Lecture

Molecular Identification and Phylogenetic Basics



Gulzhanay Kamshybayeva, PhD
Senior Lecturer

Detection, identification and relatedness

Detection: is the target sequence present?

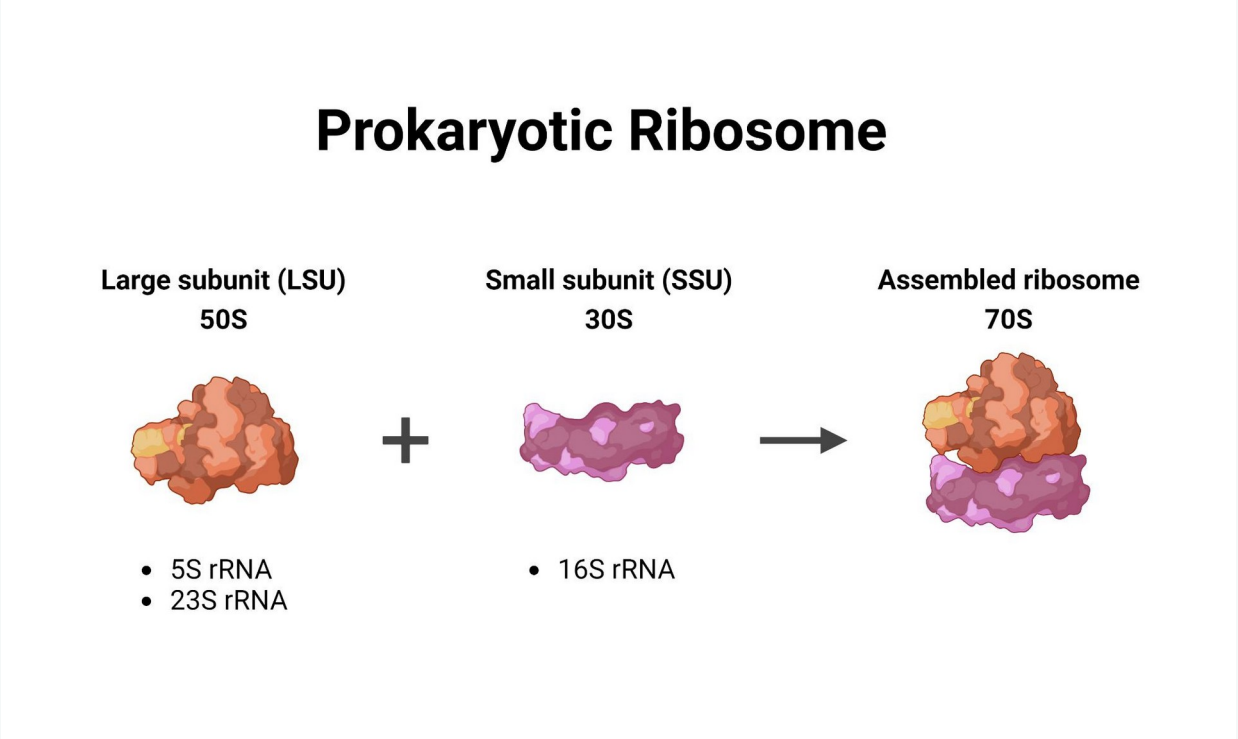
Identification: which taxon best fits the evidence?

Phylogeny: what evolutionary relationships does the dataset support?

Function requires additional evidence.

