Molecular diagnoses for high-throughput taxonomy

Ambrosio Torres, Amrita Srivathsan, and Rudolf Meier

July 11, 2023

Introduction

- Taxonomists generally diagnose taxa and identify species relying on morpho-anatomical comparisons, especially by trying to propose primary homology hypotheses.
- ”Integrative taxonomy” has increased the quality and reproducibility of species’ characterization and enabled rapid assessments of biodiversity (Ahrens, 2023*)
- However, many species descriptions based on integrative methods lack diagnoses for all the different data types and only present the barcode for the type or consensus sequences for the species (i.e., they lack molecular diagnoses. Brower, 2010*; Ahrens *et al*., 2021*; Meier *et al*., 2022*).

Available tools for Molecular Diagnoses

- Cladistic Haplotype Aggregation (`CHA`. Brower, 1999*)
- `Spider` (Brown *et al.*, 2012*)
- `FASTACHAR` (Merckelbach & Borges, 2020*)
- `QUIDDICH` (Kühn & Hasse, 2020*)
- `DeSignate` (Hütter *et al.*, 2020*)
- `MOLD` (Fedosov *et al.*, 2022*)

Received: 8 April 2021 | Revised: 19 January 2022 | Accepted: 21 January 2022
DOI: 10.1111/1755-0998.13590

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

MOLD, a novel software to compile accurate and reliable DNA diagnoses for taxonomic descriptions

Alexander Fedosov[1,2] | Guillaume Achaz[2,3,4] | Andrey Gontchar[5] | Nicolas Puillandre[2]

But...

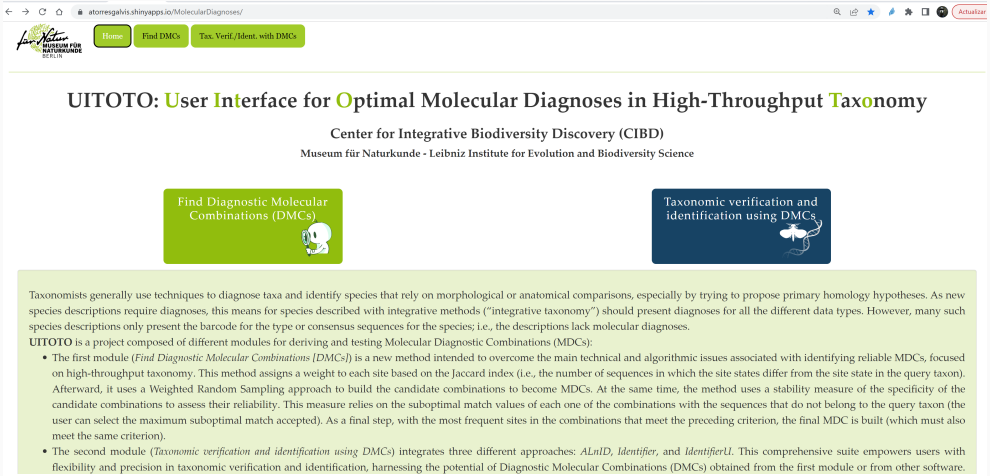
- Most of these tools are very unfriendly/difficult to use. Moreover, its operation is not intuitive, and in some cases, subjective.
- Very Time-consuming (especially for high-throughput taxonomy == thousands of sequences and hundreds of new species).
- Resulting Molecular Diagnostic Combinations (MDCs) with low reliability or unuseful.
- Lack of identification/verification tools.

Objective

- Develop a new method for deriving and testing Molecular Diagnostic Combinations (MDCs) focused in high-throughput taxonomy.

UITOTO (available on: https://atorresgalvis.shinyapps.io/MolecularDiagnoses/)

- A new method for deriving and testing Molecular Diagnostic Combinations (MDCs).



Molecular diagnoses for high-throughput taxonomy

Composition of UITOTO

Module 1*

Find Diagnostic Molecular Combinations (DMCs).

