WILEY-BLACKWELL

Cladistics 27 (2011) 171–180

Cladistics

10.1111/j.1096-0031.2010.00329.x

SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information

Gaurav Vaidya[1], David J. Lohman[1,†] and Rudolf Meier[2,*]

[1]*Department of Biological Sciences and* [2]*University Scholars Programme, National University of Singapore, 14 Science Drive 4, Singapore 117543, Singapore*

Accepted 29 April 2010

Abstract

We present SequenceMatrix, software that is designed to facilitate the assembly and analysis of multi-gene datasets. Genes are concatenated by dragging and dropping FASTA, NEXUS, or TNT files with aligned sequences into the program window. A multi-gene dataset is concatenated and displayed in a spreadsheet; each sequence is represented by a cell that provides information on sequence length, number of indels, the number of ambiguous bases (''Ns''), and the availability of codon information. Alternatively, GenBank numbers for the sequences can be displayed and exported. Matrices with hundreds of genes and taxa can be concatenated within minutes and exported in TNT, NEXUS, or PHYLIP formats, preserving both character set and codon information for TNT and NEXUS files. SequenceMatrix also creates taxon sets listing taxa with a minimum number of characters or gene fragments, which helps assess preliminary datasets. Entire taxa, whole gene fragments, or individual sequences for a particular gene and species can be excluded from export. Data matrices can be re-split into their component genes and the gene fragments can be exported as individual gene files. SequenceMatrix also includes two tools that help to identify sequences that may have been compromised through laboratory contamination or data management error. One tool lists identical or near-identical sequences within genes, while the other compares the pairwise distance pattern of one gene against the pattern for all remaining genes combined. SequenceMatrix is Java-based and compatible with the Microsoft Windows, Apple MacOS X and Linux operating systems. The software is freely available from http://code.google.com/p/sequencematrix/.

Modern phylogenetic analyses typically infer relationships using multi-gene datasets. Many software packages are capable of concatenating individual character and gene files into such sets (e.g. Maddison and Maddison, 2001, 2009; Jones and Blaxter, 2006; Roure et al., 2007; Goloboff et al., 2008; Smith and Dunn, 2008), but the concatenation tools are generally not particularly user-friendly, often do not preserve character set or codon position information, have limitations on the number of partitions that can be concatenated, and/or make it difficult for the user to check for concatenation errors. This has led to the undesirable situation that many multi-locus datasets are only assembled toward the end of a project although evaluating preliminary datasets is important for exploring their phylogenetic signal, assessing the effects of missing data, monitoring the progress of a project and/or identifying sequences that may have been compromised through laboratory contamination.

We introduce the concatenation tool SequenceMatrix, which has the following desirable properties. (i) Concatenation is fast and intuitive, so that even very large datasets with several hundred taxa and genes can be assembled quickly. The graphical user interface allows users to check whether the concatenation captured all input data. (ii) Information from the individual gene files such as character set ranges and codon positions are preserved and exported with the concatenated matrix.

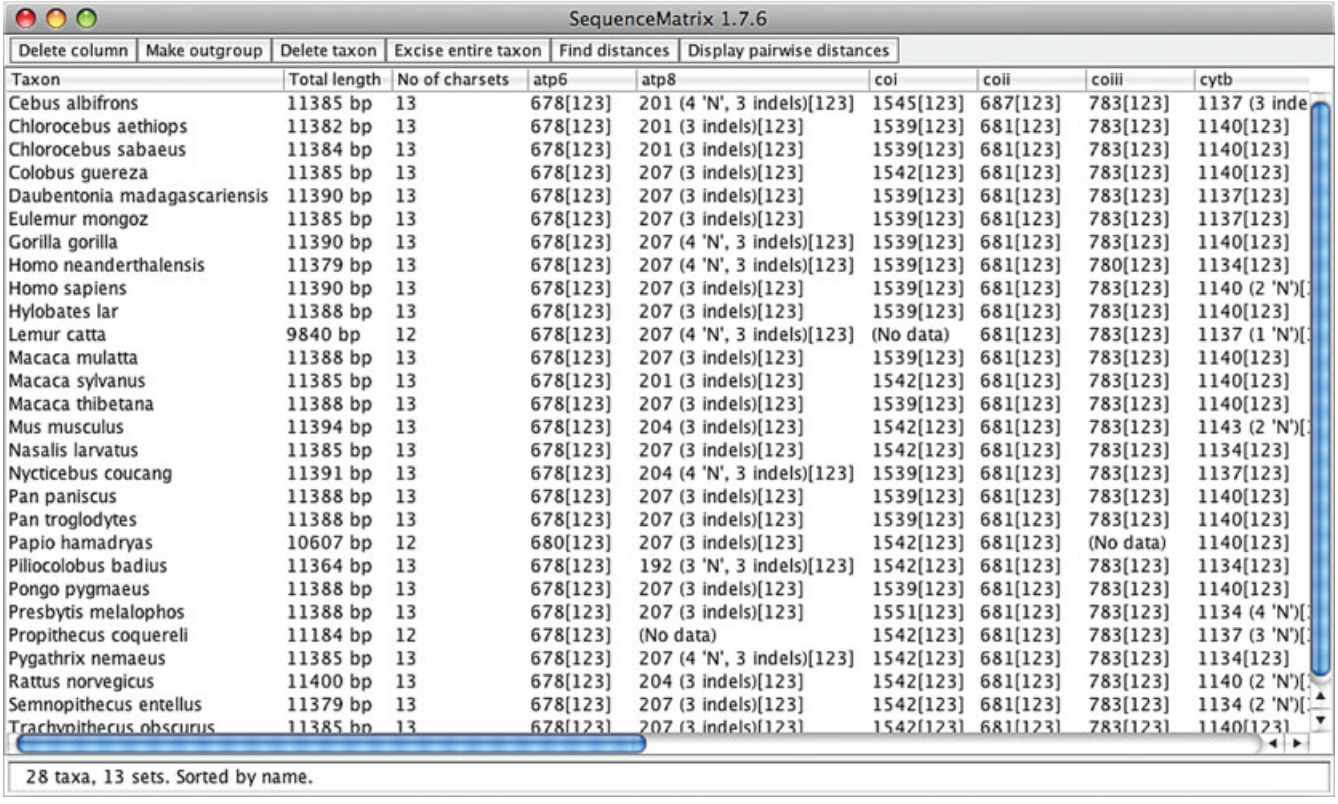
*Corresponding author:
E-mail address: meier@nus.edu.sg
†Present address: Department of Biology, The City College of New York, The City University of New York, Convent Avenue at 138th Street, New York, NY 10031, USA.

(iii) The program generates taxon and character sets according to the user's specifications. (iv) Users can exclude individual sequences from the export so that their effects on phylogenetic signal can be studied. (v) SequenceMatrix can generate a data summary and GenBank tables for use in publications. (vi) The program allows "reverse-splitting" a concatenated character matrix back into its components. (vii) SequenceMatrix assists with the discovery of unusually similar or divergent sequences that may be the result of laboratory contamination or errors in data management.

