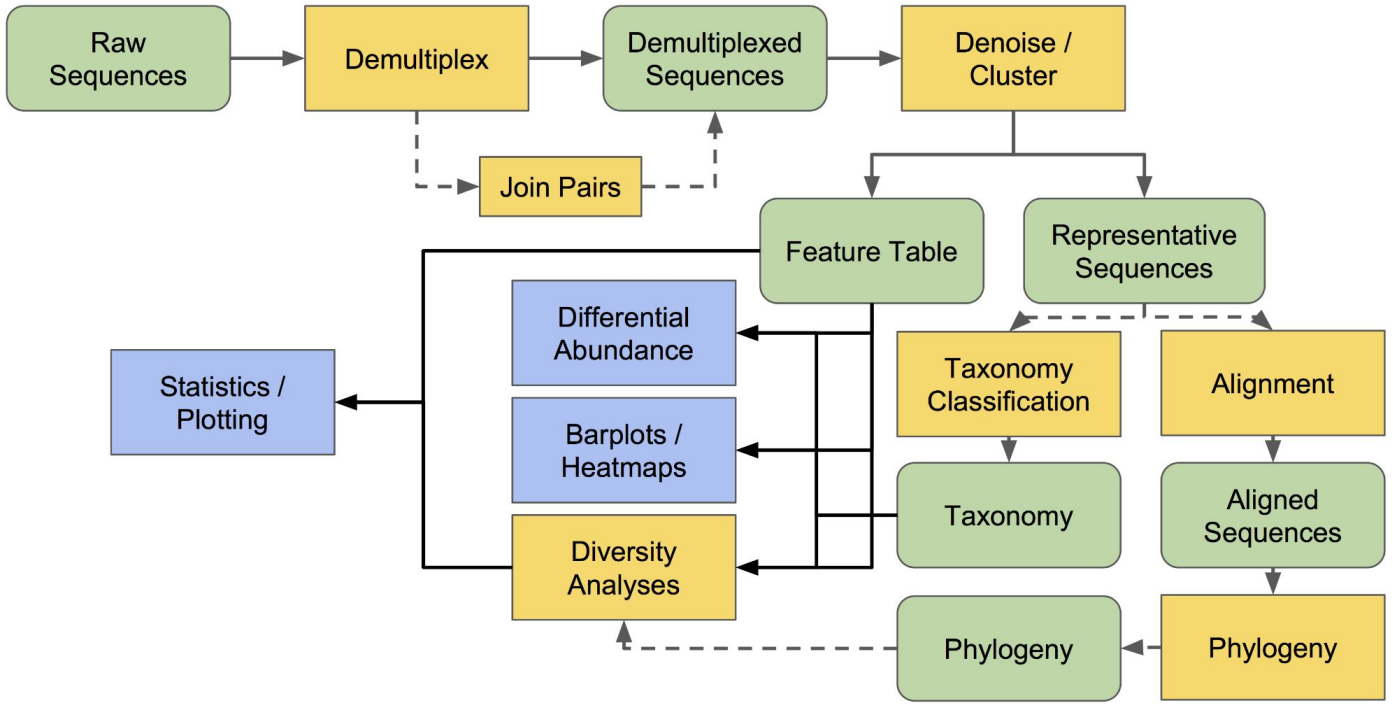
Taxonomy Assignment

Presenter: Hoang Son

18/7/2024

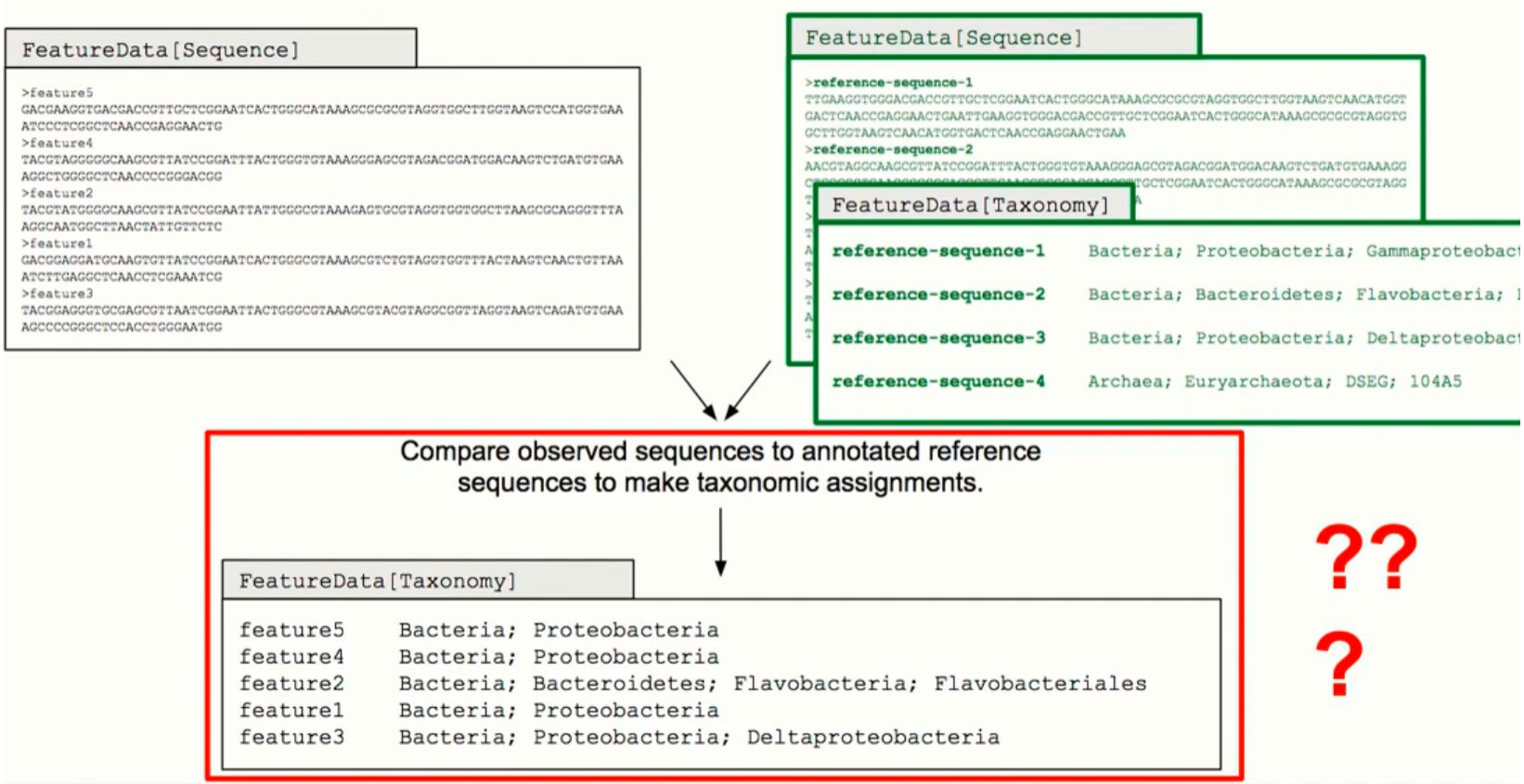
Review the previous steps



https://docs.qiime2.org/2020.8/tutorials/overview/#conceptual-overview-of-qiime-2

What is a taxonomic assignment?

