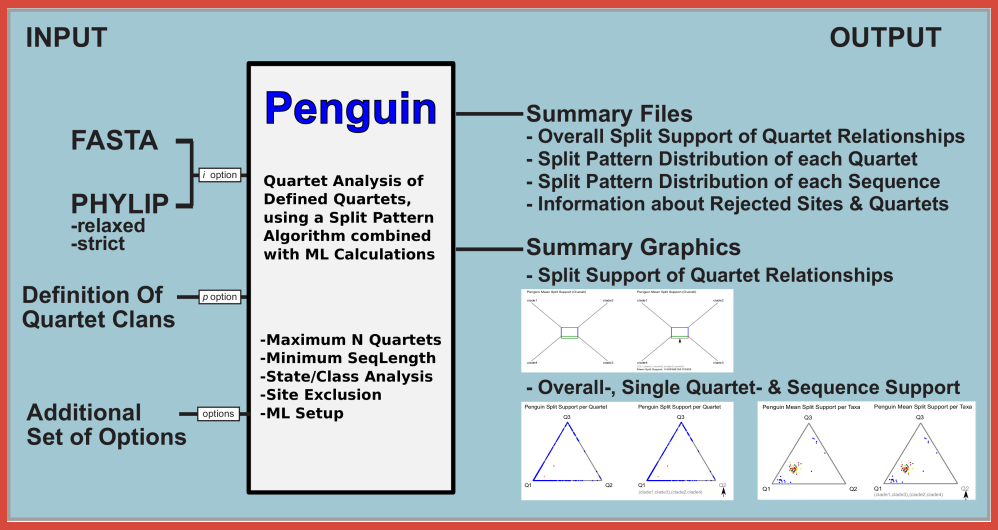
Patrick Kück



PENGUIN

May 2016

Contents

List of Figures

List of Tables

1 Features

1.1 What PENGUIN does and why you should use it

PENGUIN is designed to perform site pattern analyses for quartets of aligned nucleotide and amino acid sequences using observed and expected split-supporting site-patterns by consideration of two different topological directive transformations for the inner branch of each quartet relationship (Figure 1). A single site position supports a split represented by the inner branch of a quartet relationship if two sequences share the same character which is not present in the other two sequences. The distinction between two different topological directions for character transformation along the internal branch of a given quartet enables a classification of split-supporting positions into potentially phylogenetically informative and non-informative patterns. Therefore, split clans of a given quartet topology are evaluated depending on the assumed evolutionary direction. For the identification of potentially convergent, misleading split support for a given directive quartet relationship, the Maximum Likelihood (ML) approach of the P4 package (Foster, 2004) is used to estimate the expected number of convergently evolved split patterns based on branch length and model parameter optimisation. To use the PENGUIN implemented *PhyQuart* algorithm it is assumed that no *a priori* knowledge about the placement of the root exists. Therefore, for each quartet all possible topologies and spins are tested to find the best phylogenetic signal based exclusively on putative apomorphic character states. The polarised quartet topology with the highest phylogenetic signal is assumed to be the correct one.

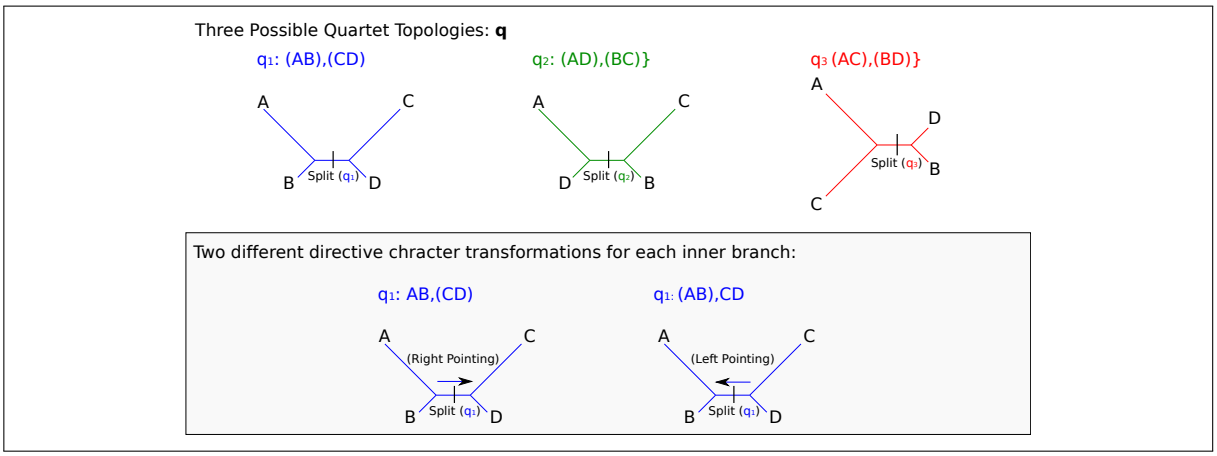


Figure 1: Definition of polarity (direction of character transformation) along the internal branch of a given quartet tree, e.g. (AB),(CD). (AB),CD indicate the direction is towards CD and towards AB for (AB),CD. Assuming polarities enables a classification of split-support into potentially phylogenetically informative (apomorphic) and uninformative (plesiomorphic) patterns.

1.2 Penguin Input

PENGUIN reads files of multiple nucleotide and amino acid sequence alignments in FASTA and PHYLIP format. If the alignment consists of more than four sequences, an input file comprising four predefined clans of one or more taxa each must be provided in plain TEXT format. If not otherwise specified, PENGUIN analyses all possible four-taxon combinations between the four predefined taxa clans. PENGUIN does not allow multiple taxon records within given input file(s), while sequences of disaccording taxa names between a predefined clan definition file and a corresponding multiple sequence alignment are just left unanalysed. Sequence sites with indel ('-'), ambiguity or missing characters are always excluded from the analysis. Under default, PENGUIN excludes all forbidden site positions (positions with gaps, ambiguities or missing characters) separately for each sequence quartet of a given alignment. Alternatively, site exclusion can be performed on the complete sequence alignment in advance of the quartet establishment.

1.3 Penguin Output

PENGUIN calculates summarised output information of identified split support for each possible quartet relationship between four-taxon or taxa clanes. Obtained discrepancies in topological split support of the three possible quartet topologies are also presented as split network and triangle graph. A further vector graphic shows the number and percentage distribution of best and second best resolved quartet relationships.

1.4 Implementation & Usage

PENGUIN is a command-line driven program written in PERL and works on WindowsPCs, Macs and Linux operating systems. Therefore, It can easily be integrated into automated pipelines of phylogenetic studies. Furthermore, results issued by PENGUIN can be used to find clusters of genes and/or taxa with similar split properties. All calculations of the PENGUIN software program are very fast and can be easily executed on a normal desktop computer, even if data sets consist of phylogenomic data.

