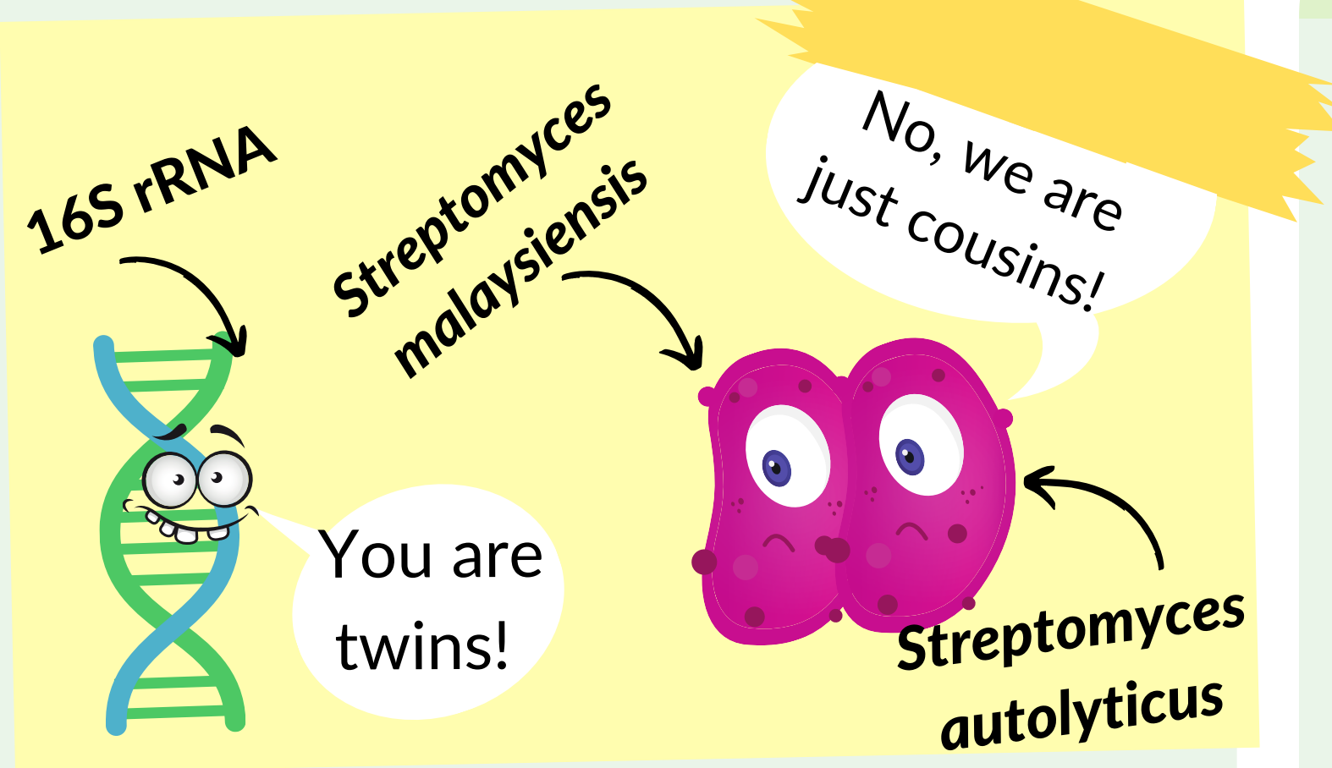
Improved and Extended Multilocus Sequence Typing (MLST) Scheme for *Streptomyces* Reveals Complex Taxonomic Structure

Angelika Kiepas, Paul A Hoskisson, Leighton Pritchard
Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street, Glasgow G4 0RE, Scotland, UK

Introduction

Streptomyces species produce over 60% of all clinically-approved bioactive compounds[1]. Continuing discoveries of new natural products suggest that *Streptomyces* genomes remain a promising source for novel antibiotics[2]. Comparative genomics and pangenomics are powerful tools for inferring genes involved in the synthesis of novel antibiotics from closely related genomic sequences. The contested nature of *Streptomyces* [3] taxonomy means that relying on existing assigned taxa may be misleading for pangenomics.



