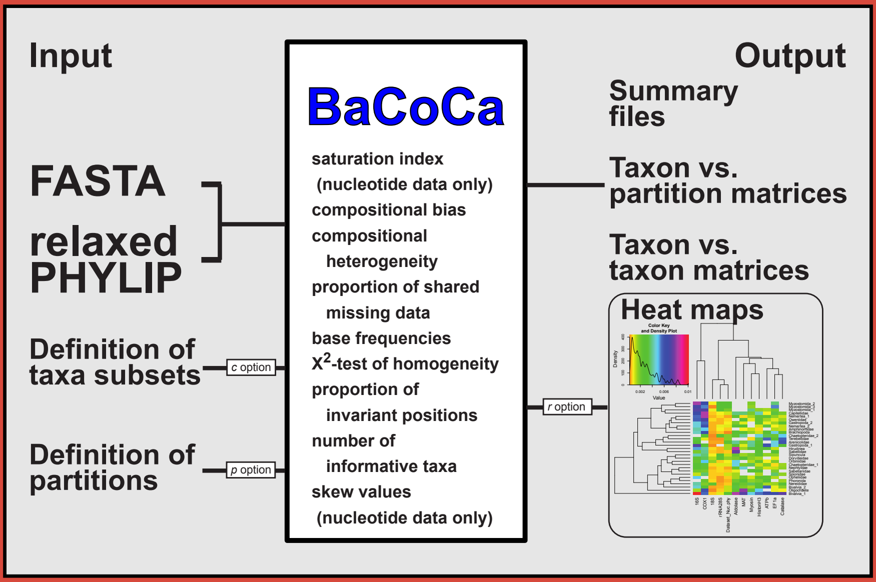
FORSCHUNGS museum KOENIG

Patrick Kück & Torsten H. Struck


Input
FASTA
relaxed PHYLIP
Definition of taxa subsets
c option
Definition of partitions
p option
BaCoCa
saturation index (nucleotide data only)
compositional bias
compositional heterogeneity
proportion of shared missing data
base frequencies
X^2-test of homogeneity
proportion of invariant positions
number of informative taxa
skew values (nucleotide data only)
Output
Summary files
Taxon vs. partition matrices
Taxon vs. taxon matrices
r option
Heat maps

BaCoCa

November 2016

Contents

List of Figures

List of Tables

1 Features

BaCoCa (**Ba**se **Co**mposition **Ca**lculator) is designed to perform multiple statistical analyses on multiple nucleotide and amino-acid sequence alignments. The results of the BaCoCa analyses can be used for a detailed and statistical comprehensive data evaluation. Furthermore, the results can help to identify phylogenetic sequence biases which can lead to incorrect tree reconstructions. The program can handle hundreds of user specified gene and taxon partitions of a single sequence input file in one process run. BaCoCa is a command-line driven program written in PERL and works on WindowsPCs, Macs and Linux operating systems. Therefore, it can be easily integrated into automatic pipeline processes of phylogenomic studies. Results issued by BaCoCa can be optionally extended through further analyses using statistical R packages. For example, heat map analyses of taxon versus gene matrices can be used to find clusters of genes and/or taxa with similar properties (Figure 3). Furthermore, all calculations of the BaCoCa software program are very fast and can be easily executed on a normal desktop computer, even if data sets consist of phylogenomic data. Table 1 gives an overview all BaCoCa implemented calculations. A more detailed discription of all BaCoCa analyses is given in section 3.

Table 1: Overview of BaCoCa implemented calculations.

Calculations	**Amino-Acid Data**	**Nucleotide Data**
Base Frequencies	yes	yes
Chi-square Test of Homogeneity	yes	yes
Saturation (C-value)	no	yes
Invariant Positions	yes	yes
Missing Data Overlap	yes	yes
Number Informative Taxa	yes	yes
Compositional Heterogeneity (RCFV-value)	yes	yes
Skew-values	no	yes
Compositional Bias	yes	yes

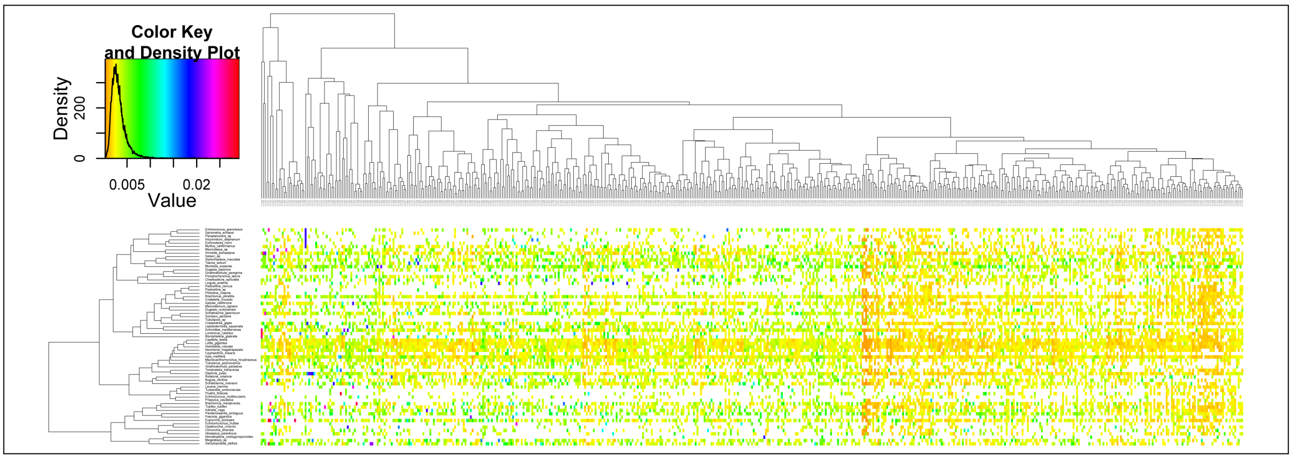


Figure 1: Example heat map of a BaCoCa compositional heterogeneity result file (generated optionally with R) of 559 genes and 65 taxa

2 Usage/Options

For using BaCoCa open the terminal of your operating system. Move through your directory path to the folder where BaCoCa is placed. To execute BaCoCa a PERL interpreter must be installed on the current run system. Linux and Mac systems do normally not need a subsequent installation because the PERL interpreter is a standard tool of these systems in advance, but have to install the additional PERL Statistics::R package to use the "-r" option of BaCoCa.vX.X.r.pl which allows the generation of result heat maps by using R. For the usage of BaCoCa.vX.X.pl which has not included the "-r" option, additional PERL or R packages have not to be installed.