Fast and intuitive concatenation with a graphical user interface

Concatenation in SequenceMatrix proceeds by dragging and dropping input files into an overview window or manually importing files ("Import" → "Add sequences"). The input files must contain aligned sequences in FASTA, NEXUS (Maddison et al., 1997), or TNT (Goloboff et al., 2008) format. The data are displayed in a spreadsheet-like grid, allowing the user to check for concatenation errors. Taxa are displayed in rows and genes in columns (Fig. 1). Rather than displaying the actual data in each cell, the cells indicate whether codon information is available (symbol: "[N123]"), and list sequence length, number of indels, and number of ambiguous bases ("N" in IUPAC) for each sequence (Fig. 1). This information can be exported ("Export" → "Export table as tab-delimited"), which helps to flag sequences that require closer inspection. For example, poor quality sequences often have a large number of of ambiguous bases ("Ns"), misaligned protein-encoding sequences have a large number of indels that are not multiples of three, and the alignment of a gene with a large number of indels may require closer scrutiny. Two additional columns in the spreadsheet specify the number of available character sets and the total number of available nucleotides for each taxon. The list of taxa can be sorted by name, species epithet, number of character sets with data, or total length of available sequence ("Sequences" → "Sort by name", etc.). In addition, one taxon can be designated as the reference taxon; it is then positioned and exported as the first taxon in the matrix (right-click on taxon name → "Make this row the reference taxon").

Using this approach, SequenceMatrix can be used to assemble very large data matrices in relatively little time. In preparation for the release of the software, we tested the program extensively. We assembled all molecular



Taxon	Total length	No of charsets	atp6	atp8	coi	coii	coiii	cytb
Cebus albifrons	11385 bp	13	678[123]	201 (4 'N', 3 indels)[123]	1545[123]	687[123]	783[123]	1137 (3 inde
Chlorocebus aethiops	11382 bp	13	678[123]	201 (3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Chlorocebus sabaeus	11384 bp	13	678[123]	201 (3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Colobus guereza	11385 bp	13	678[123]	207 (3 indels)[123]	1542[123]	681[123]	783[123]	1140[123]
Daubentonia madagascariensis	11390 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1137[123]
Eulemur mongoz	11385 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1137[123]
Gorilla gorilla	11390 bp	13	678[123]	207 (4 'N', 3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Homo neanderthalensis	11379 bp	13	678[123]	207 (4 'N', 3 indels)[123]	1539[123]	681[123]	780[123]	1134[123]
Homo sapiens	11390 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1140 (2 'N')
Hylobates lar	11388 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Lemur catta	9840 bp	12	678[123]	207 (4 'N', 3 indels)[123]	(No data)	681[123]	783[123]	1137 (1 'N')
Macaca mulatta	11388 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Macaca sylvanus	11385 bp	13	678[123]	201 (3 indels)[123]	1542[123]	681[123]	783[123]	1140[123]
Macaca thibetana	11388 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Mus musculus	11394 bp	13	678[123]	204 (3 indels)[123]	1542[123]	681[123]	783[123]	1143 (2 'N')
Nasalis larvatus	11385 bp	13	678[123]	207 (3 indels)[123]	1542[123]	681[123]	783[123]	1134[123]
Nycticebus coucang	11391 bp	13	678[123]	204 (4 'N', 3 indels)[123]	1539[123]	681[123]	783[123]	1137[123]
Pan paniscus	11388 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Pan troglodytes	11388 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Papio hamadryas	10607 bp	12	680[123]	207 (3 indels)[123]	1542[123]	681[123]	(No data)	1140[123]
Piliocolobus badius	11364 bp	13	678[123]	192 (3 'N', 3 indels)[123]	1542[123]	681[123]	783[123]	1134[123]
Pongo pygmaeus	11388 bp	13	678[123]	207 (3 indels)[123]	1539[123]	681[123]	783[123]	1140[123]
Presbytis melalophos	11388 bp	13	678[123]	207 (3 indels)[123]	1551[123]	681[123]	783[123]	1134 (4 'N')
Propithecus coquereli	11184 bp	12	678[123]	(No data)	1542[123]	681[123]	783[123]	1137 (3 'N')
Pygathrix nemaeus	11385 bp	13	678[123]	207 (4 'N', 3 indels)[123]	1542[123]	681[123]	783[123]	1134[123]
Rattus norvegicus	11400 bp	13	678[123]	204 (3 indels)[123]	1542[123]	681[123]	783[123]	1140 (2 'N')
Semnopithecus entellus	11379 bp	13	678[123]	207 (3 indels)[123]	1542[123]	681[123]	783[123]	1134 (2 'N')
Trachypithecus obscurus	11385 bp	13	678[123]	207 (3 indels)[123]	1542[123]	681[123]	783[123]	1140[123]

Fig. 1. SequenceMatrix interface showing summaries of sequence length, character sets, codon positions, number of indels, ambiguous bases (Ns), and total sequence length.

data from our recent phylogenetic analyses into a single matrix with the codon information included for the protein-encoding genes (Meier and Baker, 2002; Meier and Wiegmann, 2002; Laamanen et al., 2005; Kutty et al., 2007, 2008, 2010; Petersen et al., 2007; Su et al., 2007, 2008; Yeo et al., 2007; Ang et al., 2008; Lohman et al., 2008, 2010; Puniamoorthy et al., 2008, 2009; Balke et al., 2009; Huang et al., 2009; Wowor et al., 2009; Lim et al., 2010). Concatenation took 3 min on a standard desktop computer running Windows XP and yielded a matrix with 715 taxa and 81 050 characters in 92 character sets (48 sets with codon information). We successfully exported this matrix into NEXUS and TNT formats. We then added two unpublished datasets (including one supermatrix) to yield a matrix with 1011 taxa and 185 character sets (211 807 characters), which was also exported successfully. Lastly, we added 43 062 COI sequences (Meier et al., 2008) to yield a matrix with 44 073 taxa and 186 character sets. The matrix could be exported, but due to its size subsequent analyses could not be executed on a regular desktop computer in either TNT (Goloboff et al., 2008) or PAUP* (Swofford, 2004).

Additional information on import files

SequenceMatrix uses the names of the input files as the name of the data partitions. These names are displayed in the column header and used as the character set name (Fig. 1). In the spreadsheet, the gene partitions are arranged in alphabetical order. The gene order can be changed by renaming the input files prior to import (e.g. by adding numbers in front of the gene names). SequenceMatrix accepts aligned sequences in FASTA, NEXUS, and TNT formats as exported from sequence editing and alignment software such as Sequencher, MAFFT, Muscle, SeaView, ClustalW, Mesquite, and MacClade. Any number of different formats can be used for the concatenation of a single matrix. However, in order to avoid compatibility problems (e.g. NEXUS files exported from SeaView), we recommend using simple formats such as aligned FASTA, although most NEXUS and TNT files will also import without problems.

SequenceMatrix can process codon information for protein-encoding genes and export it in the concatenated matrix as long as NEXUS's "BEGIN CODONS" command (CodonPosSet) is used to specify the positions in the input files (do not place this information in the ASSUMPTIONS block and only include one CodonPosSet; see sample files). Input files with codon positions can be obtained by using either MacClade (Maddison and Maddison, 2001) or Mesquite (Maddison and Maddison, 2009) and files generated by these software packages are generally compatible with SequenceMatrix. If problems arise, they can be resolved by exporting the file in Mesquite as "Simplified NEXUS"; in this case, the codon information must be manually added using the "addendum to file" export box of Mesquite. Once the codon information has been successfully parsed by SequenceMatrix, the data cells will show the symbol [N123].

