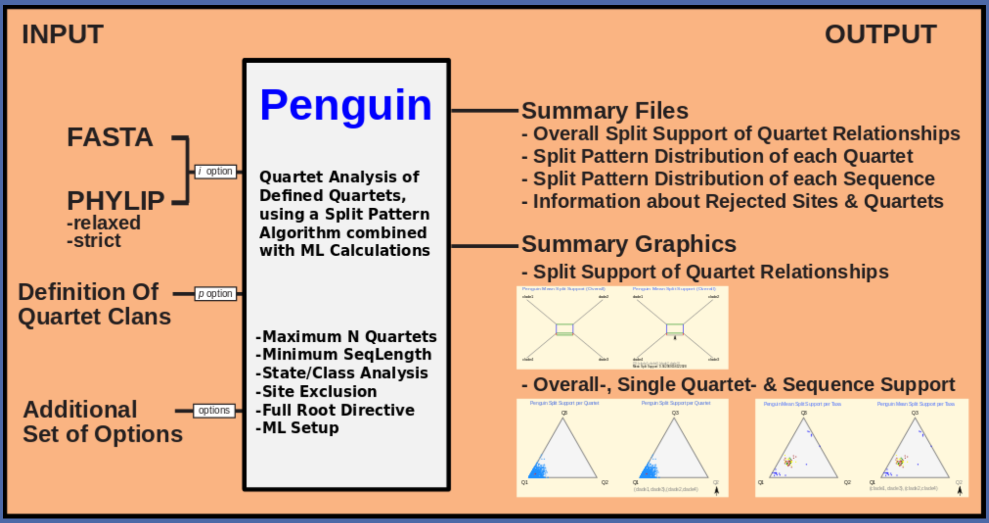
FORSCHUNGS museum KOENIG

Patrick Kück, Nathan I. Seidel



INPUT

FASTA

PHYLIP
-relaxed
-strict

i option

Definition Of Quartet Clans

p option

Additional Set of Options

options

Penguin

Quartet Analysis of Defined Quartets, using a Split Pattern Algorithm combined with ML Calculations

-Maximum N Quartets
-Minimum SeqLength
-State/Class Analysis
-Site Exclusion
-Full Root Directive
-ML Setup

OUTPUT

Summary Files
- Overall Split Support of Quartet Relationships
- Split Pattern Distribution of each Quartet
- Split Pattern Distribution of each Sequence
- Information about Rejected Sites & Quartets

Summary Graphics
- Split Support of Quartet Relationships
- Overall-, Single Quartet- & Sequence Support

PENGUIN v2.0

November 2020

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 support for a given directive quartet relationship, the Maximum Likelihood (ML) approach of the P4 package (**?**) is used to estimate the expected number of convergent 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 polarized quartet topology with the highest phylogenetic signal is assumed to be the correct one.

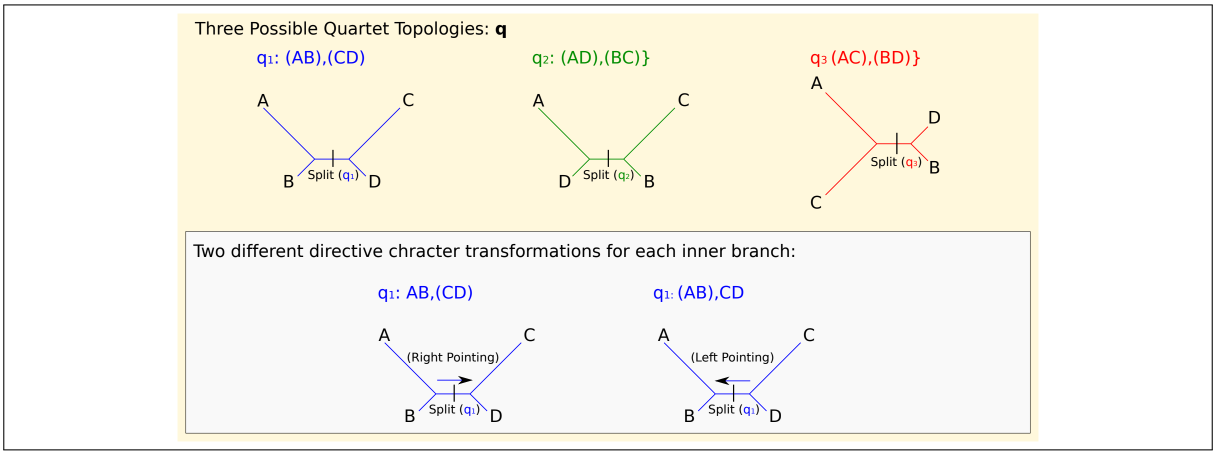


Figure 1: Definition of polarity (direction of character transformation) along the internal branch of a given quartet tree, e.g. (AB),(CD). (AB),CD indicate the direction is towards CD and towards AB for (AB),CD. Assuming polarities enables a classification of split-support into potentially phylogenetically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

Given a set of sequences characterized into four different clans, PENGUIN performs site analyses for all or a maximum number of defined quartets between sequences (each representing one of four clans). Once all sequence-quartet combinations are analysed, aggregate mean and median support of polarized 4-clan trees are individually extracted from the corresponding set of supporting sequence-quartets (PhyQuart-mapping). Normalised for each polarized 4-clan tree, the two aggregated values (mean and median) are taken as final support measures, whereby normalised support ranks between zero and one with the lowest tree support of a sequence-quartet analysis set to zero. Aggregate (mean and median) support of each of the six different polarized 4-clan trees forms the basis for the graphical result output of the PhyQuart-mapping process and the evaluation of larger multiple-clan trees given a number of defined clans greater than four. If an outgroup-clan is specified, aggregate support is printed for the three outgroup-clan polarized 4-clan trees. Without the definition of an outgroup-clan, best aggregate sequence-quartet support of both tree polarizations are individually combined for each of the three unpolarized 4-clan trees.

Additionally to the phylogenetic evaluation of polarized 4-clan trees, the Penguin approach can be used for supertree reconstruction or evaluation based on a pairwise compatibility matrix of inferred polarized support given different 4-clan combinations of an unlimited number of predefined multiple-taxon clans (Icebreaker process). Given an outgroup-clan, only outgroup-clan including 4-clan combinations are analysed. Otherwise, all possible 4-clan analyses are conducted. The evaluation of multiple-clan trees consists the generation of all possible rooted topologies (with an outgroup-clan defined only outgroup-clan rooted trees are generated) for a set of defined clans and the overall polarized support measure of each topology based on maximum-bicliques of pairwise polarized compatible 4-clan support. After the tree analysis, icebreaker evaluated multiple-clan relationships are sorted in descending order according to their assigned support and final support of the best evaluated multiple-clan tree is calculated from the normalised score difference between the best-three supported trees.

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 at least four predefined clans of one or more sequences each must be provided in plain TEXT format. If not otherwise specified, PENGUIN analyses all possible sequence-quartet combinations up to (per default specified) 10,000 quartets between the four predefined taxa clans of each analysis relevant 4-clan combination. There is no upper limit for the number of defined clans. Nevertheless, due to the strongly increasing number of different tree possibilities and thus computation time in relation to the number of specified clans, we suggest to define at maximum ten clans. PENGUIN does not allow multiple sequence-name records within given input file(s), while sequences of disaccording sequence-names between a predefined clan definition file and a corresponding multiple sequence alignment are just left un-analysed. Sequence sites with indel ('-'), ambiguity or missing characters are always excluded from the analysis. PENGUIN excludes all forbidden site positions (positions with gaps, ambiguities or missing characters) separately for each sequence quartet of a given alignment.