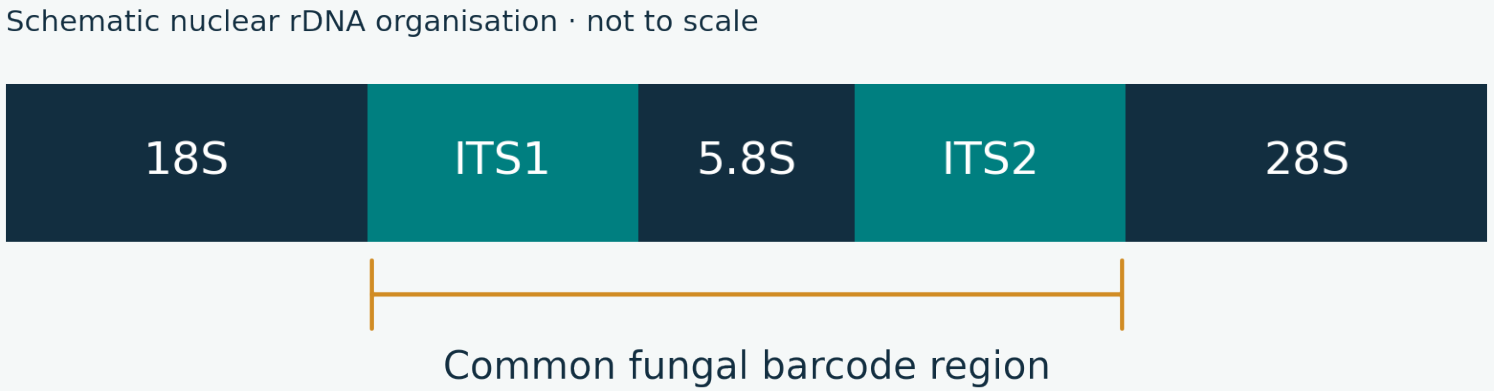
The 16S rRNA gene

A DNA marker encoding the small-subunit ribosomal RNA.
Commonly used for bacterial and archaeal identification.
Full-length sequences are roughly 1.5 kb.
Some distinct species have very similar sequences.



ITS as a fungal barcode

A widely used primary fungal barcode



Close species may require additional, lineage-appropriate loci.