SequenceMatrix is derived from the SpeciesIdentifier software package, which directly reads sequences in FASTA format and was designed to capture sequence data without additional information (Meier et al., 2006). The software thus interprets very little extraneous information that may be included in an input file. For example, in a typical NEXUS file, SequenceMatrix finds the "DATA" block and skips commands until it finds "MATRIX". It will assume that the data starts here and will parse the data until it encounters a ";" (note that unusual commands placed between the "DATA" and "MATRIX" may cause error messages). SequenceMatrix will furthermore parse the information from the character and codon set commands; additional information and character sets placed in the "ASSUMPTIONS" or other blocks are ignored. Similarly, SequenceMatrix identifies the data block in TNT files using "xread" and ";" and only the data and information in the "xgroup" and "agroup" are parsed. In FASTA files, the ">" sign is interpreted as a delimiter between taxon-specific sequences. Should SequenceMatrix fail to parse a particular input file, a warning message is displayed. Concatenation of the remaining input files will proceed, but the data from the failed file will be missing. These can then be added after resolving the compatibility issue.

Matching taxon names across input files

Concatenation requires that SequenceMatrix matches taxon names across gene files because it needs to identify which data from gene A should be concatenated with which data for gene B. SequenceMatrix has two taxon-name matching modes. For those input files for which both modes can be used it will ask the user to choose between "sequence name" and "species name" after parsing the data. The "sequence name" mode uses the full sequence name for concatenation. Users who would like to use this mode need to use identical taxon names or taxon abbreviations in all input files, i.e. the sequence names should not include additional information (e.g. gene names). The other name-matching mode in SequenceMatrix ("species name") is designed for the assembly of matrices that include GenBank data. In this mode, SequenceMatrix identifies species names using the following method. It searches for a string of two words in the sequence title that entirely consists of letters. The first word must be capitalized and followed by a space, while the second word has to be entirely composed of lower-case letters. In this parsing mode, a complex

sequence name such as "> gi|62 946 235|gb|AY972793.1| Daubentonia madagascariensis isolate PR01017 cytochrome c oxidase subunit I (COXI) gene, partial cds; mitochondrial" is simplified to *Daubentonia madagascariensis*. If a second file including 16S data is added, SequenceMatrix will identify that "> gi|3 818 542|gb|AF072414.1|AF072414 Daubentonia madagascariensis 16S ribosomal RNA gene, mitochondrial gene for mitochondrial RNA, partial sequence" also belongs to *Daubentonia madagascariensis* and the two sequences will be concatenated under this name. Users who would like to use this name parsing mode should ensure that no taxon name includes numbers, underscores, or other atypical characters. Note also that under the "species name" mode sequences for different subspecies will be fused under one species name (see PrimatesGenbank example).

The different parsing modes can yield very different matrices. For example, in the previously mentioned COI dataset with 43,062 sequences (Meier et al., 2008), each sequence has a unique name including its GenBank number. Accordingly, the matrix will include 43 062 taxa if the "sequence name" parsing mode is used. However, only 13 068 species are represented by these 43 062 sequences, i.e. using the "species name" mode a matrix is generated that has 13 068 taxa. Each species will be represented by only one sequence and SequenceMatrix will by default only retain the longest sequence available for each species.

Changing taxon names

Different spellings for the same taxon name in different gene files lead to concatenation errors, causing data from different genes for the same taxon to appear in different rows. With most concatenation software, the user has to inspect the concatenated matrix carefully and correct spelling errors in the input gene files before re-concatenating. In SequenceMatrix, spelling mistakes can be corrected by double-clicking on the cell displaying the name that needs correction. Once a spelling change renders the name identical to an existing name in another row of the matrix, the two rows are fused and the data are concatenated. A warning message alerts the user if fusing rows requires replacing data. Two examples for such "spelling mistakes" are included in the sample files for "PrimatesGenbank" included in the SequenceMatrix release. Here, the same species is either named *Pygathrix roxellana* or *Rhinopithecus roxellana*. The initial concatenation yields a matrix where the species is represented by two rows. Once *Pygathrix roxellana* is renamed to *Rhinopithecus roxellana* (or *vice versa*), SequenceMatrix combines the two rows into a single row. Spelling errors can be found quickly in SequenceMatrix because the taxa can be sorted by the amount of available data ("Sequences" → "Sort by number of character sets"); taxa with misspelled names will be data-deficient because they only have data from the input file with the misspelling.

Leading and trailing gaps ("terminal gaps")

The inappropriate coding of gaps at the beginning and end of sequences as data can lead to error when analysing concatenated datasets. For example, under its default settings, TNT will treat all gaps coded with a "–" as a 5th character state, and leading and trailing gaps will be interpreted as data if "–" is used. Sequence Matrix will therefore query the user whether leading and trailing gaps should be recoded as "?".

Quality control

Because data concatenation can easily result in errors, it is important that users scrutinize concatenated datasets for problems. One quality control strategy in SequenceMatrix is to compare the data statistics for a character set before and after concatenation. Statistics such as sequence length, number of constant and variable sites, and tree lengths should be identical for the input file and the same data within the concatenated matrix. To facilitate this check, SequenceMatrix automatically exports concatenated matrices with character sets and taxon sets for each gene. The user can then check the accuracy of the concatenation by comparing statistics obtained from a single gene input file with the same statistics obtained from the concatenated matrix after including only this gene in the analysis. The taxon sets that are created for each gene are particularly useful for this purpose. They allow users to exclude all taxa that lack data for the gene being tested. If these taxa are not excluded, gene trees may be difficult to find because species without data will behave like "wildcard taxa" (Nixon et al., 1992). A second option for quality control rests on the ability of SequenceMatrix to re-split concatenated matrices into their original character sets and to export them into individual gene files (see below). By using this feature, the user can again check whether the input and output files contain data with identical properties.

We urge all users to carry out the concatenation checks described in this section in case our testing overlooked a flaw in SequenceMatrix. Our tests were based on a variety of datasets that were assembled at very different times, using different concatenation techniques, and varied greatly in size (Meier and Baker, 2002; Meier and Wiegmann, 2002; Laamanen et al., 2005; Kutty et al., 2007, 2008, 2010; Petersen et al., 2007; Su et al., 2007, 2008; Yeo et al., 2007; Ang et al., 2008; Lohman et al., 2008, 2010; Puniamoorthy et al., 2008, 2009; Balke et al., 2009; Huang et al., 2009; Wowor et al., 2009; Lim et al., 2010). However, ultimately it

remains the responsibility of the user to test for correct concatenation. A relatively easy way to scrutinize a concatenated matrix is to use a batch file in, for example, PAUP* (Swofford, 2004) that records statistics such as the number of constant and variable characters in the input file and the same variables for the same character set embedded in the concatenated matrix. An example of such a batch file is included in the sample files provided with the program (''Scathophagidae_Concatenation_check_batch.nex''). Similar automated checks can be carried out in TNT, but they may require xgroup replacements due to limitations in the number of permissible groups (see below).