MLST is used for genomic classification by comparing internal loci[4]. The current canonical *Streptomyces* MLST scheme provided by pubMLST comprises six markers (16S rRNA, atpD, gyrB, recA, rpoB and trpB) and 236 sequence types (STs; only two new STs were reported since 2016)[5].

SCAN ME

Video

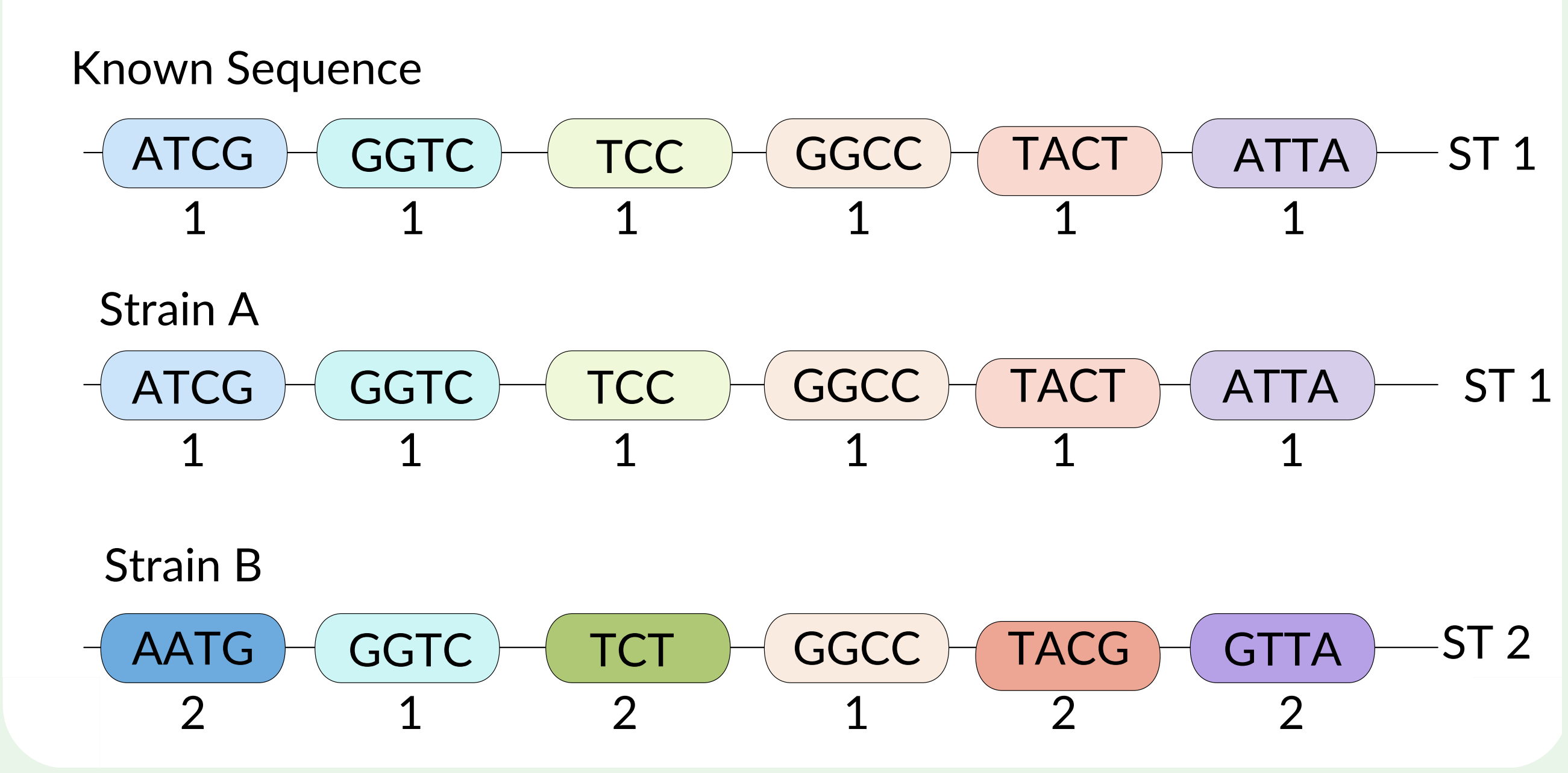


Figure 1. In MLST each marker variant sequence is assigned a unique number. For a single isolate, these numbers are combined to produce a profile, and each unique profile is assigned a ST.

