



OPEN

SUBJECT AREAS:

EVOLUTIONARY
BIOLOGY

DATA MINING

SOFTWARE

MOLECULAR EVOLUTION

TreeParser-Aided Klee Diagrams Display Taxonomic Clusters in DNA Barcode and Nuclear Gene Datasets

Mark Y. Stoeckle & Cameron Coffran

Program for the Human Environment, The Rockefeller University, New York, NY 10065.

Received

14 June 2013

Accepted

23 August 2013

Published

11 September 2013

Correspondence and requests for materials should be addressed to M.Y.S. (mark.stoeckle@rockefeller.edu)

Indicator vector analysis of a nucleotide sequence alignment generates a compact heat map, called a Klee diagram, with potential insight into clustering patterns in evolution. However, so far this approach has examined only mitochondrial cytochrome *c* oxidase I (COI) DNA barcode sequences. To further explore, we developed TreeParser, a freely-available web-based program that sorts a sequence alignment according to a phylogenetic tree generated from the dataset. We applied TreeParser to nuclear gene and COI barcode alignments from birds and butterflies. Distinct blocks in the resulting Klee diagrams corresponded to species and higher-level taxonomic divisions in both groups, and this enabled graphic comparison of phylogenetic information in nuclear and mitochondrial genes. Our results demonstrate TreeParser-aided Klee diagrams objectively display taxonomic clusters in nucleotide sequence alignments. This approach may help establish taxonomy in poorly studied groups and investigate higher-level clustering which appears widespread but not well understood.

Comparing nucleotide sequences from different organisms helps understand evolution. Applications range from reconstructing the earliest branches on the Tree of Life to mapping the routes and timing of human expansion out of Africa^{1–3}. Standard approaches evaluate homologous nucleotide or amino acid positions across a sequence alignment to infer the probable order of divergences, and display results in a tree diagram of evolutionary history^{4,5}. Phylogenetic methods generally emphasize branching order—the sequence of events along each branch—and less so timing across divisions. As a result, coincident divergences involving multiple boughs may be overlooked. Specific methods designed to detect clustering have been applied to species delimitation and viral evolution^{6–9}. This relatively limited focus to date likely reflects the commonly-held view that higher taxa are arbitrary demarcations of the taxonomic hierarchy rather than indicators of evolutionary processes^{10,11}.

Matrix heat maps help visualize clustering in complex datasets and can compress hundreds of thousands of data points into single-page displays^{12,13}. Applications range from evaluating social networks to identifying diagnostic gene expression profiles in tumors and brain scan patterns associated with schizophrenia^{14–18}. Matrix rows and columns are sorted, typically by hierarchical clustering, and the rearranged matrix is colorized as a heat map. Clusters of correlated inputs show up as “hot blocks” along the diagonal. Matrices may be *asymmetric*, e.g., a gene expression profile with genes sorted along one axis and cell types along the other, or *symmetric*, with identical inputs along both axes (e.g.^{19,20}).

A symmetric matrix heat map approach to comparative nucleotide sequence analysis using indicator vector correlations is recently described^{21,22}. Indicator vectors are digital transformations of nucleotide sequences in vector space; correlations are roughly inversely proportional to p-distances. Unlike simple p-distance methods, scaling of correlations is relative rather than absolute and vectors can represent multiple sequences. Indicator vector analysis generates a Klee diagram, a colorized heat map of the correlation matrix. Taxonomy-ordered Klee diagrams may offer new insights into evolution^{22–26}. However, to date this approach has only been applied to mitochondrial COI barcode sequences and is limited by the need for an accurate taxonomic list which is not readily available for most groups. Here we describe TreeParser, a web-based software that sorts a nucleotide sequence alignment according to a phylogenetic tree generated from the same dataset, facilitating an otherwise time-consuming step in this analytic pipeline. To assess potential utility, we apply TreeParser-indicator vector analysis to mitochondrial and nuclear gene datasets and examine clustering in the resulting Klee diagrams.



Results

TreeParser, Klee performance. TreeParser run times on the web were less than 5 s for alignments with 5,000 or fewer sequences. Larger files containing 7,500 and 10,000 sequences and were sorted in 14 s and 26 s, respectively. TreeParser outputs closely followed template trees. Differences reflected topology-equivalent branch rotations and alternate ordering of identical sequences (Supplementary Figs. S1, 2)²⁷. Klee diagrams required approximately two to three minutes on a desktop machine.

