Taxonomic Assignment and DNA Metabarcoding

Dr. Marie Louise Davey
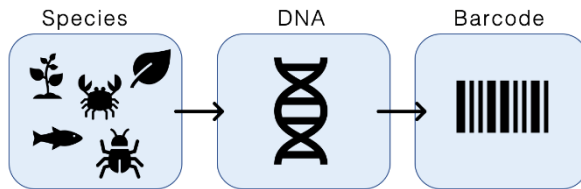
Why assign taxonomy at all?

Not strictly necessary to answer alpha and betadiversity questions

- Detecting shifts in community composition and genetic diversity doesn't require taxonomic assignments
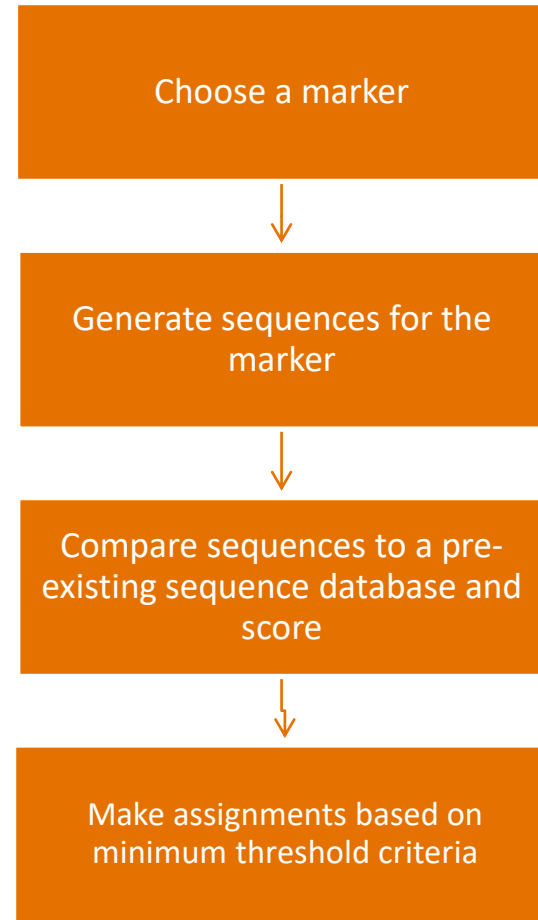
Assigning taxonomy links sequences to a wealth of pre-existing information

- Linking sequences to species improves interpretation and explanation of patterns in alpha and beta diversity

NINA

Taxonomic Assignment

Species   DNA   Barcode

Species 1
Species 2
Species 3

Choose a marker

Generate sequences for the marker

Compare sequences to a pre-existing sequence database and score

Make assignments based on minimum threshold criteria

NINA

Markers

Marker choice impacts taxonomic assignment

- No marker is perfect
  - Markers have taxonomic bias
  - Discriminating power varies between markers and taxonomic groups
  - Database quality, availability, and completeness varies between markers
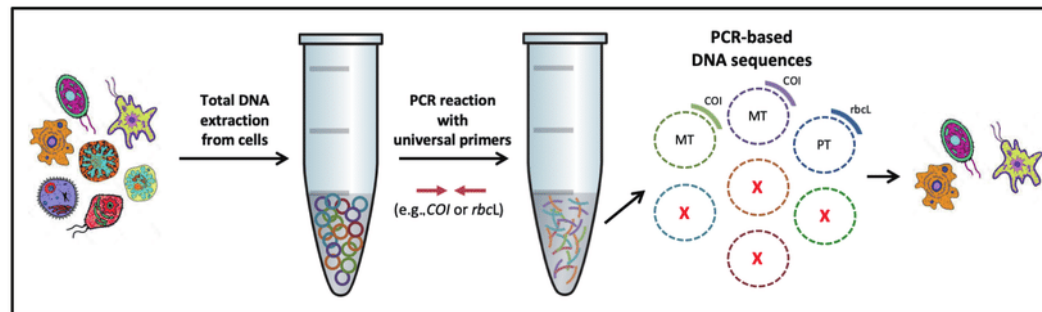
Effect of marker bias

Marker Choice

Coverage and quality of DNA barcode references for Central and Northern European Odonata