Aim

With the recent increase in available *Streptomyces* sequences we can now ask:

- How do STs map onto *Streptomyces* taxonomy determined from genome sequences?
- What does this tell us about the taxonomic structure of *Streptomyces*.

Methods

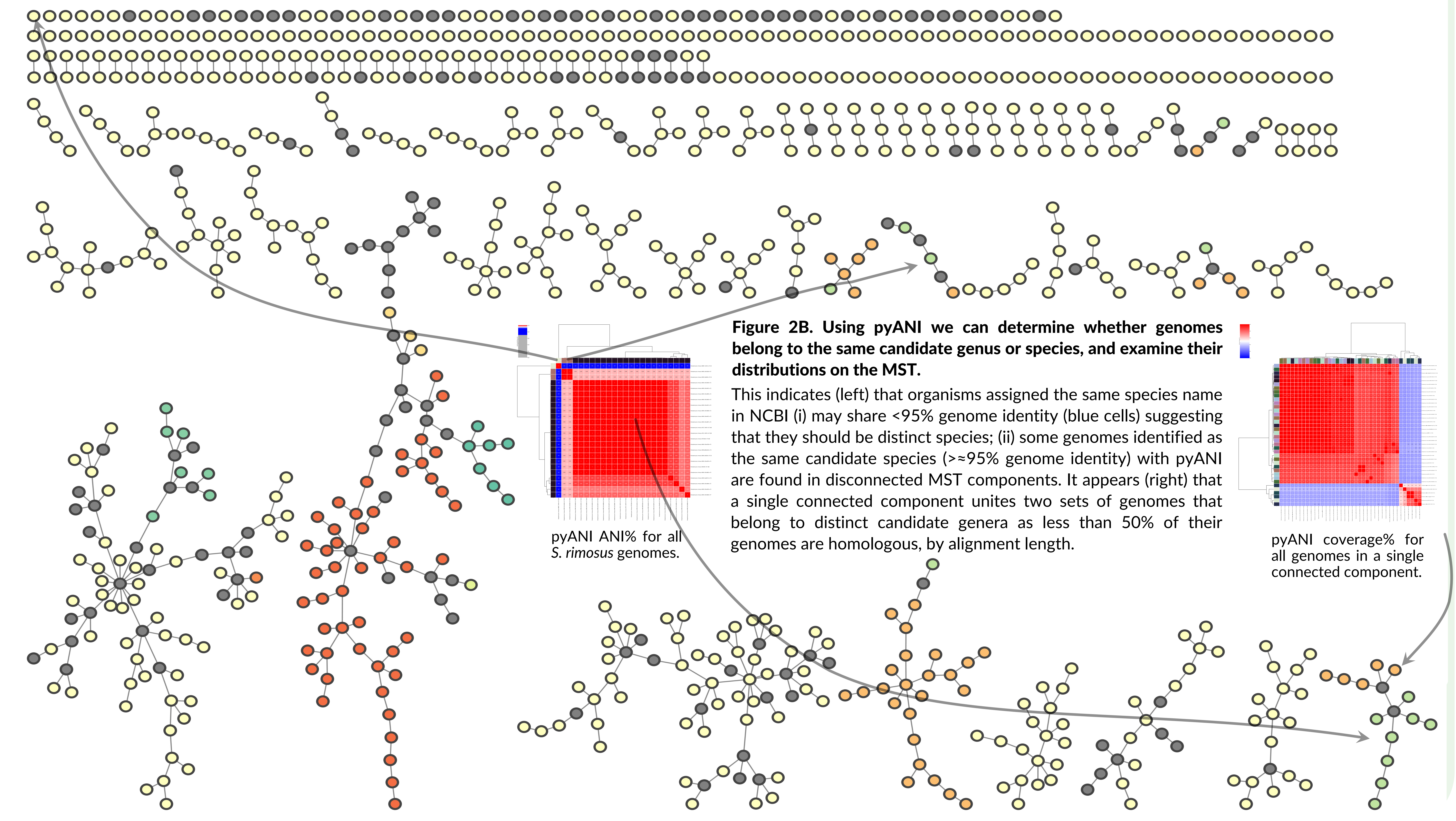
All 2276 available *Streptomyces* genome sequences were downloaded from NCBI[6] on the 8th July 2021.