***Astraptēs fulgerator* COI barcodes.** The skipper butterfly *A. fulgerator* from northwestern Costa Rica is proposed to represent ten cryptic species based on differences in caterpillar morphology, food plants, and COI barcodes²⁸. The putative species, which have modest sequence differences (average nearest neighbor distance, 1.76% K2P; range 0.32%–5.41%), formed discrete blocks of high correlation along the diagonal in TreeParser-ordered Klee diagram (Fig. 1). Exceptions were INGCUP and HIHAMP, which differ by 1–2 nucleotides and were not clearly demarcated. Whether or not these constitute valid species has been questioned^{29,30}.

***Setophaga* warbler COI barcodes.** The *Setophaga* wood warblers are one of the youngest groups of songbirds, an “explosive radiation” of largely North American species that diversified in the past 5–10 million years³¹. A Klee diagram of the TreeParser-ordered alignment, which included 24 of the 25 *Setophaga* species in North America, displayed distinct blocks of high correlation corresponding to species (Fig. 1). Expected exceptions were two species pairs known to share barcodes either due to ongoing hybridization (*S. townsendi/occidentalis*) or recent divergence (*S. americana/pitiayumi*). It has been proposed that the latter pair represent a single species³².

Tyrannid flycatchers and allies recombination activating gene 1 (RAG-1). This published dataset includes representatives of nearly all (93%) Tyrannides genera³³. Individual species are represented by single sequences. A Klee diagram of the TreeParser-ordered FASTA file displayed discrete blocks of higher correlation along the diagonal that corresponded to the revised Tyrannides phylogeny, including four of five families and several subfamilies (Fig. 2a). Some groups were “split” or “lumped” in the RAG-1 Klee. For example, two Tyrannidae subfamilies appeared as a single block, and family

Tityridae was split into independent blocks corresponding to subfamily divisions (Fig. 2b).

Comparison of avian RAG-1, COI. The avian RAG-1 Klee showed strongly demarcated blocks reflecting major phylogenetic divisions of birds (Fig. 3)^{34,35}. Short mitochondrial sequences such as COI barcodes are generally considered to lack sufficient information for evolutionary analysis above the species level^{36–38}. Thus it was of note that much of RAG-1 Klee structure was mirrored in the COI diagram, although the discontinuities were less marked (Fig. 3).

Butterfly elongation factor 1 α (EF-1), COI. These published datasets included sequences from 89 species representing five of seven recognized butterfly families, and include 15 subfamilies and 52 genera³⁹. Clusters corresponding to recognized taxonomic divisions were evident in both the EF-1 and COI Klees, including family Lycaenidae and subfamilies within Nymphalidae and Papilionidae (Fig. 4). In the Klee diagram generated from concatenated EF-1 and COI sequences, three additional families emerged as discrete blocks.

Discussion

Heat map analysis requires an organized matrix. In this study, phylogeny-ordered alignments enabled Klee heat map visualization of evolutionary sequence clusters. To generate Klee diagrams, we previously sorted sequence alignments by hand according to a taxonomic list or a phylogenetic tree. This was not optimal even for small datasets, as errors were unpredictable and hard to identify and correct. For large datasets, manual reordering was simply not feasible—a computational approach was needed. To enable automated sorting we developed the TreeParser software described in this paper. The results demonstrate that TreeParser sorts a nucleotide sequence FASTA file according to a phylogenetic tree generated from the same data. The stand-alone web version accepts standard format files and requires no additional software. In this report MEGA NJ algorithm was used to produce template trees⁴⁰. Any phylogenetic software that generates a standard format Newick tree file⁴¹ could be utilized by converting the Newick file to text format in MEGA before uploading to TreeParser. However, it is likely optimal to use a distance-based method such as NJ to create the template, given that indicator vector correlations are most closely related to Hamming or p-distances²¹. Thus distance-based NJ ordering is expected to closely follow