Amino acid and morphological data

SequenceMatrix was designed for sequence data and will export all matrices as if they contained DNA sequences. The concatenation of amino acid data is usually successful although it is still in the experimental stage and sample files are therefore not included in this release. We tested this feature by translating protein-encoding genes from our recent publications into amino acid sequences, assembling a concatenated data matrix, and confirming the accuracy of the matrix by comparing input files with the same data in the concatenated matrix. We also successfully assembled the Metazoa dataset of Dunn et al. (2008) using this procedure. In order to be imported into SequenceMatrix, amino acid sequences cannot have stop symbols and all polymorphisms must be re-coded as missing data. We alert the user to the preliminary nature of this feature by exporting the concatenated matrix as if it contained DNA; that is, before the matrix can be analysed, the data specification needs to be changed. The concatenation of morphological data is currently also still in the experimental phase and features such as polymorphism coding are not supported. Note that when morphological data are combined with DNA sequence data, the concatenated matrix will include IUPAC nucleotide ambiguity codes (i.e. K, R, W, Y, etc.) that need to be changed into polymorphisms through search and replace. Full support for amino acid and morphological data is planned for the near future but many morphological datasets can already be accommodated by re-coding the states using the symbols used for DNA sequences (ACTG).

Export of concatenated matrix with character sets and codon positions

SequenceMatrix will export concatenated matrices in TNT, NEXUS (interleaved, non-interleaved, ''naked'' for use in GARLI), and PHYLIP (using long taxon names for analyses with RAxML on CIPRES server) formats using the taxon and gene order in the program's user interface (''Export''). The taxon order can be changed using the sorting tool and the gene order manipulated by renaming gene files before concatenation. SequenceMatrix will also export codon information, and the software automatically re-calculates the character numbers for the character and codon sets as additional genes are concatenated. All taxon and character sets are available for export to NEXUS and TNT formats. If exported to a NEXUS file, the codon information is specified in the CODONS block where all first, second, and third positions are represented by one set respectively. Given that a user may want only to selectively include, for example, the third positions of COI, SequenceMatrix will also provide codon sets for each gene in the ASSUMPTION block (as COI_pos1, COI_pos2, COI_pos3). In TNT, the codon sets are specified as regular character sets by enumerating the characters that belong to the different positions. Note that TNT only supports directly up to 32 taxon and 32 character sets. Additional sets exported by SequenceMatrix are therefore placed as ''comments'' at the beginning of the matrix. In order to use these sets, one of the 32 active sets must be replaced. Note that this replacement should only involve the set, while the xgroup or agroup name should not be changed.

User-defined taxon sets

The desire to assemble multi-gene matrices quickly, even when only preliminary data were available for a project, largely motivated the creation of SequenceMatrix. Preliminary matrices are typically incomplete and analyses of such datasets require that the user exclude taxa that are very data-deficient. SequenceMatrix addresses this problem by creating taxon sets for taxa that meet minimum-data requirements set by the user (''Export'' → ''Taxonset settings''). For example, a user may want to test whether a tree based on only those taxa that already have data for five partitions or 5000 nt is stable and well supported. SequenceMatrix can be told to generate such a taxon set. Because concatenation is not time-consuming in SequenceMatrix, the user and/or principal investigator of a laboratory can follow the progress of a multi-locus sequencing project by getting an overview of sequencing progress through scrutinizing the phylogenetic signal in preliminary matrices or the export of a tab-delimited spreadsheet that specifies how much data are available for the different taxa (''Export'' → ''Export table as tab-delimited'').

Excluding data from export

Another feature that allows the user to explore preliminary data is the ''excise'' option. Currently,

exploring the effect of individual sequences on the result of phylogenetic analyses is difficult because most phylogenetic software only allows the exclusion of entire character sets or taxa; excluding a single sequence from analysis requires the removal of the sequence from the input file and subsequent re-assembly of the matrix. SequenceMatrix facilitates assessing the phylogenetic influence of single sequences by allowing their exclusion from exported files. Once the user double-clicks a cell in the user interface, the message ''EXCISED'' will appear and the cell will be coloured grey, indicating that this sequence has been selected for exclusion at the time of export. Once the matrix is exported, a warning message will alert the user of the impending exclusion. Excised sequences can be restored by double-clicking on the same cell. Using these tools, taxon and data subsets of a complex dataset can be rapidly exported in a variety of different formats to explore their influence on analysis results without editing the original input files.

Export of concatenated matrix as a spreadsheet with or without GenBank numbers

Checking on the progress of a multi-locus project is facilitated by allowing the user to export the spreadsheet overview as a table that reveals which sequences are still missing and provides quality information on existing sequences (sequence length, number of ''Ns''). Alternatively, SequenceMatrix can be used to generate a table with GenBank numbers (''Sequences'' → ''Sort by name, but display as GI numbers''; see PrimatesGenBank example). However, these numbers are only available when input files are used that still contain the GI numbers. This is not the case for concatenated matrices that have already been exported in NEXUS or TNT format, i.e. users who intend to use the GI table feature need to keep the input files with GI numbers. For newly generated sequences, the GenBank numbers should be downloaded once they become available. After the download, they can be concatenated in SequenceMatrix. Once concatenated, a table with all GenBank numbers can be exported. As the quality of the alignment does not matter for obtaining the GenBank table, any quick alignment procedure can be used.

Reverse-splitting concatenated matrices

Concatenated matrices can be split back into their component partitions by dragging a multi-gene dataset that contains character set information into SequenceMatrix. The software will then query whether a reverse-split is requested. After the split, the partitions are displayed as columns in the spreadsheet and the character set names are used as column titles. The reverse-split is fast for small datasets, but more time consuming for matrices with many taxa and character/codon sets. For example, the previously mentioned dataset with 715 taxa, 81,050 characters, and 92 character sets (48 with codon sets) reverse-splits in approximately 5 min. Exporting the partitions into FASTA, TNT, or NEXUS formats using the column headers as file names took less than 1 min [''Export'' → ''Export sequences (one file per column)'']. Note that codon information is only preserved when exporting partitions to NEXUS or TNT files and that splitting an already exported concatenated matrix back into its components only works consistently for matrices that have been concatenated by SequenceMatrix. For all other matrices, it is important to remember that when importing NEXUS files, SequenceMatrix only parses the data, the ''BEGIN SETS'', and the ''BEGIN CODONS'' blocks; in TNT files, only the data and ''xgroup'' block are imported. Matrices that contain character set and codon information in other blocks are not compatible. For TNT matrices, SequenceMatrix will read more than the ''official'' 32 xgroups. It will also read the additional xgroups that SequenceMatrix exports as ''comments'' to the header of the concatenated matrix. Note that capturing codon information in concatenated TNT files is also possible, but it requires that the xgroup names for groups containing codon information are not changed. Should the reverse split fail for a concatenated NEXUS file, the easiest way to create a compatible matrix is to use Mesquite's ''Simplified NEXUS'' export option (Maddison and Maddison, 2009). This will convert the matrix into a readable format. However, this file will lack all character set and codon information, but this information can be easily added using the ''addendum to file'' window of Mesquite's ''Simplified NEXUS'' export.

If character sets overlap, reverse-splitting and representing a concatenated matrix in a spreadsheet is difficult. For example, a user may have one character set for all 28S data and two additional character sets specifying the indel-free and ''gappy'' parts of the gene. All three sets cannot be simultaneously represented in a spreadsheet because the latter two sets overlap with the set containing all 28S data. Therefore, either only the full 28S sequence set or the two 28S subsets can be specified in the ''BEGIN SETS'' block if the matrix is supposed to be reverse-split. One convenient way to keep overlapping character set information without interfering with the ability to reverse split a matrix is to place them in an ASSUMPTIONS block. Note, however, that SequenceMatrix will not save the ASSUMPTIONS block when a newly concatenated matrix is exported. The user must manually copy set information from the ASSUMPTIONS block into the newly exported file. Because SequenceMatrix will not parse this block, these sets will not interfere with the

reverse split but can be used during the analysis of the dataset.

