# Documentation for **AliStat**

## Quantifying the Sparseness of Multiple Sequence Alignments used in Phylogenetic and Phylogenomic Studies

May, 2017

# Disclaimer

# Credits

The software *AliStat* was developed by the bioinformatics and phylogenomics team in Ecosystem Sciences, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australia.

This project is collaboration between Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australia, and ZFMK, Zentrum für Molekulare Biodiversitätsforschung (zmb), Bonn, Germany.

# Contents

# 1. Terminologies

**Ambiguous characters**

For DNA nucleotides, all the characters which are not considered in the coding types selected by the users are regarded as ambiguous characters. For example, if the selected coding type is “C, T, R (i.e. C, T, AG)”, then the characters other than A, C, G, T, R will be regarded as ambiguous characters, like ‘-’ (i.e. gap), Y, S, W, K, M, B, D, H, V, etc.

For amino acids, we only consider the standard amino acids. Thus, all the characters except A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y are considered as ambiguous characters, like ‘-’ (i.e. gap), B, J, O, U, X, Z, etc.

**$C_a$ - Completeness score for the alignment**

$C_a$ = Total number of unambiguous characters / (number of sequences * length of alignment)

**$C_r$ - Completeness scores for individual sequences**

$C_r$ = Number of unambiguous characters in the sequence / alignment length

**$C_c$ - Completeness scores for individual sites**

$C_c$ = Number of unambiguous characters in the column / number of sequences

**$C_{ij}$ - Completeness scores for pairs of sequences**

$C_{ij}$ = Number of columns in which the corresponding characters of both *i*- and *j*-th sequence are unambiguous / length of alignment. By definition, $C_{ii}$ is always 1.

**$I_{ij}$ - Incompleteness scores for pairs of sequences**

$$I_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 - C_{ij} & \text{otherwise} \end{cases}$$

**$P_{ij}$ - P-distance for pairs of sequences**

Let $W_{ij}$ be the set of columns in which the corresponding characters of both *i*- and *j*-th sequence in the alignment are unambiguous.

$P_{ij}$ = ( Number of substitutions between *i*- and *j*-th sequence along the columns in $W_{ij}$ ) / $|W_{ij}|$

$P_{ij}$ is undefined if $|W_{ij}| = 0$, and $P_{ii}$ is always 0.

## 2. Installation of AliStat

The software was written in C++, and it has been tested under linux and MacOS platform. You need to have C++ compiler installed in the machine in order to compile the source codes. The compilation steps are shown as follows:

```
$ tar -zxvf AliStat_1.7.tar.gz

$ cd AliStat_1.7

$ make
```

Then an executable file named *AliStat* will appear.

## 3. Usage of AliStat

```
Syntax: ./AliStat <alignment file> <data type> [other options]
        ./AliStat -h
```

| | |
|---|---|
| <alignment file> | Multiple alignment file in FASTA format<br><br>The gap is represented by hyphen (i.e. "-")<br><br>An example of a nucleotide alignment file:<br><br>`>Seq_1`<br>`-----CGTACCTCTATTGTATGTTTGATGCCTCTATTGTATGTTTGATGC`<br>`>Seq_2`<br>`-----CGTACATCTGTTGCATGATTAATGGATCTGTTGCATGATTAATGG`<br>`>Seq_3`<br>`-----CGTAC--------------------CTCTATTGTATGTTTGATGC`<br>`>Seq_4`<br>`-----CGTACATCTGTTGCATGATTAATGG--------------------`<br>`>Seq_5`<br>`CGTAC-----ATCTGTTGCATGATTAATGGCTCTATTGTATGTTTGATGC` |
| <data type> | The data type of the alignment file.<br><br>1 - Nucleotide;<br>2 - Amino acid;<br>3 - Mixture of nucleotide and amino acid<br><br>For 3, the user has to specify the data type for each partition inside the *partition file*. For details, please check the option '-p'. |