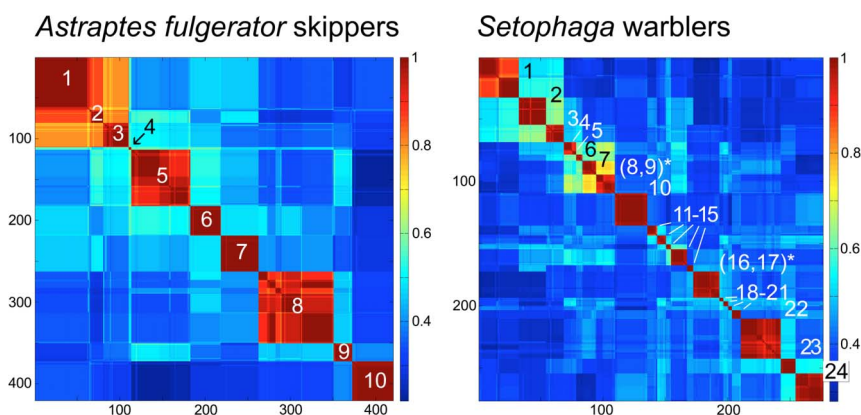


Figure 1 | Species-level clusters in butterflies and birds. At left, skipper butterfly *Astraptēs fulgerator* COI barcode Klee diagram generated from TreeParser ordered alignment ($n = 420$) with correlation scale at right of diagram. Sequence clusters appear as blocks of high correlation along the diagonal and correspond to the 10 provisional species (1. INGCUP, 2. HIHAMP, 3. FABOV, 4. BYTTNER, 5. YESENN, 6. LONCHO, 7. LOHAMP, 8. SENNOV, 9. CELT, 10. TRIGO). Block sizes reflect number of sequences per species ($n = 3–88$). At right, *Setophaga* warblers COI barcode Klee generated from TreeParser-ordered alignment ($n = 276$; 3–32 per species). Blocks along the diagonal correspond to species; species with shared blocks are marked with an asterisk (1. *petechiae*, 2. *striata*, 3. *pennsylvanica*, 4. *negrescens*, 5. *graciae*, 6. *discolor*, 7. *virens*, 8. *occidentalis*, * 9. *townsendi*, * 10. *magnolia*, 11. *tigrina*, 12. *castanea*, 13. *dominica*, 14. *palmarum*, 15. *citrina*, 16. *americana*, * 17. *pitiayumi*, * 18. *cerulea*, 19. *pinus*, 20. *kirtlandii*, 21. *fusca*, 22. *coronata*, 23. *caerulescens*, 24. *ruticilla*).

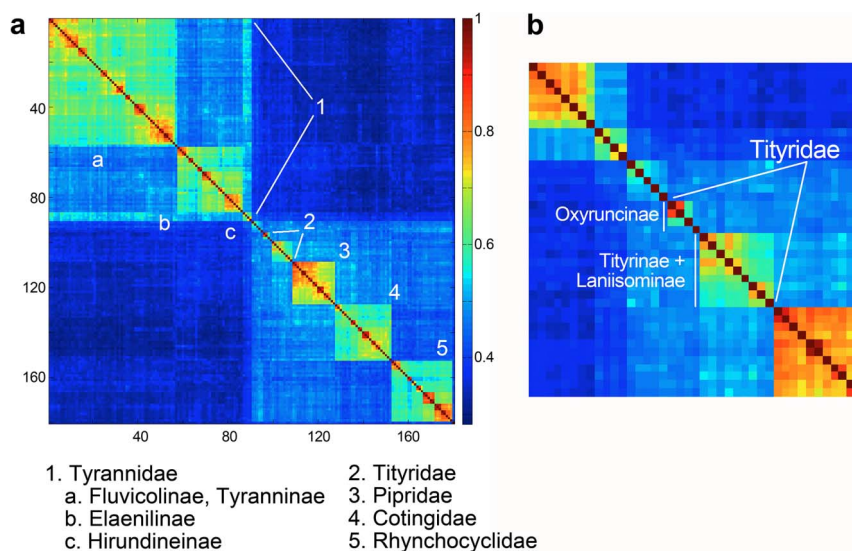


Figure 2 | Higher-level clusters in suborder Tyrannides (tyrannid flycatchers and allies) nuclear gene RAG-1. a) Klee diagram generated from TreeParser-ordered alignment ($n = 180$) is shown. Sequence clusters visible as blocks of high correlation along the diagonal correspond to taxonomic groups listed at bottom. b) Klee detail showing Tityridae and subfamilies.

indicator vector correlations. The repeated finding of coherent clusters in NJ-ordered Klee diagrams supports this approach.

An alternative to TreeParser is available in SeaView sequence analysis software⁴², which includes a utility that reorders a FASTA file according to a phylogenetic tree. For persons familiar with SeaView, this may be an attractive option. Advantages to TreeParser are that it is designed to work with the widely-used MEGA software and the stand-alone web version requires no additional software installation.