2 Pre-Installations

To execute PENGUIN, a PERL interpreter as well as P4, a Python package for Maximum Likelihood and Bayesian analysis of molecular sequences, must be installed on the current run system.

2.1 PERL Installation

Linux and Mac systems normally have a PERL interpreter pre-installed. However, Windows users have to install a PERL interpreter ex post. We would recommend the ActivePerl interpreter which can be downloaded for free under:

- http://activeperl.softonic.de/

2.2 P4 Installation

PENGUIN needs several P4 Python script routines to calculate the Maximum Likelihood expected number of site pattern frequencies and to determine the expected number of convergent split pattern for given 4-taxa constraint topologies. For the installation of P4 please follow the documentation and installation instructions under:

- http://p4.nhm.ac.uk/

3 Usage/Options

For using PENGUIN open the terminal of your operating system. Move through your directory path to the folder where PENGUIN is placed. PENGUIN can be directly started by the command line in one row which simplifies the implementation of PENGUIN into complex process pipelines. Move through your directory path to the folder where PENGUIN is located and type the name of the PENGUIN version, followed by a blank and the required options with a dash (-) sign in front of each.

- user@linux:~$ perl Penguin_v1.0.pl -h <enter> ↪ help menu
- user@linux:~$ perl Penguin_v1.0.pl -i *path/infile* <enter> ↪ start PENGUIN under default

Note: Windows users don't have to put "perl" in front of the program name. Then press <enter>. Make sure you write the input options correctly, for example "-i" and not "- i". Otherwise PENGUIN will not start working but instead open the Help menu.

3.1 Options

PENGUIN knows several input file options. It stops and opens the Help menu if an unknown option is encountered. PENGUIN checks each input file according to correct format and forbidden sequence and structure characters. This subsection gives a short explanation for possible input file options and accepted file formats. Notice that not supported file formats cause PENGUIN to abort.

Table 1: Overview of features and options

Info options	**Command**			**Example**
Help menu	h		perl	Penguin_v1.0.pl -h
Preface	P		perl	Penguin_v1.0.pl -P
Parameter options		**Default**		
Sequence Alignment Input File	i	*None*	-i	MSA_infile.phy
Clan Definition Input File	p	*None*	-p	Taxa_Clans.txt
Min. Sequence Length for Quartets	M	500	-M	<Integer>
Max. Number of Generated Quartets	I	10000	-I	<Integer>
Translate Site States to Class States	c	No	-c	
Exclude Unallowed States of Complete Alignment	N	Per Quartet	-N	
ML Substitution Model Parameter	m	GTR *or* WAG	-m	<Integer>
ML α Shape Start Parameter	a	0.5	-a	<Float>
ML pINV Start Parameter	I	0.3	-I	<Float>
Elimination of Script Queries	u	No	-u	

3.1.1 Sequence Input File (-i) Option

To define the sequence input file, type "-i *infile*". The name of the sequence input file has to be given with file format suffix (e.g. .fas). PENGUIN can handle two different types of sequence input file formats, namely FASTA (.fas) and relaxed PHYLIP (.phy). All sequences of the sequence input file must have equal sequence lengths. Sequence names are allowed to consist of alphanumeric signs, underscores ("_") and in case of FASTA files also blanks, other signs are not allowed. Sequences are allowed to consist of signs covered by the universal DNA/RNA or amino acid code, code corresponding ambiguity characters, "?", and "-".

FASTA (.fas) Format PENGUIN is able to read sequences of FASTA files either if they are in one line or with line interruptions (blocks). Sequence names have to be in one line and have to start with an ">"! Each line has to end with a line break. Table 2 gives an example of both acceptable FASTA formats.

Table 2: Known FASTA format in non-interleaved (format 1) and interleaved format (format 2).

FASTA format 1

```
>Name_sequence_1
AGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCTAGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCT
>Name_sequence_2
AGCTCCGGCCCTTTG–AGA–GTGTCCTTTCCTAGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCT
⋮
>Name_sequence_n
AGCTCCCGTCCTTTGGAGAGGTGTCCTTTCCTAGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCT
```

FASTA format 2

```
>Name_sequence_1
AGCTGTCCTTTCTTG–AGA–GTGTCCTTTCCTAGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCT
AGCTGTCCTTTCTTG–AGA–GTGTCCTTTCCTAGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCT
⋮
>Name_sequence_n
AGCTGTCCTTTCTTG–AGA–GTGTCCTTTCCTAGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCT
AGCTGTCCTTTCTTG–AGA–GTGTCCTTTCCTAGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCT
```

FASTA files can be produced by various programs, including `MAFFT` (Katoh *et al.*, 2005; Katoh and Toh, 2008, 2010; Katoh and Standley, 2013), `MUSCLE` (Edgar, 2004), or `T-COFFEE` (Notredame *et al.*, 2000; Notredame, 2002). To convert aligned sequence files from other alignment formats to FASTA, software tools like `T-COFFEE` (Notredame *et al.*, 2000; Notredame, 2002), `FASconCAT` (Kück and Meusemann, 2010) or `FASconCAT-G` (Kück and Longo, 2014) can be used. `FASconCAT` (Kück and Meusemann, 2010) and `FASconCAT-G` (Kück and Longo, 2014) can be further used for gene concatenation. Please, read the respective manuals and/or publications for further details.

Relaxed PHYLIP (.phy) Format Each line has to end with a line break. Table 3 shows a typical PHYLIP file in interleaved format, but non-interleaved format is allowed as well. Sequence names are allowed to contain more than ten signs at maximum and have to be separated from the following sequence by a white space.

Table 3: Example of a relaxed interleaved PHYLIP formatted input file.

PHYLIP format (Interleaved)

6 40				
Name_sequence_1	AGGGCCCTTG	CGCTTGGCCC	CGCTTGGCCC	AGGGCCCTTG
Name_sequence_2	AGGGCCCTTG	CGCCTCCCCC	CGCTTGGCCC	AGGGCCCTTG
Name_sequence_n	AGGCCCCTTG	CGCCGCCCGG	CGCTTGGCCC	AGGGCCCTTG
<line break>				
	ATTTCCCTTG	GGCTTCCCCC	CGCTTGGCCC	AGGGCCCTTG
	ATTTCCCTTG	GGGGGCCTCC	CGCTTGGCCC	AGGGCCCTTG
	ATCTCCCTTG	GGCCGGGGGC	CGCTTGGCCC	AGGGCCCTTG

Extant approaches which can automatedly produce aligned sequences in PHYLIP format are e.g. the

PHYLIP package (Felsenstein, 1993) T-COFFEE (Notredame *et al.*, 2000; Notredame, 2002). To convert aligned sequence files in FASTA format software tools like T-COFFEE (Notredame *et al.*, 2000; Notredame, 2002), FASconCAT (Kück and Meusemann, 2010) or FASconCAT-G (Kück and Longo, 2014) can be used. FASconCAT (Kück and Meusemann, 2010) and FASconCAT-G (Kück and Longo, 2014) can be further used for gene concatenation. Please, read the respective manuals and/or publications for further details.