### Other options:

| | |
|---|---|
| -c <coding type> | The coding type in the alignment.<br>This option is only valid when <data type> = 1.<br>The nucleotides in the alignment will be converted into the corresponding characters according to the coding type the user selected.<br>The possible coding types are:<br>`0  - A, C, G, T [default];`<br>`1  - C, T, R (i.e. C, T, AG);`<br>`2  - A, G, Y (i.e. A, G, CT);`<br>`3  - A, T, S (i.e. A, T, CG);`<br>`4  - C, G, W (i.e. C, G, AT);`<br>`5  - A, C, K (i.e. A, C, GT);`<br>`6  - G, T, M (i.e. G, T, AC);`<br>`7  - K, M    (i.e. GT,   AC);`<br>`8  - S, W    (i.e. GC,   AT);`<br>`9  - R, Y    (i.e. AG,   CT);`<br>`10 - A, B    (i.e. A,   CGT);`<br>`11 - C, D    (i.e. C,   AGT);`<br>`12 - G, H    (i.e. G,   ACT);`<br>`13 - T, V    (i.e. T,   ACG);` |
| -o <output prefix> | Prefix for output files<br>(default: <alignment file> w/o .ext) |
| -n <FILE> | Only consider sequences with names listed in FILE<br>An example of the file (with a sequence name on each line):<br>`Seq_2`<br>`Seq_4`<br>`Seq_5` |
| -t <$n_1$,$n_2$,...> | Only output the tables $n_1$, $n_2$, ...<br>1 - Completeness scores for individual sequences ($C_r$)<br>2 - Completeness scores for individual sites ($C_c$)<br>3 - Distribution of completeness scores for individual sites ($C_c$)<br>4 - Matrix with completeness scores for pairs of sequences ($C_{ij}$)<br>(Together with the heatmaps for $C_{ij}$ scores)<br>5 - Matrix with incompleteness scores for pairs of sequences ($I_{ij}$)<br>6 - Table with completeness scores ($C_{ij}$) and incompleteness scores ($I_{ij}$)<br>By default, the program does not output any table or the heat map image. For details about the tables, please refer to the Section “Output of AliStat”. |

| | |
|---|---|
| -p <FILE> | Specify the partitions in the *partition file*<br><br>For <data type> = 1 or 2, partition file format:<br><br>"<partition name 1>=<start pos>-<end pos>, ..."<br>"<partition name 2>= ... "<br><br>An example:<br><br>`part1=1-50,60-100`<br>`part2=101-200`<br><br>For <data type> = 3, partition file format:<br><br>"<NU/AA>, <partition name 1>=<start pos>-<end pos>, ..."<br>"<NU/AA>, <partition name 2>= ..."<br><br>An example:<br><br>`NU,part1=1-50,60-100`<br>`AA,part2=101-200`<br><br>For nucleotide data, in order to consider specific codon position:<br><br>"<partition name>=<start pos>-<end pos>, ... \3", or<br>"<NU>, <partition name>=<start pos>-<end pos>, ... \3"<br><br>For example:<br><br>`NU,codon1=1-7\3`<br>`NU,codon2=2-8\3`<br>`NU,codon3=3-9\3` |
| -s <$n_3$,$n_4$> | Sliding window analysis: window size = $n_3$; step size = $n_4$<br><br>The first sliding window: [1, $n_3$]<br>The second sliding window: [1+$n_4$, $n_3$+$n_4$]<br>……<br><br>Each sliding window will be regarded as a partition<br><br>This option cannot be used with '-p' at the same time |
| -r <row\|col\|both> | Reorder the rows/columns (or both) of the alignment according to the $C_r$/$C_c$ scores<br><br>All the tables will be displayed according to the reordered alignment<br><br>To output the reordered alignment, please also use the option '-m' |

| | |
|---|---|
| -m <$n_5$> | Mask the alignment; $0 \leq n_5 \leq 1$<br>Output the (reordered) alignment with columns $C_c \geq n_5$ in the file 'Mask.fa', the (reordered) alignment with columns $C_c < n_5$ in the file 'Disc.fa', and the (reordered) alignment with an extra row in the first line to indicate whether the column is masked in the file 'Stat.fa'.<br>Special case: if no <*$n_5$*> is supplied, whole alignment is outputted in the file 'Mask.fa'; if <*$n_5$*> is 0, the alignment with columns $C_c \geq 0$ is outputted in the file 'Mask.fa'. |
| -i <$n_6$> | Generate heat map image for $C_{ij}$ scores of sequence pairs<br>1 - Triangular heat map<br>2 - Rectangular heat map<br>3 - Both<br>(if no <*n6*>, then both triangular and rectangular heat map files are outputted) |
| -d | Report the p distances between sequences in the file with extension '.p-dist.csv'<br>By default, this option is disabled<br>Computation of p distances may take long time for large number of sequences<br>For more information, please refer to the Section “Output of AliStat”. |
| -h | The help page |