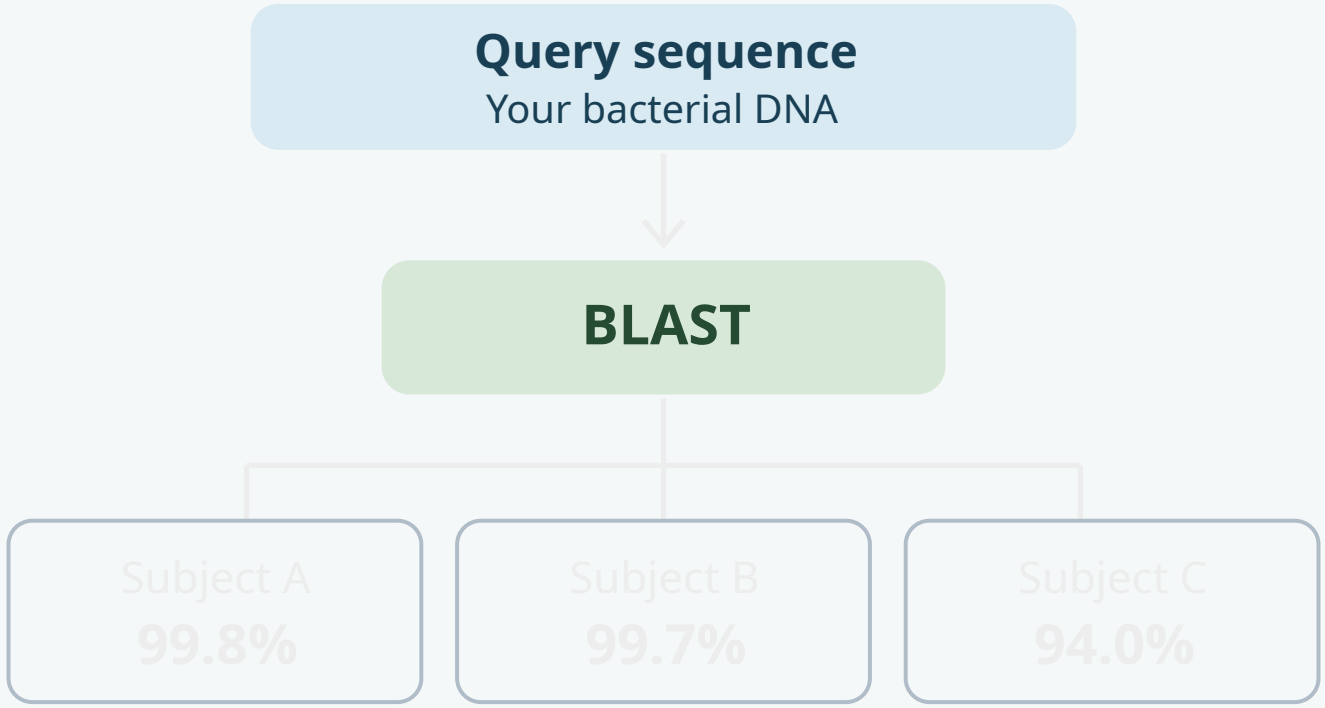
What BLAST compares

Query: the sequence you submit.

Subject: a sequence in the selected database.

BLAST finds local sequence alignments.

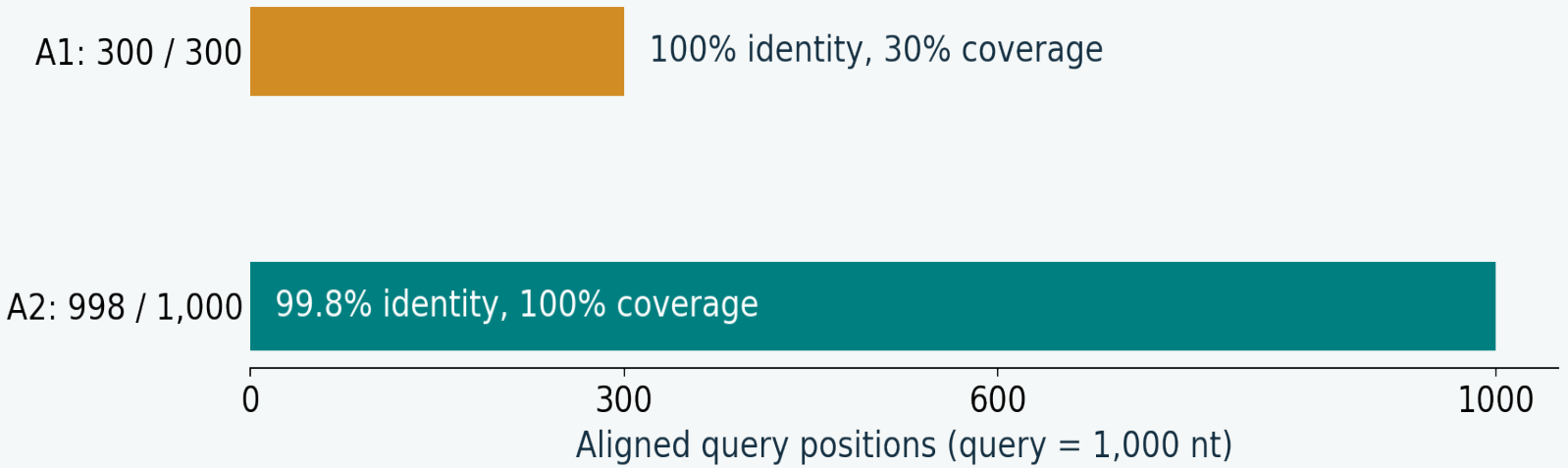
The result depends on the database and search settings.



Identity and query coverage

Identity: agreement within the aligned region

Coverage: fraction of the query represented



Case A teaching data: one ungapped alignment per reference

E-value and biological interpretation

E-value: expected chance matches with this score or better.

Smaller values support statistical significance of the match.

Database size and query properties affect the value.

A displayed zero does not prove species identity.