3.1.2 Clan Definition File (-p Option)

Besides taking a four-taxon alignment, PENGUIN can also calculate quartet support between four user-defined clans. The four clans have to be defined via the clans definition file given by the "-p" option. The definition file has to be in plane .txt format. Taxa clans are defined in separate lines. Each clan has to start with a user specified clan name (alphanumeric signs and underscore are allowed!), followed by a comma, followed by comma separated taxon names. Be aware that all taxon names are identical to the corresponding sequence names of the sequence input file (case sensitive!). Blanks between comma-separated taxon names are not allowed! A taxon name can only be included in a single clan. Table 4 shows the correct format of a typical definition file.

Table 4: Example of a typical clan definition file.

Clan_Definition_File.txt
Name_Clan_1,Taxon_1,Taxon_2,Taxon_3,Taxon_4
Name_Clan_2,Taxon_a,Taxon_b,Taxon_c
Name_Clan_3,Taxon_5,Taxon_6,Taxon_7,Taxon_8,Taxon_9
Name_Clan_4,Taxon_V,Taxon_X,Taxon_Y,Taxon_Z

3.1.3 Definition of Minimum Allowed Sequence Length for Quartets (-M Option)

After rejecting all ambiguity characters and, optionally, Indel event ('-') containing site positions, PENGUIN analyses only quartet sequences if the remaining sequence length of a given set of four taxa is at least as long as the minimum number of defined site positions. The default value of the minimum allowed sequence length is set to 2000 basepairs (bp). Quartet sequences below this value are not analysed. Instead, Pinguin prints a warning prompt to the terminal window naming the affected 4-taxa combination and skips to the next quartet analysis.

- **Default Value: 2000 bp**
- **Command: -M** <**Integer**>
- **Example : -M 4000** (Set Value To 4000 bp)

Note: The higher the number of site pattern for a given 4-taxa quartet, the better the PENGUIN implemented quartet split calculation. Therefore, we suggest rather a higher threshold value for the minimum allowed sequence length than a smaller value.

3.1.4 Definition of Maximum Number of Generated Quartets (-l Option)

PENGUIN generates all possible quartets between defined taxa sequences given a clan definition file (see section 3.1.2 Clan Definition File (-p Option)). The number of generated quartets increases strongly with the number of taxon sequences defined for each clan. Therefore, the total computation time can be very time consuming if the number of quartets is high and the quartet assigned sequence lengths large. Table 5 gives an insight about computation time for a single quartet with different sequence lengths. To avoid overstrain computation times PENGUIN allows the definition of a maximum limit of generated quartets. The default value is set to 10,000 quartets. PENGUIN stops if the number of possible quartets

exceeds the maximum limit of allowed quartets with a terminal user request with three possible user options:

1. To proceed with the total number of possible quartets type <enter>
2. To proceed with a random selection of maxima allowed quartets type <r> <enter>
3. To quit PENGUIN type <q> <enter>

Table 5: Examples of quartet computation times for different sequence lengths using a normal desktop computer (2.8 GHz Intel Core i7).

Data Type	Quartet Sequence Length (bp)	Computation Time (sec)
Nucleotide		
	30,000	≈ 8
	5,000	≈ 4
	700	≈ 1
Amino Acid		
	2,000	≈ 100
	500	≈ 80

- **Default Value: 10000 bp**
- **Command: -l <Integer>**
- **Example : -l 20000** (Set Value To 20000 bp)

Note: To avoid terminal request in pipeline processes set the maximum number of allowed quartets to sufficiently high, e.g. -M 10000000000000.

3.1.5 Elimination of Script Queries (-u Option)

If the maximum number of possible single 4-taxon quartet combinations increases the number of allowed quartet analyses, PENGUIN stops under default with a user query how to proceed further (see '-l' option). This query would stop automatic process pipelines. To oppose the script query type -u. PENGUIN will draw by random the maximum number of allowed 4-taxa combinations from the pool of all possible quartet combinations.

- **Default: Start Queries**
- **Command: -u**
- **Example : -u** (Elimination of Script Queries)

3.1.6 Translate Site States to Class States (-c Option)

PENGUIN analysis quartets in default by using IUPAC standard nucleotide coded site pattern distributions. To analyse quartets in a more conservative way PENGUIN offers the possibility to reduce number of possible site characters from 4 (nucleotide data) to 2 by summarising characters to purines (R) and pyrimidines (Y) and for amino acid data from 20 to 2 by summarising characters to hydrophobic (R) and hydrophilic (Y) classes. To compress site states to class states type -c.

- **Default: No Translation**
- **Command: -c**
- **Example : -c** (Compress Site States to Class States)

3.1.7 Exclude Unallowed Character Sites of Complete Alignment (-N Option)

PENGUIN performs quartet analyses using only sequence site positions without ambiguity and indel characters. Under default insertion/deletion events (indels ↪ '-') and/or ambiguity state including site positions are excluded individually for each given quartet. With the -N option, site positions with unallowed characters are excluded among the complete sequence alignment and not individually for each single set of quartet sequences.

- **Default: Reject Indel Site Positions Individually**
- **Command: -N**
- **Example : -N** (Reject Indel Site Positions With Unallowed Characters Over The Entire Alignment)

Note: The exclusion of unallowed characters over the complete sequence alignment also applies to quartet sequences which don't share an ambiguity character at that site position.

3.1.8 Definition of ML Substitution Model Parameter (-m Option)

PENGUIN uses P4 ML estimation to calculate the expected number of split pattern and to infer the number of convergent split pattern for the three possible quartet topologies of a given quartet. To do so, P4 uses a defined substitution model of sequence evolution. Under default the GTR model is used for nucleotide data and the WAG model for amino acids. Alternatively, PENGUIN provides optionally four other nucleotide and three other amino acid substitution models (Table 6).

Table 6: Implemented Maximum Likelihood substitution models for nucleotide and amino acid data.

Data Type	Model	Code
Nucleotide		
	General Time-Reversible	GTR (Default)
	Hasegawa, Kishino & Yano 1985	HKY
	Kimura 2-Parameter Model	K2P
	Felsenstein 1981	F81
	Jukes and Cantor 1969	JC
Amino Acid		
	Whelan & Goldman	wag (Default)
	Jones, Taylor, & Thornton	jtt
	Adachi & Hasegawa 1996	mtrev24
	Dayhoff, Schwartz & Orcutt 1978	d78

- **Default: GTR**
- **Command: -m** <**String**>
- **Example : -m JC** (Change GTR to JC)

Note: The specified substitution model has to be congruent to the given type of sequence data. Otherwise, PENGUIN uses the sequence type assigned default model. If so, a warning is printed to the terminal and the common information output file (Penguin_result_info.txt).