- Taxonomic assignment is the process of classifying unknown DNA sequences into OTUs/ASVs.
- Sequences are compared against a reference database of complete genomes.
- Based on the degree of similarity, sequences are assigned to the most closely matching OTU/ASV in the database.


FeatureData[Sequence]
>feature5
GACGAAGGTGACGACCGTTGCTCCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
FeatureData[Sequence]
>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACATGGT
GACTCAACCGAGGAACTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCAACATGGTGACTCAACCGAGGAACTGAA
>reference-sequence-2
AACGTAGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAGG
FeatureData[Taxonomy]
reference-sequence-1 Bacteria; Proteobacteria; Gammaproteobact
reference-sequence-2 Bacteria; Bacteroidetes; Flavobacteria;
reference-sequence-3 Bacteria; Proteobacteria; Deltaproteobact
reference-sequence-4 Archaea; Euryarchaeota; DSEG; 104A5
Compare observed sequences to annotated reference sequences to make taxonomic assignments.
FeatureData[Taxonomy]
feature5 Bacteria; Proteobacteria
feature4 Bacteria; Proteobacteria
feature2 Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales
feature1 Bacteria; Proteobacteria
feature3 Bacteria; Proteobacteria; Deltaproteobacteria
??
?

Strategies for taxonomic assignment

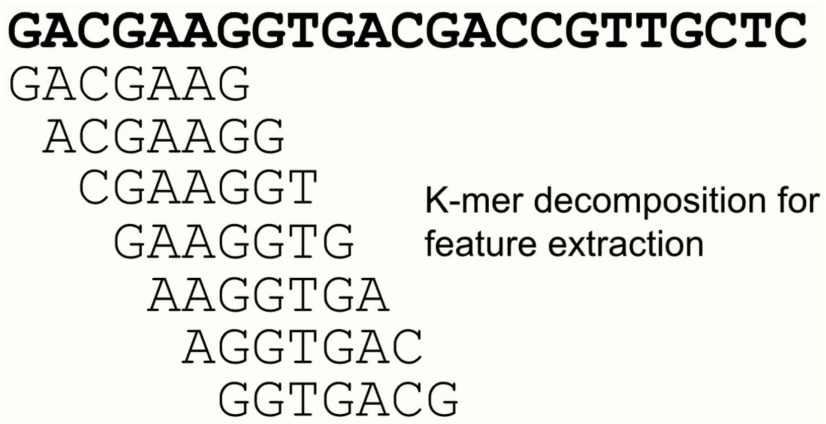
1. **Alignment-Based Methods:** Directly compare unknown sequences to a reference database of known sequences.

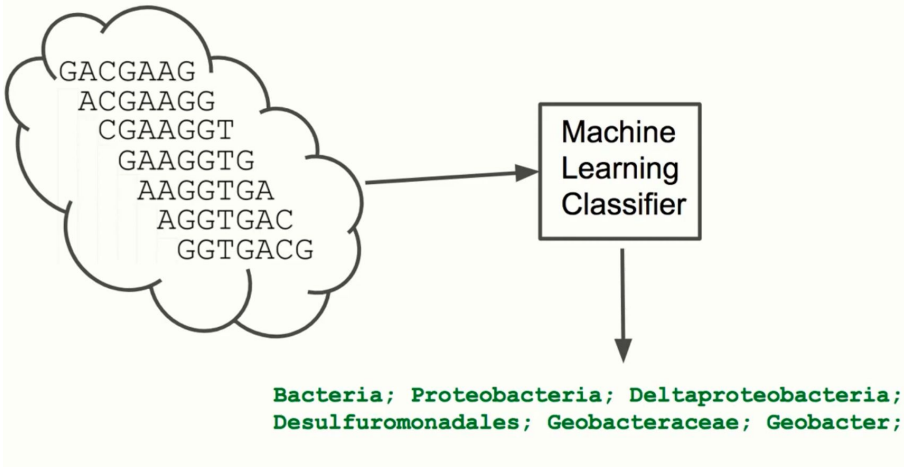
- They find the N most similar sequences (top hits) and assign a consensus taxonomy based on the most frequent classification among those top hits.
- Benefit: No pre-training needed, they leverage the reference database for classification.

```
Query
      GACGAAG
ACTGAGACGAAGAATGTGCTGAT
ATGTGTGCTGTGATGTCTGTGTA
TGTGTATGGCTGTGATGCTGATA
```

Strategies for taxonomic assignment

2. **Machine Learning-Based Methods:** Train a machine learning model on a pre-classified dataset to predict the taxonomy of new sequences.
- Benefit: This method offers speed and efficiency, particularly for large datasets.





QIIME 2 and Taxonomic Assignment

Within q2-feature-classifier, two **alignment-based methods** are available for taxonomic assignment:

Consensus approach:

- classify-consensus-blast
- classify-consensus-vsearch

Your sequence	Your ref. database	BLAST top hits	consensus
Kingdom	`Bacteria`	`Bacteria`	✅
Phyla	`Proteobacteria`	`Proteobacteria`	✅
Class	`Gammaproteobacteria`	`Gammaproteobacteria`	✅
Order	`Legionellales`	`Legionellales`	✅
Genus	`Legionellaceae`	`Legionellaceae`	✅
Species	`Legionella`	---	❌

QIIME 2 and Taxonomic Assignment

Within q2-feature-classifier, two **alignment-based methods** are available for taxonomic assignment:

- **classify-consensus-blast**: BLAST+ local alignment

- Focuses on finding regions of high similarity between your query sequence and the reference sequences.
- Selects the first N hits that exceed a certain percent identity threshold (perc-identity) to your query sequence.
- This means it just takes the first N that meet the minimum similarity criteria

QIIME 2 and Taxonomic Assignment

Within q2-feature-classifier, two **alignment-based methods** are available for taxonomic assignment:

- **classify-consensus-vsearch**: VSEARCH global alignment
- Unlike classify-consensus-blast, this method searches the entire reference database before choosing the top N hits, not the first N hits.

QIIME 2 and Taxonomic Assignment

Inputs (required):

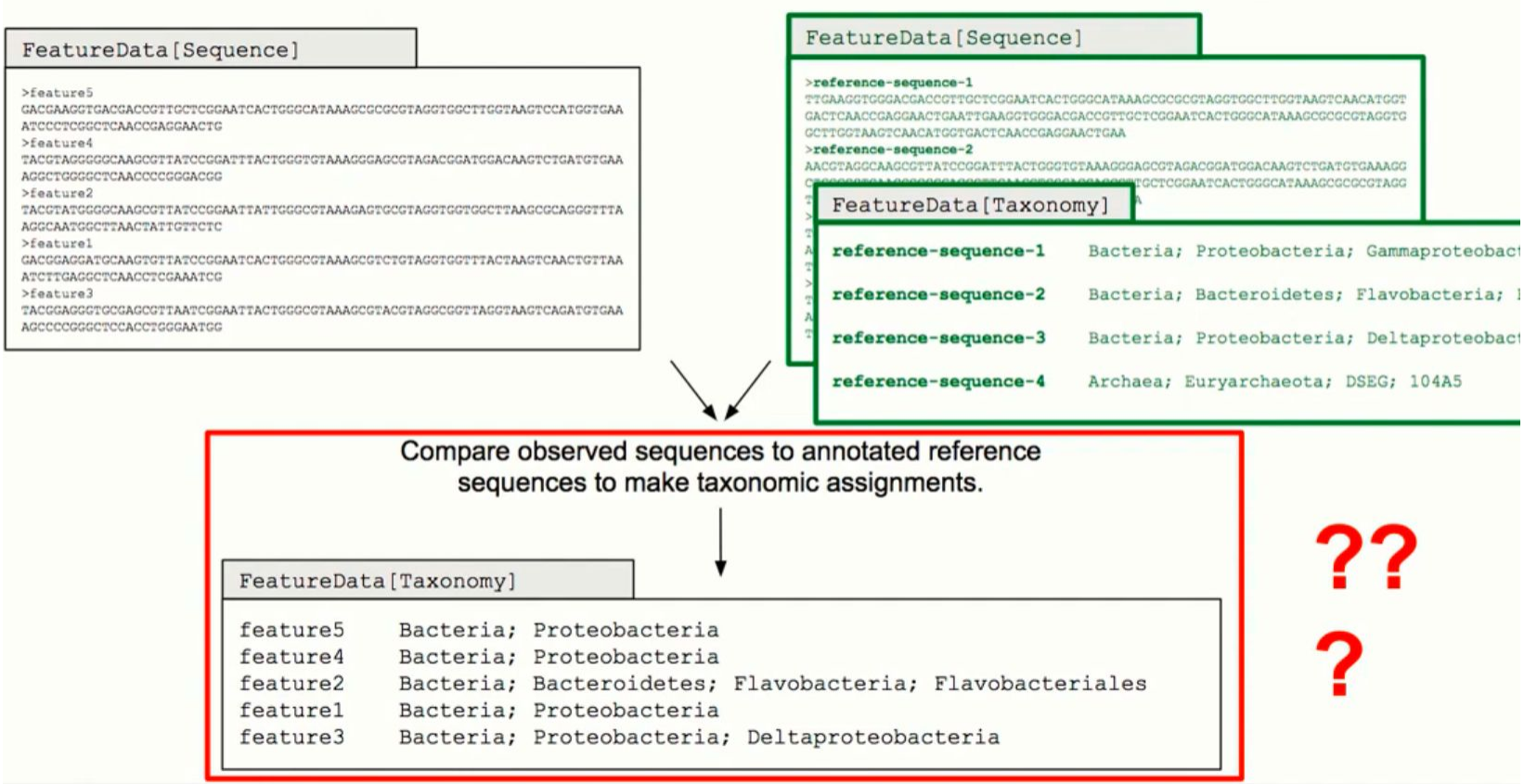
--i-query: Your sequences in FeatureData[Sequence] format.

--i-reference-reads: The reference database containing known sequences in FeatureData[Sequence] format.

--i-reference-taxonomy: The taxonomic labels corresponding to the reference sequences in FeatureData[Taxonomy] format.

Outputs:

--o-classification: FeatureData[Taxonomy] Taxonomy classifications of query sequences.


FeatureData[Sequence]
>feature5
GACGAAGGTGACGACCGTTGCTCCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
FeatureData[Sequence]
>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACATGGT
GACTCAACCGAGGAACTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCAACATGGTGACTCAACCGAGGAACTGAA
>reference-sequence-2
AACGTAGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAGG
FeatureData[Taxonomy]
reference-sequence-1 Bacteria; Proteobacteria; Gammaproteobact
reference-sequence-2 Bacteria; Bacteroidetes; Flavobacteria;
reference-sequence-3 Bacteria; Proteobacteria; Deltaproteobact
reference-sequence-4 Archaea; Euryarchaeota; DSEG; 104A5
Compare observed sequences to annotated reference sequences to make taxonomic assignments.
FeatureData[Taxonomy]
feature5 Bacteria; Proteobacteria
feature4 Bacteria; Proteobacteria
feature2 Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales
feature1 Bacteria; Proteobacteria
feature3 Bacteria; Proteobacteria; Deltaproteobacteria
??
?

QIIME 2 and Taxonomic Assignment

Key Parameters:

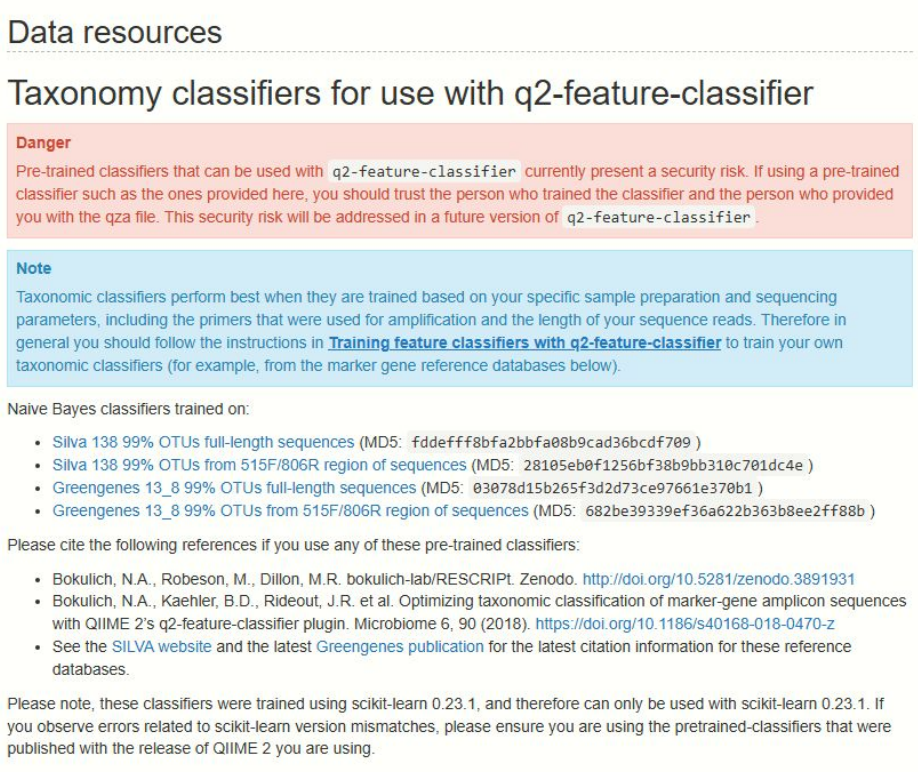
--p-maxaccepts (default: 10): controls how many similar sequences from the reference database are used for assigning taxonomy.

--p-perc-identity (default: 0.8): sets the minimum percentage match required between your sequence and a reference sequence for it to be considered a hit. (0.0-1.0)

--p-min-consensus (default: 0.51): sets the minimum proportion of top hits that need to agree on the same classification for it to be assigned to your sequence. (0.5-1.0)

QIIME 2 and Taxonomic Assignment

QIIME 2's **classify-sklearn** offers a wide array of classification algorithms, allowing for even more customization of the taxonomic assignment process.



Data resources

Taxonomy classifiers for use with q2-feature-classifier

Danger

Pre-trained classifiers that can be used with `q2-feature-classifier` currently present a security risk. If using a pre-trained classifier such as the ones provided here, you should trust the person who trained the classifier and the person who provided you with the qza file. This security risk will be addressed in a future version of `q2-feature-classifier`.

Note

Taxonomic classifiers perform best when they are trained based on your specific sample preparation and sequencing parameters, including the primers that were used for amplification and the length of your sequence reads. Therefore in general you should follow the instructions in **Training feature classifiers with q2-feature-classifier** to train your own taxonomic classifiers (for example, from the marker gene reference databases below).

Naive Bayes classifiers trained on:

- Silva 138 99% OTUs full-length sequences (MD5: `fddefff8bfa2bbfa08b9cad36bcdf709`)
- Silva 138 99% OTUs from 515F/806R region of sequences (MD5: `28105eb0f1256bf38b9bb310c701dc4e`)
- Greengenes 13_8 99% OTUs full-length sequences (MD5: `03078d15b265f3d2d73ce97661e370b1`)
- Greengenes 13_8 99% OTUs from 515F/806R region of sequences (MD5: `682be39339ef36a622b363b8ee2ff88b`)

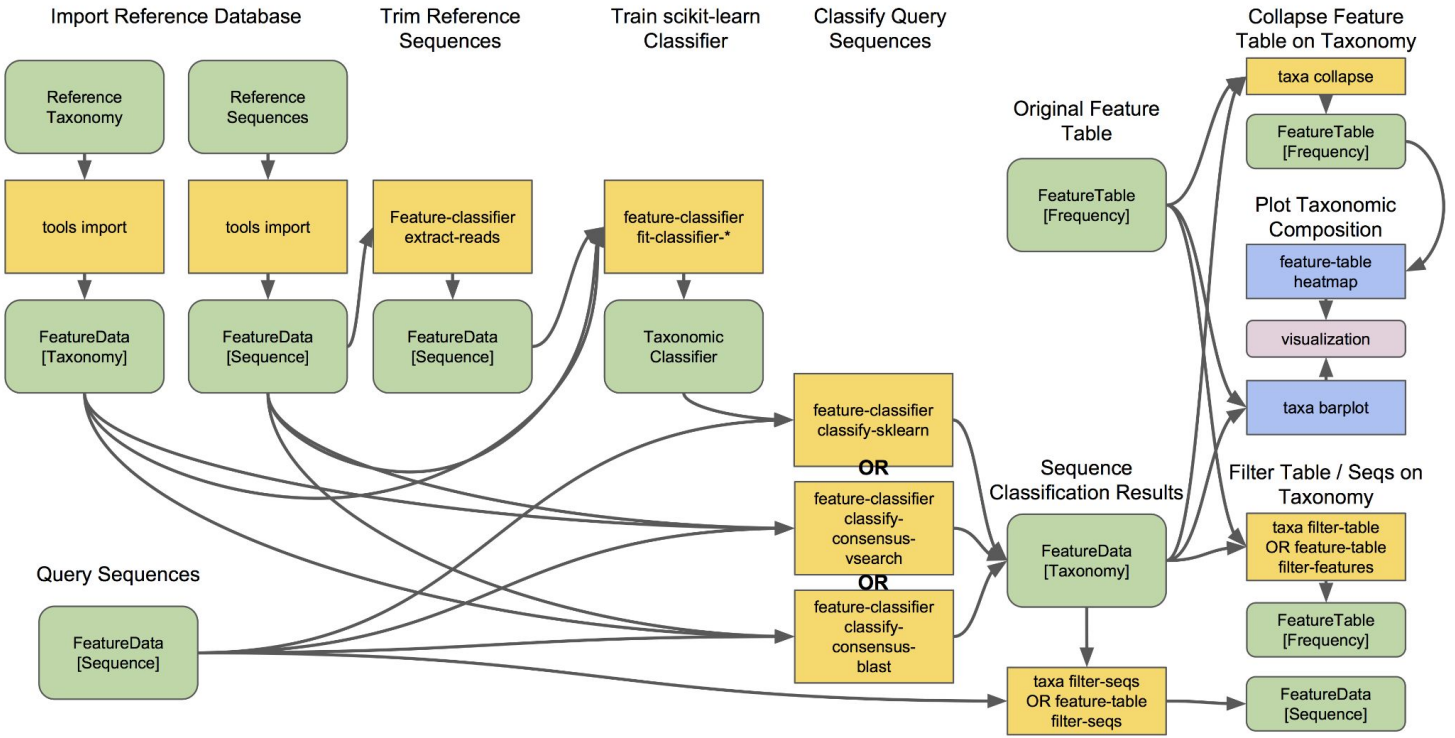
Please cite the following references if you use any of these pre-trained classifiers:

- Bokulich, N.A., Robeson, M., Dillon, M.R. bokulich-lab/RESCRIPt. Zenodo. http://doi.org/10.5281/zenodo.3891931
- Bokulich, N.A., Kaehler, B.D., Rideout, J.R. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6, 90 (2018). https://doi.org/10.1186/s40168-018-0470-z
- See the SILVA website and the latest Greengenes publication for the latest citation information for these reference databases.

Please note, these classifiers were trained using scikit-learn 0.23.1, and therefore can only be used with scikit-learn 0.23.1. If you observe errors related to scikit-learn version mismatches, please ensure you are using the pretrained-classifiers that were published with the release of QIIME 2 you are using.

https://docs.qiime2.org/2020.8/data-resources/#data-resources

Taxonomy classification workflow



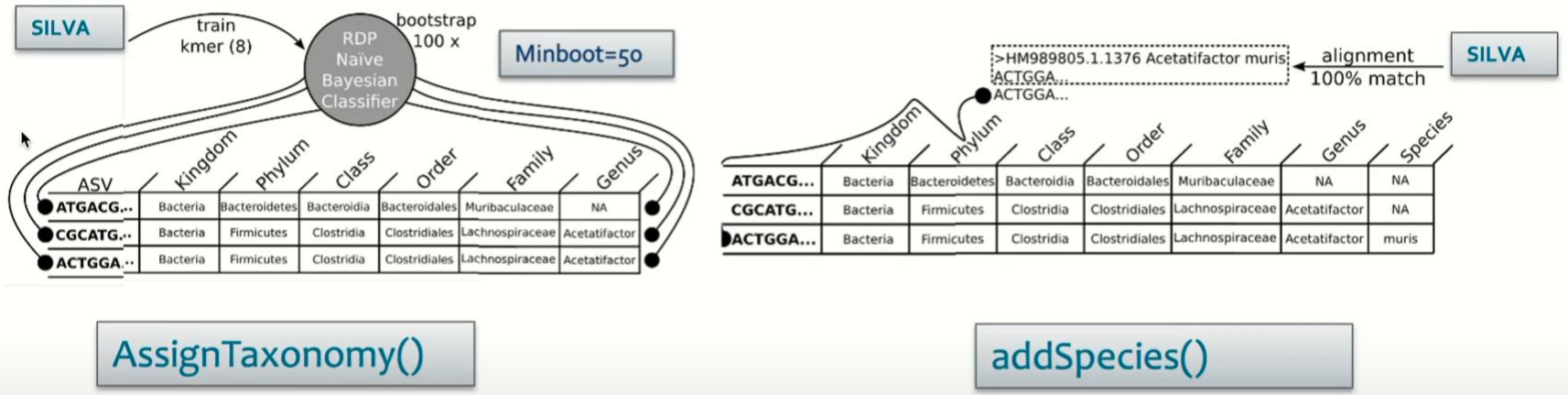
https://docs.qiime2.org/2020.8/tutorials/overview/#taxonomy-classification-and-taxonomic-analyses

Example with DADA2

Input file:

- As input you need to give the ASV table saved as .Rda file
- Uses SILVA reference files by default if no reference files were given
  - To use other reference databases, download the DADA2 supported reference file and bring it to Chipster
  - Find DADA2 supported reference databases: https://benjjneb.github.io/dada2/training.html

Example with DADA2



Images by Antonio Sousa, 16S rRNA gene amplicon - upstream data analysis,
https://igcbioinformatics.github.io/biomeshinycourse/pages/dada2/Biodata.ptCrashCourses.html

Example with DADA2

Output file:

taxonomy-assignment-matrix.Rda

taxa_seqtab_combined.tsv

- If "Combine the taxonomy and the sequence table" is set to yes, both of those tables are combined to taxa_seqtab_combined.tsv
- The names of the ASVs are renamed for visualization purposes. The .Rda object has still the full DNA sequences

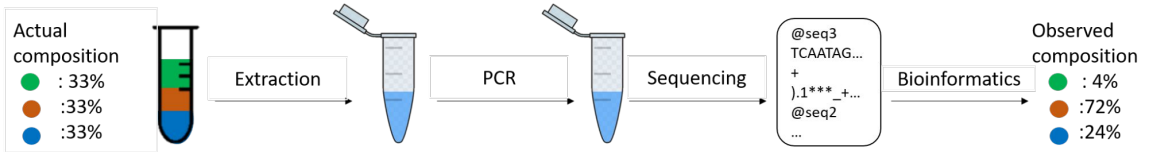
Showing the first 100 of 218 rows. View in full screen to see all rows. Full Screen

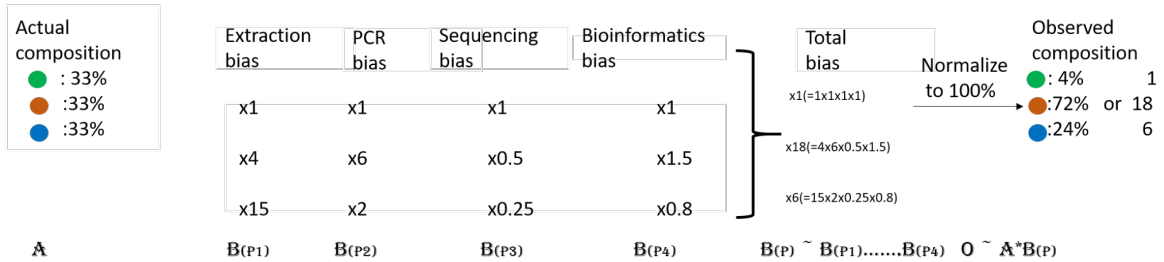
	Kingdom	Phylum	Class	Order	Family	Genus
ASV1	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Muribaculaceae	NA
ASV2	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Muribaculaceae	NA
ASV3	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Muribaculaceae	NA
ASV4	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Muribaculaceae	NA
ASV5	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
ASV6	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Muribaculaceae	NA
ASV7	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Muribaculaceae	NA
ASV8	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
ASV9	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Muribaculaceae	NA
ASV10	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Muribaculaceae	NA

Abundance bias

Abundance bias can arise from various factors throughout the metagenomic workflow:

- DNA Extraction.
- PCR Amplification
- Sequencing Errors
- Reference Database



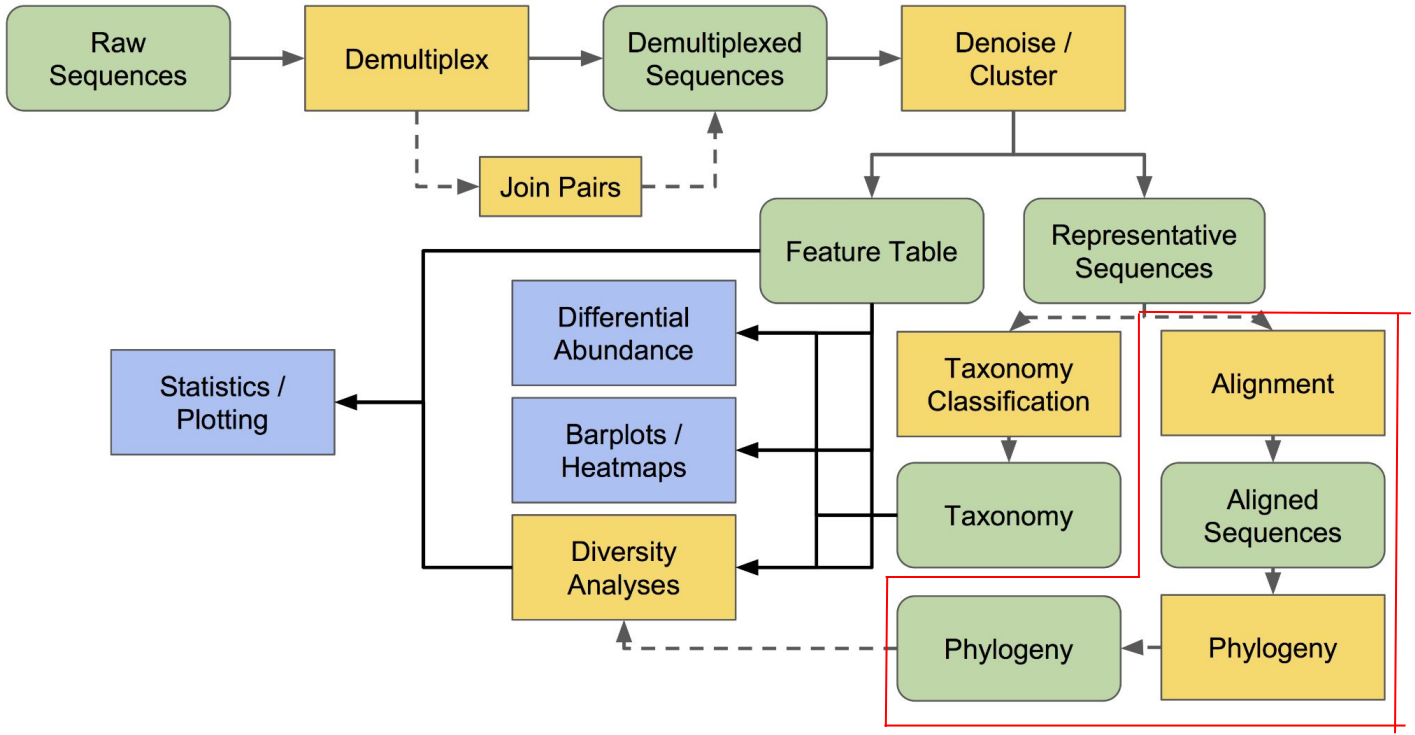


Phylogenetic tree construction

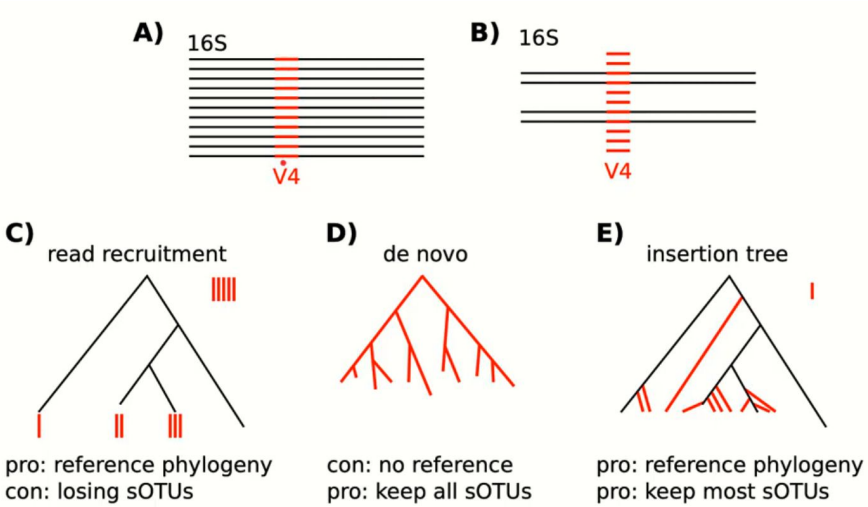
Presenter: Hoang Son

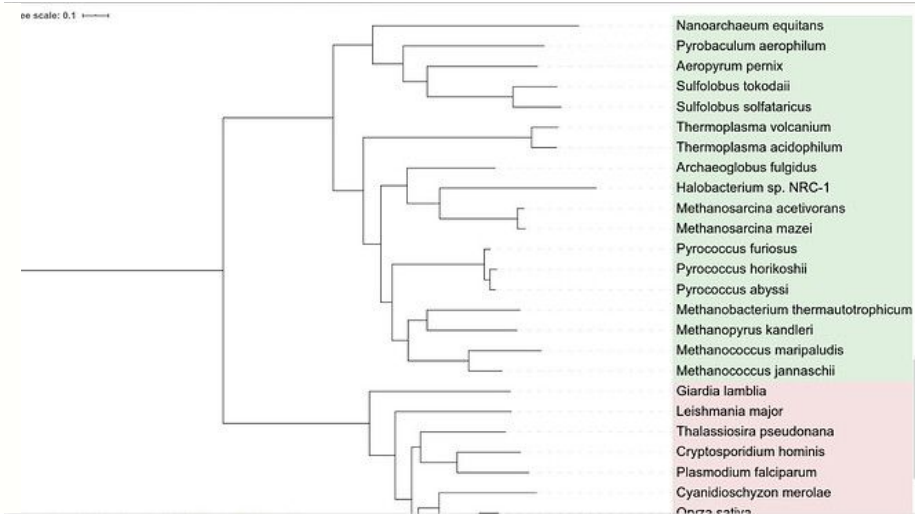
18/7/2024

Review the previous steps



https://docs.qiime2.org/2020.8/tutorials/overview/#conceptual-overview-of-qiime-2


A)
16S
V4
B)
16S
V4
C)
read recruitment
pro: reference phylogeny
con: losing sOTUs
D)
de novo
con: no reference
pro: keep all sOTUs
E)
insertion tree
pro: reference phylogeny
pro: keep most sOTUs


ee scale: 0.1
Nanoarchaeum equitans
Pyrobaculum aerophilum
Aeropyrum pernix
Sulfolobus tokodaii
Sulfolobus solfataricus
Thermoplasma volcanium
Thermoplasma acidophilum
Archaeoglobus fulgidus
Halobacterium sp. NRC-1
Methanosarcina acetivorans
Methanosarcina mazei
Pyrococcus furiosus
Pyrococcus horikoshii
Pyrococcus abyssi
Methanobacterium thermautotrophicum
Methanopyrus kandleri
Methanococcus maripaludis
Methanococcus jannaschii
Giardia lamblia
Leishmania major
Thalassiosira pseudonana
Cryptosporidium hominis
Plasmodium falciparum
Cyanidioschyzon merolae

Feature	Character-Based Methods	Distance-Based Methods
Data Used	Individual characters (nucleotides, amino acids, morphological traits)	Overall genetic or morphological distances between pairs of organisms
Analysis Focus	How characters change (gained, lost, modified) throughout evolution	Pairwise similarity based on distance calculations
Tree Building Goal	Find the tree with the fewest evolutionary changes to explain the data	Build a tree that reflects the pairwise distances
Advantages	More detailed evolutionary information	Faster and computationally simpler
Disadvantages	Can be computationally expensive for large datasets	May lose information by condensing data into distances

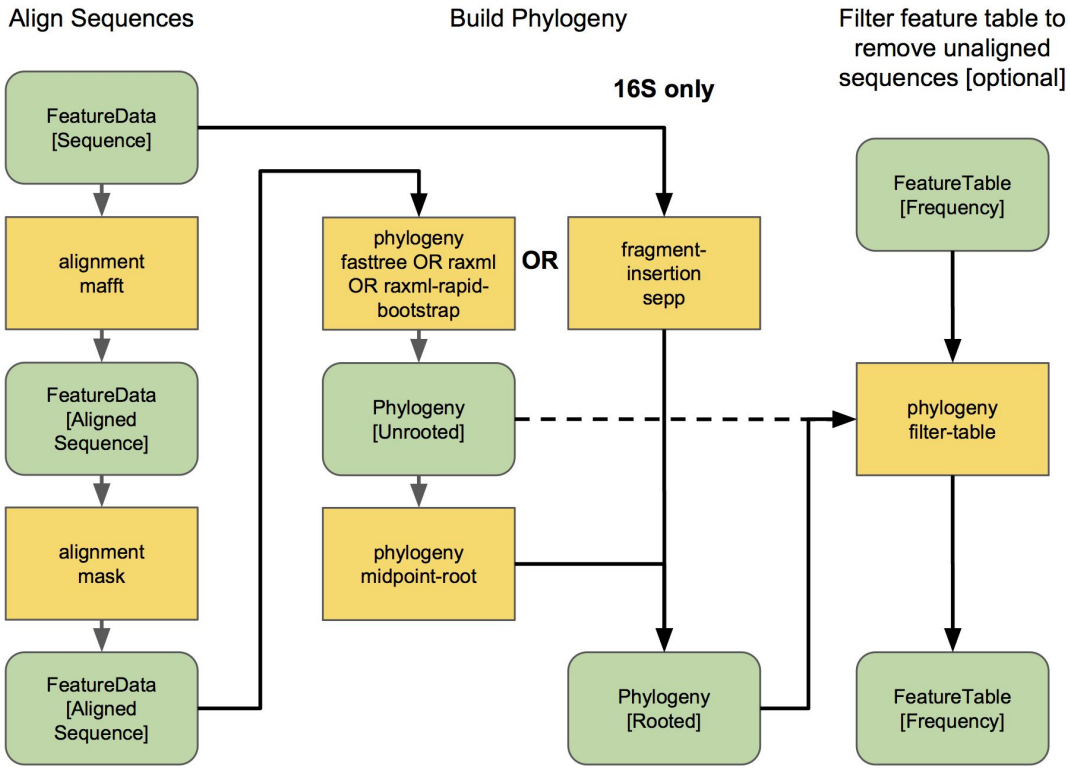
Feature	Character-Based Methods	Distance-Based Methods
Examples of Methods	Parsimony, Maximum Likelihood	Neighbor-Joining, UPGMA
Tools	Molecular Evolutionary Genetics Analysis (MEGA), FastTree2, RAxML, IQTree	MEGA, rapidNJ
Visualization	Molecular Evolutionary Genetics Analysis (MEGA), iTOL, ETE Toolkit TreeViewer, Microreact	

Sequence alignment and phylogeny building

There are two phylogeny -based approaches:

1. A reference-based fragment insertion approach

2. A de novo approach



https://docs.qiime2.org/2024.5/tutorials/overview/#sequence-alignment-and-phylogeny-building

Sequence alignment and phylogeny building

1. Reference-Based Fragment Insertion (Ideal Choice):

A well-established family tree with known relatives (reference phylogeny).

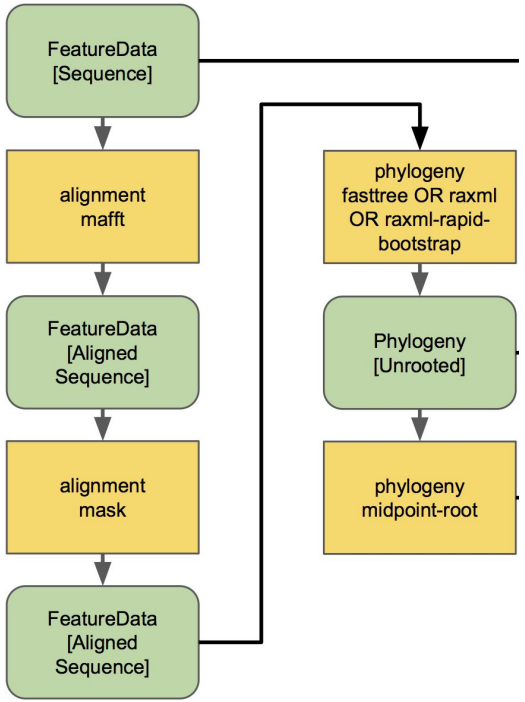
- **Method:** new sequences are inserted into the existing tree at the most appropriate positions based on their similarity to existing branches.
- **Benefits:**
  - Efficient and reliable for closely related sequences.
  - Leverages existing knowledge about evolutionary relationships.
- **Limitations:**
  - Requires a well-curated reference phylogeny that accurately reflects your sequences' ancestry.
  - Sequences too dissimilar to existing branches might not be accurately place

Sequence alignment and phylogeny building

2. De Novo Phylogeny

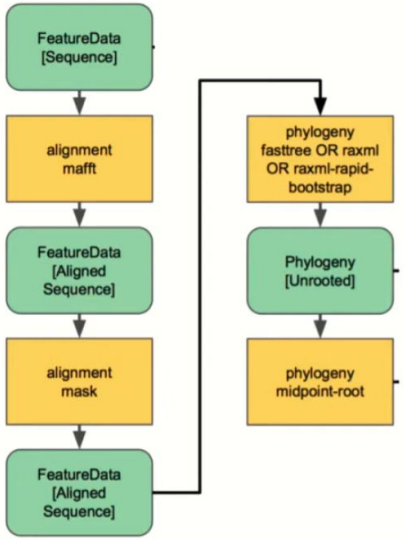
Building a family tree from scratch.

- **Method:** Your sequences are directly aligned and analyzed to infer evolutionary relationships.
- **Benefits:** Applicable to diverse sequences without requiring a pre-existing reference tree.
- **Limitations:**
  - Requires longer sequences for robust results (short reads might lack sufficient information).
  - More computationally intensive than fragment insertion.