The **Statistics::R package** can be easily installed on Linux systems like Ubuntu v12.04 or higher and Mac systems by typing:

- **user@linux**:˜$ sudo apt-get install libstatistics-R-perl <enter>

To install the **full PERL CPAN packages** type:

- **user@mac**:˜$ sudo cpan App::cpanminus <enter>

To install the **R-basic packages** and the **R-gplots library** on Linux or Mac operating systems type:

- **user@linux**:˜$ sudo apt-get install r-base <enter>
- **user@linux**:˜$ sudo apt-get install r-cran-gplots <enter>

To install the **full R-package list** type:

- **user@mac**:˜$ sudo cpanm Statistics::R <enter>

Unfortunately, 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/

The additional PERL package Statistics::R can be installed via the ActivePerl package manager. R can be installed from:

- http://cran.r-project.org/bin/windows/base/

2.1 Start BaCoCa via single command line

BaCoCa can be directly started by the command line in one row which simplifies the implementation of BaCoCa into complex process pipelines. Move through your directory path to the folder where BaCoCa is located and type the name of the BaCoCa version, followed by a blank and the demand options with a minus (-) sign in front of each. Note that Windows users don't have to put "perl" in front of program name. Then press <enter>. Make sure you write the input options correctly, for example "-i" and not "-i". Otherwise BaCoCa will not start working but instead open the Synopsis menu.

- user@linux:˜$ perl BaCoCa_v1.0.pl -help <enter> ↪ help menu
- user@linux:˜$ perl BaCoCa_v1.0.pl -i *path/infile* <enter> ↪ start BaCoCa under default

I

2.2 Input File Options

BaCoCa knows several input file options. It ignores commands if an unknown option is encountered. BaCoCa 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 format. Notice that wrong file format allocates BaCoCa to stop all running processes and to abort with a particular error prompt.

2.2.1 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). BaCoCa 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, ambiguity characters, ”?”, ”X”, and ”-”.

FASTA (.fas) format BaCoCa 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 formats 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

Extant approaches which can automatedly produce aligned sequences in FASTA format are e.g. `MAFFT` (**????**), `MUSCLE` (**?**), or `T-COFFEE` (**??**). To convert aligned sequence files in FASTA format software tools like `T-COFFEE` (**??**) or `FASconCAT` (**?**) can be used. FASconCAT (**?**) 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 (**??**) or FASconCAT (**?**) can be used. FASconCAT (**?**) can be further used for gene concatenation. Please, read the respective manuals and/or publications for further details.

2.2.2 Subclade Definition File (-c Option)

Besides the execution of BaCoCa over all taxon sequences (default), the script can simultaneously analyse different alignment parameters of one or multiple user pre-defined taxon subclades. Taxon subclades have to be defined via a subclade definition file given by the "-c" option. The subclade definition file has to be in plane .txt format. Taxon subclades are defined in separate lines. Each subclade has to start with a user specified clade name (alphanumeric signs and underscore are allowed!), followed by a comma, followed by comma separated subclade taxon names. Be aware that all taxon names of defined subclades are identical to the corresponding sequence names of the sequence input file (case sensitive!). Blanks between comma seperated taxon names are not allowed! A taxon name can be defined in multiple taxon subclades, but should appear only once per subclade. Table 4 shows the correct format of a typical subclade definition file.

Table 4: Example of a typical subclade definition file.

Subclade_Definition_File.txt
Name_Subclade_1,Taxon_1,Taxon_2,Taxon_3,Taxon_4
Name_Subclade_2,Taxon_1,Taxon_3,Taxon_5,Taxon_6,Taxon_9
Name_Subclade_3,Taxon_1,Taxon_7,Taxon_8,Taxon_9
⋮
Name_Subclade_n,Taxon_X,Taxon_Y,Taxon_Z

2.2.3 Gene Partition File (-p Option)

Under default, BaCoCa analyses complete sequence data. To investigate single genes or specific partition ranges of complete sequence data in addition to complete sequences, single gene ranges have to be defined via the "-p" option followed by the name of the gene partition file in plain .txt format. Multiple gene ranges can be defined in separeted lines. Overlapping ranges between single defined partitions are possible. Each line has to start with a user specified partition name (alphanumeric signs and underscore are allowed!), followed by a equal sign, followed by the startposition of the partition, followed by a minus sign, followed by the endposition of the partition. Blanks are allowed ! Be aware, that the startposition of a gene partition is lower as its endposition. Otherwise BaCoCa will abort with an error prompt. Partitioning based on codon positions is not implemented at the moment. Table 5 and **??** show correct formats of a typical gene partition file.

Table 5: Example formats of possible gene partition files.

Typical_Gene_Partition_File.txt				
Name_Partition_1	=	1	-	100
Name_Partition_2	=	101	-	949
Name_Partition_3	=	930	-	1000
⋮				
Name_Partition_n	=	Startposition_X	-	Endposition_Y
RAxML adapted Gene_Partition_File.txt				
AnyAlphanumericName, Name_Partition_1	=	1	-	100
AnyAlphanumericName, Name_Partition_2	=	101	-	949
AnyAlphanumericName, Name_Partition_3	=	930	-	1000
⋮				
AnyAlphanumericName, Name_Partition_n	=	Startposition_X	-	Endposition_Y

2.2.4 Generating R Heat Maps (-r Option)

Many of the BaCoCa infiles can also optionally printed out as Heat Maps with the ”-r” command by using `R` packages. For this option the `R-basic` package as well as the have additional `R-gplots` library have to be installed. To install the basic `R` packages and the `R-gplots` library on Linux or Mac operating systems from the internet, we recommend to use the following terminal command lines:

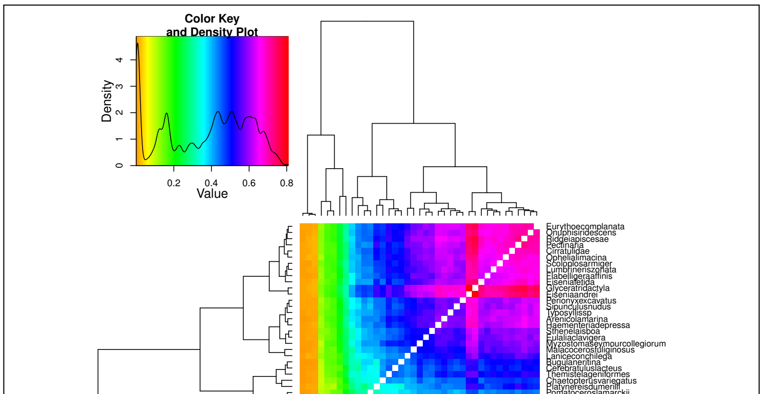
- user@linux:˜$ sudo apt-get install r-base <enter>
- user@linux:˜$ sudo apt-get install r-cran-gplots <enter>

. . . or the installation of the full CPAN packages for Perl by typing. . .