Case A: competing reference matches

Case A teaching data · 1,000 nt query · negative controls

Hit	Reference	Matched	Aligned nt	Identity	Coverage
A1	G alpha	300	300	100%	30%
A2	G alpha	998	1,000	99.8%	100%
A3	G beta	997	1,000	99.7%	100%
A4	Uncultured	999	1,000	99.9%	100%

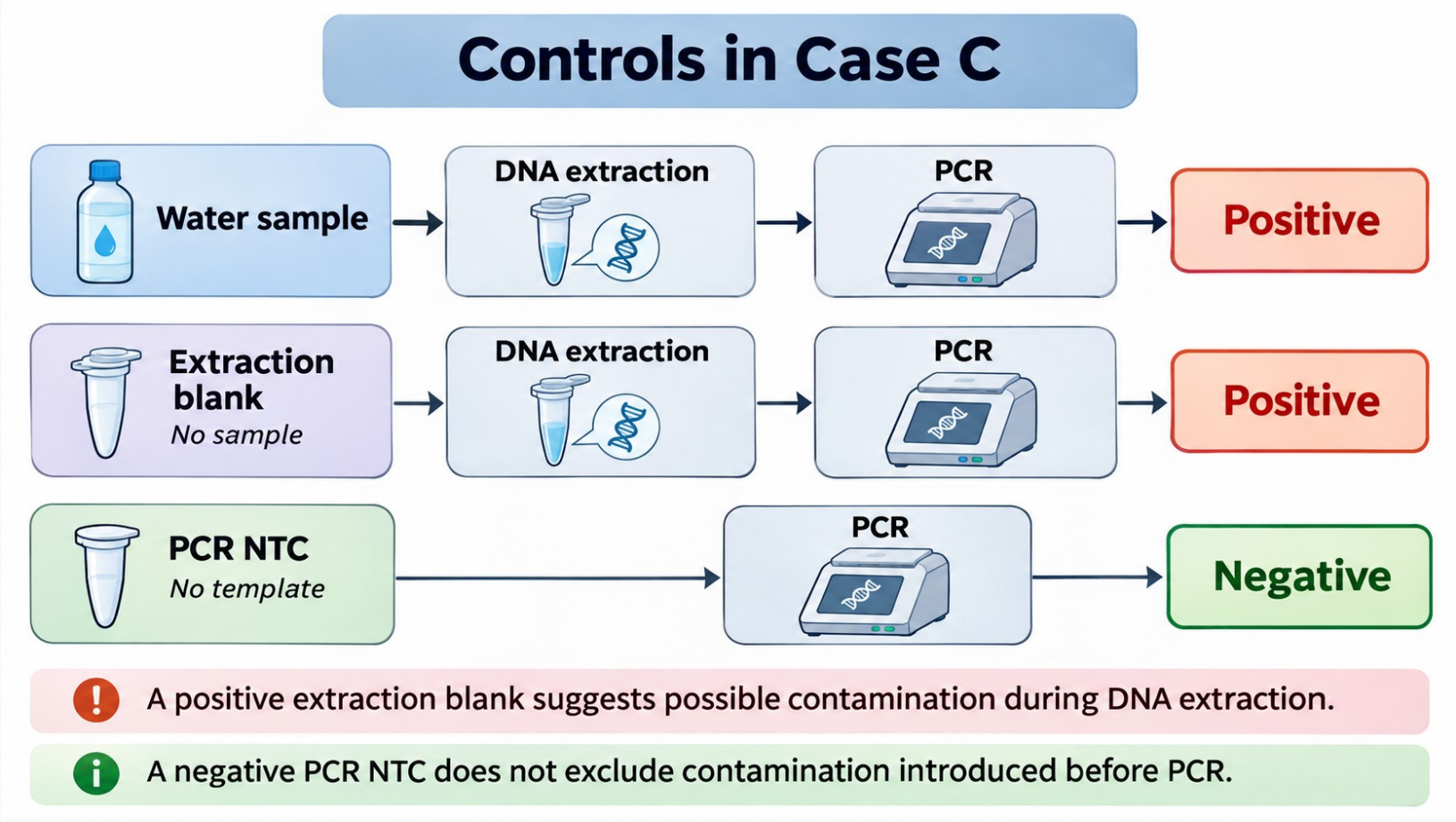
Genus G is provisionally supported. What would separate alpha from beta?

Controls and repeat measurements

Extraction blank: contamination introduced during extraction.