`OpDMC` function

Module 2

Taxonomic verification and identification using DMCs. Three different approaches:

Align + Identify*

`ALnID` function

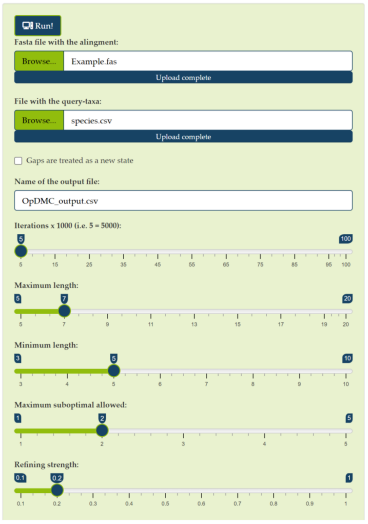
Aligned sequences*

`Identifier` function

Alignment-free*

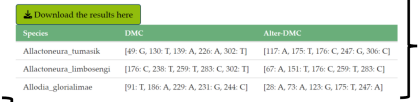
`IdentifierU` function

UITOTO-Module 1 (Find DMCs)



Search parameters

Scripter



Results

```
1  >MC112_hapMC112_SMH|Parempheriella_defectiva
2  AAATAAATGTTGATATAAAATAGGATCACCACCTCCAGCGGGGTCAAAGAAGCTTGTATTTAAAT
3  >ZRCBDP0040965_hapZRCBDP0040965_SMH|Integricypta_shirinae
4  AAATAAATGTTGATATAAAATTGGATCTCCTCCTCCGGCGGGGTCAAAAAATGAGGTGTTTAAAT
5  >ZRCBDP0047054_hapZRCBDP0047054_SMH|Integricypta_teosoonkimae
6  AAATAAATGTTGGTATAAAATTGGATCACCTCCTCCTGCAGGATCAAAAAATGAAGTATTTAAAT
7  >ZRCBDP0047056_hapZRCBDP0047056_SMH|Integricypta_fergusondavie
8  AAATAAATGTTGATATAAAATTGGGTCCCCTCCTCCTGCAGGGTCAAAAAAAGATGTATTTAAAT
9  >ZRCBDP0047057_hapZRCBDP0047057_SMH|Epicypta_foomaosheng
```

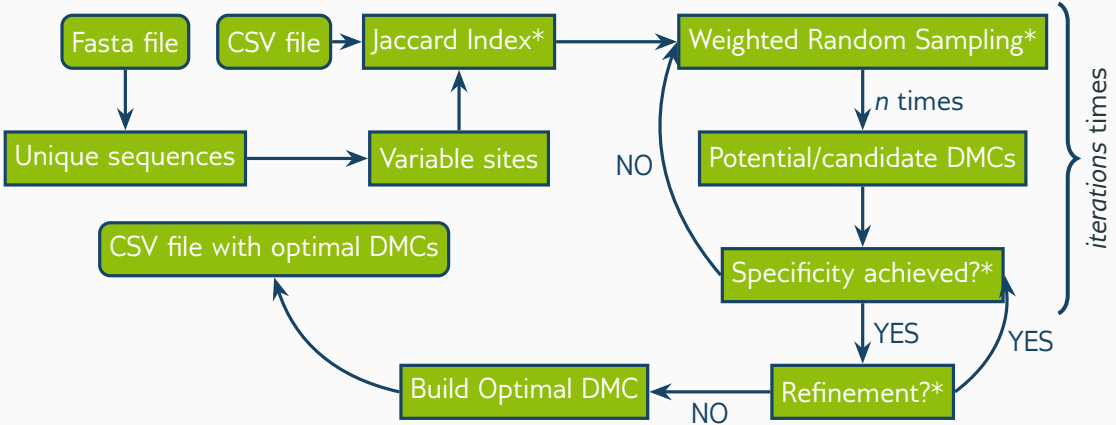
Example.fas

```
1  Allactoneura_tumasik
2  Allactoneura_limbosengi
3  Allodia_glorialimae
4
```

Species.csv

Molecular diagnoses for high-throughput taxonomy

Flowchart of Module 1 (Finding DMCs)



UITOTO-Module 2.1 (Align + Identify)

File with the available DMCs (e.g. Output from the OpDMC approach):

Browse... OpDMC_outputNew.csv

Upload complete

Fasta file with sequences used for obtaining the DMCs included in the previous file:

Browse... Example.fas

Upload complete

Fasta file with specimens to be identified (it should contain at least one specimen):

Browse... Unaligned.fas

Upload complete

Maximum suboptimal allowed: 0 1 5

Name of the output file: IdentificationOutput.csv

Name of the final alignment file: None

Pairwise sequence alignment settings:

Run!