We applied the TreeParser-indicator vector-Klee pipeline to mitochondrial and nuclear genes from invertebrate and vertebrate species. In each case there were strong congruences between clusters and taxonomic groups. The skipper butterfly *A. fulgerator* COI Klee displayed eight of the ten putative species as distinct

blocks (Fig. 1), a visual representation of the typically shallow evolutionary histories within animal species as compared to greater distances among even close relatives^{43–45}. A large set of closely-related *Setophaga* warbler species formed similarly distinct blocks in the COI Klee (Fig. 1). In tyrannid flycatchers, the nuclear RAG-1 Klee discontinuities corresponded to recently revised family and subfamily groups (Fig. 2), providing a condensed snapshot of higher-level phylogeny³³. With a broader set of avian species, a RAG-1 Klee vividly displayed major taxonomic divisions of birds (Fig. 3). A COI Klee generated from the same set of species demonstrated congruent blocks of high correlation, although less strongly demarcated. Applied to butterfly COI and nuclear EF-1 sequence alignments, Klee diagrams revealed families and subfamilies as distinct blocks (Fig. 4).

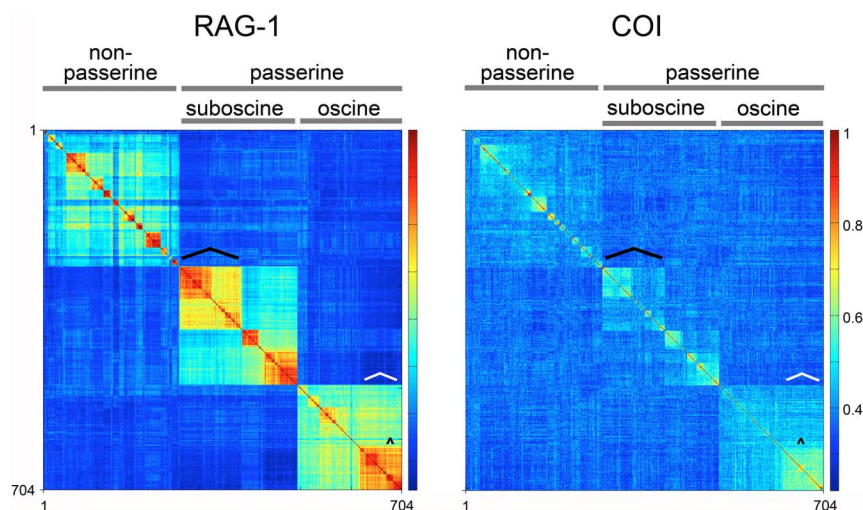


Figure 3 | Comparison of higher-order avian taxonomic clusters in RAG-1, COI. TreeParser-ordered Klee diagrams for bird species with both RAG-1 and COI barcode sequences are shown ($n = 704$). To facilitate comparison, RAG-1 Klee was rotated to more closely match arrangement of species in COI diagram. Major taxonomic divisions are labeled at top. Large and small black brackets indicate positions of Tyrannides (cf. Fig. 2a) and Parulidae wood warblers including *Setophaga* spp. (cf. Fig. 1), respectively. White bracket at lower right of each diagram indicates position of the multi-family New World songbird radiation informally referred to as “nine-primaried oscines^{45,46}.”

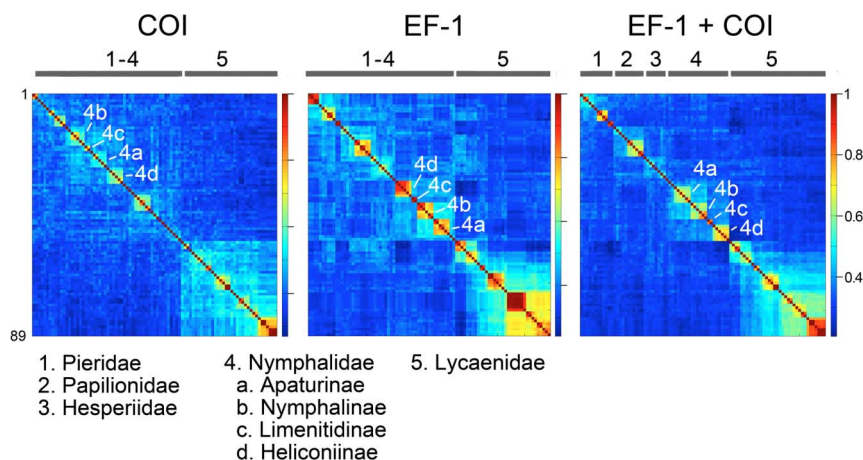


Figure 4 | Butterfly family and subfamily clusters in mitochondrial COI and nuclear EF-1. TreeParser-ordered Klee diagrams representing five of seven butterfly families are shown ($n = 89$ species). Each Klee follows the NJ tree for that dataset; EF-1 and EF-1 + COI Klees are rotated to more closely match the order in COI Klee. Bar at top indicates positions of families in each diagram and selected clusters representing Nymphalidae subfamilies are marked. Correlation scale is at right and taxonomic groups are listed at bottom.