No-template control: contamination in the PCR setup.

Positive control: assay performance under the tested conditions.



Repeat reads do not create independent isolates.

Case C and the origin of the DNA

Sample: clean 800 nt 16S sequence.

Extraction blank: the same 800 nt sequence.

PCR NTC: negative. Positive control: expected result.

Three PCRs from the same extract agree.

The organism’s presence in the water remains unconfirmed.

Reference	Matched positions	Identity
Species alpha	798/800	99,75%
Species beta	796/800	99,5%

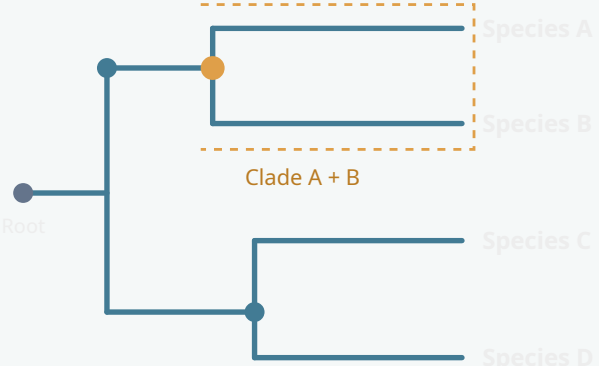
Alignment before the tree

Synthetic fragment · columns compare corresponding positions

Sequence	1	2	3	4	5	6	7	8
U	A	T	G	C	C	A	T	G
A	A	T	G	C	C	A	T	G
B	A	T	G	T	C	A	T	G
C	A	T	G	T	T	A	T	G

Which positions vary? What must an alignment column represent?

Tree anatomy

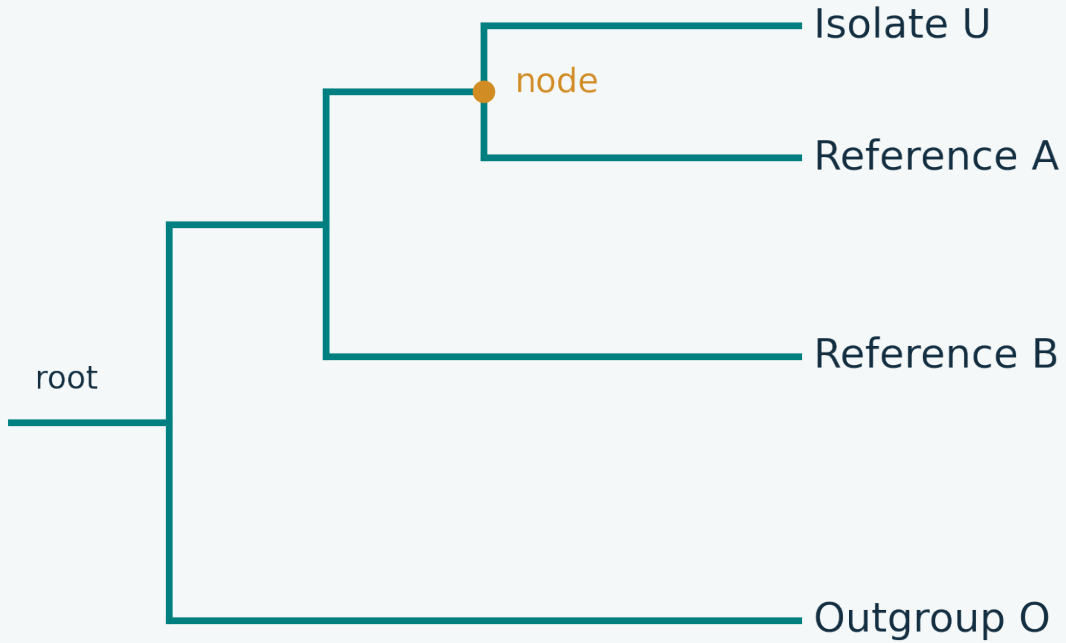


Tips represent the analysed sequences.

Internal nodes represent inferred common ancestors.

