MNHN-Tree-Tools

Thomas Haschka
Journées calcul et données

15/12/2021

MNHN-Tree-Tools

- Extended Applications:
  - Clustering of almost any dataset of nucleic sequences.
  - “Phylogenetic” guidance, Tree of Life building
  - Proven on RNA barcodes 16S/18S ribosome.
  - COVID19 seqeunces and tree

MNHN-Tree-Tools

- Is further a plethora of sequence database management tools
  - i.e. containing codes for:

    Kmerization, PCA, Duplicate Removal, Enzyme digestion, Consensus Calculation, etc…
  - C api for rapid algorithm creation to be applied on a FASTA files.

MNHN-Tree-Tools

- Consists of 100 page Manual
- 1 published article
  - Haschka et al, Bioinformatics 2021 https://doi.org/10.1093/bioinformatics/btab430
- In depth algorithmic supplement
- A website:
  - http://treetools.haschka.net

How it Works: Workflow

- FASTA file containing sequences with reasonable consensus
- Re-express sequences in the FASTA file as kmer frequency vectors

Sequence to k-mer frequencies

- K=2   sequence = ACCCTA

-
  ```
  AA AC AG AT    0 1 0 0
  CC CA CG CT    2 0 0 1
  GG GC GA GT    0 0 0 0
  TT TA TC TG    0 1 0 0
  ```
- Vector has 16 components
- => 0,1,0,0,2,0,0,1,0,0,0,0,0,1,0,0

How it Works: Workflow

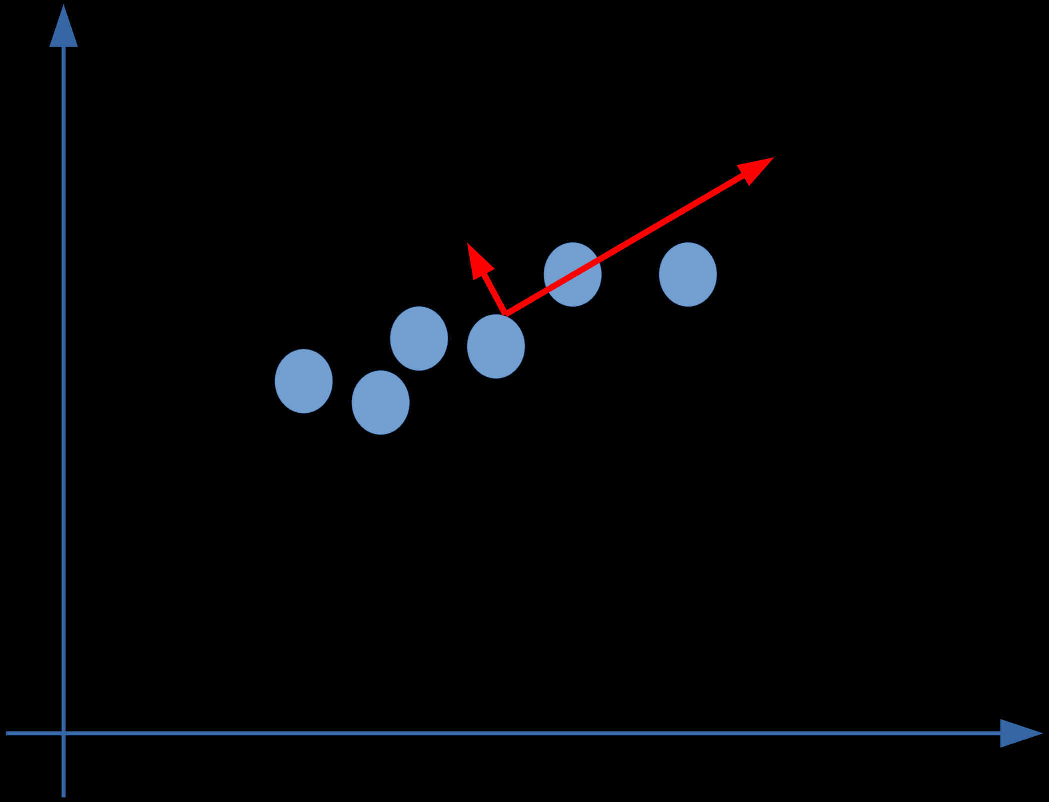
- FASTA file containing sequences with reasonable consensus
- Re-express sequences in the FASTA file as kmer frequency vectors
- Optional Dimensionality Reduction and Feature Selection:
  - Perform Principal Component Analyses PCA
  - Perform Dimensional Reduction using a neural network Autoencoder (tested, not published)

Principal Component Analyses

- The number of k-mers grows with 4^k
- Principal component analyses allows us to

  reduce the dimensionality to the k-mers with the highest variance.
- Feature selection method.

Basic Idea Behind It

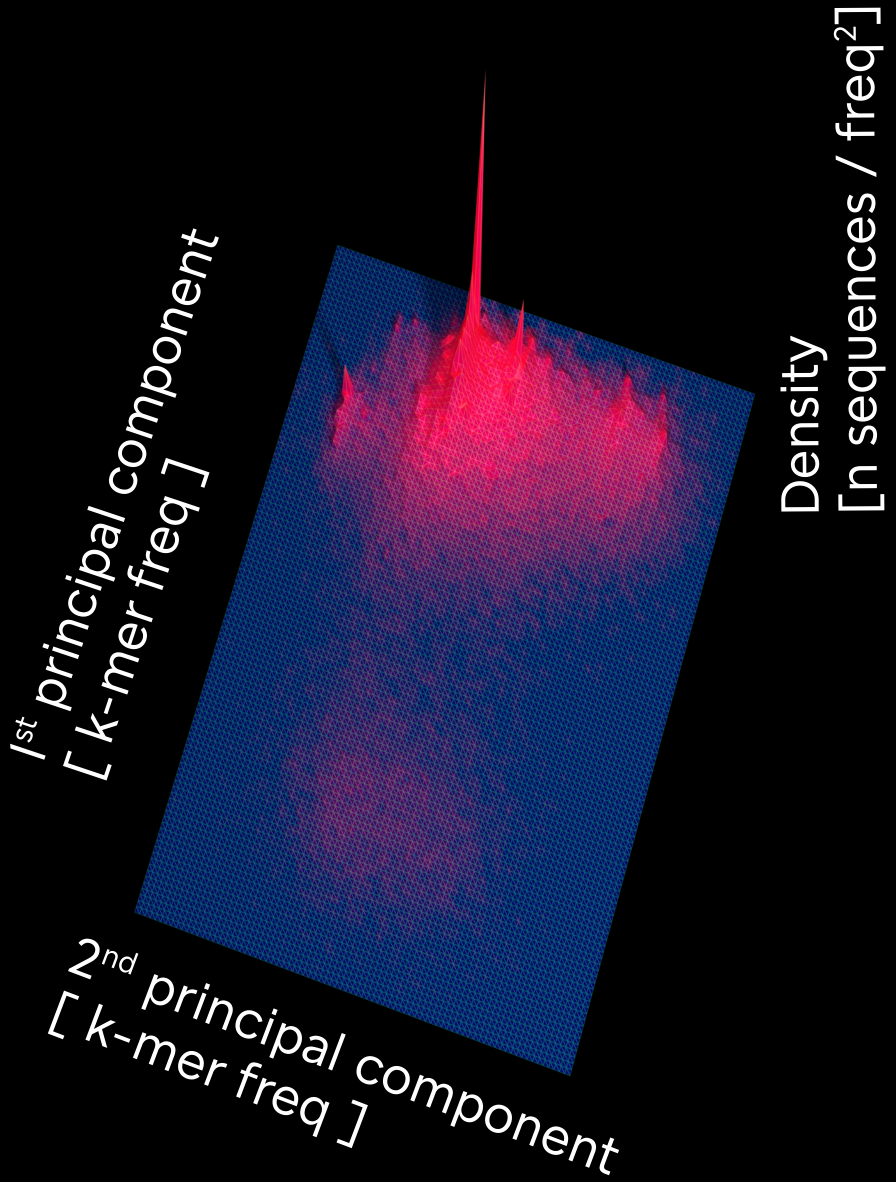


You have linear correlated data:

Transform into the red coordinate system.