673 16S rRNA, 813 atpD, 576 gyrB, 890 recA, 873 rpoB and 784 trpB new allele variants were identified with MLST tool[7].

Streptomyces taxon boundaries were assessed with pyANI [8] (%ID >≈95%, %coverage >≈50%).

Results

Figure 2A. Minimum spanning tree (MST) with 852 STs and 292 connected components describing all sequenced *Streptomyces* genomes, and all STs from the pubMLST database. Each node represents a unique ST, and each edge corresponds to traversing from one ST to other by making up to five marker changes. This division of *Streptomyces* into 292 components that share no marker alleles with each other implies a set of natural divisions between groups of isolates. There are 150 pubMLST STs without any representative genome (grey nodes). Using pyANI it was determined that some connected components describe a single candidate genus (single node colours within a connected component), and some components represent more than one genus (multiple node colours within a connected component).



pyANI ANI% for all *S. rimosus* genomes.

Figure 2B. Using pyANI we can determine whether genomes belong to the same candidate genus or species, and examine their distributions on the MST.

This indicates (left) that organisms assigned the same species name in NCBI (i) may share <95% genome identity (blue cells) suggesting that they should be distinct species; (ii) some genomes identified as the same candidate species (>≈95% genome identity) with pyANI are found in disconnected MST components. It appears (right) that a single connected component unites two sets of genomes that belong to distinct candidate genera as less than 50% of their genomes are homologous, by alignment length.

pyANI coverage% for all genomes in a single connected component.