A clade contains an ancestor and all its descendants.

Rooting gives direction to the relationships.



Synthetic topology. Branch lengths have no quantitative scale.

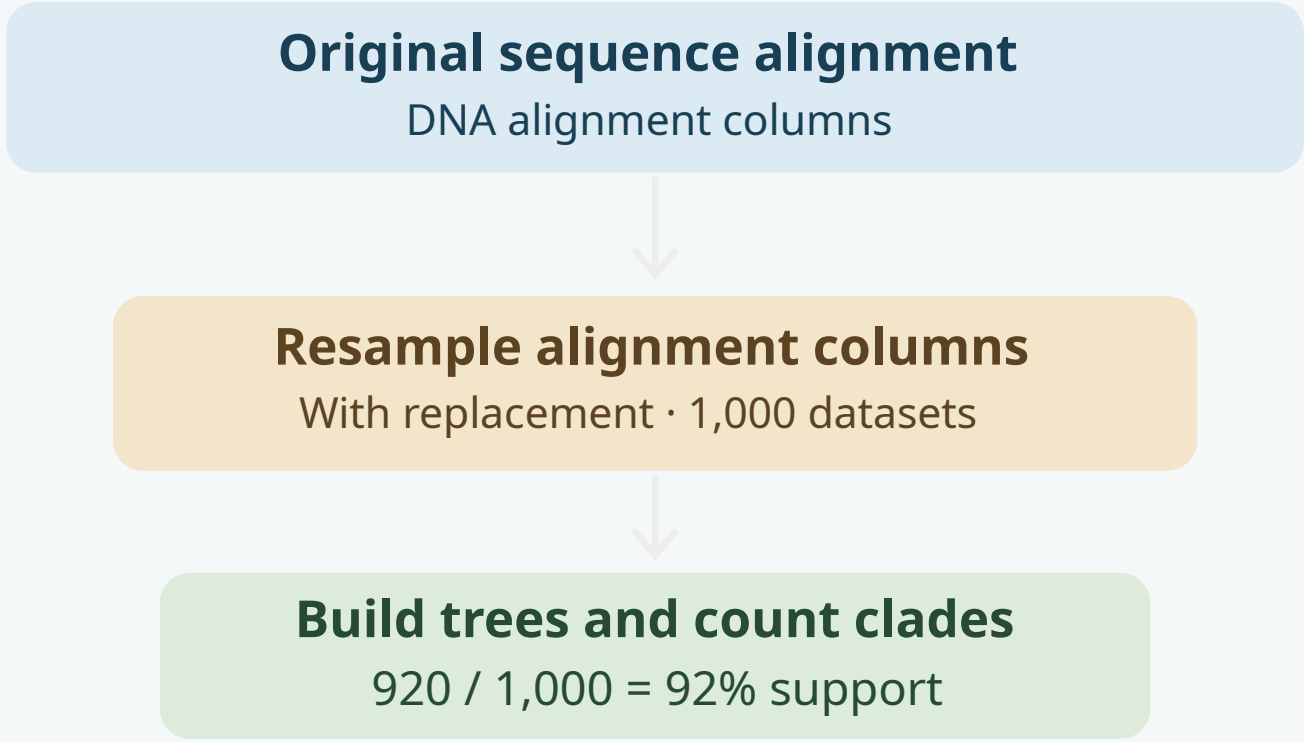
Bootstrap support

Resample alignment columns with replacement.

Infer trees from the resampled datasets.

Count how often a grouping reappears.

Support describes resampling stability, not species certainty.