- user@mac:˜$ sudo cpan App::cpanminus <enter>

If the installation of the required packages has been performed successfully, BaCoCa will generate result heat maps of all subresults within the following folders:

- BaCoCa_Results/compositional_bias/
- BaCoCa_Results/skew_values/
- BaCoCa_Results/taxon_basefrequencies_all_partions/
- BaCoCa_Results/missing_data_overlap/



Short Explanation Of Taxon To Taxon Specific Heat Maps The different colors in a heat map reflect a different value. In our example the colors range from orange via yellow, green and blue to red and correspond to values from 0 to

1 (herein reflecting the proportion of shared missing data between two taxa). White indicates that the cell of the heat map is undefined (i.e., no value is present for this pair of taxa). Both the y- and the x-axis show taxa in this example. Due the hierarchical clustering taxa with similar values in a column (or row for that matter) are grouped together more closely. For example, the three taxa, for which complete genomes are available (*Capitella teleta*, *Helobdella robusta* and *Lottia gigantea*), group together as each of them show very low proportions of shared missing data with the other taxa due to the fact that all genes of the analysis are present in these three. The results of the hierarchical clustering is also shown by the trees displayed on the y- and x-axis. Finally, a density plot is provided in the upper left corner showing the distribution of the values from 0 to 1 in the analysed data set.

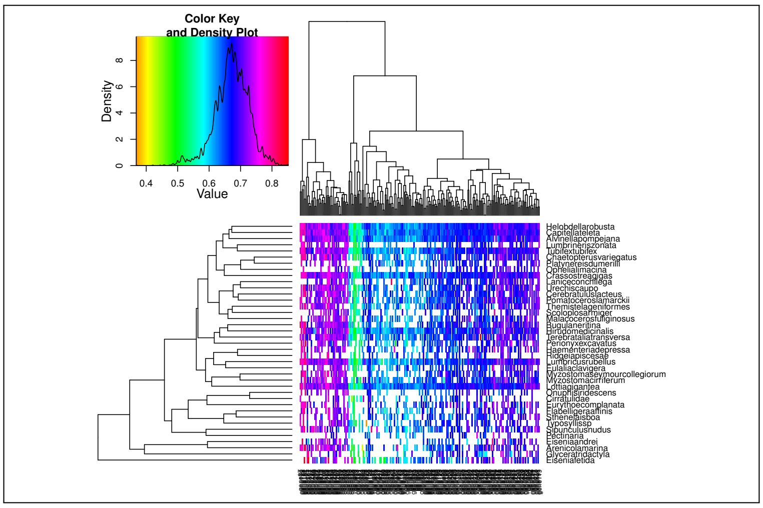


Figure 3: Example taxon to gene specific heat map of a BaCoCa hydrophilic base frequency result file of 231 genes and 39 taxa

Short Explanation Of Taxon To Gene Specific Heat Maps As above the different colors in a heat map reflect a different value. In our example the colors range from orange via yellow, green and blue to red and correspond to values from below 0.4 to above 0.8 (herein reflecting the proportion of hydrophilic amino-acids of a taxon in a partition). White indicates that the cell of the heat map is undefined (i.e., no value is present for this pair of taxon and partition as the corresponding taxon is lacking in the partition). The y-axis shows the taxa and the x-axis the partitions in this example. Due the hierarchical clustering taxa with similar values in their rows are grouped together more closely. For example, the three taxa *Capitella teleta*, *Helobdella robusta* and *Alvinella pompejana*) group together as each of them has similar proportions of hydrophilic amino-acids across the different partitions. The results of the hierarchical clustering is also shown by the tree displayed on left-hand side. Similarly, the tree above the heat map shows the results of the hierarchical clustering of the partitions. In this clustering partition with similar values in their columns (i.e., across different taxa) are grouped together. Finally, a density plot is provided in the upper left corner showing the distribution of the proportion of hydrophilic amino-acids in the analysed data set.

Note: At least two gene partitions have to be defined to use the ”-r” option. A heat map cannot be generated if BaCoCa is used for analysing only the complete dataset. As R does not generate heat maps when all values in the matrix have the same value (e.g., all are 1), BaCoCa does not generate heat maps in these cases.

2.2.5 Testfiles

The datasets are based on empirical data including the test files for partitions and clades. The amino-acid dataset is from **?** and the nucleotide dataset from **?**.

3 BaCoCa Calculations & Output

BaCoCa performs multiple statistical analyses on multiple nucleotide and amino-acid sequence alignments for complete data sets as well as for user specified taxon subclades and gene partitions. A summarized overview of all calculations is given in table 1. All results are summarized within the newly generated folder BaCoCa_Results and, depending on the calculation values, further divided into specific result subfolders:

- BaCoCa_Results/
  - chisquare_test_homogeneity_taxa/
  - compositional_bias/
  - c_value_calculations/
  - invariant_alignment_positions/
    - fasta/
    - svg/
    - txt/
  - missing_data_overlap/
  - skew_values/
  - taxon_basefrequencies_all_partions/
  - taxon_basefrequencies_single_partions/

All result files are printed into specific subfolders, except of the summarized result file *summarized_taxon_base_frequen* which includes summarized results for most of the calculations of complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 4). This file is printed directly into the main result folder BaCoCa_Results/. Moreover, a plain-text documentation (explanation_guide.txt) of the output files and structure are provided along-side the result files. The following subsections give a more detailed overview of single BaCoCa calculations and print outs.

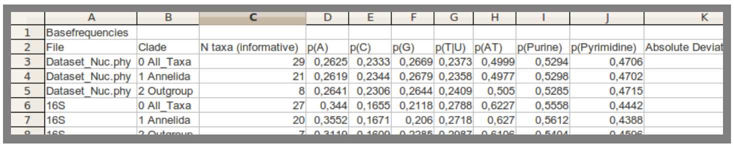
Note: Microsoft Excel versions older than Microsoft Excel 2007 cannot handle output files with more than 256 columns while newer versions have no such problems (Microsoft Excel 2007 and higher can handle 16.384 columns ↪ A to XFD). Open Office does not have this problem.

3.1 Base Frequencies

Calculation of base frequencies for nucleotide and amino-acid character states, ambiguity character states, indel events (-), missing data (?), summarised frequencies of uninformative character states (ambiguities, indel events, and missing data), summarised frequencies of informative character states (including taxa whose sequence consist of at least one informative character state), and invariant sequence positions. Furthermore, summarised frequencies are printed for AT, purine (A/G), and pyrimidine (C/T) content (nucleotide data) as well as for positively (R/K), negatively (D/E) and neutral electrically charged side chains (charge), polar and non-polar (A/G/W/M/V/I/L/F/P) charged side chains, and hydrophobic (A/W/M/I/L/F/P) and hydrophilic charged side chains (amino-acid data). BaCoCa results of base frequencies are printed to:

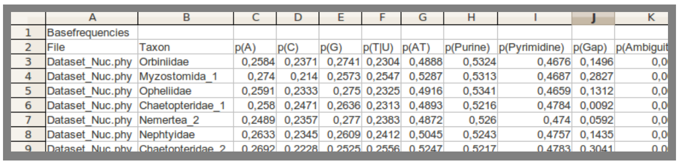
- Summary file (summarized_taxon_base_frequencies.txt) over all taxa of complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 4)
- Summary files (taxon_base_frequencies_single_partitions/∗_summarized_taxon_base_frequencies.txt) for single taxa of complete sequence data (see Figure 5)

- Summary files (taxon_base_frequencies_all_partitions/∗_frequencies_all_partitions.txt) for individual frequencies like purine, pyrimidine, AT-content (nucleotide data) or amino-acid states for each taxon vs. complete sequence data and single gene partitions ↪ useful for taxa-gene heat maps and hierarchical clustering (see Figure 6). If the -r option is used, heat maps will be generated (see 2.2.4 Generation R Heat Maps (-r Option) for examples).