3.1.9 Definition of ML α Shape Start Parameter (-a Option)

To calculate the expected number of split pattern and to infer the number of convergent split pattern for the three possible quartet topologies of a given quartet the PENGUIN implemented P4 ML calculations

uses an α shape parameter of rate heterogeneity as start value to optimise branch lengths of a given constraint topology and data set. To change the α shape start parameter use the -a option followed by a float number reflecting the desired start parameter.

- **Default: 0.5**
- **Command: -a <Float>**
- **Example : -a 1.0** (Change α to 1.0)

Note: To use P4 without estimation of among-site rate variation (ASRV) set $\alpha = 100$ ('-a 100'). P4 will use just one rate category instead of four and the proportion of invariabel sites will be set to $I = 0.0$ independent of given paramter value under '-I' option.

3.1.10 Definition of ML pINV Start Parameter (-I Option)

To calculate the expected number of split pattern and to infer the number of convergent split pattern for the three possible quartet topologies of a given quartet the PENGUIN implemented P4 ML calculations uses an extra proportion of invariabel sites as start value to optimise branch lengths of a given constraint topology and data set. To change the the proportion of invariabel site parameter use the -I option followed by a float number reflecting the desired start parameter.

- **Default: 0.3**
- **Command: -I <Float>**
- **Example : -I 0.1** (Change pINV to 0.1)

Note: If P4 is used without estimation of among-site rate variation (ASRV) set ('-a 100'; see section '-a option'). P4 will use just one rate category instead of four and the proportion of invariabel sites will be set to $I = 0.0$ independent of given paramter value under '-I' option.

4 Output/Result Files

After analysing all single quartets PENGUIN prints following split result information of performed quartet analyses as graphic (SVG) and/or text (TXT) formatted output files:

- **Single split support** values of the three possible 4 clan relationships **for each analysed quartet** ↪ SVG & TXT
- **Mean and median split support** values of the three possible clan relationships **overall** analysed quartets ↪ SVG & TXT
- **Mean split support** values of the three possible clan relationships **for each analysed taxon** sequence ↪ SVG & TXT
- **Median split support** values of the three possible clan relationships **for each analysed taxon** sequence ↪ SVG
- **Number of observed best and second best split support for each of the three possible clan relationships** overall single quartet analyses ↪ SVG
- **General information** on analysed 4 taxa sequence combinations, like number of rejected site positions, subsequently excluded quartet combinations, and number of best quartet split support ↪ TXT

According to the respective type format, Output files (Table 7 and 14) are printed to one of two new generated result folders, named **Penguin_SVG** & **Penguin_TXT**.

4.1 Text Formatted Result Output Files

Table 7 lists text (TXT) formatted output file names, printed to result folder **Penguin_TXT** .

Table 7: Name and information content of text formatted output files.

Output File (.txt)	Information Content
Penguin_result_info	General analysis info & main split results
Penguin_qsingle_spl_sup	List of single quartet split support for possible clan relationships
Penguin_tsingle_mean_sup	List of mean taxon split support for possible clan relationships
Penguin_tsingle_median_sup	List of median taxon split support for possible clan relationships
Penguin_qsingle_rej_pos	List of rejected site positions in single quartet analyses

4.1.1 General Analysis Info & Main Split Results (*Penguin_result_info.txt)*

This result file gives an general overview about chosen PENGUIN parameter setup values (Table 8). It contains also information about identified sequence states, taxa assignments, and overall quartet performance (Table 9), as well as identified mean and median split support values, observed from all single quartet analyses between sequences of defined clans for each of the three possible clan relationships (Table 10).

Table 8: Example output of the Penguin_result_info.txt parameter setup section. A maximum number of 5000 single quartet analyses has been performed with split pattern analyses of state characters in combination with ML estimations under the GTR model of sequence evolution and defined α shape and invariabel site proportion start values.

PENGUIN Setup	
Indel/Amb Sites:	rejected (single)
Pattern Handling:	states
Substitution Model:	GTR
Start Alpha (ML):	0.5
Start pINV (ML):	0.3
Maximum Limit Quartets:	5000
Minimum Sequence Length:	4500
Clan Definition Inputfile:	*clan_infile*.txt
MSA Quartet Inputfile:	*alignment_inputfile*.fas

Table 9: Example output of the Penguin_result_info.txt sequence info section. Example input alignment consists of nucleotide data with sequence length of 11687 base pairs. Identified clan names of the clan definition input file (e.g. *clan_3*) are listed according to their internally assigned clan number (e.g. *clan_3* has been coded as "Clan 3"). The total number of single performed quartet analyses is 5000. Mean number of remaining site positions after site exclusion of unallowed character states is 4599 base pairs. The mean length of rejected quartets due to reduced sequence lengths below the minimum number of allowed sequence states is 4426.

MSA Sequence Type:	nuc
MSA Sequence Length:	11687
READ IN Clan File:	*clan_infile*.txt
Defined Clans:	
Clan 1:	*clan_1*
Clan 2:	*clan_2*
Clan 3:	*clan_3*
Clan 4:	*clan_4*
N Single Quartet-Analyses:	5000
Mean Remaining Site Positions:	4599
Mean Sites / Rejected Quartet:	4426

Table 10: Example output of the Penguin_result_info.txt overall split support result section. Overall mean and median split support value of each possible clan relationship ($Q1$, $Q2$, $Q3$) based on split values of all single performed quartet analyses.

Overall Split Signal (**Mean**):	
Q1 (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.41547204286773
Q2 (*clan_1*,*clan_3*),(*clan_2*,*clan_4*):	0.0158977209470447
Q3 (*clan_1*,*clan_4*),(*clan_2*,*clan_3*):	0.568630236185225
Overall Split Signal (**Median**):	
Q1 (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.345102028049545
Q2 (*clan_1*,*clan_3*),(*clan_2*,*clan_4*):	0
Q3 (*clan_1*,*clan_4*),(*clan_2*,*clan_3*):	0.654897971950455

4.1.2 List of Single Quartet Split Support Values (*Penguin_qsingle_spl_sup.txt)*

This result file lists all single quartet identified split support values for each of the three possible clan relationships ($Q1$, $Q2$, $Q3$) in relation to each other. Each line presents split support scores given a single analysed set of four taxon sequences. An example is given in Table 11.