Detecting laboratory contamination

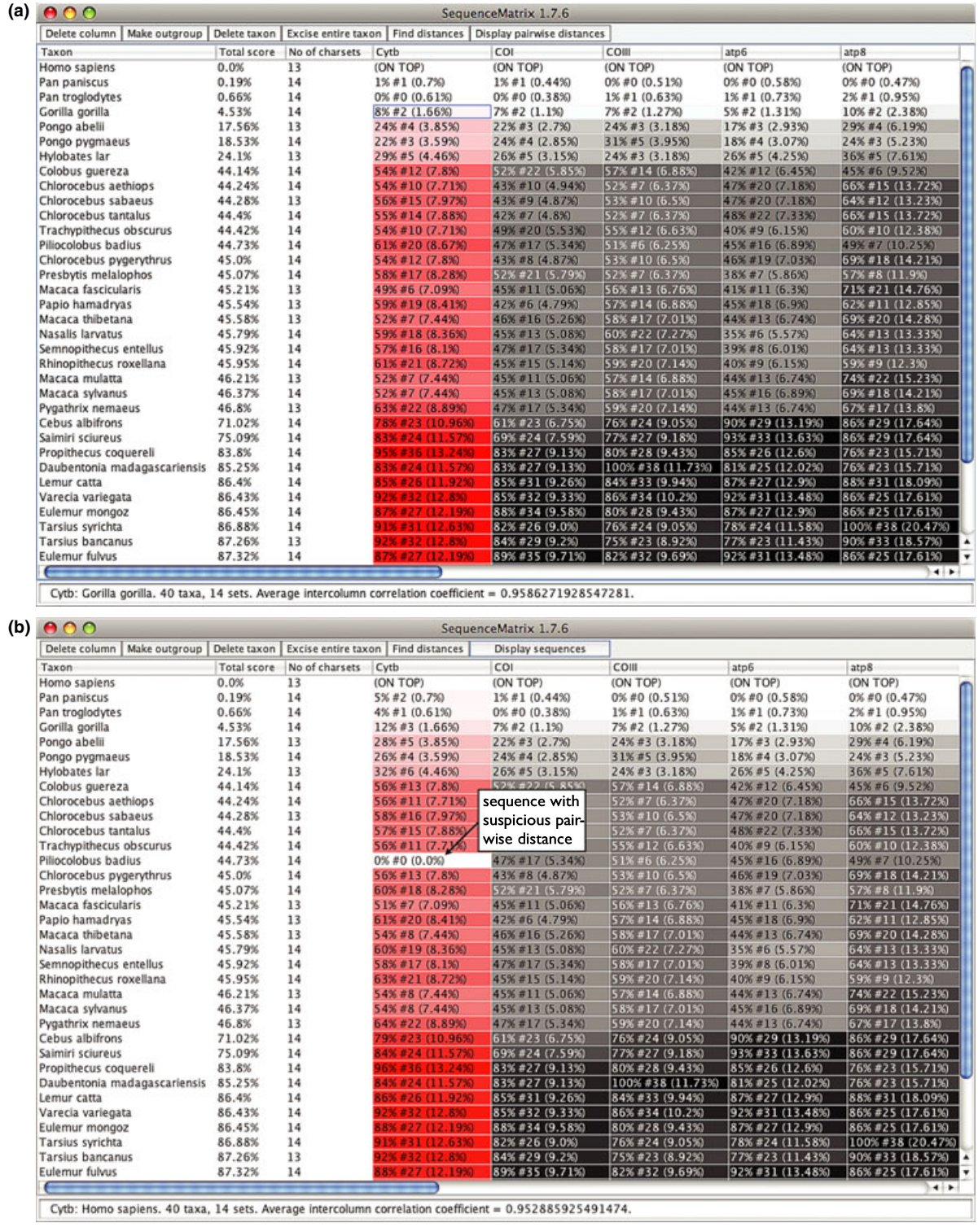
SequenceMatrix has two features that help detect DNA sequences that might have been compromised through laboratory error or mistakes in data management. Two contamination scenarios are particularly common. Genomic DNA from one species finds its way into the genomic DNA or some PCR reactions for another species. The two species will then have identical sequences for all or a few genes, respectively. SequenceMatrix includes a search function that allows the user to search for sequences that are identical or fall below a specified pair-wise distance threshold within gene columns (''Find distances'' button). Of course, small distances are not necessarily due to contamination because some genes evolve slowly, and closely related species can have identical sequences. But scrutiny is warranted if genes with similar evolutionary properties yield very different genetic distances for the same taxa.

Unfortunately, it is more difficult to detect laboratory contamination if the contaminating species is not included in the study. In order to detect those cases, SequenceMatrix provides a tool for examining pair-wise distances across genes (''View'' → ''As pair-wise distances'' or ''Display pair-wise distances'' button). The rationale for this tool is as follows. Overall, one would expect a reasonably good correlation between the pair-wise distances for different genes. For example, if several sequences from species A and B are more similar to each other than either is to sequences from species C, one would be suspicious of a novel gene sequence from species A that is more similar to C than it is to B; i.e. cross contamination of DNA should yield a pattern whereby the contaminated sequence for species A is unexpectedly similar or dissimilar to the remaining sequences. But how can one determine what is ''unexpected''? This can be accomplished by comparing the genetic distances of the gene under scrutiny with the genetic distances of all remaining genes combined. The distances for the latter yield the expected pattern against which the scrutinized gene can be compared. This means that, for a multi-locus dataset with ten genes, ten different analyses, should be carried out. In each analysis, the distance matrix for one gene is compared with the combined distances for the remaining nine genes.

This is the rationale behind the ''Display pair-wise distances'' tool in SequenceMatrix. In this mode, distances can be expressed as either uncorrected pair-wise distances, K2P distances, or transversion-only distances. After activating this mode, SequenceMatrix will automatically reorder all taxa from the smallest to largest distance given the selected reference taxon (Fig. 2). The information in each individual cell is replaced with the genetic distance between the reference taxon and every other taxon, the rank order of similarity with regard to the reference taxon, and a rescaled representation of distance to the reference taxon in which the most divergent sequence is assigned a value of 100%. In addition, the background colour of the cell is adjusted to reflect its genetic distance from the reference taxon. The overall taxon order is calculated based on all genes except for the one gene that is under scrutiny; the cells representing that gene are tinted with shades of red. Suspicious sequences for this gene are those that have unusually high or low distances relative to the surrounding taxa that were placed there based on the data for all remaining genes (see example in Fig. 2). Such a suspicious sequence may require additional scrutiny through BLASTing or re-sequencing. Note that for full exploration of a dataset, each gene has to be inspected separately (right-click on the centre of a cell for the gene and select ''Display pair-wise distances''). It is also important to change the reference taxon repeatedly and to select the most appropriate distance measure (uncorrected, K2P, transversions only). This tool generally works well for matrices consisting of closely related species, as the patterns are difficult to interpret for distantly related taxa where saturation makes distances particularly unreliable.