# 4. Output of AliStat

## 4.1 Table 1 - completeness scores for individual sequences ($C_r$)

Name of the file: *<output prefix>.table1.csv*, or *<output prefix>.<partition name>.table1.csv* if partition information is provided.

$C_r$ provides the completeness score for individual sequences in the alignment (considering only the positions specified for the partition if the partition exists).

$C_r$ = Number of unambiguous characters in the sequence / alignment length (or partition length)

Referring to the alignment example before, the corresponding table 1 is as follows:

```
Seq ID       Sequence     Sites        Cr
1            Seq_1        45           0.9
2            Seq_2        45           0.9
3            Seq_3        25           0.5
4            Seq_4        25           0.5
5            Seq_5        45           0.9
```

Note: Comma is used as the delimiter inside the file.

## 4.2 Table 2 - completeness scores for individual sites ($C_c$)

Name of the file: *<output prefix>.table2.csv*, or *<output prefix>.<partition name>.table2.csv* if partition information is provided.

$C_c$ provides the completeness score for individual sites in an alignment (considering only the positions specified for the partition if the partition exists).

$C_c$ = Number of unambiguous characters in the column / number of sequences

Referring to the alignment example before, the corresponding table 2 is as follows:

```
Site ID      Cc
1            0.2
2            0.2
3            0.2
4            0.2
5            0.2
6            0.8
7            0.8
8            0.8
9            0.8
10           0.8
.....
```

Note: Comma is used as the delimiter inside the file.

## 4.3 Table 3 - Distribution of C scores for individual sites ($C_c$)

Name of the file: *<output prefix>.table3.csv*, or *<output prefix>.<partition name>.table3.csv* if partition information is provided

This table shows the distribution of $C_c$ scores.

Referring to the alignment example before, the corresponding table 3 is as follows:

```
Here X represents the number of ambiguous characters at a given site

X       Cc      Sites   Proportion      Cumulative
0       1       0       0               0
1       0.8     45      0.9             0.9
2       0.6     0       0               0.9
3       0.4     0       0               0.9
4       0.2     5       0.1             1
5       0       0       0               1
```

Note: Comma is used as the delimiter inside the file.

## 4.4 Table 4 - Matrix with C scores for pairs of sequences ($C_{ij}$)

Name of the file: *<output prefix>.table4.dis*, or *<output prefix>.<partition name>.table4.dis* if partition information is provided

$C_{ij}$ provides the completeness score for each pairs of sequences in an alignment (considering only the positions specified for the partition if the partition exists).

$C_{ij}$ = Number of columns in which the corresponding characters of both *i*- and *j*-th sequence are unambiguous / length of alignment (or length of partition)

By definition, $C_{ii}$ is always 1.

Referring to the alignment example before, the corresponding table 4 is as follows:

```
5
Seq_1   1       0.9     0.5     0.5     0.8
Seq_2   0.9     1       0.5     0.5     0.8
Seq_3   0.5     0.5     1       0.1     0.4
Seq_4   0.5     0.5     0.1     1       0.4
Seq_5   0.8     0.8     0.4     0.4     1
```

Note: Tab is used as the delimiter inside the file.

The number in the first line is the total number of sequences.

Starting from the second line, it shows a symmetric square matrix.

The first column states the name of the sequence for each row of the matrix, and the order of the sequences for the columns is always the same as that for the rows.