More about alignment settings

Type of pairwise sequence alignment:

- (•) Global (Needleman-Wunsch)
- ( ) Local (Smith-Waterman)
- ( ) Overlap (ends-free)
- ( ) Global-Local
- ( ) Local-Global

Cost for opening a gap in the alignment: 0 10 15

Incremental cost for gap extension: 0 5 15

Maximum accepted Levenshtein distance: 100 300 600

Align against the reference/target matrix using:

- ( ) First taxon
- (•) All sequences available

Parameters

Download the ALnID.R Script here

- Note 1: after a run, it is advisable to refresh the webpage before starting a new run with different settings.
- Note 2: this tool will use the column with the name "DMC" (without spaces) in your file of DMCs. Please be sure about having a column with that name.
- Note 3: if you have a huge number (e.g. >500) of specimens to be identified, you should consider using the script version of this tool. Especially if the "All sequences available" option is used, which is much more time-consuming.

To do that, download the previous script and use the following two commands in an R session (make sure you have the "seqinr" and "Biostrings" packages installed in R):

```
source("ALnID.R")

ALnID("OpDMC_outputNew.csv", "Example.fas", "Unaligned.fas", subopt = 1, type = "global", gapOpening = 10, gapExtension = 4, AcceDist = 300, against = "All", OutName =
"IdentificationOutput.csv", AligName = "None")
```

Scripter

Download the results here

SpecimenNumber	SpecimenName	MatchesWith
1	myceto_ZRCBDP0041017_UrbanForest	-Epicypta_foomaosheng[5/5 [1]]
2	myceto_ZRCBDP0047814_SwampForest	-Neoempheria_chantek[4/5 [55]]-Neoempheria_xinjiapo[4/5 [77]]
3	myceto_ZRCBDP0047920_SwampForest	-Allodia_glorialimae[4/5 [35]]
4	myceto_ZRCBDP0082301_Rainforest	Unidentified
5	myceto_ZRCBDP0103977_Mangrove	-Manota_bukittimah[4/5 [102]]-Neoempheria_mandai[4/5 [108]]

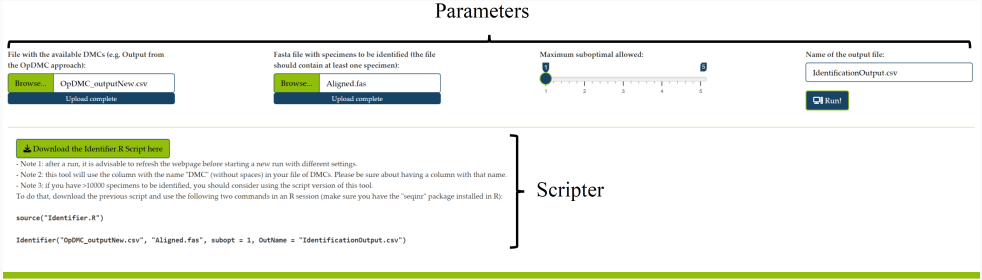
Results

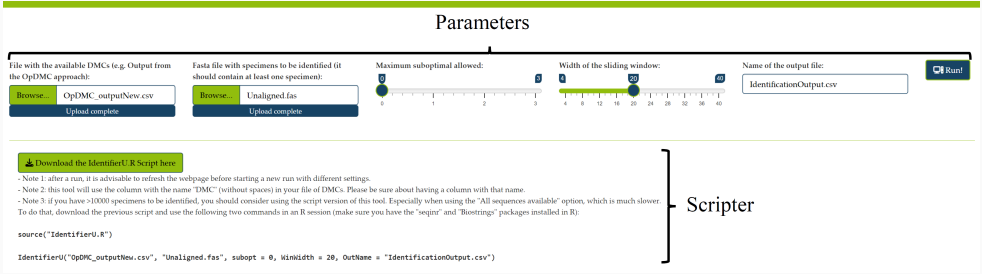
	A	B	C	D
1	Species	DMC	Alter-DMC	
2	1 Allactoneura_tumasik	[49: G, 176: C, 226: A, 247: G, 302: T]	[176: C, 177: T, 184: A, 226: A, 226: A]	
3	2 Allactoneura_limbosengi	[176: C, 238: T, 247: A, 283: C, 302: T]	[67: A, 176: C, 238: T, 247: A, 283: C]	
4	3 Allodia_glorialimae	[160: G, 186: A, 229: A, 231: G, 244: C]	[37: T, 91: T, 186: A, 244: C, 301: T]	
5	4 Allodia_limtzepengi	[37: G, 94: A, 115: G, 172: G, 226: A]	[37: G, 73: A, 109: G, 120: G, 238: T]	
6	5 Allodia_murphyi	[94: A, 160: T, 238: T, 286: A, 289: G]	[115: A, 238: T, 268: A, 286: A, 289: G]	
7	6 Allodia_teopohlengi	[28: G, 103: G, 127: G, 208: G, 226: A]	[28: G, 46: G, 130: T, 208: G, 303: A]	
8	7 Aspidionia_cheesweeleeae	[31: A, 100: A, 124: G, 181: C, 301: G]	[31: A, 37: A, 43: A, 124: G, 301: G]	
9	8 Aspidionia_fatimahae	[34: G, 162: A, 190: G, 199: G, 273: G]	[34: G, 162: A, 165: A, 199: G, 273: G]	
10	9 Aspidionia_janetjesudasonae	[19: G, 76: A, 160: T, 188: C, 189: T]	[28: C, 78: G, 160: T, 188: C, 259: T]	
11	10 Azana_demeijeri	[85: G, 109: G, 112: G, 141: T, 166: G]	[52: T, 97: T, 142: T, 166: G, 166: G]	