1.3 Penguin Output

PENGUIN calculates summarised output information of identified support for each possible sequence-quartet relationship and 4-clan combination as well as for the 100 (default) best supported multiple-clan trees. Obtained discrepancies in topological support of the three possible quartet topologies are also presented as split network (for each 4-clan combination) and triangle graph (4-clan and multiple-clan analysis). A further vector graphic shows the number and percentage distribution of best and second best resolved quartet relationships (for each 4-clan combination).

1.4 Implementation & Usage

PENGUIN is a command-line driven program written in `C++` and works on Macs and Linux operating systems, offering the opportunity of process parallelization using OpenMP as application programming interface (API). Thus, 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. PENGUIN does not run on WindowsPCs, because of installation incompatibilities with requisite P4 python packages. Calculations of the PENGUIN software program are generally very fast and can be easily executed on a normal desktop computer, even if data sets consist of phylogenomic data. However, the computation time increases strongly with the number of defined clans and the number of possible sequence-quartet combinations. In such cases, we recommend to execute PENGUIN on a computational stronger server system.

2 Pre-Installations

To execute PENGUIN, a gcc/g++ compiler version 8.4 or higher, openMP as well as P4, a Python package for Maximum Likelihood and Bayesian analysis of molecular sequences, must be installed on the current run system. The first two system-requirements should be pre-installed on modern operating systems. The P4 package has to be installed manually.

2.1 Compilation of PENGUIN Binaries

PENGUIN binaries can be compiled via a shell script (compile_penguin.sh), provided with the PENGUIN binaries. To compile PENGUIN binaries with the shell script move to the path where the binaries and the shell script are located and type:

- user@linux:˜$ sh compile_penguin.sh <enter>

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 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:˜$ ./penguin -h <enter> ↪ help menu
- user@linux:˜$ ./penguin -i *path/infile* -p *path/clanfile* <enter> ↪ start PENGUIN under default

Note: Make sure you write the input options correctly, for example ”-i” and not ”- i”. Otherwise PENGUIN will not start working but instead opens the Help menu.

3.1 Options

Table 1: Overview of features and options

Info options	**Command**		**Example**	
Help menu	h		perl	./penguin -h
Preface	P		perl	./penguin -P
Parameter options		**default**		
Sequence alignment input file	i	*none*	-i	*filename*.phy
Clan-definition input file	p	*none*	-p	*filename*.txt
Min. sequence length for quartets	M	500	-M	<Integer>
Max. number of quartets per 4-clan combination	l	10000	-l	<Integer>
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>
Outgroup-uninformative 4-clan analysis	r	outgroup-informative	-r	
Translate site states to class States	c	keep site states	-c	
Outgroup-clan classification	o	*none*	-o	<string>
Number of computational threads	t	1	-t	<integer>
Number of print issued best multiple-clan tree support	b	100	-b	<integer>
Elimination of script queries	u	No	-u	
Specification of a customized output path	pre	PENGUIN home directory	-pre	<path_dir>
Further analysis of already existing sub-results	restart	*none*	-restart	
Removal of single sequence-quartet depending output for restart	slim	*none*	-slim	

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.

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 as well as code corresponding ambiguity characters, ”?”, ”X”, 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 (**????**), MUSCLE (**?**), or T-COFFEE (**??**). To convert aligned sequence files from other alignment formats to FASTA, software tools like T-COFFEE (**??**), FASconCAT (**?**) or FASconCAT-G (**?**) can be used. FASconCAT (**?**) and FASconCAT-G (**?**) 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` (**?**) `T-COFFEE` (**??**). To convert aligned sequence files in FASTA format software tools like `T-COFFEE` (**??**), `FASconCAT` (**?**) or `FASconCAT-G` (**?**) can be used. `FASconCAT` (**?**) and `FASconCAT-G` (**?**) 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 quartet-sequence alignment, PENGUIN can also calculate 4-clan and multiple-clan tree support of user-defined clans. The 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. 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 sequence-names. Be aware that all sequence-names are identical to the corresponding sequence-names of the sequence input file (case sensitive!). Blanks between comma-separated sequence-names are not allowed! A sequence-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,Sequence_name_1,Sequence_name_2,Sequence_name_3,Sequence_name_4
Name_Clan_2,Sequence_name_a,Sequence_name_b,Sequence_name_c
Name_Clan_3,Sequence_name_5,Sequence_name_6,Sequence_name_7,Sequence_name_8,Sequence_name_9
Name_Clan_4,Sequence_name_V,Sequence_name_X,Sequence-name_Y,Sequence-name_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 sequence-quartets if the remaining sequence length of a given set of four sequences 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 sequence-quartet and skips to the next sequence-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 sequence-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 can generate all possible quartets between defined sequences given a clan definition file (see section 3.1.2: -p option). The number of generated sequence-quartets increases strongly with the number of sequences defined for each clan. Therefore, the total computation time can be very time consuming if the number of sequence-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 (based on the usage of a single thread). To avoid overstrain computation times PENGUIN allows the definition of a maximum limit of generated sequence-quartets with the default value set to 10,000. PENGUIN stops if the number of possible sequence-quartets exceeds the maximum limit of allowed quartets with a terminal user request offering 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 sufficiently high, e.g. -M 10000000000000.

3.1.5 Elimination of Script Queries (-u Option)

If the maximum number of possible sequence-quartet combinations increases the number of allowed sequence-quartet analyses, PENGUIN stops under default with a user query how to proceed further (see section 3.1.4: -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 sequence-quartets in default by using IUPAC standard nucleotide coded site pattern distributions. To analyse sequence-quartets in a more conservative way PENGUIN offers the possibility to reduce number of possible site characters from four (A, C, G, T) to two by summarising nucleotide characters to purines (R) and pyrimidines (Y) and for amino acid data from 20 to two by summarising characters to hydrophobic (R) and hydrophilic (Y) classes. To compress site states to class states type -c.

- **Default: keep original characters**
- **Command: -c**
- **Example : -c** (translate site states to class states)

3.1.7 Outgroup-Clan Definition (-o Option)

Use the '-o' option to perform a PENGUIN outgroup-informative split analysis based on a defined outgroup-clan. Without the definition of an outgroup-clan, each clan is respectively analysed to be the potential outgroup clan. To perform an outgroup-uninformative 4-clan analysis (see section 3.1.11) use the '-r' option without the specification of an outgroup-clan.

- **Default: no outgroup-clan defined**
- **Command: -o <string>**
- **Example : -o outgroup_clan** (define clan outgroup_clan as outgroup)

Note: Due to a lack of testing yet, we blocked the possibility of analysing each clan as potential outgroup in the actual PENGUIN beta version. Given the beta version, the classification of an outgroup clan is mandatory for the phylogenetic analysis of more than four clans.