Sample files

The release contains several sets of sample files for testing SequenceMatrix and we recommend that they be used for an initial exploration of the software. Input files are provided in FASTA, TNT, and NEXUS formats. ''Primates'', ''Sepsidae'', and ''Scathophagidae'' provide examples for straightforward concatenation with some of the TNT and NEXUS input files having codon information. ''PrimatesGenBank'' is more complex because the sequence names include full GenBank information. This example can be used to explore the following features of SequenceMatrix. (i) Importing the files in the ''species name'' or ''sequence name'' mode will yield different datasets. (ii) When using the ''species name'' mode, the user will be alerted that there are several sequences for the same gene and species and that only the longest will be used. An inspection of the input file reveals that the sequences come from different subspecies, but in the ''species name'' mode subspecies differences are not recognized. (iii) Two species have variant spellings across the input files (*Pygathrix roxellana* and *Rhinopithecus roxellana*; *Procolobus badius* and *Piliocolobus badius*). Unifying the spelling within SequenceMatrix's taxon name cell will cause the respective taxon rows to fuse. For the



Taxon	Total score	No of charsets	Cytb	COI	COIII	atp6	atp8
Homo sapiens	0.0%	13	(ON TOP)	(ON TOP)	(ON TOP)	(ON TOP)	(ON TOP)
Pan paniscus	0.19%	14	1% #1 (0.7%)	1% #1 (0.44%)	0% #0 (0.51%)	0% #0 (0.58%)	0% #0 (0.47%)
Pan troglodytes	0.66%	14	0% #0 (0.61%)	0% #0 (0.38%)	1% #1 (0.63%)	1% #1 (0.73%)	2% #1 (0.95%)
Gorilla gorilla	4.53%	14	8% #2 (1.66%)	7% #2 (1.1%)	7% #2 (1.27%)	5% #2 (1.31%)	10% #2 (2.38%)
Pongo abelii	17.56%	13	24% #4 (3.85%)	22% #3 (2.7%)	24% #3 (3.18%)	17% #3 (2.93%)	29% #4 (6.19%)
Pongo pygmaeus	18.53%	14	22% #3 (3.59%)	24% #4 (2.85%)	31% #5 (3.95%)	18% #4 (3.07%)	24% #3 (5.23%)
Hylobates lar	24.1%	13	29% #5 (4.46%)	26% #5 (3.15%)	24% #3 (3.18%)	26% #5 (4.25%)	36% #5 (7.61%)
Colobus guereza	44.14%	14	54% #12 (7.8%)	52% #22 (5.85%)	57% #14 (6.88%)	42% #12 (6.45%)	45% #6 (9.52%)
Chlorocebus aethiops	44.24%	14	54% #10 (7.71%)	43% #10 (4.94%)	52% #7 (6.37%)	47% #20 (7.18%)	66% #15 (13.72%)
Chlorocebus sabaeus	44.28%	13	56% #15 (7.97%)	43% #9 (4.87%)	53% #10 (6.5%)	47% #20 (7.18%)	64% #12 (13.23%)
Chlorocebus tantalus	44.4%	14	55% #14 (7.88%)	42% #7 (4.8%)	52% #7 (6.37%)	48% #22 (7.33%)	66% #15 (13.72%)
Trachypithecus obscurus	44.42%	14	54% #10 (7.71%)	49% #20 (5.53%)	55% #12 (6.63%)	40% #9 (6.15%)	60% #10 (12.38%)
Piliocolobus badius	44.73%	14	61% #20 (8.67%)	47% #17 (5.34%)	51% #6 (6.25%)	45% #16 (6.89%)	49% #7 (10.25%)
Chlorocebus pygerythrus	45.0%	14	54% #12 (7.8%)	43% #8 (4.87%)	53% #10 (6.5%)	46% #19 (7.03%)	69% #18 (14.21%)
Presbytis melalophos	45.07%	14	58% #17 (8.28%)	52% #21 (5.79%)	52% #7 (6.37%)	38% #7 (5.86%)	57% #8 (11.9%)
Macaca fascicularis	45.21%	13	49% #6 (7.09%)	45% #11 (5.06%)	56% #13 (6.76%)	41% #11 (6.3%)	71% #21 (14.76%)
Papio hamadryas	45.54%	13	59% #19 (8.41%)	42% #6 (4.79%)	57% #14 (6.88%)	45% #18 (6.9%)	62% #11 (12.85%)
Macaca thibetana	45.58%	13	52% #7 (7.44%)	46% #16 (5.26%)	58% #17 (7.01%)	44% #13 (6.74%)	69% #20 (14.28%)
Nasalis larvatus	45.79%	14	59% #18 (8.36%)	45% #13 (5.08%)	60% #22 (7.27%)	35% #6 (5.57%)	64% #13 (13.33%)
Semnopithecus entellus	45.92%	14	57% #16 (8.1%)	47% #17 (5.34%)	58% #17 (7.01%)	39% #8 (6.01%)	64% #13 (13.33%)
Rhinopithecus roxellana	45.95%	14	61% #21 (8.72%)	45% #15 (5.14%)	59% #20 (7.14%)	40% #9 (6.15%)	59% #9 (12.3%)
Macaca mulatta	46.21%	13	52% #7 (7.44%)	45% #11 (5.06%)	57% #14 (6.88%)	44% #13 (6.74%)	74% #22 (15.23%)
Macaca sylvanus	46.37%	14	52% #7 (7.44%)	45% #13 (5.08%)	58% #17 (7.01%)	45% #16 (6.89%)	69% #18 (14.21%)
Pygathrix nemaeus	46.8%	13	63% #22 (8.89%)	47% #17 (5.34%)	59% #20 (7.14%)	44% #13 (6.74%)	67% #17 (13.8%)
Cebus albifrons	71.02%	14	78% #23 (10.96%)	61% #23 (6.75%)	76% #24 (9.05%)	90% #29 (13.19%)	86% #29 (17.64%)
Saimiri sciureus	75.09%	14	83% #24 (11.57%)	69% #24 (7.59%)	77% #27 (9.18%)	93% #33 (13.63%)	86% #29 (17.64%)
Propithecus coquereli	83.8%	14	95% #36 (13.24%)	83% #27 (9.13%)	80% #28 (9.43%)	85% #26 (12.6%)	76% #23 (15.71%)
Daubentonia madagascariensis	85.25%	14	83% #24 (11.57%)	83% #27 (9.13%)	100% #38 (11.73%)	81% #25 (12.02%)	76% #23 (15.71%)
Lemur catta	86.4%	14	85% #26 (11.92%)	85% #31 (9.26%)	84% #33 (9.94%)	87% #27 (12.9%)	88% #31 (18.09%)
Varecia variegata	86.43%	14	92% #32 (12.8%)	85% #32 (9.33%)	86% #34 (10.2%)	92% #31 (13.48%)	86% #25 (17.61%)
Eulemur mongoz	86.45%	14	87% #27 (12.19%)	88% #34 (9.58%)	80% #28 (9.43%)	87% #27 (12.9%)	86% #25 (17.61%)
Tarsius syrichta	86.88%	14	91% #31 (12.63%)	82% #26 (9.0%)	76% #24 (9.05%)	78% #24 (11.58%)	100% #38 (20.47%)
Tarsius bancanus	87.26%	13	92% #32 (12.8%)	84% #29 (9.2%)	75% #23 (8.92%)	77% #23 (11.43%)	90% #33 (18.57%)
Eulemur fulvus	87.32%	14	87% #27 (12.19%)	89% #35 (9.71%)	82% #32 (9.69%)	92% #31 (13.48%)	86% #25 (17.61%)