In addition to congruences, differences between Klee clusters and named taxonomic divisions suggest possible areas that could benefit from further attention (e.g. Fig. 2b). Indicator vector-Klee analysis may point to groups meriting formal taxonomic names, such as the New World passerine radiation of “nine-primaried oscines”^{46,47}, which appeared as a densely correlated block in both RAG-1 and COI Klees (Fig. 3).

Several limitations to this analytic approach were encountered. The initial version of TreeParser had difficulty finding unique IDs in some files, reflecting the diversity of sequence headers. To circumvent this problem we modified the program and web portal, adding an option of using the entire sequence header as an identifier. Regarding indicator vector analysis, alignments with large gaps or numbers of missing characters produced distorted Klee diagrams. This was addressed by filtering alignments for full-length sequences and setting indicator vector bp parameters to exclude regions with missing data. It should be noted that all datasets in this study were protein coding regions. It may be of interest to test this approach on alignments of introns, ribosomal genes, or other non-coding sequences that contain gaps.

More generally, although not relevant to above examples, we encountered limitations to analyzing large files at multiple steps in the pipeline: alignment, tree generation, and indicator vector-Klee analysis. The computing challenges to generating alignments and phylogenetic trees for large datasets are well known (e.g.⁴⁸). Using higher capacity hardware we have been able to generate phylogeny-informative Klee diagrams for alignments as large as 11,000 sequences (Supplementary Fig. S3). It should be noted that TreeParser sorted this relatively large dataset on our standard server without difficulty.

Although it is possible to construct an accurate evolutionary branching diagram for just a few taxa, clustering is likely evident only if many closely related organisms are analyzed. DNA barcode libraries are an attractive resource given the breadth of taxonomic coverage. Drawbacks are reliance on a single gene and the paucity of phylogenetic signal in short mitochondrial DNA sequences^{36–38,49}. In this study, higher-level COI clusters were concordant with those in nuclear or combined gene analysis and with established taxonomy (Figs. 2–4; see also^{22,24}). These results suggest DNA barcode Klee analysis could help establish a taxonomic framework, which even if incomplete, could be useful particularly for groups less well known than butterflies or birds. It should be straightforward to test if these findings are generally applicable by analyzing other animal groups

with large datasets of mitochondrial and nuclear genes in GenBank or Barcode of Life Datasystems (BOLD)⁵⁰. Unlike animals, green plants (Viridiplantae) do not show strong intraspecific clustering in organellar genes including the standard plant barcode loci, *rbcL* and *matK*^{51,52}. Given this apparent dichotomy, it would be of interest to apply TreeParser-indicator vector-Klee analysis to examine higher-level structure in land plants.

The present findings support the re-emerging view that clustering is a widespread evolutionary pattern not limited to species-level differences. For example, Barraclough and colleagues recently proposed that that higher-level diversity is comprised of “evolutionary significant units worthy of scientific study” and put forth a mechanism by which such units could arise⁵³. However to date there is no broadly-applicable method other than expert opinion to define clusters above species level and thus a lack of objective data for model testing. Our results demonstrate indicator vector-Klee heat map analysis delineates higher-level structure in nucleotide sequence alignments. Analyzing additional datasets as outlined above will help determine the generality of clustering and the utility of this approach in investigating underlying mechanisms.

In summary, TreeParser-indicator vector-Klee software visualizes evolutionary clusters in nucleotide sequence datasets. This approach provides a condensed snapshot of a sequence alignment and should help investigate the structure of higher-level diversity which is not well understood.

Methods

Datasets. DNA barcode sequences were downloaded from BOLD project “EPAF *Astrartes fulgerator* complex”^{28,50}. Sequences were aligned with MUSCLE in MEGA and trimmed to include 648 base pair (bp) corresponding to nucleotides 52–699 of mouse mitochondrial genome^{40,54}. Those representing the ten putative species and containing at least 600 bp (positions 42 to 642) were selected for further analysis ($n = 420$). The sequence alignment and a MEGA-generated Kimura-2-parameter (K2P) neighbor-joining (NJ) tree file in text format were uploaded to TreeParser, producing an output FASTA file that followed the order of terminals in the tree. A Klee diagram was generated by indicator vector analysis with parameters $n = 1$ sequence/vector and $bp\ window = 42–642$.

Setophaga warbler COI DNA barcode sequences were downloaded from GenBank using search terms “setophaga[organism] AND BARCODE[keyword]”, aligned in MEGA, and trimmed to COI barcode region as described above. Sequences containing at least positions 100–600 were selected for further analysis ($n = 276$). A K2P NJ tree text file and FASTA alignment were uploaded to TreeParser, and the re-ordered alignment was used generate a Klee diagram, with parameters $n = 1$ sequence/vector and $bp\ window = 100–600$.

RAG-1 sequences from suborder Tyrannides (tyrant flycatchers, cotingas, manakins, and their allies)³³ were downloaded from GenBank PopSet and aligned in MEGA using MUSCLE ($n = 180$). The alignment contained 1,183 variable and 1,689



conserved positions. To facilitate desktop indicator vector analysis, conserved positions were deleted using MEGA export function. The condensed alignment was reordered with TreeParser according to a K2P NJ text file as described above. A Klee diagram was generated with parameters $n = 1$ sequence/vector and bp window = 1–1183.

To compare clustering in avian RAG-1 and COI, all avian RAG-1 sequences in GenBank (search terms “aves[organism] AND (rag-1[gene name] OR rag1[gene name])”) were downloaded and aligned in MEGA using MUSCLE. These were filtered to exclude short sequences, multiple sequences per species, conserved positions as described above, and positions with gaps in more than 90% of sequences. The resulting alignment contained 595 bp. Sequences from those species also represented in a published avian COI BARCODE dataset⁵⁵ were selected for further analysis, as were the corresponding COI BARCODEs ($n = 704$). K2P NJ tree files produced in MEGA and their respective alignments were uploaded to TreeParser. Klee diagrams were generated from rearranged FASTA files with parameters $n = 1$ sequence/vector, and bp window = 1–595 (RAG-1) or 100–600 (COI).

To examine higher-level patterns in butterfly genes, datasets of EF-1 (1066 bp) and COI (1101 bp) sequences ($n = 89$ sequences, 1 per species)³⁹ were downloaded from GenBank PopSet, aligned in MEGA, and used to generate TreeParser-ordered Klee diagrams. For combined analysis, a FASTA file of concatenated EF-1 and COI sequences was condensed by removing invariant positions as described above (final size 864 bp). A Klee diagram was generated from the TreeParser re-ordered alignment with bp window = 1–864.

TreeParser software. TreeParser is designed to work with FASTA files downloaded from GenBank or BOLD and with phylogenetic tree text files generated by MEGA. Programming language PHP version 5 was chosen for web compatibility and ease of use. The software and step-by-step instructions on running TreeParser and generating Klee diagrams are posted on the web at <http://phe.rockefeller.edu/barcode/klee.php>. The web version, hosted on a Linux server running Apache at the address above, requires no additional software. The source code, designed to be downloaded and run locally, is available at http://phe.rockefeller.edu/barcode/klee_sourcecode/tree_parser.tar.gz.

TreeParser accepts two files: a FASTA-formatted alignment of nucleotide sequences and a text format tree file generated from the alignment using MEGA.

Once files are uploaded, TreeParser performs the following algorithm:

1. Search the tree and FASTA files for the unique ID of each nucleotide sequence (represented as a regular expression).
2. Obtain two lists, using the unique ID of the particular sequence as the index of each fragment:
 - a. Tree list: An ordered list of all sequences in the template tree text file.
 - b. FASTA list: A list of nucleotide sequences constructed by splitting up the FASTA file into blocks. Each block consists of a unique header and its sequence.
3. Loop through the Tree list and search the FASTA list for each entry.
4. Construct an array consisting of reordered FASTA blocks corresponding to the order of the Tree list file.
5. Check the new array for any missing values from either the Tree or FASTA list.
6. Write the new array to an output FASTA file.
7. Generate a secondary log file.

The output FASTA file is identical in content to the original, but reordered in accordance with the template tree. This file can then be passed directly into indicator vector software to construct a Klee diagram. The output log file records the number of matched sequences and lists any missing values from the FASTA or Tree lists.

Indicator vector analysis. This was performed as described²¹ using updated software available at http://phe.rockefeller.edu/barcode/klee_sourcecode/Indicator_Vector_Klee_v1.tar.gz.

