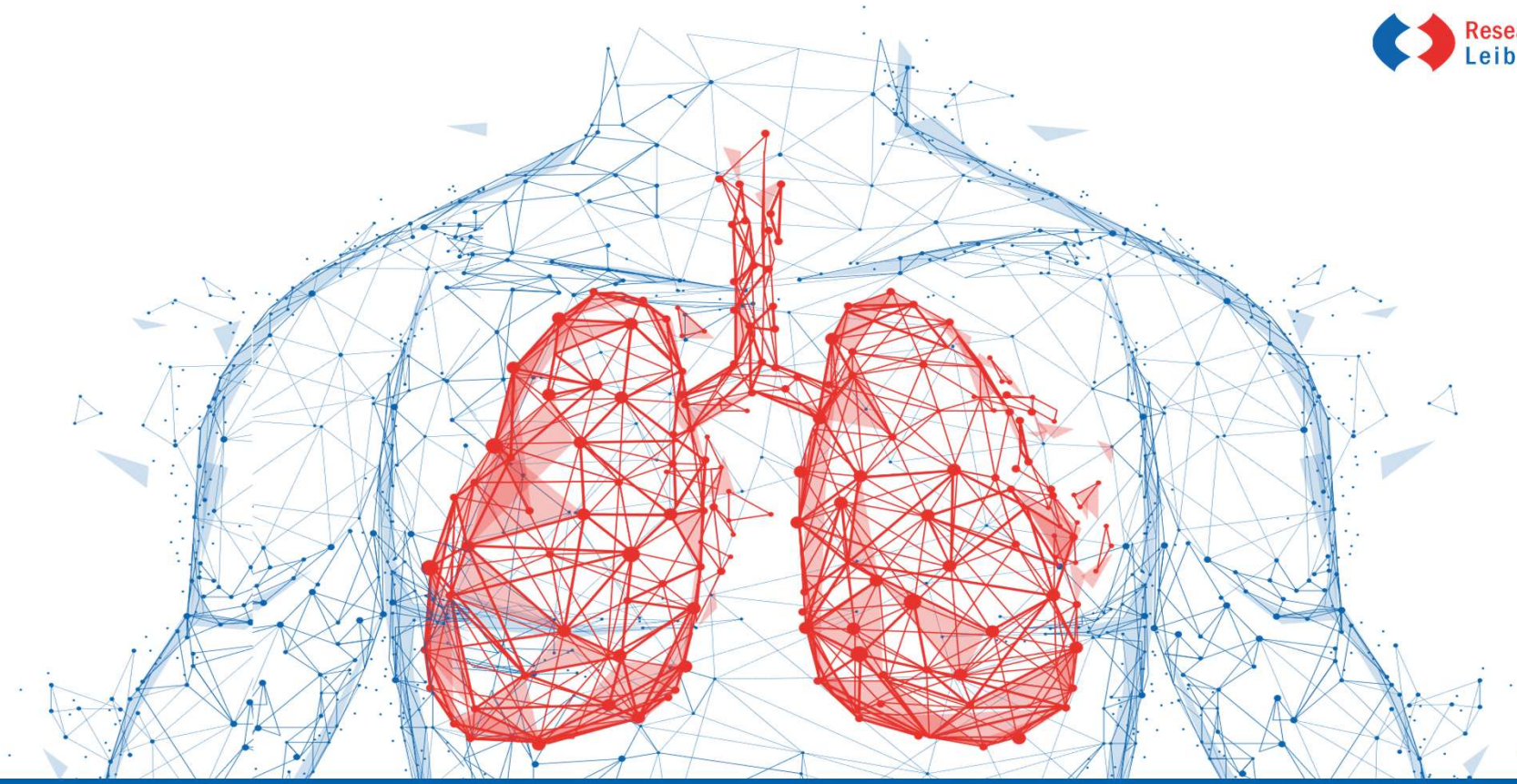
Leibniz Association

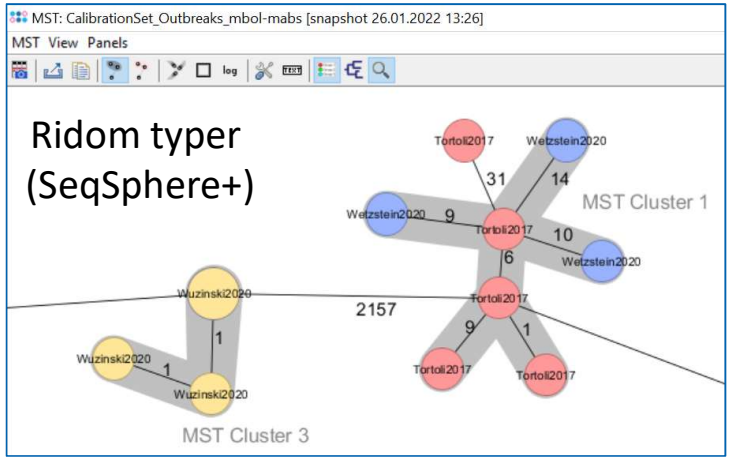
Research Center Borstel
Leibniz Lung Center

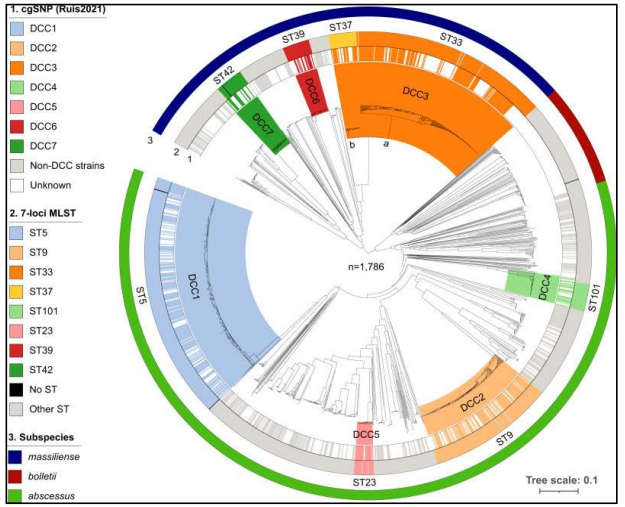


Cluster analysis: Which threshold should I use?

CLUSTER ANALYSIS

- **Cluster analysis = Grouping of isolates**
  - E.g. based on a specific genetic threshold
- Reason for cluster analysis
  - Outbreak investigation (identify likely transmission chain)
  - Surveillance/population genomics → detect circulating lineages or dominant clones
  - Hospital epidemiology → detect nosocomial spread
  - Environmental studies → identify common contamination sources





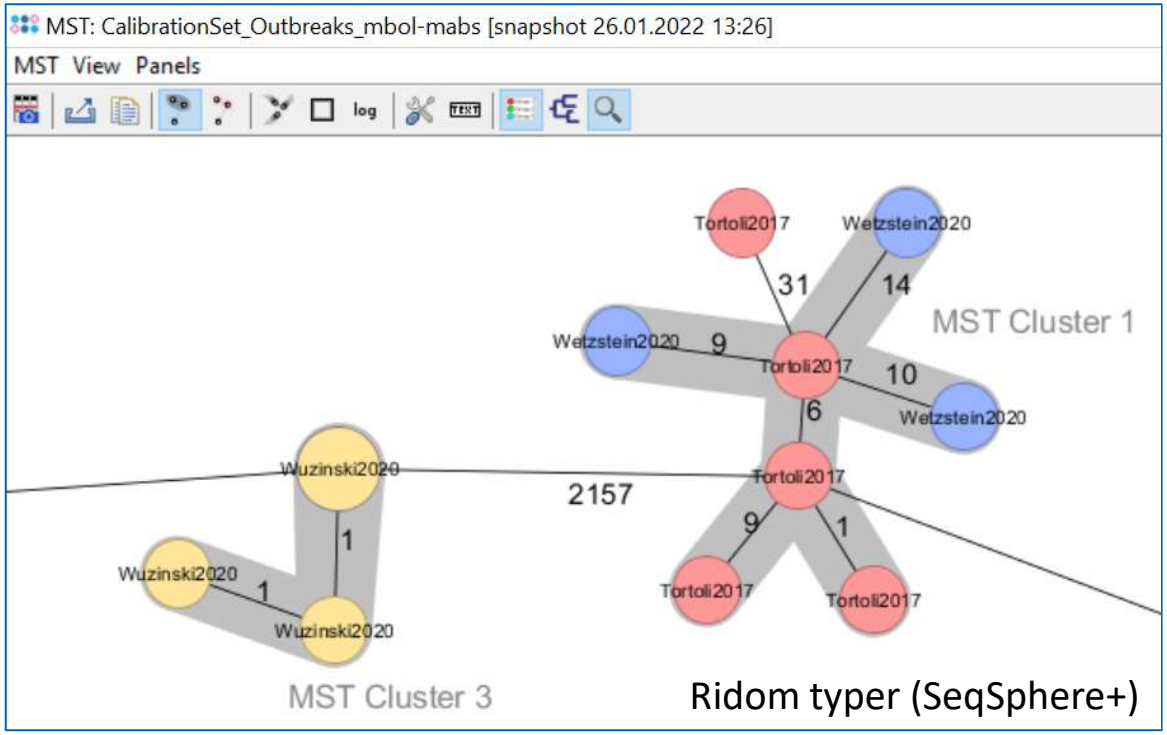
Different patient isolates clustering together with very low distances:
person-to-person transmission
or **transmission from the same environmental source**
or **contamination (pseudo-outbreaks)**

CLUSTER ANALYSIS

- **Cluster analysis** = Grouping of isolates
- **Cluster threshold** = genetic threshold to group isolates

Depends on

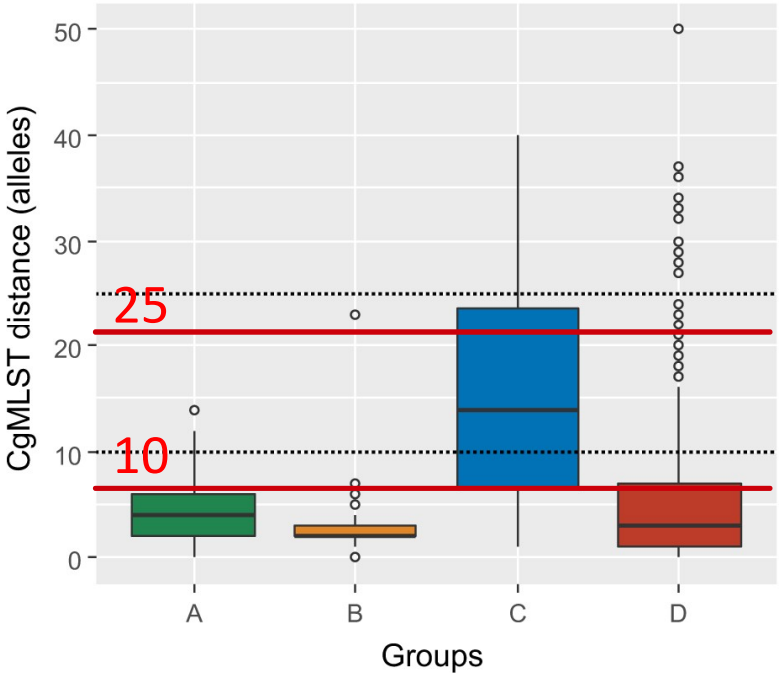
- Species
- Mutation/recombination rate
- Hypermutators
- Typing method
- Sampling timeframe
- Epidemiological question



Cluster threshold: 25

WHICH THRESHOLD SHOULD I USE?

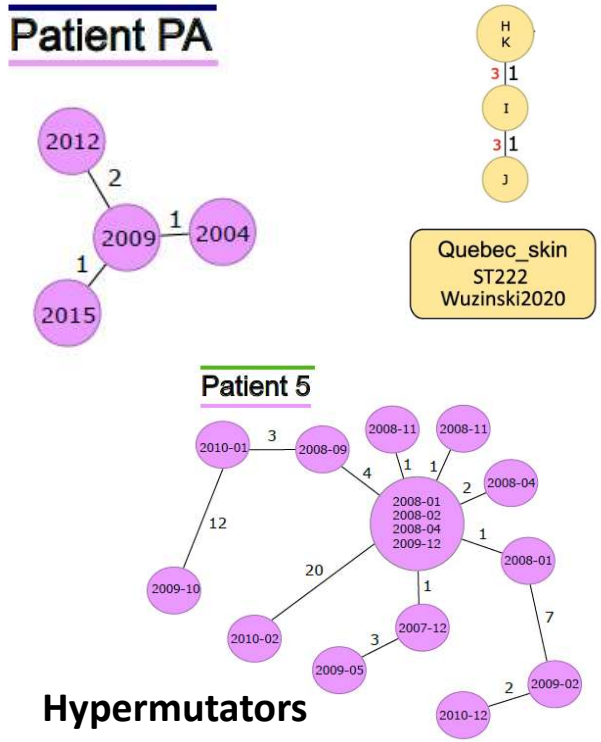
Compare isolates with known epidemiological links



Groups

- A: Extra-pulmonary outbreaks
- B: Clustered pulmonary isolates - Transmission supported by epi data
- C: Clustered pulmonary isolates - Transmission not supported by epi data
- D: Diversity within individuals (CF patients)

Group	Links	Isolates	Patients	Studies	Clusters
A	444	37	37	2	3
B	43	20	20	3	6
C	23	21	21	2	7
D	1150	291	69	3	/



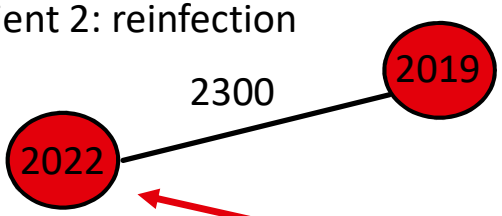
99% pairwise comparisons between epi-linked isolates < 25 alleles
90% pairwise comparisons between epi-linked isolates < 10 alleles

Diricks et. al 2022

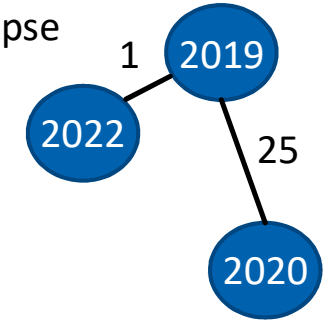
WHICH THRESHOLD SHOULD I USE

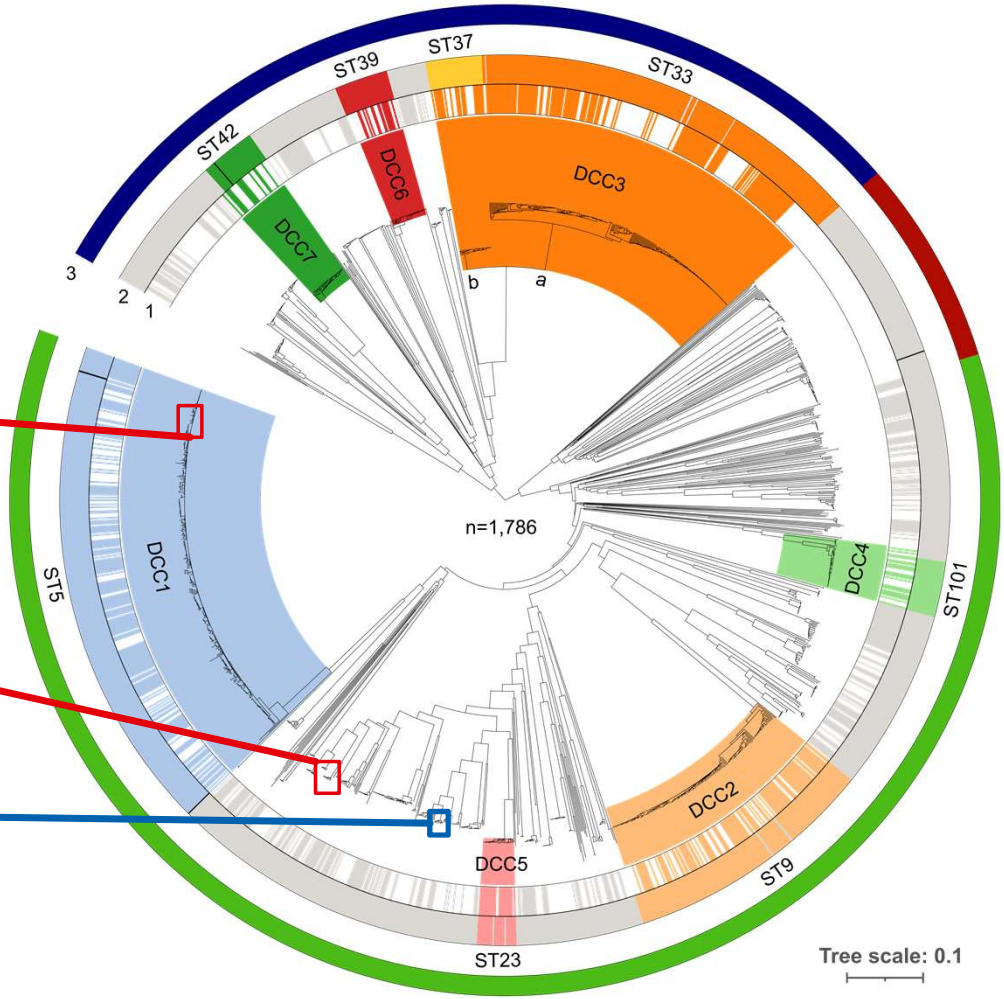
- Reinfection vs relapse
  - Same ST (*M. abscessus*)?
  - Do they cluster together in phylogeny?

Patient 2: reinfection

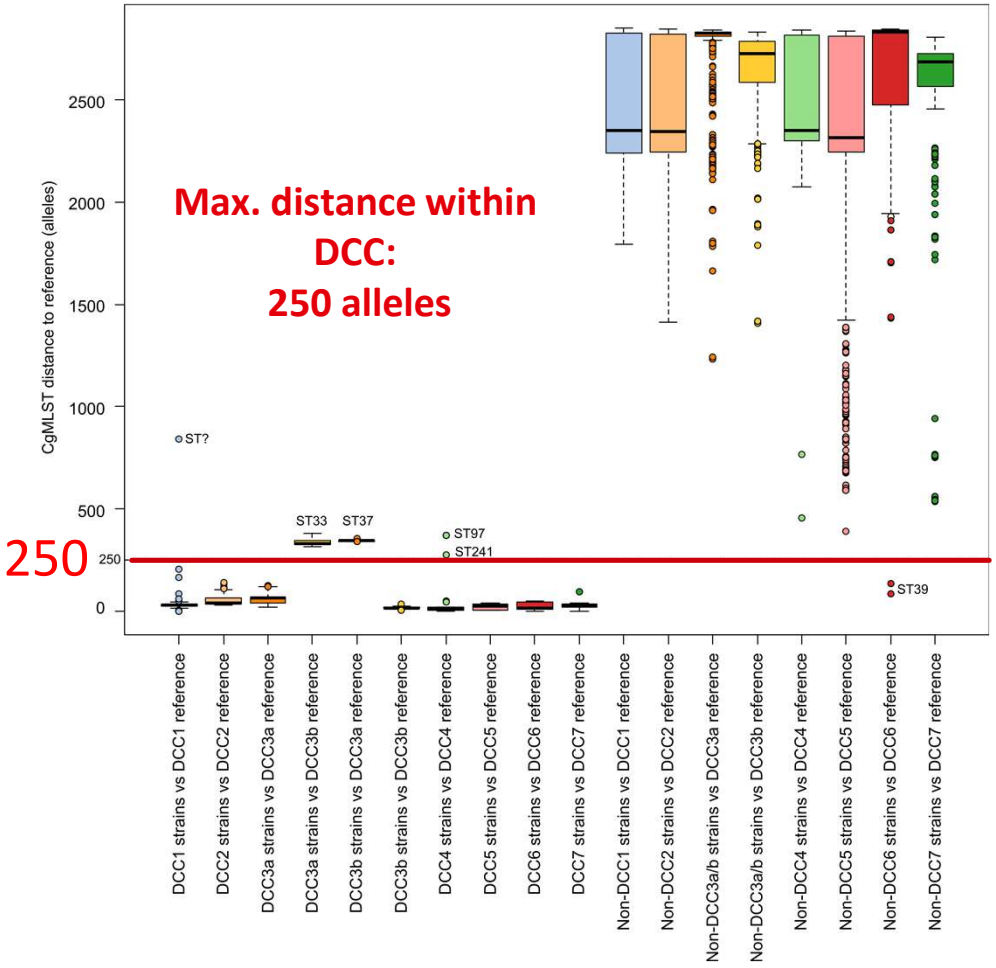


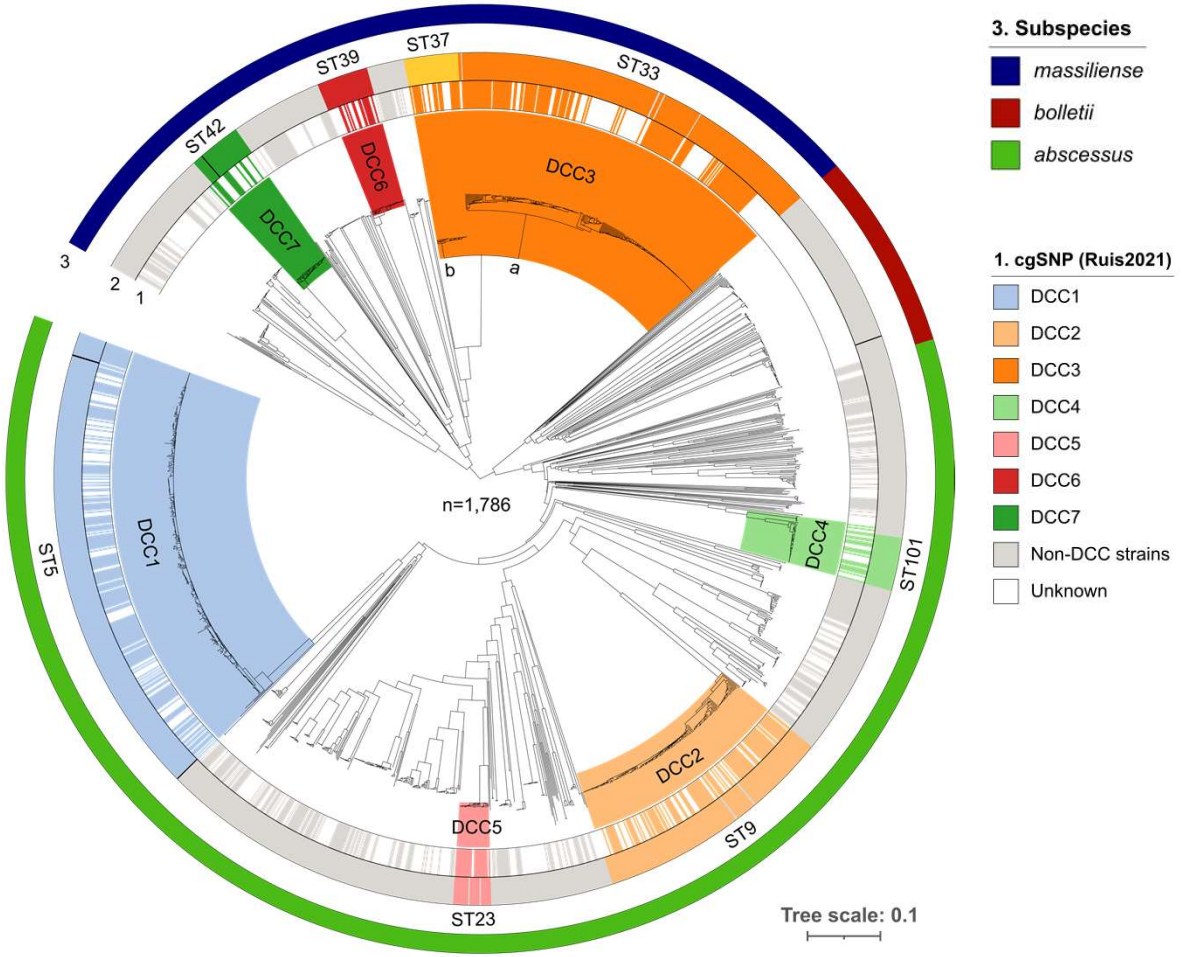
Patient 2: relapse





DCC CLASSIFICATION THRESHOLD

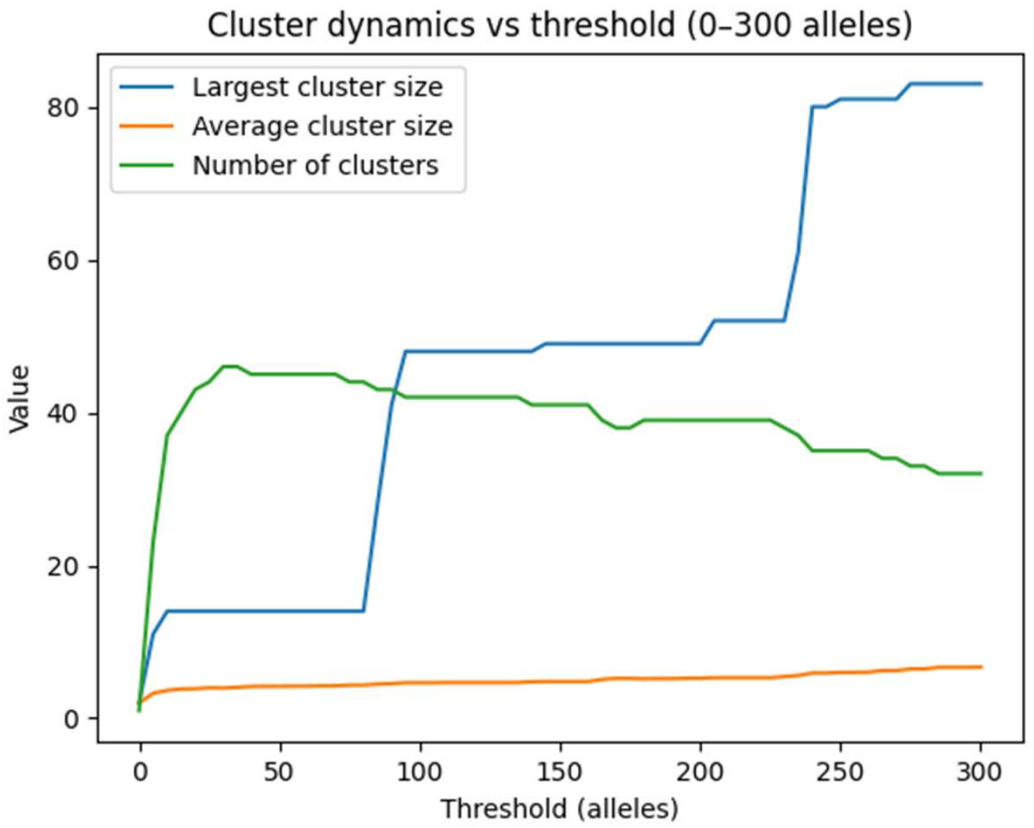




Diricks et. al 2022,

WHICH THRESHOLDS SHOULD I USE

- Use published thresholds (e.g. 25 or 10 for *M. abscessus*)
- Compare known epidemiologically related samples
- Put your samples into context
  - Do they belong to global clones?
  - Is it a hypermutator?
- Visualise genetic distances and groups
- Try different thresholds
  - What happens if I would choose another one?
- Use thresholds not as a fixed cut-off but only as a tool to guide epidemiological investigations



CLUSTER ANALYSIS TOOLS

- GUI: pubMLST + Grapetree
  - Only allows to visually recognize groups, does not define them
- Desktop GUI: Ridom Typer (SeqSphere+) (https://www.ridom.de/ridom-typer/)
  - Allows to visualise clusters using user-defined thresholds
- CLI: MTBseq (https://github.com/ngs-fzb/MTBseq_source)
  - By default outputs groups of isolates with a distance threshold of 12 SNPs (can be changed)
- CLI: FastBaps (https://github.com/gtonkinhill/fastbaps)
  - Does not use a fixed threshold, clusters based on bayesian population genetics
- CLI: PopPUNK (https://github.com/bacpop/PopPUNK)
  - Does not use a fixed threshold, cluster based on k-mer distances in the core and accessory genome
- CLI: Reportree (https://github.com/insapathogenomics/ReporTree)
  - Supports dynamic cluster definition across multiple genetic distance thresholds