Taxon	Total score	No of charsets	Cytb	COI	COIII	atp6	atp8
Homo sapiens	0.0%	13	(ON TOP)	(ON TOP)	(ON TOP)	(ON TOP)	(ON TOP)
Pan paniscus	0.19%	14	5% #2 (0.7%)	1% #1 (0.44%)	0% #0 (0.51%)	0% #0 (0.58%)	0% #0 (0.47%)
Pan troglodytes	0.66%	14	4% #1 (0.61%)	0% #0 (0.38%)	1% #1 (0.63%)	1% #1 (0.73%)	2% #1 (0.95%)
Gorilla gorilla	4.53%	14	12% #3 (1.66%)	7% #2 (1.1%)	7% #2 (1.27%)	5% #2 (1.31%)	10% #2 (2.38%)
Pongo abelii	17.56%	13	28% #5 (3.85%)	22% #3 (2.7%)	24% #3 (3.18%)	17% #3 (2.93%)	29% #4 (6.19%)
Pongo pygmaeus	18.53%	14	26% #4 (3.59%)	24% #4 (2.85%)	31% #5 (3.95%)	18% #4 (3.07%)	24% #3 (5.23%)
Hylobates lar	24.1%	13	32% #6 (4.46%)	26% #5 (3.15%)	24% #3 (3.18%)	26% #5 (4.25%)	36% #5 (7.61%)
Colobus guereza	44.14%	14	56% #13 (7.8%)	52% #22 (5.85%)	57% #14 (6.88%)	42% #12 (6.45%)	45% #6 (9.52%)
Chlorocebus aethiops	44.24%	14	56% #11 (7.71%)	[illegible]	52% #7 (6.37%)	47% #20 (7.18%)	66% #15 (13.72%)
Chlorocebus sabaeus	44.28%	13	58% #16 (7.97%)	[illegible]	53% #10 (6.5%)	47% #20 (7.18%)	64% #12 (13.23%)
Chlorocebus tantalus	44.4%	14	57% #15 (7.88%)	[illegible]	52% #7 (6.37%)	48% #22 (7.33%)	66% #15 (13.72%)
Trachypithecus obscurus	44.42%	14	56% #11 (7.71%)	[illegible]	55% #12 (6.63%)	40% #9 (6.15%)	60% #10 (12.38%)
Piliocolobus badius	44.73%	14	0% #0 (0.0%)	47% #17 (5.34%)	51% #6 (6.25%)	45% #16 (6.89%)	49% #7 (10.25%)
Chlorocebus pygerythrus	45.0%	14	56% #13 (7.8%)	43% #8 (4.87%)	53% #10 (6.5%)	46% #19 (7.03%)	69% #18 (14.21%)
Presbytis melalophos	45.07%	14	60% #18 (8.28%)	52% #21 (5.79%)	52% #7 (6.37%)	38% #7 (5.86%)	57% #8 (11.9%)
Macaca fascicularis	45.21%	13	51% #7 (7.09%)	45% #11 (5.06%)	56% #13 (6.76%)	41% #11 (6.3%)	71% #21 (14.76%)
Papio hamadryas	45.54%	13	61% #20 (8.41%)	42% #6 (4.79%)	57% #14 (6.88%)	45% #18 (6.9%)	62% #11 (12.85%)
Macaca thibetana	45.58%	13	54% #8 (7.44%)	46% #16 (5.26%)	58% #17 (7.01%)	44% #13 (6.74%)	69% #20 (14.28%)
Nasalis larvatus	45.79%	14	60% #19 (8.36%)	45% #13 (5.08%)	60% #22 (7.27%)	35% #6 (5.57%)	64% #13 (13.33%)
Semnopithecus entellus	45.92%	14	58% #17 (8.1%)	47% #17 (5.34%)	58% #17 (7.01%)	39% #8 (6.01%)	64% #13 (13.33%)
Rhinopithecus roxellana	45.95%	14	63% #21 (8.72%)	45% #15 (5.14%)	59% #20 (7.14%)	40% #9 (6.15%)	59% #9 (12.3%)
Macaca mulatta	46.21%	13	54% #8 (7.44%)	45% #11 (5.06%)	57% #14 (6.88%)	44% #13 (6.74%)	74% #22 (15.23%)
Macaca sylvanus	46.37%	14	54% #8 (7.44%)	45% #13 (5.08%)	58% #17 (7.01%)	45% #16 (6.89%)	69% #18 (14.21%)
Pygathrix nemaeus	46.8%	13	64% #22 (8.89%)	47% #17 (5.34%)	59% #20 (7.14%)	44% #13 (6.74%)	67% #17 (13.8%)
Cebus albifrons	71.02%	14	79% #23 (10.96%)	61% #23 (6.75%)	76% #24 (9.05%)	90% #29 (13.19%)	86% #29 (17.64%)
Saimiri sciureus	75.09%	14	84% #24 (11.57%)	69% #24 (7.59%)	77% #27 (9.18%)	93% #33 (13.63%)	86% #29 (17.64%)
Propithecus coquereli	83.8%	14	96% #36 (13.24%)	83% #27 (9.13%)	80% #28 (9.43%)	85% #26 (12.6%)	76% #23 (15.71%)
Daubentonia madagascariensis	85.25%	14	84% #24 (11.57%)	83% #27 (9.13%)	100% #38 (11.73%)	81% #25 (12.02%)	76% #23 (15.71%)
Lemur catta	86.4%	14	86% #26 (11.92%)	85% #31 (9.26%)	84% #33 (9.94%)	87% #27 (12.9%)	88% #31 (18.09%)
Varecia variegata	86.43%	14	92% #32 (12.8%)	85% #32 (9.33%)	86% #34 (10.2%)	92% #31 (13.48%)	86% #25 (17.61%)
Eulemur mongoz	86.45%	14	88% #27 (12.19%)	88% #34 (9.58%)	80% #28 (9.43%)	87% #27 (12.9%)	86% #25 (17.61%)
Tarsius syrichta	86.88%	14	91% #31 (12.63%)	82% #26 (9.0%)	76% #24 (9.05%)	78% #24 (11.58%)	100% #38 (20.47%)
Tarsius bancanus	87.26%	13	92% #32 (12.8%)	84% #29 (9.2%)	75% #23 (8.92%)	77% #23 (11.43%)	90% #33 (18.57%)
Eulemur fulvus	87.32%	14	88% #27 (12.19%)	89% #35 (9.71%)	82% #32 (9.69%)	92% #31 (13.48%)	86% #25 (17.61%)

Fig. 2. "Pair-wise distance mode": gene cytochrome *b* (cytb) is screened for contamination (column in red). (a) No sign of contamination. (b) Cytb for *Piliocolobus badius* is probably due to contamination because its distance to the reference taxon, *Homo sapiens*, is unexpectedly low compared with the signal in all other genes.

Scathophagidae dataset, we also include a NEXUS batch file that can be used for a concatenation check. When executed in PAUP* within the folder that includes the concatenated matrix and the input files, it will generate a series of log files that include the data characteristics in the input file and the same data within the matrix. For all four examples, we also include output files in TNT and NEXUS formats.

Compatibility and installation

SequenceMatrix is written in Java and can therefore be used on many platforms, including Microsoft Windows (XP or 7), Apple Macintosh (different versions of the OS), and Linux. The program can be downloaded from http://code.google.com/p/sequencematrix/. The program is "stand-alone"; no installer file is needed. On MacOS X, the "SequenceMatrix" application can be executed by double-clicking on it. On Windows, drag the "SequenceMatrix.jar" and "SequenceMatrix.bat" into a folder and double-click on either to start the program. On desktop computers, relatively little memory is assigned to Java by default. However, for the concatenation and reverse-splitting of large matrices more memory is needed. We therefore recommend the using the batch file ("SequenceMatrix.bat"), which will assign additional memory to Java.