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 sequence-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 2001	wag (Default)
	Le & Gascuel 2008	LG
	Henikoff & Henikoff 1992	BLOSSUM62
	Jones, Taylor, & Thornton 1992	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 sequence-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 invariable sites will be set to $I = 0.0$ independent of given parameter 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 sequence-quartet the PENGUIN implemented P4 ML calculations uses an extra proportion of invariable sites as start value to optimise branch lengths of a given constraint topology and data set. To change the proportion of invariable 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)

3.1.11 Outgroup-Uninformative 4-clan Analysis (-r Option)

Under default, PENGUIN calculates outgroup-informative tree support for each polarized 4-clan tree of a 4-clan combination. If an outgroup-clan is specified, only sequence-quartet of outgroup-clan including sequences are analysed. The total quartet support of possible 4-clan relationships is then calculated by a separate summarization of polarized split scores until all single sequence-quartet analyses of a 4-clan combination have been performed. For the final determination of total quartet support for each possible 4-clan relationship, PENGUIN only uses quartet information with respect to the totally better supported polarized assumption of a given tree hypothesis. These scores are further used to determine the support of multiple-clan trees (with number of clans greater than four). Using the -r option without the definition of an outgroup-clan induces a summarisation of each of the two (mean or median) polarized sequence-quartet support values of an outgroup-uninformative 4-clan tree for each of the three possible sister-pair relationships. Therefore, the overall sampling of individually identified sequence-quartet support of each outgroup-uninformative 4-clan tree is based on a mixture of polarized 4-clan relationships corresponding mean or median supported polarized quartet-trees. Thus, the -r option (which only works for a defined set of four clans without specification of an outgroup-clan) is an appropriate measure of 4-clan tree support of pairwise related sister-clans, sharing an outgroup-uninformative signal along the internal branch connecting both groups.

- **Default: outgroup-informative**
- **Command: -r**
- **Example : -r** (outgroup-uninformative)

Note: The '-r' option is restricted to a single 4-clan analysis without additional specification of an outgroup-clan (see section 3.1.7: -o option). 4-clan combinations of a multiple-clan analysis (with number of clans greater than four) are always outgroup-informative analysed.

3.1.12 Number of Computational Threads (-t Option)

The process time of a PENGUIN analysis strongly depends on the number of single investigated sequence-quartets in conjunction with the number of utilized threads. PENGUIN offers the opportunity of process parallelization using OpenMP, an implementation of multithreading with the number of specified threads running concurrently. To start processing in parallel the minimum number of specified thread availability is two (default). The upper bound of multithreading is merely restricted by the total number of available CPU's. To change the number of threads use the -t option followed by an integer number.

- **Default: 2**
- **Command: -t <integer>**
- **Example : -t 4** (change number of threads to 4)

3.1.13 Number of Print issued Best Multiple-clan Tree Support (-b Option)

Both the number of 4-clan analyses and the number of multiple-clan tree analyses increase strongly with the number of defined clans. For example, with one clan defined as outgroup the number of rooted multiple-clan tree evaluations for eight clans is 10,395. After the Icebreaker process of outgroup-informative multiple-clan evaluation, PENGUIN prints a list of individual tree support of the best 100 topologies (default) in descending order. To change number of listed trees use the -b option followed by an integer number.

- **Default: 100**
- **Command: -b**
- **Example : -b 10** (change number of listed multiple-clan tree support to 10)

3.1.14 Path Specification of the Output-Folder (-pre Option)

The -pre option redirects all PENGUIN output from its home directory to the specified (existing or yet nonexistent) path. PENGUIN builds the new path structure if the specified path does not exist.

- **Default: PENGUIN_home_directory/**
- **Command: -pre**
- **Example : -pre new_path** (change PENGUIN output from home directory to specified path)

3.1.15 Further Analysis of already existing Sub-Results (-restart Option)

The -restart option allows both, a more extensive analysis based on an already existing set of analysed sequence-quartets and the re-analysis of a system aborted PENGUIN process. In both cases, preexisting sub-results of already analysed sequence-quartets (which could have already caused long computation time) are read again and further processed by PENGUIN without significant loss of time.

- **Default: start new analysis**
- **Command: -restart**

- **Example : -restart** (start of a forward analysis based on already existing sub-results)

Note: The -restart option needs the same PENGUIN parameter settings as used for the initial analysis. Be aware that -restart is specified as first parameter in the command line.

3.1.16 Removal of single Sequence-Quartet depending Output for Restart (-slim Option)

For each sequence-quartet analyses, PENGUIN stores both, the corresponding P4 output file of ML expected site-pattern frequencies for each polarized sequence-quartet tree (subfolder P4_Results/) and the related support calculation file (subfolder clanX_clanY_clanW_clanZ/TXT/single_quartet_calculations/). Thus, the overall data volume of both folders can get very large. Nevertheless, the information content of both subfolders is mandatory for using the -restart option (section 3.1.15). The -slim option induces the disposal of both data files after their use in the PENGUIN process.

- **Default: start new analysis**
- **Command: -slim**
- **Example : -slim** (removal of sequence-quartet individual sub-results)

Note: The -slim option does not enable a subsequent analysis based on the -restart option. Be aware that -restart is specified as first or second parameter in the command line.

4 Output/Result Files

4.1 Results of a 4-Clan Combination (PhyQuart-Mapping)

After analysing all sequence-quartets, PENGUIN prints for each 4-clan combination following split result information as graphic (.svg) and/or text (.txt) formatted output files:

- **Single support** of the three output relevant 4-clan relationships **for each analysed sequence-quartet** ↪ .svg & .txt
- **Mean and median support** of the three output relevant 4-clan relationships **overall analysed sequence-quartets** ↪ .svg & .txt
- **Mean and median support** of the three output relevant 4-clan relationships **for each analysed sequence** ↪ .svg & .txt
- **Distribution of observed support order** for each of the three output relevant 4-clan relationships **overall single sequence-quartet analyses** ↪ .svg & .txt
- **General information** on analysed sequence-quartet combinations, like number of rejected site positions or subsequently excluded sequence-quartets ↪ .txt

Which three 4-clan trees are relevant for the support output depends on the specified parameter options. PENGUIN prints support info of the three outgroup-clan polarized 4-clan trees if an outgroup-clan (-o option) has been specified. Instead, if a 4-clan tree evaluation has been performed without the specification of an outgroup-clan or is based on an outgroup-uninformative split analysis (-r option), support info is given for the three-best supported polarized 4-clan trees or the three outgroup-uninformative supported 4-clan trees, respectively. According to the respective type format, output files (Table 7 and 15) are printed to **SVG** and **TXT** subfolders of individually 4-clan corresponding main folders (entitled by the respective names of the four clans). For example, given a 4-clan combination of the four clans c1, c2, c3, and c4, the path of both 4-clan result folders would be /c1_c2_c3_c4/TXT and /c1_c2_c3_c4/SVG.

4.1.1 Text Formatted Result Output Files

Table 7 lists text (.txt) formatted output file names of the result folder **TXT** .

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

Output File (.txt)	**Information Content**
result_info	General analysis info & main support results
qsingle_spl_sup	List of single quartet support for each of the three output relevant 4-clan relationships
tsingle_mean_sup	List of mean taxon support for each of the three output relevant 4-clan relationships
tsingle_median_sup	List of median taxon support for each of the three output relevant 4-clan relationships
qsingle_rej_pos	List of rejected site positions in single sequence-quartet analyses
best_tree_order_distribution	List of identified signal strength order for each of the three output relevant 4-clan relationships

General Analysis Info & Main Split Results (*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 support of either the three outgroup-clan polarized 4-clan trees or the three outgroup-uninformative best supported 4-clan topologies, 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 *result_info.txt* parameter setup section (part I). A maximum number of 5,000 single sequence-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 invariable site proportion (pINV) start values. The minimum number of allowed sequence-quartet lengths after individual exclusion of unallowed charater site positions is set to 4,500 sites positons.

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 *result_info.txt* sequence info section (part II). Example input alignment consists of nucleotide data with sequence length of 11,687 base pairs. Identified clan names of the clan definition input file (*clan_1* to *clan_4*) are listed according to their internally assigned clan number ("Clan 1" to "Clan 4"). The total number of single performed quartet analyses is 5,000. Mean number of remaining site positions after site exclusion of unallowed character states is 4,599 base pairs. The mean length of rejected sequence-quartets due to reduced sequence lengths below the minimum number of allowed sequence states is 4,426.

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 *result_info.txt* overall support result section. Overall mean and median support of each o the three output relevant 4-clan relationships ($Q1$, $Q2$, $Q3$) based on single performed sequence-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

List of Single Quartet Support (*qsingle_spl_sup.txt)* This result file lists all single quartet identified support for each of the three output relevant 4-clan relationships ($Q1$, $Q2$, $Q3$) in relation to each other. Each line presents 4-clan tree support given a single analysed sequence-quartet. An example is given in Table 11.

Table 11: Example output of the *qsingle_spl_sup.txt* result file, listing support of each sequence-quartet analysis (q_n) of each of the three output relevant 4-clan trees ($Q1$, $Q2$, $Q3$) in relation to each other.

Quartet Number	Seq. Combination	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 $Q2\ldots$
Split Analysis q_2:	T1A:T2A:T3A:T4B	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.4411	Split Support $Q2\ldots$
Split Analysis q_3:	T1A:T2A:T3A:T4C	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.2490	Split Support $Q2\ldots$
Split Analysis q_4:	T1A:T2A:T3A:T4D	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.7543	Split Support $Q2\ldots$
$\vdots$				
Split Analysis q_n:	T1n:T2n:T3n:T4n	Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.9107	Split Support $Q2\ldots$

List of Mean Sequence Support (*tsingle_mean_sup.txt)* Each line of this result file lists the mean identified support of each of the three output relevant 4-clan relationships ($Q1$, $Q2$, $Q3$) in respect of a single sequence (s_n). An example is given in Table 12.

Table 12: Example output of the *tsingle_spl_sup.txt* result file. For each analysed sequence (s_n), this file lists the mean support of each of the three output relevant 4-clan trees ($Q1$, $Q2$, $Q3$) in relation to each other. For example, split analysis s_1 lists the mean support of $Q1$, $Q2$, $Q3$ given all sequence-quartet analyses with sequence T1A (assigned to *clan_1*).

Quartet Number	Seq.	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	$Q2 \ldots$
Split Analysis s_2:	T1B	Clan: *clan_1* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.4154	$Q2 \ldots$
Split Analysis s_3:	T1C	Clan: *clan_1* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.5609	$Q2 \ldots$
Split Analysis s_4:	T2A	Clan: *clan_2* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.1245	$Q2 \ldots$
⋮				
Split Analysis s_n:	T4n	Clan: *clan_4* Mean Split Support $Q1$ (*clan_1*,*clan_2*),(*clan_3*,*clan_4*):	0.4154	…

List of Median Sequence Support (*tsingle_median_sup.txt)* As described in section 4.1.3, just for median support.

List of Rejected Site Positions & Excluded Quartets (*qsingle_rej_pos.txt)* For each sequence-quartet, this file lists the number of rejected site positions of unallowed character states, the number of remaining sites, and sequence-quartets which have been rejected from the analysis in respect of too short sequence lengths. An example is given in Table 13.

Table 13: Example output of the *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. A summarised report of rejected sequence-quartets is also given at the end of the file.

Quartet Number	Seq. Combination	N Rejected Quartet Positions	N Remaining Quartet Positions	Rejected
Split Analysis q_1:	T1A:T2A:T3A:T4A	Rejected Quartet Positions: 7185	Remaining Quartet Positions 4502	
Split Analysis q_2:	T1A:T2A:T3A:T4B	Rejected Quartet Positions: 7281	Remaining Quartet Positions 4406	rejected
Split Analysis q_3:	T1A:T2A:T3A:T4C	Rejected Quartet Positions: 7143	Remaining Quartet Positions 4544	
Split Analysis q_4:	T1A:T2A:T3A:T4D	Rejected Quartet Positions: 7241	Remaining Quartet Positions 4446	rejected
⋮				
Split Analysis q_z:	T1n:T2n:T3n:T4n	Rejected Quartet Positions: 7099	Remaining Quartet Positions 4588	
Rejected Quartets:				
q_2:	T1A:T2A:T3A:T4B			
⋮				
q_z:	T1x:T2y:T3w:T4z			

Distribution of Sequence-Quartet Support (*best_tree_order_distribution.txt)* PENGUIN lists single sequence-quartet support for each of the three output relevant 4-clan trees in relation to sequence-quartet corresponding support ranks of each 4-clan tree. An example is given in Table 14.

4.1.2 SVG Formatted Graphic Result Output Files

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

Table 14: Example output of the *best_tree_order_distribution.txt* result file, listing for each of the three output relevant 4-clan trees the corresponding sequence-quartet support in relation to sequence-quartet respective supported tree rankings (best, 2nd best, 3rd best)).

```
Quartet topology (clan_1,clan_2),(clan_3,clan_4)

Best split score: 2 sequence-quartets.
T1A:T2A:T3A:T4A    0.911043
T1A:T2A:T3A:T4B    0.783917

2nd Best split score: 1 sequence-quartet.
T1A:T2A:T3A:T4C    0.497854

3rd Best split score: 1 sequence-quartet.
T1A:T2A:T3A:T4D    0.205877

Quartet topology (clan_1,clan_3),(clan_2,clan_4)

Best split score: 1 sequence-quartet.
T1A:T2A:T3A:T4D    0.655877
:
:
Quartet topology (clan_1,clan_4),(clan_2,clan_3)

Best split score: 0 sequence-quartets.
:
:
3rd Best split score: 2 sequence-quartets.
T1A:T2A:T3A:T4A    0.211443
T1A:T2A:T3A:T4B    0.284917
```

Table 15: Name and information content of SVG formatted output files.

Output File (.svg)	**Information Content**
split_qmean_sup	Mean support graphic for each of the three output relevant 4-clan trees
split_qmedian_sup	Median support graphic for each of the three output relevant 4-clan trees
triangle_qsingle_split_sup_mean	Triangle graphic, presenting support of mean analysed quartet-sequences
triangle_qsingle_split_sup_median	Triangle graphic, presenting support of median analysed quartet-sequences
triangle_tsingle_mean_sup	Triangle graphic, presenting mean support of each sequence
triangle_tsingle_median_sup	Triangle graphic, presenting median support of each sequence
best_tree_order_distribution*	Number of quartets ordered by signal strength for the three outgroup relevant 4-clan trees
spin_tree_order_distribution**	Number of quartets ordered by spin signal strength the three outgroup relevant 4-clan trees
* with -r option; ** without -r option	

Note: Most 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.

Mean Support SVG Output (*split_qmean_sup.svg*) Split-Network graph of mean support for each of the three output relevant 4-clan relationships ($Q1$, $Q2$, $Q3$), based on all single processed sequence-quartet analyses. The best supported split is presented proportionally to its scoring by horizontal lines, the second best support by vertical lines above the most inner branch, and the lowest support by vertical lines below the internal branch (Figure 2).

