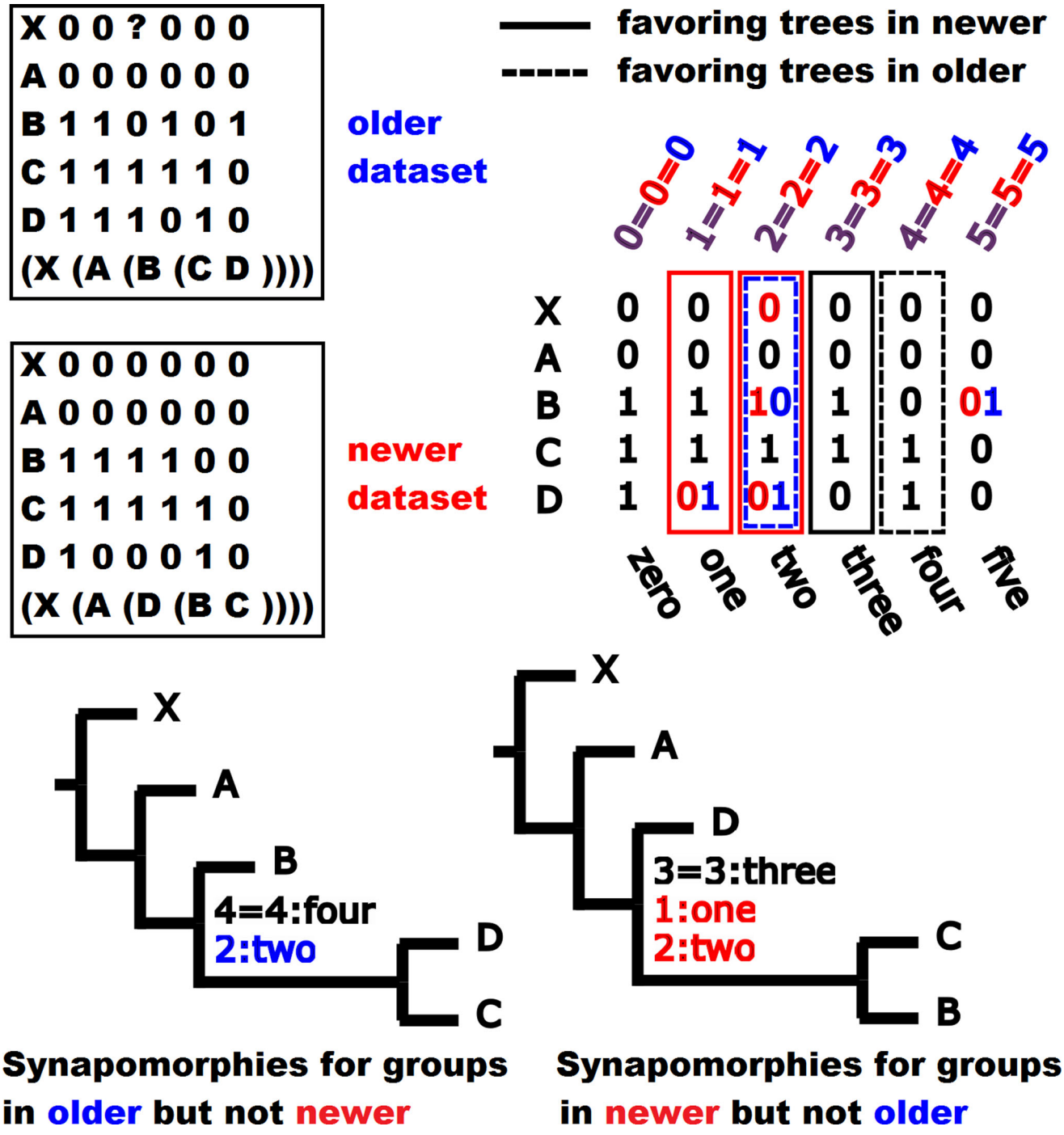

Figure 6. Example comparison between two datasets showing how characters are highlighted that favour the results of `older` or `newer` datasets. Rectangular frames highlight characters responsible for length differences in `older` and `newer` trees. Frame colour indicates the scoring that favours tree(s) with the `older` scoring (blue) or with the `newer` scoring (red). Black frames indicating support in both scorings. The frame lines (dashed, solid) indicate support for older tree(s) (dashed) and newer tree(s) (solid). In this example, the assessment of characters that generate different results after comparing character steps in trees from `older` and `newer` datasets is similar to the assessment obtained by listing synapomorphies for clades that are not present in both trees.

the `older` and the `newer` scoring, on the tree(s) for the `newer` dataset, and each internal node i of the tree is assigned the set of states ($S_{s,i}$) shared in the `newer` scoring by all their descendant terminals (an empty set if no states shared by all terminals), as well as the states of the up-pass optimization ($S_{o,i}$ and $S_{n,i}$, for `older`

and newer scoring). A character may become a synapomorphy for the group in the newer scoring (i.e. $S_{o,node} \cap S_{o,anc} \neq \emptyset$ and $S_{n,node} \cap S_{n,anc} = \emptyset$) either because the ancestral state set changed ($S_{o,anc} \neq S_{n,anc}$) or because the descendant state changed ($S_{o,node} \neq S_{n,node}$), or both. If the ancestral state changed in the newer scoring, the initial list of nodes to combine will contain each successive sister group g of the ancestor (for as long as $S_{o,i}$ in the successive ancestors i of the node differs from $S_{n,i}$) for which $S_{s,g} \neq S_{o,g}$. If the node states changed in the newer scoring, and there is a state present in older sets for node and ancestor which is not in the newer state for the node (transforming the character into a synapomorphy), the initial list will also contain the node itself. The final list is formed by adding each successive immediate descendant node d, of the nodes in the initial list, as long as the same condition (i.e. $S_{o,d} \neq S_{n,d}$ and $S_{s,d} = \emptyset$) is fulfilled; terminal nodes and internal nodes for which $S_{s,d} \neq \emptyset$ are added to the final list of nodes to combine.

In the case of internal nodes for which $S_{s,d} \neq \emptyset$, the changes back to older scoring are effected simultaneously for all the terminals d descended from the node (instead of tried in separate individual combinations). This greatly reduces the number of cells differing in both older and newer scoring that have to be combined, restricting it to the surrounding nodes likely to have had an actual influence on the character becoming a synapomorphy in newer scoring. For every combination of nodes in the final list of nodes to combine, the corresponding matrix cells are changed to the state in the older scoring, and the character is reoptimized. If the character thus changed is not a synapomorphy, all the changed cells are added to the list of critical cells. The changed cells are then set again to the states in the newer scoring, and the next combination is tried. The cell combinations are tried in an orderly fashion, first trying all single cell combinations, then all 2-cell combination, all the way to the maximum possible n-cell combinations. When the character stops acting as a synapomorphy with a combination of a given number of cells, the combinations of larger numbers of cells are not tried (although the rest of combinations of the same number of cells is still checked).

In the case of multiple trees for each dataset, TNT will consider those synapomorphies that are present in every one of the trees in the newer datafile when mapped with the newer scoring, and absent in at least one of the newer trees when mapped with the older scoring. Thus, the process just described is skipped for those trees of the newer datafile where the scoring as in older still has the character as a synapomorphy for the group (for there is no way, on that tree, to turn off the synapomorphy by switching some cell scorings back to the older state). Note that this process is done only in one direction (i.e. checking only the characters which, in the newer results but not the older, are synapomorphies for the groups present in the newer results but not the older), thus identifying the cells that when changed from the older to the newer scoring contribute to producing the newer result. There may be some cells that, when changed from the older to the newer scoring, contribute to producing the older result; those are not identified.

Once this process is completed, the scorings and trees are inverted for older and newer, and the process is repeated. This second step identifies the cells that, when changed from the scoring in newer back to older, contribute to producing the results in older. Again, this process will not identify the cells that, when changed from the newer back to the older scoring, would contribute towards producing the results in newer.

Method 2 finds critical cells by comparing tree lengths, reducing the number of combinations to try by comparing the older and newer mappings of the character. The strategy here uses a list of the branches in the newer tree along which there are changes under the older scoring but not under the newer. A change in a character constitutes a synapomorphy, but to take into account ambiguity in optimization, the list contains those nodes for which no possible most-parsimonious reconstruction of the newer scoring has a change along the branch and some most-parsimonious reconstruction of the older scoring does (these branches are identified by comparing the states of the first and second pass of the optimization, checking whether the state-set of the ancestor of the node contains any states not present in the preliminary state set of the node). For each combination of nodes in this list, the scoring in older is changed to the scoring in newer; in the case of internal nodes, successive descendants (as long as the final state set in older and newer scoring differ, i.e. $S_{o,d} \neq S_{n,d}$) are changed as well. The changed character (with a scoring intermediate between older and newer) is now optimized, and its length checked on all the trees.