Acknowledgements

We thank the members of the Evolutionary Biology Laboratory in Singapore for testing this program extensively and offering suggestions for its improvement. This project was funded by grant R377-000-040-112 from the Ministry of Education (Singapore).

References

Ang, Y., Puniamoorthy, N., Meier, R., 2008. Secondarily reduced foreleg armature in *Perochaeta dikowi* sp.n. (Diptera: Cyclorrhapha: Sepsidae) due to a novel mounting technique. Syst. Entomol. 33, 552–559.

Balke, M., Ribera, I., Hendrich, L., Miller, M.A., Sagata, K., Posman, A., Vogler, A.P., Meier, R., 2009. New Guinea highland origin of a widespread arthropod supertramp. Proc. R. Soc. Lond. B Biol. Sci. 276, 2359–2367.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sorensen, M.V., Haddock, S.H.D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q., Giribet, G., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.

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

Huang, D.W., Meier, R., Todd, P.A., Chou, L.M., 2009. More evidence for pervasive paraphyly in scleractinian corals: systematic study of Southeast Asian Faviidae (Cnidaria; Scleractinia) based on molecular and morphological data. Mol. Phylogenet. Evol. 50, 102–116.

Jones, M., Blaxter, M., 2006. TaxMan: a taxonomic database manager. BMC Bioinformatics 7, 536.

Kutty, S.N., Bernasconi, M.V., Sifner, F., Meier, R., 2007. Sensitivity analysis, molecular systematics and natural history evolution of Scathophagidae (Diptera: Cyclorrhapha: Calyptratae). Cladistics 23, 64–83.

Kutty, S.N., Pape, T., Pont, A., Wiegmann, B.M., Meier, R., 2008. The Muscoidea (Diptera: Calyptratae) are paraphyletic: evidence from four mitochondrial and four nuclear genes. Mol. Phylogenet. Evol. 49, 639–652.

Kutty, S.N., Pape, T., Wiegmann, B.M., Meier, R., 2010. Molecular phylogeny of the Calyptratae (Diptera: Cyclorrhapha) with an emphasis on the superfamily Oestroidea and the position of Mystacinobiidae and McAlpine"s Fly. Syst. Entomol., in press, doi: 10.1111/j.1365-3113.2010.00536.x.

Laamanen, T.R., Meier, R., Miller, M.A., Hille, A., Wiegmann, B.M., 2005. Phylogenetic analysis of *Themira* (Sepsidae: Diptera): sensitivity analysis, alignment, and indel treatment in a multigene study. Cladistics 21, 258–271.

Lim, G.S., Hwang, W.S., Kutty, S.N., Meier, R., Grootaert, P., 2010. Mitochondrial and nuclear markers support the monophyly of Dolichopodidae and suggest a rapid origin of the subfamilies (Diptera: Empidoidea). Syst. Entomol. 35, 59–70.

Lohman, D.J., Peggie, D., Pierce, N.E., Meier, R., 2008. Phylogeography and genetic diversity of a widespread Old World butterfly, *Lampides boeticus* (Lepidoptera: Lycaenidae). BMC Evol. Biol. 8, 301.

Lohman, D.J., Ingram, K.K., Prawiradilaga, D.M., Winker, K., Sheldon, F.H., Moyle, R.G., Ng, P.K.L., Ong, P.S., Wang, L.K., Brailee, T.M., Astuti, D., Meier, R., 2010. Cryptic genetic diversity in "widespread" Southeast Asian bird species suggests that Philippine avian endemism is gravely underestimated. Biol. Conserv. 143, 1885–1890.

Maddison, D.R., Maddison, W.P., 2001. MacClade 4: Analysis of Phylogeny and Character Evolution. Sinauer Associates, Sunderland, MA.

Maddison, W.P., Maddison, D.R. 2009. Mesquite: a modular system for evolutionary analysis. Version 2.72. http://mesquiteproject.org.

Maddison, D.R., Swofford, D.L., Maddison, W.P., 1997. NEXUS: an extensible file format for systematic information. Syst. Biol. 46, 590–621.

Meier, R., Baker, R., 2002. A cladistic analysis of Diopsidae (Diptera) based on morphological and DNA sequence data. Insect Syst. Evol. 33, 325–336.

Meier, R., Wiegmann, B.M., 2002. A phylogenetic analysis of Coelopidae (Diptera) based on morphological and DNA sequence data. Mol. Phylogenet. Evol. 25, 393–407.

Meier, R., Kwong, S., Vaidya, G., Ng, P.K.L., 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst. Biol. 55, 715–728.

Meier, R., Zhang, G.Y., Ali, F., 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Syst. Biol. 57, 809–813.

Nixon, K.C., Wheeler, Q.D., 1992. Extinction and the origin of species. In: Novacek, M.J., Wheeler, Q.D. (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 117–143.

Petersen, F.T., Meier, R., Kutty, S.N., Wiegmann, B.M., 2007. The phylogeny and evolution of host choice in the Hippoboscoidea (Diptera) as reconstructed using four molecular markers. Mol. Phylogenet. Evol. 45, 111–122.

Puniamoorthy, N., Su, K.F.Y., Meier, R., 2008. Bending for love: losses and gains of sexual dimorphisms are strictly correlated with

changes in the mounting position of sepsid flies (Sepsidae: Diptera). BMC Evol. Biol. 8, 155.

Puniamoorthy, N., Ismail, M.R.B., Tan, D.S.H., Meier, R., 2009. From kissing to belly stridulation: comparative analysis reveals surprising diversity, rapid evolution, and much homoplasy in the mating behaviour of 27 species of sepsid flies (Diptera: Sepsidae). J. Evol. Biol. 22, 2146–2156.

Roure, B., Rodriguez-Ezpeleta, N., Philippe, H., 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. 7, Suppl. 1, S2.

Smith, S.A., Dunn, C.W., 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24, 715–716.

Su, K.F., Meier, R., Jackson, R.R., Harland, D.P., Li, D., 2007. Convergent evolution of eye ultrastructure and divergent evolution of vision-mediated predatory behaviour in jumping spiders. J. Evol. Biol. 20, 1478–1489.

Su, F.-Y.K., Kutty, S.N., Meier, R., 2008. Morphology versus molecules: the phylogenetic relationships of Sepsidae (Diptera: Cyclorrhapha) based on morphology and DNA sequence data from ten genes. Cladistics 24, 902–916.

Swofford, D.L., 2004. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods). Sinauer Associates, Inc., Sunderland, MA.

Wowor, D., Muthu, V., Meier, R., Balke, M., Cai, Y.X., Ng, P.K.L., 2009. Evolution of life history traits in Asian freshwater prawns of the genus *Macrobrachium* (Crustacea: Decapoda: Palaemonidae) based on multilocus molecular phylogenetic analysis. Mol. Phylogenet. Evol. 52, 340–350.

Yeo, D.C.J., Shih, H.T., Meier, R., Ng, P.K.L., 2007. Phylogeny and biogeography of the freshwater crab genus *Johora* (Crustacea: Brachyura: Potamidae) from the Malay Peninsula, and the origins of its insular fauna. Zool. Scr. 36, 255–269.