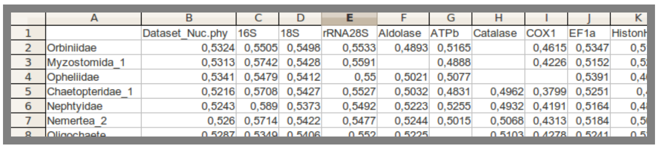
	A	B	C	D	E	F	G	H	I	J	K
1	Basefrequencies										
2	File	Clade	N taxa (informative)	p(A)	p(C)	p(G)	p(T|U)	p(AT)	p(Purine)	p(Pyrimidine)	Absolute Deviat
3	Dataset_Nuc.phy	0 All_Taxa	29	0,2625	0,2333	0,2669	0,2373	0,4999	0,5294	0,4706	
4	Dataset_Nuc.phy	1 Annelida	21	0,2619	0,2344	0,2679	0,2358	0,4977	0,5298	0,4702	
5	Dataset_Nuc.phy	2 Outgroup	8	0,2641	0,2306	0,2644	0,2409	0,505	0,5285	0,4715	
6	16S	0 All_Taxa	27	0,344	0,1655	0,2118	0,2788	0,6227	0,5558	0,4442	
7	16S	1 Annelida	20	0,3552	0,1671	0,206	0,2718	0,627	0,5612	0,4388	
8	16S	2 Outgroup	7	0,3119	0,1609	0,2285	0,2987	0,6106	0,5404	0,4596	

Figure 4: Example cut out of BaCoCa summarized frequency output file for single frequencies, separately summarized over all taxa for complete sequence data, pre-defined taxon subclades, and gene partitions



	A	B	C	D	E	F	G	H	I	J	K
1	Basefrequencies										
2	File	Taxon	p(A)	p(C)	p(G)	p(T|U)	p(AT)	p(Purine)	p(Pyrimidine)	p(Gap)	p(Ambiguit
3	Dataset_Nuc.phy	Orbiniidae	0,2584	0,2371	0,2741	0,2304	0,4888	0,5324	0,4676	0,1496	0,0
4	Dataset_Nuc.phy	Myzostomida_1	0,274	0,214	0,2573	0,2547	0,5287	0,5313	0,4687	0,2827	0,0
5	Dataset_Nuc.phy	Opheliidae	0,2591	0,2333	0,275	0,2325	0,4916	0,5341	0,4659	0,1312	0,0
6	Dataset_Nuc.phy	Chaetopteridae_1	0,258	0,2471	0,2636	0,2313	0,4893	0,5216	0,4784	0,0092	0,0
7	Dataset_Nuc.phy	Nemertea_2	0,2489	0,2357	0,277	0,2383	0,4872	0,526	0,474	0,0592	0,0
8	Dataset_Nuc.phy	Nephtyidae	0,2633	0,2345	0,2609	0,2412	0,5045	0,5243	0,4757	0,1435	0,0
9	Dataset_Nuc.phy	Chaetopteridae_2	0,2692	0,2228	0,2525	0,2556	0,5247	0,5217	0,4783	0,3041	0,0

Figure 5: Example cut out of BaCoCa frequency output file for single taxa of complete sequence data



	A	B	C	D	E	F	G	H	I	J	K
1		Dataset_Nuc.phy	16S	18S	rRNA28S	Aldolase	ATPb	Catalase	COX1	EF1a	HistonH
2	Orbiniidae	0,5324	0,5505	0,5498	0,5533	0,4893	0,5165		0,4615	0,5347	0,5
3	Myzostomida_1	0,5313	0,5742	0,5428	0,5591		0,4888		0,4226	0,5152	0,5
4	Opheliidae	0,5341	0,5479	0,5412	0,55	0,5021	0,5077			0,5391	0,4
5	Chaetopteridae_1	0,5216	0,5708	0,5427	0,5527	0,5032	0,4831	0,4962	0,3799	0,5251	0,
6	Nephtyidae	0,5243	0,589	0,5373	0,5492	0,5223	0,5255	0,4932	0,4191	0,5164	0,4
7	Nemertea_2	0,526	0,5714	0,5422	0,5477	0,5244	0,5015	0,5068	0,4313	0,5184	0,5
8	Oligochaete	0,5287	0,5349	0,5406	0,552	0,5225		0,5103	0,4278	0,5241	0,5

Figure 6: Example cut out of BaCoCa frequency output file for purine frequencies for each taxon vs. complete sequence data and single gene partitions

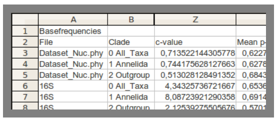
3.2 Saturation (C Value)

C-value calculation of **?** (only for nucleotide data). The saturation value gauges the convergence behavior of the plot of ti/tv ratios against the uncorrected p distance. In case of saturation the curve of ti/tv converges upon on value with increasing p. Hence, the standard deviations of the p distances over all pairwise taxa comparisons will increase, while the standard deviation of ti/tv will decrease. The formula to calculate the C-value is:

$$C = \frac{\sigma(\frac{Ti}{Tv})}{\sigma(p)} \tag{1}$$

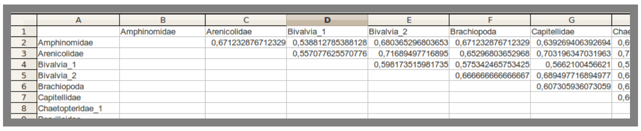
BaCoCa results of saturation calculations are printed to:

- Summary file (summarized_taxon_base_frequencies.txt) ↪ Average and standard deviation of ti/tv ratio and p distance as well as C-value over all taxa of complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 7)
- Summary files (c_value_calculations/results_pairwise_comparisons∗.txt) of pairwise comparisons of ti/tv and p distances between all taxa and taxon subclades vs each defined partition and the concatenated dataset ↪ useful for taxa-gene heat maps and hierarchical clustering (see Figure 8)



	A	B	Z	
1	Basefrequencies			
2	File	Clade	c-value	Mean p-
3	Dataset_Nuc.phy	0 All_Taxa	0,713522144305778	0,6227
4	Dataset_Nuc.phy	1 Annelida	0,744175628127663	0,6278
5	Dataset_Nuc.phy	2 Outgroup	0,513028128491352	0,6843
6	16S	0 All_Taxa	4,34325736721667	0,6536
7	16S	1 Annelida	8,08723921290358	0,6914
8	16S	2 Outgroup	2,12539275505676	0,5701