If the length difference for the modified character between older and newer trees ($\mathtt{Length}_{\mathtt{mod,oldtree}} - \mathtt{Length}_{\mathtt{mod,newtree}}$) is larger than the length difference between trees for the older scoring, the cells changed are added to the list of critical cells. As in the case of vanishing synapomorphies, the cell combinations are tried in increasing numbers of changed cells, skipping combinations of larger numbers of cells once a combination that improves the length difference is found. In the implementation in TNT, given that combinations of many

Figure 7. Critical cells (those responsible for different results) are circled for the same datasets shown in Figure 6. As in frames for entire characters (see Fig. 6), a red solid-line circle indicates a cell that (with `newer` scoring) contributes to the preferred results of the `newer` dataset. A blue dashed-line circle indicates a cell that (with `older` scoring) contributes to the preferred results of the `older` dataset. The cells that favour the `older` results with the `newer` scoring, or the `newer` results with the `older` scoring, are not shown. Note that some cells with different character states in opposing datasets (e.g. character 'five' in taxon B) do not generate any difference in results (i.e. they have the same number of steps `older` and `newer` trees and, thus, do not generate synapomorphies supporting conflicting results).

elements may be very numerous, when there are 20 or more nodes with synapomorphies in `older` but not in `newer`, the maximum number of nodes combined together for changing is 12 (125,970 combinations). As in the previous method, this identifies the differing cells that, in the `newer` scoring, favour the results of the `newer` dataset (it does not identify the differing cells that, in the `newer` scoring, favour the results of the `older` dataset). Once the process described above is complete, it is repeated by switching `older` and `newer`, to find the differing cells that, in the `older` scoring, favour the results of the `older` dataset.

Many critical cells identified by Methods 1 and 2 are the same. By default both methods in TNT identify cells as critical with a user option to report either one or both. Extending the matrix example used in Figure 6, critical states are circled using the same colour coding scheme as we employed for character frames (Fig. 7). Both methods properly identify three of the five differing character-state scores as providing critical support for differing results. Two of the altered cells are not critical and do not favour either tree. For example, character 'five' for taxon B is scored like other taxa (state 0) in the `newer` dataset, eliminating an autapomorphy in the `older` matrix. This change reduces the length of both `newer` and `older` trees, and thus favours neither tree.

Lists of wildcards. The comparisons of tree topologies calculate the consensus of the trees in the datasets, including (by default) all taxa in each of the dataset. Some analyses may have wildcards, identified with specific processes. If these wildcards are included in the consensus trees, they may well obscure comparisons. For that reason, it is possible to provide TNT with a list of taxa to exclude from consensus comparisons. This list should be in the form of a (simple-text) list of taxon names, included in a file. The name of the file is given to the `dcomp` command, after the name of the second dataset to compare. This option is only available via commands. This option was used in the final comparison between the datasets of Dembo *et al.* (2015) and Dembo (2016), as three taxa ('H_floresiensis', 'African_H_erectus' and 'S_tchadensis') had very different positions in the possible trees for the two datasets (see Supplemental material).

Bremer supports. Besides cell scoring differences linked to alternative topologies for each dataset, degree of support for various groups may also differ. In this case, for simplicity, only characters are considered (i.e. without distinguishing cell changes within characters). Logging cell changes across branchpoint support would require a table showing every cell in the character support for every group, which would be challenging to obtain and not obviously effective. The effect of adding/removing a taxon to/from the `older` dataset is also shown (preceded with an asterisk; see Context of scoring changes, and bremer_differences.svg files in Supplemental material).

Calculation and display. Bremer support is calculated by quick approximation via TBR branch swapping (Goloboff & Farris 2001), which records the cost of moves that violate the monophyly of each group. The value of sub-optimal rearrangements is automatically determined by finding the minimum value of suboptimality to collapse all groups in the `newer` set of trees and characters. This is an approximation only, as a larger suboptimal value may be required to precisely calculate support values for trees in `newer` using `older` characters. The supports are calculated as combined Bremer supports (Goloboff 2014), or $CBS = (1 - 0.36R)^{1/A}$, where R is 'relative' Bremer support (Goloboff & Farris 2001) and A is 'absolute' Bremer support (Bremer 1994). The resulting values of support vary between 0 and 1. In simple cases,

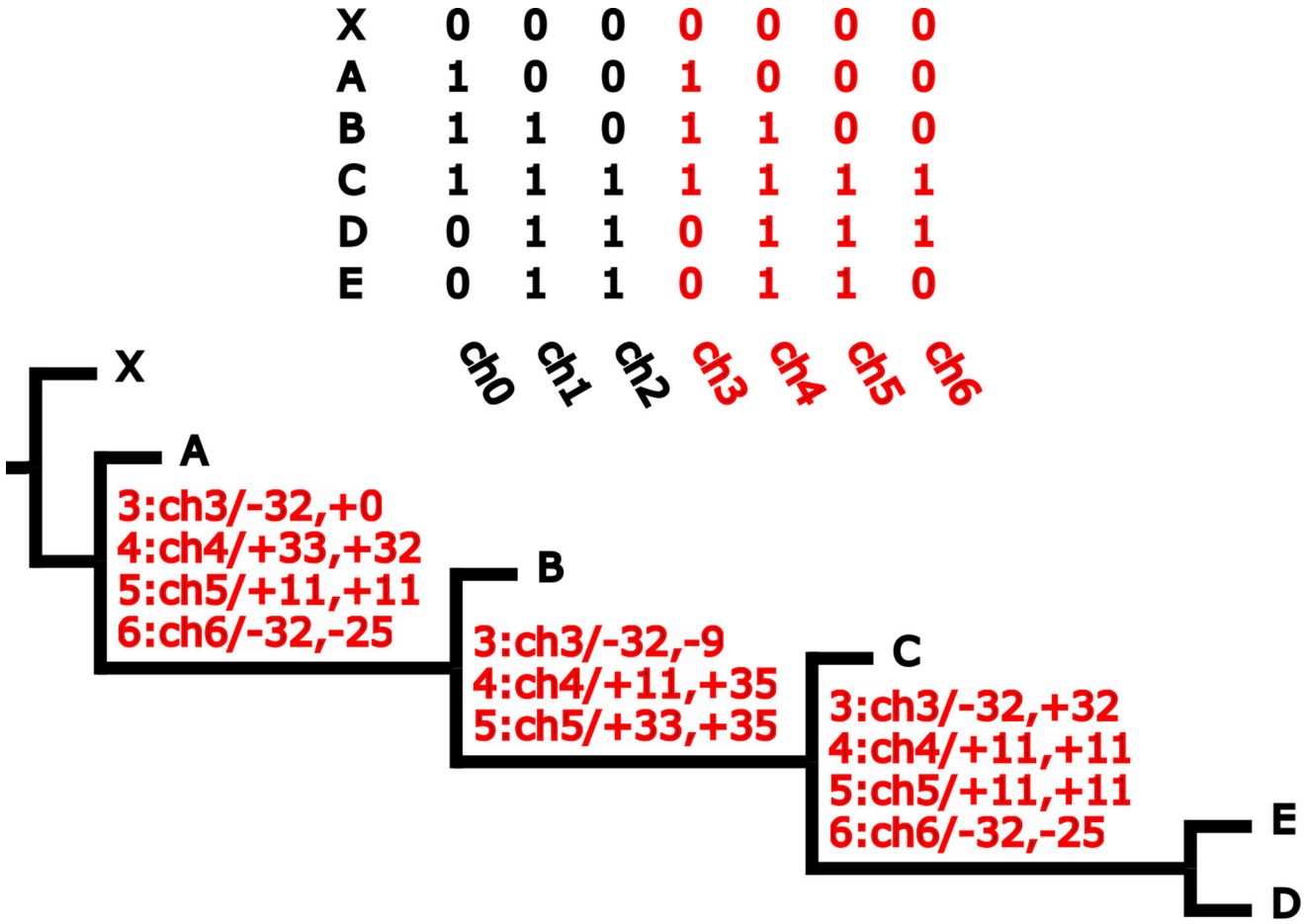

Figure 8. Two matrices compared wherein the first contains three characters (*ch0–ch2, black*) and the second includes an additional four characters (*ch3–ch6,* red). Difference in clade support are shown after adding (first and last) each of the new characters above the tree branches. Adding character *ch3* first *decreases* support for clade ED, whereas adding it last *increases* support.

the results not only are similar to resampling (bootstrapping or jackknifing), but are better than either relative or absolute Bremer support and are obtained faster (via TBR approximation) than resampling. In the case of implied weights (where A can be smaller than unity), the numerator of the exponent considers the cost (h_o) of adding the first step of homoplasy under the given concavity value when CBS = $(1 - 0.36R)^{0.25}h_o/A$.