## 4.5 Table 5 - Matrix with incompleteness scores for pairs of sequences ($I_{ij}$)

Name of the file: *<output prefix>.table5.dis*, or *<output prefix>.<partition name>.table5.dis* if partition information is provided

$I_{ij}$ provides the incompleteness score for each pairs of sequences in an alignment (considering only the positions specified for the partition if the partition exists).

$$I_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 - C_{ij} & \text{otherwise} \end{cases}$$

Referring to the alignment example before, the corresponding table 5 is as follows:

```
5
Seq_1  0      0.1    0.5    0.5    0.2
Seq_2  0.1    0      0.5    0.5    0.2
Seq_3  0.5    0.5    0      0.9    0.6
Seq_4  0.5    0.5    0.9    0      0.6
Seq_5  0.2    0.2    0.6    0.6    0
```

Note: Tab is used as the delimiter inside the file.

The number in the first line is the total number of sequences.

Starting from the second line, it shows a symmetric square matrix.

The first column states the name of the sequence for each row of the matrix, and the order of the sequences for the columns is always the same as that for the rows.

The format of this file is suitable as a distance matrix for some phylogenetic software tools, like PHYLIP.

## 4.6 Table 6 - Table with completeness scores ($C_{ij}$) and incompleteness scores ($I_{ij}$)

Name of the file: *<output prefix>.table6.csv*, or *<output prefix>.<partition name>.table6.csv* if partition information is provided

$C_{ij}$ provides the completeness score for each pairs of sequences in an alignment (considering only the positions specified for the partition if the partition exists).

$I_{ij}$ provides the incompleteness score for each pairs of sequences in an alignment (considering only the positions specified for the partition if the partition exists).

Referring to the alignment example before, the corresponding table 6 is as follows:

```
taxon 1       taxon 2       Cij    Iij
Seq_1         Seq_1         1      0
Seq_1         Seq_2         0.9    0.1
Seq_1         Seq_3         0.5    0.5
Seq_1         Seq_4         0.5    0.5
Seq_1         Seq_5         0.8    0.2
Seq_2         Seq_1         0.9    0.1
Seq_2         Seq_2         1      0
Seq_2         Seq_3         0.5    0.5
Seq_2         Seq_4         0.5    0.5
…….
```

Note: Comma is used as the delimiter inside the file.

## 4.7 Summary statistics of the alignment

Name of the file: *<output prefix>.summary.txt*, or *<output prefix>.<partition name>.summary.txt* if partition information is provided

This file shows the summary statistics of the alignment (considering only the regions specified for the partition if the partition exists), which includes the following statistics:

- $C_a$ provides the completeness score for the alignment.

  $C_a$ = Total number of unambiguous characters / (number of sequences * length of alignment)

- Maximum and minimum values of $C_r$ (i.e. completeness score for individual sequences in the alignment)

- Maximum and minimum values of $C_c$ (i.e. completeness score for individual sites in the alignment)

- Maximum and minimum values of $C_{ij}$, where $i \neq j$ (i.e. completeness score for each pairs of sequences in the alignment)

As an example, referring to the alignment before (i.e. the example file name: test.fa), the corresponding summary statistics is as follows:

```
Name of input file ............................................ test.fa
Type of data .................................................. nucleotides
Characters used ............................................... ACGT
Number of sequences in the alignment    ....................... 5
Number of sites in the alignment    ........................... 50
Pairs of sequences ............................................ 10
Completeness (C) score for the alignment    (Ca) .............. 0.74
Maximum C-score for individual sequences (Cr_max) ............. 0.9
Minimum C-score for individual sequences (Cr_min) ............. 0.5
Maximum C-score for individual sites (Cc_max) ................. 0.8
Minimum C-score for individual sites (Cc_min) ................. 0.2
Maximum C-score for pairs of sequences (Cij_max, i!=j) ........ 0.9
Minimum C-score for pairs of sequences (Cij_min, i!=j) ........ 0.1

C scores for individual sequences (Cr) stored in .............. test.table1.csv
C scores for individual sites (Cc) stored in .................. test.table2.csv
Distribution of C scores for individual sites (Cc) in ......... test.table3.csv
Matrix with C scores for individual pairs of seqs (Cij) in .... test.table4.dis
Matrix with incompleteness scores (Iij = 1.0 - Cij) in ........ test.table5.dis
Table with C scores (Cij) and incompleteness scores (Iij) in .. test.table6.csv
```

## 4.8 P-distance table – p-distance for pairs of sequences ($P_{ij}$)

This table will be outputted only when the p-distance option (i.e. "*-d*") is used.