OpDMC_outputNew.csv

Molecular diagnoses for high-throughput taxonomy

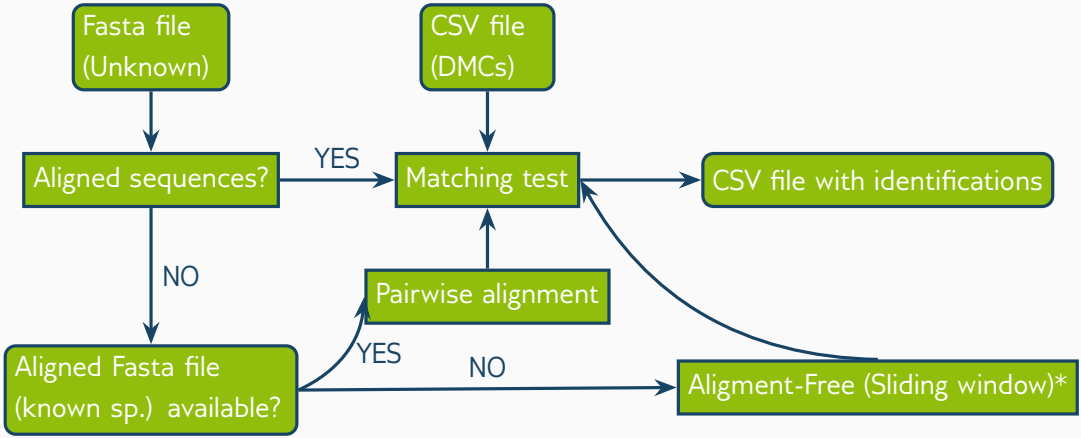
UITOTO-Modules 2.2 (Aligned sequences) and 2.3 (Alignment-free)





Molecular diagnoses for high-throughput taxonomy

Flowchart of Module 2 (Taxonomic verification/identification using DMCs)



Testing UITOTO

- Mycetophilidae (Diptera): 1311 sequences, 118 species, 313 bp.
- We obtained DMCs of minimum defined length (= 5) for all species. Maximum suboptimality = 2.
- DMCs were tested against another Mycetophilidae dataset of 1721 sequences: $\sim$96% of the sequences were identified using the obtained DMCs.
- DMCs were tested against a Mycetophilidae GenBank dataset with $\sim$35077 sequences: $\sim$99.55% did not match with any of the DMCs.
  The Levenshtein distance to those sequences that matched with the DMCs ranged from 29 to 200 (mean of 187).

Comparison of UITOTO with other software tools

Software tool	Site types*	Site scoring	Combination scoring	Verif./Ident.
UITOTO	1 & 5	JI & WRS	Subopt. & freq.	Yes
MOLD*	1 & 5	JI & limit values	Simulations	No
DeSignate*	1, 4 & '5'	JI & polymorphisms	JI= 1 [length= 2]	No
FASTACHAR*	1 & 2	–	–	No
QUIDDICH*	1 – 3	–	–	No
Spider*	1	–	–	No

JI= Jaccard Index; freq= frequencies; Subopt.= Suboptimality; Verif./Ident.= Verification/Identification tool; WRS= Weighted Random Sampling.

Some theoretical and philosophical concerns regarding molecular diagnoses

- **Outdated molecular diagnoses:** "The more taxa are known, the more characters and character states become evident. [Morphological] Diagnoses of taxa then need to be updated to consider newly enclosed, new taxa [...]" (Ahrens, 2023*)
- **Non-universal scope of molecular diagnoses:** all taxonomic hypotheses/categorization is limited by a hierarchical scope.
- **There is no standard about the length of the DMCs:** the same occurs in morphology (especially in Geometric Morphometrics). Moreover, it is easier to standardize molecular diagnoses than morphological diagnoses.

Some theoretical and philosophical concerns regarding molecular diagnoses

- **Composite diagnostic molecular characters are not homologous:** "This relaxed concept of homology has been introduced because [...] the widespread use of landmark configurations [...] not necessarily meet the more rigorous homology criteria of evolutionary biologists." (Palci and Lee, 2019*)
  Only this year three different papers related to the homology concept have been published in *Cladistics*:
  - **Ballego-Campos *et al.*, 2023**. A unified view of homology*
  - **Brower, 2023**. Hierarchies, classifications, cladograms and phylogeny*
  - **Göpel and Richter, 2023**. Homologues and homology and their related terms in phylogenetic systematics*

Acknowledgements

- Leshon Lee and Emily Hartop

Contact: *atorresgalvis@gmail.com* and *Ambrosio.TorresGalvis@mfn.berlin*

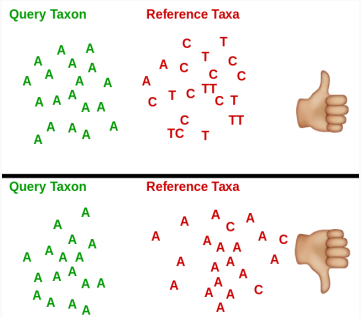
UITOTO: https://atorresgalvis.shinyapps.io/MolecularDiagnoses/

Supplementary material (Jaccard Index [JI] and Weighted Random Sampling [WRS])

JI measures the similarities between finite sample sets. It is formally defined as the size of the intersection divided by the size of the union of the sample sets. In this sense, *1 - Jaccard Index = Jaccard dissimilarity Index*. And it is represented as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Then, a WRS approach is used: sites are weighted (the higher their Jaccard dissimilarity Index, the higher their weight), and the probability of each site being selected is determined by its relative weight.



Module 1 flowchart

Supplementary material (Specificity check)

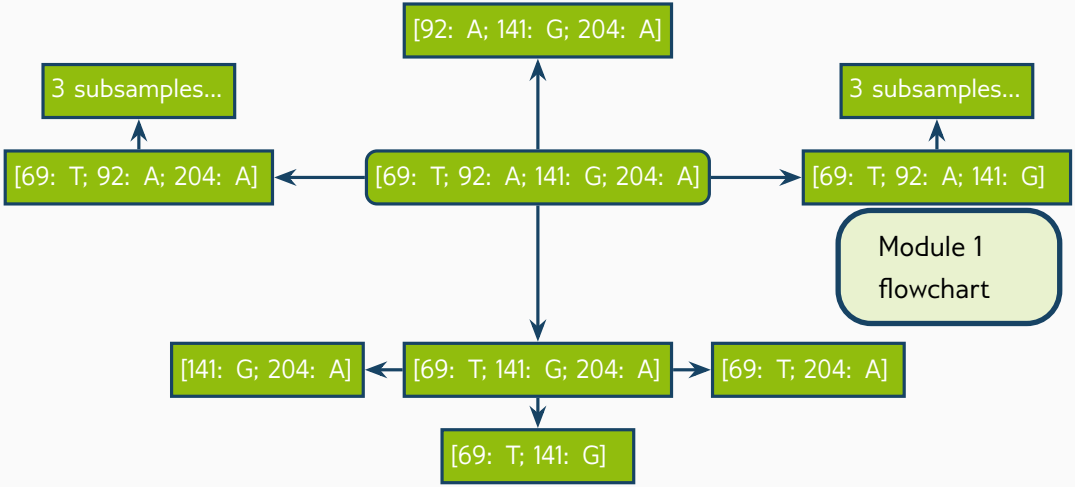
	DMC 1	DMC 2
Query	[49: G, 176: C, 226: A, 247: G, 302: T]	[37: G, 94: A, 115: G, 172: G, 226: A]
Query	[49: G, 176: C, 226: A, 247: G, 302: T]	[37: G, 94: A, 115: G, 172: G, 226: A]
Query	[49: G, 176: C, 226: A, 247: G, 302: T]	[37: G, 94: A, 115: G, 172: G, 226: A]
Reference	[49: G, 176: A, 226: A, 247: G, 302: T]	[37: G, 94: T, 115: A, 172: G, 226: T]
Reference	[49: G, 176: A, 226: A, 247: G, 302: T]	[37: C, 94: A, 115: A, 172: C, 226: T]
Reference	[49: G, 176: A, 226: A, 247: G, 302: T]	[37: G, 94: T, 115: A, 172: G, 226: T]
Reference	[49: G, 176: A, 226: T, 247: G, 302: T]	[37: C, 94: A, 115: A, 172: G, 226: T]
Reference	[49: C, 176: C, 226: T, 247: G, 302: T]	[37: C, 94: A, 115: A, 172: G, 226: T]
Reference	[49: C, 176: A, 226: T, 247: G, 302: T]	[37: C, 94: A, 115: A, 172: C, 226: T]
	Maximum suboptimality: 4/5	**Maximum suboptimality: 2/5**

The extra matches needed to break the specificity of the DMCs are counted. This approach is based on the Bremer support (Bremer, 1994*).

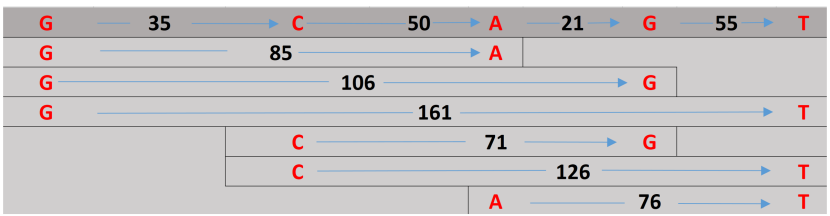
Module 1 flowchart

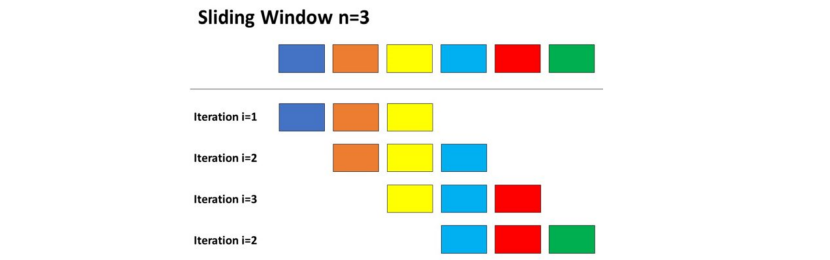
Supplementary material (Refinement)

The refinement strength is the proportion of subsamples to be tested.



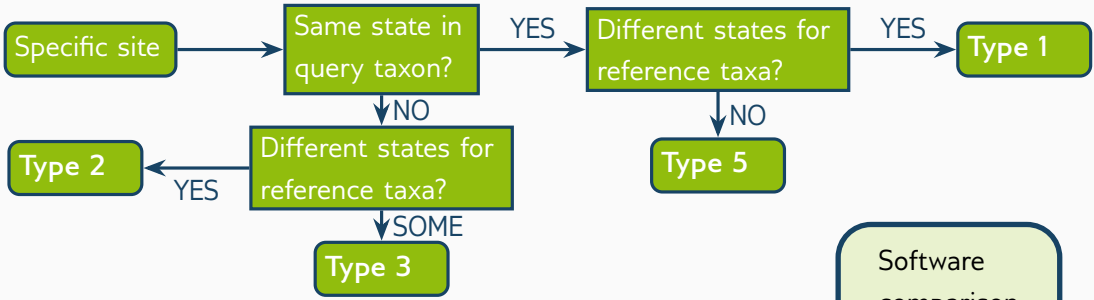
Supplementary material (Alignment-Free)





Module 2 flowchart

Supplementary material (Site types [Kühn and Haase, 2020*])



- Type 1: Query taxon= **A**; Reference taxa= **T, C**
- Type 2: Query taxon= **A, T**; Reference taxa= **C, G**
- Type 3: Query taxon= **C, T**; Reference taxa= **T, G**
- Type 4: an extension of type 1 for pairwise comparisons.
- Type 5: Query taxon= **T**; Reference taxa= **T, G** (Fedosov *et al.*, 2022*)