Table 11: Example output of the Penguin_qsingle_spl_sup.txt result file, listing split support values of each quartet analysis (q_n) of each of the three possible clan relationships ($Q1$, $Q2$, $Q3$) in relation to each other.

Quartet Number	Seq. Combination	Split Support of Clan Relationship $Q1$, $Q2$, $Q3$		
Split Analysis q_1:	T1A:T2A:T3A:T4A	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.1375	Split Support $Q2$...
Split Analysis q_2:	T1A:T2A:T3A:T4B	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.4411	Split Support $Q2$...
Split Analysis q_3:	T1A:T2A:T3A:T4C	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.2490	Split Support $Q2$...
Split Analysis q_4:	T1A:T2A:T3A:T4D	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.7543	Split Support $Q2$...
⋮				
Split Analysis q_n:	T1n:T2n:T3n:T4n	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.9107	Split Support $Q2$...

4.1.3 List of Mean Sequence Split Support Values (*Penguin_tsingle_mean_sup.txt)*

Each line of this result file presents the mean identified split support score for each of the three possible quartet relationships ($Q1$, $Q2$, $Q3$), given all quartet analyses of a single taxon sequence (s_n). An example is given in Table 12.

Table 12: Example output of the Penguin_tsingle_spl_sup.txt result file, listing for each analysed sequence (s_n) the identified mean split support value for each of the three possible clan relationships ($Q1$, $Q2$, $Q3$) in relation to each other. For example, split analysis s_1 lists mean split support values of all quartet analyses with sequence T1A (defined as *clan_1* sequence) for $Q1$, $Q2$, $Q3$.

Quartet Number	Seq.	Split Support of Clan Relationship $Q1$, $Q2$, $Q3$		
Split Analysis s_1:	T1A	Clan: *clan_1* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.3494	$Q2$ ...
Split Analysis s_2:	T1B	Clan: *clan_1* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.4154	$Q2$ ...
Split Analysis s_3:	T1C	Clan: *clan_1* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.5609	$Q2$ ...
Split Analysis s_4:	T2A	Clan: *clan_2* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.1245	$Q2$ ...
⋮				
Split Analysis s_n:	T4n	Clan: *clan_4* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.4154	...

4.1.4 List of Median Sequence Split Support Values (*Penguin_tsingle_median_sup.txt)*

As described in section 4.1.3, just for median split scores.

4.1.5 List of Rejected Site Positions & Excluded Quartets (*Penguin_qsingle_rej_pos.txt)*

Each line of this result file presents the mean identified split support score for each of the three possible quartet relationships ($Q1$, $Q2$, $Q3$), given all quartet analyses of a single taxon sequence (s_n). An example is given in Table 13.

Table 13: Example output of the Penguin_qsingle_rej_pos.txt result file, listing for each analysed quartet (q_n) the number of rejected site positions of unallowed character states, the remaining number of sites, and if remaining number of sites is below minimum defined sequence length, and therefore, has been rejected. Rejected quartets are further listed separately after site statements have been listed for each single quartet.

Quartet Number	Seq. Combination	N Rejected Quartet Positions	N Remaining Quartet Positions	If Below Allowed Sequence Lengths...
Split Analysis q_1:	T1A:T2A:T3A:T4A	Rejected Quartet Positions: 7185	Remaining Quartet Positions 4502	
Split Analysis q_1:	T1A:T2A:T3A:T4B	Rejected Quartet Positions: 7281	Remaining Quartet Positions 4406	rejected
Split Analysis q_1:	T1A:T2A:T3A:T4C	Rejected Quartet Positions: 7143	Remaining Quartet Positions 4544	
Split Analysis q_1:	T1A:T2A:T3A:T4D	Rejected Quartet Positions: 7241	Remaining Quartet Positions 4446	rejected
⋮				
Split Analysis q_1:	T1n:T2n:T3n:T4n	Rejected Quartet Positions: 7099	Remaining Quartet Positions 4588	
Rejected Quartets:				
q_4:	T1A:T2A:T3A:T4D			
⋮				
q_z:	T1x:T2y:T3w:T4z			

Note: The information output file about rejected site positions and quartets below allowed sequence length after site exclusion (Penguin_qsingle_rej_pos.txt) is only printed if site exclusion of unallowed character states take place separately for each 4-sequence quartet combination. If site exclusion happens in one go along the complete sequence alignment (see "-N" option), the number of overall sequences excluded site positions is only listed in the General analysis info & main split results output file (Penguin_result_info.txt).

4.2 SVG Formatted Graphic Result Output Files

Following SVG formatted output files are printed to result folder **Penguin_SVG** :

Table 14: Name and information content of SVG formatted output files.

Output File (.svg)	Information Content
Penguin_split_qmean_sup	Mean split support graphic for each clan relationship
Penguin_split_qmedian_sup	Median split support graphic for each clan relationship
Penguin_split_qnumber_best	Number of quartets supporting best for each clan relationship
Penguin_split_qnumber_second_best	Number of quartets supporting second best for each clan relationship
Penguin_triangle_qsingle_spl_sup	Triangle graphic, presenting split support values of analysed quartet
Penguin_triangle_tsingle_mea_sup	Triangle graphic, presenting mean split support of each sequence
Penguin_triangle_tsingle_mea_sup	Triangle graphic, presenting median split support of each sequence

Note: All SVG output files are interactive vector graphics, meaning that additional data information will become visible if the mouse curser points on specific data points. Data points with additional information are single described in following graphic subsections. Furthermore, not all vector applications support interactive vector graphics. All example figures are posed by the `Gapplin` vector application tool for Mac OS.

4.2.1 Mean Split Support SVG Output (*Penguin_split_qmean_sup.svg)*

Split-Network graph of mean split support values for each of the three possible quartet relationships ($Q1$, $Q2$, $Q3$), based on all single processed quartet analyses. The best supported split is presented proportionally to its scoring by horizontal lines, the second best split support by vertical lines above the most inner branch, and the lowest split support by vertical lines below the internal branch (Figure 2).