Name of the file: *<output prefix>.p-dist.csv*, or *<output prefix>.<partition name>.p-dist.csv* if partition information is provided

$P_{ij}$ provides the p-distance for each pairs of sequences in the alignment (considering only the positions specified for the partition if the partition exists).

Let $W_{ij}$ be the set of columns in which the corresponding characters of both *i*- and *j*-th sequence in the alignment are unambiguous.

$P_{ij}$ = ( Number of substitutions between *i*- and *j*-th sequence along the columns in $W_{ij}$ ) / $|W_{ij}|$

$P_{ij}$ is undefined if $|W_{ij}| = 0$, and $P_{ii}$ is always 0.

Referring to the alignment example before, the corresponding p-distance table is as follows:

```
5
Seq_1  0      0.27   0      0.24   0.15
Seq_2  0.27   0      0.24   0      0.15
Seq_3  0      0.24   0      0      0
Seq_4  0.24   0      0      0      0
Seq_5  0.15   0.15   0      0      0
```

Note: Comma is used as the delimiter inside the file.

The number in the first line is the total number of sequences.

Starting from the second line, it shows a symmetric square matrix.

The first column states the name of the sequence for each row of the matrix, and the order of the sequences for the columns is always the same as that for the rows.

## 4.9 Masked and discarded (reordered) alignment

The marked and the discarded alignment will be outputted only when the option "*-m*" is used.

The masked alignment with columns Cc >= n7 will be outputted in the file '*Mask.fa*' (or '*[partition name].Mask.fa*' if a partition file is used).

The discarded alignment with columns Cc < n7 will be outputted in the file 'Disk.fa' (or '*[partition name].Disc.fa*' if a partition file is used).

If the reorder option "*-r*" is also used, then the rows/columns (or both) of alignment will be sorted according $C_r$/$C_c$ (or both) scores.

The original alignment with an extra line in the beginning to show which columns are included in the masked alignment, will be outputted in the file '*Stat.fa*' (or '*[partition name].Stat.fa*' if a partition file is used).

Referring to the alignment example before, when using the option "-m 0.7", the file '*Stat.fa*' will be as follows:

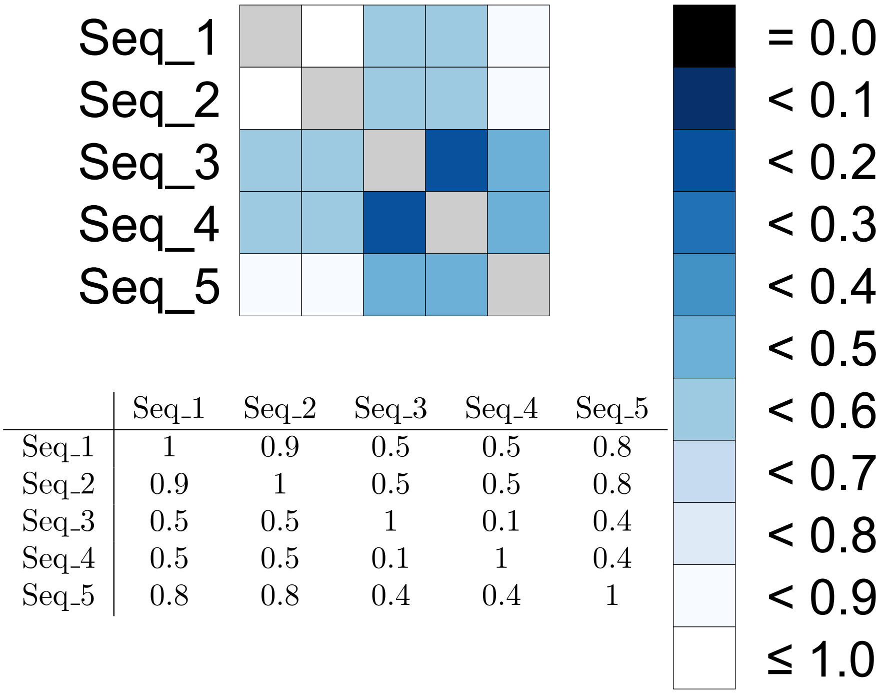
```
>Mask
EEEEEIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
>Seq_1
-----CGTACCTCTATTGTATGTTTGATGCCTCTATTGTATGTTTGATGC
>Seq_2
-----CGTACATCTGTTGCATGATTAATGGATCTGTTGCATGATTAATGG
>Seq_3
-----CGTAC--------------------CTCTATTGTATGTTTGATGC
>Seq_4
-----CGTACATCTGTTGCATGATTAATGG--------------------
>Seq_5
CGTAC-----ATCTGTTGCATGATTAATGGCTCTATTGTATGTTTGATGC
```

Note: 'E' means the column is excluded and 'I' means the column is included in the masked alignment.

## 4.10 Heatmap

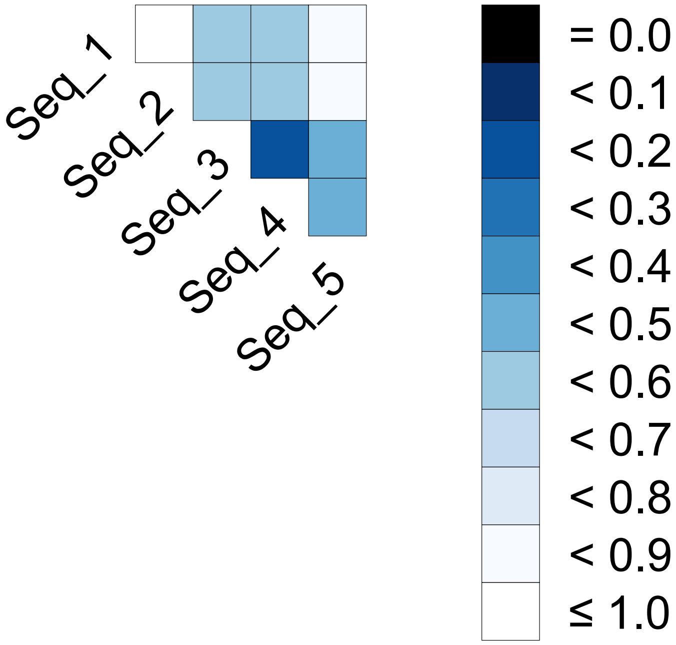
The heatmap files can visually display the $C_{ij}$ values and they are in SVG image format. The rectangular heatmap will be outputted in the file with extension '*.full.svg*', while the triangular one will be outputted in the file with extension '*.triangular.svg*'.

Referring to the alignment example before, the following is rectangular heatmap for table 4:



| | Seq_1 | Seq_2 | Seq_3 | Seq_4 | Seq_5 |
|---|---|---|---|---|---|
| Seq_1 | 1 | 0.9 | 0.5 | 0.5 | 0.8 |
| Seq_2 | 0.9 | 1 | 0.5 | 0.5 | 0.8 |
| Seq_3 | 0.5 | 0.5 | 1 | 0.1 | 0.4 |
| Seq_4 | 0.5 | 0.5 | 0.1 | 1 | 0.4 |
| Seq_5 | 0.8 | 0.8 | 0.4 | 0.4 | 1 |

And the corresponding triangular heatmap table is:



## 5. More examples

Sequence alignment file: M_Ento.fas (Source: http://doi.org/10.5061/dryad.mk8p7 [1])

Data type: Amino acid

Partition file: M_Ento.partition

In this example, we consider three partitions (as stated inside the file *M_Ento.partition*) :

| *Partition name* | *Positions* |
|---|---|
| gene409969g | 1 - 120 |
| gene409979g | 121-290 |
| gene409986g | 291-460 |

Command to run *AliStat* with the option to reorder both the rows and the columns of the alignment according to the Cr and Cc scores, respectively:

```
$./AliStat M_Ento.fas 2 -p M_Ento.partition -r both
```

Then for each partition, the tables 1-4 and the summary file are outputted.

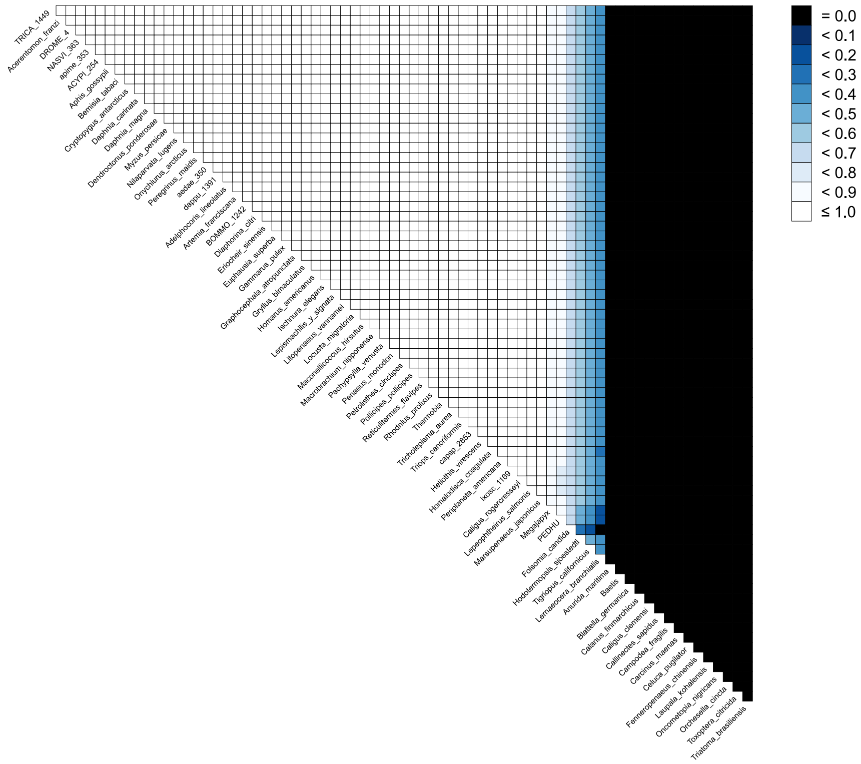
According to the summary files, the completeness scores of the partitions are:

| Partition | Completeness score for the partition ($C_a$) |
|---|---|
| gene409969g | 0.725231 |
| gene409979g | 0.790033 |
| gene409986g | 0.362173 |

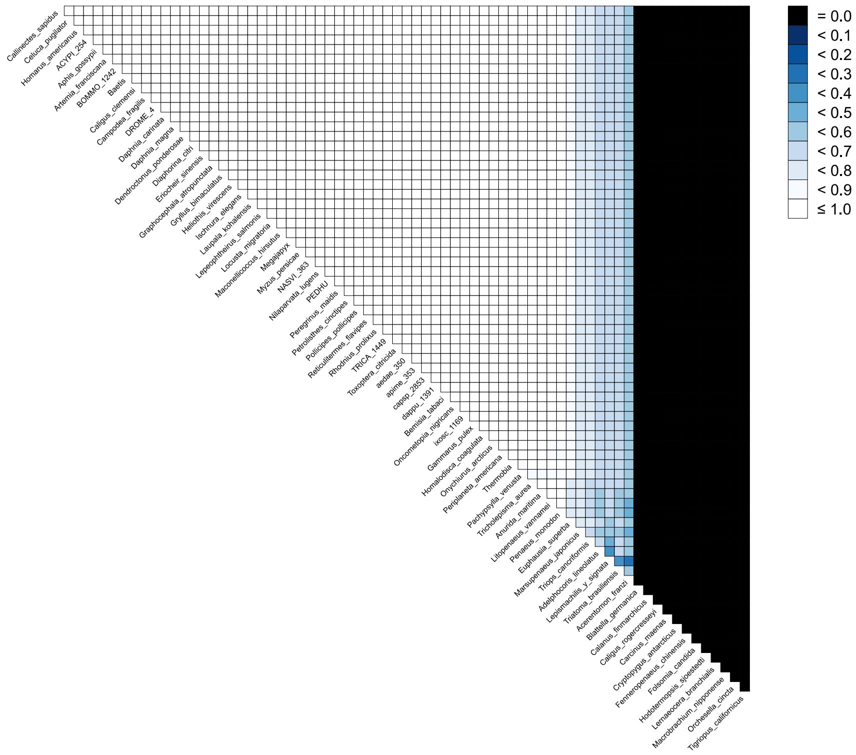
It indicates that the partition *gene409979g* has highest level of completeness, while the partition gene409986g has lowest level of completeness.