Computer hardware. MEGA (nucleotide sequence alignment, neighbor-joining) and MATLAB 2009a (indicator vector-Klee) analyses were performed on Mac Mini desktop (Mac OSX 10.7.4, 2.5 GHz Intel Core i5 processor, 8 GB RAM).

1. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090 (1977).
2. Pace, N. R. Mapping the Tree of Life: progress and prospects. *Microbiol. Mol. Biol. Rev.* **73**, 565–576 (2009).
3. Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. The great human expansion. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17758–17764 (2012).
4. Hillis, D. M., Moritz, C. & Mable, B. K. *Molecular Systematics, Second Edition* (Sinauer Associates, Sunderland, 1996).
5. Nei, M. & Kumar, S. *Molecular Evolution And Phylogenetics* (Oxford University Press, New York, 2000).
6. Pons, J. *et al.* Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.* **55**, 595–609 (2006).
7. Archer, J. & Robertson, D. CTree: comparison of clusters between phylogenetic trees made easy. *Bioinformatics* **23**, 2952–2953 (2007).
8. Prosperi, M. C. F. *et al.* A novel methodology for large-scale phylogeny partition. *Nat. Commun.* **2**, 321 (2011).
9. Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M. & Holmes, C. Phylogeny and the origin of HIV-1. *Nature* **410**, 1047–1048 (2001).
10. Wheeler, W. C. *Systematics: A Course Of Lectures* (Wiley-Blackwell, Oxford, 2012).
11. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer Associates, Sunderland, 2004).
12. Wilkinson, L. & Friendly, M. The history of the cluster heat map. *Am. Stat.* **63**, 179–184 (2009).
13. Weinstein, J. N. A postgenomic visual icon. *Science* **319**, 1772–1773 (2008).
14. Gove, R. *et al.* NetVisia: Heat map & matrix visualization of dynamic social network statistics & content. *SocialCom/PASSAT* **2011**, 19–26 (2011).
15. Chiaretti, S. *et al.* Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* **103**, 2771–2778 (2004).
16. Elashoff, M. R. *et al.* Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients. *BMC Med. Genomics* **4**, 26 (2011).
17. Ahfeldt, T. *et al.* Programming human pluripotent stem cells into white and brown adipocytes. *Nat. Cell Biol.* **14**, 209–219 (2012).
18. Yu, Q. *et al.* Modular organization of functional network connectivity in healthy controls and patients with schizophrenia during the resting state. *Front. Syst. Neurosci.* **5**, 103 (2012).
19. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2013).
20. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2013).
21. Sirovich, L., Stoeckle, M. Y. & Zhang, Y. A scalable method for analysis and display of DNA sequences. *PLoS ONE* **4**, e7051 (2009).
22. Sirovich, L., Stoeckle, M. Y. & Zhang, Y. A structural analysis of biodiversity. *PLoS ONE* **5**, e2966 (2010).
23. Bucklin, A., Steinke, D. & Blanco-Bercial, L. DNA barcoding of marine metazoa. *Ann. Rev. Mar. Sci.* **3**, 471–508 (2011).
24. Bucklin, A. *et al.* A census of zooplankton of the global ocean. pp. 247–266 In McIntyre, A. ed., *Life In The World's Oceans*, Wiley-Blackwell (2010).
25. Costa, F. O. *et al.* A ranking system for reference libraries of DNA barcodes: application to marine fishes from Portugal. *PLoS ONE* **7**, e35858 (2012).
26. Raupach, M. J. *et al.* Molecular identification of Central European ground beetles (Coleoptera: Carabidae) using nuclear rDNA expansion segments and DNA barcodes. *Front. Zool.* **7**, 26 (2010).
27. Baum, D. Reading a phylogenetic tree: the meaning of monophyletic groups. *Nat. Educ.* **1**, 41956 (2008).
28. Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astrartes fuligator*. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 14812–14817 (2004).
29. Nielsen, R. & Matz, M. Statistical approaches for DNA barcoding. *Syst. Biol.* **55**, 162–169 (2006).
30. Brower, A. V. Z. Problems with DNA barcodes for species delimitation: ‘ten species’ of *Astrartes fuligator* reassessed (Lepidoptera: Hesperidae). *Syst. Biodivers.* **4**, 127–132 (2006).
31. Lovette, I. J. & Bermingham, E. Explosive speciation in the New World *Dendroica* warblers. *Proc. R. Soc. B* **266**, 1629–1636 (1999).
32. Lovette, I. J. & Bermingham, E. Mitochondrial perspective on the phylogenetic relationships of the *Parula* wood-warblers. *Auk* **118**, 211–215 (2001).
33. Tello, J. G., Moyle, R. G., Marchese, D. J. & Cracraft, J. Phylogeny and phylogenetic classification of the tyrant flycatchers, cotingas, and their allies (Aves: Tyrannidae). *Cladistics* **25**, 429–467 (2009).
34. Dickinson, E. C. ed. *Howard And Moore Complete Checklist Of The Birds Of The World, Third Edition*. Princeton: Princeton University Press. 1056 p. (2003).
35. Clements, J. F. *The Clements Checklist Of Birds Of The World, Sixth Edition*. Ithaca: Comstock Publishing Associates. 864 p. (2007).
36. Ballard, W. O. & Rand, D. M. The population biology of mitochondrial DNA and its phylogenetic implications. *Ann. Rev. Ecol. Syst.* **36**, 621–642 (2005).
37. Hajibabei, M., Singer, G. A. C. & Hickey, D. A. Benchmarking DNA barcodes: an assessment using available primate sequences. *Genome* **49**, 851–854 (2011).
38. Waters, J. M., Rowe, D. L., Burrridge, C. P. & Wallis, G. P. Gene trees versus species trees: reassessing life-history evolution in a freshwater fish radiation. *Syst. Biol.* **59**, 504–517 (2010).
39. Kim, M. I. *et al.* Phylogenetic relationships of true butterflies (Lepidoptera: Papilionoidea) inferred from COI, 16S rRNA and EF-1 α sequences. *Mol. Cells* **30**, 409–435 (2010).
40. Kumar, S., Tamura, K. & Nei, M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5**, 150–163 (2004).
41. Olsen, G. Newick tree format standard. (1990) http://evolution.genetics.washington.edu/phylib/newick_doc.html. Accessed March 5, 2013.
42. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
43. Brown, W. M., George, M. & Wilson, A. C. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 1967–1971 (1979).