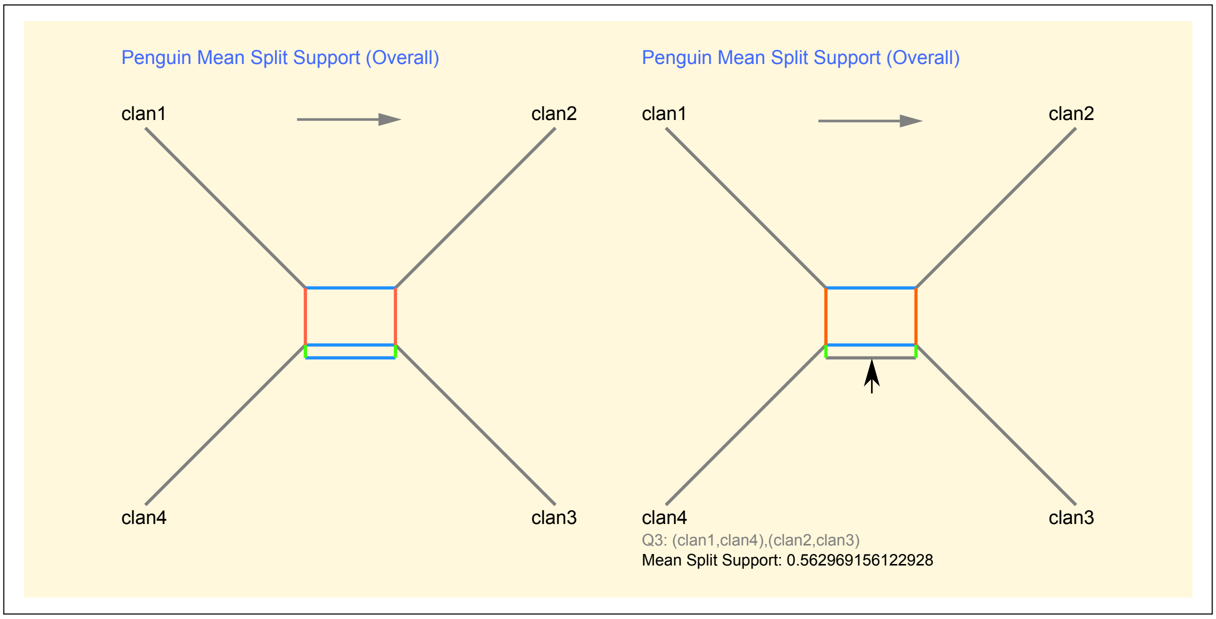
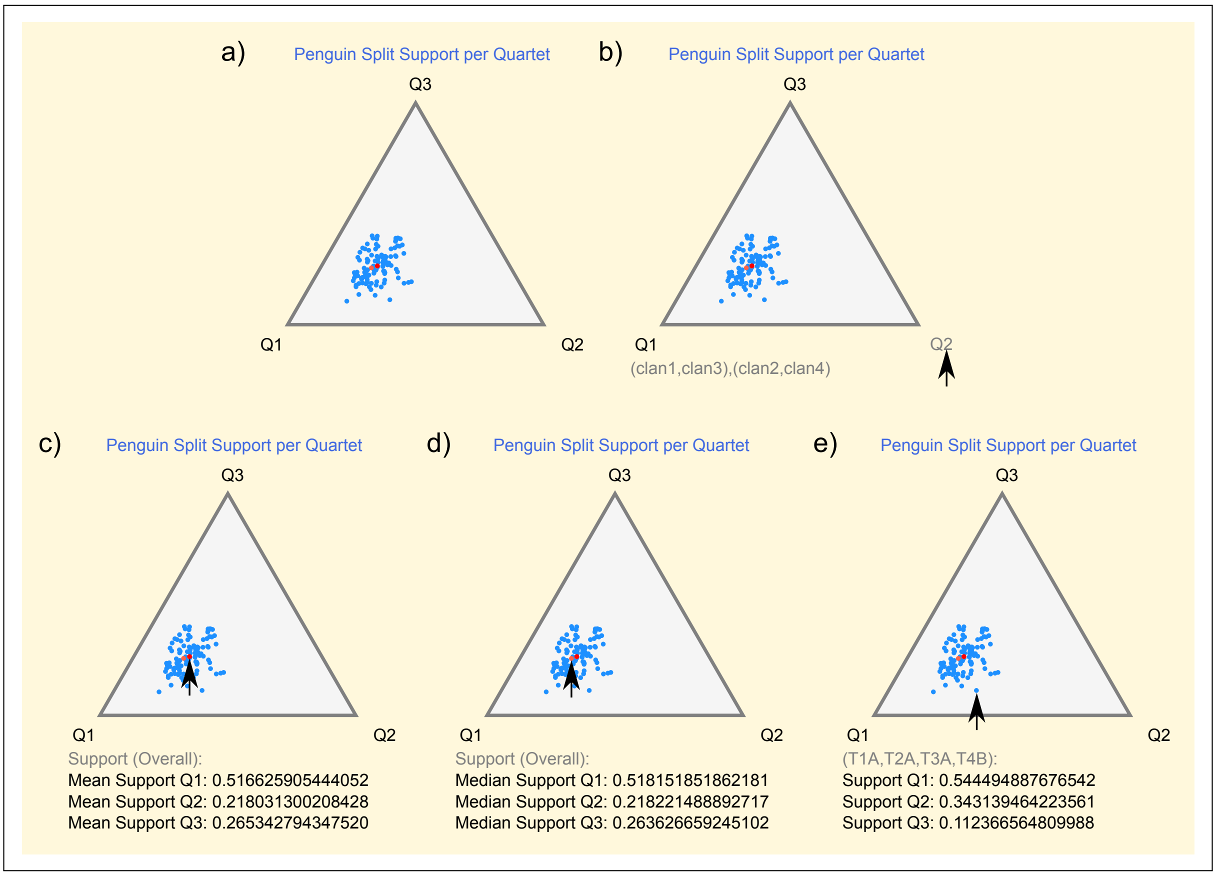

Penguin Mean Split Support (Overall)
clan1
clan2
clan4
clan3
Penguin Mean Split Support (Overall)
clan1
clan2
clan4
clan3
Q3: (clan1,clan4),(clan2,clan3)
Mean Split Support: 0.562969156122928

Figure 2: Example Split-Network output presenting normalised mean support for each of the three output relevant 4-clan relationships, obtained from all single single sequence-quartet analyses. The best mean support is shown by horizontal lines (blue), the second best mean split by vertical lines above the inner branch (red), and the lowest support below the internal branch (green) (figure left). Single 4-clan relationships and assigned mean support values are highlighted if a mouse courser is pointed on a support corresponding split line (figure right). Without the -r option, root directive assumption of character alteration along the innermost branch are shown by an arrow above each network graph.

Median Split Support (*split_qmedian_sup.svg*) Split-Network graph of median support for each of the three output relevant 4-clan relationships ($Q1$, $Q2$, $Q3$), based on all single processed sequence-quartet analyses. Split vector graph with the same presentation as shown in Figure 2 for mean support.

Mean Single Sequence-Quartet Support (*triangle_qsingle_spl_sup_mean.svg*) Triangle plot of single support of each analysed quartet combination for each of the three output relevant mean supported 4-clan relationships ($Q1$, $Q2$, $Q3$). Identified support ratio for each of the three trees of each single sequence-quartet is coded by a blue dot. The overall mean support is marked by a red dot (Figure 3). Additional data information is highlighted if a mouse courser points on an assigned data point (Figure 3b, c, d, e). The higher the support for a given 4-clan relationship in relation to the other two 4-clan trees, the closer is the sequence-quartet support to the best supported 4-clan tree assigned corner ($Q1$, $Q2$, or $Q3$) of the triangle.