Figure 7: Example cut out of BaCoCa summarized frequency output file for single C-values, separately summarized over all taxa for complete sequence data, pre-defined taxon subclades, and gene partitions



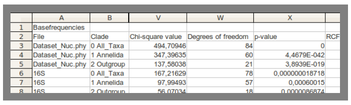
	A	B	C	D	E	F	G	
1		Amphinomidae	Arenicolidae	Bivalvia_1	Bivalvia_2	Brachiopoda	Capitellidae	Chae
2	Amphinomidae		0,671232876712329	0,538812785388128	0,680365296803653	0,671232876712329	0,639269406392694	0,6
3	Arenicolidae			0,557077625570776	0,71689497716895	0,652968036529680	0,703196347031963	0,7
4	Bivalvia_1				0,598173515981735	0,575342465753425	0,5662100456621	0,5
5	Bivalvia_2					0,666666666666667	0,689497716894977	0,6
6	Brachiopoda						0,607305936073059	0,6
7	Capitellidae							0,6
8	Chaetopteridae_1							

Figure 8: Example cut out of BaCoCa taxon specific C-value p distances calculations for each taxon vs. complete sequence data and single gene partitions

3.3 Chi-square Test of Homogeneity

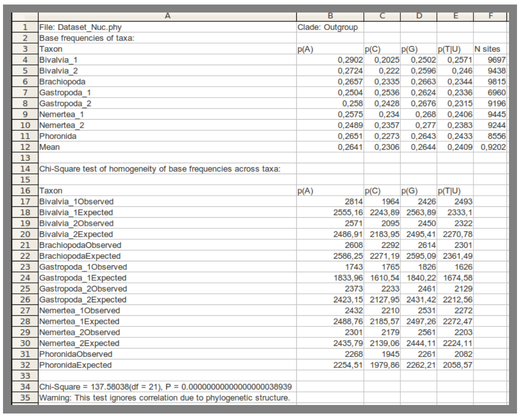
Chi-Square test of homogeneity of base frequencies across taxa is performed. BaCoCa results of base homogeneity tests are printed to:

- Summary file (summarized_taxon_base_frequencies.txt) ↪ Output of chi-square value, degrees of freedom, and p-value over all taxa of complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 9
- Summary files (chisquare_test_homogeneity_taxa/∗.txt) ↪ Detailed overview of chi-square calculations for complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 10)



	A	B	V	W	X	
1	Basefrequencies					
2	File	Clade	Chi-square value	Degrees of freedom	p-value	RCF
3	Dataset_Nuc.phy	0 All_Taxa	494,70946	84	0	
4	Dataset_Nuc.phy	1 Annelida	347,39635	60	4,4679E-042	
5	Dataset_Nuc.phy	2 Outgroup	137,58038	21	3,8939E-019	
6	16S	0 All_Taxa	167,21629	78	0,000000018718	
7	16S	1 Annelida	97,99493	57	0,00060015	
8	16S	2 Outgroup	56,07034	18	0,0000086874	

Figure 9: Example cut out of BaCoCa output summary frequency file for chi-square calculations (Chi-square value, Degrees of freedom, p-value) of complete sequence data, pre-defined taxon subclades, and gene partitions



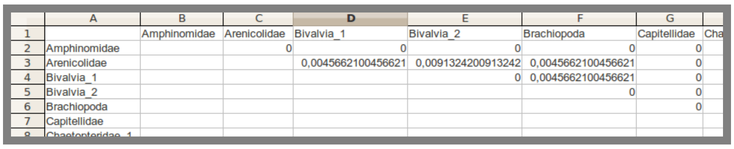
	A	B	C	D	E	F
1	File: Dataset_Nuc.phy	Clade: Outgroup				
2	Base frequencies of taxa:					
3	Taxon	p(A)	p(C)	p(G)	p(T|U)	N sites
4	Bivalvia_1	0,2902	0,2025	0,2502	0,2571	9697
5	Bivalvia_2	0,2724	0,222	0,2596	0,246	9438
6	Brachiopoda	0,2657	0,2335	0,2663	0,2344	9815
7	Gastropoda_1	0,2504	0,2536	0,2624	0,2336	6960
8	Gastropoda_2	0,258	0,2428	0,2676	0,2315	9196
9	Nemertea_1	0,2575	0,234	0,268	0,2406	9445
10	Nemertea_2	0,2489	0,2357	0,277	0,2383	9244
11	Phoronida	0,2651	0,2273	0,2643	0,2433	8556
12	Mean	0,2641	0,2306	0,2644	0,2409	0,9202
13						
14	Chi-Square test of homogeneity of base frequencies across taxa:					
15						
16	Taxon	p(A)	p(C)	p(G)	p(T|U)	
17	Bivalvia_1Observed	2814	1964	2426	2493	
18	Bivalvia_1Expected	2555,16	2243,89	2563,89	2333,1	
19	Bivalvia_2Observed	2571	2095	2450	2322	
20	Bivalvia_2Expected	2486,91	2183,95	2495,41	2270,78	
21	BrachiopodaObserved	2608	2292	2614	2301	
22	BrachiopodaExpected	2586,25	2271,19	2595,09	2361,49	
23	Gastropoda_1Observed	1743	1765	1826	1626	
24	Gastropoda_1Expected	1833,96	1610,54	1840,22	1674,58	
25	Gastropoda_2Observed	2373	2233	2461	2129	
26	Gastropoda_2Expected	2423,15	2127,95	2431,42	2212,56	
27	Nemertea_1Observed	2432	2210	2531	2272	
28	Nemertea_1Expected	2488,76	2185,57	2497,26	2272,47	
29	Nemertea_2Observed	2301	2179	2561	2203	
30	Nemertea_2Expected	2435,79	2139,06	2444,11	2224,11	
31	PhoronidaObserved	2268	1945	2261	2082	
32	PhoronidaExpected	2254,51	1979,86	2262,21	2058,57	
33						
34	Chi-Square = 137.58038(df = 21), P = 0.0000000000000000038939					
35	Warning: This test ignores correlation due to phylogenetic structure.					

Figure 10: Example of detailed BaCoCa output summary file for chi-square calculations of complete sequence data

3.4 Missing Data Overlap

Recent publications (**????**) showed that non-random distrubition of missing data might substantially influence Maximum Likelihood analyses. Therefore, BaCoCa calculates for each pairwise taxon comparison the percentage of sequence positions in which both taxa have either an indel event (-), an ambiguity character state, or a missing character state (?) at the same position (negative overlap) as well as the percentage of shared data in which both taxa do NOT have an indel event (-), an ambiguity character state, or a missing character state (?) at the same position (positive overlap). BaCoCa results of base homogeneity tests are printed to:

- Summary files (missing_data_overlap/∗.txt) ↪ representing positive and negative missing data overlaps between taxa of complete sequence data, pre-defined taxon subclades, and gene partitions as taxon vs. taxon matrix (see Figure 11). If the -r option is used, heat maps will be generated (see 2.2.4 Generation R Heat Maps (-r Option) for examples).