Using combined Bremer supports, support differences are recorded for each group with the character scored as in `older` and `newer` datasets (with corresponding sign indicating decreasing or increasing support). A tree-diagram of the consensus of newer trees (saved to `bremer_differences.svg`) displays the differences in support above a certain threshold (by default, 0.05) resulting from changing the scoring for each character from `older` to `newer`. The majority of changes has no appreciable effect on the support of most groups, and then showing only differences above 0.05 highlights the most relevant changes. In the diagram, a colour code is used to indicate whether the characters are shared by both matrices (black, with some differences in scoring), or present only in `older` (blue) or `newer` (red) dataset. Optionally, the effect on the supports for each group and character can be displayed in the form of a table (checking the corresponding box in `Show changes in Support` panel of the dialog for comparing and combining datasets, or with the | option for the `dcomp` command).

In addition to checking the influence of character changes on group supports, TNT also compares the support (for the equivalent groups) before and after addition of taxa present in `newer` but not `older` datasets.

Context of scoring changes. Whether a character supports or contradicts a given group can depend on the context provided by other characters (Goloboff *et al.* 2003). It is possible that a character that (individually) contradicts a given group provides a synapomorphy for that group in the context of the entire matrix. The same problem affects, of course, all comparisons attempting

to determine characters (or individual scorings) that favour or contradict a given result; the classifications provided by `dcomp` are therefore intended as heuristic approximations to guide potential re-examinations of characters (or individual scorings) most likely to affect general results. In the case of comparisons of Bremer support, changing the scoring of a character from `older` to `newer` may increase the support for a given group if none of the other characters differing in both datasets have been changed, but decrease it otherwise. Although modification of the scoring from `older` to `newer` may differently affect the supports depending on the exact sequence (relative to changes in the other characters), for simplicity TNT calculates only the changes in support when the character is changed first and last (i.e. the two extremes, possibly also the two extremes of changes in support, although this may not be so), displaying them in that same order. In the case of taxa, a single value is calculated, the modification of the support when the taxon is added first (i.e. without any of the other taxa unique to `newer` added yet).

An example where the sequence of changes in scoring has the effect just described is shown in Fig. 8. The `older` matrix has three characters (ch0–ch2) in black (shared by the two matrices). The `newer` matrix preserves the same three characters and character states but adds four characters (ch3–c6). For group DE, the only character able to function as a synapomorphy in `older` is ch0; which is in conflict with ch1–ch2. Without them it does not (by itself) support group DE (as it supports that group with a reversal). In the `newer` matrix, character ch3 is identical to ch0, and ch6, contradicting group DE. If ch3 is added first, then ch0 and ch3 are in conflict with ch1–ch2; the group DE is lost. If the character is added last, the two reversals (in ch0 and ch3) outnumber the single character (ch6) contradicting group DE, and the group is again present. Thus, adding ch3 first decreases the support by 0.32; adding ch3 last increases it by 0.32.

The differences just described are the result of character interactions. In many cases, a character will provide either support or contradiction for a given group, regardless of other characters in the matrix. To some extent, therefore, the degree to which changes in support vary in the same direction when adding/rescoring a character first or last is an indication of whether the character supports the group by itself or whether the support depends on the interaction with other characters. An example is character (ch6), which contradicts group DE and does not depend on any other characters. Ch6 always decreases support for group DE, regardless of whether ch6 is added first (–0.32) or last (–0.25).

Related work

The only related work of which we are aware, `phenotools` (an R package; Eliason *et al.* 2019), is focused on recognizing synonymous characters and reporting similar character-state scorings shared by two matrices. That program incorporates more elaborate routines for recognizing wording variants than the string-matching Needleman–Wunsch algorithm in the present implementation. `phenotools` may be useful for automatically harvesting data from many morphology-based datasets without the need for expert knowledge or user input. The emphasis in the TNT implementation is to detect and analyse differences in datasets responsible for differing phylogenetic results. This requires character and character-state matching to pinpoint differences in input and tree-length calculations to assess effects on branch-point support. Quantitative comparisons of this sort are beyond hand calculation or cursory estimation. The methods outlined here complement and extend, rather than shortcut, the work of taxonomic experts.

Discussion

An empirical example: basal radiation of dinosaurs

Recent competing hypotheses over the basal relationships of the three major clades of dinosaurs (Ornithischia, Sauropodomorpha, Theropoda; Fig. 2) provide an opportunity to compare very similar datasets yielding conflicting results. Baron *et al.* (2017a; hereafter 'BEA') argued for a closer relationship between Theropoda and Ornithischia (as Ornithoscelida), whereas in response Langer *et al.* (2017; hereafter 'LEA') generated the traditional pairing of Sauropodomorpha and Theropoda (as Saurischia) as sister clade to Ornithischia.

Although BEA stated "For 130 years, dinosaurs have been divided into two distinct clades – Ornithischia and Saurischia" (p. 501), that is not entirely accurate. 'Clade' implies monophyly and 'distinct' implies an association with widely recognized synapomorphies. The statement comports better with the fairly recent phylogenetic definition of Dinosauria as anchor to a clean a 'node-stem triplet' (Sereno 1999b, c), and only Ornithischia could be described as morphologically distinctive (probably as a result of a sparse early fossil record).

During nineteenth and twentieth centuries, in contrast, a colourful array of ideas on dinosaur phylogeny emerged within a traditional taxonomic framework based on a smidgen of the fossil material available today. As briefly reviewed below, monophyly was not

the only criterion for recognizing higher taxa such as Saurischia.

News of the taxonomic reshuffling by BEA was billed as a "bombshell discovery" that would "shake dinosaur paleontology to its core" (Brusatte 2017, p. 390). In a matter of months, however, LEA challenged their results using the same characters with a suite of altered character-state scores listed in Supplementary material. The original authors (Baron *et al.* 2017b) and colleagues (Parry *et al.* 2017) responded to that rebuttal by examining partitions of the original dataset and by using alternative methods of phylogenetic inference (Parry *et al.* 2017).

Now, some four years later, the controversy has subsided with no resolution over the evidence supporting opposing arguments. The particular characters and character-state differences critical to either basal configuration – a subset of characters and ~2500 character-state changes made by LEA – have not been identified. In this regard, we have advanced little in the past thirty years, since the explicit recognition of character delineation, coding and scoring as the 'bête noire' (Pogue & Mickevich 1990) or 'black box' (Patterson & Johnson 1997) of morphology-based phylogenetics. The ability to measure and characterize differences between relevant character data in overlapping datasets is the fundamental challenge of comparative cladistics (Sereno 2009).

Historical background

Seeley's (1888) division of Dinosauria into Saurischia and Ornithischia initiated what eventually became a traditional two-part pre-cladistic scheme for dinosaurs, with few characters to justify their union in Dinosauria and only primitive features (e.g. triradiate hip design) characterizing one of the two named subgroups (Saurischia). Seeley's Saurischia also excluded Aves, which now is its most diverse subgroup as the taxon has been reconceptualized a century later during the cladistic era. In Seeley's day, new dinosaur finds were pouring in from North America and elsewhere, revolutionizing the fossils and genera to be considered (Marsh 1895). Baur (1887) had identified the three clades we still recognize today (Fig. 2), and he preferred that tripartite arrangement to Seeley's pair in his summary of early classificatory schemes (Baur 1891).

Thus, it was not unusual for twentieth century systematists to regard Saurischia, as well as Dinosauria, as non-monophyletic ('artificial') taxa that arose from a pool of hypothetical basal archosaurs called 'thecodonts'. In his classic treatise on reptilian osteology and taxonomy, Romer (1956, p. 609) wrote that the "morphological 'boundary' between the two orders [Saurischia, Ornithischia] in the Triassic … is far from clear, and it is further possible that the saurischians are polyphyletic, derived from two or more related thecodont lines". In a later addendum (Romer 1968, p. 137), he wrote that "It is not improbable that the Saurischia had a di- or polyphyletic origin, and that coelurosaurs came from a discrete thecodont stock".