Median Single Sequence-Quartet Support (*triangle_qsingle_spl_sup_mean.svg*) Triangle plot of single support of each analysed quartet combination for each of the three output relevant median supported 4-clan relationships ($Q1$, $Q2$, $Q3$). Identified support ratio for each of the three trees of each single sequence-quartet is coded by a blue dot. The overall median support is marked by an orange dot (Figure 3). Additional data information is highlighted if a mouse courser points on an assigned data point (Figure 3b, c, d, e). The higher the support for a given 4-clan relationship in relation to the other two 4-clan trees, the closer is the sequence-quartet support to the best supported 4-clan tree assigned corner ($Q1$, $Q2$, or $Q3$) of the triangle.


a)
Penguin Split Support per Quartet
Q3
Q1
Q2
b)
Penguin Split Support per Quartet
Q3
Q1
Q2
(clan1,clan3),(clan2,clan4)
c)
Penguin Split Support per Quartet
Q3
Q1
Q2
d)
Penguin Split Support per Quartet
Q3
Q1
Q2
e)
Penguin Split Support per Quartet
Q3
Q1
Q2

```
Support (Overall):
Mean Support Q1: 0.516625905444052
Mean Support Q2: 0.218031300208428
Mean Support Q3: 0.265342794347520
```

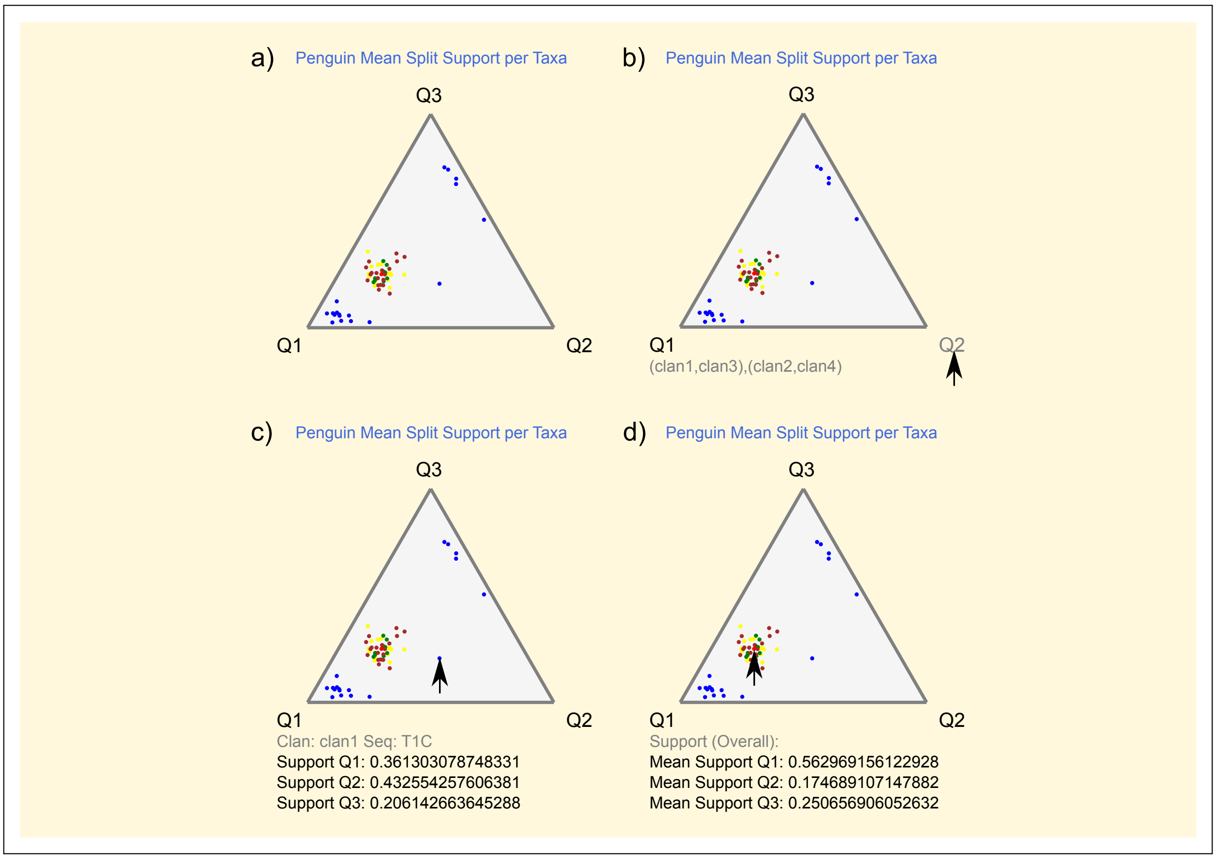
```
Support (Overall):
Median Support Q1: 0.518151851862181
Median Support Q2: 0.218221488892717
Median Support Q3: 0.263626659245102
```

```
(T1A,T2A,T3A,T4B):
Support Q1: 0.544494887676542
Support Q2: 0.343139464223561
Support Q3: 0.112366564809988
```

Figure 3: Example Triangle-Graph output presenting normalised mean or median, and single sequence-quartet support for each of the three output relevant mean supported 4-clan trees ($Q1$, $Q2$, $Q3$). Single quartet split scores are presented by blue dots, mean 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 support, d) highlighted median support, e) highlighted single sequence-quartet support.

Note: Single support for each of the three possible clan relationships ($Q1$, $Q2$, $Q3$) are given in relation to each other. The higher the difference between both values, as closer is the support plotted next to the highest supported clan relationship.

Mean Sequence Support (*triangle_tsingle_mea_sup.svg*) Triangle plot of mean supported 4-clan trees scores of each analysed sequence due to the sequence corresponding single support, observed from sequence corresponding sequence-quartet analyses. Mean sequence support are colored 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 a mouse courser points on an assigned data point (Figure 4b, c, d). The higher the support for a given 4-clan relationship, the closer is the support value to one of the three quartet relationship assigned corners ($Q1$, $Q2$, $Q3$) of the triangle.



Clan: clan1 Seq: T1C
Support Q1: 0.361303078748331
Support Q2: 0.432554257606381
Support Q3: 0.206142663645288

Support (Overall):
Mean Support Q1: 0.562969156122928
Mean Support Q2: 0.174689107147882
Mean Support Q3: 0.250656906052632

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 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 overall mean support.

Median Sequence Support (*triangle_tsingle_med_sup.svg)* Triangle plot of median support of each analysed sequence due to the sequence corresponding single support, observed from sequence corresponding sequence-quartet analyses. Like described for mean triangle graph, median sequence support values are colored based on given clan membership. First defined clan: blue dots, second: yellow, third: green, fourth: purple (Figure 4). Overall median support is marked as orange 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.

N Best Quartet Support (*Penguin_best_tree_order_distribution.svg)* Barchart graph on the total number and distribution of signal strength of single quartets whose split value support one of the three output relevant 4-clan trees ($Q1$, $Q2$, $Q3$), see Figure 5, left.

N Best Quartet Support (*Penguin_spin_tree_order_distribution.svg)* Barchart graph on the total number and distribution of signal strength of single quartets whose split value support one of the three output relevant 4-clan trees ($Q1$, $Q2$, $Q3$), see Figure 5, right.

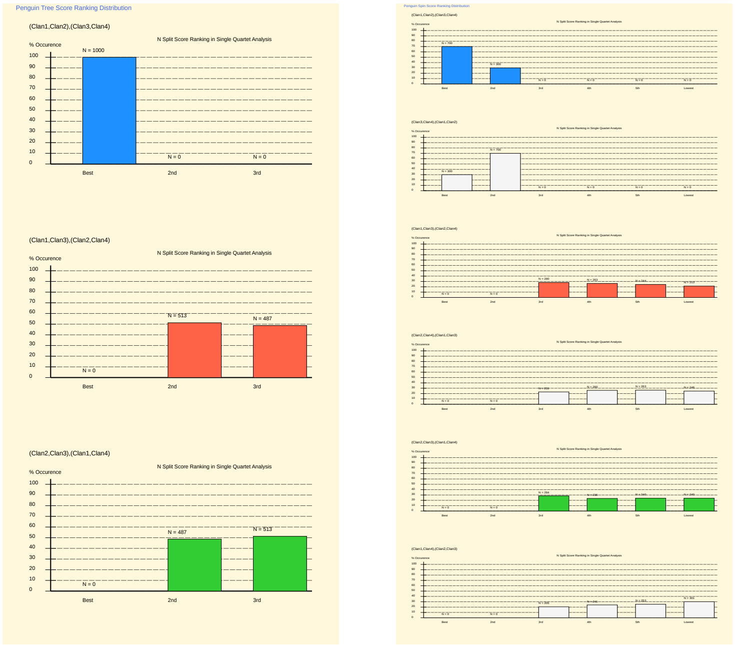


Figure 5: Example SVG-Barchart output presenting the total number and distribution of best split score values for each of the three output relevant 4-clan relationships (left) and (except for the -r option) additionally for each of the two different assumed polarisations of the three possible 4-clan trees, obtained from all single single quartet analyses. Each barchart shows the distribution of signal strength for a specific topological assumption. The number and distribution of support strength observed for each topological assumption compared to other split scores is shown along the x-axis from left to right (best support $\hookrightarrow$ lowest support) as well as top down, presenting the topological assumption with the highest overall observed split signal strength as uppermost barchart. Without the -r option, scoring order of the two root directive split assumptions of character alteration (polarisation) along the innermost branch of one of the three possible 4-clan relationships is shown as additional graph output (right). The single quartet inferred distribution of the in total lower supported polarisation of each clan relationship is shown as white barcharts. Thereby, tree polarisation is always coded by a barchart assigned newick string, showing the the direction of character alteration always from left to right, e.g. (Clan1,Clan2),(Clan3, Clan4) means that character alteration along the innermost branch points from (Clan1,Clan2) to (Clan3,Clan4). In the opposite case, the newick string would be ordered (Clan3,Clan4),(Clan1, Clan2). The order of clan relationships left and right to the innermost comma of a newick string output has no additional root information if generated with the -r option.

4.2 Results of a Multiple-Clan Analysis (Icebreaker)

After analysing the support of all outgroup-informative multiple-clan trees (with number of defined clans greater than four) based on single inferred (mean or median) polarized 4-clan tree support of each possible 4-clan combination, PENGUIN prints following multiple-clan split result information as graphic (.svg) and/or text (.txt and .tre) formatted output files:

- **Newick string** of the (mean and median) **best** supported multiple-clan tree $\hookrightarrow$ .tre (main folder)
- **Main Icebreaker result file**, listing the **final** support of the best multiple-clan tree based on the support ratio (c) of the best-three topologies (main folder) $\hookrightarrow$ .txt
- **Multiple-clan tree support** of the **best** (mean and median) supported tree based on the support ratio of the best-three topologies (triangle graph) $\hookrightarrow$ .svg
- **Multiple-clan tree support** of the **best-three** (mean and median) topologies normalised to each other $\hookrightarrow$ .txt

- **Multiple-clan tree support** of a **n specified number** of best (mean and median) supported trees ↪ .txt
- **Pairwise compatibility matrix** of (mean and median) polarized 4-clan tree support ↪ .txt

Except for the newick string of the best multiple-clan tree and the main Icebreaker result file, which are directly printed to the main result folder, output files (Table 16) are printed to the result subfolder ICEBREAKER.

Table 16: Name and information content of text and graphical formatted output files (Icebreaker).

Output File (.txt, .tre, and .svg)	**Information Content**
Penguin_final_results_mean_*outgroup*	Final support (c) of the (mean) best tree due to conflict calculation of the best-three trees
Penguin_final_results_median_*outgroup*	Final support (c) of the (median) best tree due to conflict calculation of the best-three trees
mclan_support_matrix_mean_∗_best_tree	Newick string of the mean best supported multiple-clan tree
mclan_support_matrix_median_∗_best_tree	Newick string of the median best supported multiple-clan tree
mclan_support_matrix_mean_∗_top_topology_scores	Normalised mean support of the best-three multiple-clan trees
mclan_support_matrix_median_∗_top_topology_scores	Normalised median support of the best-three multiple-clan trees
mclan_support_matrix_mean_∗_all_topology_scores	List of n defined best mean supported trees given unnormalised matrix support
mclan_support_matrix_median_∗_all_topology_scores	List of n defined best median supported trees given unnormalised matrix support
mclan_support_matrix_mean	Pairwise compatibility matrix of single mean inferred polarized 4-clan tree support
mclan_support_matrix_median	Pairwise compatibility matrix of single median inferred polarized 4-clan tree support
mclan_support_matrix_mean_∗_best_sup_ratio	Triangle graphic, presenting mean support ratio of best-three multiple-clan trees

4.2.1 Text and SVG Formatted Result Output Files

Final support of the best multiple-clan tree (*Penguin_final_results_∗.txt*) Implies the final score (c) of the best mean or median supported multiple-clan tree in relation to the two next best supported topologies (Table 19). This value is based on the support ratio of the normalized score difference between the best (S_1') and 2^{nd} best tree (S_2') and the best (S_1') and the 3^{rd} best tree (S_3') as final measure (Eq. 1).

$$c := \frac{S_1' - S_2'}{S_1' - S_3'} \tag{1}$$

Normalized support values (S_1', S_2', S_3') are thereby listed for each of the best-three topologies at the end of the file. A (currently) fixed threshold value of 0.6 is further used as a simple guide value for the evaluation of best tree signal strength. Depending on c, *Penguin* scores provide either very strong (score $\geq$ 0.8) support, strong (0.6 $\leq$ score $<$ 0.8) support, moderate (0.4 $\leq$ score $<$ 0.6) conflict, or strong conflict ($<$ 0.4) (Tab. 17), whereby best trees with support within the moderate conflict range should be considered to be rather conflicted and thus not conclusively supported.

Table 17: Summary of PENGUIN aggregate support ranges as approximate guide for result interpretation.

0.8 ≥	score	≤ 1.0	→	very strong support
0.6 ≥	score	≤ 0.8	→	strong support
0.4 ≥	score	≤ 0.6	→	moderate conflict
0.0 ≥	score	≤ 0.4	→	strong conflict

Newick string of the best supported multiple-clan tree (*∗_best_tree.tre*) Implies the best mean or median supported multiple-clan tree as newick string (Table 19).

Table 18: Example output of the best (mean or median) supported multiple-clan tree (in this example based on six clans), rooted either by the specified outgroup-clan (-o option) or by PENGUIN itself based on support evaluation of the pairwise compatibility matrix if the outgroup-clan is left unspecified.

(Outgroup_Clan, (Clan_1, ((Clan_2, Clan_3), (Clan_4, Clan_5))));

Table 19: Example output of the best (mean or median) supported multiple-clan tree (in this example based on six clans), rooted either by the specified outgroup-clan (-o option) or by PENGUIN itself based on support evaluation of the pairwise compatibility matrix if the outgroup-clan is left unspecified.

(Outgroup_Clan, (Clan_1, ((Clan_2, Clan_3), (Clan_4, Clan_5))));

Best-three supported multiple-clan trees (∗_*top_topology_scores.txt)* Includes normalised (mean or median) support of the best-three multiple-clan trees, rooted either with the -o option defined outgroup-clan (-o option) or by PENGUIN itself based on support evaluation of the pairwise compatibility matrix if the outgroup-clan is left unspecified. (Table 20).

Table 20: Example output of the (rooted) best-three (mean or median) supported multiple-clan trees in descending order, with single support normalised to each other.

(Outgroup_Clan, (Clan_1, ((Clan_2, Clan_3), (Clan_4, Clan_5))))	0.58077
(Outgroup_Clan, (Clan_1, ((Clan_2, Clan_4), (Clan_3, Clan_5))))	0.29184
(Outgroup_Clan, (Clan_1, (Clan_2, (Clan_3, (Clan_4, Clan_5)))))	0.127389

List of n defined best supported trees (∗_*all_topology_scores.txt)* Includes unnormalised (mean or median) support of the n specified (default: $n = 100$ ↪ -b option) best supported multiple-clan trees (in descrnding order), rooted either with the -o option defined outgroup-clan (-o option) or by PENGUIN itself based on support evaluation of the pairwise compatibility matrix if the outgroup-clan is left unspecified. (Table 21).

Table 21: Example output of the n specified best (mean or median) supported multiple-clan trees in descending order, with single support unnormalised to each other.

(Outgroup_Clan, (Clan_1, ((Clan_2, Clan_3), (Clan_4, Clan_5))))	0.268235
(Outgroup_Clan, (Clan_1, ((Clan_2, Clan_4), (Clan_3, Clan_5))))	0.13479
(Outgroup_Clan, (Clan_1, (Clan_2, (Clan_3, (Clan_4, Clan_5)))))	0.0588362
⋮	
(Outgroup_Clan, (Clan_2, (Clan_1, (Clan_3, (Clan_4, Clan_5)))))	0.0548362

Compatibility matrix of supported 4-clan trees (*mclan_support_matrix.txt*) For each analysed 4-clan combination of a set of multiple clans greater than four, single identified (mean or median) support values of the six polarized 4-clan relationships are entered into a polarized pairwise compatibility matrix. The matrix assignment of single 4-clan support values of each 4-clan combination is thereby based on a further subdivision into rooted-triplets (R). Based on rooted-triplets in conjunction with polarized 4-clan tree support, a polarized support matrix is established across each rooted triplet-clan (y-axis) and its triplet corresponding pair of stronger derived clans (x-axis) (Table 22).

Table 22: Example output of a (mean or median) inferred pairwise compatibility matrix with polarized 4-clan tree support established across each rooted triplet-clan (y-axis) and its triplet corresponding pair of stronger derived clans (x-axis).

	Clan_1|Clan_2	Clan_1|Clan_3	...	Clan_x|Clan_z
R|Clan_1	0	0	...	0.0294181
R|Clan_2	0	0.144145	...	0.0673949
R|Clan_3	0.0643949	0	...	0.0373949
⋮	⋮	⋮		⋮
R|Clan_x	0.231053	0.144145	...	0
R|Clan_z	0.0294181	0.144145	...	0

Support ratio of the best-three multiple-clan trees (∗_*best_sup_ratio.svg)* Visualizes the support ratio of normalised mean or median support of the best-three multiple-clan trees by a dot within a triangle graph, whereas each corner of the triangle represents one of the three best supported multiple-clan trees. The final support ration between these three trees is taken as the final measure of the best supported tree with its score highlighted by the corresponding dot color (Figure 6).

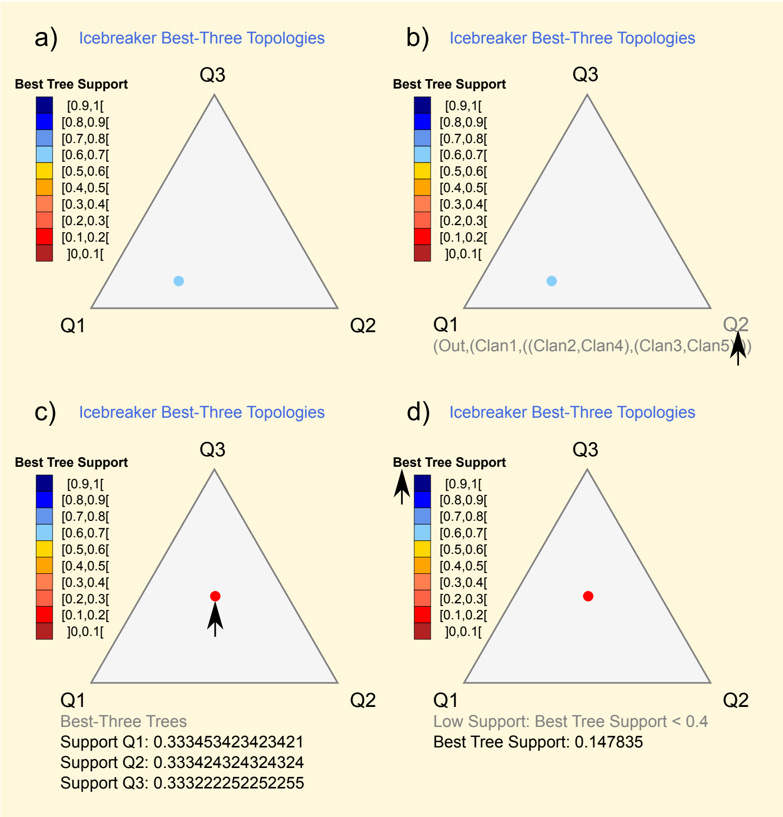


Figure 6: Example Triangle-Graph output presenting normalised mean support for each of the best-three multiple-clan relationships ($Q1$, $Q2$, $Q3$). The support ratio of these three trees is presented by a dot, colored by corresponding score value. a) without additional information, b) phylogenetic clan relationship of corner defined quartet trees ($Q1$, $Q2$, $Q3$), c) highlighted best-three multiple-clan individual support, d) highlighted support ratio of the best-three trees.

4.3 Single Sequence-Quartet Output (PhyQuart)

Under default, PENGUIN stores both, the corresponding P4 output file of ML expected site-pattern frequencies for each polarized sequence-quartet tree (subfolder P4_Results/) and the related support calculation file (subfolder clanX_clanY_clanW_clanZ/TXT/single_quartet_calculations/). The files in both subfolders enable an extended analysis based on already analysed sequence-quartet combinations without loosing much computation time (see -restart option at section 3.1.15). Nevertheless, since both files are generated for each sequence-quartet, the data amount of both subfolders can get very large. Thus, we implemented the -slim option (section 3.1.16) with which both file types are continuously removed from the analysis.

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 (**?**) 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 7.

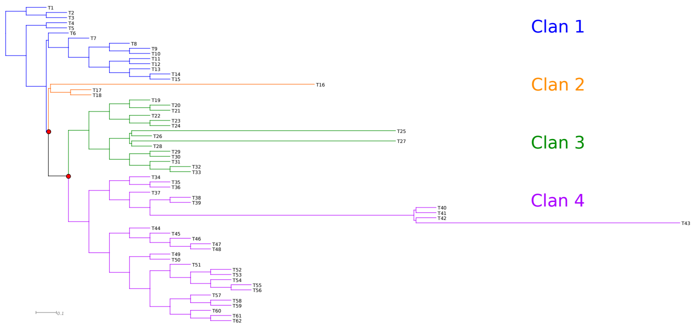


Figure 7: Example MSA input data underlying topological relationship between 61 sequences, divided into four different clans. Internal branch assigned support analysis based on defined sequence clans is flanked by red dots.

Note: 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.

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

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.

7 Copyright