Matthias Geiger[1], Stephan Koblmüller[2], Giacomo Assandri[3], Andreas Chovanec[4], Torbjørn Ekrem[6], Iris Fischer[5,7,8], Andrea Galimberti[9], Michał Grabowski[10], Elisabeth Haring[5,7,8], Axel Hausmann[11], Lars Hendrich[11], Stefan Koch[12], Tomasz Mamos[10], Udo Rothe[13], Björn Rulik[1], Tomasz Rewicz[10], Marcia Sittenthaler[7], Elisabeth Stur[6], Grzegorz Tończyk[10], Lukas Zangl[2,14,15] and Jerome Moriniere[16]

Marker Choice

Coverage and quality of DNA barcode references for Central and Northern European Odonata

- COI marker was selected for barcoding

- >80% of European species have a pre-existing publicly available barcode

- 88% of species tested could be resolved using the marker
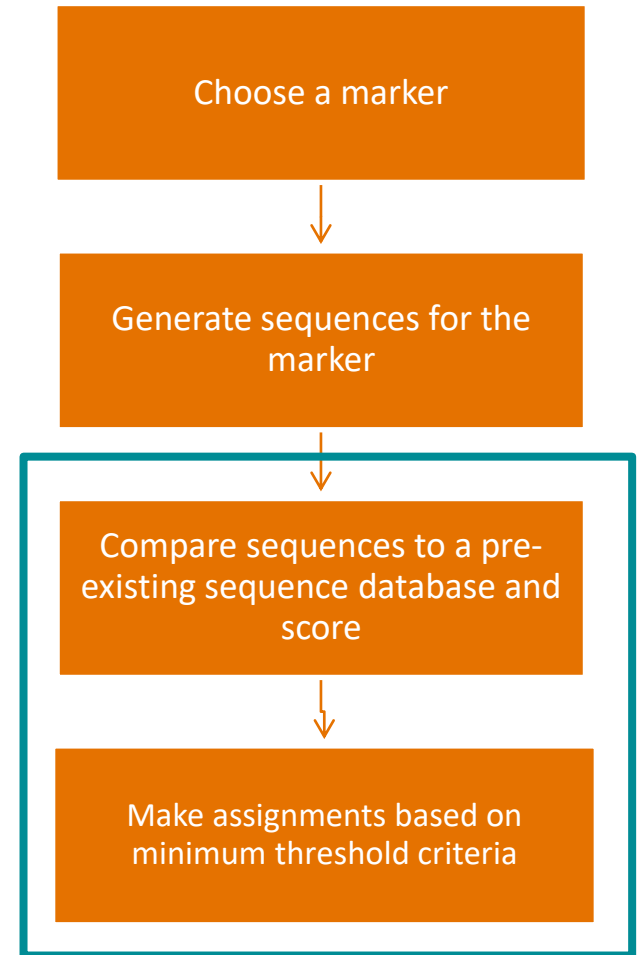
Databases

Taxonomic assignment quality is highly dependent on database accuracy and completeness

- Misidentified sequences create identification errors and low-quality assignments
- Missing reference sequences reduce resolution of taxonomic assignments, or results in misidentifications

Quality of taxonomic assignments can often be improved by creating custom-curated reference databases

NINA

Methods for Taxonomic Assignment

- Diverse algorithms are used for comparing sequences to databases and scoring the results
  - Alignment based
    - BLAST, vsearch, OBItools
  - Phylogenetic based
    - HmmUFOtu, TIPP, DECARD, SAP
  - Kmer-based machine learning approaches
    - RDP, UTAX, SINTAX

Choose a marker

↓

Generate sequences for the marker

↓

Compare sequences to a pre-existing sequence database and score

↓

Make assignments based on minimum threshold criteria

NINA

Alignment based taxonomic assignment

Alignment strategy can be local (BLAST) or global (vsearch)

- Local alignments begin by checking a small piece of the query sequence against the reference database and then expanding the match to find areas of high similarity
- Global alignments find the best match in the reference database across the entire length of the query sequence

Output is typically an alignment score, percent identity, and coverage score

- These are used as criteria for assigning taxonomy

NINA

Alignment based taxonomic assignment

```
Global Alignment:
--AGATCCGGATGGT--GTGACATGCGAT--AAG--AGGCGTT
  ||| |   |   |   |||||| ||||||  |||  |   | ||
GTCCATCTG--TCTTGGGTGAC-TGCGATACAAGTTA--CCTT


Local Alignment:
--AGATCCGGATGGT--GTGACATGCGATA--AG--AGGCGTT
                  |||||| |||||||
GTCCATCTG--TCTTGGGTGAC-TGCGATACAAGTTA--CCTT
```

62% similarity

93% similarity
32% coverage

NINA

Alignment based taxonomic assignment

Pros	Cons
• Extremely well developed infrastructure • Easily applied to custom databases • Computationally inexpensive, particularly for large databases • High assignment rate	• Typically lower taxonomic accuracy, lower precision, and lower taxonomic sensitivity than other methods • Thresholds for successful assignment must be set by the user and require a priori knowledge • No systematic assignment at higher taxonomic levels (unless paired with LCA)

NINA

Criteria for taxonomic assignments

Threshold Cutoffs

- A minimum 'score' must be met for a successful assignment
1. Identity based
   - reference sequence taxonomy is assigned for the best match exceeding the threshold value (often used in BLAST)
2. Confidence intervals
   - Proportion of kmers matching a given taxonomic assignment at a given taxonomic rank (ex/ SINTAX, RDP)

Lowest Common Ancestor (LCA) consensus

- Assigns consensus taxonomy based on a group of pre-defined 'best hits' for each sequence
- Typically requires a separate additional step to calculate the LCA (ex/ LCA*, BASTA, MEGAN)

NINA

Lowest Common Ancestor Calculations

Phylum

Class

Order — Classified to Order

Family

Genus — Classified to Genus

Species

Best Hits   LCA

www.nina.no

NINA
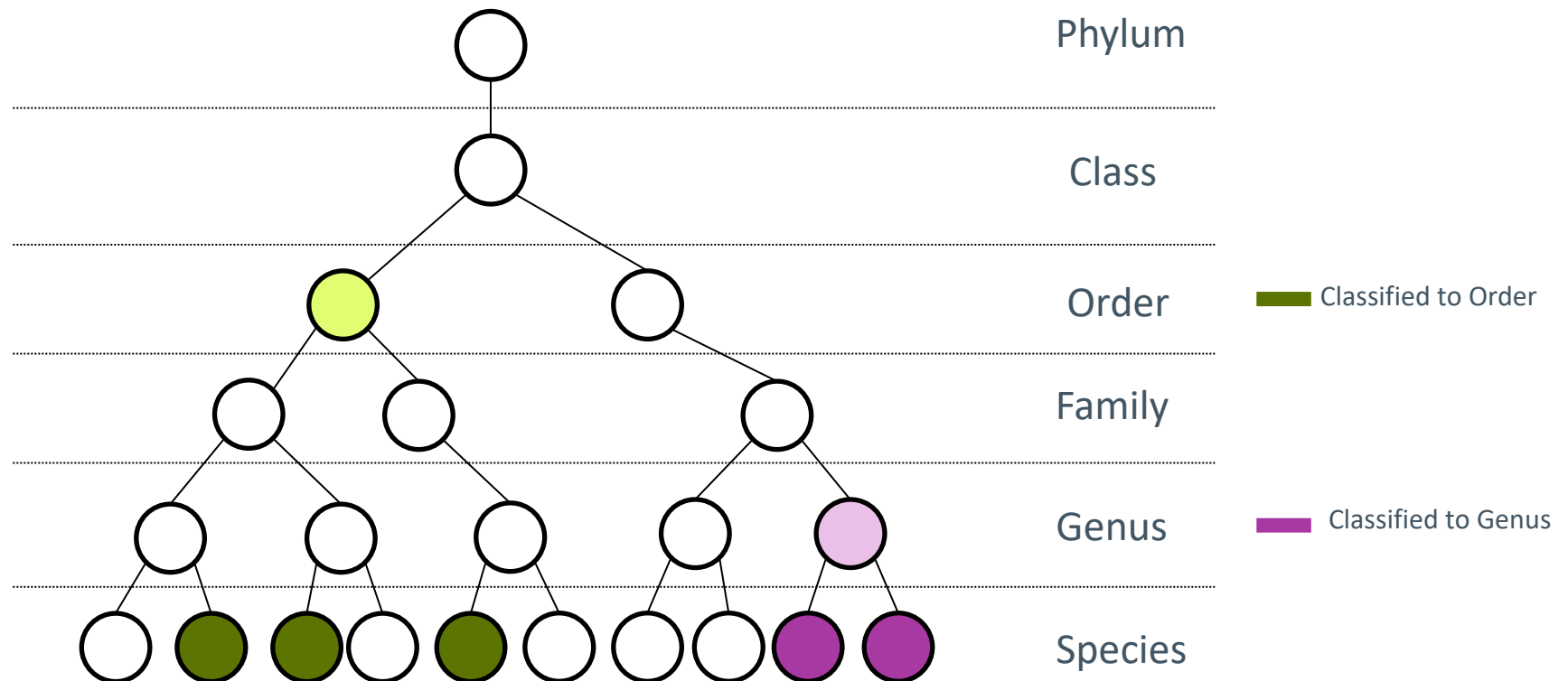
Phylogenetic based taxonomic assignment

Assignment of reads on a known reference tree, often followed by clustering or LCA analysis

1. Placement on the reference tree is typically done using the pplacer algorithm, which is based on HMMER alignments of the query sequences with the reference database. This is often followed by clustering or LCA analysis

2. Alignment based searches (BLAST) are used to identify relevant reference sequences, and MCMC or bootstrapping are used to evaluate a series of phylogenetic trees generated from the query + references

Output is typically a taxonomy string with accompanying support statistics

NINA

Phylogenetic based taxonomic assignment

Phylogenetic based taxonomic assignment

Pros	Cons
• Assignment thresholds are flexible, and can vary between taxonomic groups • Provides phylogenetic data about sequences • Good taxonomic precision and sensitivity	• Requires a known reference phylogeny • Many algorithms require that the target gene region is alignable across the entire target group • Computationally expensive • Taxonomic accuracy can be compromised by an incomplete or poor reference phylogeny

NINA

Machine-learning based taxonomic assignment

Sequences are given a best classification based on a training set

1. Reads are broken into k-mers (sequence fragments of fixed length)
2. K-mers are compared to a training set of reference sequences with known taxonomies
3. Proportion of k-mers assigned to a given taxonomic level are used as a confidence score

Output is typically a taxonomy string with accompanying support statistics for each taxonomic level

NINA

Machine-learning based taxonomic assignment

Pros	Cons
• Provides confidence intervals for all taxonomic levels • Assignment thresholds are flexible, and can vary between taxonomic groups • High taxonomic precision and sensitivity	• Can be computationally intensive • Often requires establishment and benchmarking of a custom training set for the classification algorithm • Database insufficiencies can create inaccuracy

NINA

Uncertainty in Taxonomic Assignments

- One of the biggest sources of variation in metabarcoding results is bioinformatics
  - Of this variation, taxonomic assignments have the biggest impact
  - Quantifying this uncertainty is helpful in providing good monitoring data

NINA
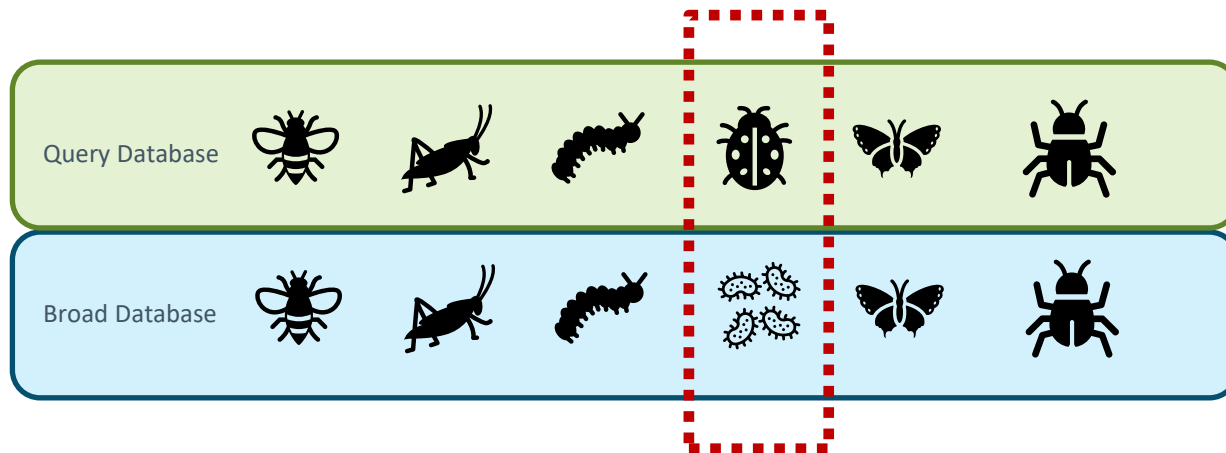
Reccomendation 1:
Probabilistic Assignment Tools

- Algorithm provides statistical confidence estimates at each taxonomic level
 - Enables use of flexible thresholds

	Kingdom	Phylum	Class	Order	Family	Genus	Species
🐝							
🐛							
🐛							
🐞							
🦋							
🐜							
🦗							

100%

0%
Confidence

NINA
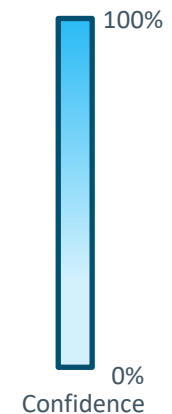
Reccomendation 2:
External Database Controls

- ## Identify potential database contamination
  - ### BLAST training set/database against a broader database (ex/ GenBank)

Reccomendation 3:
Internal Database Controls

- ## Identify groups with poor marker performance
 - ### BLAST training set against itself
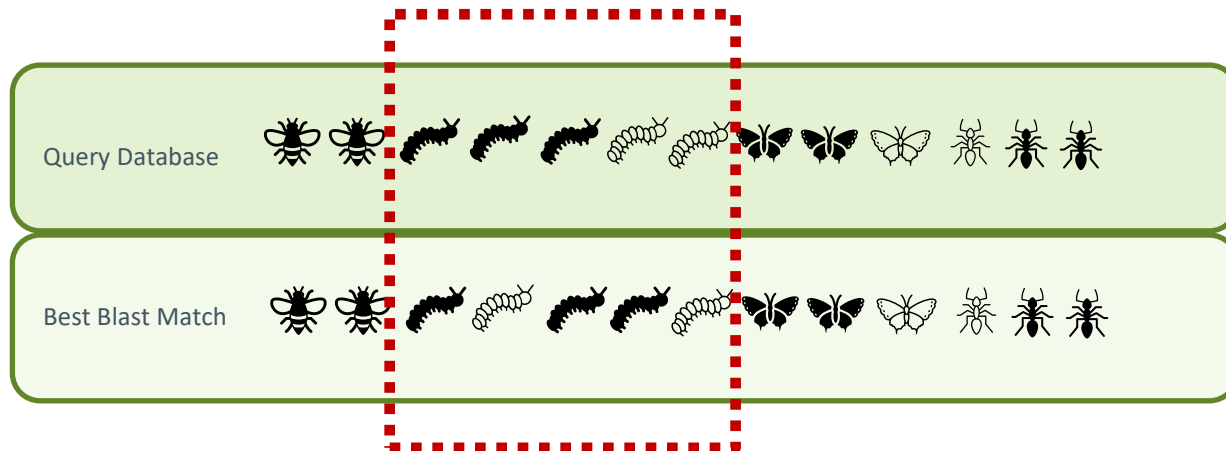
Reccomendation 4:
Assess Database Coverage