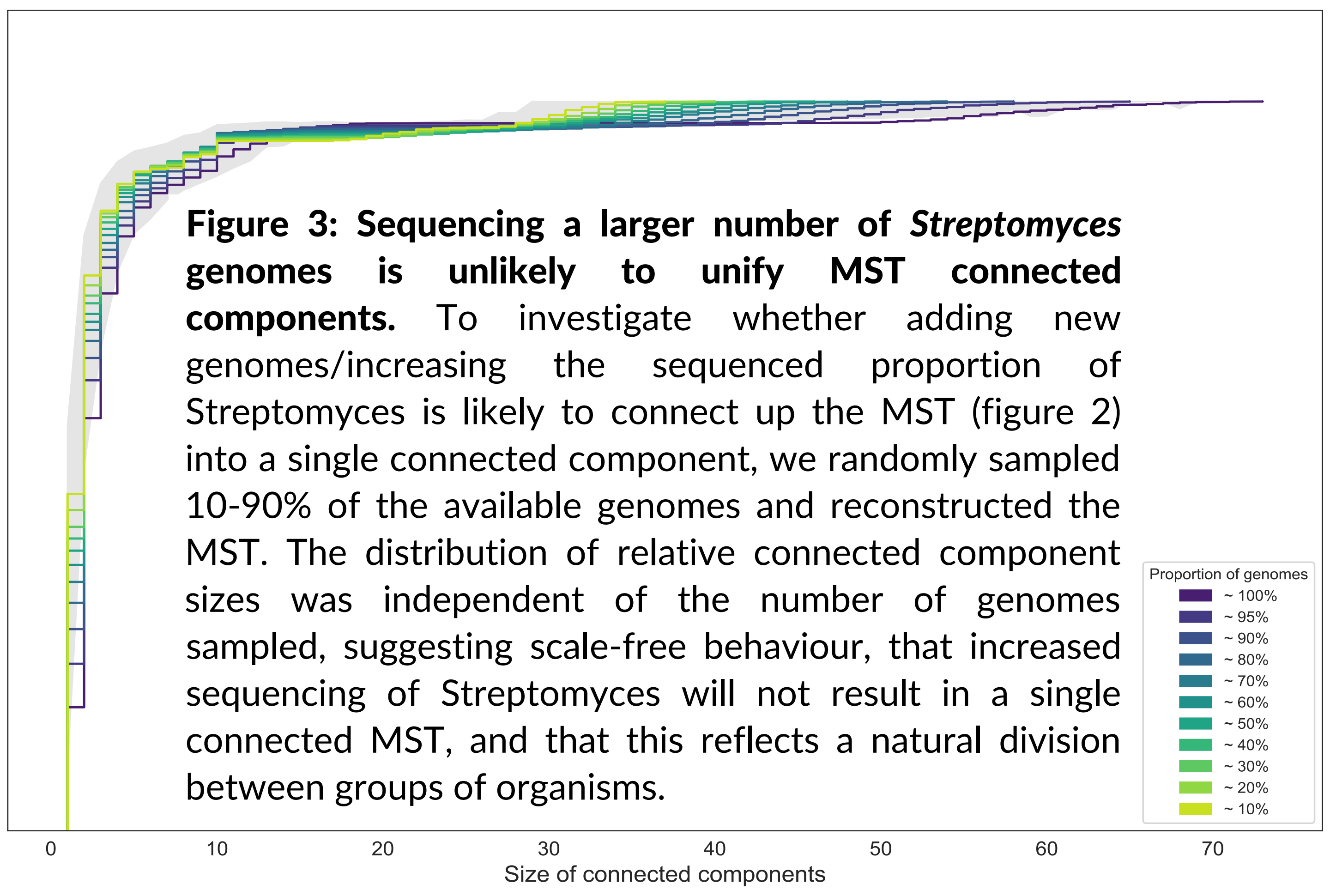


Figure 3: Sequencing a larger number of *Streptomyces* genomes is unlikely to unify MST connected components. To investigate whether adding new genomes/increasing the sequenced proportion of Streptomyces is likely to connect up the MST (figure 2) into a single connected component, we randomly sampled 10-90% of the available genomes and reconstructed the MST. The distribution of relative connected component sizes was independent of the number of genomes sampled, suggesting scale-free behaviour, that increased sequencing of Streptomyces will not result in a single connected MST, and that this reflects a natural division between groups of organisms.

Conclusions

- Multiple different candidate genera can be present in the same connected group of STs.
- Isolates that belong to the same candidate genus or species can be split across disconnected groups of STs.
- Currently assigned species names in NCBI do not reflect genomic difference and may incorrectly group organisms.
- The current set of MLST markers is unlikely to produce a fully-connected MST independent of number of sequenced genomes.

SCAN ME

Interactive Graph

References

[1] Procopio *et al.* (2012) *Braz J Infect Dis* doi:10.1016/j.bjid.2012.08.014
[2] Maiti *et al.* (2020) *Scientific reports* doi: 10.1038/s41598-020-66984-w
[3] Labeda *et al.* (2012) *Springer* doi: 10.1007/s10482-011-9656-0
[4] Maiden *et al.* (1998) *PNAS* doi: 10.1073/pnas.95.6.3140
[5] Jolley *et al.* (2018) *Wellcome Open Research* doi: 10.12688/wellcomeopenres.14826.1
[6] Sayers et al. (2020) Nucleic Acids Res. doi: 10.1093/nar/gkaa892
[7] https://github.com/tseemann/mlst
[8] Pritchard et al. (2016) Analytical Methods doi:10.1039/c5ay02550h

Acknowledgements

We acknowledge the University of Strathclyde Research Excellence Award for PhD funding and a Microbiology Society Conference Grant 2022 for enabling me to attend.