For most of the twentieth century, thus, there was little logical basis for, or resolution of, pre-cladistic phylogenetic relationships between basal archosaurs and dinosaurs and among basal dinosaurian clades. Ornithischians, alone, were widely regarded as monophyletic on the basis of a derived pelvic architecture (Romer 1956, 1968). By the end of the 1970s, an expanding suite of 'loose ends' had appeared – basal dinosaurs (e.g. *Herrerasaurus*, *Pisanosaurus*) and dinosaurian outgroups (e.g. *Lagosuchus*, *Lagerpeton*), some fragmentary and most from Late Triassic beds in South America. Their position among the major clades was uncertain.

The overarching traditional understanding of dinosaurs remained the same, as best captured in a major revision of Romer's text by Robert Carroll, one of his last students. Dinosaurs and Saurischia may or may not be monophyletic, although "ornithischians … unquestionably have a common ancestry, established on the basis of a host of shared derived characters" (Carroll 1988, p. 289). Substantive osteological evidence was not apparent that could unite dinosaurs as a whole or link together any two of its three clear subgroups, Ornithischia, Sauropodomorpha and Theropoda.

One of the earliest morphology-based cladistic analyses argued otherwise, that both Dinosauria and Saurischia are monophyletic, and that Saurischia also includes Aves (Gauthier 1986). It was a bold hypothesis, one departing from historical consensus. In the 35 years since Gauthier's paper, dinosaurian monophyly and a basal split into Ornithischia and Saurischia has been almost universally upheld (Padian 2017), albeit only weakly, with basal 'loose end' taxa multiplying in number and sometimes jumping from one basal position to another (all comparable analyses through 2019 in Table 1). During this span, the first relatively complete skulls and skeletons of the earliest ornithischians, sauropodomorphs and theropods were discovered, filling in proverbial missing 'gaps' to the point where their very identity remains controversial. *Eoraptor*, a basal sauropodomorph (Sereno *et al.* 2012), was first described as a theropod (Sereno *et al.* 1993). Its contemporary, the basal theropod *Eodromaeus* (Martínez *et al.* 2011), was mistaken as another *Eoraptor* specimen until fully prepared (Sereno 2012).

BEA's disbanding Saurischia in favour of Theropoda plus Ornithischia, while unorthodox without question, is

less radical than it may seem at first. From the outset, cladistic evidence supporting Dinosauria and Saurischia was limited (Gauthier 1986) and variously reinterpreted. Key results might be overturned with minor adjustment in analytic settings, inclusion/exclusion of taxa or rescoring of character data. And that is exactly what LEA showed by reanalysing BEA's dataset with minimal changes to character-state scores.

BEA's resurrection of Huxley's taxon 'Ornithoscelida' for the new theropod-ornithischian clade, in addition, is problematic, because Huxley clearly included the sauropod *Cetiosaurus* within Ornithoscelida (Huxley 1870, pp. 32, 35), which then overlaps in content Owen's earlier taxon Dinosauria. At the same time, BEA's narrowing of Saurischia to include only sauropodomorphs would generate confusion given the century-long tradition of incorporating both sauropodomorphs and nonavian theropods within this taxon (Holtz 2017). LEA (in their supplementary material) also critiqued the suggested taxonomy of BEA, although they did not mention the taxonomic breadth of Huxley's original use of Ornithoscelida.

Competing hypotheses

Baron *et al.* (2017a). BEA compiled a taxon-character matrix composed of 74 taxa and 457 characters (Table 1). Of the 457 characters, BEA reported that 63 were original to the analysis and most of the remainder were either taken, modified or combined from eight published studies, with citations given for all characters in their supplementary material. Thirty-nine (9%) of the 457 characters were ordered with reasons given in their supplemental material.

Claiming to have decisively overturned Sauropodomorpha plus Theropoda as Saurischia, BEA stated that their analysis "strongly supported" the clade Ornithoscelida (Theropoda plus Ornithischia) with 21 unambiguous synapomorphies (Baron *et al.* 2017a, p. 502). Furthermore, 20 additional steps were required to reconfigure a traditional Saurischia, leading the authors to conclude that there exists "strong support to our recovery of a paraphyletic Saurischia and a monophyletic Ornithoscelida" (Baron *et al.* 2017a, p. 504).

At face value, this would seem to be an insurmountable amount of phylogenetic evidence for ornithoscelidan monophyly. Given the fragmentary nature of many basal taxa, however, the Bremer support value (decay index) for Ornithoscelida was only 4, although higher than either Saurischia (2) or Dinosauria (1). Excluding four of the most incomplete basal taxa (*Diodorus*, *Saltopus*, *Agnosphytis*, *Euceolophysis*) did not raise Bremer support values substantially, which remain low (1–4). Thus, either there are other unstable terminal taxa at basal nodes or the characters in that region are particularly homoplasious – or both.

Langer *et al.* (2017). LEA generated a rapid response to the BEA analysis by palaeontologists who in earlier studies had resolved, if weakly, a monophyletic Saurischia (Theropoda + Sauropodomorpha). LEA used the same 457 characters as BEA but revised some character-state scores and added nine terminal taxa. Their modified BEA matrix weakly favoured the traditional basal split of Dinosauria into Saurischia and Ornithischia, neither hypothesis is significantly different under a Templeton test and the third resolution, an ornithischian-sauropodomorph clade, is only four steps longer. LEA concluded that their corrected BEA dataset alone is insufficient to satisfactorily resolve the basal split within Dinosauria.

LEA also discovered a large number of uninformative characters in the BEA dataset (53, ∼12% of characters), and a lower number of extra steps (15 rather than 20) needed to recompose the traditional Saurischia (given the presence of more sub-optimal trees than initially reported). LEA did not detail the effect of the nine terminal taxa they added to the matrix, although none of these are controversial and some are quite completely known (*Allosaurus*, *Buriolestes*, *Ceratosaurus*, *Daemonosaurus*, *Elaphrosaurus*, *Eoabelisaurus*, *Ixelerpeton*, *Piatnitzkysaurus*, *Scutellosaurus*). Their basal resolution within Dinosauria, however, did not depend on inclusion of new terminal taxa.

LEA mentioned that they made character-state changes but did not indicate the scale of their scoring modification in the main text (all scoring changes were indicated in supplementary material). As Baron *et al.* (2017b) noted in response, approximately 2500 character states were changed, or approximately 10% of the data. Differences in character scoring on the order of 10% would not be unexpected between researchers or research groups, given the complexity of character delineation and supporting fossil evidence. LEA (supplementary material) also commented individually on each of the 21 unambiguous synapomorphies supporting Ornithoscelida in BEA. Many were regarded as problematic in structure or as scored, yet all were retained in the re-analysis. In the final analysis of the re-scored matrix, saurischian monophyly was weakly supported, with only two additional steps needed for recovery of Ornithoscelida.

Resolution of basal splits. We agree with LEA that no resolution among the three basal clades of dinosaurs is strongly supported. Major differences in character delineation, selection and scoring, furthermore, are apparent in even a cursory examination of the datasets in

previous studies (Table 1). "Critical evaluation of characters – how they are defined and scored, whether they are independent from one another, [and] how different authors have used them" was posited to be the only way forward, according to LEA (Langer *et al.* 2017, E2).

Those aims are central to comparative cladistics. BEA *vs* LEA is a relatively simple comparison, as both studies use the same characters. More than 1000 character-state scores were altered by LEA, many of which have no discernible phylogenetic effect. Others favour one or the other hypothesis, and so the argument rests with those characters and character-state scores. Below we implement basic automated character and character-state matching between datasets in order to identify and quantify key differences between these datasets that are responsible for their conflicting results.

Matrix comparison

Comparison in TNT. We used the new implementations in TNT to search for, and colour code, character and character-state differences between BEA and LEA matrices. BEA and LEA have the same set of 457 characters in the same order, but the LEA matrix incorporates nine additional terminal taxa. Character-state cells for those nine taxa, thus, comprise 'unshared data' with no counterpart in the opposing analysis. The remaining character state cells comprise 'shared data' that have matched cells in both matrices (Sereno 2009, table 6).

We label character-state scores either as 'extra' or 'shared', respectively, for the nine terminal taxa present only in the LEA matrix and all other character-state cells. For the 'extra' character-state cells in the LEA matrix, our implementation scores those as 'missing' in the BEA matrix. Character-state differences between the two matrices are of two kinds: (1) 'character-state disparity', when a positively coded state (integer) is matched with an ambiguous state (question mark for missing, too transformed or polymorphic); (2)'character-state conflict', when differing positively coded character states are matched (Sereno 2009, table 6). These character-state differences are logged quantitatively as 50% (0.5) or 100% (1.0) discordant, respectively.

Character comparison in TNT. Most of the scoring differences have little or no effect on the phylogenetic results, with many characters in both matrices supporting the results of both analyses. For the early dinosaur *Eoraptor*, for example, character 1 (skull proportions) has state 1 (preorbital length less than 45% of basal skull length) in BEA but state 0 (preorbital length more than 45% of basal skull length) in LEA (Fig. 9, grey arrow). This scoring change does not alter the status of character 1 as a synapomorphy for any of the groups in conflict in both analyses, nor does it influence length differences in trees derived from either BEA or LEA matrices.

Some characters of LEA (e.g. 4, 9, 11) favour the results of BEA, and some characters of BEA (e.g. 10, 46) favour the hypothesis of LEA (character numbers in Fig. 9 follows TNT's convention for numbering the first character '0', corresponding to the character labelled '1' in LEA and BEA).

Frames that surround a column of states identify characters that generate phylogenetic support for a particular hypothesis (Figs 6, 9). There are 153 framed characters that are of particular interest in this regard (see Supplemental material). We show character frames among the first 50 characters (Fig. 9; see Supplemental material for the full colour-coded matrix).

The matrices use the same 457 characters, which would generate a character similarity index (CSI) of 1.00. Fifty-three characters in BEA, however, are uninformative. After the scoring changes in LEA, 17 of those 53 characters are now informative. With uninformative characters excluded, the CSI in a comparison of BEA and LEA decreases to 0.92, or about 8% discordant. That value, which is quite high for matrix comparisons, is owed to the fact that LEA accepted all of BEA characters, limiting changes to character state scores.

Character-state comparison in TNT. We summarize character-state differences graphically for the first 50 characters (Fig. 9; see Supplemental material for the full colour-coded change matrix). Cells for the nine unique terminal taxa in LEA are scored as 'all-missing' in BEA. As in all of the figures, blue is used for the earlier hypothesis (BEA), and red is used for the later hypothesis (LEA). For character-state disparity (positive character state *vs* ?), the single positive character-state score from one of the analyses is shown in the appropriate colour. For character-state conflict (opposing positive character states), both positive character states are shown in respective colours. Opposing character states for cells that generate phylogenetic differences are circled with appropriate colours (blue supporting earlier, red supporting later). Some are circled with one colour, as the changed character state only provides group support in one analysis, and some are circled with both colours, because the respective states support different groups in opposing analyses.

A total of 8050 cells have different character-state scores between BEA and LEA matrices. Of the 8050 cell differences, approximately 58% (4684 cells) comprise character-state disparity, where a positive character state (0, 1, 2, 3) opposes an ambiguous score (?, –, 0/1). LEA has positive states in 1419 cells that are scored as ambiguous in BEA, and BEA has positive states in 963

Figure 9. Comparison of characters 1–50 in the matrices for the analysis of basal relationships within Dinosauria by Baron *et al.* (2017a) and Langer *et al.* (2017), with circled cells highlighting critical differences (following Figs 6, 7). Different character-state scores without circles do not contribute to differing results. The grey arrow (pointing to a cell for skull proportions in *Eoraptor)*, for example, highlights a character-state score (black 1, red 0) that differs between the matrices of the older study (Baron *et al.* 2017a) and its reconsideration (Langer *et al.* 2017), respectively, but does not contribute to differing phylogenetic results.

cells that are scored as ambiguous in LEA. The remaining (2302 cells) correspond to taxa added in LEA.

The remaining state differences (approximately 42% or 3366 cells) involve character-state conflicts, or cells that have different positive states. Of these character-state conflicts, approximately 39% (3109 cells) generate tree support differences. Resolution of basal branch points within Dinosauria, in other words, hinges on the

Table 2. 'Keystone character map' with data output from TNT summarizing 42 characters at the base of Dinosauria that include all potential and standing synapomorphies and characters with differential tree lengths in competing hypotheses (BEA, LEA) for resolution between the three major clades of dinosaurs (Ornithischia, Sauropodomorpha, Theropoda). Both analyses employ the same 457 characters with character-state scoring differences that generate in BEA Ornithoscelida (Ornithischia + Theropoda) and Herrerasauridae + Sauropodomorpha ('Saurischia' *sensu* BEA) and in LEA Saurischia (Sauropodomorpha + Theropoda), with the early dinosaur *Eoraptor* allied with either Theropoda (BEA) or Sauropodomorpha (LEA). A character 'favours' trees for one dataset if its maximum length in output trees is less than its minimum length in output trees of the opposing dataset. 'Synapomorphies' are present in character maps for the group indicated. Character numbers in TNT are less one, as the initial character is labelled 0 rather than 1. Only those synapomorphies that occur in all most parsimonious trees for each dataset are indicated (note that the diagrams in Supplemental material, for simplicity, use only the strict consensus for each dataset, which may incorrectly add or subtract some synapomorphies). **Abbreviations: 1,** Ornithoscelida in BEA data; **2,** Ornithoscelida in LEA data; **3,** Saurischia in LEA data; **4,** Saurischia (*sensu* LEA) in BEA data; **5,** Herrerasauridae + Sauropodomorpha (= 'Saurischia' *sensu* BEA) in BEA data; **6,** Herrerasauridae + Sauropodomorpha (= 'Saurischia' *sensu* BEA) in LEA data; **BEA,** Baron *et al.* (2017a); **LEA,** Langer *et al.* (2017); **ph**, phalanx; **pr-ab,** presence-absence.

			Favours BEA trees		Favours LEA trees		Synapomorphies						
No.	Original character no.	TNT character no.	BEA data	LEA data	LEA data	BEA data	1	2	3	4	5	6	Short character description
1	26	25	X	—	—	—	—	—	—	—	X	—	Subnarial foramen position
2	30	29	X	—	—	—	—	—	—	—	X	—	Antorbital fenestra shape
3	35	34	X	—	—	—	X	—	—	—	—	—	Maxilla, lateral ridge
4	47	46	—	—	X	X	—	—	—	X	—	—	Lacrimal lateral overhang
5	54	53	X	—	—	—	X	—	—	—	—	—	Jugal border, antorbital fenestra
6	56	55	—	—	X	—	—	—	X	—	—	—	Jugal, anterior ramus shape
7	76	75	X	—	—	—	X	—	—	—	—	—	Quadrate inclination (lateral)
8	88	87	X	—	—	—	X	—	—	X	—	—	Paroccipital process shape
9	89	88	—	X	—	—	—	—	—	—	X	—	Opisthotic ventral process length
10	90	89	X	—	—	—	X	—	—	—	—	—	Posttemporal foramen position
11	97	96	X	—	—	—	X	—	—	—	—	—	Supraoccipital shape
12	100	99	X	—	—	—	X	—	—	—	—	—	Parabasisphenoid fossa shape
13	141	140	—	—	—	—	—	X	—	—	—	—	Surangular, anterior foramen
14	145	144	X	—	—	—	X	—	—	—	—	—	Retroarticular process pr-ab/shape
15	180	179	X	—	—	—	—	—	—	—	X	—	Tooth crown shape/ornamentation
16	190	189	—	—	X	X	—	—	X	X	—	—	Cervical epipophyses
17	222	221	X	—	—	—	X	—	—	—	—	—	Dorsosacral vertebral number
18	224	223	X	—	—	—	—	—	—	X	X	—	Primordial sacral 1 rib end shape
19	241	240	—	—	X	—	X	—	—	—	—	—	Scapular blade shape
20	247	246	X	—	—	—	—	—	—	—	X	—	Coracoid postglenoid notch
21	256	255	X	—	—	—	X	—	—	—	—	—	Humeral shaft curvature
22	279	278	—	—	X	X	—	—	—	X	—	—	Manual digit V phalanges
23	285	284	—	—	X	—	—	—	X	—	—	—	Metacarpal 1 distal condyles size/offset
24	286	285	—	—	—	—	—	—	—	—	X	X	Manual digit I-1 ph, distal end rotated
25	290	289	—	—	X	—	—	—	X	—	—	—	Metacarpal 4, shaft width
26	308	307	X	—	—	—	X	—	—	—	—	—	Iliac acetabular flange, pr-ab/size
27	321	320	—	—	X	—	—	—	—	X	X	—	Ischial distal end shape
28	325	324	—	X	—	—	—	X	—	—		—	Ischial proximal articulations joined
29	329	328	—	—	X	—	—	—	X	—		—	Ischium median contact size
30	330	329	—	—	X	—	—	—	—	—	X	—	Ischial midshaft, cross-section
31	360	359	—	—	—	—	X	—	—	—		—	Femoral shaft curvature
32	370	369	X	—	X	—	X	—	—	—		—	Femoral anterior trochanter form
33	372	371	X	—	—	—	X	—	—	—		—	Femoral trochanter cleft
34	402	401	—	—	—	—	—	—	—	X	X	—	Tibial distal end width

(*Continued*)

Table 2. (*Continued*).

			Favours BEA trees		Favours LEA trees		Synapomorphies						
No.	Original character no.	TNT character no.	BEA data	LEA data	LEA data	BEA data	1	2	3	4	5	6	Short character description
35	417	416	X	—	X	—	—	—	X	—	X	—	Astragalar fossa, proximal surface
36	424	423	—	—	—	—	X	—	—	—	—	—	Calcaneum shape
37	427	426	—	—	X	X	—	—	X	X	—	—	Distal tarsal 4 posterior process
38	435	434	—	—	X	—	X	—	—	—	—	—	Metatarsal 1 proximal end position
39	438	437	X	—	—	—	X	—	—	—	—	—	Distal tarsal-metatarsal fusion
40	440	439	X	—	—	—	—	—	—	X	X	—	Metatarsal 4, proximal lateral flange
41	446	445	X	—	X	—	—	—	—	X	X	—	Pedal digit I length
42	450	449	—	—	—	—	—	—	—	—	—	X	Pedal digit I-ungual shape

interpretation of 39% of all cells that exhibit differences, or just over 3000 cells. The remaining ~5000 cells that have different character-state scores, including all character-state disparity, seem to have no (or little) influence on differing phylogenetic results.

In sum, given a total cell count of 75,862 (457 characters in 83 terminal taxa), approximately 10.6% (8050 cells) were assigned different character states. Subtracting for character-state disparity (0.5) and conflict (1.0) across all cells we calculate a character-state similarity indices (CSSI) of 0.865 and 0.855, the former counting differences in character states for all 457 characters and the latter including only differences between characters that are informative in both analyses. Thus, the matrices are substantially similar, differing in their character states (CSSI) by only ~14%.

Many of the cells scored for LEA's added terminal taxa notably influence the difference in results between the analyses (Fig. 9, lower portion of the matrix), although they were not determinative for traditional saurischian monophyly (Sauropodomorpha + Theropoda + Herrerasauridae). A majority, but not all, scoring changes by LEA in original and added terminal taxa favour their results regarding Saurischia. A prior hypothesis of relationships, thus, could have influenced some, but not all, character-state rescoring, by LEA.

Keystone characters. The focus of controversy between BEA and LEA analyses is the resolution between the three major clades at the base of Dinosauria. Scoring differences, therefore, can be isolated that support these conflicting basal resolutions. TNT recognizes 42 of these 'keystone characters' (LEA, Supplementary material) within the full set of 457 characters, or about 8% of the dataset (Table 2). These 42 characters either have fewer steps in one of the matrices on minimum length trees or are optimized as synapomorphies in either dataset for one of three clades in the opposing analyses: Ornithoscelida (Ornithischia + Theropoda) and 'Saurischia' (Herrerasauridae + Sauropodomorpha) in BEA, and traditional Saurischia in LEA.

BEA reported 21 unambiguous synapomorphies for Ornithoscelida and 20 additional steps needed to recompose Saurischia. Three of those characters, however, are not unambiguous ornithoscelidan synapomorphies, and a full set of optimal trees reducing to 15 the number of additional steps to recompose a traditional Saurischia (Langer *et al.* 2017). BEA's dataset also includes 10 potential synapomorphies for LEA's traditional Saurischia. LEA's dataset includes seven synapomorphies for Saurischia as well as two potential synapomorphies supporting Ornithoscelida (Table 2).

Homoplasy in keystone characters is substantial, as Bremer support values for the preferred resolutions in BEA (Ornithoscelida) and LEA (Saurischia) are four and two, respectively. The scoring change matrix of 42 keystone characters likely captures most of the character-state differences between these analyses that lead to different results at the base of Dinosauria. Most of the 457 characters and 74 (or 83) terminal taxa are irrelevant to resolution at the base of Dinosauria but contribute to an imposing and homoplasious dataset that cannot be analysed manually.

Conclusions

More than a decade ago, when morphology-based phylogenetic analyses tended to have smaller datasets, controversy arose over the early branch points within Dinosauria. A comparative analysis at that time (Sereno 2007b) showed strong character selection at work. Approximately 50% of the characters relevant to basal nodes in one analysis (Sereno 1999a) were excluded by a later analysis claiming new results (Langer & Benton 2006). Among the shared characters in each matrix, furthermore, 40% showed different character-state scores for the same terminal taxa. Character selection and character-state scoring in these and many previous phylogenetic controversies are central to differing results, necessitating a more rigorous approach to reporting and analysing the sources responsible for differing phylogenetic results (Sereno 2009).

A decade later, the problem of basal dinosaur relationships has reached a new and still controversial landmark, wherein opposing analyses (Baron *et al.* 2017a; Langer *et al.* 2017) using a very much expanded taxon-character matrix draw very different conclusions about the relationships between the three major clades of dinosaurs. Langer *et al.* (2017, E2) concluded that "a more critical evaluation of characters – how they are defined and scored, whether they are independent from one another, how different authors have used them – is the best tool for untangling the roots of the dinosaur family tree". These very issues, nonetheless, were amply demonstrated long ago as the major factors colouring phylogenetic results at the base of Dinosauria (Sereno 2007b). A quantitative comparative approach is needed that specifically compares characters and character states between competing analyses.

In this study, we described an implementation in TNT to make comparisons of morphology-based datasets. We compared, in particular, competing datasets for early branch points within Dinosauria (Baron *et al.* 2017a; Langer *et al.* 2017), the second a relatively simple modification of the first. The controversy generated a range of interpretations and subsequent commentary

(Baron *et al.* 2017b; Parry *et al.* 2017), although very little quantitative analysis of scoring differences. We determined that about 21% (∼8000) of matrix cells differed between the analyses with approximately 39% of those (∼3000 cells) of phylogenetic impact. Scoring changes responsible for the phylogenetic differences are concentrated in 153 characters (34% of character data) with even fewer characters and character states involved at basal nodes.

It would be very interesting to compare the BEA matrix to recent and largely independent matrices for basal dinosaurs that maintain a traditional Saurischia (Cabreira *et al.* 2016; Marsh *et al.* 2019; Table 1). These comparisons, however, would be more challenging than our BEA–LEA comparison, given greater differences between the matrices in character formulation.

Short of manually editing and re-formulating character statements to instil uniformity, comparisons to, and among, the more disparate datasets in Table 1 will require more elaborate matching implementations.

Morphology-based phylogenetic analysis is hamstrung by the inability to effectively compare taxon-character matrices. Dataset comparisons show that considerable differences in character selection and character-state scoring are the norm for competing morphology-based phylogenies (Sereno & Brusatte 2009). Given the size of many datasets, computer-assisted implementation must be employed to locate and summarize differences in character selection and character-state scores between analyses, much as computer assistance revolutionized the determination and characterization of preferred trees in the 1980s. The implementation described here provides an initial step in that direction.

Acknowledgements

We thank Diego Pol for suggesting this collaboration and, with Santiago Catalano, for testing the new algorithms and options on many datasets and for suggesting several useful modifications and features. Comments by Jonah Choiniere and reviewers Michael Lee and anonymous also improved many aspects of the manuscript. We also thank editor Paul Barrett for his assistance in the submission and review process. This research was supported by a grant from the Consejo Nacional de Investigaciones Científicas y Técnicas (PUE 0070 to PA, PIP 110 to C. Szumik).

Supplemental material

Supplemental material for this article can be accessed here: https://doi.org/10.1080/14772019.2021.1970038.

ORCID

Pablo A. Goloboff http://orcid.org/0000-0002-1979-3982

References

Baron, M. G., & **Barrett, P. M.** 2017. A dinosaur missing-link? *Chilesaurus* and the early evolution of ornithischian dinosaurs. *Biology Letters*, **13** 20170220. doi: 10.1098/rsbl.2017.0220.

Baron, M. G., Norman, D. B. & **Barrett, P. M.** 2017a. A new hypothesis of dinosaur relationships and early dinosaur evolution. *Nature*, **543**, 501–506. doi: 10.1038/nature24012.

Baron, M. G., Norman, D. B. & **Barrett, P. M.** 2017b. Baron *et al.* reply. *Nature*, **551**, E4–5.

Baur, G. 1887. On the phylogenetic arrangement of the Sauropsida. *Journal of Morphology*, **1**, 95–104.

Baur, G. 1891. Remarks on the reptiles generally called Dinosauria. *American Naturalist*, **25**, 434–453.

Bremer, K. 1994. Branch support and tree stability. *Cladistics*, **10**, 295–304.

Brusatte, S. L. 2017. Evolution: uprooting the dinosaur family tree. *Current Biology*, **27**, 390–392.

Cabreira, S. F., Kellner, A. W. A., Dias-da-Silva, S., Roberto da Silva, L., Bronzati, M., Marsola, J. C. A., Müller, R. T., Bittencourt, J. S., Batista, B. J., Raugust, T., Carrilho, R., Brodt, A. & **Langer, M. C.** 2016. A unique Late Triassic dinosauromorph assemblage reveals dinosaur ancestral anatomy and diet. *Current Biology*, **26**, 3090–3095.

Carroll, R. L. 1988. *Vertebrate paleontology and evolution*. W. H. Freeman & Company, New York, 698 pp.

Catalano, S. A., Goloboff, P. A. & **Giannini, N. P.** 2010. Phylogenetic morphometrics (I): the use of landmark data in a phylogenetic framework. *Cladistics*, **26**, 539–549.

Dahdul, W. M., Balhoff, J. P., Blackburn, D. C., Diehl, A. D., Haendel, M. A., Hall, B. K., Lapp, H., Lundberg, J. G., Mungall, C. J., Ringwald, M., Segerdell, E., Van Slyke, C. E., Vickaryous, M. K., Westerfield, M. & **Mabee, P. M.** 2012. A unified anatomy ontology of the vertebrate skeletal system. *PLoS ONE*, **7**(12), e51070. doi: 10.1371/journal.pone.0051070

Dahdul, W., Balhoff, J., Engeman, J., Grande, T., Hilton, E., Kothari, C., Lapp, H., Lundberg, J., Midford, P., Vision, T., Westerfield, M. & **Mabee, P.** 2010. Evolutionary characters, phenotypes and ontologies: Curating data from the systematic biology literature. *PLoS ONE*, **5**, e10708. doi:10.1371/journal.pone.0010708

Deans, A. R, Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B., Cooper, L. D., Courtot, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T. A., Dettai, A., Diogo, R., Druzinsky, R. E., Dumontier, M., Franz, N. M., Friedrich, F., Gkoutos, G. V., Haendel, M., Harmon, L. J., Hayamizu, T. F., He, Y., Hines, H. M., Ibrahim, N., Jackson, L. M., Jaiswal, P., James-Zorn, C., Köhler, S., Lecointre, G., Lapp, H., Lawrence, C. J., Le Novère, N., Lundberg, J. G., Macklin, J., Mast, A. R., Midford, P., Mikó, I.,

Mungall, C. J., **Oellrich, A.**, **Osumi-Sutherland, D.**, **Parkinson, H.**, **Ramírez, M. J.**, **Richter, S.**, **Robinson, P. N.**, **Ruttenberg, A.**, **Schulz, K. S.**, **Segerdell, E.**, **Seltmann, K. C.**, **Sharkey, M. J.**, **Smith, A. D.**, **Smith, B.**, **Specht, C. D.**, **Squires, R. B.**, **Thacker, R. W.**, **Thessen, A.**, **Fernandez-Triana, J.**, **Vihinen, M.**, **Vize, P. D.**, **Vogt, L.**, **Wall, C. E.**, **Walls, R. L.**, **Westerfeld, M.**, **Wharton, R. A.**, **Wirkner, C. S.**, **Woolley, J. B.**, **Yoder, M. J.**, **Zorn, A. M.** & **Mabee, P.** 2015. Finding our way through phenotypes. *PLoS Biology*, **13**, 1–9. doi: 10.1371/journal.pbio.1002033

Dececchi, T. A., **Balhoff, J. P.**, **Lapp, H.** & **Mabee, P. M.** 2015. Towards synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Systematic Biology*, **64**, 936–952. doi:10.1093/sysbio/syv031

Dembo, M. 2016. *Exploring morphological phylogenetics of fossil hominins.* PhD thesis, Department of Archaeology, Faculty of Environment, Simon Fraser University, 235 pp.

Dembo, M., **Matzke, N. J.**, **Mooers, A. Ø.** & **Collard, M.** 2015. Bayesian analysis of a morphological supermatrix sheds light on controversial fossil hominin relationships. *Proceedings of the Royal Society of London, Series B*, 28220150943. doi:10.1098/rspb.2015.0943

Eliason, C. M., **Edwards, S. V.** & **Clarke, J.** 2019. phenotools: an R package for visualizing and analyzing phenomic datasets. *Methods in Ecology and Evolution*, **10**, 1393–1400. doi:10.1111/2041-210X.13217

Excurra, M. D. & **Brussatte, S. L.** 2011. Taxonomic and phylogenetic reassessment of the early neotheropod dinosaur *Camposaurus arizonensis* from the Late Triassic of North America. *Palaeontology*, **54**, 763–772.

Gauthier, J. A. 1986. Saurischian monophyly and the origin of birds. *Memoirs of the California Academy of Sciences*, **8**, 1–55.

Goloboff, P. A. 1993. A reanalysis of mygalomorph spider families. *American Museum Novitates*, **3056**, 1–32.

Goloboff, P. A. 1997. Self-weighted optimization: character state reconstructions and tree searches under implied transformation costs. *Cladistics*, **13**, 225–245. doi:10.1111/j.1096-0031.1997.tb00317.x

Goloboff, P. A. 2013. Extended implied weighting. *Cladistics*, **30**, 260–272.

Goloboff, P. A. 2014. Oblong, a program to analyse phylogenomic data sets with millions of characters, requiring negligible amounts of RAM. *Cladistics*, **30**, 273–281.

Goloboff, P. A. & **Farris, J. S.** 2001. Methods for quick consensus estimation. *Cladistics*, **17**, S26–S34.

Goloboff, P. A., **Farris, J. S.**, **Källersjö, M.**, **Oxelmann, B.**, **Ramírez, M. J.** & **Szumik, C. A.** 2003. Improvements to resampling measures of group support. *Cladistics*, **19**, 324–332.

Goloboff, P. A., **Farris, J. S.** & **Nixon, K.** 2008. TNT, a free program for phylogenetic analysis. *Cladistics*, **24**, 774–786.

Goloboff, P. A. & **Catalano, S. A.** 2012. GB-to-TNT: facilitating creation of matrices from GenBank and diagnosis of results in TNT. *Cladistics*, **28**, 503–513. doi: 10.1111/j.1096-0031.2012.00400.x

Goloboff, P. A. & **Catalano, S. A.** 2016. TNT version 1.5, including a full implementation of geometric morphometrics. *Cladistics*, **32**, 221–238.

Goloboff, P. A. & **Morales, M.** 2020. A phylogenetic C interpreter for TNT. *Bioinformatics*, **13**, 3988–3995.

Goloboff, P., De Laet, J., Ríos-Tamayo, D. & **Szumik, C.** 2021. A reconsideration of inapplicable characters, and an approximation with step-matrix recoding. *Cladistics*. doi: 10.1111/cla.12456

Harris, S. R., **Pisani, D.**, **Gower, D. J.** & **Wilkinson, M.** 2007. Investigating stagnation in morphological phylogenetics using consensus data. *Systematic Biology*, **56**, 125–129.

Hawkins, J. A. 2000. A survey of primary homology assessment: different botanists perceive and define characters in different ways. Pp. 22–53 in R. Scotland & R. T. Pennington (eds) *Homology and systematics*. Special Volume Series **58**. Taylor & Francis, London.

Holtz, T. R. 2017. Share names for dinosaur divisions. *Nature*, **545**, 30. doi:10.1038/545030d

Huxley, T. H. 1870. On the classification of the Dinosauria, with observations on the Dinosauria of the Trias. *Proceedings of the Geological Society*, **26**, 32–51.

Irmis, R. B., **Nesbitt, S. J.**, **Padian, K.**, **Smith, N. D.**, **Turner, A. H.**, **Woody, D.** & **Downs, A.** 2007. A Late Triassic dinosauromorph assemblage from New Mexico and the rise of dinosaurs. *Science*, **317**, 358–361.

Kluge, A. & **Farris, J.** 1969. Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, **18**, 1–32.

Langer, M. C. 2004. 2. Basal Saurischia. Pp. 25–46 in D. B. Weishampel, P. Dodson & H. Osmólska (eds) *The Dinosauria*. Second edition. University of California Press, Berkeley. doi:10.1525/9780520941434-007/html

Langer, M. C. 2014. The origins of Dinosauria: much ado about nothing. *Palaeontology*, **214**, 1–10.

Langer, M. C. & **Benton, M. J.** 2006. Early dinosaurs: a phylogenetic study. *Journal of Systematic Palaeontology*, **4**, 309–358.

Langer, M. C., **Ezcurra, M. D.**, **Rauhut, O. W. M.**, **Benton, M. J.**, **Knoll, F.**, **McPhee, B. W.**, **Novas, F. E.**, **Pol, D.** & **Brusatte, S. L.** 2017. Untangling the dinosaur family tree. *Nature*, **551**, E1–E3. doi:10.1038/nature24011

Marsh, A. D., **Parker, W. G.**, **Langer, M. C.** & **Nesbitt, S. J.** 2019. Redescription of the holotype specimen of *Chindesaurus bryansmalli* Long and Murry, 1995 (Dinosauria, Theropoda), from Petrified Forest National Park, Arizona. *Journal of Vertebrate Paleontology*, **39**, e1645682. doi:10.1080/02724634.2019.1645682

Marsh, O. C. 1895. On the affinities and classification of the dinosaurian reptiles. *American Journal of Science*, **50**, 483–498.

Martínez, R. N., **Sereno, P. C.**, **Alcober, O. A.**, **Colombi, C. E.**, **Renne, P. R.**, **Montañez, I. P.** & **Currie, B. S.** 2011. A basal dinosaur from the dawn of the dinosaur era in southwestern Pangaea. *Science*, **331**, 206–210.

Müller, R. T. & **Dias-da-Silva, S.** 2019. Taxon sampling and character coding deeply impact branches in phylogenetic trees of dinosaurs. *Historical Biology*, **31**, 1089–1092.

Needleman, S. & **Wunsch, C.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

Nesbitt, S. J. 2011. Early evolution of archosaurs: relationships and the origin of major clades. *Bulletin of the American Museum of Natural History*, **352**, 1–292.

Nesbitt, S. J., **Irmis, R. B.**, **Parker, W. G.**, **Smith, N. D.**, **Turner, A. H.** & **Rowe, T.** 2009. Hindlimb osteology and

the distribution of basal dinosauromorphs from the Late Triassic of North America. *Journal of Vertebrate Paleontology*, **29**, 498–516.

Novas, F. E. 1992. Phylogenetic relationships of the basal dinosaurs, the Herrerasauridae. *Palaeontology*, **35**, 51–62.

Novas, F. E. 1993. New information on the systematics and postcranial skeleton of *Herrerasaurus ischigualastensis* (Theropoda: Herrerasauridae) from the Ischigualasto Formation (Upper Triassic) of Argentina. *Journal of Vertebrate Paleontology*, **13**, 400–423.

O'Leary, M. A., **Alphonse, K.**, **Arce, M.**, **Cavaliere, D.**, **Cirranello, A.**, **Dietterich, T.**, **Julius, M.**, **Law, E.**, **Passarotti, M.**, **Reft, A.**, **Robalino, J.**, **Simmons, N.**, **Smith, S.**, **Stevenson, D.**, **Theriot, E.**, **Velazco, P.**, **Walls, R.**, **Yu, M.** & **Daly, M.** 2018. Crowds replicate performance of scientific experts scoring phylogenetic matrices of phenotypes. *Systematic Biology*, **67**, 49–60.

Padian, K. 2017. Dividing the dinosaurs. *Nature*, **543**, 494–495.

Palci, A. & **Lee, M. S. Y.** 2018. Geometric morphometrics, homology and cladistics: review and recommendations. *Cladistics*, **35**, 230–242.

Parry, L. A., **Baron, M. G.** & **Vinther, J.** 2017. Multiple optimality criteria support Ornithoscelida. *Royal Society Open Science*, **4**, 170833. doi:10.1098/rsos.170833

Patterson, C. & **Johnson, G. D.** 1997.The data, the matrix, and the message: Comments on Begle's "Relationships of the osmeroid fishes". *Systematic Biology*, **46**, 358–365.

Pogue, M. G. & **Mickevich, M. F.** 1990. Character definitions and character state definitions: the bête noire of phylogenetic inference. *Cladistics*, **6**, 319–361.

Ramírez, M. J., **Coddington, J. A.**, **Maddison, W. P.**, **Midford, P. E.**, **Prendini, L.**, **Miller, J.**, **Griswold, C. E.**, **Hormiga, G.**, **Sierwald, P.**, **Scharff, N.**, **Benjamin, S. P.** & **Wheeler, W. C.** 2007. Linking of digital images to phylogenetic data matrices using a morphological ontology. *Systematic Biology*, **56**, 283–294.

Rauhut, O. A. 2003. The interrelationships and evolution of basal theropods. *Special Papers in Palaeontology*, **69**, 1–213.

Richter, S. & **Wirkner, C.** 2014. A research program for evolutionary morphology. *Zoological Systematics and Evolutionary Research*, **52**, 338–350.

Romer, A. S. 1956. *Osteology of the reptiles*. University of Chicago Press, Chicago, 772 pp.

Romer, A. S. 1968. *Notes and comments on vertebrate paleontology*. University of Chicago Press, Chicago, 304 pp.

Seeley, H. G. 1888. On the classification of the fossil animals commonly named Dinosauria. *Proceedings of the Royal Society of London*, **43**, 165–171.

Sereno, P. C. 1999a. The evolution of dinosaurs. *Science*, **284**, 2137–2147.

Sereno, P. C. 1999b. A rationale for dinosaurian taxonomy. *Journal of Vertebrate Paleontology*, **19**, 788–790.

Sereno, P. C. 1999c. Definitions in phylogenetic taxonomy: critique and rationale. *Systematic Biology*, **48**, 329–351.

Sereno, P. C. 2007a. Logical basis for morphological characters in phylogenetics. *Cladistics*, **23**, 565–587.

Sereno, P. C. 2007b. Phylogenetic relationships of early dinosaurs: a comparative report. *Historical Biology*, **19**, 145–155.

Sereno, P. C. 2009. Comparative cladistics. *Cladistics*, **25**, 624–659.

Sereno, P. C. 2012. Preface. *Journal of Vertebrate Paleontology*, **32**, 1–9.

Sereno, P. C., **Forster, C. A.**, **Rogers, R. R.** & **Monetta, A. M.** 1993. Primitive dinosaur skeleton from Argentina and the early evolution of Dinosauria. *Nature*, **361**, 64–66.

Sereno, P. C. & **Brusatte, S. L.** 2009. Comparative assessment of tyrannosaurid interrelationships, *Journal of Systematic Palaeontology*, **7**, 455–470.

Sereno, P. C., **Martínez, R.** & **Alcober, O. A.** 2012. Osteology of *Eoraptor lunensis* (Dinosauria, Sauropodomorpha). *Memoir of the Society of Vertebrate Paleontology*, **12**, 83–179.

Sues, H.-D., **Nesbitt, S. J.**, **Berman, D. S.** & **Henrici, A.** 2011. A late-surviving basal theropod dinosaur from the latest Triassic of North America. *Proceedings of the Royal Society of London*, *Series B*, **278**, 3459–3464.

Vogt, L., **Bartolomaeus, T.** & **Giribet, G.** 2010. The linguistic problem of morphology: structure versus homology and the standardization of morphological data. *Cladistics*, **26**, 301–325. doi:10.1111/j.1096-0031.2009.00286.x

Whitlock, J. A. & **Wilson, J. A.** 2017. Character distribution maps: a visualization method for comparative cladistics. *Zoological Journal of the Linnean Society*, **94**, 490–499. doi:10.1111/azo.12006

Associate Editor: Jonah Choiniere