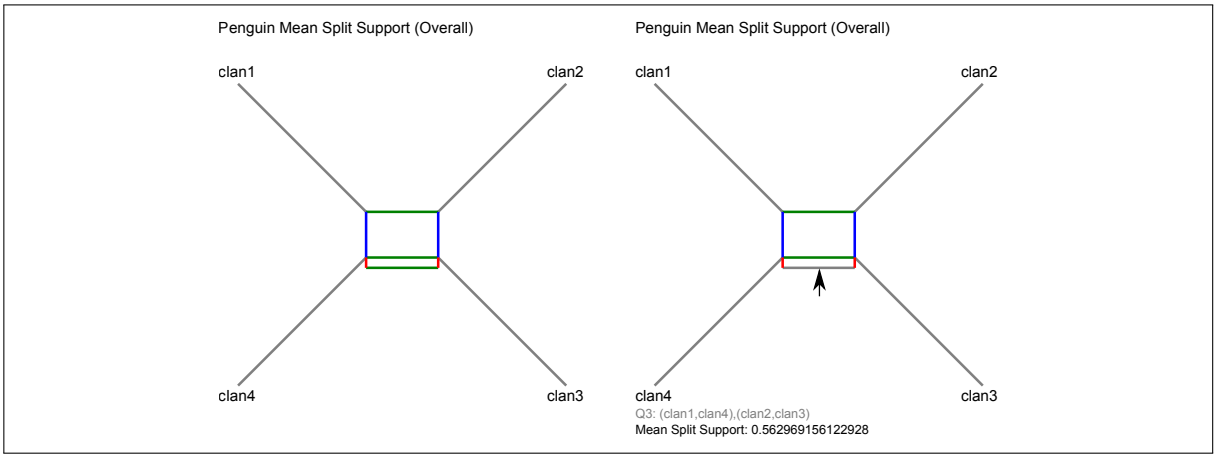


Figure 2: Example Split-Network output presenting normalised mean score values for each of the three possible clan relationships, obtained from all single single quartet analyses. Highest mean split score is shown by horizontal lines (blue), the second highest mean split by vertical lines above the inner branch (red), and the lowest split support below the internal branch (green) (figure left). Single clan relationships and assigned mean split support values are highlighted if mouse courser is pointed on corresponding split line (figure right).

4.2.2 Median Split Support SVG Output (textitPenguin_split_qmedian_sup.svg)

Split-Network graph of median split support values for each of the three possible quartet relationships ($Q1$, $Q2$, $Q3$), based on all single processed quartet analyses. Split vector graph with the same presentation as shown in Figure 2 for mean split support values.

4.2.3 N Best Quartet Support SVG Output (*Penguin_split_qnumber_best.svg)*

Network graph on the total number of single quartets whose highest split value support a possible quartet relationships ($Q1$, $Q2$, $Q3$). Split vector graph with the same presentation as shown in Figure 2 for mean split support values.

4.2.4 N 2^{nd} Best Quartet Support SVG Output (*Penguin_split_qnumber_second_best.svg)*

Network graph on the total number of single quartets whose second highest split value support a possible quartet relationships ($Q1$, $Q2$, $Q3$). Split vector graph with the same presentation as shown in Figure 2 for mean split support values.

4.2.5 Single Quartet Split Support SVG Output (*Penguin_triangle_qsingle_spl_sup.svg)*

Triangle plot of single split support scores of each analysed quartet combination for each of the three possible clan relationships ($Q1$, $Q2$, $Q3$). Identified split support of each single quartet result is coded by a blue dot. The overall mean support is marked by a red dot, the median support by an orange dot (Figure 3). Additional data information is highlighted if mouse courser points on assigned data point (Figure 3b, c, d, e). The higher the support for a given clan relationship, the closer is the support value to one of the three quartet relationship assigned corners ($Q1$, $Q2$, $Q3$) of the triangle.

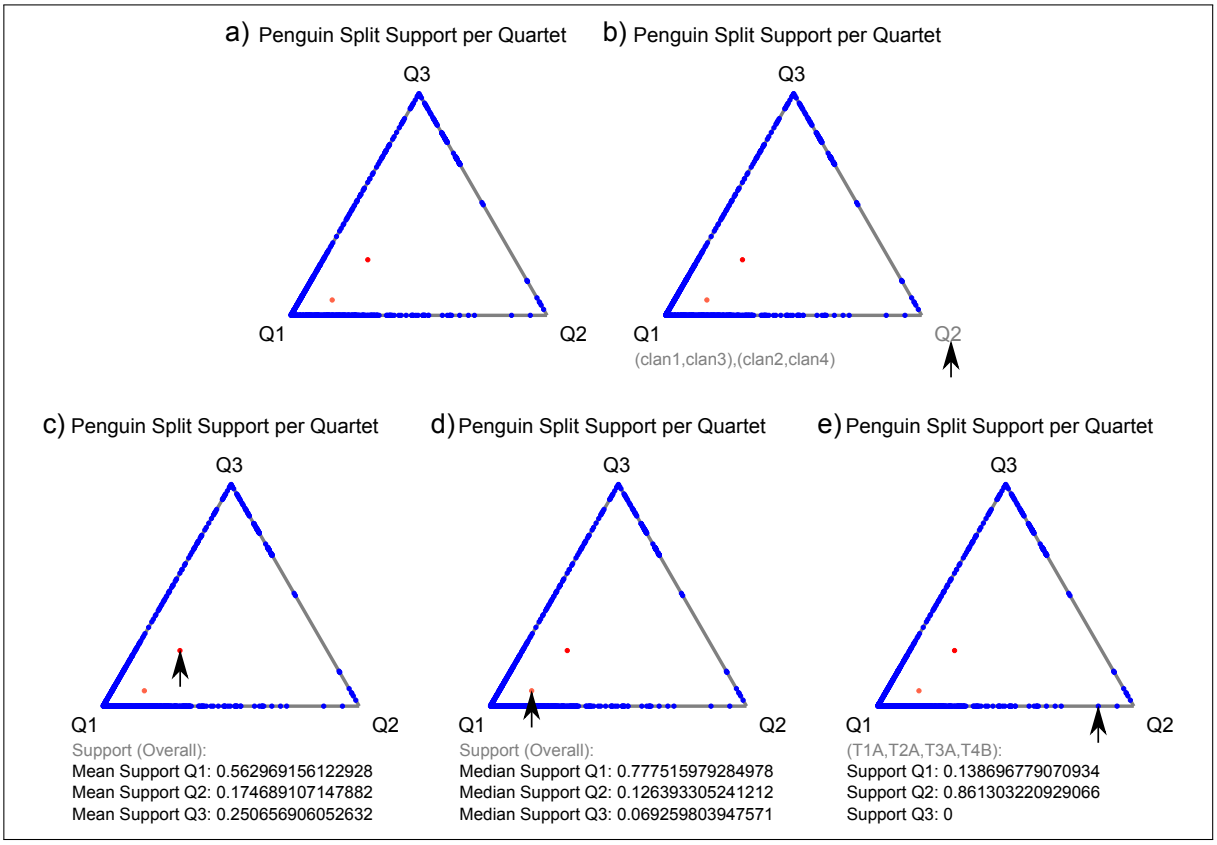


Figure 3: Example Triangle-Graph output presenting normalised mean, median, and single quartet score values for each of the three possible clan relationships ($Q1$, $Q2$, $Q3$). Single quartet split scores are presented by blue dots, mean split support by a red dot, and median support by an orange dot. SVG graphic is interactive. Additional data information available if mouse courser is pointed on data corresponding vector sections. a) without additional information, b) phylogenetic clan relationship of corner defined quartet trees ($Q1$, $Q2$, $Q3$), c) highlighted mean split support, d) highlighted median split support, e) highlighted single quartet split support values.

Note: Single split support values for each of the three possible clan relationships ($Q1$, $Q2$, $Q3$) are given in relation to each other, whereas the lowest split support value is set to zero (Figure 3e). Therefore, single quartet support values are always plotted on the triangle edge between corresponding best and second best split support of a clan relationship. The higher the difference between both values, as closer is the support plotted next to the highest supported clan relationship.

4.2.6 Mean Sequence Split Support SVG Output (*Penguin_triangle_tsingle_mea_sup.svg)*

Triangle plot of mean support scores of each analysed sequence due to the sequence corresponding single split support values, observed from sequence corresponding quartet analyses. Mean sequence support values are coloured based on given clan membership. First defined clan: blue dots, second: yellow, third: green, fourth: purple (Figure 4). Overall mean support is marked as red dot. Additional data information is highlighted if mouse courser points on assigned data point (Figure 4b, c, d). The higher the support for a given clan relationship, the closer is the support value to one of the three quartet relationship assigned corners ($Q1$, $Q2$, $Q3$) of the triangle.

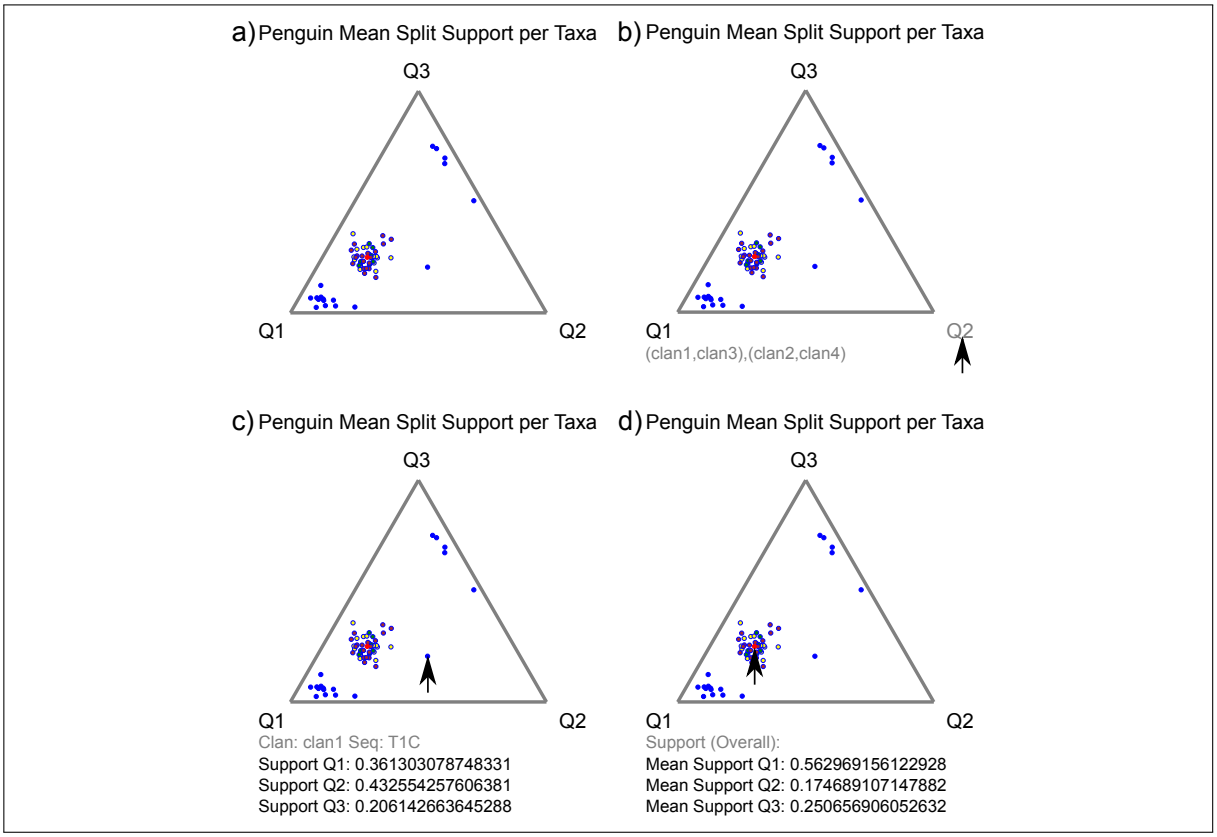


Figure 4: Example Triangle-Graph output presenting normalised mean split scores of single sequences for each of the three possible clan relationships ($Q1$, $Q2$, $Q3$). Single mean split scores observed for a specific sequence via all single quartet analyses are presented by dots, coloured by corresponding clan membership (First defined clan: blue, second clan: yellow, third clan: green, fourth clan: purple). Mean split support overall single quartet analyses is presented as red dot. a) without additional information, b) phylogenetic clan relationship of corner defined quartet trees ($Q1$, $Q2$, $Q3$), c) highlighted mean sequence support, d) highlighted mean split support.

4.2.7 Median Sequence Split Support SVG Output (*Penguin_triangle_tsingle_med_sup.svg)*

Triangle plot of median support scores of each analysed sequence due to the sequence corresponding single split support values, observed from sequence corresponding quartet analyses. Like described for mean triangle graph, median sequence support values are coloured based on given clan membership. First defined clan: blue dots, second: yellow, third: green, fourth: purple (Figure 4). Overall median support is marked as red dot. Additional data information is highlighted if mouse courser points on assigned data point (Figure 4b, c, d). The higher the support for a given clan relationship, the closer is the support value to one of the three quartet relationship assigned corners ($Q1$, $Q2$, $Q3$) of the triangle.

5 Example Input & Output Files

In addition to the actual PENGUIN software script, the compressed download file 'Penguin.zip' includes an example input file package in an own subfolder named 'example_setup'. This example setup consists of an input alignment file, including 61 correctly aligned sequences (named from T1 to T61) of 30,000bp length and without any unallowed character states. The nucleotide data set has been simulated with the INDELible software (Fletcher and Yang, 2009) using the GTR model and a mixed-distribution model of among-site rate variation. Given the corresponding clan definition file, sequences are divided into four clans (named 'clan1' to 'clan4').

- clan1,T1,. . . ,T15 (15 taxa)

- clan2,T16,. . . ,T18 (3 sequences)
- clan3,T19,. . . ,T33 (15 sequences)
- clan4,T34,. . . ,T61 (28 sequences)

The data underlying topology, the four defined sequence clans of the example clan definition file, and the internal branch assigned split analysis of corresponding clan relationships ($Q1$, $Q2$, $Q3$) are displayed in Figure 5. A simplified flowchart of input, main analysis, and SVG output processes is given by Figure 6.