Sequence alignment and phylogeny building

Step-by-step commands >>




```
qiime alignment mafft \
  --i-sequences filtered-sequences-2.qza \
  --o-alignment aligned-rep-seqs.qza
```

```
qiime alignment mask \
  --i-alignment aligned-rep-seqs.qza \
  --o-masked-alignment masked-aligned-rep-seqs.qza
```

```
qiime phylogeny fasttree \
  --i-alignment masked-aligned-rep-seqs.qza \
  --o-tree unrooted-tree.qza
```

```
qiime phylogeny midpoint-root \
  --i-tree unrooted-tree.qza \
  --o-rooted-tree rooted-tree.qza
```

```
qiime phylogeny align-to-tree-mafft-fasttree \
    --i-sequences filtered-sequences-2.qza \
    --output-dir mafft-fasttree-output
```

<< Pipeline command

Sequence alignment and phylogeny building

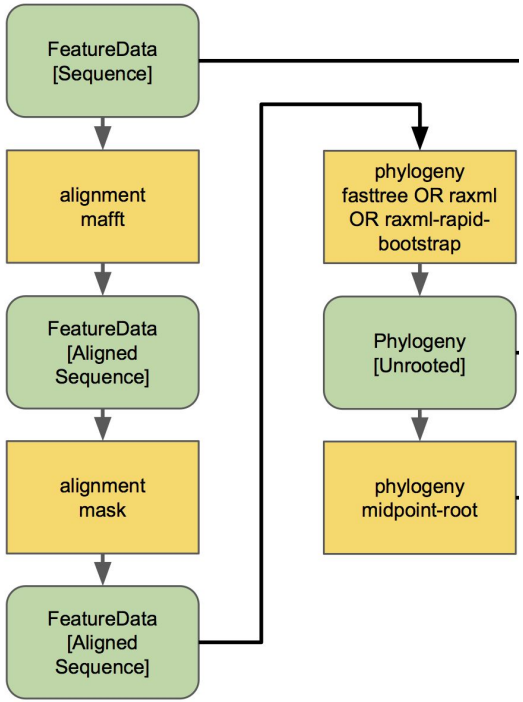
2. De Novo Phylogeny

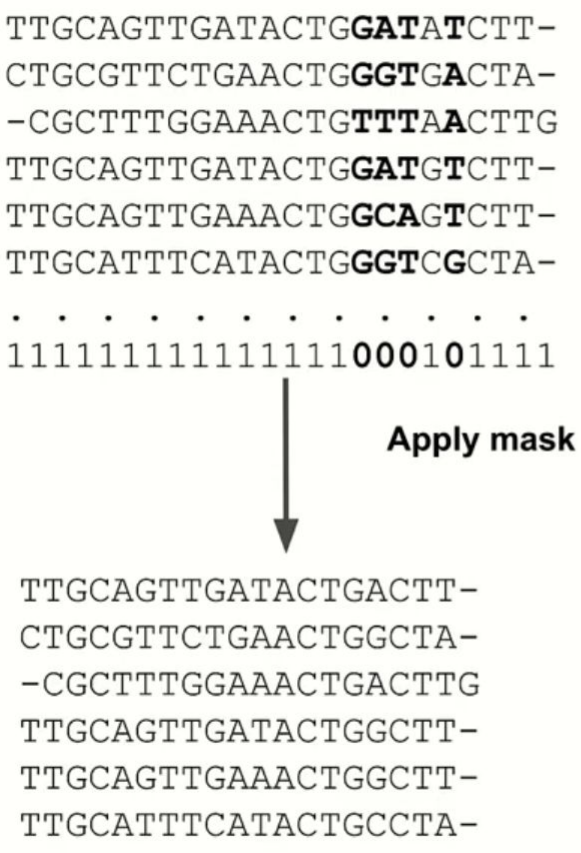
q2-alignment
&
q2-phylogeny

plugin commands

- qiime alignment …
  - mafft …
  - mask …
- qiime phylogeny …
  - fasttree …
  - raxml …
  - raxml-rapid-bootstrap …
  - iqtree …
  - iqtree-ultrafast-bootstrap
  - midpoint-root …
  - align-to-tree-fasttree …
  - align-to-tree-raxml …
  - align-to-tree-iqtree …

pipelines

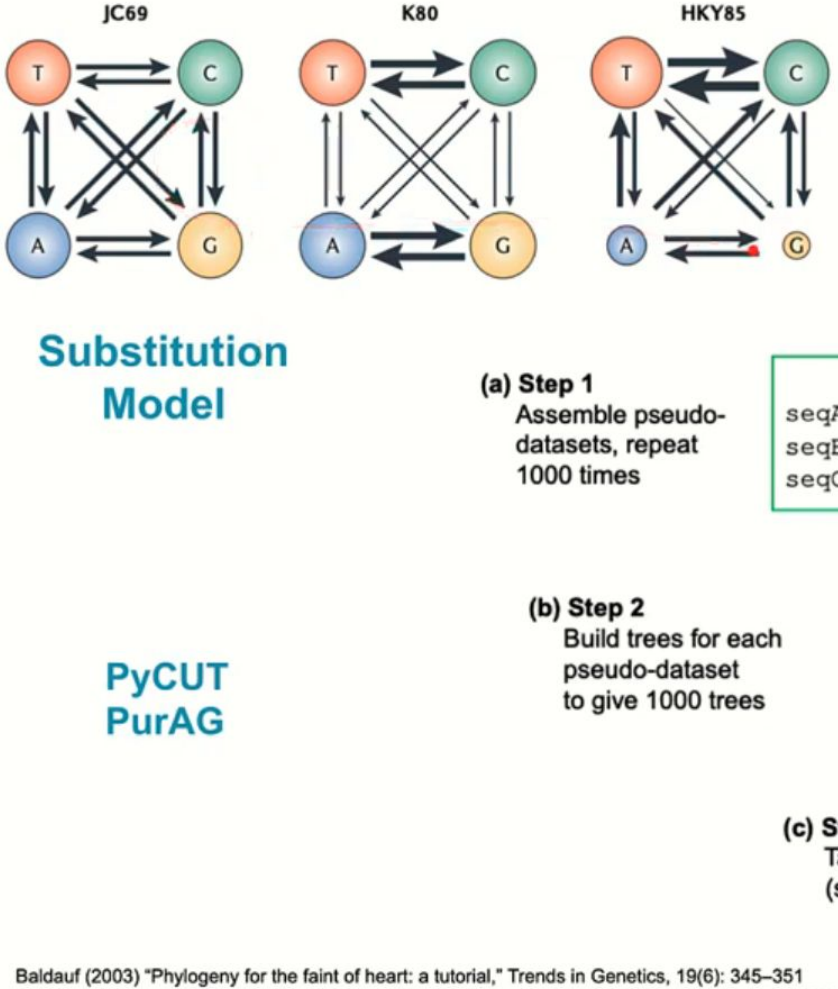




Why “mask” alignment columns?

- Proposed by David Lane (1991)
  - “Lane mask”
- Remove errors introduced by alignment heuristics
  - i.e. incorrect statements of positional homology.
- Eliminate phylogenetically uninformative / misleading sites
- Remove repetitive / homopolymeric regions (TATATATA / AAAAAAAAA)
- These obfuscate phylogenetic inference.

*** Reminder: Sequences can be aligned differently by different programs, and the resulting inferred phylogenies can differ substantially! Often additional algorithmic and/or manual curation is required. Can affect masking!*



Baldauf (2003) "Phylogeny for the faint of heart: a tutorial," Trends in Genetics, 19(6): 345–351
Yang & Rannala (2012) "Molecular phylogenetics: principles and practice," Nat. Rev. Genet., 13(5): 1–12

Thank you