Since *AliStat* was run with the option to reorder both the rows and the columns of the alignment according to the $C_r$ and $C_c$ scores respectively, we can see that the resulting sets of heatmap are displayed nicely in a sorted manner.

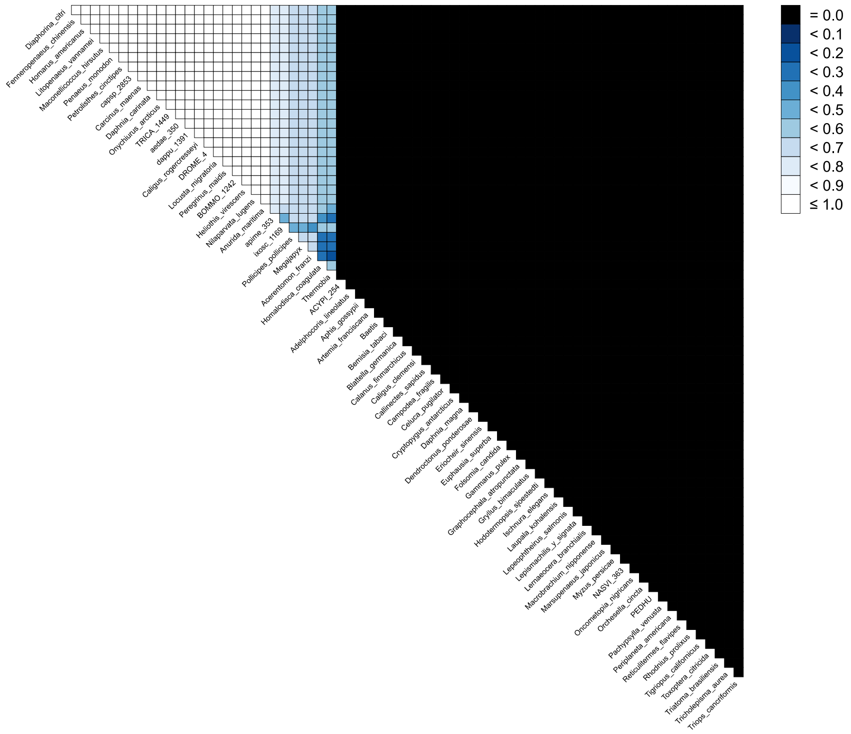
The corresponding heatmap for partition *gene409969g*:



The corresponding heatmap for partition *gene409979g*:



The corresponding heatmap for partition *gene409986g*:



According to these sets of heatmap, there is a large proportion of black colour for the partitions *gene409969g* and *gene409979g*, which indicates that many pairs of sequences overlap completely with each other within these two partitions. On the other hand, for the partition *gene409986g*, there is a large proportion of red colour, which indicates that many pairs of sequences however do not have any overlap with each other within that partition.

## 6. Contact person

Dr Lars Jermiin

Email: lars.jermiin@anu.edu.au

Dr Thomas Wong

Email: thomas.wong@anu.edu.au

Research School of Biology, Australian National University, ACT 2601, Australia

## 7. References

1. Dell'Ampio E., Meusemann K., U. Szucsich N., S. Peters R., Meyer B., Borner J., Petersen M., J. Aberer A., Stamatakis A., G. Walzl M., Quang Minh B., von Haeseler A., Ebersberger I., Pass G., Misof B. (2013) Decisive Datasets in Phylogenomics: Lessons from Studies on the Phylogenetic Relationships of Primarily Wingless Insects. *Molecular Biology and Evolution* doi: 10.1093/molbev/mst196