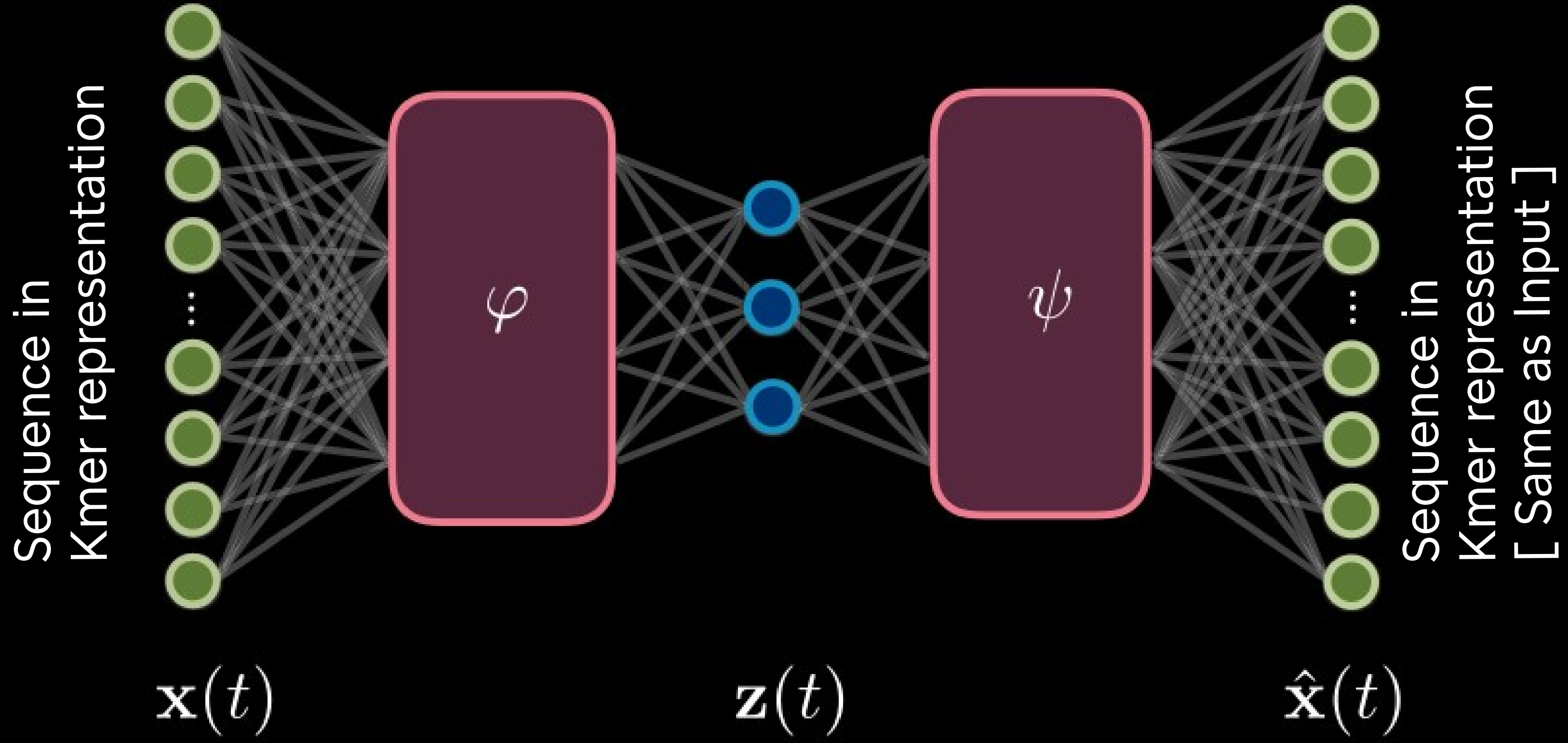
forget about the tiny component and keep only the dimensions with large components.

Direction : eigenvectors
Length: eigenvalues
of the co-variance matrix



Dataset
reduced from
1024 to 2 dimensions

Autoencoder



Sequence in
Kmer representation

Sequence in
Kmer representation
[Same as Input]

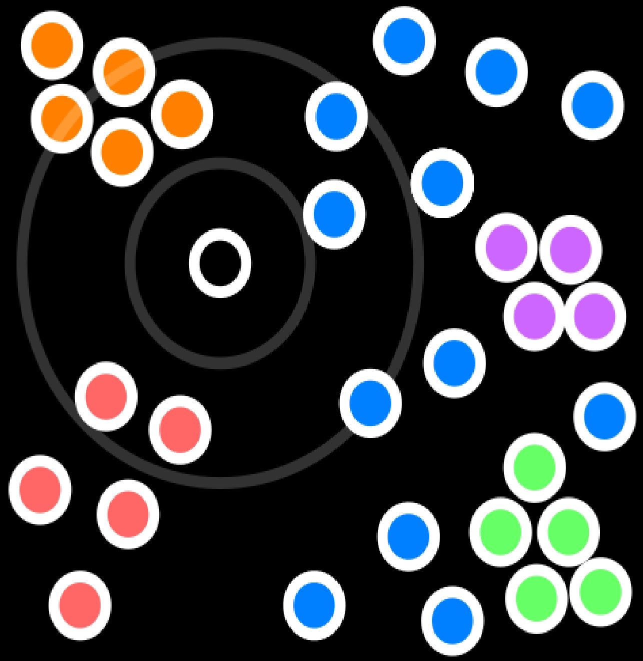
Compressed non-linearly
reduced coordinates that contain
enough information for Kmer reconstruction

How it Works: Workflow

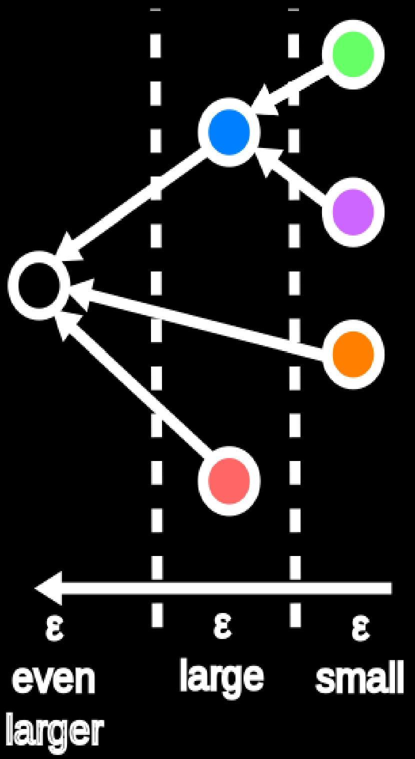
- FASTA file containing sequences
  with reasonable consensus
- Re-express sequences in the FASTA file as kmer frequency vectors
- Optional Dimensionality Reduction and Feature Selection:
  - Perform Principal Component Analyses PCA
  - Perform Dimensional Reduction using
    a neural network Autoencoder (tested, not published)
- Adaptive Clustering and Tree-Building

How do you do this

Sequence space



Phylogenetic tree built from sequence space



DBSCAN Algorithm

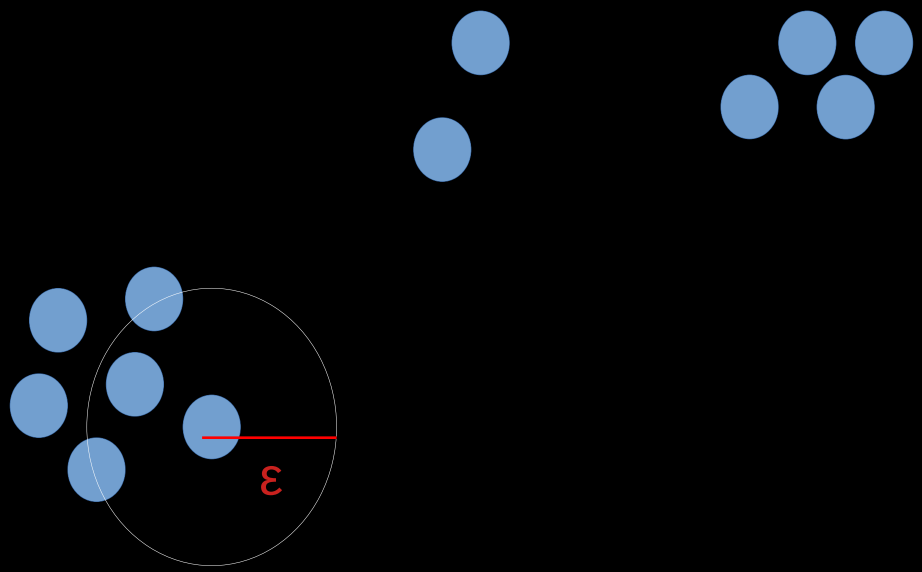
- Published in 1996
- Density Based Algorithm for Discovering Clusters in Large Spatial Datasets with noise.
- Finds members of a density connected region

$$\rho > \rho(\varepsilon, minpts) = minpts \,/\, V(\varepsilon)$$

Number of a sequences within radius ε

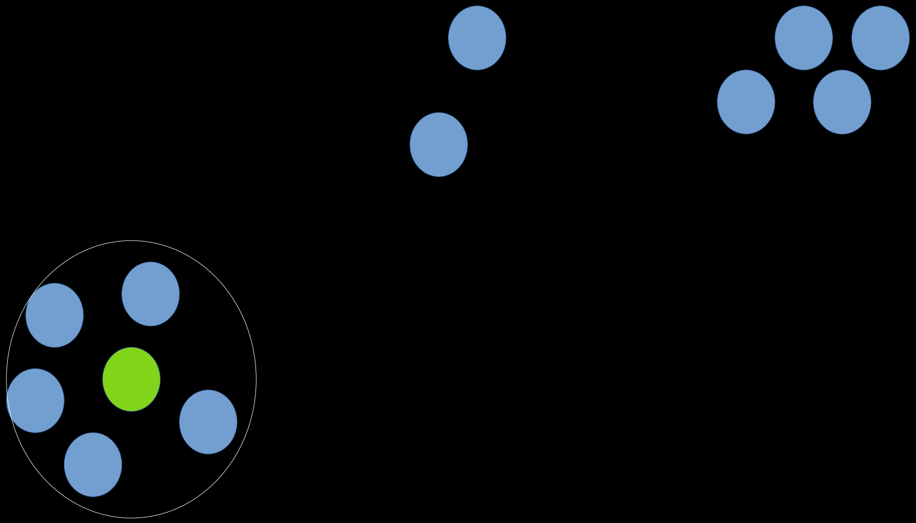
Volume of a ball with radius ε

How it works



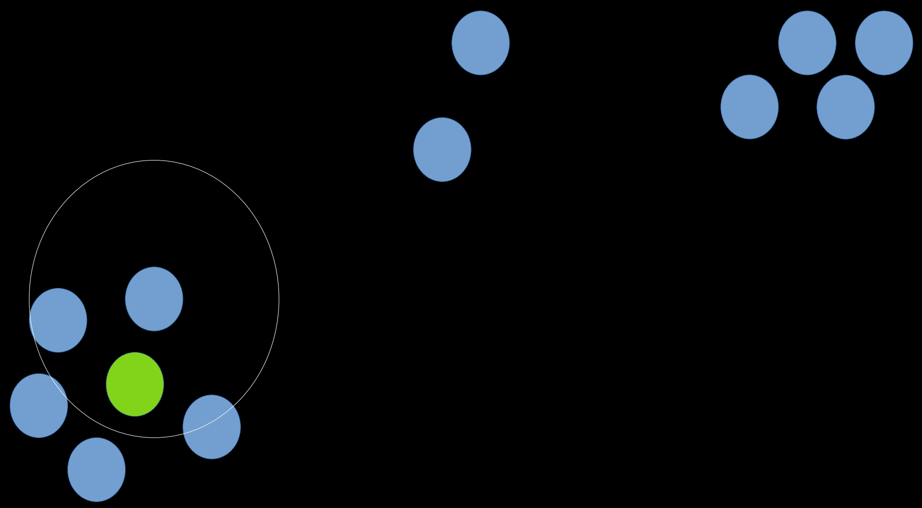
Run with minpts 3

How it works



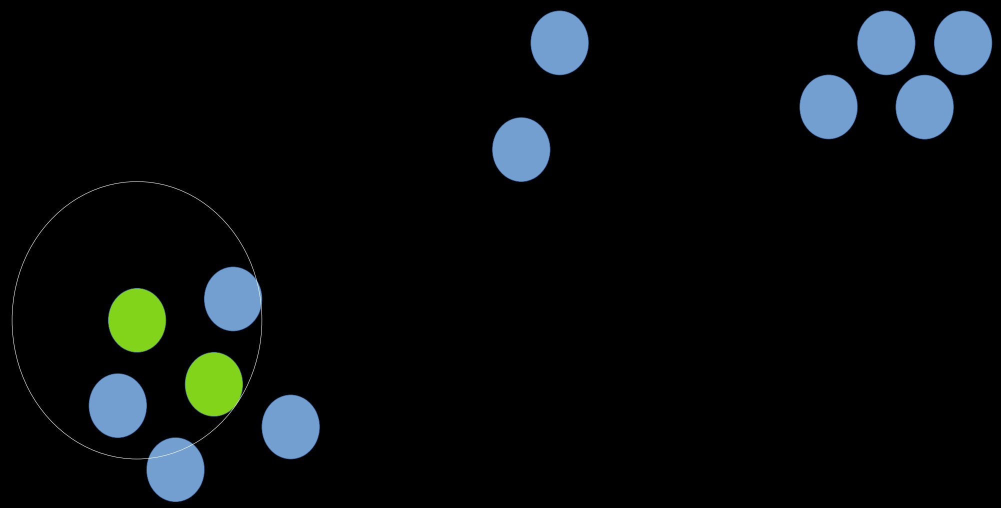
Run with minpts 3

How it works



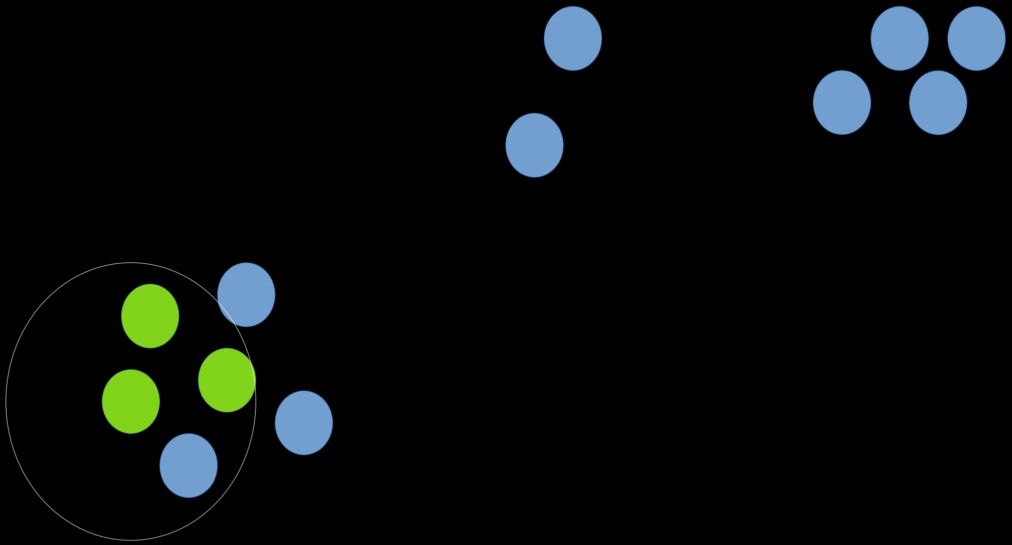
Run with minpts 3

How it works



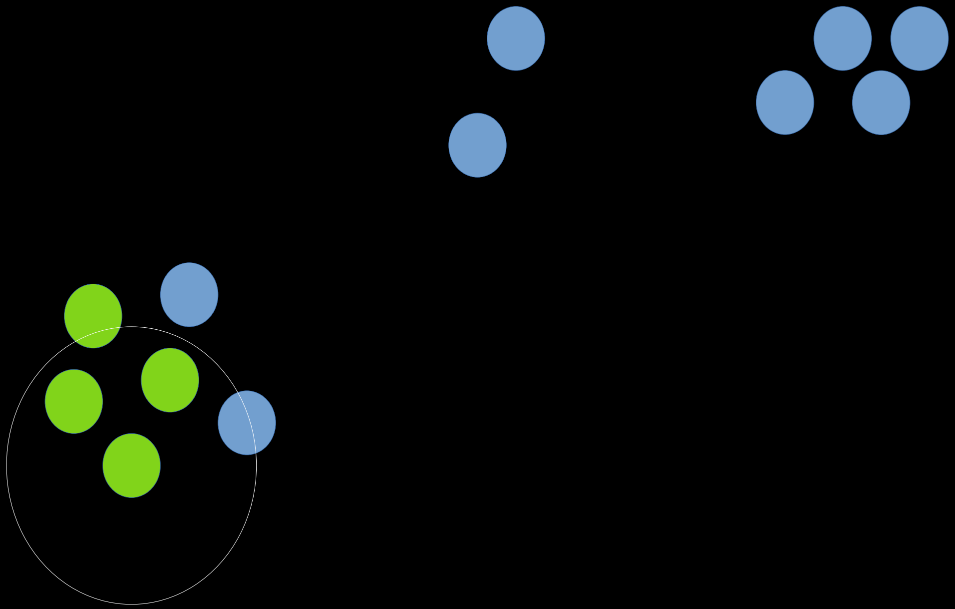
Run with minpts 3

How it works



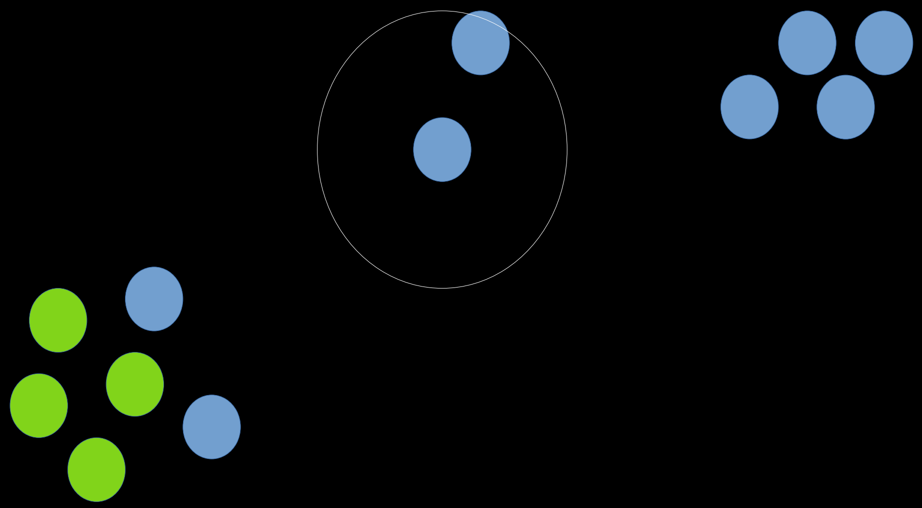
Run with minpts 3

How it works



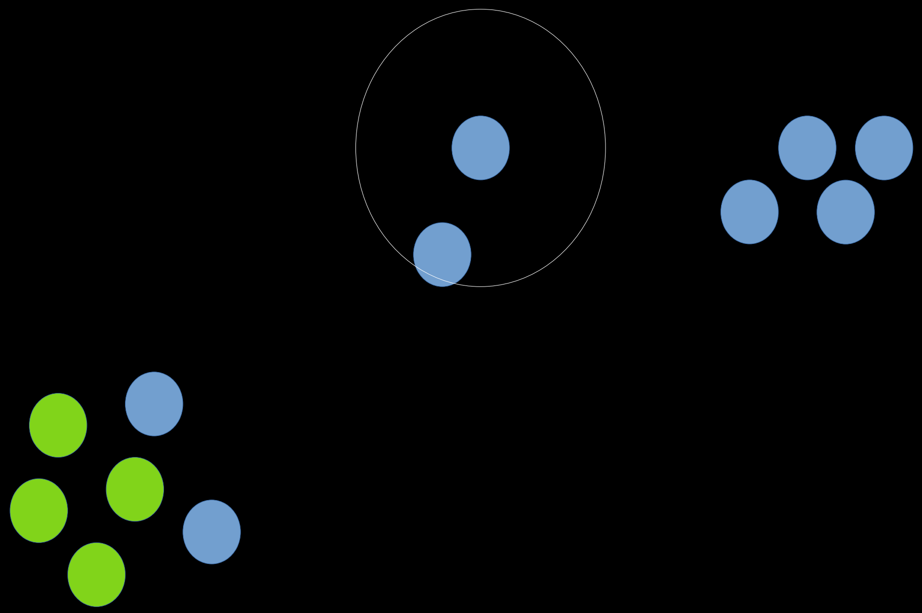
Run with minpts 3

How it works



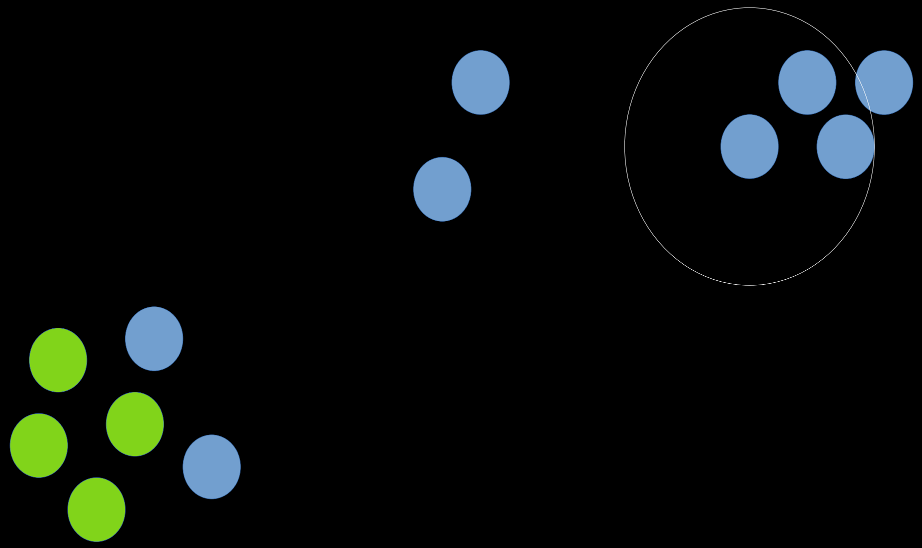
Run with minpts 3

How it works



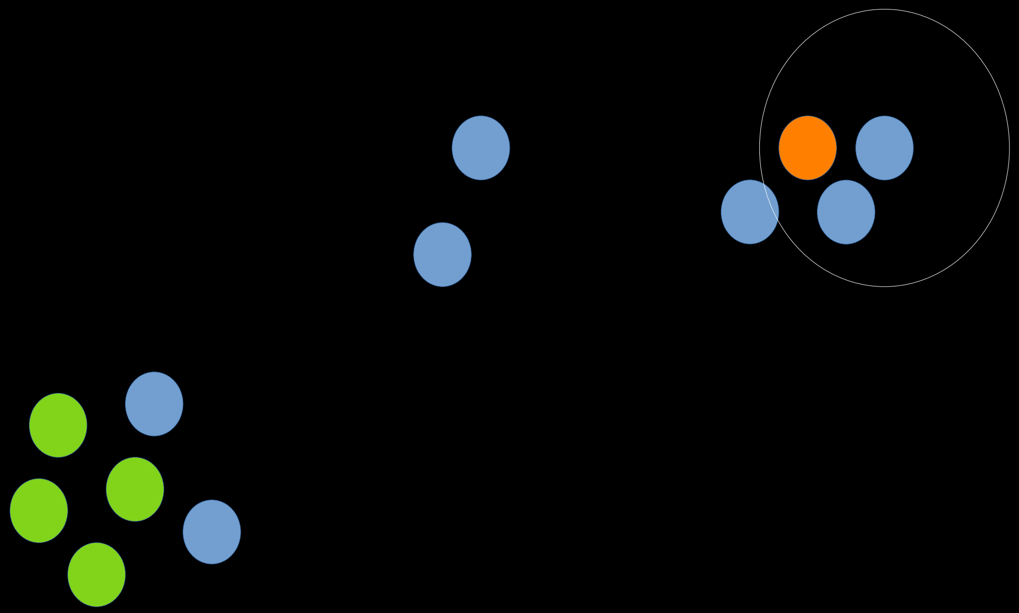
Run with minpts 3

How it works



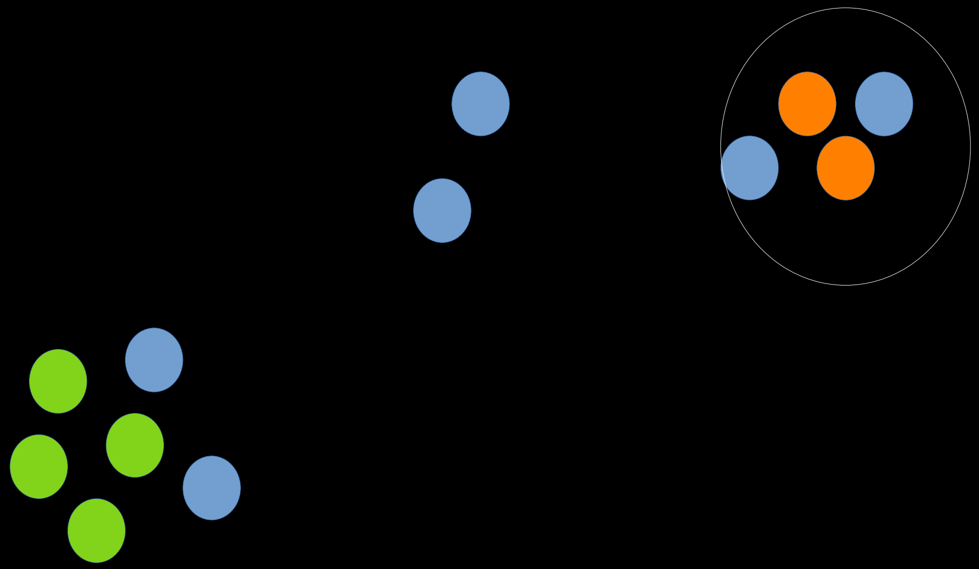
Run with minpts 3

How it works



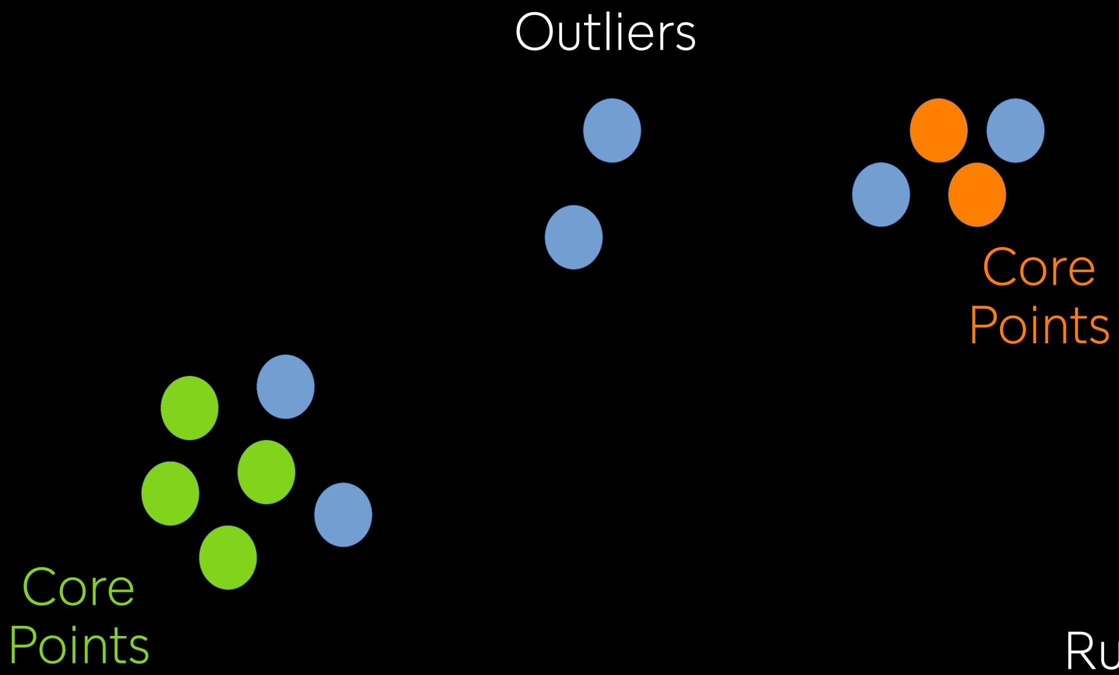
Run with minpts 3

How it works



Run with minpts 3

How it works

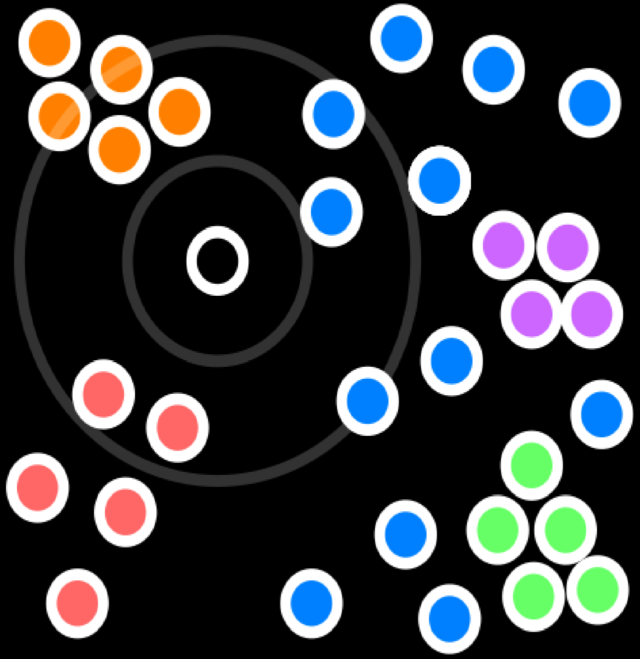


Run with minpts 3

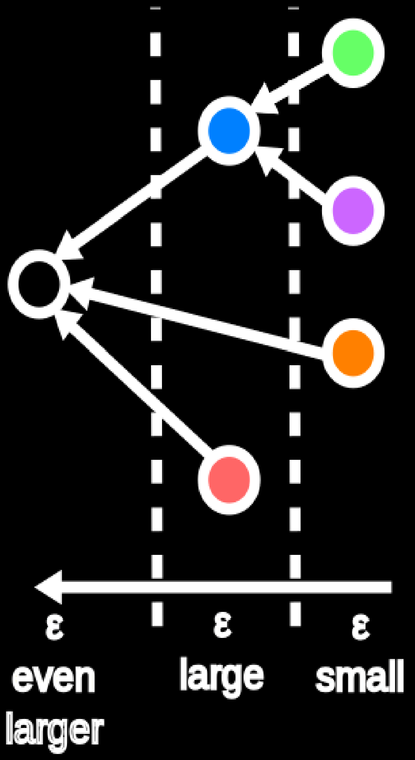
Now imagine what happens if you increase ε

And apply it on this

Sequence space

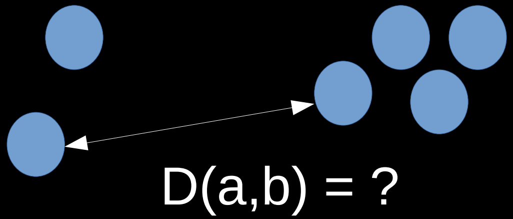


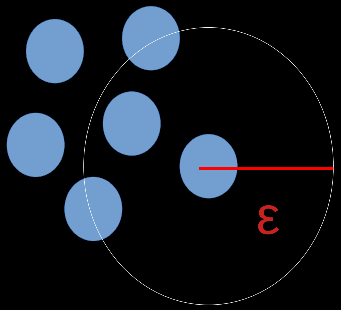
Phylogenetic tree built from sequence space



The importance of distance

- What is a space of sequences
- What is a distance between two sequences

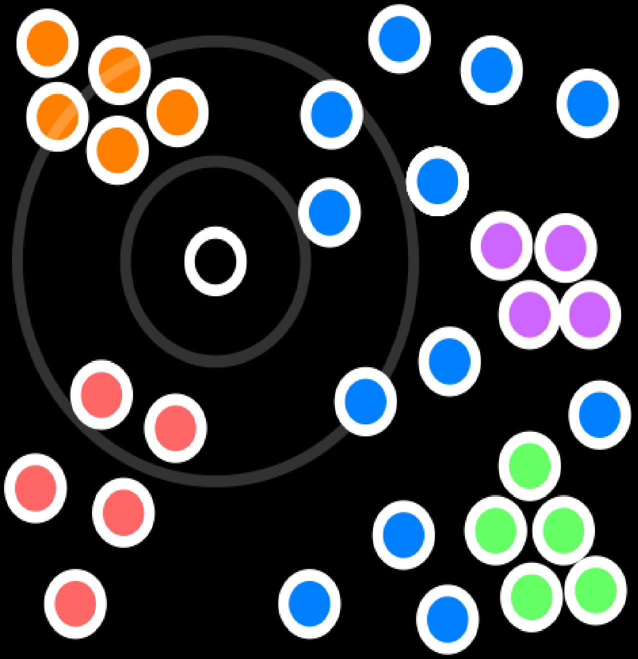




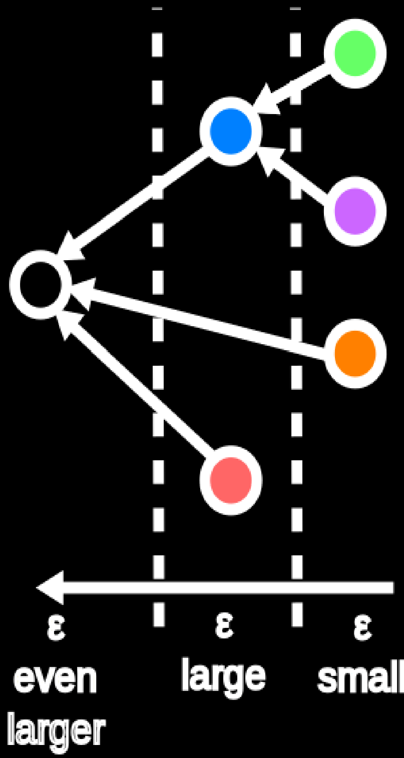
- K-mer direct (L1 / L2)
- K-mer + PCA + L1 / L2
- Autoencoder + L1 / L2
- Smith Waterman

Building a tree by embedding “new” high density sequence clusters in “old” low density sequence clusters

Sequence space

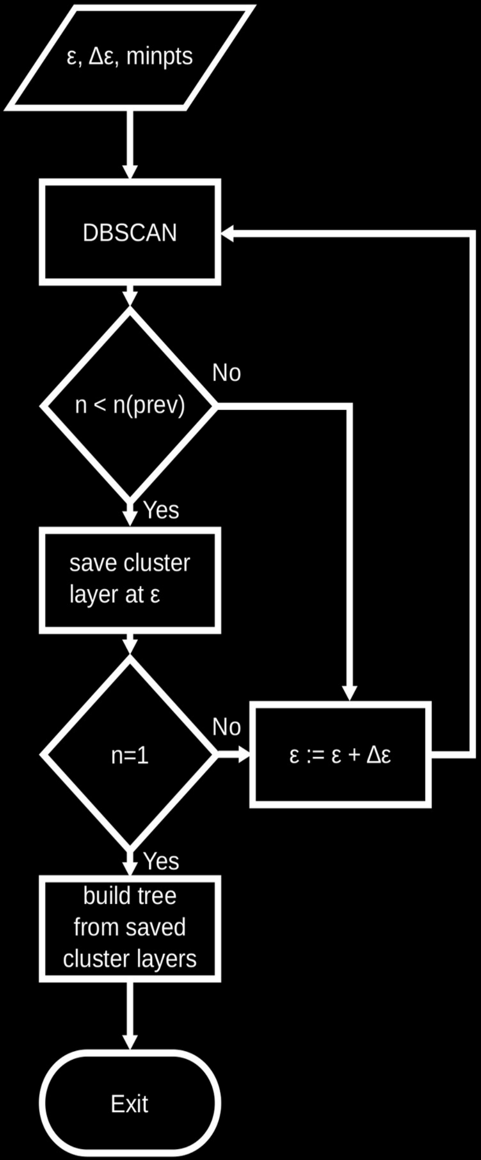


Phylogenetic tree built from sequence space



Treebuilding Algorithm

Algorithm



ε, $\Delta\varepsilon$, minpts

DBSCAN

$n < n(prev)$

No

Yes

save cluster layer at ε

$n=1$

No

Yes

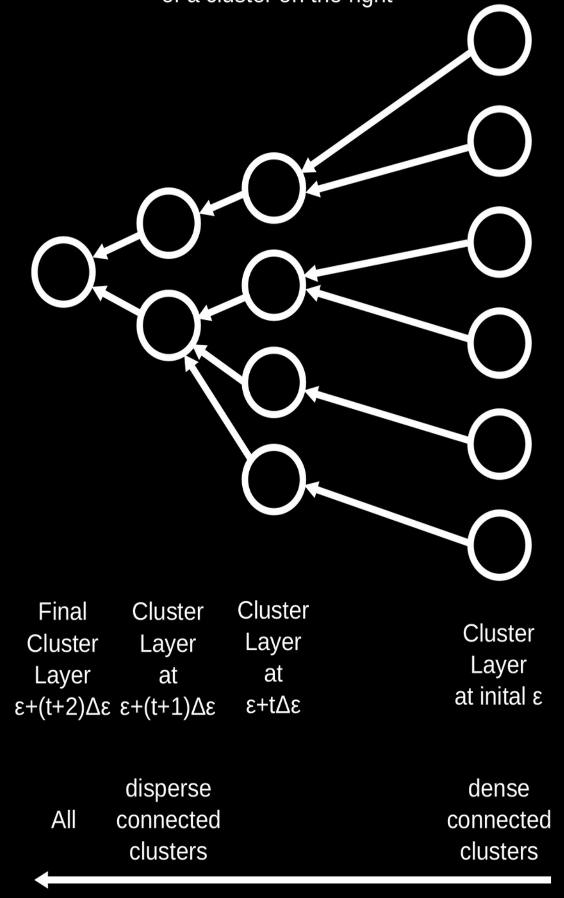
$\varepsilon := \varepsilon + \Delta\varepsilon$

build tree from saved cluster layers

Exit

Treebuilding

Create a connection if a cluster in a layer to the left contains 80% of the sequences of a cluster on the right



Final Cluster Layer $\varepsilon+(t+2)\Delta\varepsilon$

Cluster Layer at $\varepsilon+(t+1)\Delta\varepsilon$

Cluster Layer at $\varepsilon+t\Delta\varepsilon$

Cluster Layer at inital ε

All

disperse connected clusters

dense connected clusters

ε increases
density decreases

Multilevel parallelization

- SIMD Vectoriziation (using C Intel Intrinsics) for
  - K-mer calculation
  - Principal Component Analyses
  - DBSCAN pairwise distance calculation

Multilevel parallelization

- Multithreaded:
  - K-mer calculation
  - Principal Component Analyses
  - Tree Building Algorithm : Runs multiple DBSCAN processes in parallel for different ε radius's

Multilevel parallelization

- Multi GPU and Mulitnode parallelization
  - DBSCAN processes that use the Smith Waterman distance for pairwise distance calculations
  - Multiple pairwise distances are calculated on multiple GPUs in parallel across multiple machines
  - Saturation reached at 4 nodes with 4 GPUs each on the ROMEO supercomputer of the University of Reims Champagne Ardenne

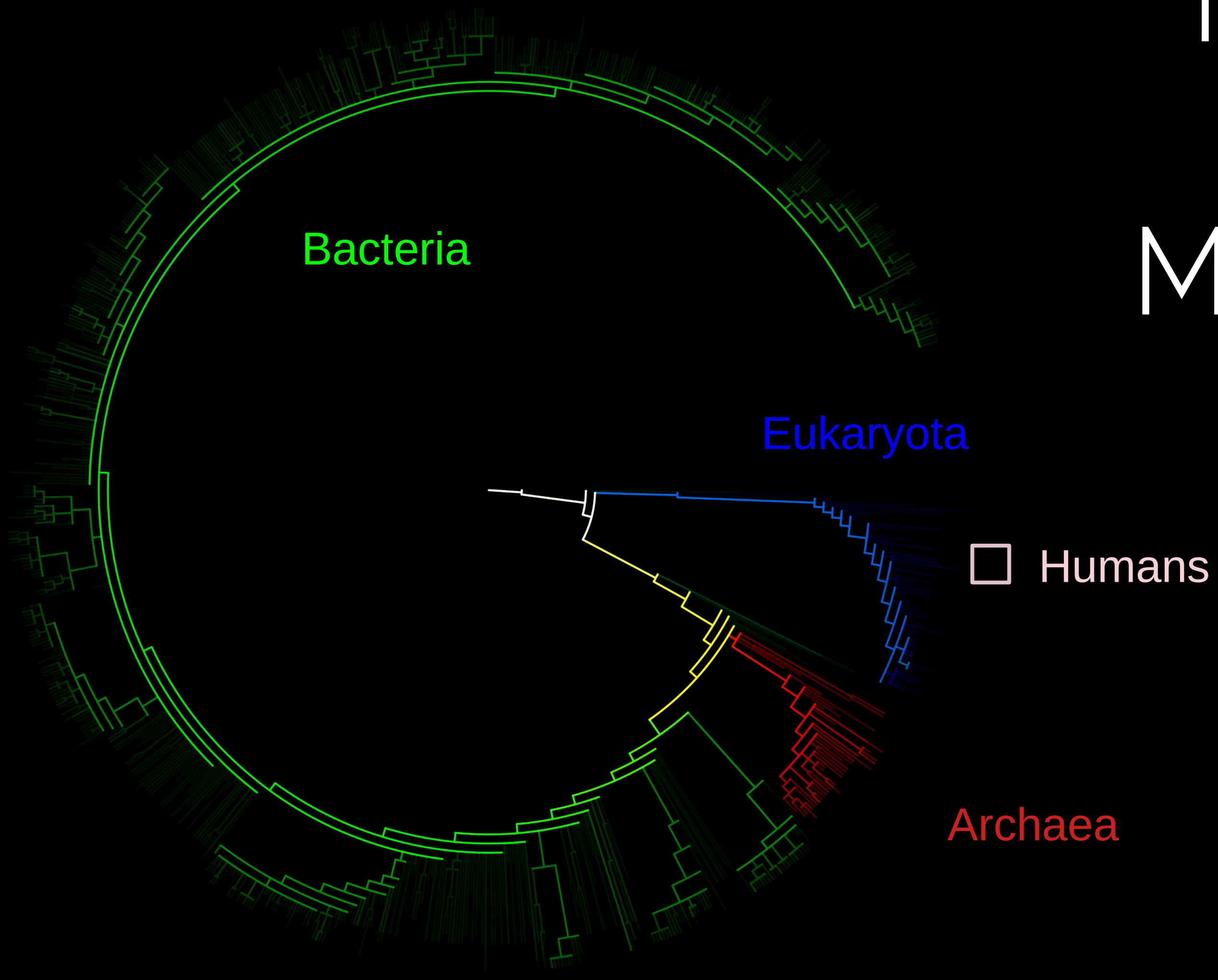
Multilevel Parallization

- Allows:
  - inferring trees for 100 000 sequences using the Smith Waterman distance metric
  - Inferring trees for 2 000 000+ sequences using a k-mer PCA based approach.

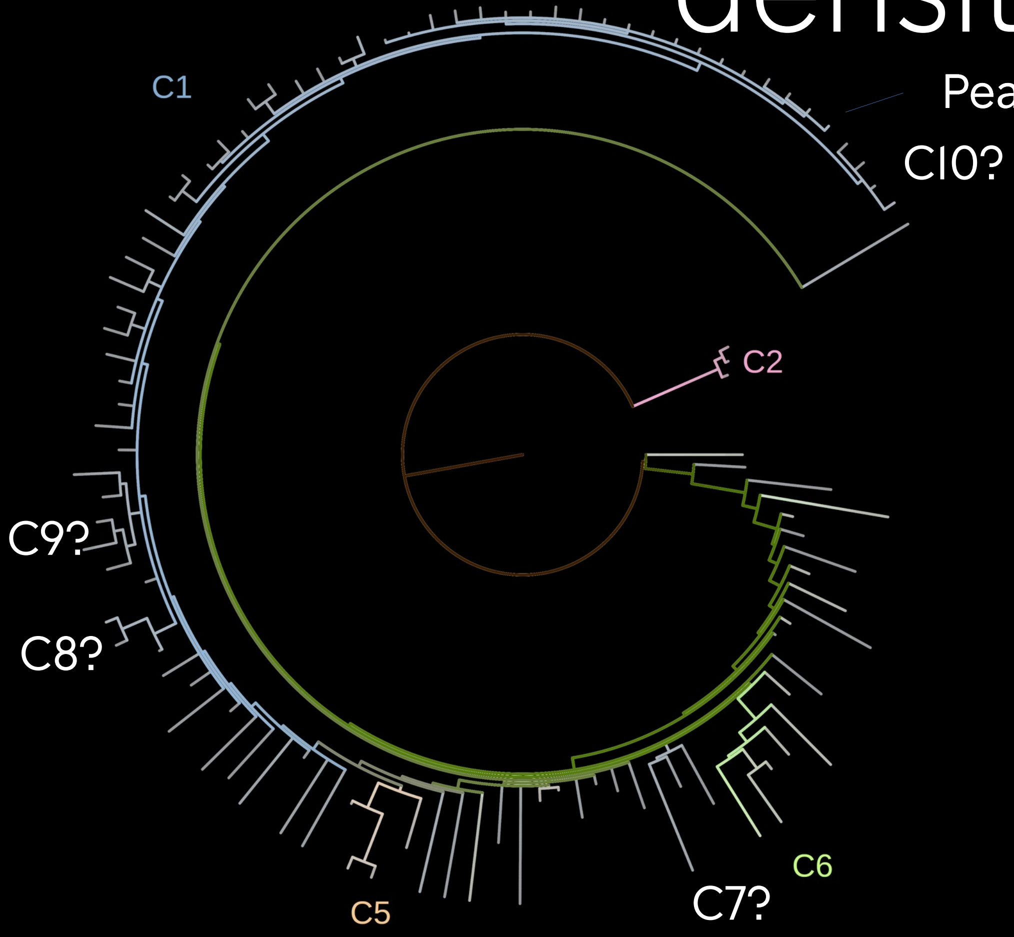
Tree of Life and DNA Barcoding

- We took the SILVA dataset, containing more than 2 million sequences of the 16S and 18S ribosomes.
- And investigated what happened.

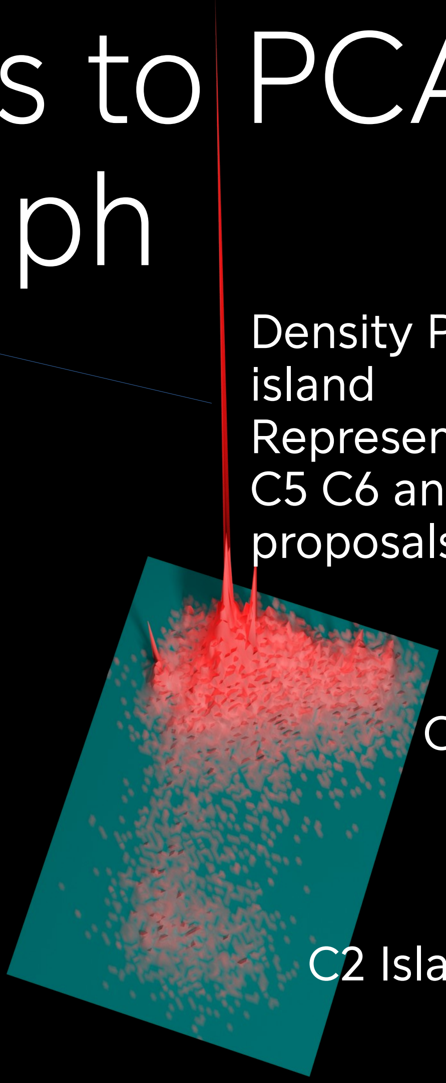
Tree of Life from MNHN-Tree-Tools



Tree corresponds to PCA density graph



Peak in CI



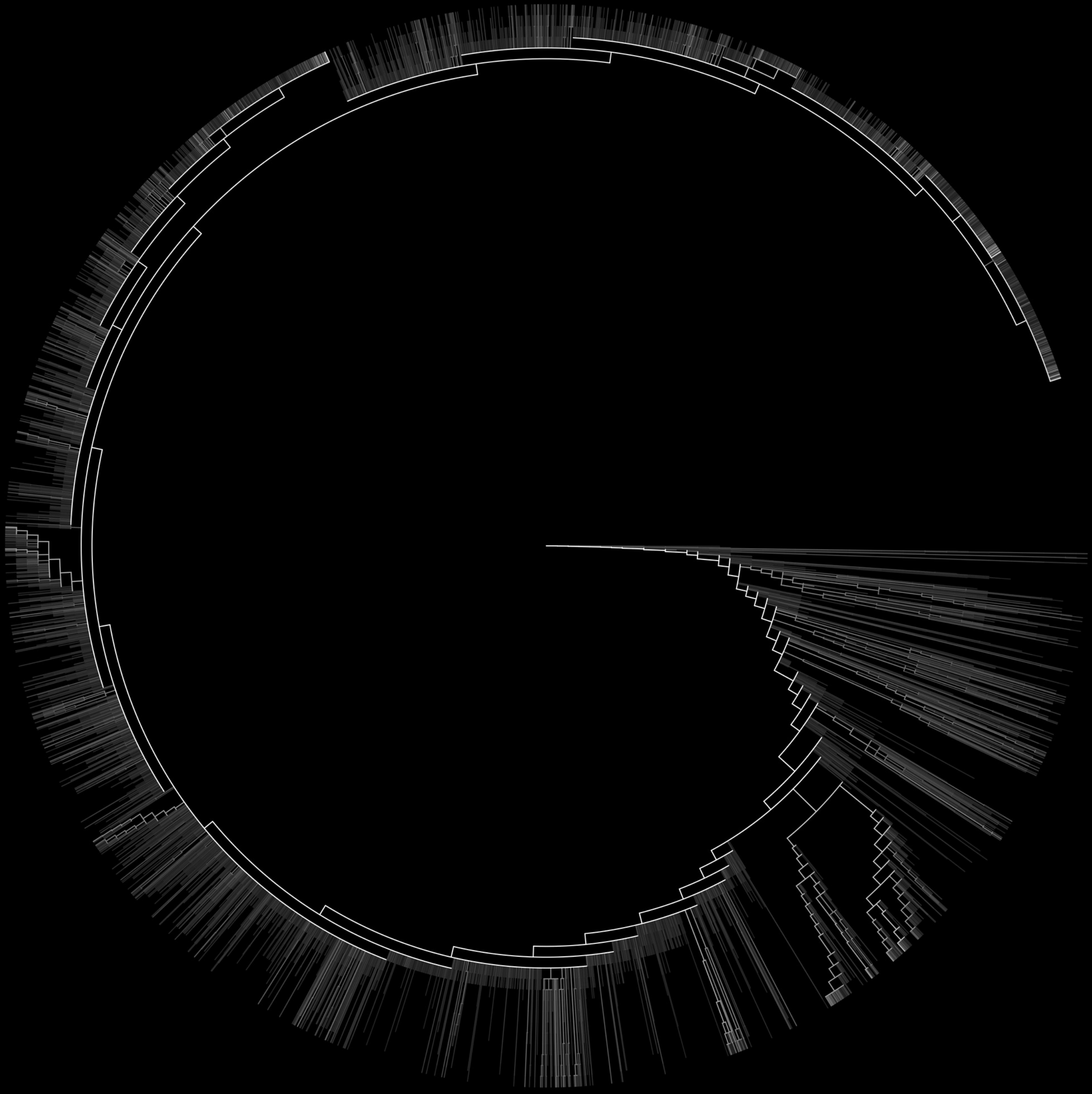
Density Peaks in CI island
Represent groups C5 C6 and new proposals

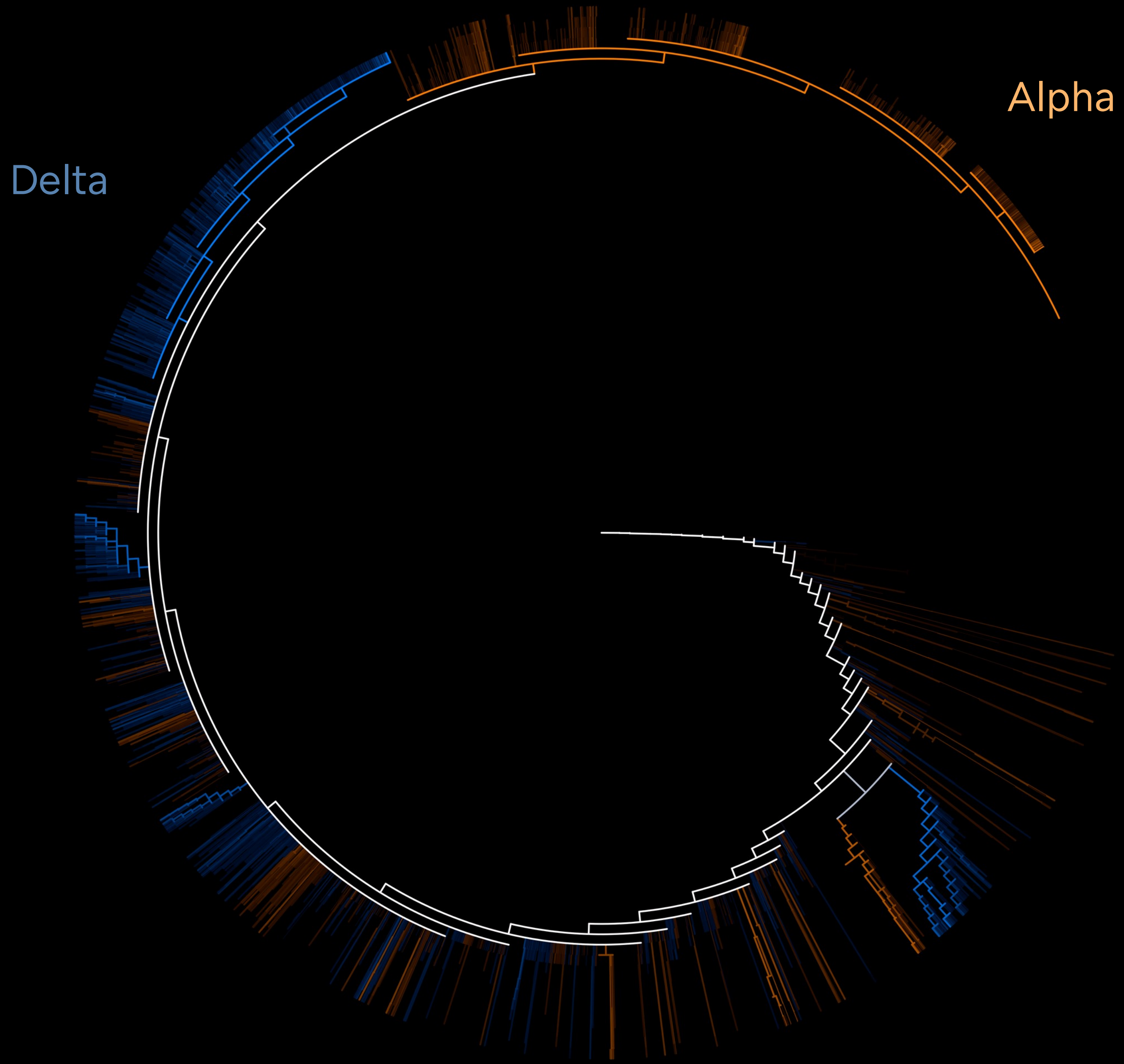
SARS-CoV-2 sequences

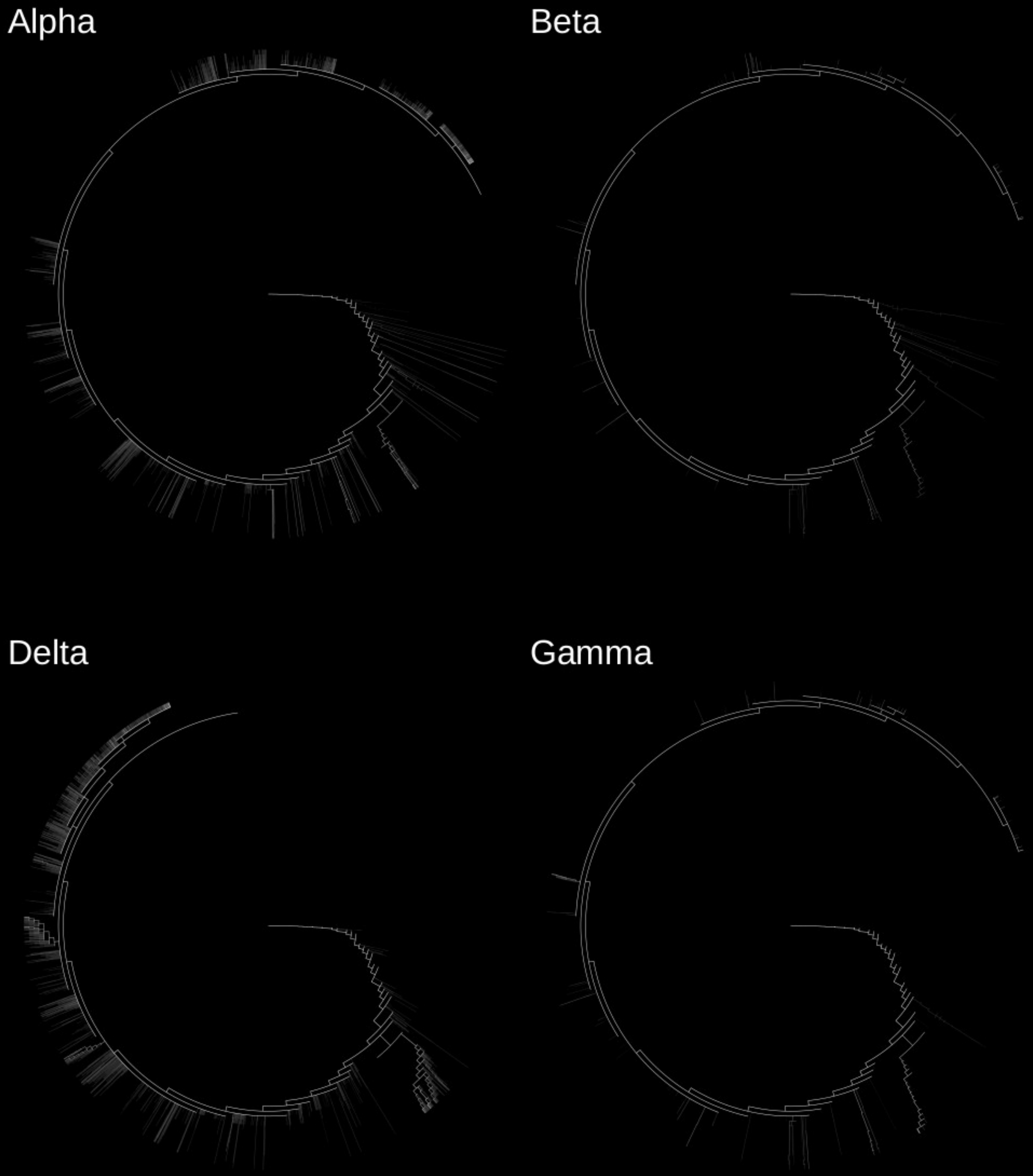
- Downloaded 1 442 669 covid19 sequences from:

  https://www.ncbi.nlm.nih.gov/sars-cov-2/

- Selection critera were:
  - 29 500bp < Sequence Length < 32 300
  - Sequences are from human hosts



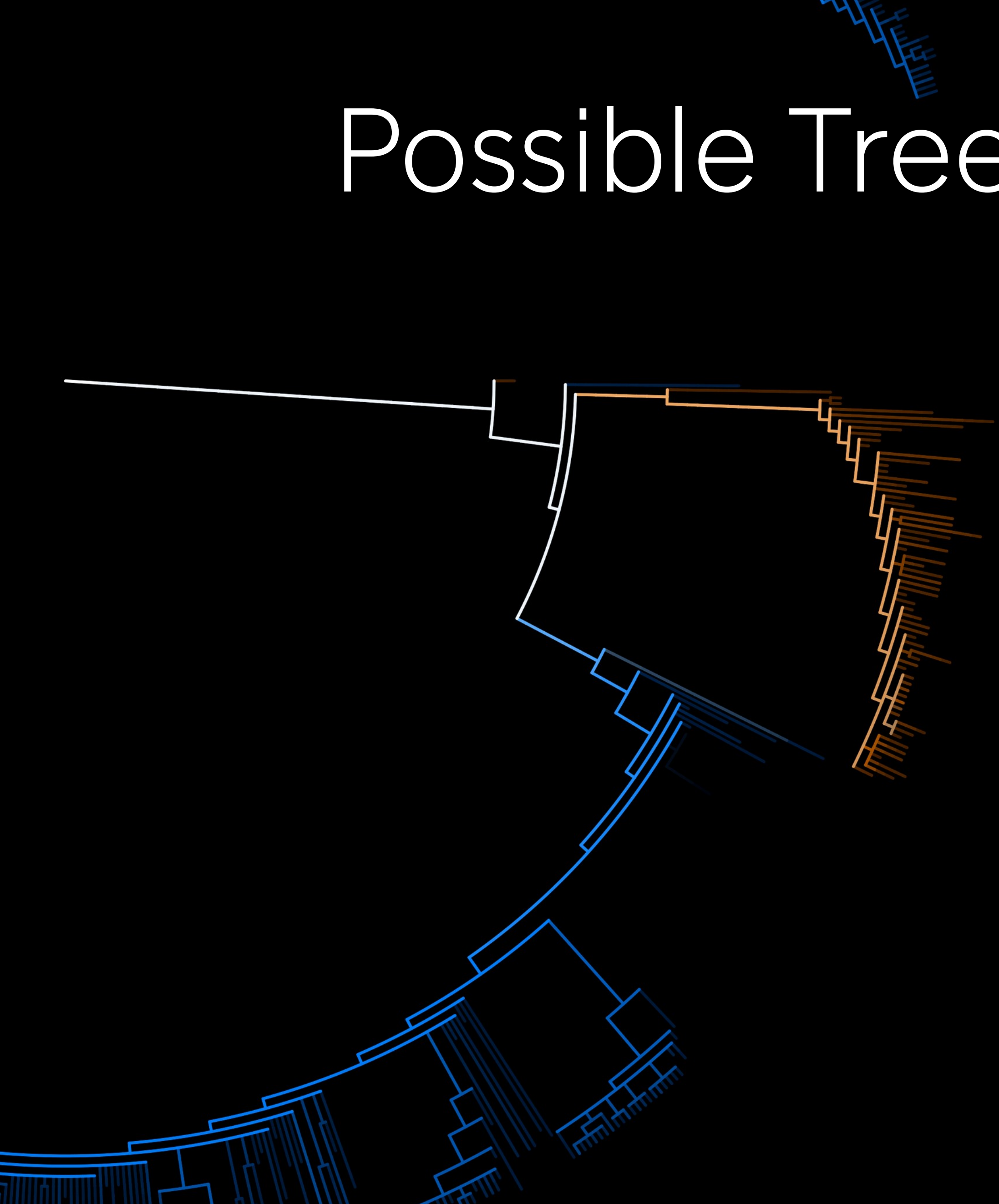

Alpha
Delta



Few Beta and Gamma sequences.

Beta and Gamma seem to be close to Alpha and further from Delta

Possible Tree Properties



- Each node is a cluster discovered by DBSCAN.
- Dense clusters are on the outside of the tree.
- Dispersed clusters are on the inside (at the root) of the tree.
- Tree length corresponds to density change and allows for inference of evolution distance.
- Sequences / Sequence number etc are accessible for each node / cluster etc.

Thanks

- To the Muséum National d'Histoire Naturelle for funding of this project
- To Julien Mozziconnacci, Christophe Escudé and Luïc Ponger for animated discussions and paper editing help.
- To you for having me at the JCAD