44. Avise, J. C. *et al.* Intraspecific phylogeography: the mitochondrial bridge between population genetics and systematics. *Ann. Rev. Ecol. Syst.* **18**, 489–522 (1987).
45. Moore, W. S. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution* **49**, 718–726 (1995).
46. Klicka, J., Johnson, K. P. & Lanyon, S. M. New World nine-primaried oscine relationships: constructing a mitochondrial DNA framework. *Auk* **117**, 321–336 (2000).
47. Barker, F. K., Cibois, A., Schikler, P., Feinstein, J. & Cracraft, J. Phylogeny and diversification of the largest avian radiation. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11040–11045 (2004).
48. Sanderson, M. J. & Driskell, A. C. The challenge of constructing large phylogenetic trees. *Trends Plant Sci.* **8**, 374–379 (2003).
49. Springer, M. S. *et al.* Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. *Mol. Biol. Evol.* **18**, 132–143 (2001).
50. Ratnasingham, S. & Hebert, P. D. N. BOLD: The Barcode of Life Data System. *Mol. Ecol. Notes* **7**, 355–364 (2007).
51. Hollingsworth, P. M., Graham, S. W. & Little, D. A. Choosing and using a plant DNA barcode. *PLoS ONE* **6**, e19254 (2011).
52. Stoeckle, M. Y. *et al.* Commercial teas highlight plant DNA barcode identification successes and obstacles. *Sci. Rep.* **1**, 42; DOI:10.1038/srep00042 (2011).
53. Barraclough, T. G. Evolving entities: towards a unified framework for understanding diversity at the species and higher levels. *Phil. Trans. R. Soc. Lond. B* **365**, 1801–1813 (2010).
54. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792–1797 (2004).
55. Stoeckle, M. Y. & Kerr, K. C. R. Frequency matrix approach demonstrates high sequence quality in avian BARCODEs and highlights cryptic pseudogenes. *PLoS ONE* **78**, e43992 (2012).

Acknowledgments

We thank Jesse Ausubel for helpful discussions.

Author contributions

M.Y.S. designed the study, C.C. wrote the computer software; M.Y.S. and C.C. wrote the manuscript.

Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Stoeckle, M.Y. & Coffran, C. TreeParser-Aided Klee Diagrams Display Taxonomic Clusters in DNA Barcode and Nuclear Gene Datasets. *Sci. Rep.* **3**, 2635; DOI:10.1038/srep02635 (2013).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>