	A	B	C	D	E	F	G	
1		Amphinomidae	Arenicolidae	Bivalvia_1	Bivalvia_2	Brachiopoda	Capitellidae	Cha
2	Amphinomidae		0	0	0	0	0	
3	Arenicolidae			0,0045662100456621	0,0091324200913242	0,0045662100456621	0	
4	Bivalvia_1				0	0,0045662100456621	0	
5	Bivalvia_2					0	0	
6	Brachiopoda						0	
7	Capitellidae							
8	Chaetopteridae_1							

Figure 11: Example of BaCoCa output file of negative missing data overlap between complete sequence data

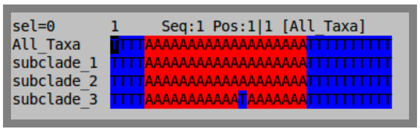
3.5 Invariant Positions

Determination of invariant and variable sequence positions classified into invariant nucleotide or amino-acid character states (state), into purine and pyrimidine invariant positions (class) under nucleotide data, as well as into positively, negatively and neutral electrically charged side chains (charge), polar and non-polar charged side chains (polarity), and hydrophobic and hydrophilic charged side chains (structure) under amino-acid data. BaCoCa results of invariant positions are printed as graphic, text and fasta files for each classification to:

- Summary files (invariant_alignment_positions/fas/∗_invariant_msa_positions.fas) ↪ Invariable positions of each classification for complete sequence data, pre-defined taxon subclades, and gene partitions as fasta file in which a ”T” encodes for invariant sequence positions and an ”A” encodes for variabel sequence positions. This allows another way for graphical display of invariabel and variable site positions which can be easily opened even for very large datasets (see Figure 12)

- Summary files (invariant_alignment_positions/svg/∗_invariant_msa_positions.svg) ↪ Invariable positions of each classification for complete sequence data, pre-defined taxon subclades, and gene partitions as svg file in which different colors encode for invariant and variabel sequence positions (see Figure 13)

- Summary files (invariant_alignment_positions/txt/∗_invariant_msa_positions.txt) ↪ Invariable positions of each classification for complete sequence data, pre-defined taxon subclades, and gene partitions as text file in which a ”1” encodes for invariant sequence positions and an ”0” encodes for variabel sequence positions (see Figure 14)



```
sel=0       1     Seq:1 Pos:1|1 [All_Taxa]
All_Taxa    TTTTAAAAAAAAAAAAAAAAAAATTTTTTTTTT
subclade_1  TTTTAAAAAAAAAAAAAAAAAAATTTTTTTTTT
subclade_2  TTTTAAAAAAAAAAAAAAAAAAATTTTTTTTTT
subclade_3  TTTTAAAAAAAAAAATAAAAAAATTTTTTTTTT
```

Figure 12: Example cut out of BaCoCa FASTA output of invariabel purine or pyrimidine states (class invariabel ↪ T; class variable ↪ A) for complete and partitioned nc data

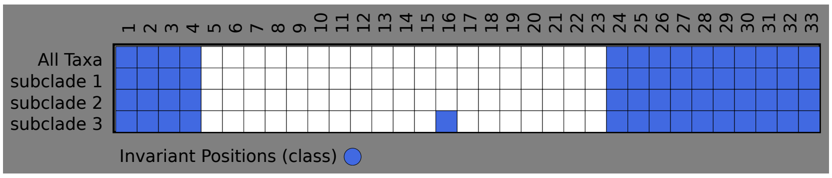
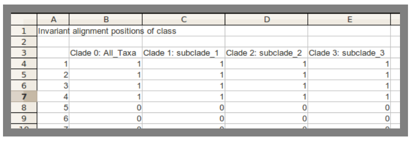


Figure 13: Example of BaCoCa SVG output of invariabel purine or pyrimidine states (class ↪ colored) for complete and partitioned nc data



	A	B	C	D	E
1	Invariant alignment positions of class				
2					
3		Clade 0: All_Taxa	Clade 1: subclade_1	Clade 2: subclade_2	Clade 3: subclade_3
4	1	1	1	1	1
5	2	1	1	1	1
6	3	1	1	1	1
7	4	1	1	1	1
8	5	0	0	0	0
9	6	0	0	0	0

Figure 14: Example of BaCoCa FASTA output of invariabel purine or pyrimidine states (class) for complete and partitioned nc data

3.6 Compositional Heterogeneity and Bias (RCFV Value)

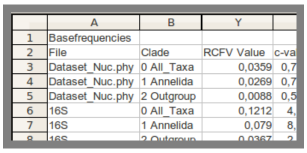
RCFV values to assess compositional heterogeneity (**?**). RCFV is calculated using the following formula:

$$\sum_{i=1}^{n}\sum_{j=A}^{TorU/W}\frac{|\mu_{ij}-\overline{\mu_j}|}{n} \tag{2}$$

i = taxon; j = nucleotide or amino-acid; μ = frequency

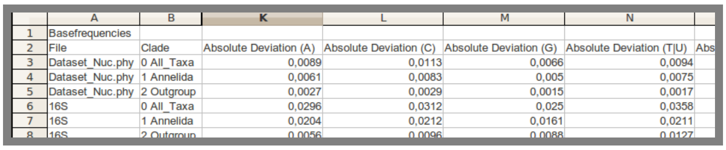
Thus, RCFV measures the absolute deviation from the mean for each amino-acid or nucleotide and taxon and sums these up over all taxa and amino-acids/nucleotides. The higher the RCFV value is, the higher is the degree of compositional heterogeneity in that partition. Summing up only over all amino-acids/nucleotides for each taxon results in the taxon-specific RCFV value. On the other hand, summing up only over all taxa for each amino-acid/nucleotide results in the character state-specific RCFV (csRCFV) value. Besides individual character states this is also done for classes of character states such as purines or polar amino-acids as well as for ambiguity characters. In specific, BaCoCa in the moment calculates the following character state-specific RCFV (csRCFV) values for ambiguity, indel event (-), AT, purine and pyrimidine states (nucleotide data) as well as hydrophobic and hydrophilic, polar and non-polar, positive, neutral and negative, ambiguity (X), and indel event (-) for amino-acid data. BaCoCa results of calculated RCFV values are printed to:

- Summary file (summarized_taxon_base_frequencies.txt) for all taxa of complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 15 & 16)
- Summary files (taxon_base_frequencies_single_partitions/∗_summarized_taxon_base_frequencies.txt) for single taxa of complete sequence data (see Figure 17 & 18)
- Summary files (taxon_base_frequencies_all_partitions/RCFV_frequencies_all_partitions.txt) of taxon specific RCFV values for each taxon vs. complete sequence data and single gene partitions ↪ useful for taxa-gene heat maps and hierarchical clustering (see Figure 19). If the -r option is used, heat maps will be generated (see 2.2.4 Generation R Heat Maps (-r Option) for examples).
- Summary files (compositional_bias/∗_absolute_deviation_frequencies_all_partitions.txt) for individual frequencies by taxon vs each defined partition and the concatenated dataset ↪ useful for taxa-gene heat maps and hierarchical clustering (see Figure 20). If the -r option is used, heat maps will be generated (see 2.2.4 Generation R Heat Maps (-r Option) for examples).



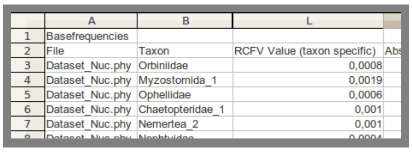
	A	B	Y	
1	Basefrequencies			
2	File	Clade	RCFV Value	c-va
3	Dataset_Nuc.phy	0 All_Taxa	0,0359	0,7
4	Dataset_Nuc.phy	1 Annelida	0,0269	0,7
5	Dataset_Nuc.phy	2 Outgroup	0,0088	0,5
6	16S	0 All_Taxa	0,1212	4,
7	16S	1 Annelida	0,079	8,
8	16S	2 Outgroup	0,0367	2,

Figure 15: Example cut out of BaCoCa summarized frequency output file showing single RCFV-values, separately summarized over all taxa for complete sequence data, pre-defined taxon subclades, and gene partitions



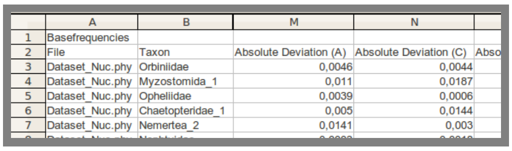
	A	B	K	L	M	N	
1	Basefrequencies						
2	File	Clade	Absolute Deviation (A)	Absolute Deviation (C)	Absolute Deviation (G)	Absolute Deviation (T|U)	Abs
3	Dataset_Nuc.phy	0 All_Taxa	0,0089	0,0113	0,0066	0,0094	
4	Dataset_Nuc.phy	1 Annelida	0,0061	0,0083	0,005	0,0075	
5	Dataset_Nuc.phy	2 Outgroup	0,0027	0,0029	0,0015	0,0017	
6	16S	0 All_Taxa	0,0296	0,0312	0,025	0,0358	
7	16S	1 Annelida	0,0204	0,0212	0,0161	0,0211	
8	16S	2 Outgroup	0,0056	0,0096	0,0088	0,0127	

Figure 16: Example cut out of BaCoCa summarized frequency output file showing single absolute deviation values, separately summarized over all taxa for complete sequence data, pre-defined taxon subclades, and gene partitions



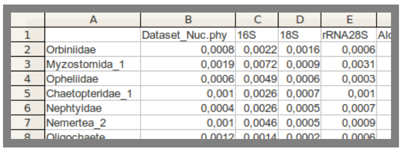
	A	B	L	
1	Basefrequencies			
2	File	Taxon	RCFV Value (taxon specific)	Abs
3	Dataset_Nuc.phy	Orbiniidae	0,0008	
4	Dataset_Nuc.phy	Myzostomida_1	0,0019	
5	Dataset_Nuc.phy	Opheliidae	0,0006	
6	Dataset_Nuc.phy	Chaetopteridae_1	0,001	
7	Dataset_Nuc.phy	Nemertea_2	0,001	
8	Dataset_Nuc.phy	Nephtyidae	0,0004	

Figure 17: Example cut out of BaCoCa frequency output file for RCFV-values of single taxa of complete sequence data



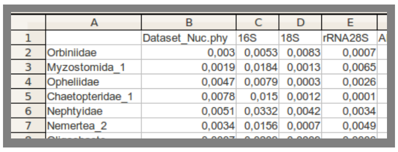
	A	B	M	N	
1	Basefrequencies				
2	File	Taxon	Absolute Deviation (A)	Absolute Deviation (C)	Abso
3	Dataset_Nuc.phy	Orbiniidae	0,0046	0,0044	
4	Dataset_Nuc.phy	Myzostomida_1	0,011	0,0187	
5	Dataset_Nuc.phy	Opheliidae	0,0039	0,0006	
6	Dataset_Nuc.phy	Chaetopteridae_1	0,005	0,0144	
7	Dataset_Nuc.phy	Nemertea_2	0,0141	0,003	
8	Dataset_Nuc.phy	Nephtyidae	[illegible]	[illegible]	

Figure 18: Example cut out of BaCoCa frequency output file for absolute deviation values of single taxa of complete sequence data



	A	B	C	D	E	
1		Dataset_Nuc.phy	16S	18S	rRNA28S	Alo
2	Orbiniidae	0,0008	0,0022	0,0016	0,0006	
3	Myzostomida_1	0,0019	0,0072	0,0009	0,0031	
4	Opheliidae	0,0006	0,0049	0,0006	0,0003	
5	Chaetopteridae_1	0,001	0,0026	0,0007	0,001	
6	Nephtyidae	0,0004	0,0026	0,0005	0,0007	
7	Nemertea_2	0,001	0,0046	0,0005	0,0009	
8	Oligochaete	0,0012	0,0014	0,0002	0,0006	

Figure 19: Example cut out of BaCoCa taxon specific RCFV values for each taxon vs. complete sequence data and single gene partitions



	A	B	C	D	E	
1		Dataset_Nuc.phy	16S	18S	rRNA28S	A
2	Orbiniidae	0,003	0,0053	0,0083	0,0007	
3	Myzostomida_1	0,0019	0,0184	0,0013	0,0065	
4	Opheliidae	0,0047	0,0079	0,0003	0,0026	
5	Chaetopteridae_1	0,0078	0,015	0,0012	0,0001	
6	Nephtyidae	0,0051	0,0332	0,0042	0,0034	
7	Nemertea_2	0,0034	0,0156	0,0007	0,0049	
8	Oligochaete	[illegible]	[illegible]	[illegible]	[illegible]	

Figure 20: Example cut out of BaCoCa taxon specific absolute deviation values for each taxon vs. complete sequence data and single gene partitions

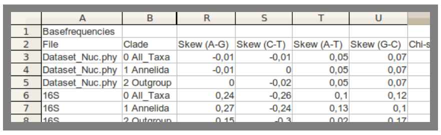
3.7 Skew Values

Calculation of skew values for A/G, C/T. A/T, G/C between nucleotide sequences. The skew values assess either strand-specific biases (i.e. A/T and G/C; **?**) or biases within purines and pyrimidines (i.e. A/G and C/T; **?**). The formula to calculate the skew values is the same for all four and here the one for the A/T skew is provided as an example:

$$A/TSkew = \frac{\mu_A - \mu_T}{\mu_A + \mu_T} \tag{3}$$

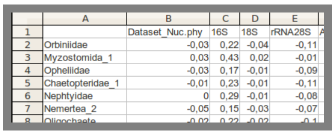
BaCoCa results of calculated skew values are printed to:

- Summary file (summarized_taxon_base_frequencies.txt) for all taxa of complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 21)
- Summary file (skew_values/$*$_skew_values_all_partitions.txt) for each taxon of complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 22). If the -r option is used, heat maps will be generated (see 2.2.4 Generation R Heat Maps (-r Option) for examples).



	A	B	R	S	T	U	
1	Basefrequencies						
2	File	Clade	Skew (A-G)	Skew (C-T)	Skew (A-T)	Skew (G-C)	Chi-s
3	Dataset_Nuc.phy	0 All_Taxa	-0,01	-0,01	0,05	0,07	
4	Dataset_Nuc.phy	1 Annelida	-0,01	0	0,05	0,07	
5	Dataset_Nuc.phy	2 Outgroup	0	-0,02	0,05	0,07	
6	16S	0 All_Taxa	0,24	-0,26	0,1	0,12	
7	16S	1 Annelida	0,27	-0,24	0,13	0,1	
8	16S	2 Outgroup	0,15	-0,3	0,02	0,17	

Figure 21: Example cut out of BaCoCa summarized frequency output file showing single skew values, separately summarized over all taxa for complete sequence data, pre-defined taxon subclades, and gene partitions



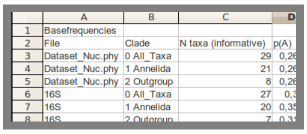
	A	B	C	D	E	
1		Dataset_Nuc.phy	16S	18S	rRNA28S	A
2	Orbiniidae	-0,03	0,22	-0,04	-0,11	
3	Myzostomida_1	0,03	0,43	0,02	-0,01	
4	Opheliidae	-0,03	0,17	-0,01	-0,09	
5	Chaetopteridae_1	-0,01	0,23	-0,01	-0,11	
6	Nephtyidae	0	0,29	-0,01	-0,08	
7	Nemertea_2	-0,05	0,15	-0,03	-0,07	
8	Oligochaete	-0,02	0,22	-0,02	-0,1	

Figure 22: Example cut out of BaCoCa calculated A/G skew values of single taxa for complete sequence data and gene partitions

3.8 Number Informative Taxa

Absolute Numbers of informative taxa (taxa with at least one informative character state for analysed sequence data or gene partition). BaCoCa results of inferred informative taxa are printed to:

- Summary file (summarized_taxon_base_frequencies.txt) summarized for all taxa of complete sequence data, pre-defined taxon subclades, and gene partitions (see Figure 23)



	A	B	C	D
1	Basefrequencies			
2	File	Clade	N taxa (informative)	p(A)
3	Dataset_Nuc.phy	0 All_Taxa	29	0,26
4	Dataset_Nuc.phy	1 Annelida	21	0,26
5	Dataset_Nuc.phy	2 Outgroup	8	0,26
6	16S	0 All_Taxa	27	0,3
7	16S	1 Annelida	20	0,35
8	16S	2 Outgroup	7	0,31

Figure 23: Example cut out of BaCoCa summarized frequency output file showing absolute numbers of informative taxa, summarized for complete sequence data, defined taxon subclades, and gene partitions

3.9 Abbreviations Used In Output Files

Table 6: List of abbreviations or symbols which are used in the BaCoCa output result files.

Abbreviation	Definition
A	Adenine (nucleotide, purine), Alanine (amino-acid, hydrophobic, nonpolar, neutral)
AMB	Ambiguity characters (e.g. N ↪ nucleotide data, X ↪ amino-acid data)
AT	Adenine & Thymin/Uracil content (nucleotide data)
C	Cytosine (nucleotide, pyrimidine), Cysteine (amino-acid, hydrophilic, polar, neutral)
c-value	Saturation value (only nucleotide data, see section 3.2)
charge	Invariant amino-acid positions determined by electrical charge
D	Aspartic acid (amino-acid, hydrophilic, polar, negative)
E	Glutamic acid (amino-acid, hydrophilic, polar, negative)
F	Phenylalanine (amino-acid, hydrophobic, nonpolar, neutral)
G	Guanine (nucleotide, purine), Glycine (amino-acid, hydrophilic, nonpolar, neutral)
GAP	Indel Events (-)
H	Histidine (amino-acid, hydrophilic, polar, neutral)
hydrophilic	Amino-acids with hydrophilic side chain
hydrophobic	Amino-acids with hydrophobic side chain
I	Isoleucin (amino-acid, hydrophobic, nonpolar, neutral)
K	Lysine (amino-acid, hydrophilic, polar, positive)
L	Leucin (amino-acid, hydrophobic, nonpolar, neutral)
M	Methionine (amino-acid, hydrophobic, nonpolar, neutral)
MIS	Missing data (?)
N	Asparagine (amino-acid, hydrophilic, polar, neutral)
negative	Amino-acids with electrically negative charged side chains
neutral	Amino-acids with electrically neutral charged side chains
nonpolar	Amino-acids with nonpolar uncharged side chains
P	Proline (amino-acid, hydrophobic, nonpolar, neutral)
p(...)	proportion of (...)
p_inv(...)	Invariabel proportion of (...)
p-value	Probability of obtaining a test statistic assuming that the null hypothesis is true
polar	Amino-acids with polar uncharged side chains
polarity	Invariant amino-acid positions determined by polarity
positive	Amino-acids with electrically positive charged side chains
purine	Heterocyclic aromatic compound (part of Guanine, Adenine)
pyrimidine	Aromatic heterocyclic organic compound (part of Cytosine, Thymine/Uracil)
Q	Glutamine (amino-acid, hydrophilic, polar, neutral)
R	Arginine (amino-acid, hydrophilic, polar, positive)
RCFV	Value to assess compositional heterogeneity (see section 3.6)
S	Serine (amino-acid, hydrophilic, polar, neutral)
skew value	Assess of either strand-specific biases or biases within purines and pyrimidines
state	Invariant positions determined by character state
structure	Invariant amino-acid positions determined by hydrophilicity
T	Thymine (nucleotide, DNA, pyrimidine), Threonine (amino-acid, hydrophilic, polar, neutral)
U	Uracil (nucleotide, RNA, pyrimidine)
V	Valine (amino-acid, hydrophilic, nonpolar, neutral)
W	Tryptophan (amino-acid, hydrophobic, nonpolar, neutral)
X	Ambiguity state (amino-acid data)
Y	Tyrosine (amino-acid, hydrophilic, polar, neutral)

4 License/Help-Desk/Citation

BaCoCa was developed by Patrick Kück and Torsten Struck and has been written in Perl by Patrick Kück in 20012/13. 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 BaCoCa feel free and write an email to patrick_kueck@web.de which is the official help desk email account for the software. For further free downloadable programs from our institute visit:
http://software.zfmk.de.

If you use BaCoCa please contact P. Kück or T. Struck until the manuscript adressing BaCoCa is published

5 Copyright