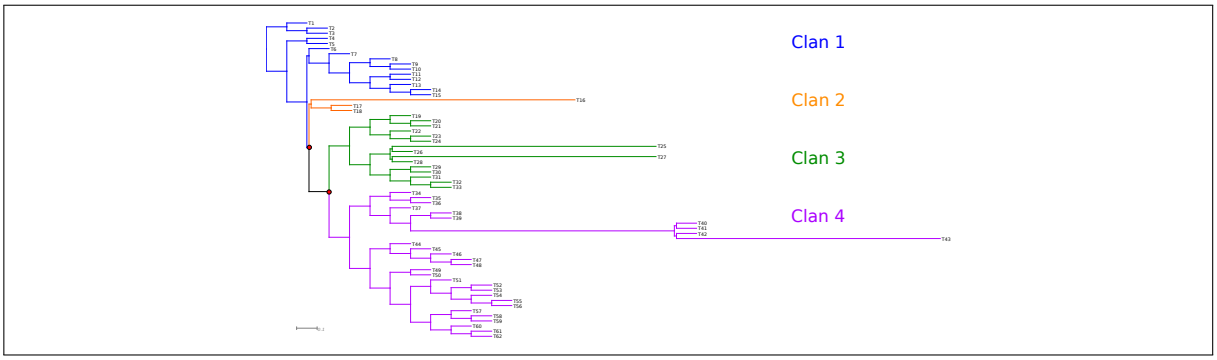


Figure 5: Example MSA input data underlying topological relationship between 61 sequences, divided into four different clans. Internal branch assigned split support analysis based on defined sequence clans is flanked by red dots.

Note: Split support analyses of other existing or non-existing internal branch relationships can be easily performed by changing the clan assignments in the example clan definition file.

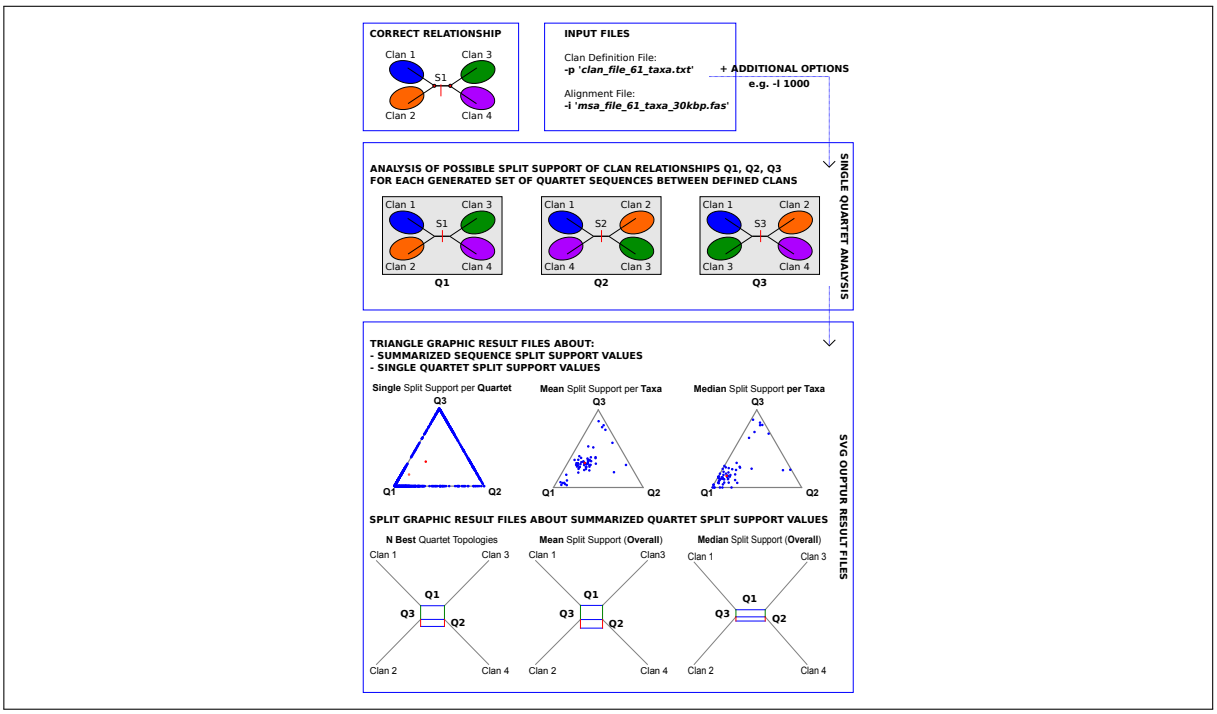


Figure 6: Simplified flowchart of input, main analysis, and SVG output processes.

6 License/Help-Desk/Citation

PENGUIN was developed and has been written in Perl by Patrick Kück in 2015. It is implemented in Perl and a free software. It can be distributed and/or modified under the terms of the GNU General Public

License as published by the Free Software Foundation; either 2 of the license, or (at your option) any later version.
This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.
You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139, USA.

If you have any problems, error-reports or other questions about PENGUIN feel free and write an email to patrick_kueck@web.de which is the official help desk email account for the software. For other open source software visit also:

http://nhm.ac.uk/

https://www.zfmk.de/en/research/research-centres-and-groups/software

If you use PENGUIN please contact Patrick Kück until the manuscript addressing PENGUIN is published.

References

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**(5), 1792–1797.

Felsenstein, J. (1993). *PHYLIP: phylogenetic inference package*. Department of Genetics, University of Washington, Seattle, USA, version 3.5c edition.

Fletcher, W. and Yang, Z. (2009). INDELible: A flexible simulator of biological sequence evolution. *Mol Biol Evol*, **26**(8), 1879–1888.

Foster, P. G. (2004). Modeling compositional heterogeneity. *Syst Biol*, **53**, 485–495.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30**(4), 772–780.

Katoh, K. and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*, **9**(4), 286–298.

Katoh, K. and Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, **26**(15), 1899–1900.

Katoh, K., Kuma, K.-i., Hiroyuki, T., and Miyata, T. (2005). MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33**(2), 511–518.

Kück, P. and Longo, G. C. (2014). FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front Zool*, **11**, 81.

Kück, P. and Meusemann, K. (2010). FASconCAT: Convenient handling of data matrices. *Mol Phylogenet Evol*, **56**, 1115–1118.

Notredame, C. (2002). Recent progresses in multiple sequence alignment: a survey. *Pharmacogenomics*, **3**(1), 1–14.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-COFFEE: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–217.

7 Copyright