- Identify groups in the target geographic area with poor representation in the database
  - Compare database to local species checklists
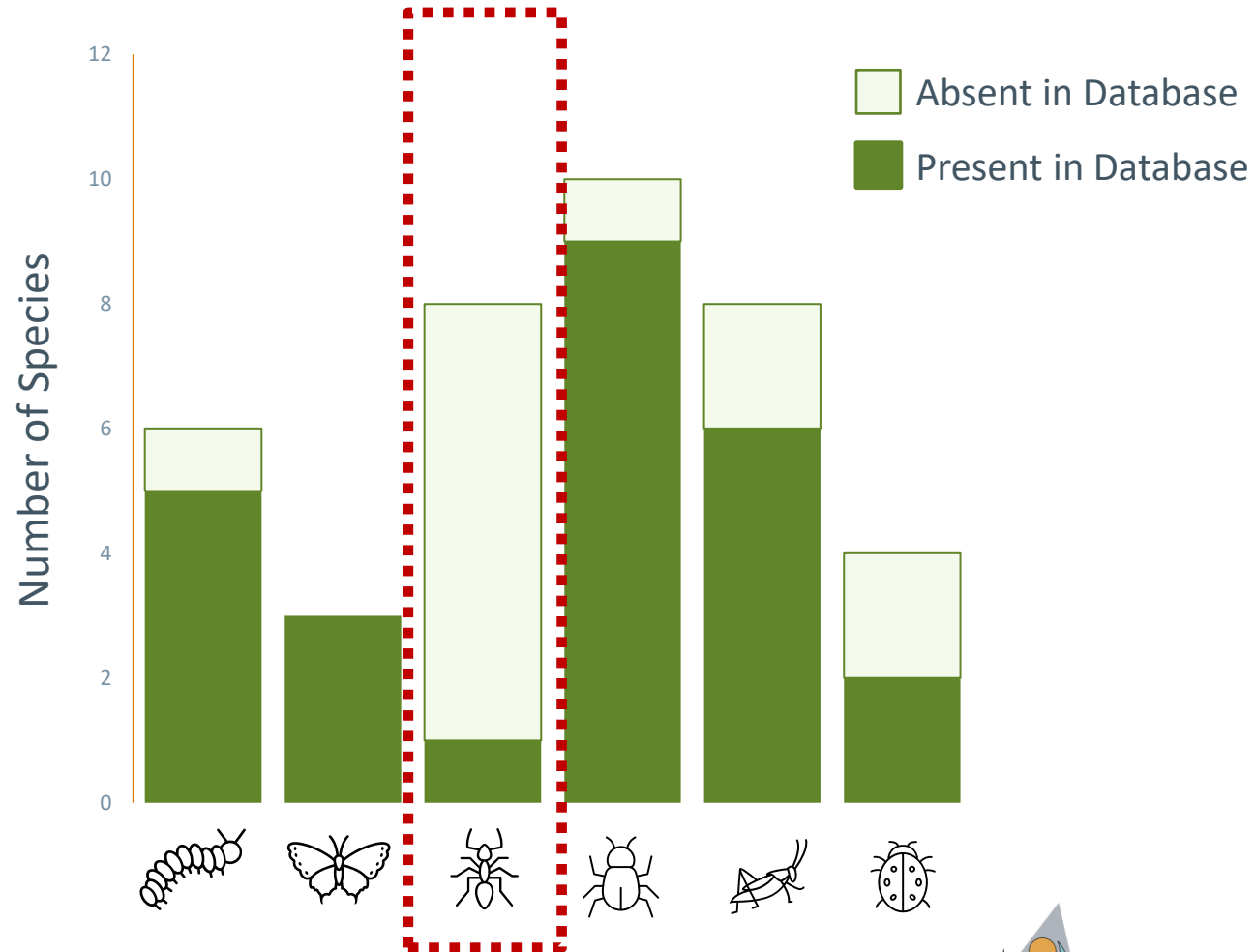
Reccomendation 5:
Assess geographic 'plausability'

- ### Identify taxonomic assignments that are unlikely for the focal geographic area
  - #### Compare assignments to local species checklists
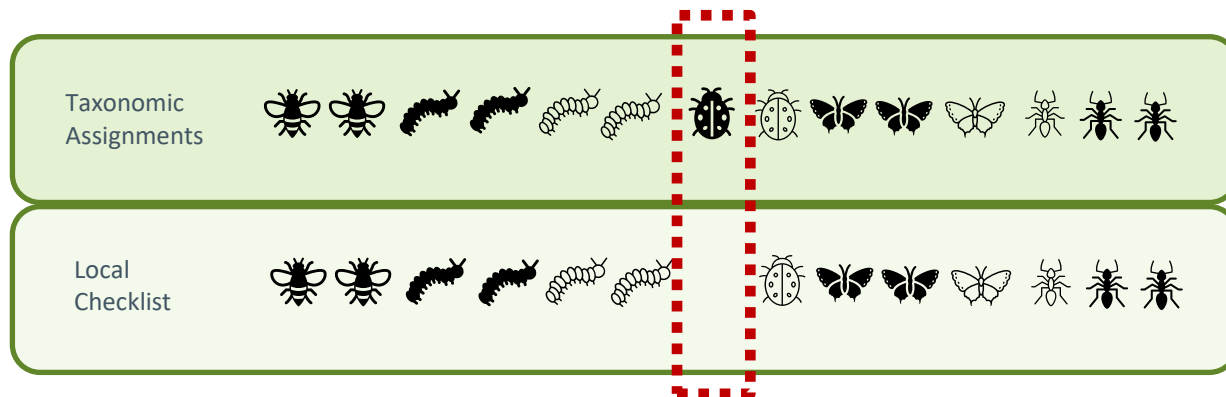
Report Uncertainty

- Use these criteria to categorize uncertainty in taxonomic assignments or flag individual cases for further investigation

	🐝	🐛	🐛	🐞	🦋	🦋	🐜	🐜	🦗
Algorithm Confidence Score									
Database Contamination									
Marker Performance									
Database Coverage									
Geographic Plausability									

NINA

How to decide?

- Find a method that balances your community's diversity with computational time

- Look for existing database resources, particularly for machine learning approaches

- Evaluate the importance of taxonomy to your conclusions

- Be critical, report uncertainty in taxonomic assignments

NINA