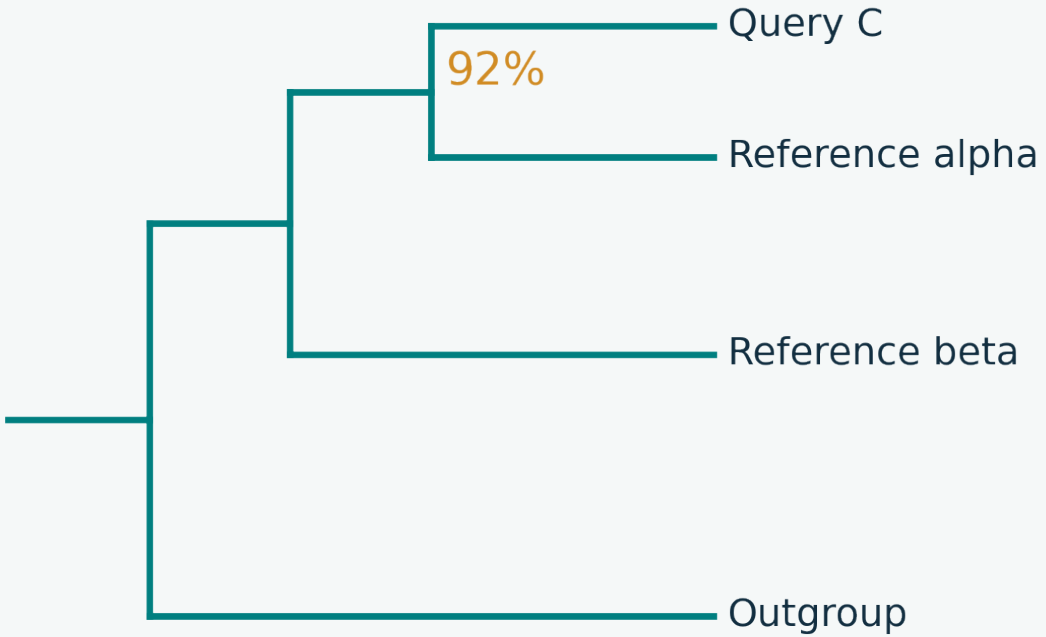
Case C and the 92 percent grouping

Schematic of the grouping reported in Case C.

The query groups with reference alpha at 92%.

Support concerns this marker and analysis.

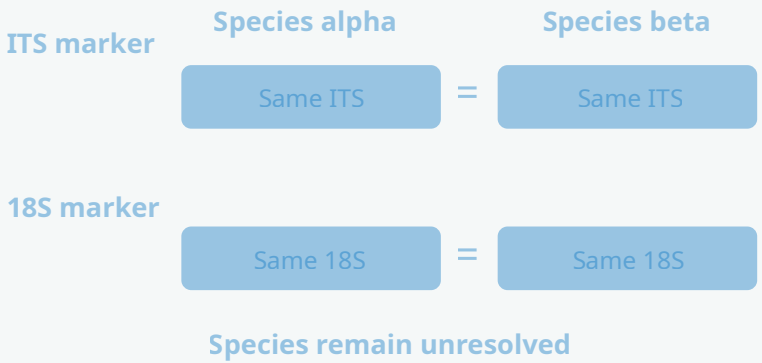
The extraction blank still contains the same sequence.



Teaching schematic. Branch lengths have no scale.

Case B and a useful next test

ITS: alpha 600/600, beta 600/600, gamma 594/600.
18S: all three candidates match 900/900.
Provisional genus F assignment; alpha/beta unresolved.



Choose evidence that can distinguish the candidates.

Reference	Matched / Aligned	Identity
Genus F species alpha	600/600	100%
Genus F species beta	600/600	100%
Genus F species gamma	594/600	99%

An identification report

Question and sample provenance

Marker, sequence length and quality checks

Search details and competing references

Tree method and support, if used

Conclusion, limitation and confirmation step

Exit ticket

Why can 100% identity leave a species unresolved?

What does 92% bootstrap support describe?

Which next step addresses your case’s uncertainty?

Sources and further reading

NCBI
https://www.ncbi.nlm.nih.gov/books/NBK62051/

Johnson 2019
https://pmc.ncbi.nlm.nih.gov/articles/PMC6834636/

Schoch 2012
https://pmc.ncbi.nlm.nih.gov/articles/PMC3341068/

Salter 2014
https://pmc.ncbi.nlm.nih.gov/articles/PMC4228153/

EMBL-EBI
https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics/