

University of
Strathclyde
Science

Resolving *Streptomyces* taxonomy: conflicts between
whole-genome, core gene, MLST and 16S classifications

PhD Thesis

Angelika Beata Kiepas

PhD Student


Strathclyde Institute of Pharmacy and Biomedical Sciences

University of Strathclyde, Glasgow

March 10, 2025

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed: 

Date: March 10, 2025

Note to the reader

I strongly recommend reading this thesis in its digital format. Due to the number of figures containing large amount of data points, some graphs may be difficult to interpret in a printed copy, despite efforts to optimise them for readability on paper. The digital version allows for zooming in and out, providing a clearer view of the plots and data, ensuring that the reader can fully appreciate the details presented.

As a result of the work carried out for this project, a first-author paper has been published. The core content of this paper, including the methodology, results, and discussion, is presented in Chapter 2 of this thesis. Additionally, excerpts from the introduction and conclusions are incorporated into Chapters 1 and 5, respectively. The reference for this publication and its pre-print is listed below.

Kiepas, A. B., Hoskisson, P. A., & Pritchard, L. (2024). 16s rRNA phylogeny and clustering is not a reliable proxy for genome-based taxonomy in *Streptomyces*. *Microbial Genomics*, 10(9), 001287

Kiepas, A. B., Hoskisson, P. A., & Pritchard, L. (2023). 16s rRNA phylogeny and clustering is not a reliable proxy for genome-based taxonomy in *Streptomyces*. *bioRxiv*, 2023-08

Acknowledgements

I once read: 'Good supervisors can take you to incredible heights. They help you learn to fly, providing the wind beneath you, and providing a net for when you fall.' With that analogy in mind, I would like to express my heartfelt gratitude to my supervisors, Dr Leighton Pritchard and Prof Paul Hoskisson, whose thoughtful and wise guidance has not only helped me soar to new heights but also guided me to fly higher and provided a safety net for me at each step of my PhD journey.

Leighton, I will forever be grateful to you for placing your trust in me by giving me the opportunity to work on this PhD project, despite my lack of prior coding experience. Thank you for introducing me to the world of bioinformatics and for your patience when I struggled to grasp even the simplest concepts. I look forward to using and building upon this skill set in the near future. I would also like to thank you for inspiring me during our meetings and for providing feedback that helped shape me into the computational biologist I am today. Words cannot fully express my gratitude for your unwavering support and for always making time to listen to me (or letting me cry my eyes out in your office), especially during the most challenging times. You are a wonderful human being and a true friend.

Paul, I can't thank you enough for believing in me and helping me become the scientist I am today. I am deeply grateful for your guidance and for our meetings, where

I discovered just how exciting microbes are. Your incredible microbiology knowledge and expertise in *Streptomyces* have shaped the development of my research over the past four years. I cannot thank you enough for your support, both academically and personally, especially during the toughest times. You are a truly wonderful person.

I am also grateful to past and present PhD students and postdoc colleagues from HW601 who made me feel welcome after I finally got my desk on-site in my final year, and for all the wonderful memories that we have created together. Special thanks to John Munnoch for always finding the time to listen to me. Our conversations were more than just words; they helped me see hope when I needed it most. And thank you, too, for stepping in when I couldn't eat—sharing those lunches meant I didn't carry the guilt of wasting food. I would also like to thank all members of the Random Generator Bioinformatics Journal Club, especially Leighton for starting it, Dr Morgan Freeny for delicious cakes and Tom Harris for taking over and continuing its legacy. I will miss these sessions deeply.

A special thank you to Jasmine Thomson for being the incredible friend I never knew I needed. You are truly the most selfless person I've ever met. You "adopted" me at school when I could barely speak English, guided me through some of the strangest times in my life, and understood me without words (literally). I wouldn't be where I am today without your help, and I'm forever grateful.

Big thanks to Ania Brakowska, my best friend from boarding school. If there's one thing I really miss about Poland, it's you. I can't even put into words how grateful I am for all the times you checked in on me, especially when you had your own stuff going on. Anyone who has you in their life is lucky.

I would also like to thank my closest family who supported me at every turn of my PhD journey. Mum, thank you for believing in me and allowing me to chase my dreams while you stayed at home, cooking my dinners and ironing my clothes. After I moved out, I realised how much work you put into this, allowing me to concentrate on my career. I would especially like to thank my brother, Hubert, who put his own education on hold to provide financial support when things were tough. I cannot express enough how much it means to me, and I promise to make you proud.

Besides my friends, family, and other sources of inspiration that have supported me throughout this PhD journey, I would like to express my gratitude to my purrfect cats, Penguin, Wasabi, and Mishka. Doing my PhD during lockdown was mentally challenging, but with your company, I never felt alone in my office. You were also the reason for countless edits to this thesis, with constant walks across my keyboard, and I wouldn't have it any other way.

To my love, Chaz—my biggest cheerleader and best friend. Before I started this PhD journey, I often wondered if I was good enough. But you knew from the very beginning that not only was I good enough, but that I would succeed. Words can't capture just how much your support has meant to me over these long four years. You were always there to reassure me that things would be okay. When I felt like I was falling apart (and falling apart wasn't an option), you held me together. Thank you for being my rock, for the countless times you provided a shoulder to cry on, and for all those late-night trips to get sweets and Coca-Cola—they meant more than you know. I'm also so grateful for how you took care of me; I never had to cook dinner or make my own lunch (even though I had the nerve to give some away!). Your love and support

made this journey so much easier. Writing this thesis was mentally exhausting, and you made sure to keep me entertained. Even though you're a total gym freak, you put it aside to make sure I was staying sane, and our weekly beer runs to Willy Wastle's were much needed. It's wild to think that when I started this PhD, we weren't even living together, and now not only do we share a home, but we're engaged! You banned me from wedding planning until this thesis was done... so now that it's finished, do I finally have the green light? Always and forever.

Finally, I would like to thank my parents-in-law to be. Thank you for everything you have done for me and Chaz. Kenny will be thrilled to hear this has finally been submitted.

*”Kiedym cie żegnał, usta me milczały, I nie wiedziałem, jakie słowo rzucić,
Wiec wszystkie słowa przy mnie pozostały, A serce zbiegło i nie chce
powrócić.”*

– Adam Asnyk

This thesis is dedicated to my late grandmother, Urszula, who never saw this adventure. Grandma, I would give anything to tell you all about this journey. You will always be in my heart because there you are still alive.

”For he who knows not mathematics cannot know any other science; what is more, he cannot discover his own ignorance, or find its proper remedy.”

– Roger Bacon

Abstract

Accurate taxonomy is central to comparative genomics, our understanding of microbial evolution, and many aspects of applied microbiology. The current taxonomy of *Streptomyces* is under active revision, and resolving this could provide a foundational framework for genome-based investigation of this important genus, the source of 60% of clinically-approved antibiotics.

Several approaches have been proposed for taxonomic classification, such as single, multi-gene, and whole-genome phylogenies, Multilocus-Sequence Typing (MLST) and Average Nucleotide Identity (ANI). However, these approaches often disagree in detail, in part due to the varying amounts of genomic data they use to define species.

Reconstructing phylogeny for 14,239 *Streptomyces* 16S sequences suggests three major groups but does not place the same species consistently within the tree. Through ANI analysis, I demonstrate that 16S zero-radius Operational Taxonomic Units (zOTUs) are often inconsistent with ANI-based taxonomy.

Using six-gene MLST and by updating the canonical pubMLST scheme with 568 new sequence types from all *Streptomyces* genomes, I find that MLST subdivides *Streptomyces* into 278 distinct groups. Using ANI, I establish provisional species and genus boundaries within and between MLST subgroups, concluding that the current MLST scheme does not align with genome-based taxonomy, with multiple potential

misclassifications.

Using a core gene tree constructed from single-copy orthologues across *Streptomyces* genomes, I obtain a robust phylogeny that recapitulates the three-group structure suggested by 16S and propose that horizontal gene transfer may be common in *Streptomyces* even for highly conserved genes. By combining this core gene tree with ANI, I propose quantitative thresholds for subdividing *Streptomyces* into new taxa, identifying distinct groups.

Taken together, my analyses support extensive reclassification of *Streptomyces*, and provide a principled basis for making informed decisions about which methods to prefer for determining *Streptomyces* taxonomy, and for subdividing these genomes into groups that support comparative genomics analyses for antibiotic discovery.

Contents

Note to the reader	ii
Acknowledgements	iii
Abstract	ix
List of Figures	xvii
List of Tables	xxvi
1 General Introduction	2
1.1 Antimicrobial Resistance Crisis	2
1.2 A brief history of natural bioactive compound discovery	9
1.2.1 Empirical drug discovery	9
1.2.2 The influence of genomics on drug discovery	24
1.2.3 A brief history of genome sequencing	24
1.2.4 Drug discovery through pangenomic analyses	31
1.3 Taxonomy Across Time: Past, Present, Future	40
1.3.1 Origins of Taxonomic Classification in Biology	40
1.3.2 Brief history of bacterial taxonomic classification	41

1.3.3	Taxonomic classification in the genomic era	44
1.4	Phylogenetic trees: best-effort attempt to reconstruct evolutionary history	63
1.4.1	Overview of phylogenetic trees	63
1.4.2	Types of clades: monophyletic, paraphyletic and polyphyletic . .	67
1.4.3	Bifurcation <i>versus</i> multifurcation	70
1.4.4	The impact of horizontal gene transfer on phylogenetic inference and the bifurcation model	73
1.5	Using graph theory to analyse bacterial classifications	74
1.5.1	Weighted graphs	77
1.5.2	Disconnected graph	80
1.5.3	Minimum Spanning Tree	82
1.5.4	Hamming distance	86
1.6	The Genus <i>Streptomyces</i> : A Promising Source of Novel Antibiotics . . .	89
1.7	Taxonomic incongruence in the genus <i>Streptomyces</i>	99
1.8	Thesis outline	107
2	16S rRNA phylogeny and clustering is not a reliable proxy for genome- based taxonomy in <i>Streptomyces</i>	110
2.1	Introduction	110
2.1.1	Motivation	110
2.1.2	Public 16S databases	111
2.1.3	Considerations for construction of a 16S phylogenetic tree	112
2.1.4	Best practices	129

2.1.5	Aims and objectives	130
2.2	Methodology	132
2.2.1	Data summary and availability	132
2.2.2	Acquisition of 16S rRNA <i>Streptomyces</i> sequences from major 16S rRNA databases	132
2.2.3	Selection of full-length <i>Streptomyces</i> 16S rRNA sequences	136
2.2.4	LPSN Nomenclature Validation	136
2.2.5	Elimination of nomenclature disagreements at higher taxonomic ranks	137
2.2.6	Removing redundant and ambiguous sequences	137
2.2.7	Clustering of complete 16S rRNA <i>Streptomyces</i> sequences	142
2.2.8	Phylogenetic reconsuction	142
2.2.9	Assessment of unique 16S rRNA sequences from <i>Streptomyces</i> genomes	143
2.2.10	Network analysis of genomes based on shared 16S rRNA sequences	146
2.3	Results and Discussion	146
2.3.1	Public 16S <i>Streptomyces</i> sequence databases include records with low-quality or redundant sequence, or that have issues with taxo- nomic nomenclature	146
2.3.2	16S percentage sequence identity thresholds do not reliably delin- eate existing <i>Streptomyces</i> species assignments	152
2.3.3	A comprehensive <i>Streptomyces</i> 16S phylogeny	157

2.3.4	Whole-genome sequence classification indicates that distinct <i>Streptomyces</i> species can share identical full-length 16S sequences . . .	172
-------	--	-----

3 Updated Multilocus Sequence Typing (MLST) scheme for *Streptomyces* reveals a complex taxonomic structure 192

3.1	Introduction	192
3.1.1	Motivation	192
3.1.2	Aims and Objectives	193
3.2	Methodology	194
3.2.1	Data retrieval and availability	194
3.2.2	Filtration of <i>Streptomyces</i> genomes	195
3.2.3	Identification of novel allele sequences and ST assignment	197
3.2.4	Scheme refinement	198
3.2.5	Visualisation of MLST scheme: Minimum Spanning Tree	199
3.2.6	Genome Quality Assessment	199
3.2.7	Influence of genome sampling on the connectivity of MST	200
3.2.8	Empirical non-parametric network test	201
3.2.9	ANI analysis	202
3.2.10	MLSA Phylogenetic reconstruction from MLST markers	203
3.2.11	Sensitivity test	204
3.2.12	Representation of <i>Streptomyces</i> in sister genera	205
3.3	Results	205
3.3.1	Updated scheme	205

3.3.2	Graph based analysis of STs	215
3.3.3	Connectivity of MST	223
3.3.4	Empirical test	229
3.3.5	Comparing MLST divisions and whole-genome sequence classification landscapes	234
3.3.6	ANIm analysis of genomes assigned the same species designations in NCBI	253
3.3.7	MLSA Phylogeny	257
3.3.8	Sensitivity test	260
3.3.9	Sister genera	268
3.4	Discussion	268
3.4.1	The canonical pubMLST Scheme is incomplete	268
3.4.2	The current set of MLST markers can lead to genus-level misclassifications	273
3.4.3	Graph based analysis of STs subdivides <i>Streptomyces</i> into distinct groups and which likely represent biologically-meaningful divisions	276
3.4.4	MLST subgraphs do not generally correspond to <i>Streptomyces</i> taxa	278
3.4.5	Inconsistencies between MLST divisions and NCBI nomenclature	281
3.4.6	Clades in MLSA phylogeny largely do not correspond with MLST subgraphs	284

4 Genomic insights into *Streptomyces* phylogeny, taxonomy and struc-

ture	286
4.1 Introduction	286
4.1.1 Motivation	286
4.1.2 Aims and Objectives	287
4.2 Methodology	288
4.2.1 Data retrieval and availability	288
4.2.2 Taxonomic sampling	289
4.2.3 Nomenclature status	289
4.2.4 Identification of Orthogroups	290
4.2.5 Single Copy Orthologue phylogeny	290
4.2.6 Testing congruence of ANIm taxonomic boundaries and SCOG phylogeny and identification of genus boundaries	291
4.2.7 SCOG location on the chromosome	293
4.2.8 Distribution of SCOGs on the phylogeny	294
4.3 Results	294
4.3.1 Representative set of genomes	294
4.3.2 General features of <i>Streptomyces</i> pangenome	296
4.3.3 Single Gene Copy Orthologue Phylogeny	298
4.3.4 ANIm reveals significant genomic diversity within <i>Streptomyces</i> .	301
4.3.5 Identification of genus boundaries for <i>Streptomyces</i>	304
4.3.6 Location of SCOGs on chromosome	316
4.3.7 Conservation of SCOGs nucleotide variants across the phylogenetic tree	321

4.4	Discussion	342
4.4.1	Representative set of high-quality <i>Streptomyces</i> genomes reveals a small core genome	342
4.4.2	A highly resolved core genome phylogeny supports conclusions of widespread misclassification in <i>Streptomyces</i>	345
4.4.3	New genus boundary threshold reveals 79 distinct <i>Streptomyces</i> groups	348
4.4.4	Single-Copy Orthologues in <i>Streptomyces</i> predominantly reside in the core chromosomal region with extensions into the arms . . .	353
4.4.5	Phylogenetic distribution of Single-Copy Orthologue nucleotide variants reveals non-monophyletic patterns in <i>Streptomyces</i> . . .	359
5	Conclusions and Recommendations	363
5.1	Conclusions	363
5.2	To rename or not to rename - this is the question	370
5.3	Future work	371
5.3.1	Is horizontal gene transfer occurring in the core genome of <i>Strep- tomyces</i> ?	371
5.3.2	Explore the impact of HGT of single-copy orthologues in the context of <i>Streptomyces</i>	372
5.3.3	Exploring the bioactive potential of <i>Streptomyces</i>	373
A	Appendix 1	375
	Bibliography	376

List of Figures

1.1	Examples of how antibiotic resistance spreads in a population.	4
1.2	Annual mortality estimates	6
1.3	Timeline illustrating the introduction of new antibiotic classes into clinical use.	21
1.4	Timeline of key developments in bacterial genome sequencing.	26
1.5	Steps in DNA sequencing across three generations	27
1.6	Annual submissions of archaeal and bacterial genomes to NCBI.	28
1.7	Chemical structure of humimycin A	30
1.8	The pangenome concept.	33
1.9	Types of homologous genes.	35
1.10	Challenges of ortholog identification.	37
1.11	Concept of Multilocus Sequence Typing	51
1.12	Summary Statistics of GenBank and WGS (Whole Genome Shotgun) sequence submissions in NCBI.	53
1.13	Preliminary data comparing ANI methods using synthetic sequences with known true identities, as presented by Dr. Leighton Pritchard at the genomeRxiv meeting	62

1.14	Structure of evolutionary trees.	64
1.15	Ernst Haeckel Tree of Life.	66
1.16	Types of taxonomic clades.	69
1.17	Bifurcating <i>versus</i> multifurcation tree.	72
1.18	Basic graph concepts.	76
1.19	Visual representation of complete and non-complete graph.	79
1.20	Visual representation of a disconnected graph.	81
1.21	Example of a minimum spanning tree.	83
1.22	Visual representation of algorithms used in finding MSTs.	85
1.23	Example use of Hamming distance calculation between MLST profiles.	88
1.24	<i>Streptomyces</i> life cycle.	91
1.25	Key discoveries of secondary metabolites produced by members of the genus <i>Streptomyces</i>	94
1.26	Most consumed antimicrobial classes by humans and food-producing animals (mg/kg), 2019	95
1.27	Most sold antibiotic classes for food-producing animals in 2022.	96
1.28	Diverse roles of <i>Streptomyces</i> species in agriculture.	98
1.29	Maximum-likelihood tree of 16S rRNA sequences showing the placement of <i>Streptomyces caelicus</i> based on five different copies from the same isolate.	103
1.30	The core-genome phylogeny of 218 <i>Streptomyces</i> species.	106
2.1	Illustrative representation of multiple sequence alignment.	119

2.2	Schematic workflow for construction of the full-length 16S rRNA <i>Streptomyces</i> phylogeny.	135
2.3	The number of ambiguity bases per 16S rRNA sequence	139
2.4	Schematic representation of the pipeline used to filter publicly available <i>Streptomyces</i> genomes.	145
2.5	Sankey plot showing counts of taxonomic names in source 16S databases.	151
2.6	Cluster sizes.	155
2.7	Cluster taxID abundance.	156
2.8	Maximum-likelihood tree of the genus <i>Streptomyces</i> constructed from 9,049 full-length 16S rRNA sequences.	159
2.9	Maximum-likelihood tree of the genus <i>Streptomyces</i> showing branches with transfer bootstrap expectation support of $\geq 50\%$	160
2.10	Maximum-likelihood tree of the genus <i>Streptomyces</i> showing distribution of <i>Streptomyces albulus</i> , <i>Streptomyces lydicus</i> and <i>Streptomyces venezuelae</i>	162
2.11	Maximum-likelihood tree of the genus <i>Streptomyces</i> showing distribution of <i>Streptomyces lavendulae</i> , <i>Streptomyces rimosus</i> and <i>Streptomyces scabiei</i>	163
2.12	Maximum-likelihood tree of the genus <i>Streptomyces</i> showing distribution of <i>Streptomyces albus</i> and <i>Streptomyces griseus</i>	164
2.13	Maximum-likelihood tree of the genus <i>Streptomyces</i> showing distribution of <i>Streptomyces clavuligerus</i> and <i>Streptomyces coelicolor</i>	165
2.14	Distribution of members of the novel <i>Wenjunlia</i> genus on the ML tree. .	167
2.15	Distribution of members of the novel <i>Actinacidiphila</i> genus on the ML tree.	168

2.16	Distribution of members of the novel <i>Mangrovactinospora</i> genus on the ML tree.	169
2.17	Distribution of members of the novel <i>Phaeacidiphilus</i> genus on the ML tree.	170
2.18	Distribution of members of the novel <i>Streptantibioticus</i> genus on the ML tree.	171
2.19	Intragenomic 16S rRNA heterogeneity within 1,369 <i>Streptomyces</i> genomes.	174
2.20	Intragenomic 16S rRNA heterogeneity within <i>Streptomyces</i> genomes at assembly level complete and chromosome.	175
2.21	Names assigned to genomes at NCBI	178
2.22	Network graph with 1369 genomes and 709 connected components. . . .	180
2.23	Heatmaps of ANIm coverage (left), and ANIm identity (right) for three example connected components from Figure 2.22	182
2.24	Network graph with 1369 genomes and 709 connected components. . . .	185
2.25	Network graph with 1369 genomes and 709 connected components. . . .	186
2.26	Scatterplots showing genome coverage for pairwise ANI comparisons for genomes sharing identical full-length and ambiguity base-free 16S sequences.	190
2.27	Scatterplots showing genome identity for pairwise ANI comparisons for genomes sharing identical full-length and ambiguity base free 16S sequences.	191
3.1	Schematic representation of the pipeline used to update the pubMLST <i>Streptomyces</i> scheme.	196
3.2	History of submission of STs for <i>Streptomyces</i> scheme in pubMLST. . .	207
3.3	Overall distribution of NCBI assigned species with novel and existing STs.	209

3.4	MST with 852 STs and 278 connected components describing all <i>Streptomyces</i> genomes, and all STs from the pubMLST database.	216
3.5	Distribution of connected component sizes showing that smaller connected components are more frequent, while larger connected components are less common.	217
3.6	Distribution of connections (degrees) of STs in the MST representation of MLST scheme for <i>Streptomyces</i>	218
3.7	MST of the updated pubMLST <i>Streptomyces</i> scheme showing number of STs connections (degrees).	219
3.8	Minimum Spanning Tree of the updated pubMLST <i>Streptomyces</i> scheme showing GenBank represented STs and non-GenBank represented STs.).	222
3.9	Scatter plot showing the relationship between the number of randomly sampled genomes 10-90% and number of disjoint graphs.	224
3.10	The distribution of relative connected component sizes.	225
3.11	Relationship between the number of randomly sampled genomes 10-90% and number of disjoint graphs from the artificial scheme.	227
3.12	Distribution of relative connected component sizes from randomly samples 10-90% genomes from the artificial scheme.	228
3.13	Genomes sharing identical STs can be assigned different taxonomic names in NCBI.	237
3.14	Scatter plots for genome coverage for pairwise ANIm comparisons for genomes sharing identical STs.	238

3.15	Scatter plots for genome identity for pairwise ANIm comparisons for genomes sharing identical STs.	239
3.16	Exemplar ANIm genome coverage and ANIm identity analysis of genomes found in the same group of connected STs.	241
3.17	ANIm genome coverage analysis of genomes found in the same group of connected STs.	243
3.18	ANIm genome identity analysis of genomes found in the same group of connected STs.	244
3.19	Minimum Spanning Tree of the updated pubMLST <i>Streptomyces</i> scheme, showing unique candidate genera per connected component.	246
3.20	Minimum Spanning Tree of the updated pubMLST <i>Streptomyces</i> scheme showing unique candidate species per connected component.	248
3.21	Multiple STs are used to describe single <i>Streptomyces</i> species ($\geq 50\%$ genome coverage; $\geq 95\%$ genome identity).	249
3.22	Multiple NCBI names are used to describe single <i>Streptomyces</i> species ($\geq 50\%$ genome coverage; $\geq 95\%$ genome identity).	251
3.23	ANIm genome identity for an exemplar case where species designations in NCBI do not match, despite sharing $\geq 95\%$ genome identity.	252
3.24	ANIm coverage analysis of genomes sharing the same name other than <i>Streptomyces sp.</i> in NCBI.	255
3.25	ANIm identity analysis of genomes sharing the same name other than <i>Streptomyces sp.</i> in NCBI.	256

3.26	Midpoint rooted ML tree of six concatenated full-length allele sequences with mapped MST congruence.	259
3.27	Heatmap of the sensitivity test results for <i>16S</i> marker gene.	262
3.28	Heatmap of the sensitivity test results for <i>atpD</i> marker gene.	263
3.29	Heatmap of the sensitivity test results for <i>recA</i> marker gene.	264
3.30	Heatmap of the sensitivity test results for <i>gyrB</i> marker gene.	265
3.31	Heatmap of the sensitivity test results for <i>trpB</i> marker gene.	266
3.32	Heatmap of the sensitivity test results for <i>rpoB</i> marker gene.	267
4.1	Maximum-likelihood tree of concatenated 117 SCOGs sequences with the distribution of genomes currently assigned <i>S. griseus</i> , <i>S. rimosus</i> and <i>S.</i> <i>clavuligerus</i>	300
4.2	Pairwise ANIm comparisons of 295 representative <i>Streptomyces</i> genomes.	303
4.3	Assessment of genome coverage thresholds on the clustering of <i>Strepto-</i> <i>myces</i> and their congruence with the SCOG phylogeny.	306
4.4	Piecewise linear regression with six segments for ANIm comparisons among 295 representative <i>Streptomyces</i> genomes does not yield clear boundary separation.	311
4.5	Piecewise linear regression with four segments for ANIm comparisons among 295 representative <i>Streptomyces</i> genomes reveals a clear boundary at 88.1% genome coverage and 55.6% genome identity.	312

4.6	Maximum-likelihood tree of concatenated 137 SCOG sequences, rooted at the midpoint, with 79 labeled distinct groups, each representing a separate candidate genus within the broader <i>Streptomyces</i> lineage. . . .	315
4.7	OriC location for genomes assembled to complete or chromosomal level in NCBI.	317
4.8	Location of SCOGs on the chromosome for genomes assembled to complete or chromosomal level in NCBI.	320
4.9	Variants of SCOG's nucleotide sequences, identified in two or more <i>Streptomyces</i> species, consistently form monophyletic clades in the core genome phylogenetic tree.	332
4.10	SCOG's repeated nucleotide sequence variants do not form monophyletic clades on the core genome tree.	336
4.11	Repeated nucleotide sequence variants of SCOG do not form monophyletic clades in the core genome tree, but are confined within the same <i>Streptomyces</i> subgroups (with genome coverage of $\geq 45.8\%$ and genome identity of $\geq 88.8\%$).	338
4.12	SCOG's repeated nucleotide sequence variants do not form monophyletic clades on the core genome tree and are shared across genomically distinct <i>Streptomyces</i> subgroups ($>45.8\%$ genome coverage; $>88.8\%$ genome identity).	340
4.13	Tripartite structure of the <i>Streptomyces coelicolor</i> A3(2) chromosome. .	355
A.1	pyANI genome coverage of genomes currently named as <i>S. griseus</i> in NCBI.	375

A.2	pyANI genome coverage of genomes currently named as <i>S. rimosus</i> in	
	NCBI.	376

List of Tables

1.1	Current list of ESKAPE pathogens.	8
1.2	Antimicrobial agents and their targets. Synthetic classes are indicated by *. Adapted from Hutchings et al., 2019.	12
1.3	A summary description of software that calculate ANI and algorithms they use.	59
1.4	Example of <i>Streptomyces</i> species sharing identical or highly identical 16S sequences.	102
2.1	Commonly used sequence quality control tools.	117
2.2	Summary description of 16S rRNA databases used in the study.	134
2.3	Summary statistics for taxonomic composition of subgraphs uniting at least two genome assemblies.	183
3.1	Examples of novel STs with corresponding genome accessions and assigned NCBI names.	208
3.2	An overview of the distribution of missing allele copies across the 1,782 analysed <i>Streptomyces</i> genomes.	211
3.3	The count of genomes with single and multiple identical copies for each marker.	214

3.4	Summary of node degrees for novel and pubMLST STs.	231
3.5	Summary of degree node connections represented and not represented in Genbank.	233
3.6	Five larges STs with representatives in GenBank and their corresponding taxonomic assignments.	236
3.7	Summary of the taxonomic composition of subgraphs uniting at least two genome assemblies.	242
3.8	Summary of sensitivity test showing number of unique STs represented in GenBank after the exclusion of each marker.	261
4.1	Summary statistics of <i>Streptomyces</i> pangenome.	297
4.2	Summary statistics for identification of optimal number of segments for piecewise linear regression.	309
4.3	Lists 52 SCOGs with at least one nucleotide variant shared by two or more <i>Streptomyces</i> genomes. It includes the total number of unique nucleotide sequence variants, the corresponding protein products (as annotated in the sequence records), and the number of conserved nucleotide sequence variants categorized as either monophyletic or non-monophyletic in the SCOG tree.	322

4.4	List of 51 SCOGs in which "repeated" nucleotide sequence variants are shared across distinct <i>Streptomyces</i> genomes and form monophyletic clades. The corresponding protein products are annotated as per the sequence records. The order of the SCOGs aligns with the arrangement of the rings in Figure 4.9, from inner to outer.	333
4.5	List of 38 SCOGs in which at least one repeated nucleotide variants are shared across distinct <i>Streptomyces</i> genomes and form non-monophyletic clades with their corresponding protein products as annotated in the sequence records. The order of the SCOGs corresponds to the order (from inner to outer) of the rings in the Figure 4.10.	337
4.6	List of 38 SCOGs, each containing at least one repeated nucleotide variant shared within the same <i>Streptomyces</i> subgroup and forming non-monophyletic clades. Corresponding protein products, as annotated in the sequence records, are also included. The order of the SCOGs follows the arrangement of rings (from inner to outer) in Figure 4.11.	339
4.7	List of 16 SCOGs in which repeated nucleotide variants are shared across distinct <i>Streptomyces</i> genomes and form non-monophyletic clades with their corresponding protein products as annotated in the sequence records. The order of the SCOGs corresponds to the order (from inner to outer) of the rings in the Figure 4.12.	341

General Introduction

1.1 Antimicrobial Resistance Crisis

Antimicrobial compounds have revolutionised modern medicine (Ventola, 2015). They are highly effective at killing bacteria, making them indispensable for treating acute bacterial infections (Nemeth et al., 2015). However, their effectiveness and, in some cases, low cost have led to their widespread and sometimes inappropriate prophylactic use to prevent the likelihood of infections. Antimicrobials were successfully introduced to assist complex cardiac surgeries, joint replacements and organ transplants (Najjar & Smink, 2015); to prevent infections in patients diagnosed with HIV or diabetes, and those that receive chemotherapy (Ventola, 2015); and to control recurrent urinary tract infections (Marschall et al., 2013). Moreover, antimicrobial agents are now widely used in agriculture for the treatment and prevention of infections in livestock, and to promote growth and improve feed efficiency (Manyi-Loh et al., 2018).

The overuse and misuse of antimicrobials in food, medicine, and agriculture have contributed to the rise of antimicrobial resistance (AMR) (Ventola, 2015) (Figure 1.1). While antimicrobial resistance genes naturally exist in environmental bacteria, they can be transferred to clinically relevant strains through horizontal gene transfer.

Additionally, spontaneous mutations may lead to resistance mechanisms, such as target site modifications, allowing bacteria to evade antimicrobial effects (Hughes & Andersson, 2017). When antibiotics are introduced into an environment, they create selective pressure that favors bacteria carrying resistance genes (Uddin et al., 2021). Over time, these resistant bacteria proliferate, eventually dominating the population.

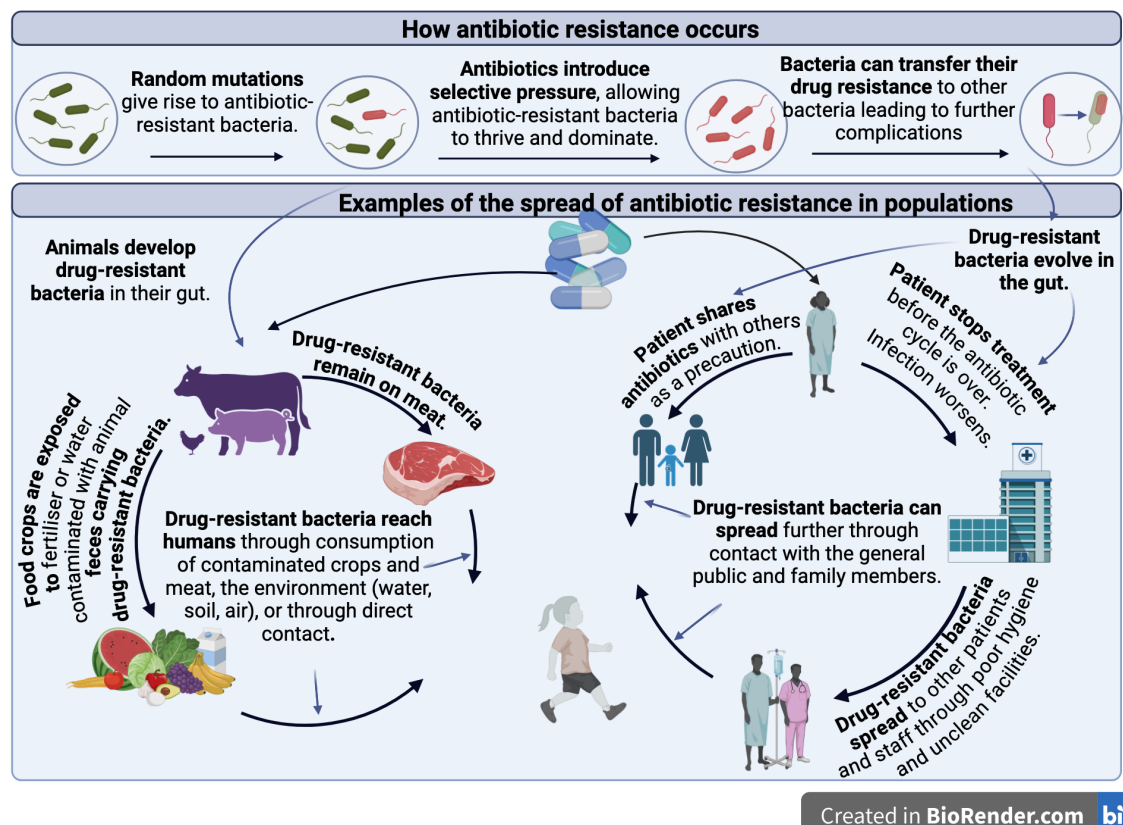


Figure 1.1: Examples of how antibiotic resistance spreads in a population. Figure adapted from Centers for Disease Control and Prevention (CDC) et al., 2013. When antibiotics are prescribed to treat acute infections in humans, stopping the course of antibiotics before the cycle is over can result in the survival and proliferation of bacteria, worsening the infection. This can result in patients needing hospital admission for more intensive treatment. In healthcare settings, inadequate hygiene and contaminated surfaces can facilitate the transmission of infections between patients, potentially spreading them to the broader community (Michael et al., 2014). Moreover, patients may occasionally share antibiotics with others as a precautionary measure, despite medical advice against this practice, which can contribute to the emergence of drug-resistant bacteria (Alhomoud et al., 2017). These resistant bacteria can then propagate through human contact, posing significant public health challenges (Michael et al., 2014).

Animals treated with antibiotics, either to cure infections or as a preventive measure, can evolve resistant bacteria in their gut flora (Upadhayay & Vishwa, 2014). These resistant bacteria can enter the environment through animal waste, when used as fertilisers exposing food crops to antibiotic-resistant bacteria (Founou et al., 2016). Additionally, resistant bacteria can persist on meat products, potentially transferring to humans through the consumption of contaminated food or direct contact with infected livestock (Founou et al., 2016).

The global crisis of AMR has emerged as a major threat to public health. The number of global direct deaths caused by antibiotic resistant pathogenic bacteria increased from 700,000 in 2014 (O'Neill, 2014) to 1.27 million in 2019 (Murray et al., 2022a). If left unchecked, AMR-related deaths are predicted to surpass the current annual mortality rate caused by cancer, reaching a staggering 10 million deaths per year by 2050 worldwide (Allcock et al., 2017) (Figure 1.2). Current annual worldwide costs of AMR are estimated to be around \$300 billion and are projected to increase to \$1 trillion in the next 30 years (Chokshi et al., 2019; WorldBankGroup, 2016). Infections caused by multi-drug resistant bacteria can no longer be treated with first-line antibiotics, leading to more expensive medicines being prescribed, prolonged duration of treatment, and hospital stays which cause additional problems for overstretched healthcare systems (Porooshat, 2019). Urgent action is required to prevent the spread of drug-resistant microbes. The loss of productivity caused by infections, prolonged hospital admissions, and premature death would negatively impact Gross Domestic Product (GDP). It is projected that this impact could lead to a decrease of 5-7% in GDP in developing countries by 2050 (Porooshat, 2019).

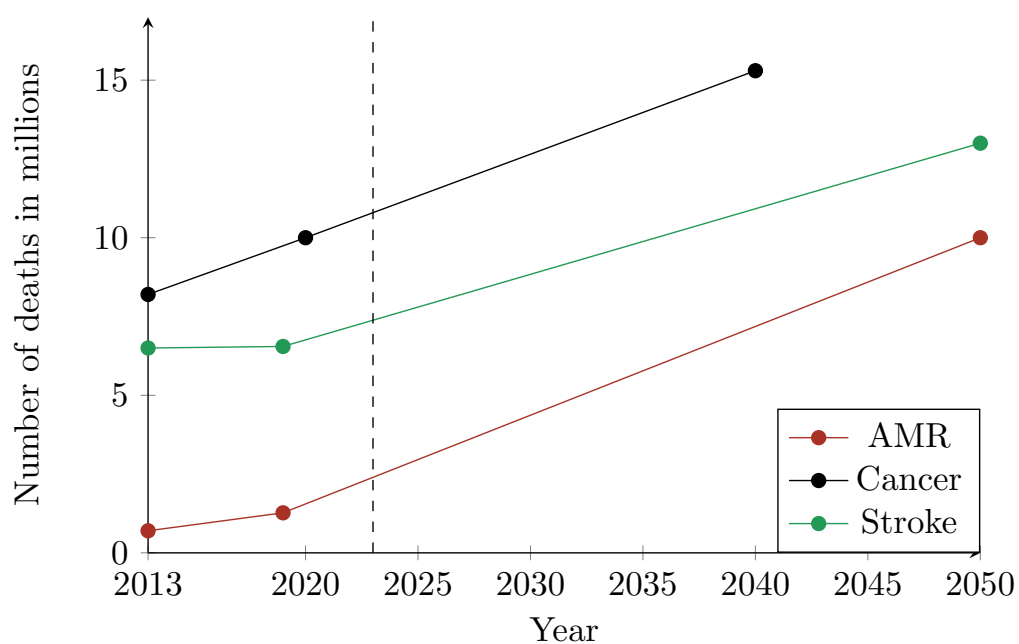


Figure 1.2: Estimates of annual number of deaths from cancer, strokes, and AMR from 2014 to 2050. The black vertical dashed line represents the current year. The data was sourced from Feigin et al., 2015; Feigin et al., 2021; Fitzmaurice et al., 2015; Murray et al., 2022a; O'Neill, 2014; Sung et al., 2021 and Allcock et al., 2017.

Bacteria can acquire AMR genes by horizontal transfer (Normark & Normark, 2002), which poses an issue especially in hospitals where bacteria gain a strong advantage from resistance as a result of being under selective pressure from multiple antibiotics (Evans et al., 2020). In environments with persistent antibiotic selection pressure, such as hospitals, bacteria may harbour large reservoirs of AMR genes. This creates conditions where AMR genes can be readily acquired by pathogens when they are present, potentially leading to treatment challenges (Kaplan, 2014; Woodford & Ellington, 2007). Among the most concerning are the ESKAPE pathogens, a group of bacteria known for their ability to "escape" the killing mechanisms of antibiotics (Rice, 2008). ESKAPE pathogens are particularly problematic in healthcare settings due to their high levels of resistance and association with severe infections. The ESKAPE pathogens were associated with more than 2.5 million deaths in 2019, contributing to both direct fatalities and indirect complications that led to death (Table 1.1).

Table 1.1: Current list of ESKAPE pathogens, their common resistances, and the number of related deaths in 2019.

Pathogen	Common Resistance	Estimated global deaths (2019)	Reference
<i>Enterococcus faecium</i>	vancomycin ampicillin	200,000	Coll et al., 2024 Murray et al., 2022b
<i>Staphylococcus aureus</i>	methicillin clindamycin	750,000	Lee et al., 2018 Murray et al., 2022b Drinkovic et al., 2001
<i>Klebsiella pneumoniae</i>	imipenem ertapenem	650,000	Navon-Venezia et al., 2017 Murray et al., 2022b
<i>Acinetobacter baumannii</i>	meropenem imipenem amoxicillin ticarcillin	450,000	Wu et al., 2023b Murray et al., 2022b
<i>Pseudomonas aeruginosa</i>	ceftazidime cefepime piperacillin gentamicin tobramycin amikacin	350,000	Riera et al., 2011 Murray et al., 2022b
<i>Enterobacter</i> spp.	ampicillin amoxicillin ticarcillin cephalothin polymyxin	180,000	Davin-Regli et al., 2019 Murray et al., 2022b Nang et al., 2019

Modern medicine heavily relies on effective antibiotic treatment both prophylactically to prevent infections and therapeutically to combat acute infections. If we lose effective antibiotics, we risk regressing to a time before their discovery. Acute infections, which are currently manageable with antibiotics, would once again become major causes of mortality. Additionally, without effective antibiotics, modern medical interventions including transplants and complex cardiac surgeries would have high mortality rates from infections (Najjar & Smink, 2015).

Given the current threat to the effectiveness of antibiotics and the declining discovery and development of novel medications (Hutchings et al., 2019), there is an urgent need for new methods and approaches to accelerate the discovery of antibiotics.

1.2 A brief history of natural bioactive compound discovery

1.2.1 Empirical drug discovery

Mankind's use of natural compounds with positive effect on health dates back to ancient times, and various plants and herbs have long been used for medical purposes. The use of medicinal plants likely evolved through an informal, empirical process of trial and error. Effective treatments, identified through observations of their effects on health, were retained and passed down through generations based on their observed properties. This is evident in common names like "deadly nightshade", "pissenlits," and "fever bush," which reflect their observed effects on health and have been passed down through generations. Herbs and plants, such as *Curcuma longa* (turmeric) (Akaberi et al., 2021), *Cannabis sativa* (marijuana) (Bridgeman & Abazia, 2017) and *Papaver somniferum* (opium poppy - a source of morphine); (Brook et al., 2017) which were commonly used

in ancient medicine, are still recognised for their potential health benefits.

Empirical exploration of medical plants and herbs was widespread through the Middle Ages. In the 19th Century Paul Ehrlich's discovery that certain chemical dyes selectively stained bacterial cells led him to hypothesise that bacteria could produce molecules capable of selectively killing other bacteria without harming other cells (Valent et al., 2016). This hypothesis was later confirmed by accidental discovery of the antibiotic called penicillin by Alexander Fleming in 1928 (Fleming, 1980). This event sparked the beginning of the Golden Age of naturally occurring antibiotics that peaked in the 1950s and extended until the late 1960s, during which many novel antibiotics were discovered (Figure 1.3). During this Golden Age, 20 of the 38 currently known antibiotic classes were identified (Coates et al., 2011; Hutchings et al., 2019, Table 1.2).

Table 1.2: Antimicrobial agents and their targets. Synthetic classes are indicated by *. Adapted from Hutchings et al., 2019.

Class	Example	Producing Organism	Target Organism	Molecular Target	Reference
Aminoglycosides	streptomycin	<i>Streptomyces griseus</i>	<i>Mycobacterium avium</i>	Protein synthesis: 30S ribosomal subunit	Waksman et al., 1946
Tetracyclines	oxytetracycline	<i>Streptomyces rimosus</i>	<i>Mycoplasma pneumoniae</i>	Protein synthesis: 30S ribosomal subunit	Petkovic et al., 2006
Amphenicols	chloramphenicol	<i>Streptomyces venezuelae</i>	<i>Staphylococcus aureus</i>	Protein synthesis: 50S ribosomal subunit	Mosher et al., 1995
Macrolides	erythromycin A	<i>Saccharopolyspora erythraeus</i>	<i>Brucella</i> spp.	Protein synthesis: 50S ribosomal subunit	Galvidis et al., 2015
Tuberactinomycins	viomycin	<i>Streptomyces puniceus</i>	<i>Mycobacterium tuberculosis</i>	Protein synthesis: 30S and 50S ribosomal subunits	Finlay et al., 1951

Continued on next page

Table 1.2 – Continued from previous page

Class	Example	Producing Organism	Target Organism	Molecular Target	Reference
Glycopeptides	vancomycin	<i>Amycolatopsis orientalis</i>	<i>Staphylococcus aureus</i>	Cell wall synthesis: D-Ala-D-Ala termini of lipid II	Xu et al., 2014
Lincosamides	lincomycin	<i>Streptomyces lincolnensis</i>	<i>Staphylococcus aureus</i>	Protein synthesis: 50S ribosomal subunit	Du et al., 2012
Ansamycins	rifamycin	<i>Amycolatopsis mediterranei</i>	<i>Mycobacterium tuberculosis</i>	Nucleic acid synthesis: RNA polymerase	Venkateswarlu et al., 1999
Cycloserines	seromycin	<i>Streptomyces garyphalus</i>	<i>Mycobacterium tuberculosis</i>	Cell wall synthesis: inhibition of alanine racemase	Kumagai et al., 2010
Streptogramins	virginiamycin	<i>Streptomyces virginiae</i>	<i>Staphylococcus</i> spp.	Protein synthesis: 50S ribosomal subunit	Mast and Wohlleben, 2014

Continued on next page

Table 1.2 – Continued from previous page

Class	Example	Producing Organism	Target Organism	Molecular Target	Reference
Phosphonates	Fosfomycin	<i>Streptomyces fradiae</i>	<i>Staphylococcus aureus</i>	Cell wall synthesis: MurA inhibition	Díez-Aguilar and Cantón, 2019
Carbapenems	thienamycin	<i>Streptomyces cattleya</i>	<i>Pseudomonas</i> spp.	Cell-wall biosynthesis inhibitor	Rodríguez et al., 2011
Lipopeptides	daptomycin	<i>Streptomyces roseosporus</i>	<i>Streptococcus pneumoniae</i>	Cell wall: cell membrane disruption	Miao et al., 2005
Liparmycins	tiacumicin B	<i>Dactylosporangium aurantiacum</i>	<i>Clostridium difficile</i>	Nucleic acid synthesis: RNA polymerase	De Simeis and Serra, 2021
Polypeptides	actinomycin D	<i>Streptomyces parvulus</i>	<i>Staphylococcus aureus</i>	Inhibits RNA synthesis	Lee et al., 2016b
Bacitracin	Bacitracin A	<i>Bacillus subtilis</i>	<i>Staphylococcus</i> spp.	Cell wall synthesis: inhibition of dephosphorylation	Johnson et al., 1945

Continued on next page

Table 1.2 – Continued from previous page

Class	Example	Producing Organism	Target Organism	Molecular Target	Reference
Polymyxins	polymyxin B	<i>Bacillus polymyxa</i>	<i>Klebsiella pneumoniae</i>	Outer cell membrane disruption	Zavascki et al., 2007
Mupirocin	mupirocin	<i>Pseudomonas fluorescens</i>	<i>Haemophilus influenzae</i>	Protein synthesis: isoleucyl t-RNA synthetase	Sutherland et al., 1985
Monobactams	nocardicin A	<i>Nocardia uniformis</i>	<i>Pseudomons</i> spp.	Bacterial cell wall integrity and synthesis	Aoki et al., 1976
Penicillins	penicillin	<i>Penicillium notatum</i>	<i>Escherichia coli</i>	Cell wall synthesis: penicillin-binding proteins	Toghueo and Boyom, 2020
Fusidic acid	Fusidic acid	<i>Fusidium coccineum</i>	<i>Staphylococcus aureus</i>	Protein synthesis: elongation factor G	Curbete and Salgado, 2016
Enniatins	enniatin B	<i>Fusarium tricinctum</i>	<i>Candida albicans</i>	Cell wall: cell membrane disruption	De Felice et al., 2023

Continued on next page

Table 1.2 – Continued from previous page

Class	Example	Producing Organism	Target Organism	Molecular Target	Reference
Cephalosporins	Cephalosporin C	<i>Acremonium chrysogenum</i>	<i>Streptococcus pneumoniae</i>	Cell wall synthesis: penicillin-binding proteins	Lin and Kück, 2022
Pleuromutilins	Retapamulin	<i>Clitopilus scyphoides</i>	<i>Streptococcus pyogenes</i>	Protein synthesis: 50S ribosomal subunit	Yang and Keam, 2008
Arsphenamines*	melarsoprol	NA	<i>Trypanosoma brucei</i>	Disruption of Cellular Membranes	Alibu et al., 2006
Sulfonamides*	Mafenide	NA	<i>Pseudomonas</i> spp.	Folate synthesis: inhibition of dihydropteroate synthetase	Acaban et al., 2024
Salicylates*	Para- aminosalicylic acid	NA	<i>Mycobacterium tuberculosis</i>	Folate synthesis: prodrug that inhibits dihydrofolate reductase	Chakraborty et al., 2013

Continued on next page

Table 1.2 – Continued from previous page

Class	Example	Producing Organism	Target Organism	Molecular Target	Reference
Sulfones*	Dapsone	NA	<i>Pneumocystis carinii</i>	Folate synthesis: inhibition of dihydropteroate synthetase	Hughes, 1998
Pyridinamides*	Isoniazid	NA	<i>Mycobacterium tuberculosis</i>	Cell wall: prodrug that inhibits the synthesis of mycolic acids	Timmins and Deretic, 2006
Nitrofurans*	Nitrofurantoin	NA	<i>Escherichia coli</i>	DNA synthesis: DNA damage	Munoz-Davila, 2014
Azoles*	Metronidazole	NA	<i>Gardnerella vaginalis</i>	DNA synthesis: DNA damage	Löfmark et al., 2010
(Fluoro)quinolones*	Ciprofloxacin	NA	<i>Pseudomonas aeruginosa</i>	DNA synthesis: inhibition of DNA gyrase, and topoisomerase IV	Scully et al., 1986

Continued on next page

Table 1.2 – Continued from previous page

Class	Example	Producing Organism	Target Organism	Molecular Target	Reference
Diaminopyrimidines*	Trimethoprim	NA	<i>Pneumocystis carinii</i>	Folate synthesis: inhibition of dihydrofolate reductase	Smilack, 1999
Ethambutol*	Ethambutol	NA	<i>Mycobacterium avium</i>	Cell wall: arabinosyl transferase inhibition	Kim et al., 2019
Thioamides*	Ethionamide	NA	<i>Mycobacterium tuberculosis</i>	Cell wall: prodrug that inhibits the synthesis of mycolic acids	DeBarber et al., 2000
Phenazines*	Clofazimine	NA	<i>Mycobacterium leprae</i>	DNA synthesis: binds to guanine bases	Cholo et al., 2012
Oxazolidinones*	Linezolid	NA	<i>Enterococcus faecium</i>	Protein synthesis: 50S ribosomal subunit	Hashemian et al., 2018

Continued on next page

Table 1.2 – Continued from previous page

Class	Example	Producing Organism	Target Organism	Molecular Target	Reference
Diarylquinolines*	Bedaquiline	NA	<i>Mycobacterium tuberculosis</i>	ATP synthesis: proton pump inhibition	Mahajan, 2013

Production of bioactive secondary metabolites enhances the biological fitness of the producing organisms (Williams et al., 1989). These compounds are synthesised as part of many processes including defence mechanisms, nutrient acquisition processes, and chemical communication strategies. They are encoded by clusters of genes known as biosynthetic gene clusters (BGCs), which typically range in size from approximately 6.2kb (Wuisan et al., 2021) to 249.7kb (Hou et al., 2023). Genes found in BGCs encode enzymes responsible for synthesis, assembly, resistance, regulation and transport (Medema et al., 2015). Because molecules synthesised by these BGCs provide selective advantage to the host, they are often spread in populations through the mechanisms of HGT (Fischbach et al., 2008; Jenke-Kodama & Dittmann, 2009). As gene clusters contain entire metabolic pathways or functional modules, enabling a single transfer event to potentially move all necessary components—such as enzymes, regulatory proteins, and transport mechanisms—required for expressing a new trait in the recipient organism. This comprehensive transfer increases the likelihood that the new trait will be active, with the opportunity to be both functional and advantageous. In particular, successful genetic exchange of BGCs between organisms can allow bacteria to exploit new ecological niches and adapt to varying environmental conditions (Fischbach et al., 2008; Ochman et al., 2005).

There are several diverse structural classes of BGCs of which polyketide synthases (PKS) and non-ribosomal peptide synthases (NRPS) attract the most attention due to their significant pharmaceutical impact. They have been found to possess antimicrobial, anticancer, antifungal, immunosuppressant, and antiparasitic properties (Baltz, 2021;

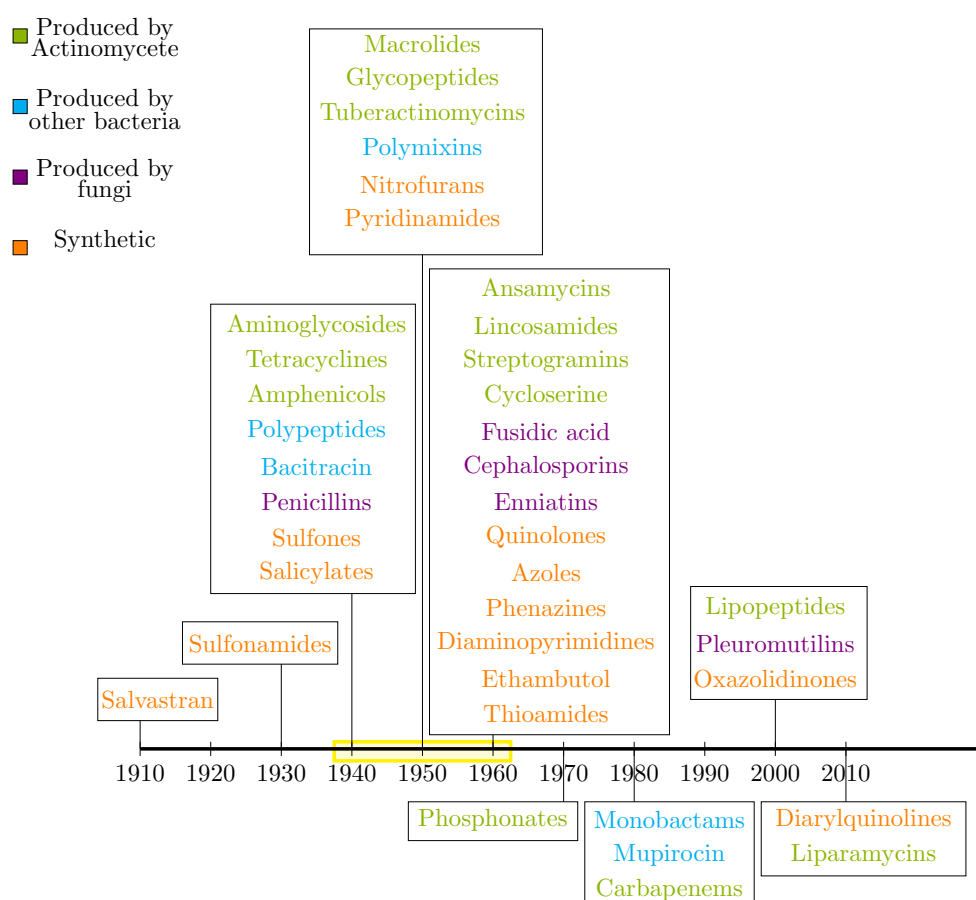


Figure 1.3: Timeline illustrating the introduction of new antibiotic classes into clinical use. The 'Golden Age' of antibiotics is highlighted in yellow, and the source of each antibiotic class is shown in a different colour. Figure adapted from Hutchings et al., 2019.

Buermans & Dunnen, 2014; Martínez-Núñez & López, 2016). The retention of BGCs is costly due to their size, and is assumed to be justified by providing a benefit of some sort to the organism.

Although many microbes including bacteria and fungi synthesise secondary metabolites, the primary source of naturally occurring bioactive compounds currently used in clinical applications is a single group of filamentous bacteria, the *Actinomycetes* (Simeis & Serra, 2021). This prolific production of secondary metabolites is attributed to their high recombination rates and frequent HGT (Ziemert et al., 2014). Such genetic flexibility and plasticity enables *Actinomycetes* to thrive in diverse environments, including soil, aquatic systems, and extreme habitats (Chevrette et al., 2019c; Seipke et al., 2012; Sivalingam et al., 2019; Ziemert et al., 2014). This adaptability also drives the extensive diversification of BGCs within this phylum (Ziemert et al., 2014).

Similarly to the search for plant-derived medicines, the discovery and identification of compounds produced by microorganisms mostly involves empirical approaches. In the Golden Age, this exploration focused on identifying compounds exhibiting desired pharmaceutical behavior *in vitro* (Singh & Barrett, 2006a; Swinney, 2020) and determining their efficacy, mechanism of action, physiochemical properties and spectrum of activity (Fidock et al., 2004; Singh & Barrett, 2006b). However, identifying compounds with *in vitro* activity is not sufficient for developing effective medicines; additional factors such as bioavailability, biodistribution, formulation potential, and toxicity must also be considered to ensure the compound's suitability and safety for human consumption (Blomme & Will, 2016).

The empirical search for candidate therapeutics is a complex, and time consuming

process, as it relies on extensive trial-and-error laboratory work which often leads to re-discoveries of already known antibiotics (Katz & Baltz, 2016). Rediscovery, the repeated identification of known compounds, is problematic because it wastes valuable resources and time without yielding new therapeutics. Traditional empirical methods screen strains without knowing in advance whether they produce new or existing compounds. Many strains may be screened, with many rediscoveries of known secondary metabolites, before a new potential therapeutic is identified. Genomics based approaches offer a promising solution by predicting the biosynthetic potential of strains before experimental screening, and help exclude those unlikely to produce new compounds (for more details see section 1.2.2). Taxonomy plays a significant role in this context, as related organisms often share similar genetic pathways that can lead to the discovery of novel bioactive compounds. By understanding the taxonomic relationships among microorganisms, we can more effectively target strains with the highest likelihood of producing novel compounds.

Further limitations arise from our inability to cultivate some microorganisms and synthesise bioactive compounds in typical laboratory conditions (Li & Vederas, 2009), which makes empirical drug discovery less effective and more difficult. This challenge is particularly problematic for organisms that typically live in extreme environments, for which their unique nutrient requirements cannot be replicated *in vitro*. Genomics can help overcome these obstacles by identifying nutrient requirements for fastidious organisms. Comparative genomics can also identify genes involved in the production of bioactive compounds with potential novel activity. These genes can be transferred into model organisms that are more suitable to laboratory cultivation and study, facilitating

the production and characterisation of their bioactive products (Shi et al., 2019). Despite its limitations, empirical drug discovery still remains widely used for discovering new bioactive compounds today (Moffat et al., 2017).

1.2.2 The influence of genomics on drug discovery

1.2.3 A brief history of genome sequencing

Since the inception of genomic sequencing, the field has undergone remarkable advancements that have impacted bacterial genomics (Figure 1.4), such as bacterial taxonomy and drug discovery. Accurate taxonomic classification often relies on genomic sequencing to distinguish bacterial species, resolve phylogenetic relationships, and refine microbial taxonomy (see Section 1.3.3). Beyond taxonomy, sequencing has also been instrumental in antibiotic discovery (see Section 1.2.3). Comparative genomics enables researchers to identify biosynthetic gene clusters (BGCs) responsible for producing antimicrobial compounds (Chu et al., 2016).

The journey of genome sequencing began with the sequencing of the first bacterial genome in 1995: the 1.8Mb genome of *Haemophilus influenzae*, achieved using the Sanger method that took over one year to complete at a cost of approximately one million US dollars (Fleischmann et al., 1995). Early methods like Sanger sequencing (Figure 1.5), the first generation of sequencing technology, laid the foundation by enabling the reading of DNA sequences with high accuracy. However, this process was labour intensive and costly, limiting its application for large-scale genomic projects (Land et al., 2015a).

The advent of second-generation sequencing technologies (Figure 1.5), also known as next-generation sequencing (NGS), revolutionised the field by significantly lowering costs and increasing data output. Techniques such as Illumina sequencing allowed for

the parallel sequencing of millions of DNA fragments, facilitating large-scale projects like collection of genotyping data for the International HapMap Project and making whole-genome sequencing more accessible (Steemers & Gunderson, 2005). According to Illumina Inc., the cost of sequencing a human genome dropped from \$150,000 in 2007 to around \$200 in 2023 (data available [here](#)). Similarly, bacterial genome sequencing costs plummeted from about \$50,000 per genome in the early 2000s to approximately \$1 per draft genome today (Land et al., 2015b).

More recently, third-generation sequencing technologies, including single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio) and nanopore sequencing by Oxford Nanopore Technologies, have further made sequencing more efficient, affordable, and accessible. These technologies offer much longer read lengths, which are particularly beneficial for assembling complex genomes and detecting structural variations (Deamer et al., 2016; Rhoads & Au, 2015). Additionally, they enable real-time sequencing, providing rapid results that are invaluable in clinical and research settings.

These continuous improvements and reductions in costs have made sequencing bacterial genomes obligatory for almost any research team. It has also led to an unprecedented growth of sequence data (Figure 1.6), profoundly impacting modern microbiology, particularly bacterial genomics (Buermans & Den Dunnen, 2014). The combination of advancements in sequencing technology and bioinformatics-driven analyses has facilitated the discovery of antibiotics through comparative genomics (see section 1.2.3), enabled the development of a more robust and accurate taxonomic framework (see Section 1.3.3 for detailed information), and enhanced our understanding of bacterial evolution and interactions.

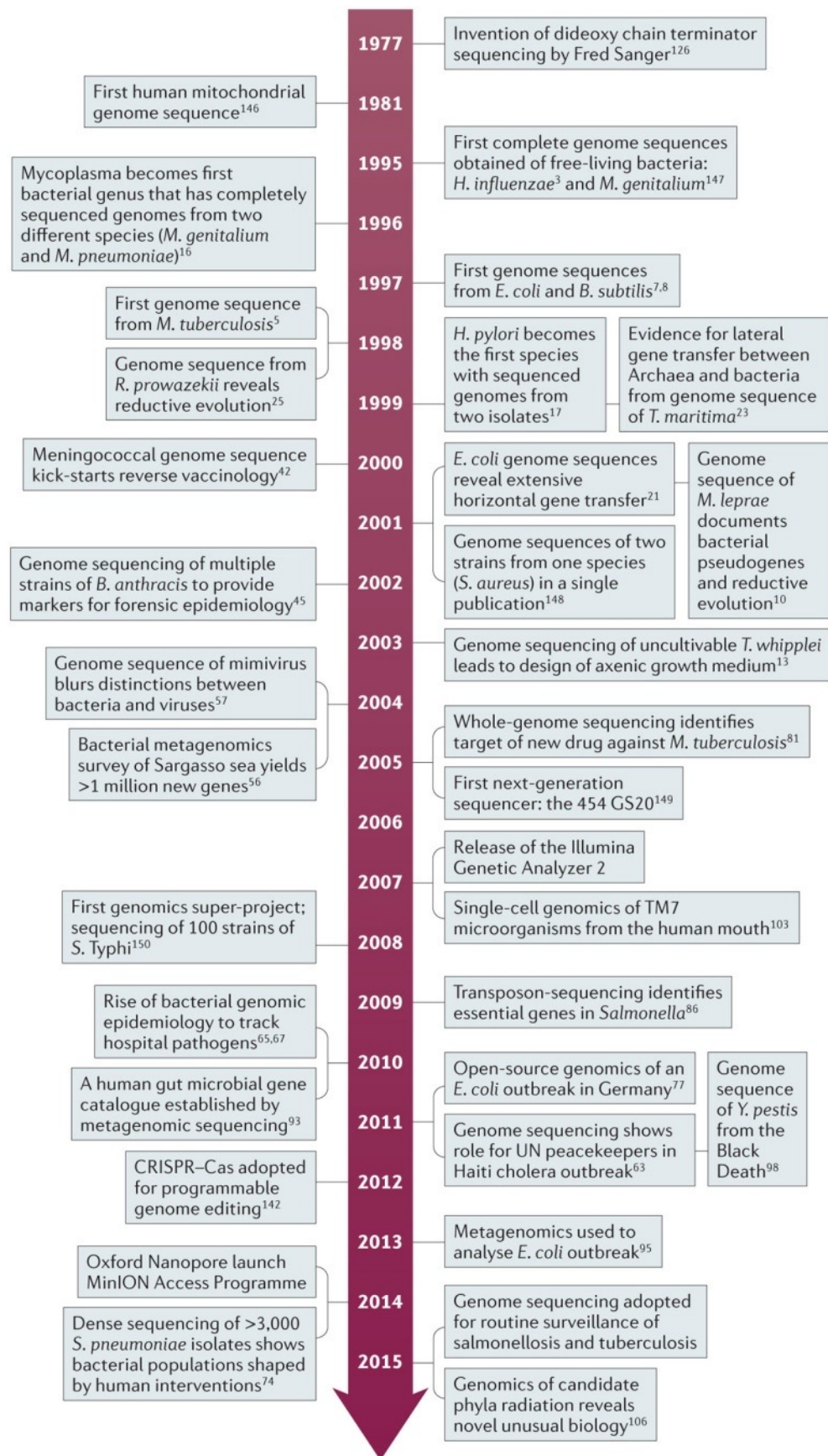


Figure 1.4: Timeline of key developments in bacterial genome sequencing. Figure reproduced from Loman and Pallen, 2015.

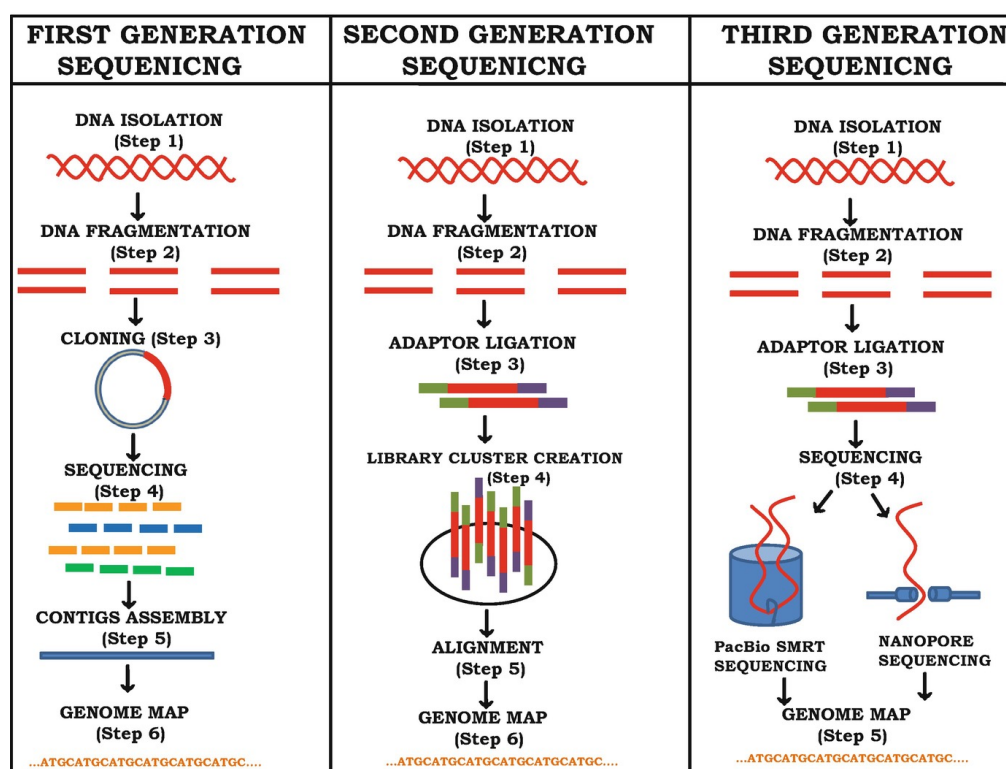


Figure 1.5: Steps in DNA sequencing across three generations. First-generation sequencing begins with DNA isolation and fragmentation, followed by cloning the fragments into vectors for amplification. The Sanger method is then used for sequencing, involving extension and termination of DNA fragments, which are separated by gel electrophoresis and read to assemble contigs and create a genome map. In second-generation sequencing, after DNA isolation and fragmentation, adaptors are ligated to the fragments. These fragments form clusters through amplification on a solid surface. High-throughput sequencing produces millions of short reads, which are aligned to a reference genome or assembled *de novo* to form contigs and a detailed genome map. Third-generation sequencing also starts with DNA isolation, fragmentation, and adaptor ligation, but it uses single-molecule technologies like nanopore or SMRT sequencing. These methods produce long reads directly from single DNA molecules without amplification, allowing for immediate alignment and the creation of a highly accurate genome map, revealing complex structural variations and providing deep genomic insights. Figure reproduced from Srivastav and Suneja, 2019.

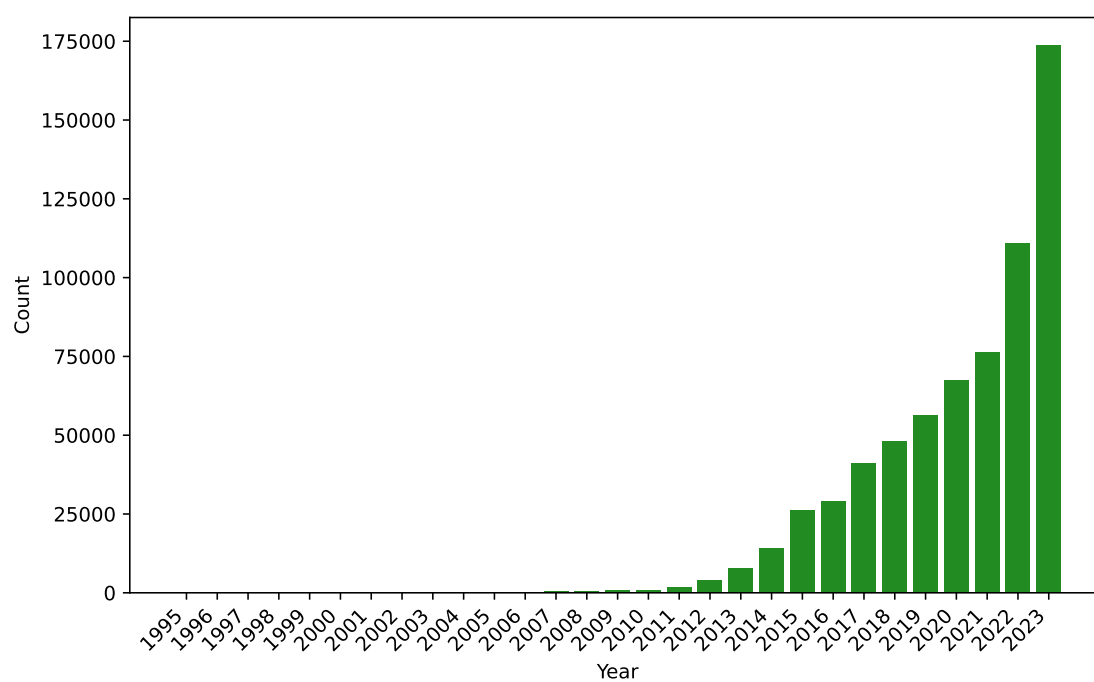


Figure 1.6: Annual submissions of archaeal and bacterial genomes to NCBI. Data sourced from [NCBI genome reports](#) downloaded on the 28th of July 2024.

Genome mining of biosynthetic gene clusters

Advances in sequencing technology and the increase in available genomic sequences enable more powerful approaches for drug discovery, potentially without the need to cultivate microorganisms or isolate compounds in laboratory settings (Buermans & Dunnen, 2014). These approaches involve identifying biosynthetic gene clusters (BGC) responsible for the synthesis of antibiotics in genomic sequences. A number of bioinformatic tools for genome mining such as antiSMASH (Medema et al., 2011), SMURF (Khaldi et al., 2010), GECCO (Carroll et al., 2021), and PRISM (Skinnider et al., 2017) have been developed. These tools, together with increased number of publicly-available genomes, are expected to accelerate discovery of novel antibiotics and avoid rediscovery of already-known secondary metabolites (Baltz, 2021). An example of successful genome mining is the identification of humamycins using antiSMASH (Figure 1.7), which were found to be active against methicillin-resistant *Staphylococcus aureus* *in vitro* (Chu et al., 2016).

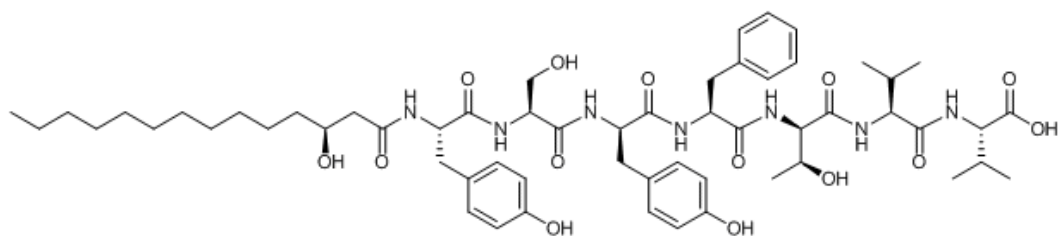


Figure 1.7: Chemical structure of humimycin A. Figure taken from Chu et al., 2016.

1.2.4 Drug discovery through pangenomic analyses

Pangenome analysis is emerging as a powerful tool for drug discovery and development, as well as for studying bacterial populations and microbial ecology. The concept of the bacterial pangenome was first introduced in 2005 in a paper detailing the complete gene composition of closely related isolates of *Streptococcus agalactiae* (Tettelin et al., 2005).

The pangenome of a group of organisms contains the complete set of genes present across those organisms (Figure 1.8). The concept divides genes into core and accessory categories. Core genes are those found in all (or nearly all) individuals within a group, while accessory genes are present in only a subset of organisms (Chung et al., 2018). The bacterial core genome is crucial for understanding relatedness among isolates, and even defining membership of an organism group, by identifying conserved and common genes (Chung et al., 2018). Accessory genes, on the other hand, can reveal unique characteristics of individual organisms or subgroups of organisms, as they provide the genetic flexibility essential for environmental adaptation and the regulation of specific metabolic pathways (Costa et al., 2020b). However, others argue that the traditional core-accessory classification of the pangenome oversimplifies its complexity by not accounting for population structure and biased genome sampling (Horesh et al., 2021). Genes may have vastly different evolutionary and ecological significance depending on lineage representation, which traditional methods fail to capture. To address this, a population structure-aware approach has been proposed to refine gene classification, providing a more accurate understanding of pan-genome evolution and its underlying

selective forces (Horesh et al., 2021).

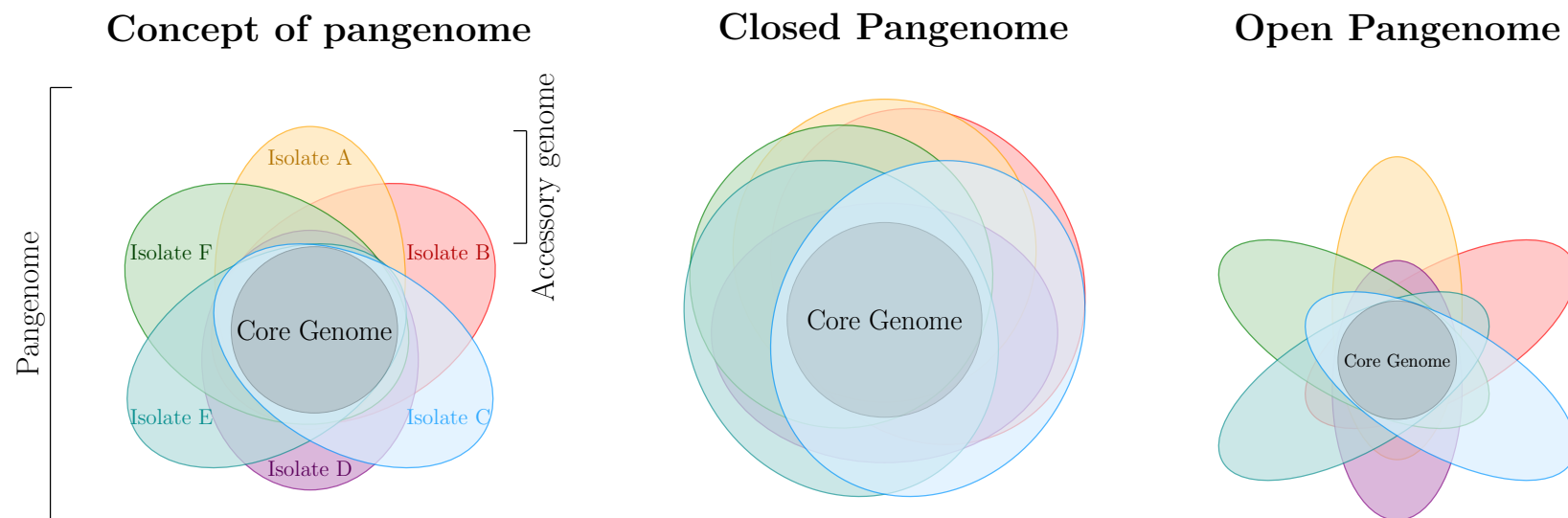


Figure 1.8: The pangenome concept. A pangenome is the entire gene composition of all strains within a phylogenetic clade. A pangenome consists of core genes that are present in all (or nearly all) strains, and accessory genes, that are present only in one genome, or a subgroup of genomes. A pangenome is described as open if the core genome is small and accessory genomes are large, and as closed if it consists of a large core genome and small accessory genomes. Figure adapted from McInerney et al., 2017

To fully comprehend core and accessory genes, it is important to understand how genes with varying sequences across different organisms can be considered related. The term "homologous" is central to this understanding. Pangenomes are predicted or calculated based on the assumption that clustered sequences are homologous, meaning they are derived from a common ancestor (Costa et al., 2020a). Therefore, identifying homology among selected sequences is essential for generating meaningful pangenomes.

To explore this further, I will focus on two types of homologous genes: paralogs and orthologs. The distinction between orthologs and paralogs is illustrated in Figure 1.9. Orthologs are genes present in different species that evolved from a common ancestral gene through speciation (Ridley, 2004). In contrast, paralogs are genes that arise from the duplication of an ancestral gene within a lineage, independent of speciation events (Fitch, 1970). As paralogs evolve from a common ancestral gene through duplication, they diverge and acquire different functions over time (Koonin, 2005). Consequently, orthologous genes are considered more reliable for capturing evolutionary histories. They are often assumed to maintain similar functions throughout evolution, and patterns of genetic divergence can aid in studying relationships among organisms (Koonin, 2005). Paralogous genes introduce complexities when attempting to calculate meaningful phylogenies or to distinguish between closely related organisms and species. Distinguishing between paralogous and orthologous genes can be important for robust and meaningful analyses.

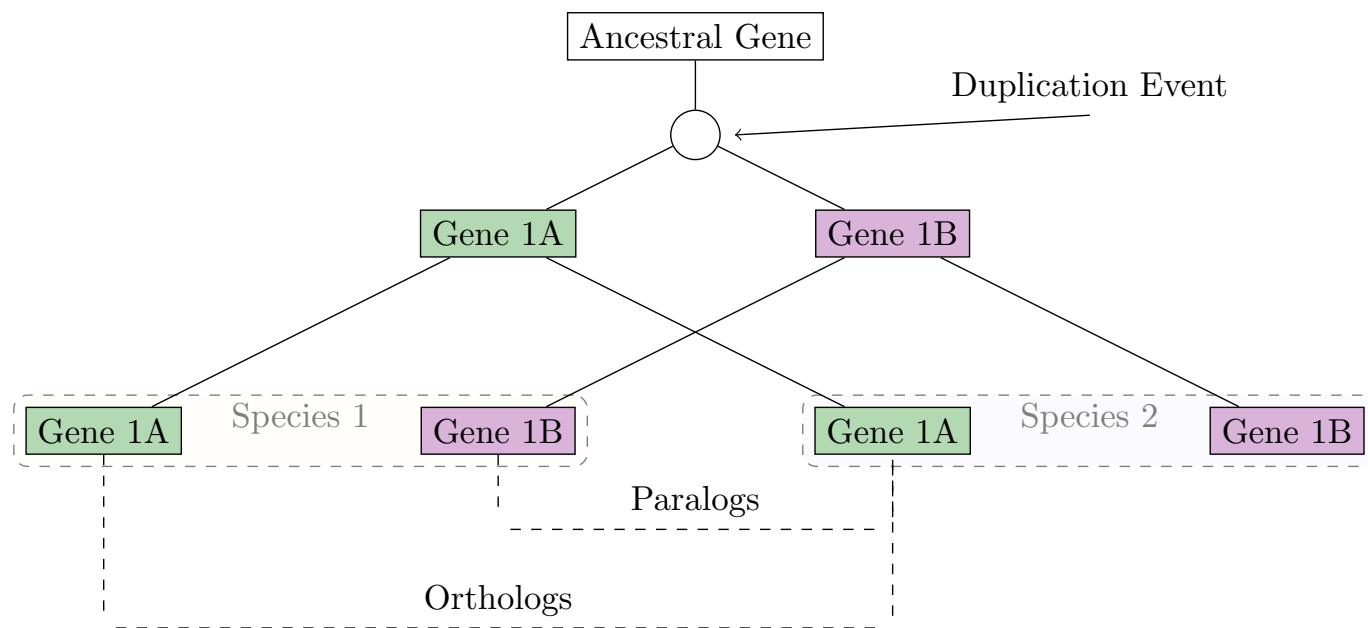


Figure 1.9: Types of homologous genes. Orthologs correspond to two genes in two different lineages that diverged through a speciation event. Paralogs are homologous genes present in two different lineages that separated by gene duplication. Figure adapted from Ridley, 2004.

The emergence of new genes from functional paralogs of central metabolic pathways could potentially reveal previously unknown and unexplored chemical compounds, often referred to as 'chemical dark matter' (Chevrette et al., 2019a). These new genes, which are derived through duplication events, might be capable of producing novel and diverse natural products. For example, *Streptomyces* species possess redundant copies of many glycolytic enzymes, such as the pyruvate kinase genes (pyk1 and pyk2). Although these paralogs perform similar biochemical functions, they have been found to distinctly modulate secondary metabolite production and fitness in *Streptomyces coelicolor* (Schniete et al., 2018a). This demonstrates how the functional diversity of paralogs can lead to variations in metabolic outputs, including the synthesis of new natural products (Chevrette et al., 2019a). By studying these paralogs, we can uncover novel biosynthetic pathways and discover previously unknown compounds.

An important consideration is that distinguishing between paralogs and orthologs can be challenging due to gene loss events during evolution, which may lead to mistakenly identifying paralogs as orthologs (Ridley, 2004). For instance, when a duplication event leads to the formation of orthologs and paralogs, subsequent gene loss in different lineages may leave only the descendants of the duplicated genes (paralogs) (see Figure 1.10). This can pose challenges, especially when studying microbial diversity through core-gene phylogenies.

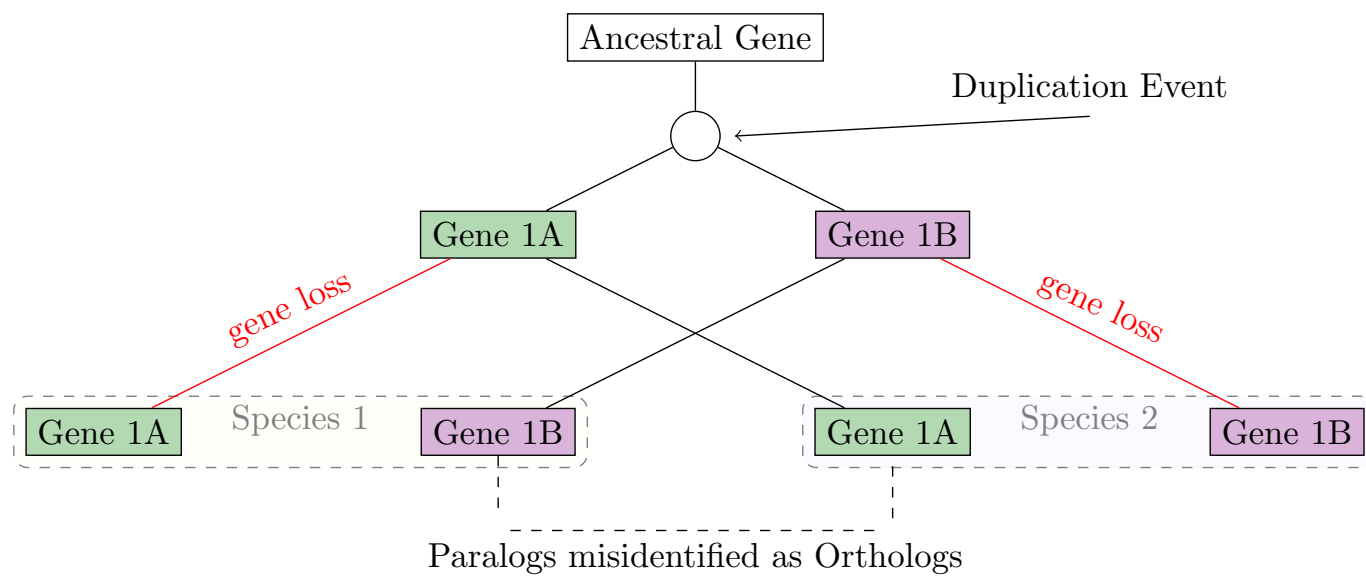


Figure 1.10: Descendants of copies of the duplicated genes can be lost in different lineages posing a challenge to orthology identification. Figure adapted from Ridley, 2004.

Pangenome analyses hold significant potential for accelerating the discovery of novel antimicrobial candidates and mitigating the effects of antimicrobial resistance (AMR). By conducting pangenome analyses, researchers can compare the complete gene sets of various organisms. This comparison can reveal unique or less common biosynthetic gene clusters (BGCs), gene families not found in commonly studied strains, and BGCs shared in a group of organisms with a common phenotype. Such BGCs may encode novel antibiotics or other bioactive compounds with potential therapeutic applications (Livingstone et al., 2018).

The significance of accessory genes can vary. A gene unique to a single organism might: (i) not provide a substantial advantage, thus offering limited insights into biological phenomena (Domingo-Sananes & McInerney, 2021); (ii) be an artifact of genome contamination, especially with poor-quality genomes, potentially misleading biological interpretations (Tonkin-Hill et al., 2023); (iii) be a recent evolutionary acquisition still in the process of fixation, offering insights into adaptive processes and evolutionary dynamics (Chevrette et al., 2019a); (iv) result from horizontal gene transfer (Tonkin-Hill et al., 2023); or (v) be involved in specific ecological interactions or niche adaptations, enhancing our understanding of the organism's unique lifestyle and environmental relationships (Domingo-Sananes & McInerney, 2021).

Closed pangenomes are characterised by a large core genome with relatively small accessory genomes, suggesting that adding more genomes is unlikely to uncover novel gene families or bioactive compounds (Figure 1.8). In contrast, open pangenomes feature a small core genome and a large accessory genome, indicating that new genome additions

could introduce new genes, potentially including those with antimicrobial activity (Costa et al., 2020b; Mohite et al., 2022; Rouli et al., 2015). Our understanding of a pangenome can be influenced by the number of genomes sequenced (sampling bias) or the specific genomes included in the analysis, leading to observations that may differ from the true diversity present in nature. For instance, if sequencing efforts are concentrated on a particular subset of the species or genus (e.g., only *Streptomyces coelicolor*), the resulting pangenome may not reflect the full variability of gene content across the entire genus of *Streptomyces*. Additionally, a pangenome that appears closed with a small set of genomes might actually reveal more openness and diversity when additional genomes are sequenced (Domingo-Sananes & McInerney, 2021). Furthermore, there are also consequences of misclassification of the input dataset. In this hypothetical example, we can consider pangenome analyses involving isolates from two different genera that were unintentionally assigned to the same genus, such as *Streptomyces griseus* and *Micromonospora aurantiaca*. These genera are distinct, with separate evolutionary histories, leading to divergent sets of core and accessory genes. If we were to compare pangenomes between these two genera, we might observe a greater apparent openness due to the large number of genes unique to each genus that are not shared. This apparent openness could mislead us into thinking that each genus's pangenome is more expansive than it truly is, when in fact, the observed diversity may primarily reflect the evolutionary distance and differences in gene content between the genera. This can also lead to redundancy or misleading results if the evolutionary divergence is not adequately considered (Tonkin-Hill et al., 2023). For a more accurate assessment of pangenome openness and to avoid biases, it's crucial to select genomes with similar evolutionary

backgrounds for comparison or to understand how evolutionary distance influences the observed pangenome structure (Mohite et al., 2022).

1.3 Taxonomy Across Time: Past, Present, Future

1.3.1 Origins of Taxonomic Classification in Biology

Taxonomic classification has a long history, tracing back to ancient times when Aristotle first classified animals into three distinct groups based on their locomotion capabilities: walking, flying and swimming (Bouteau et al., 2021). However, as ducks can walk, fly, and swim, a more comprehensive system is clearly needed.

The modern classification system and the nomenclature of organisms that we use today have their roots in the 1700s when Carl Linnaeus introduced the hierarchical classification system and binomial nomenclature (Linnaeus, 1759). The binomial nomenclature introduced by Linnaeus involved assigning latinised two part names, which consist of the generic name (genus) and the specific name (species). The aim of this standardised naming system was to eliminate confusion and ambiguity, providing scientists with a common language to communicate precisely about different living organisms. However, Linnaeus could only assign names to what he knew about, mostly macroscopic organisms. The diversity of bacteria would not become evident until the 19th century, when Ferdinand Cohn first implemented Linnaeus's taxonomic classification in bacteria in 1872 (Cohn, 1872).

The hierarchical classification system arranged organisms into categories, from more general (e.g. kingdom) to more specific (e.g. species), allowing for a structured and comprehensive understanding of the diversity of life (Kämpfer, 2012). The arrangement

of various life forms into this hierarchical systems is based on shared characteristics. Since the proposal of the Linnaean system, it has been under continuous debate which characteristics should be used for assigning taxonomy. For instance, grouping bacteria by shape alone can yield different results compared to grouping by cell wall composition, such as Gram-positive and Gram-negative bacteria. Considering a hypothetical scenario, *Bacillus subtilis* and *Escherichia coli* would be placed in the same category if classified solely by their rod shape. However, when classified by cell wall composition, *B. subtilis*, a Gram-positive bacterium, would be separated from *E. coli*, a Gram-negative bacterium. This highlights the complexity and ongoing debates in microbial taxonomy, underscoring the role of human judgment in deciding which characteristics are most informative and practical for classification. This is inevitable, as the classification of organisms evolves in response to the available data, progress of our understanding of biology and continuous advances in technology (Schleifer, 2009).

1.3.2 Brief history of bacterial taxonomic classification

Bacterial taxonomy began in the late 19th century as scientists wished to describe pathogenic bacteria (Drews, 2000). This interest was mainly driven by an improved understanding of medical science and public health concerns after an increased awareness of the involvement of microorganisms in causing diseases. During this time, many pathogenic bacteria known today, including *Bacillus* (Ehrenberg, 1834), *Vibrio cholerae* (Lippi & Gotuzzo, 2014), and *E. coli* (Escherich, 1885), were described. Ferdinand Cohn was the first person to implement Linnaeus's taxonomic classification in bacteria, classifying six genera primarily based on morphology (sphericals, short rods, threads, and spirals) in 1872 (Cohn, 1872). In the context of bacteria, the assignment of names

historically aimed to reflect specific characteristics, such as pathogenicity, nutritional preferences, susceptibility to antibiotics, and antibiotic production capabilities. In modern practice, assigned names are often interpreted as implying those characteristics (Prinzi & Moore, 2023).

As our knowledge has expanded and more biological data have become available due to advances in technology and chemistry, taxonomists have developed many new techniques for distinguishing between different bacterial organisms (Schleifer, 2009). This includes chemotaxonomy, a classification method based on the chemical composition of bacterial species, particularly the analysis of biochemical and metabolic characteristics. For example, the development of advanced chromatography and mass spectrometry has enabled detailed analysis of cell wall components such as peptidoglycan, revealing variations in its structure and composition (Schleifer & Kandler, 1972). Additionally, modern gas chromatography and fatty acid methyl ester (FAME) analysis have provided insights into variations in membrane fatty acids among different bacterial genera (Cacciapuoti et al., 1991).

The DNA era has had a significant impact on studying evolutionary relationships by enabling more efficient and reliable classification of microbes. Ben Hall and Sol Spiegelman were the first to use DNA in taxonomic classification by developing the DNA-DNA hybridisation (DDH) technique (Hall & Spiegelman, 1961). This technique involves several steps. Initially, DNA samples from different isolates are fragmented into smaller pieces, usually 1200bp. These fragmented DNA samples are then denatured and allowed to re-anneal, forming hybrid double helices. The hybridisation occurs both between DNA fragments from the same sample (self-self) and between fragments

from different samples (self-other). The degree of similarity between the two sequences is measured based on the difference in melting temperature needed to separate the strands between self-self and self-other. A higher temperature needed to separate the strands indicates more hydrogen bonds being formed, suggesting a much closer complementarity between the two sequences, whereas a lower temperature needed to separate the strands indicates fewer hydrogen bonds, suggesting that the two isolates are more distantly related. This approach provides a quantitative measure, enabling a more precise determination of the genetic relationships of microbial taxa. DDH values above 70% are considered indicative of species-level relatedness (Goris et al., 2007).

Classification of organisms based on shared characteristics (polyphasic classification), including visible phenotypes, developmental processes, behavior, and biochemical properties, as well as DNA-DNA hybridisation and 16S rRNA gene diversity, relies on extensive laboratory work. These approaches are often costly and time-consuming, and can be limited in its ability to resolve differences between closely related organisms. However, recent advances in sequencing technology have allowed for more efficient and reliable classification of microbes based on heritable genomic material (Rosselló-Móra & Amann, 2015). Several classification methods have been proposed to infer likely taxonomy by comparing genomic or protein sequences (Thompson et al., 2013). Currently used taxonomic classification methods compare organisms based on a range of different characteristics and exploit different amounts of biological information to infer taxonomic groups (eg. from a single gene to an entire genome) (Chevrette et al., 2019b).

1.3.3 Taxonomic classification in the genomic era

Single gene taxonomy

In 1977, Woese and Fox were the first to use the sequence of the 16S rRNA gene to generate a microbial phylogeny (Kapustina et al., 2021; López-Aladid et al., 2023; Woese & Fox, 1977). 16S rRNA was recognised as a good candidate for studying microbial diversity and community due to its ubiquitous distribution across all bacterial species, functional conservation, presence of conserved regions used as sites for PCR priming, and presence of nine variable regions (V1-V9) that could be used for delineating taxonomic boundaries (Clarridge, 2004). Additionally, at the time, it was widely assumed that 16S rRNA genes were not subject to HGT, further reassuring its suitability as reliable phylogenetic marker.

The preferred choice of variable 16S region for phylogenetic reconstruction has changed over time, evolving alongside advancements in our understanding of microbial diversity, the availability of new data, and improvements in sequencing technology. Initially, single variable regions like V4 were used for taxonomic identification due to their practicality in PCR amplification. Being approximately 250bp long, they allowed for easier and more cost-effective sequencing compared to longer regions. However, the V4 region was found to be poor in practice as it led to low taxonomic resolution (Edgar, 2018b). As sequencing technology evolved, it allowed for longer , combined variable regions such as V3-V4 and V4-V5 to be sequenced, offering improved taxonomic resolution (Bukin et al., 2019). This led to debates over the choice of variable regions for taxonomic classification, often resulting in different regions being more acceptable for

different genera. For instance, previous studies showed that the V2 region, but not the V3 region, is able to distinguish between *Mycobacteria* and *Nocardia* at the genus level (Chakravorty et al., 2007). V6-V9 sub-regions were found to be effective in classifying sequences belonging to the genera *Clostridium* and *Staphylococcus*, whereas V3-V5 sub-regions provided better resolution for sequences belonging to *Klebsiella* (Johnson et al., 2019).

Phylogenetic tree reconstruction based on 16S rRNA sub-regions was associated with many discrepancies when compared to polyphasic taxonomies (Edgar, 2018a). Hence, taxonomists have suggested that hypervariable regions of 16S rRNA provide enough genetic information for phylogenetic classification to be only carried out at genus-level and not species-level (Johnson et al., 2019). It was later proposed that discrepancies in phylogenetic classification can be resolved by investigation of full-length 16S rRNA sequences (Terefework et al., 1998). However, taxonomic classification based on full-length 16S rRNA sequences can still fail to provide high-resolution phylogenetic analysis. This approach often produces conflicting branch orders and lacks sufficient variation in the sequence to accurately assign taxa at the species level, for example, within *Pseudomonas* and *Actinobacteria* (Bosshard et al., 2006).

Outside their traditional use in taxonomic studies, 16S rRNA gene sequences are extensively used in metabarcoding, a technique for characterising and quantifying microbial communities in environmental samples and studying microbiome diversity (Jiang et al., 2023; Jovel et al., 2018; Santos et al., 2020). Metabarcoding is an approach used to unravel the complexities of biological communities by providing detailed insights into species diversity and ecosystem dynamics. For instance, metabarcoding of environmen-

tal samples, such as soil, can reveal changes in the rhizosphere microbial community in response to disease (Jiang et al., 2023). Similarly, metabarcoding of the human gut microbiome has uncovered associations between microbial diversity and various diseases (Jovel et al., 2018), including potential links between gut bacterial dysbiosis and neurodegenerative diseases (Singh et al., 2022). This application fundamentally relies on the taxonomic frameworks developed from traditional studies. Essentially, we have a pre-existing taxonomy of microorganisms, and metabarcoding involves matching the 16S sequences we obtain from samples to this established taxonomy to identify and classify the microbes present.

However, the accuracy of assigning taxonomy using 16S rRNA sequences in both taxonomic classification and metabarcoding studies faces additional challenges. One of which is the lack of a precise threshold to use that corresponds to a particular taxonomic circumscription level. Clustering of 16S rRNA sequences at a specified threshold is a common approach for grouping similar sequences together (Rognes et al., 2016). 16S rRNA sequence clustering thresholds were established at 95% identity for defining the same genus and 97% for the same species (Stackebrandt & Goebel, 1994), with the species threshold later revised to 98.7% (Stackebrandt, 2006). These thresholds were proposed, when only a small fraction of the currently-known 16S rRNA sequences were available (Stackebrandt, 2006; Stackebrandt & Goebel, 1994). It was later proposed to treat each unique 16S rRNA sequence as a distinct taxonomic unit (zero Operational Taxonomic Unit; zOTU) (Edgar, 2018c). However, with the recent increase in available 16S rRNA sequences, it is now known that:

1. Distinct species can share identical 16S rRNA sequences, such as *Bacillus glo-*

bisporus and *Bacillus psychrophilus* (Fox et al., 1992), potentially leading to misidentification of taxa.

2. Some species possess multiple diverse copies of the 16S rRNA gene (Eren et al., 2013), which can result in overestimation of abundance in bacterial populations. This discrepancy arises because the number of 16S gene copies in a genome does not always reflect the true abundance of a bacterial taxon in the sample.
3. Strain-level isolates of the same species, such as *E. coli K12* (Eren et al., 2013), may have different 16S sequences, which can lead to an overestimation of bacterial diversity.

Multi-gene taxonomic studies

Studying evolutionary relationships and classifying bacteria using multiple gene markers is expected to improve taxonomic resolution as using multiple genes provides a greater number of informative nucleotide sites compared to study of a single gene (Schleifer, 2009). Both approaches often utilise the use of housekeeping genes, which are involved in the maintenance of fundamental biological processes and considered essential for bacterial survival. These housekeeping genes are often highly conserved across bacterial species within the same group (e.g., genus, species), making them good candidates for studying bacterial diversity. However to be effective for bacterial typing purposes, genes must be present in many taxa while also being sufficiently variable in sequence to provide useful phylogenetic information. Examples of bacterial housekeeping genes might include those encoding enzymes involved in:

1. DNA replication, such as those involved in the repair of DNA damage (*gyrB*) or

- initiation of DNA replication (*dnaA*) (Miyoshi-Akiyama et al., 2013).
2. transcription, like those encoding for the subunits of RNA polymerase needed for the elongation of DNA (*rpoB*) (Jacobson et al., 2008).
 3. translation, genes encoding a GTPase involved in translation elongation (*elpA*) (Godoy et al., 2003).
 4. metabolism, like *atpD*, which encodes for a crucial component of the ATP synthase enzyme complex (Baldwin et al., 2009).
 5. components of the bacterial ribosome (16S and 23S) (Inoue et al., 2011; Rimbara et al., 2012)

The multilocus sequence typing (MLST) approach was first introduced in 1998 and applied to characterise an exemplar organism *Neisseria meningitidis* (Maiden et al., 1998a). This choice of organism was due to the difficulties previous methods faced with *N. meningitidis* because of its frequent recombinational exchanges among lineages. MLST enables discrimination of bacterial diversity based on the comparison of internal sequence fragments of usually 5-7 gene markers (Figure 1.11). In MLST, each marker variant is screened for identity against a database of already known marker sequences. Each variant sequence is considered to be a distinct allele and is assigned a unique allele number. For each isolate, the allele numbers are combined to produce a profile, and each unique profile is assigned a unique sequence type (ST) (Aanensen & Spratt, 2005). Bacterial isolates sharing the same ST are considered to be closely related, and it is usually assumed that MLST provides resolution at a taxonomic level below species (Pilo et al., 2021).

MLST has been popular as it can:

1. Provide fast, scalable and reproducible typing (Urwin & Maiden, 2003).
2. Be used to study bacterial relatedness and to infer population structure (Spilker et al., 2012).
3. Trace the origin and spread of AMR (Wang et al., 2021).

However, MLST has been criticised on the basis of:

1. Lack of a universal set of marker sequences suitable to study all bacterial strains, as organisms differ in composition of their core genomes, making it impossible to use the same set of markers between, and sometimes within, different genera (Jolley & Maiden, 2014; Jolley et al., 2018; Larsen et al., 2012; Thomas et al., 2006).
2. Some sets of markers might not reflect evolutionary relationships inferred from whole-genome sequences (Tsang et al., 2017).
3. Isolates can only be compared if one variation of each marker gene is present (i.e. markers must be present in single copies, and this is not always true).
4. There are no clear criteria for selection of appropriate marker genes and, there is a lack of systematic exploration of alternative markers (Schleifer, 2009).

MLSA is similar to MLST in that it studies bacterial diversity based on a set of several genes. The term MLSA was initially proposed by Gevers et al., 2005 and described as a genotypic approach for investigating wider taxonomic groups such as

genera, using sequences derived from multiple protein-coding genes. The main difference between MLST and MLSA is that MLST is a clustering-based method that relies on STs (allelic profiles) to describe the relationship between isolates, while MLSA uses genomic sequences to construct phylogenetic trees (Gevers et al., 2005; Glaeser & Kämpfer, 2015), providing enhanced resolution and the ability to resolve discrepancies in phylogenetic studies based on single-gene phylogeny (e.g., 16S rRNA) (Rong & Huang, 2010a). MLSA was proposed as a potential replacement for DNA-DNA hybridization in species delineation, and has been used to help delineate novel species (Vandamme & Peeters, 2014). However, both MLSA and MLST have also been associated with discrepancies in taxonomic classification (Jin et al., 2020). This led to the proposal of core-genome MLST (cgMLST) which expands on the principles of MLST by examining the core set of genes shared across all strains within a set of organisms (often within a species), rather than relying on a smaller set of housekeeping genes. This approach allows for a finer resolution of bacterial diversity and can provide insights into relationships at a more granular level than traditional MLST (De Been et al., 2015).

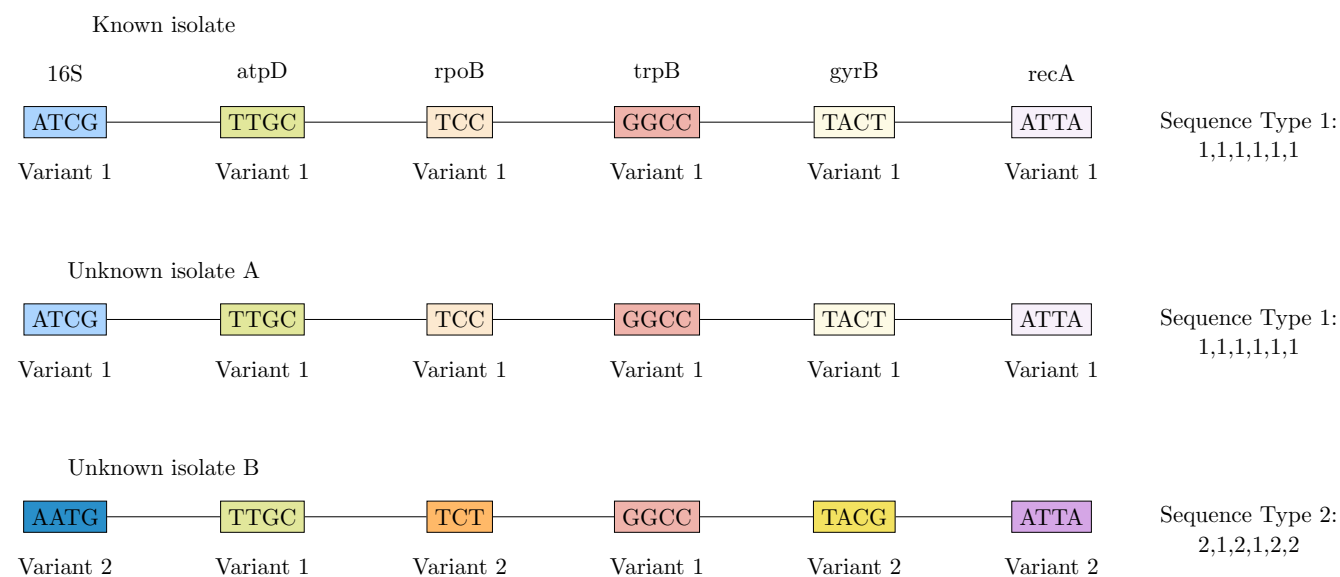


Figure 1.11: Illustration of multilocus sequence typing (MLST). Each genome contains seven housekeeping genes, each indicated by a different colour. MLST involves screening each input allele sequence (unknown isolate A and unknown isolate B) against a database of previously known allele sequences, each of which has been assigned an allele number. When a locus sequence does not match any sequence in the database, a new allele number is assigned (indicated here by a darker fill colour). The assigned allele numbers are combined into allelic profiles that are used to characterise strains. Each allelic profile is assigned a unique sequence type identifier

Whole genome taxonomy

An organism's whole genome encompasses its entire set of genes and provides a comprehensive blueprint that contributes to its function and survival, as well as highly detailed information about the organism's evolutionary history (Goldman & Landweber, 2016). The genome comprises coding regions that encode proteins and non-coding regions responsible for regulating gene expression and other biological processes whose functions may still be unknown or unclear. The genomic data provided by whole genomes serves as a valuable resource for taxonomic classification, containing all the information needed to understand complex relationships across microbial taxa (Whitman, 2015). Rapid advances in technology have enabled faster and more affordable sequencing, making genome sequences more accessible for studying bacterial diversity (Land et al., 2015b). As of April 15th, 2024, a total of 3,333,621,823 whole genome sequences have been submitted to the National Centre for Biotechnology Information (NCBI) (Sayers et al., 2021) (Figure 1.12).

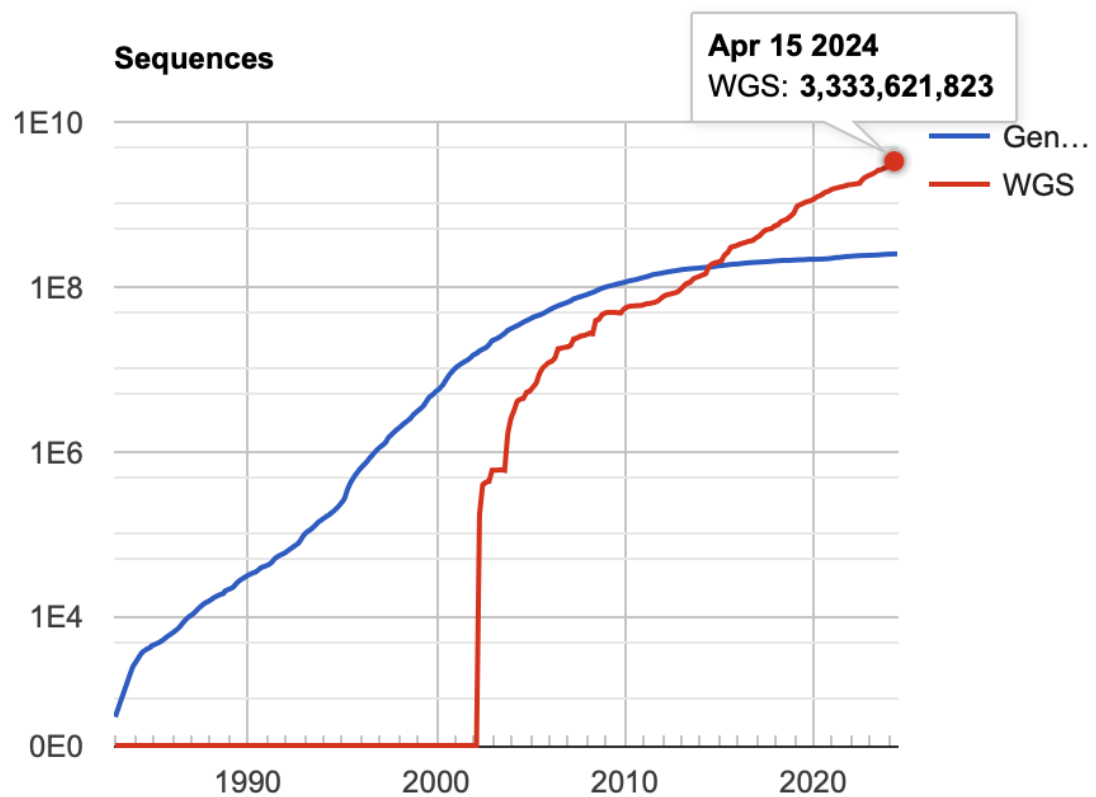


Figure 1.12: Summary Statistics of GenBank and WGS (Whole Genome Shotgun) sequence submissions in NCBI. Figure sourced from [NCBI GenBank and WGS Statistics Page](#) accessed on the 7th of August 2024.

Studying evolutionary relationships using whole genomes has gained widespread uptake for the following reasons:

- It includes more genomic variation, offering higher phylogenetic resolution and a deeper understanding of evolutionary relationships (Brumfield et al., 2020; Chevrette et al., 2019b).
- It can be used to detect misidentified and misclassified strains (Figueras et al., 2014).
- It can identify cryptic species, which are morphologically indistinguishable but genetically distinct (Jin et al., 2020).

However, there are also challenges associated with whole genome taxonomy. For instance, it can require large amounts of storage and computational resources for analysis, as analysing large genomic datasets can be time-consuming and computationally expensive (Servedio et al., 2015). Additionally, low-quality genomes (e.g., contaminated and incomplete genomes) can introduce errors and lead to biased observations (Li & Yin, 2022). For example, low-quality genomes arising from contamination or incomplete sequencing may lead to inaccurate phylogenetic trees or misinterpretations of evolutionary relationships (Lemmon et al., 2009; Schloss et al., 2011). Contamination can be detected using tools like CheckM, which helps identify and exclude problematic contigs (Parks et al., 2015). Incomplete genomes can similarly be flagged by comparing the genome size to expected values or using CheckM to assess completeness.

Several whole-genome-based classification methods have been developed. For in-

stance, inferring phylogenies from the concatenated set of genes shared between organisms that have remained in a single copy (single-copy orthologs; SCOGs) since the last common ancestor. SCOGs are useful for studying bacterial diversity as they are often involved in biological processes essential for the organism's survival and growth, rather than genes involved in niche-specific adaptations (Ciccarelli et al., 2006). Since they are present in one copy per genome, they allow for a 1-to-1 mapping, ensuring consistency and robustness in comparative analyses. Although they reflect more than one potential gene history, phylogenies inferred from SCOGs are considered to provide more reliable taxonomic classification compared to single-gene or multi-gene phylogenies, often revealing discrepancies in current taxonomic classifications. Additionally, this method has led to the proposal of novel genera such as *Denitrificimonas*, *Parapseudomonas*, and *Neopseudomonas*, which were previously classified as *Pseudomonas*, which were previously indistinguishable with single or multi-gene approaches, (Saati-Santamaría et al., 2021).

Whole-genome sequence classifications can also be based around distance methods such as Average Nucleotide Identity (ANI). ANI was initially proposed to mimic the DDH method for defining species boundaries. ANI quantifies the genetic distance between two genomes by aligning similar regions, allowing for insertion, deletions and substitutions. An ANI identity value is then calculated to represent the nucleotide identity over these aligned regions (Goris et al., 2007). A typical threshold for species boundary is 95%, which has been found to be consistent with the traditional DDH threshold of 70% (Ondov et al., 2016; Richter & Rosselló-Móra, 2009). While ANI provides a robust framework for species delineation, according to the International Code of Nomenclature of Prokaryotes

(ICNP), formal guidelines for genus description remain undefined (Parker et al., 2015). Additionally, there is no universally accepted rule for defining genus boundaries in whole-genome comparisons. Some researchers use percentage of conserved proteins (POCP) and Average Amino Acid (AAI) values to classify genera (Barco et al., 2020; Qin et al., 2014), but these are not standardised criteria, nor are they required for genus designation. This lack of standardised criteria suggests that further investigation into appropriate thresholds for genus-level classification is necessary. A critical consideration in this context is determining the threshold at which two genomes should no longer be regarded as part of the same genus. Closely related genera tend to share a higher proportion of their genetic material (Goris et al., 2007), so more closely related organisms are expected to exhibit greater alignment length. One theoretical approach suggests that a 50% genome coverage threshold could serve as a useful starting point for genus classification (Pritchard et al., 2015). The rationale is that if two genomes (A and B) share less than 50% of their genetic material by alignment length, a significant portion of one genome (e.g., genome A) may instead be more closely related to a different genome (e.g., genome C), which has no homology with genome B. For example, if genome A and genome B share only 20% of their sequence, while genome A shares 80% of its sequence with genome C—but genome B and C have no detectable homology—then classifying A and B within the same genus may not be justifiable. This suggests that below a 50% threshold, genomic continuity between two organisms weakens to the point where they may be more appropriately placed in separate genera. While the 50% genome coverage threshold is not a diagnostic measure, it provides a practical starting point for genus delineation. It offers a preliminary framework for assessing taxonomic relationships,

guiding further investigation before more sophisticated classification methods or refined thresholds can be proposed and applied.

Several bioinformatics tools for measuring ANI, such as pyANI (Pritchard et al., 2015), OrthoANI (Lee et al., 2016a), FastANI (Jain et al., 2017), Mash (Ondov et al., 2016), sourmash (Brown & Irber, 2016) and JSpecies (Richter et al., 2016), have been developed. These tools, along with the increasing number of genomic sequences, allow for more robust taxonomic classification of organisms. They have been identified discrepancies between whole-genome methods and single-gene taxonomy (Chevrette et al., 2019b), while being scalable for large datasets (Varghese et al., 2015) and identification of mislabeled genomes (Figueras et al., 2014).

ANI typically involves searching for and aligning genome sequences, followed by calculation of identity (Goris et al., 2007). Over time, many algorithms have been proposed in response to increased availability of genomic data and advances in technology. In the original concept of ANI proposed by Goris (Goris et al., 2007), fragmented query sequences of 1020bp are aligned against intact subject sequences using BLAST (Altschul et al., 1990): the ANIb algorithm. Later, in 2009, Richter and Rosselló-Móra developed the ANIm algorithm that uses MUMmer software (Kurtz et al., 2004) to align entire genome sequences identifying maximal unique matches, which are then used to calculate the final ANI value (Richter et al., 2016).

Alternative approaches and methods have been recently developed such as fastANI, Mash, sourmash and OrthoANI (Brown & Irber, 2016; Jain et al., 2017; Lee et al., 2016a; Ondov et al., 2016). FastANI, Mash, and sourmash are alignment-free methods that use MinHash sketching, a technique that approximates genome similarity by comparing

representative subsets of k-mers instead of performing full alignments. These methods calculate a Jaccard index based on shared k-mers between genome sketches, which is then converted into a Mash distance and used to estimate ANI. By avoiding computationally expensive alignments, MinHash-based approaches enable rapid, large-scale genome comparisons while maintaining high accuracy for closely related sequences. On the other hand, OrthoANI calculates the ANI value by fragmenting both genomes (query and subject) into 1020bp long fragments, identifying pairs of fragments with reciprocal best hits, and reporting values for all reciprocal best hits with the exclusion of matches with less than 1020 bp. Since ANI was first introduced in 2007, several bioinformatics tools have been developed to estimate ANI values using different algorithms, which are summarised in Table 1.3.

Table 1.3: A summary description of software that calculate ANI and algorithms they use.

software	MUMmer	BLAST	k-mer/fastANI	reference
pyANI	✓	✓	✓	(Pritchard et al., 2015)
orthoANI		✓		(Lee et al., 2016a)
dnadiff	✓			(Kurtz et al., 2004)
JSpecies	✓	✓		(Richter et al., 2016)
FastANI			✓	(Jain et al., 2017)
Mash			✓	(Ondov et al., 2016)
sourmash			✓	(Brown & Irber, 2016)

The reported ANI value can vary depending on the algorithm used, as each method handles sequence alignment and similarity calculation differently. This leaves us with choices regarding which algorithm to use. Since ANI values are approximations rather than definitive truths, there is no single "correct" method for ANI calculation, and the choice often comes down to preference. Comparisons of ANI methods have been previously done, and they are almost always benchmarked against ANIb (Ndovie et al., 2025; Yoon et al., 2017). This is likely because ANIb was the first proposed method and directly mimics DNA-DNA hybridization by fragmenting the query genome sequence and aligning it to the intact genome. However, the reliability of ANIb itself might raise questions. Although it was originally validated against DNA-DNA hybridization data (Goris et al., 2007), the true ANI value for a given comparison remains unknown, making it unclear whether ANIb is truly the best reference standard. Previous studies have shown that ANIm runs faster than ANIb while providing similar results (Yoon et al., 2017). Meanwhile, fastANI operates at a speed three orders of magnitude faster than ANIb, producing results comparable to both ANIb and ANIm for relatively highly identical sequences (Yoon et al., 2017). However, its accuracy for more distantly related genomes has shown to be less reliable (Ndovie et al., 2025). Others have argued that discrepancies in reported values are likely to be more apparent when comparing more divergent genomes, which is likely the result of different sensitivities provided by the two methods (Goris et al., 2007). A more rigorous assessment of ANI methods was recently presented by Dr. Leighton Pritchard at a genomeRXIV meeting, where preliminary data compared ANI methods using synthetic sequences with known true values (Figure

1.13). These findings suggest that while ANIm and ANIb yield similar results, ANIm is consistently closer to the true values. In contrast, fastANI produced the least reliable results. While ANIb is often the default choice, these findings indicate that ANIm may be the more accurate method for calculating ANI values.

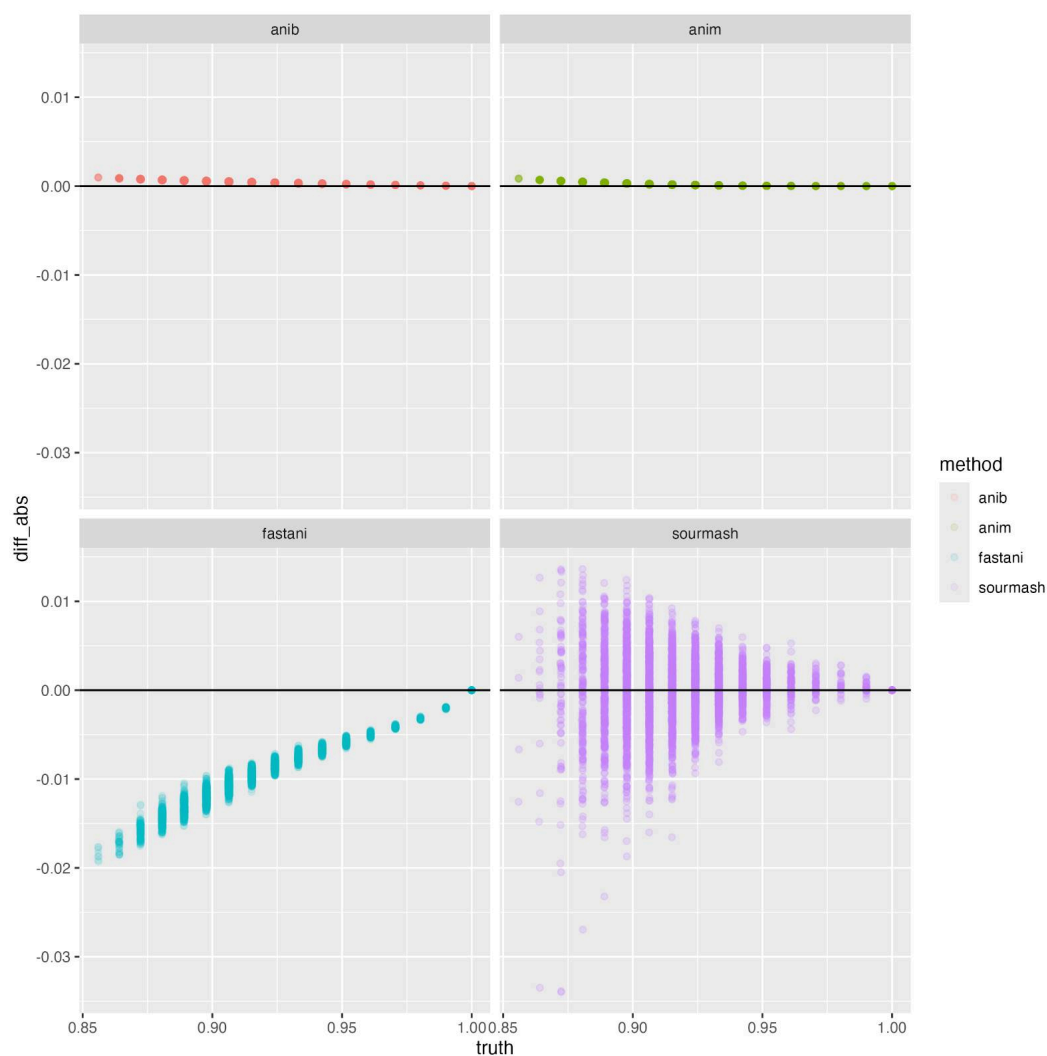


Figure 1.13: Preliminary data comparing ANI methods using synthetic sequences with known true identities, as presented by Dr. Leighton Pritchard at the genomeRxiv meeting. The results indicate that ANIm provides ANI values closest to the true identity, followed by ANIb, sourmash, and fastANI, with fastANI showing the greatest deviation.

1.4 Phylogenetic trees: best-effort attempt to reconstruct evolutionary history

1.4.1 Overview of phylogenetic trees

Phylogenetic trees provide a framework for organising and classifying organisms based on shared characteristics or genetic information, such as single genes or collections of genes, which are believed to reflect evolutionary relationships. Phylogenetic trees consist of various elements: leaves, branches, nodes and roots. Leaves represent data points or entities, such as taxa, while branches denote their evolutionary histories and relationships (Figure 1.14). Nodes indicate common ancestors where branches bifurcate, and the root - when its placement is known - represents the most recent common ancestor of all entities depicted in the tree. Interpreting a phylogenetic tree involves analysing branch lengths, which can reflect the amount of genetic change or evolutionary time, as well as the branching order (topology) - the arrangement of these branches and nodes (Gregory, 2008; Kapli et al., 2020).

Phylogenetic trees can be represented as either rooted or unrooted (Ridley, 2004). Rooted trees include a root node, which represents the most recent common ancestor of all the entities (taxa) in the tree. This root establishes a direction of evolution from the ancestor to the present. In contrast, unrooted trees lack a designated root and do not show a specific overall direction of evolutionary change (Figure 1.14). However, it can be assumed that the direction of change generally proceeds towards the leaves.

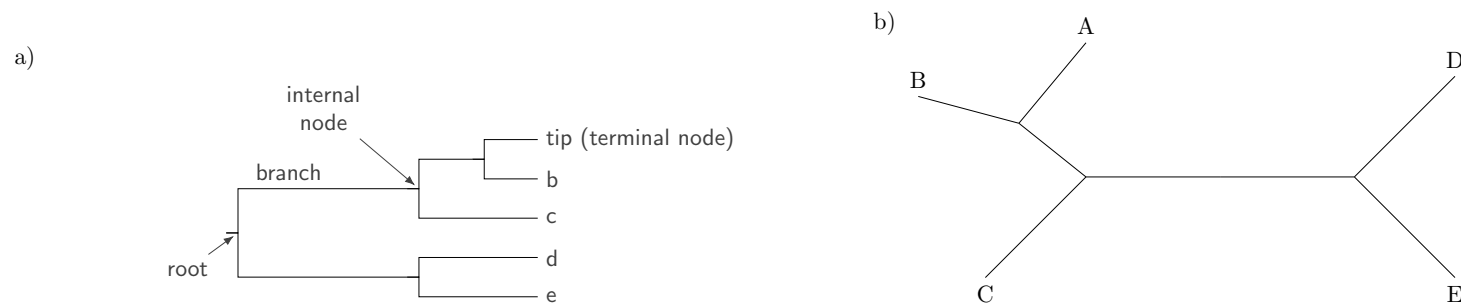


Figure 1.14: Structure of evolutionary (a) rooted and (b) unrooted trees. Terminal nodes/leaves represent a taxon, which could be species, genes, or groups of organisms. Branch ordering represents the relationships among the sequences, captured by the branching order (topology). The amount of evolutionary changes between the nodes are captured and represented as the branch lengths; the longer the branch, the more evolutionary change has occurred between the compared sequences.

Initially, phylogenetic reconstructions relied on morphological observations. Ernst Haeckel's famous 'Tree of Life' (Figure 1.15), for example, was primarily based on characteristics such as body shape, bone structure, and developmental stages (Haeckel, 1866). However, with advancements in sequencing technology, phylogenetic inference has increasingly relied on genomic data (Lemmon & Lemmon, 2013). This shift has provided more precise and comprehensive insights into evolutionary relationships, and more sophisticated methods for inferring phylogenies. Beyond their traditional role in taxonomic studies, phylogenetic trees have extended into molecular evolution. They help us understand genetic change rates in protein and DNA sequences, crucial for molecular processes, and reveal mechanisms driving genetic diversity (Lewis-Rogers et al., 2009). In epidemiology, phylogenetic trees are useful in tracking pathogen evolution, identifying resistance gene origins, and monitoring their transmission through populations (Hadfield et al., 2018; Veyrier et al., 2009). They also play a critical role in identifying horizontal gene transfer (HGT) events among bacteria and archaea. By analysing evolutionary relationships, scientists can trace how genes move between bacterial species, contributing to resistance spread and adaptation to new conditions (Nickrent et al., 2004).

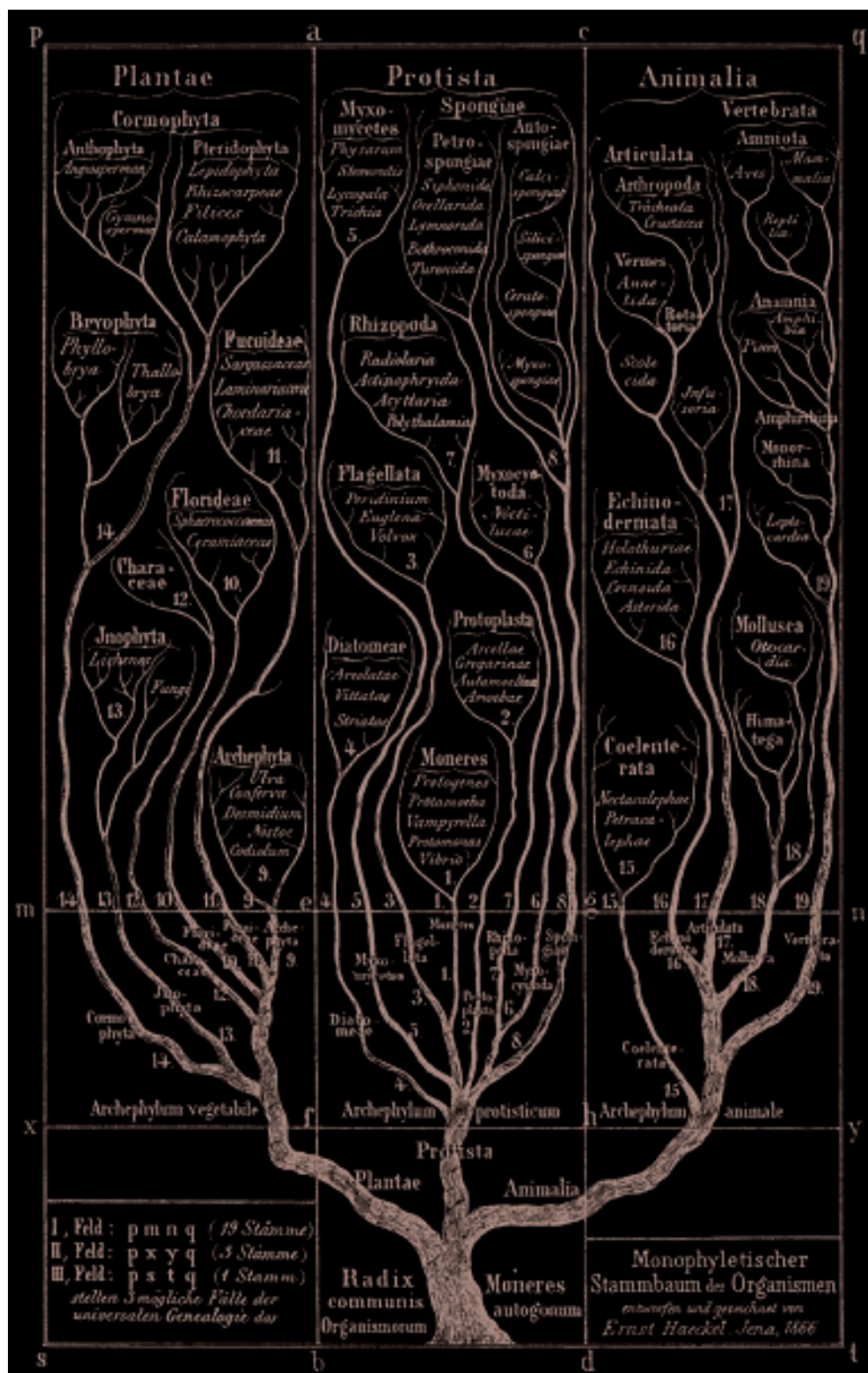


Figure 1.15: Ernst Haeckel Tree of Life. Figure reproduced from Haeckel, 1866.

1.4.2 Types of clades: monophyletic, paraphyletic and polyphyletic

Cladistics is an approach which categorises taxa based on shared characteristics. The basic concept of cladistics is that taxa sharing a greater number common features are assumed to share more recent common evolutionary history and so are grouped more closely together (Ridley, 2004). In cladistics, organisms are understood to form groups (specifically, clades) that are hypothesised based on sharing a most recent common ancestor (Hennig, 1965). There are three types of clades: monophyletic, paraphyletic and polyphyletic (Figure 1.16).

Suppose there is a set of organisms in the tree which are supposed to belong to the same species. If these organisms form a monophyletic group, it means they share a common ancestor and include all of its descendants (Figure 1.16; green). This classification indicates that the organisms not only share a common evolutionary history but are also more closely related to each other than to any organisms outside this group. Monophyletic groups are highly valued in biological classification because they are believed to accurately capture evolutionary history and shared ancestry, representing a natural and cohesive group (Velasco, 2009). They provide a comprehensive view of species adaptation and evolution over time (Ridley, 2004).

In contrast, if the organisms form a paraphyletic group, this means they share a common ancestor, but not all of its descendants are included in the group (Figure 1.16; purple). From a biological perspective, paraphyletic groups are often problematic because they may fail to accurately capture evolutionary history, presenting an incomplete representation of evolution. The acceptability of paraphyletic clades is often under

debate. Some argue that it is a natural phenomenon in nature (Hörandl & Stuessy, 2010), while others consider them to be unreliable taxonomy (Ebach et al., 2006). Those that believe that paraphyletic clades naturally occur in nature argue that these clades can form as a result of diversification. This occurs because evolving lineages might accumulate enough genetic differences leading to their exclusion of the main group, yet they might not have accumulated enough differences to be placed in a separate clade (Hörandl & Stuessy, 2010). In contrast, others believe that paraphyletic clades are formed as a result of: (i) faulty algorithms and techniques used to infer phylogenies; (ii) genes used in the phylogenetic analyses are not adequate for inferring reliable taxonomies as they do not contain sufficient or reliable phylogenetic signal (Ebach et al., 2006).

Finally, if the organisms are part of a polyphyletic group, they do not share a common ancestor, indicating multiple evolutionary origins (Ridley, 2004). Polyphyletic groups are controversial and typically not accepted as valid classifications in modern taxonomy because they undermine fundamental goals and principles of classification (Ojha et al., 2022). The most fundamental basis of modern taxonomy assumes that single taxa must share a single evolutionary lineage. However, polyphyly violates this principle by grouping taxa that do not share a common ancestor. It is also argued that polyphyly represents inaccurate evolutionary history as it disrupts biologically “true” clades by grouping evolutionarily distinct organisms. However, it’s not always the case that these clades are rejected outright, as some explain them through convergent evolution (Hernández-González et al., 2018). This phenomenon suggests that similar traits evolved independently in different lineages due to similar environmental pressures or functional constraints (Ridley, 2004).

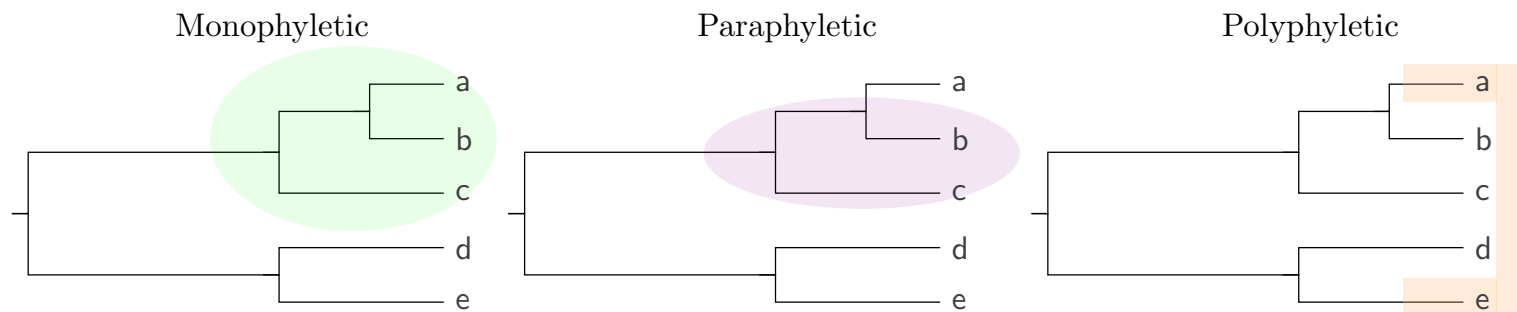


Figure 1.16: Types of taxonomic clades. A clade is considered monophyletic if it contains all descendants of the common ancestor, at the root of the clade, while paraphyletic clades include only a subset of descendants of the root common ancestor. Polyphyletic clades group together lineages that do not share a common ancestor.

1.4.3 Bifurcation *versus* multifurcation

The primary goal of phylogenetic reconstruction is to capture the evolutionary patterns and to understand the mechanism of cladogenesis - the process by which a single ancestral species diverges into two or more distinct daughter species. A common assumption in phylogenetic studies is that cladogenesis happens through a series of dichotomous branching events, where a lineage splits into exactly two distinct groups (Rokas & Carroll, 2006). This assumption is generally supported in evolutionary theory, which supports that speciation generally occurs in a stepwise fashion, with a single ancestral lineage diverging into two separate lineages over time (DeSalle & Riley, 2020). This type of branching events are represented using bifurcation trees (Figure 1.17; left), where each internal node represents a point of divergence, and each branch indicates the lineage leading to the descendant species.

Polytomous branching events can also occur, and these are depicted using multifurcation trees (Lin et al., 2011). In such events, a single node gives rise to three or more descendant branches simultaneously, resulting in a tree structure where multiple lineages emerge from a single point (Figure 1.17; right). Polytomous branching often reflects uncertainty or incomplete resolution in the evolutionary relationships among taxa. This can occur due to insufficient data, or when shared characteristics are inadequately analysed, leaving bifurcating relationships unresolved (Coyne et al., 2004; Lin et al., 2011). For this reason, multifurcating trees are often interpreted as representing unresolved evolutionary relationships, indicating that the available data may not provide sufficient resolution to determine the exact order of species divergence. However, some

biologists acknowledge that polytomous branching events can indeed occur (very rarely), when multiple new lineages arise from a single population nearly simultaneously as seen with cichlid fish. Polytomies in cichlid fish occur due to rapid speciation events, often triggered by adaptive radiation in environments like African rift lakes, where multiple species diverge from a common ancestor in a short time (Takahashi et al., 2001).

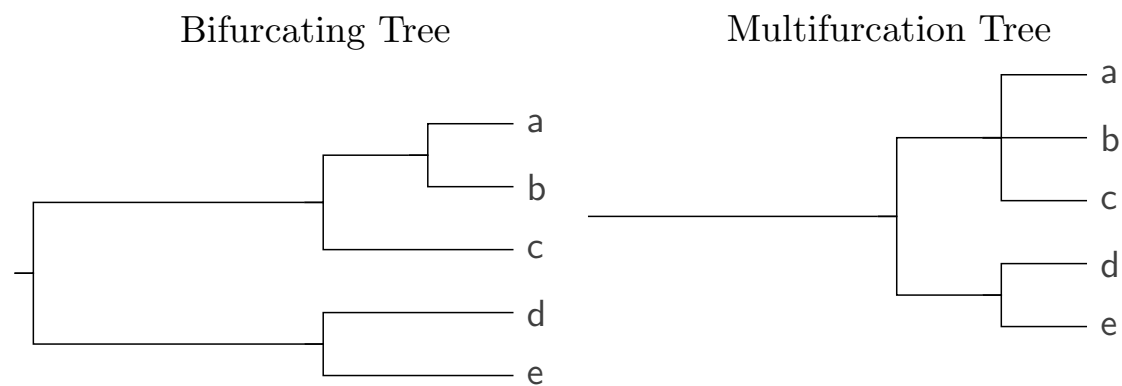


Figure 1.17: Bifurcating *versus* multifurcation tree. A bifurcation tree (left) with each internal node splitting into exactly two branches, representing a dichotomous evolutionary split. A multifurcation tree (right) where internal nodes can split into more than two branches. Multifurcations often arise due to poor resolution in the data, indicating uncertainty about the exact relationships among the taxa (soft polytomy), or represent a polytomy where multiple lineages diverge simultaneously (hard polytomy).

1.4.4 The impact of horizontal gene transfer on phylogenetic inference and the bifurcation model

Initially, bacteria were believed to reproduce solely through clonal cell division, where new bacterial cells inherit all genetic material directly from their parent cells, leading to clear evolutionary pathways of divergence from common ancestors (Smith et al., 1993). This understanding supported the bifurcation model, which suggest that evolutionary history can be mapped as a tree where each branch represents a point of divergence. However, in 1928, Griffith observed that pneumococcus bacteria could acquire pathogenic traits after contact with pathogenic strains, marking the first documented case of horizontal gene transfer (HGT). Initially thought to be a rare phenomenon, advancements in sequencing technology have revealed that HGT occurs far more frequently than previously believed (Daubin & Szöllősi, 2016). Additionally, although less common, HGT was also detected to occur distantly related lineages (Beiko et al., 2005; McDonald & Currie, 2017a). This discovery has lead to realisation that bifurcating model of evolution might be illogical, as genetic material is no longer transferred solely from parent to offspring but also between unrelated lineages (Baptiste et al., 2013). As a result, HGT introduces complexities to phylogenetic reconstructions, as genes unaffected by HGT may depict different evolutionary paths compared to those influenced by HGT, leading to inconsistencies in inferred phylogenies. Lan and Reeves (1996) had a significant impact on our understanding of how HGT affects phylogeny and our concept of bacterial species with their core genome hypothesis (CGH). The CGH suggests that a core subset of genes, present in nearly all individuals of a species,

defines its essential characteristics and is largely unaffected by HGT. These core genes are thought to be under purifying selection, preserving their functions and supporting the evolution and therefore bifurcation model. This viewpoint has again reshaped how taxonomists understand the nature of bacterial species (Riley & Lizotte-Waniewski, 2009). Some argued that reconstructing phylogenies using core gene trees could help avoid the "noise" introduced by HGT, providing a clearer and more stable framework for understanding evolutionary relationships among bacteria (Lerat et al., 2003). Although HGT has been found to affect core genes, its impact has generally been reported on a smaller scale and is unlikely to significantly alter the overall topology of core-gene trees (Saunders et al., 2005; Shi & Falkowski, 2008).

1.5 Using graph theory to analyse bacterial classifications

Graph theory, the study of graphs (also known as networks), is used to model and analyse relationships between objects. In simple terms, graphs are networks consisting of objects called nodes, or vertices (V , usually represented a shape, such as a circle) that may be connected by edges (E ; Figure 1.18) - often represented as lines. Mathematically, graphs are defined as a pair of sets (V, E) , where each edge E describes the vertices it connects (e.g. an edge connecting vertex $V1$ to vertex $V2$ is the edge $(V1, V2)$) (Diestel, 2000).

Graph theory has proven useful across various disciplines, aiding in our understanding of real-world phenomena. Graphs serve as tools to capture and represent relationships between entities that might otherwise remain hidden. Within many aspects of microbiology, graphs have been successfully applied to study protein-protein interactions (Gao

et al., 2023), gene expression (Fofana et al., 2021), evolution and population studies (Harling-Lee et al., 2022) and many other aspects of applied microbiology (Pavlopoulos et al., 2011).

Within evolutionary and microbial populations, graphs are valuable tools for representing the relationships between biological entities, such as isolates or genomes, and their structures can provide critical insights into evolutionary patterns and processes. These graphical representations can help shed light on how evolution shapes population structures. In evolutionary networks, nodes represent individual entities (organisms or isolates), while edges symbolise relationships like genetic similarity, or connections based on phenotypes, or shared traits such as 16S, or allelic sequences used in molecular typing, or whole-genome distance measures. The structure of the graph—how the nodes and edges are arranged—can reveal meaningful information about these relationships. For example, nodes with a higher number of connections may represent entities that share many characteristics or interactions with others, suggesting that they play a key role in the evolutionary landscape. Additionally, certain substructures within the graph, such as tightly connected clusters, can indicate groups of closely related isolates or evolutionary lineages. By analysing both the overall structure and specific properties of the nodes, researchers can uncover patterns that provide deeper insights into the biological characteristics, evolutionary history, and relationships of the isolates under study (Pavlopoulos et al., 2011). In the following sections, I will focus on introducing the fundamental concepts of graph theory relevant to studies carried out in this thesis by providing an empirical rather than mathematical description, focusing specifically on bacterial relationships.

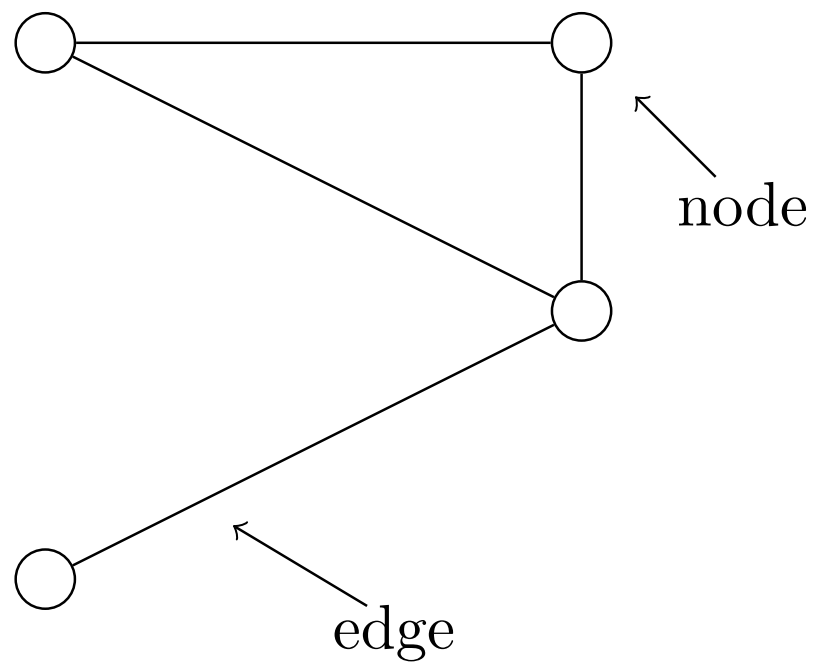


Figure 1.18: Basic graph concepts. Graphs are networks made up of objects known as nodes or vertices (V), which are often represented by shapes like circles. These nodes can be linked by connections called edges.

1.5.1 Weighted graphs

Weighted graphs are valuable tools for identifying relationships between nodes, such as genomes, where the edges are assigned weights that reflect something about their shared relationship. These weights might correspond, for example, to the number of shared genes or phenotypes between two genomes, allowing for the quantification of genetic relatedness. For instance, weights can represent variations in allele counts from an MLST profile or metrics such as pairwise genome identity. Weighted graphs are extensively used in the study of bacterial populations; for example, they have been used to analyse co-occurrence patterns within soil microbial communities (Barberán et al., 2012) and in epidemiological studies to aid in differentiating between epidemiologically related and unrelated isolates of *Enterococcus faecium* (De Been et al., 2015).

Complete (clique) and non-complete (non-clique) graph

A complete (undirected) graph is one in which every node is directly connected to every other node by a unique edge, forming what is known as a clique (Figure 1.19A). In contrast, a non-complete graph has at least one pair of nodes that is not connected by an edge, meaning it forms a non-clique (Figure 1.19B), indicating that they do not share the same relationship as the other nodes. For example, imagine you have sequenced four isolates and are trying to determine if they belong to the same species. You run an ANI analysis, adopting the widely used 95% genome identity cut-off point. In this case, if each genome shares $\geq 95\%$ genome identity with every other genome, this would be represented as a complete graph, where every node (genome) is connected by an edge. This type of graph reflects a coherent group of nodes that may be interpreted as

a clade—a group defined by common characteristics shared by all members, much like a clade is defined by sharing a single common ancestor.

Later, after isolating another four genomes, you perform the same ANI analysis. This time, not all genomes share the 95% identity threshold: for example, genome G4 and G2 share only 92% identity, while the rest of the genomes meet the 95% threshold. This situation would be represented as a non-complete graph, where certain nodes lack connections. This incomplete structure suggests inconsistencies or ambiguities in the relationships, indicating that while some isolates share certain characteristics, these are not universal across the entire group, hinting at genetic divergence or different evolutionary histories within the population.

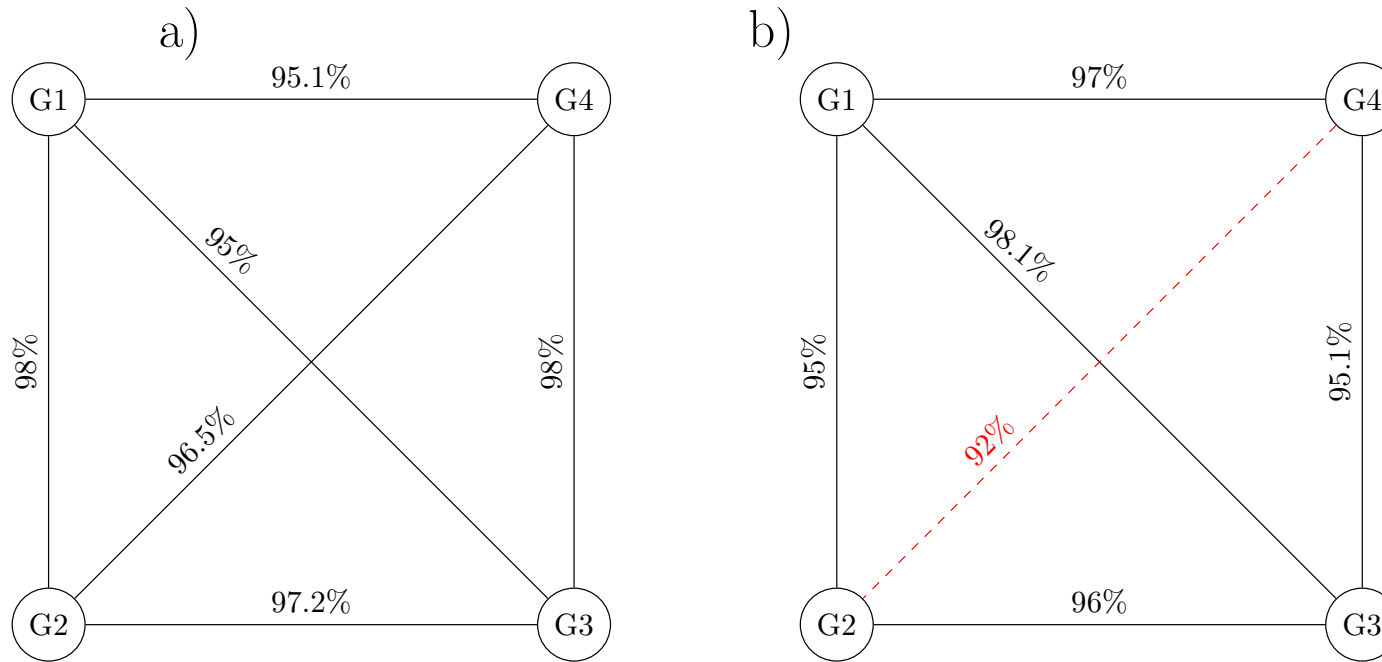


Figure 1.19: Visual comparison between a complete (a) and non-complete (b) graph. Each node represents a genome, and edges indicate genome identity, with a threshold cut-off of 95%. In the complete graph (a), every genome shares at least 95% identity with all other genomes, resulting in edges connecting every pair of nodes. In the non-complete graph (b), only genomes that meet or exceed the 95% identity threshold are connected by edges, while others are left unconnected. The red dashed line indicates a genome pair that falls below the threshold and is shown for demonstration purposes only.

1.5.2 Disconnected graph

A disconnected graph is a graph where there is not a path between every pair of nodes (Figure 1.20). This structure could suggest the presence of distinct groups within the population, potentially characterised by different variants of a particular trait (e.g., genetic sequences).

Consider a hypothetical scenario where we have four bacterial genomes isolated from a soil sample. We again used ANI comparisons with a 95% threshold, a standard for species delineation (Figure 1.20). In this example, one genome (G2), was found to share less than 95% genome identity with the other three genomes (G1, G3, G4). Conversely, the remaining three genomes each shared 95% or more genome identity with one another.

As a result, when these genomes are represented in a graph based on their ANI relationships, the graph will split into two distinct subgraphs. One subgraph will include G2, indicating that it is less similar to the other genomes. The other subgraph will cluster the three genomes that are highly similar to each other. This division illustrates that G2 is quite different from the other three genomes, which form a closely related group among themselves.

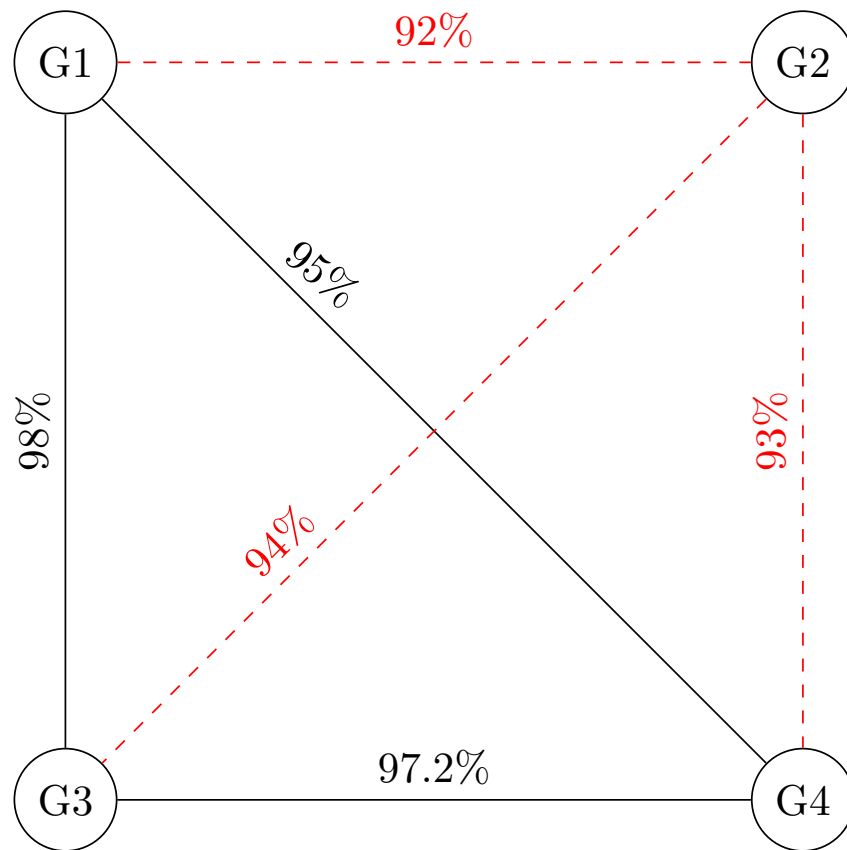


Figure 1.20: Visual representation of a disconnected graph. Each node corresponds to a genome, and edges between nodes reflect genome identity values. Only genomes with genome identity values above a 95% cut-off are connected by edges, demonstrating high genetic similarity. The graph is consequently disconnected, as some genomes fall below this similarity threshold. The red dashed lines are included for illustration purposes only and do not represent an edge.

1.5.3 Minimum Spanning Tree

Minimum Spanning Trees (MST) are a type of acyclic graph that minimises the total edge weight required to connect all the vertices (Figure 1.21). This means that the MST connects all the nodes in the graph while ensuring the smallest possible total edge weight. The resulting tree can be interpreted as representing the relationships with the minimum evolutionary distance necessary to explain them.

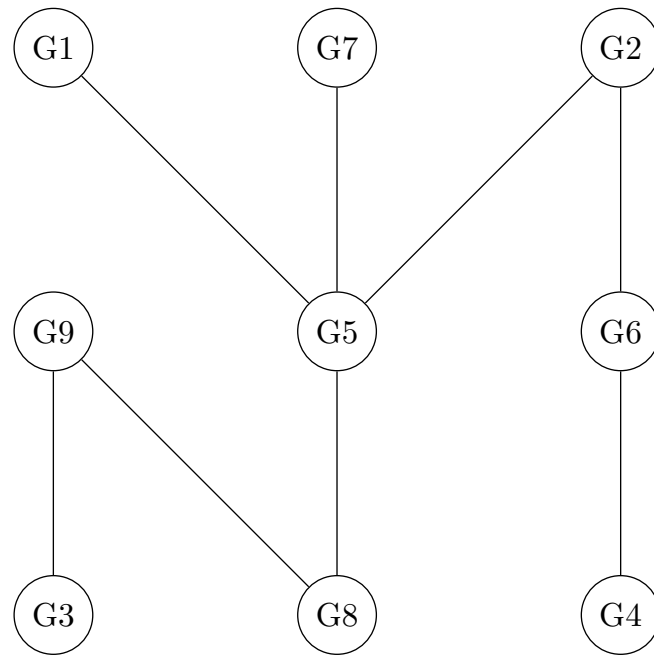
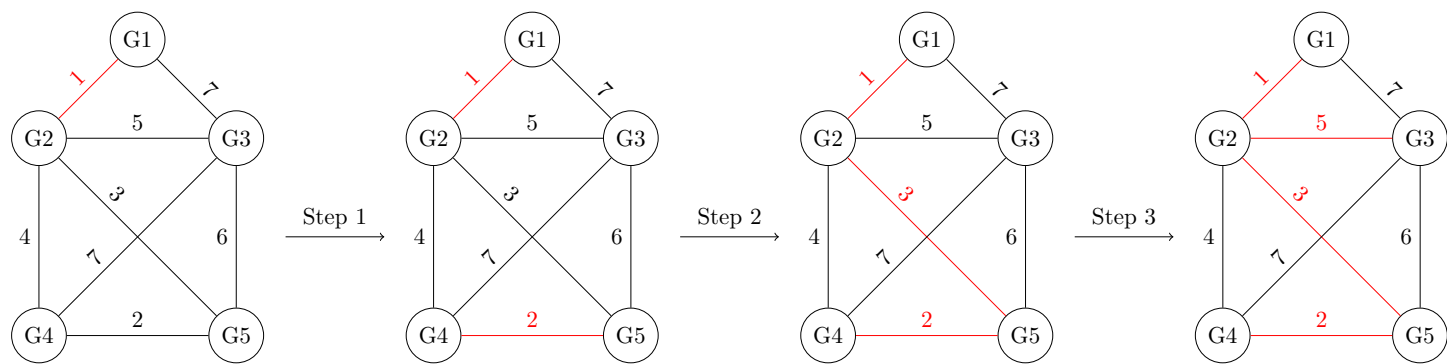


Figure 1.21: Example of a minimum spanning tree.

There are several algorithms that exist for finding MSTs. However, I will focus on Kruskal's algorithm and Prim's algorithm (Figure 1.22). Both of these algorithms are known as greedy algorithms, meaning they add one edge at a time and choose the cheapest available edge at each stage, without considering future consequences. However, they differ in how they find a minimum spanning tree and the order in which edges are added to the graph.

As shown in Figure 1.22, Kruskal's algorithm identifies the edge with the lowest weight and adds it to the graph until all nodes are connected without forming a cycle (Kruskal, 1956). Conversely, Prim's algorithm begins at any node in the graph, evaluates all neighboring edges, and selects the one with the lowest weight (Prim, 1957). The process by which these algorithms construct MSTs and in which sequence edges are added is illustrated in Figure 1.22. An important point to note here is that both algorithms may find more than one MST - these may differ in topology but have the same minimal total weight. However, this will only occur if some edges share the same weight, implying that multiple solutions are available. If, however, only one MST exists, the algorithms will yield identical MST topologies.

Kruskal’s Algorithm



Prim’s Algorithm

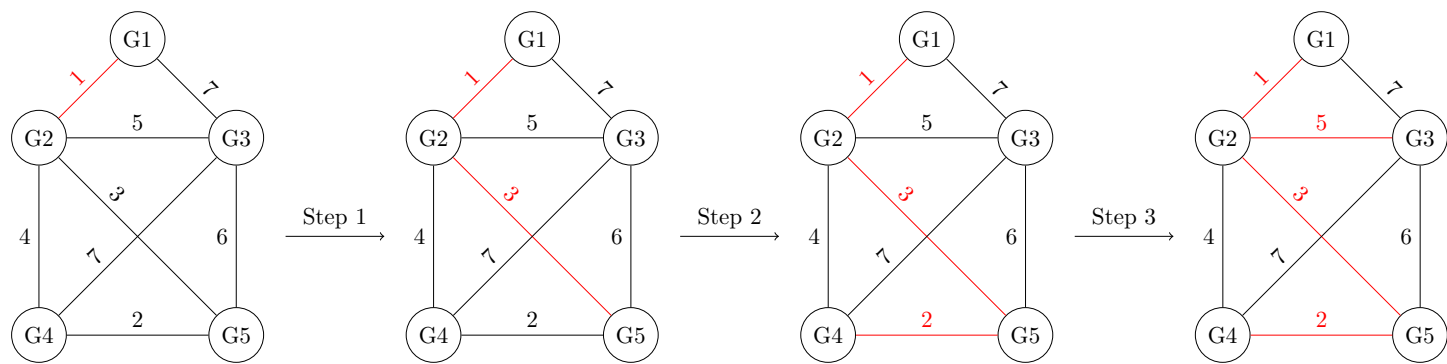


Figure 1.22: Visual representation of algorithms used in finding MSTs.

We can identify the most parsimonious network for example by establishing the shortest distance between genomes, i.e., establishing relationships between genomes sharing the fewest genomic differences. In section 1.3.3, I introduced MLST as a molecular typing method (Section 1.3.3) for studying bacterial diversity. MLST profiles can be visually represented and analysed using graph theory-based approaches like MST (Ribeiro-Gonçalves et al., 2016). In the context of MLST and MST, genomes that share the fewest differences in their alleles—i.e., those with the smallest number of distinct alleles—are positioned closer together on the MST. This reflects their closer genetic or evolutionary relationship compared to genomes that are not directly connected on the tree.

1.5.4 Hamming distance

The Hamming distance is a measure of the difference between two sequences of equal length. It quantifies the number of positions at which the corresponding symbols in each sequence differ (Pinheiro et al., 2012). In the context of MLST, Hamming distance frequently serves as a measure of genetic distance between allelic profiles. This metric calculates dissimilarity by determining the number of allele variants that differ between two isolates (Figure 1.23). A smaller Hamming distance indicates a greater number of shared markers between isolates, suggesting closer relationships, while a larger Hamming distance implies fewer shared markers, indicating that the isolates might be more distantly related.

When calculating a MST for MLST data, the Hamming distance between two organisms is commonly assigned to the corresponding edge as a weight, representing the

number of allele differences between the allelic profiles. This approach is incorporated into software tools developed for analysing MLST schemes using MSTs, such as Phyloviz (Ribeiro-Gonçalves et al., 2016), eBURST (Feil et al., 2004) and EnteroBase (The EnteroBase user’s guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity), which enhance the understanding of microbial population structures and evolutionary dynamics.

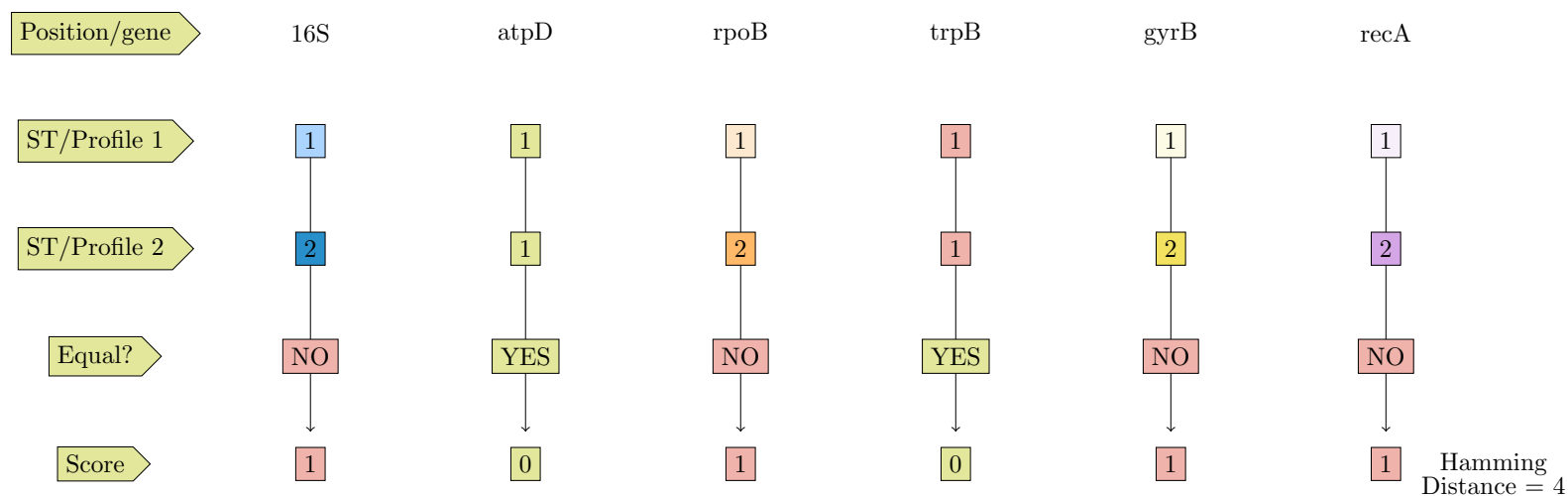


Figure 1.23: How Hamming distance is used to compare MLST profiles. Each gene's allele number in two profiles is compared: a score of 0 is given for matching alleles and 1 for mismatches. The Hamming distance is the sum of these scores, reflecting the number of differences between the two profiles.

1.6 The Genus *Streptomyces*: A Promising Source of Novel Antibiotics

The *Streptomyces* genus of Actinomycete bacteria was first proposed in 1943 by Waksman and Henrici, who described them as spore-forming microbes (Waksman & Henrici, 1943). *Streptomyces* is the largest genus of *Actinomycetota*, widely ubiquitous in nature, and distributed across diverse terrestrial and aquatic environments. These bacteria have also evolved to inhabit extreme environments, including the deep sea and deserts, and are known to engage in symbiotic relationships with animals and fungi (Risidian et al., 2021; Sivakala et al., 2021). For instance, *Candidatus Streptomyces philanthi* associates with the European beewolf (*Philanthus triangulum*), protecting its larvae from fungal infestation (Kaltenpoth et al., 2005; Kaltenpoth et al., 2006), while *Streptomyces* sp. AcH 505 has been shown to promote the mycelial growth of *Amanita muscaria* (commonly known as fly agaric) through the synthesis of auxofuran (Riedlinger et al., 2006).

Streptomyces are generally non-pathogenic, saprophytic soil microbes, with a few exceptions, such as *Streptomyces scabiei*, *Streptomyces acidiscabies*, and *Streptomyces turgidiscabies*, which are known to cause common scab disease in potatoes (Lambert & Loria, 1989; Loria et al., 2006). One of the main characteristics of streptomycetes is their complex life cycle that comprises of three developmental stages; vegetative hyphae, aerial hyphae and spore (Figure 1.24). *Streptomyces* species belong to the Gram-positive bacteria and have an average GC content of 72-75% (Subramaniam et al., 2020). The

Streptomyces genus is distinguished by its large linear chromosome, typically ranging from 6 to 15 Mb, with an average size around 8 Mb (Bury-Moné et al., 2023; Volff & Altenbuchner, 1998). This chromosome features a centrally located origin of replication (oriC) within a conserved region comprising the *dnaA-dnaN-gyrB* genes, from which replication proceeds bidirectionally (Bentley et al., 2002; Bury-Moné et al., 2023). On average *Streptomyces* have approximately 7,000 protein coding genes (Lee et al., 2020). The number of core genes (conserved across the genus) are estimated to range between 600 and 1,018 genes, and are in the central region of the chromosome, while the terminal regions are more variable, often containing genes responsible for secondary metabolite production (Bury-Moné et al., 2023). The *Streptomyces* chromosome is known for its instability, frequently undergoing recombination, which contributes to its genetic diversity and these organisms' adaptability in various conditions (Bury-Moné et al., 2023; Liroy et al., 2021).

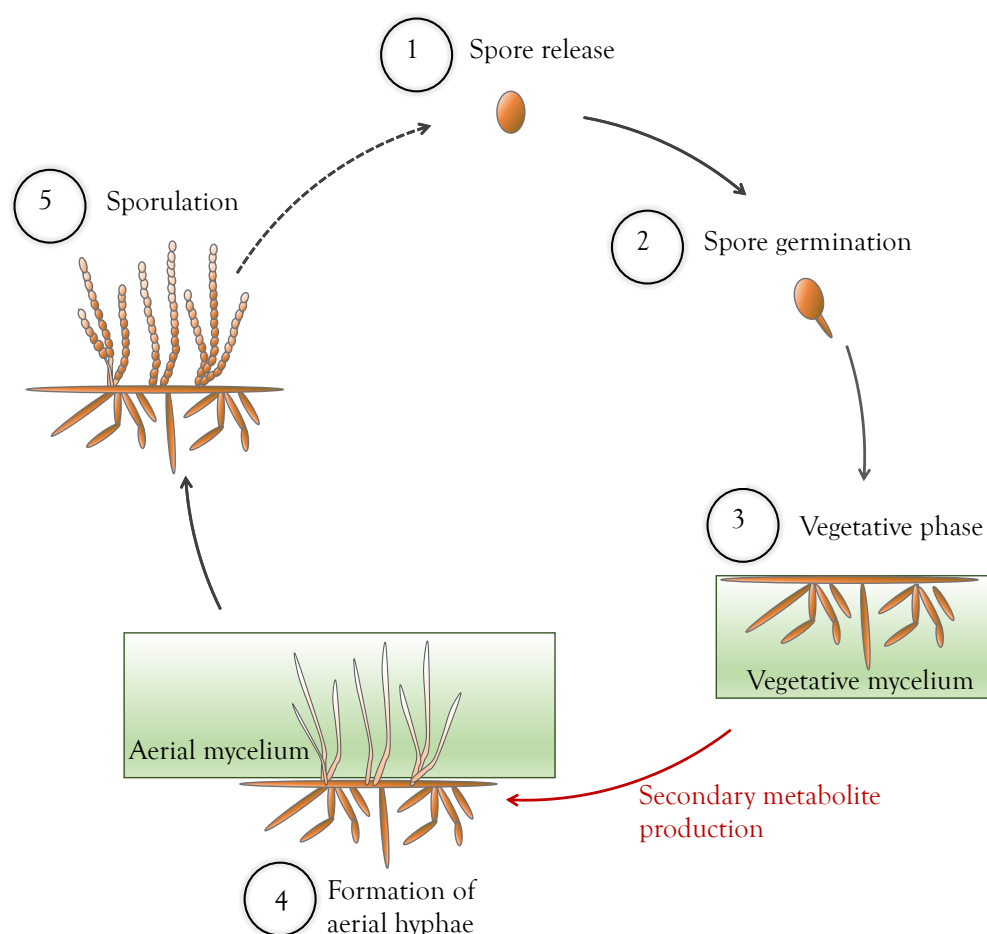


Figure 1.24: The *Streptomyces* life cycle. begins with the release of a free spores dispersed in the environment, serving as the starting point for new colonies. When conditions become favourable, such as the presence of nutrients and suitable environmental factors, the spores undergo germination, leading to development of a germ tube. Following germination, the spores give rise to vegetative hyphae, forming a network of mycelia that colonise the substrate (eg. soil) and further utilise various nutrients in the environment. As mycelia grow, they differentiate into aerial hyphae, rising above the substrate and forming a visible layer called aerial mycelium. It is during this transition phase from vegetative mycelia to aerial mycelia when *Streptomyces* synthesise secondary metabolites (red arrow). The synthesis and excretion of these natural products into the surrounding environment by microorganisms (in this case *Streptomyces*) serves as a natural defense mechanism against nearby competitors, and may also protect from other environmental stresses such as changes in temperature, light, pH, moisture or phage infection (Koskella & Brockhurst, 2014; Tyc et al., 2017). In the final stage, aerial mycelia undergoes further differentiation, leading to the formation of chains of spores which, once mature, are released to the environment to continue the life cycle. Adapted from Barka et al., 2016.

Considerable interest in the members of *Streptomyces* genus was sparked by discovery of the anticancer medication actinomycin and the antimicrobial agent streptomycin (Schatz et al., 1944; Waksman & Woodruff, 1941). Over 650 taxa have so far been identified in the genus *Streptomyces* (Labeda et al., 2012). Bioactive compounds produced by *Streptomyces* have been found to have antimicrobial, anticancer, antifungal and antiparasitic activity (Bolourian & Mojtahedi, 2018; Kamarudheen & Rao, 2018; Procópio et al., 2012; Schatz et al., 1944). Clinically important species within this genus include *Streptomyces griseus* (producer of streptomycin), *Streptomyces lydicus* (producer of natamycin, lydimycin, and streptolydigin), and *Streptomyces clavuligerus* (producer of clavulanic acid and cephamycin C), as well as *Streptomyces coelicolor* A3(2) (producer of actinorhodin and non-ribosomal peptide Calcium-Dependent Antibiotic (CDA)), a representative organism of the genus (Bentley et al., 2002).

The impact of *Streptomyces* on antibiotic use in both humans and animals is significant, as this genus is responsible for producing approximately 80% of all clinically approved antibiotics (Procópio et al., 2012). A timeline of the most important therapeutics produced by the members of *Streptomyces* genus is shown in Figure 1.25. *Streptomyces*-derived antibiotics, including clavulanic acid and various tetracyclines, are among the most widely used (Figure 1.26 and Figure 1.27). Clavulanic acid, produced by *Streptomyces clavuligerus*, is a β -lactamase inhibitor that enhances the efficacy of penicillins by inhibiting β -lactamases, enzymes produced by pathogenic bacteria such as *Staphylococcus* that would otherwise reduce the effectiveness of the penicillin (Beytur et al., 2015). Additionally, many tetracycline antibiotics are derived from various *Strep-*

tomyces species - such as chlortetracycline (produced by *Streptomyces aureofaciens*), oxytetracycline (produced by *Streptomyces rimosus*), and tetracycline (produced by *Streptomyces aureofaciens*) (Biffi et al., 1954; Darken et al., 1960; Petkovic et al., 2006).

In November 2023, the UK government published the Third UK One Health Report, which indicates that penicillins, including Amoxicillin+Clavulanic acid, are the most widely consumed antibiotics in humans, accounting for approximately 66% of all antibiotic sales, followed by tetracyclines, which represent 9% of human antibiotic use (Figure 1.26 and Figure 1.27). Furthermore, the UK Veterinary Antibiotic Resistance and Sales Surveillance Report for 2022 indicated that tetracyclines—including chlortetracycline, doxycycline, oxytetracycline, and tetracycline—were the most sold antibiotics for animal use from 2014 to 2022 (Figure 1.27).

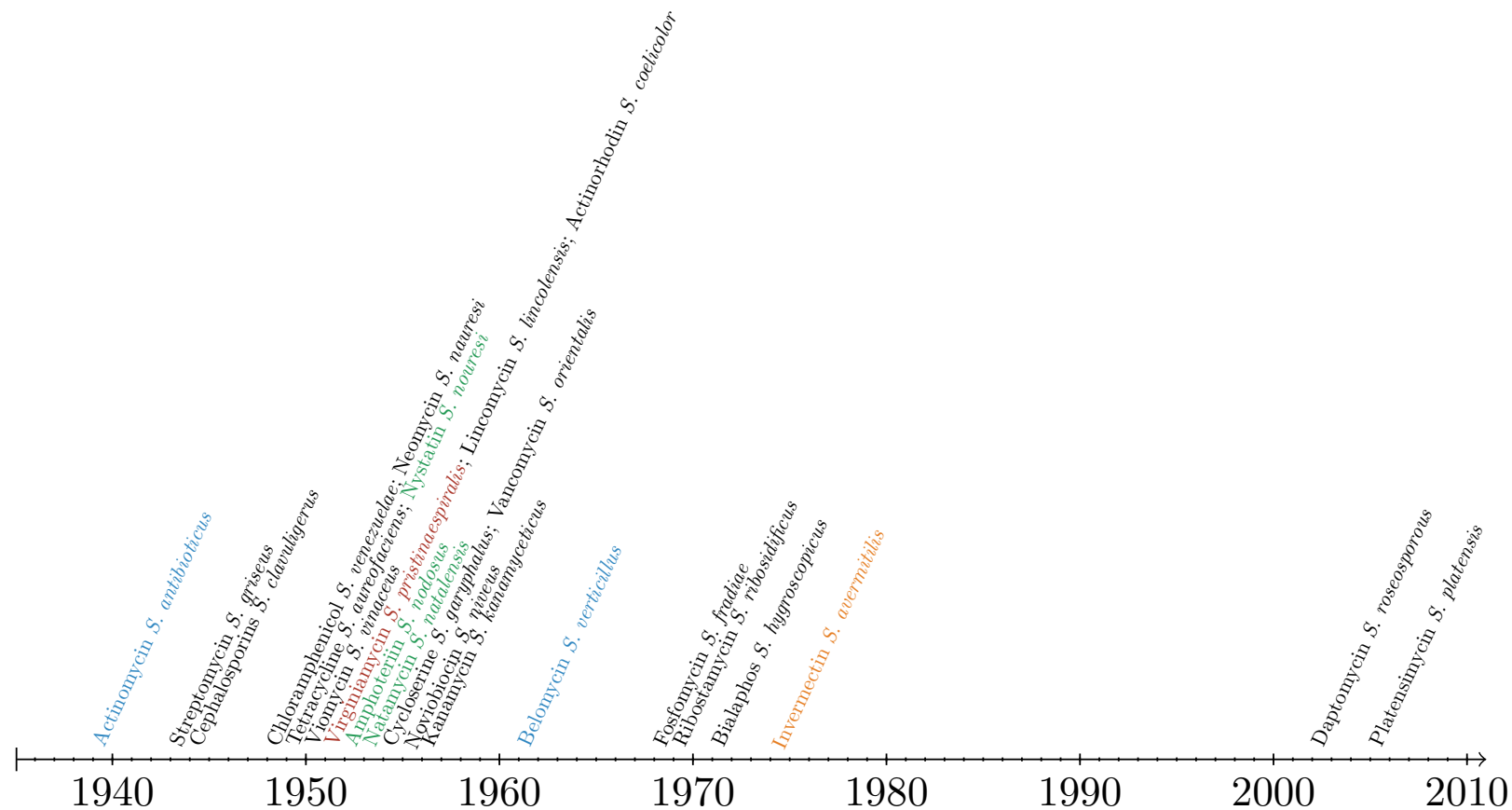


Figure 1.25: Key discoveries of secondary metabolites produced by members of the genus *Streptomyces*. Antimicrobials are shown in black, antifungals in green, antiparasitic agents in orange, anticancer in blue, and agricultural treatments in red. Figure adapted from Procópio et al., 2012.

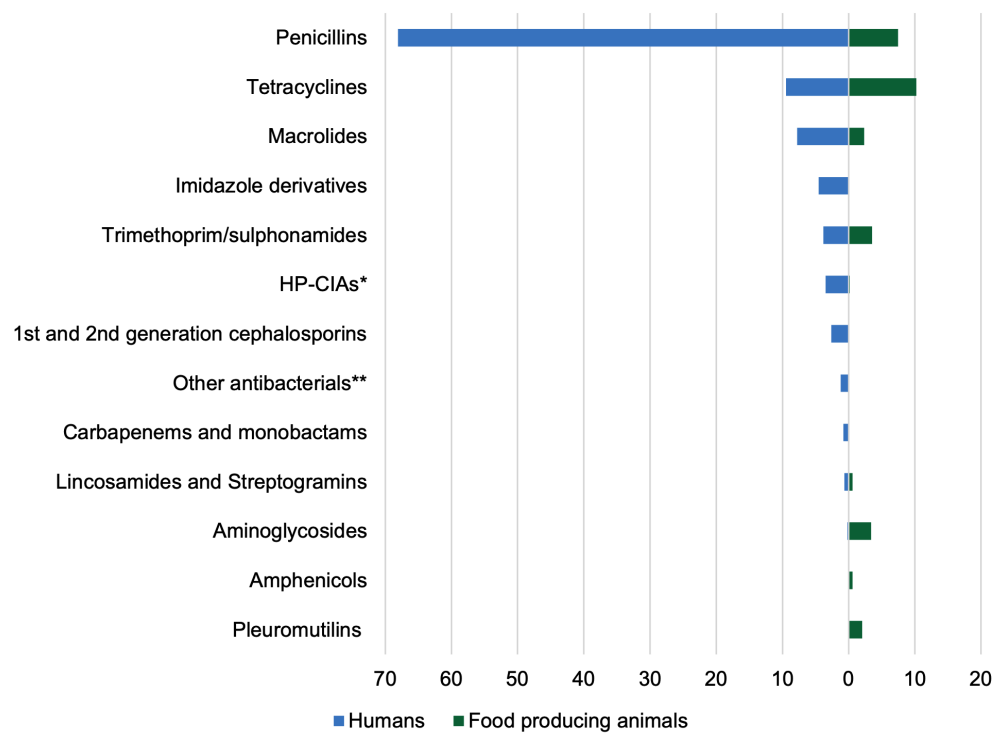


Figure 1.26: Most consumed antimicrobial classes by humans and food-producing animals (mg/kg), 2019. Figure adapted from the UK Veterinary Antibiotic Resistance and Sales Surveillance Report for 2022 (available at https://assets.publishing.service.gov.uk/media/663373da1834d96a0aa6cfd5/2779033-v1-VARSS_2022__April_2024_Update_.pdf).

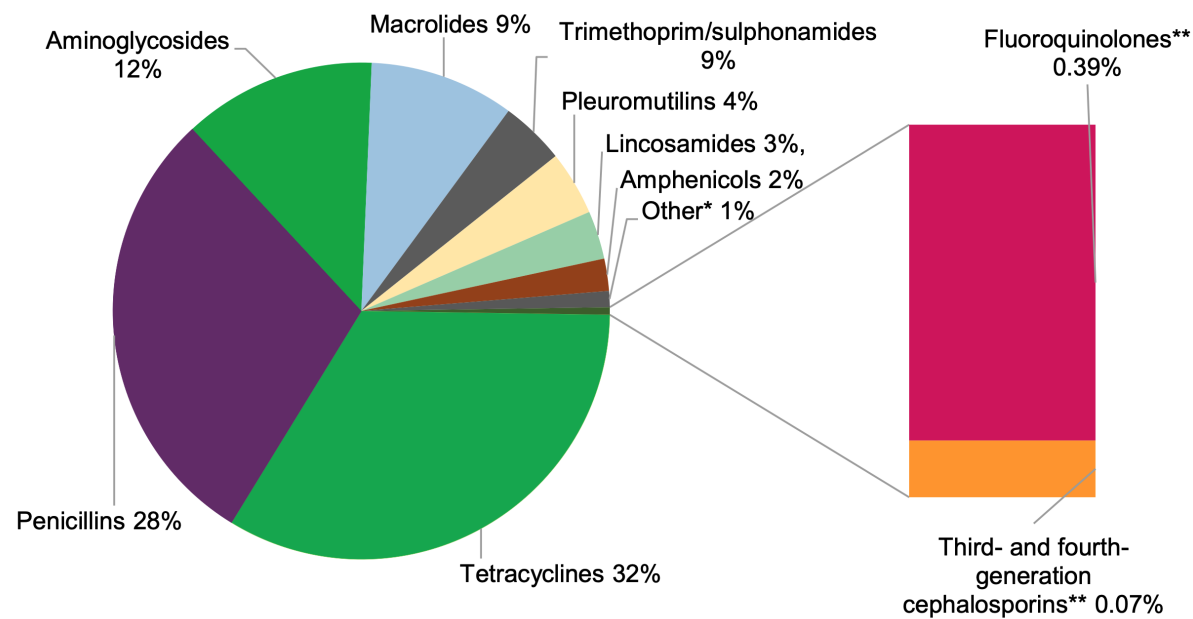


Figure 1.27: Most consumed antimicrobial classes by humans and food-producing animals (mg/kg), 2019. Figure adapted from the Third UK One Health Report (available at https://assets.publishing.service.gov.uk/media/656488f11524e60011a100f8/_2681096-v1-Third_UK_One_Health_Report.PDF).

Furthermore, *Streptomyces* are well-regarded for their agricultural benefits, as they predominantly act as commensals in the rhizosphere and produce a range of secondary metabolites effective against plant pathogens (see Figure 1.28). For example, *Streptomyces lydicus* has been approved as a biopesticide for against fungal pathogen *Rhizoctonia solani* (Yuan & Crawford, 1995).

Recent discoveries of novel candidate antimicrobials produced by *Streptomyces* include picolinamycin (Maiti et al., 2020), formacamycins (Qin et al., 2017) (which are believed to represent novel antibiotic classes), and peptide 5812-A/C (Vasilchenko et al., 2020). All three exhibit activity against a wide spectrum of multi-drug-resistant bacterial pathogens, including *Staphylococcus aureus* and *Enterococcus faecium*. Furthermore, the most comprehensive study of genome mining, conducted on 1,100 publicly available *Streptomyces* genomes, revealed a predicted ability to synthesize 34 major classes of biosynthetic gene clusters (BGCs), detecting over 1,062 non-ribosomal peptide synthetase (NRPS) and 981 type I polyketide synthase (PKS) BGCs (Komaki et al., 2018). Moreover, recent research has demonstrated that *Streptomyces* isolates, even those classified as the same species, can exhibit significant variation in the BGCs they harbor. Pangenomic analysis has further revealed that *Streptomyces* species possess an open pangenome, reflecting their remarkable genetic diversity and adaptability to various environments (Otani et al., 2022). Streptomycetes' ability to synthesise a wide variety of secondary metabolites and continuing discoveries of new natural products by this genus suggests their great potential as novel drugs reservoirs that can assist with control of AMR.

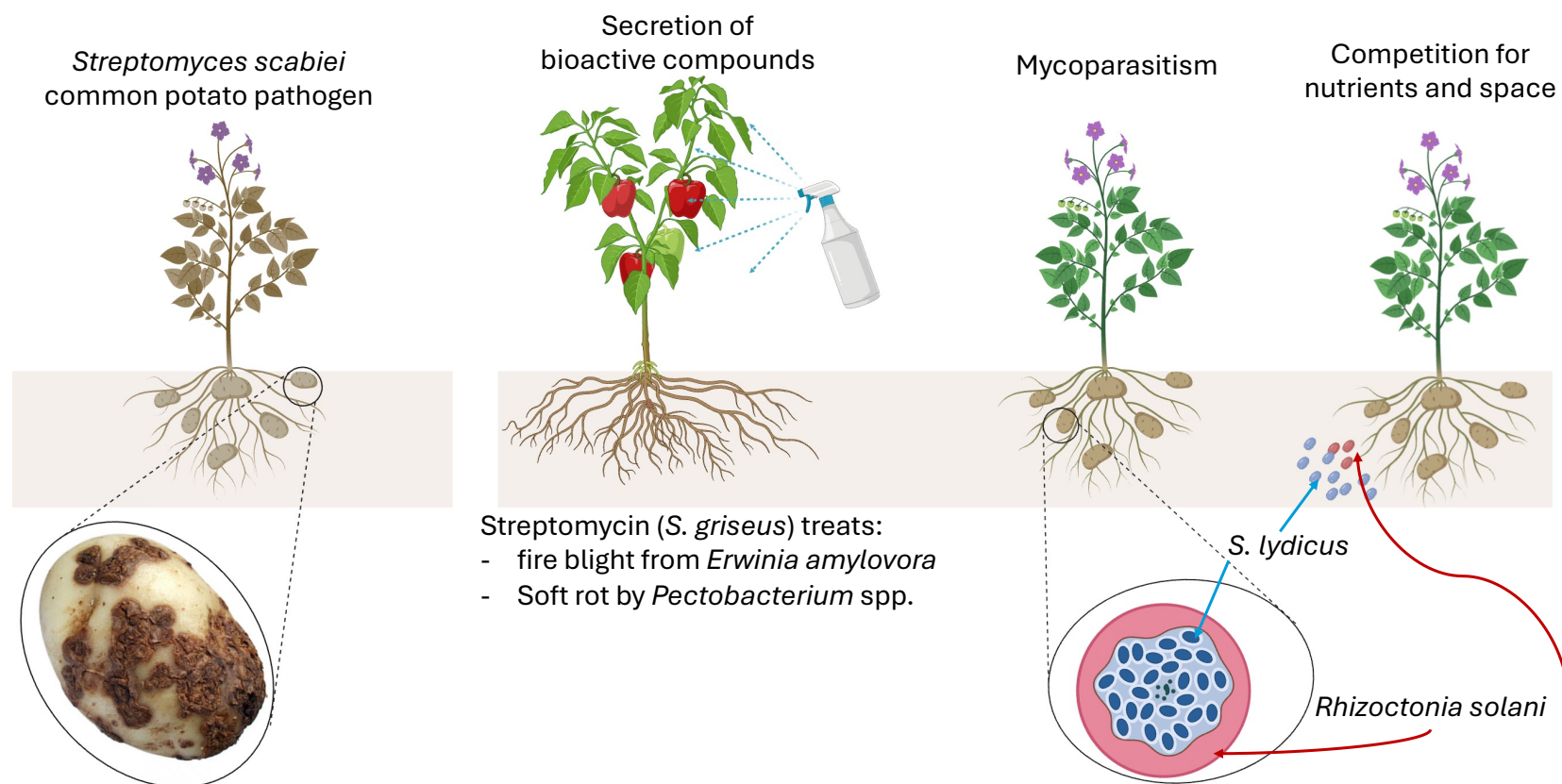


Figure 1.28: Diverse roles of *Streptomyces* species in agriculture. *Streptomyces scabiei*, a plant pathogen, is known for causing common potato scab, which can significantly reduce both the yield and quality of potatoes (left). *Streptomyces griseus* has been recognised for its production of streptomycin, an antibiotic that is vital in controlling fire blight caused by *Erwinia amylovora* and soft rot caused by *Pectobacterium* spp. (Le et al., 2022). Meanwhile, *Streptomyces lydicus* exhibits notable biocontrol activity against *Rhizoctonia solani* by colonizing the root tips of plants. It parasitises *R. solani* through mycoparasitism, competing with the pathogen for nutrients and space, thereby aiding in the suppression of soil-borne diseases. (Yuan & Crawford, 1995)

1.7 Taxonomic incongruence in the genus *Streptomyces*

Although *Streptomyces* is arguably the most extensively studied genera within the *Actinomycetota* phylum due to its pharmaceutical significance, its taxonomy remains contested, with significant species-level misclassifications (Mispelaere et al., 2024). The consequent risk of including too distantly-related genomes in the analysis poses a substantial limitation for effective pangenomic analyses aimed at identifying novel secondary metabolites with potential clinical and agricultural applications.

The taxonomic classification of *Streptomyces* has undergone significant changes over the years. Initially, members of this genus were believed to be eukaryotes due to their distinctive life cycle, which bears similarities to those of filamentous fungi (Procópio et al., 2012). When *Streptomyces* was first proposed as a bacterial genus by Waksman and Henrici, its classification was based primarily on morphological characteristics and cell wall chemotypes, such as spore color and spore chain morphology (Waksman & Henrici, 1943). This led to the genus being placed in the family *Streptomycetaceae*. With advancements in sequencing technology, *Streptomyces* species have since been extensively studied based on differences in their 16S rRNA sequences (Labeda et al., 2012).

Studying *Streptomyces* diversity through 16S rRNA sequences has resulted in numerous reclassifications both within and between genera. One of the most notable reclassifications is the frequent inclusion and exclusion of the sister genus *Kitasatospora* (Anderson & Wellington, 2001). *Kitasatospora* was first described as a distinct genus by Omura et al. in 1982. This was based on specific morphological and biochemical

characteristics that differentiated it from *Streptomyces*. Despite these morphological differences, *Kitasatospora* was included within the *Streptomyces* genus due to the high similarity of their 16S rRNA sequences (Wellington et al., 1992b). However, only five years later, *Kitasatospora* was reclassified as a distinct genus after it was demonstrated that members of *Kitasatospora* consistently formed a stable monophyletic clade separate from *Streptomyces* when the entire 16S rRNA gene sequences were analysed (Zhang et al., 1997a).

The most comprehensive study of the *Streptomyces* genus, conducted in 2012, estimated the presence of 650 taxa based on 16S rRNA sequence differences (Labeda et al., 2012). However, distinct *Streptomyces* species often share highly similar phenotypes and 16S rRNA sequences, leading to taxonomic inconsistencies. For instance, other studies found that 16S rRNA was found unable to distinguish between members of order *Streptomycetales* from those in *Frankiales*, *Catenulisporiales* and *Streptosporangiales* or from *Micrococcales*, *Streptosporangiales* and *Catenulisporales* (Verma et al., 2013). However, whole genome-based taxonomic classification of genus *Streptomyces* provided cleaner separation of suborders *Streptomycineae* and *Frankineae* and for the species within the *Mycobacteraceae* family that could not be resolved with genomic data contained only within the 16S rRNA gene (Alam et al., 2010). The examination of evolutionary relationships among *Streptomyces* species using the 16S rRNA gene and whole genome distance measures also exhibited incongruences, suggesting that current taxonomic classification of these organisms is uncertain (Chevrette et al., 2019b).

Furthermore, it is known that distinct species can share identical or highly identical 16S rRNA sequences (Table 1.4). This similarity poses a significant challenge for species-

level identification, as relying solely on 16S rRNA gene differences may lead to numerous misclassifications. If distinct species share the same or highly similar 16S sequences, distinguishing between them based on this gene alone becomes impossible.

It is also not unusual for *Streptomyces* species to contain multiple non-identical copies of the 16S rRNA gene. The presence of these divergent copies within a single species complicates taxonomic classification, as it raises the question of which, if any, sequence should be prioritised for accurate classification and identifications of bacteria. This intragenomic variability can result in discrepancies when assigning species within the *Streptomyces* genus, potentially leading to conflicting phylogenetic signals. For example, *Streptomyces caelicus* possesses five distinct 16S rRNA gene copies, each varying in its degree of similarity to other *Streptomyces* species. Consequently, these copies are placed in different clades on the phylogenetic tree (Figure 1.29).

Table 1.4: Example of *Streptomyces* species sharing identical or highly similar 16S sequences. Figure adapted from Komaki, 2021.

Similarity (%)	Different species
100	<i>S. coelescens</i> \neq <i>S. violaceoruber</i>
100	<i>S. chattanoogensis</i> \neq <i>S. lydicus</i>
100	<i>S. chrestomyceticus</i> \neq <i>S. paromomycinus</i>
100	<i>S. variabilis</i> \neq <i>S. labedae</i>
100	<i>S. fulvissimus</i> \neq <i>S. fulvorobeus</i>
100	<i>S. phaeogriseichromatogenes</i> \neq <i>S. graminearus</i>
99.97	<i>S. angustmycinicus</i> \neq <i>S. lydicamycinicus</i>
99.97	<i>S. caniferus</i> \neq <i>S. hygrosopicus</i> subsp. <i>glebosus</i>
99.97	<i>S. caniferus</i> \neq <i>S. libani</i> subsp. <i>rufus</i>
99.97	<i>S. nigrescens</i> \neq <i>S. tubercidicus</i>

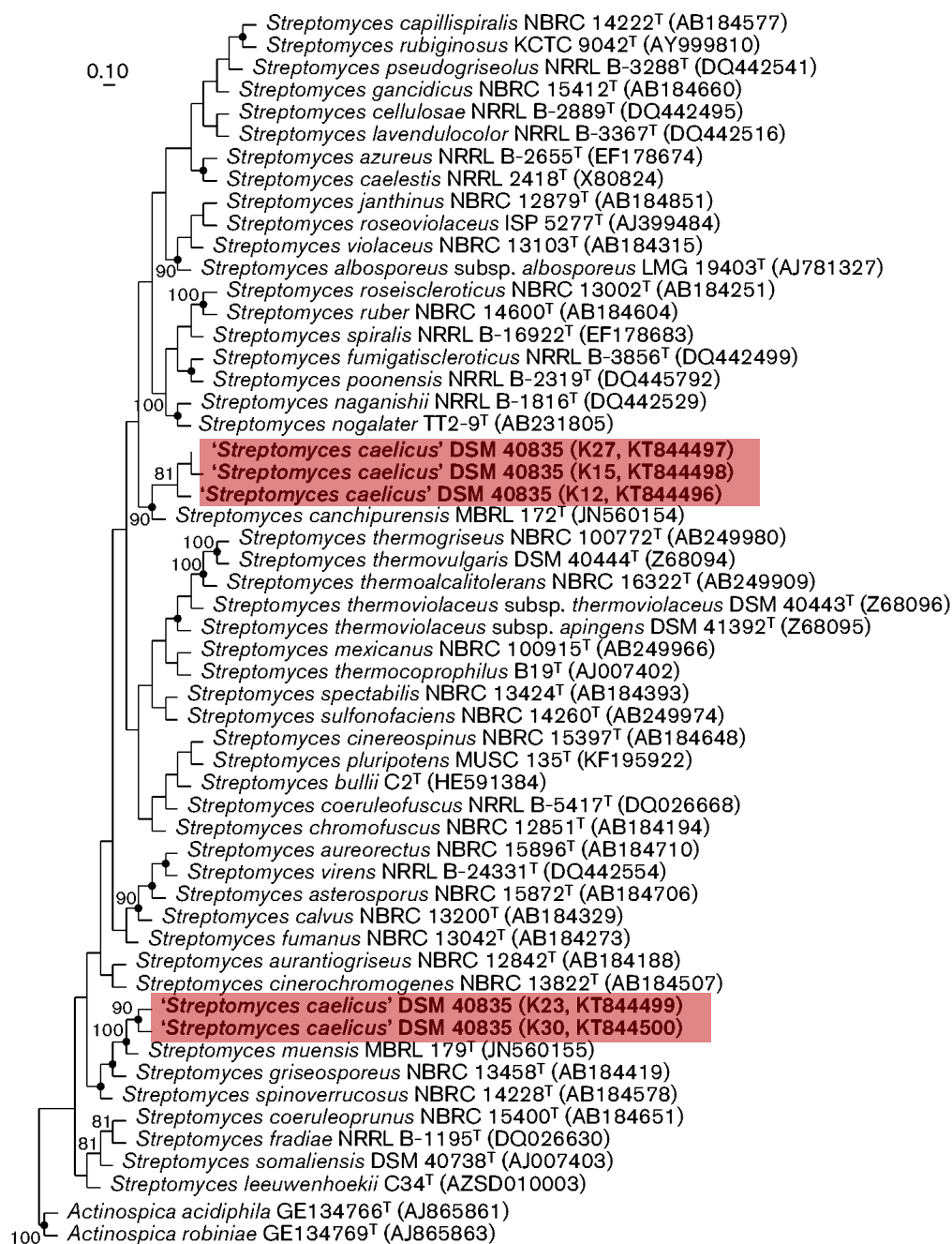


Figure 1.29: Maximum-likelihood tree of 16S rRNA sequences showing the placement of *Streptomyces caelicus* (red box) based on five different copies from the same isolate. Figure taken from Wink et al., 2017.

There have been attempts to resolve conflicts within the genus through the MLSA approach. The most commonly used genes for MLSA characterisation of species within the *Streptomyces* genus are *atpD*, *gyrB*, *recA*, *rpoB*, and *trpB* (Guo et al., 2008; Rong & Huang, 2010b). Through phylogenetic analysis focusing on variation within these genes, previously unresolved taxonomic differences between species were revealed, overcoming some limitations of 16S rRNA gene sequencing (Labeda, 2011). However, while MLSA analysis has enhanced the phylogenetic resolution for certain species like *S. griseus*, it has not completely resolved conflicts within the genus (Guo et al., 2008; Labeda, 2011). Additionally, the current canonical *Streptomyces* MLST scheme available from the pubMLST comprises six marker genes (16S rRNA, *atpD*, *gyrB*, *recA*, *trpB*, *rpoB*) and 237 STs. Despite recent increase in available genomic sequences (as mentioned in section 1.3.3; Figure 1.12), only two new STs were reported since 2016, and the resolution of the MLST *Streptomyces* scheme remains unknown (Jolley et al., 2018).

In recent years, the increasing use of whole-genome taxonomy has influenced the taxonomic classification of various bacterial lineages, including those belonging to the *Streptomyces* genus. In 2023, Nikolaidis et al. conducted the most comprehensive phylogenetic study of the *Streptomyces* genus to date, analysing 318 core protein orthologous groups shared across 218 *Streptomyces* species. This study significantly improved the topological stability of the phylogenetic tree and provided clearer insights into the placement of key strains, including pharmaceutically significant species such as *S. clavuligerus*, *S. lydicus*, *S. griseus*, *S. coelicolor*, and the plant pathogen *S. scabiei* (Figure 1.30). However, many *Streptomyces* genomes remain unexplored; as of July 8,

2021, the NCBI database listed 2,276 genomes belonging to the genus *Streptomyces* (Schoch et al., 2020). The classification of these genomes could be critical for further meaningful analyses and may yield novel insights into the diversity of *Streptomyces* genus.

The application of whole-genome taxonomic approaches has led to numerous reclassifications within the *Streptomyces* genus and even provided the foundation for the establishment of new genera. In 2022, Madhaiyan et al., 2022 reclassified several *Streptomyces* species into six newly proposed genera—*Actinacidiphila*, *Mangrovactinospora*, *Streptantibioticus*, *Wenjunlia*, *Peterkaempfera*, and *Phaeacidiphilus*—based on whole-genome analyses and phenotypic comparisons. This reclassification, which included 12 species from the *Streptomyces* genus and three from the *Streptacidiphilus* genus, expanded the family *Streptomycetaceae* to encompass 12 genera. The use of high-resolution whole-genome sequencing has allowed for a more precise subdivision of the complex *Streptomyces* genus. However, a crucial question remains: how many other genera within this group still require reclassification?

There is also a notable gap in our understanding regarding the congruence between whole-genome distance methods and whole-genome phylogenies. Better understanding of the discriminatory power of these approaches might prove useful for achieving a more robust and accurate classification system within the *Streptomyces* genus. Better understating of evolutionary relationships within the genus *Streptomyces* is necessary to avoid miscalculation of pangenomes between unrelated genomic sequences.

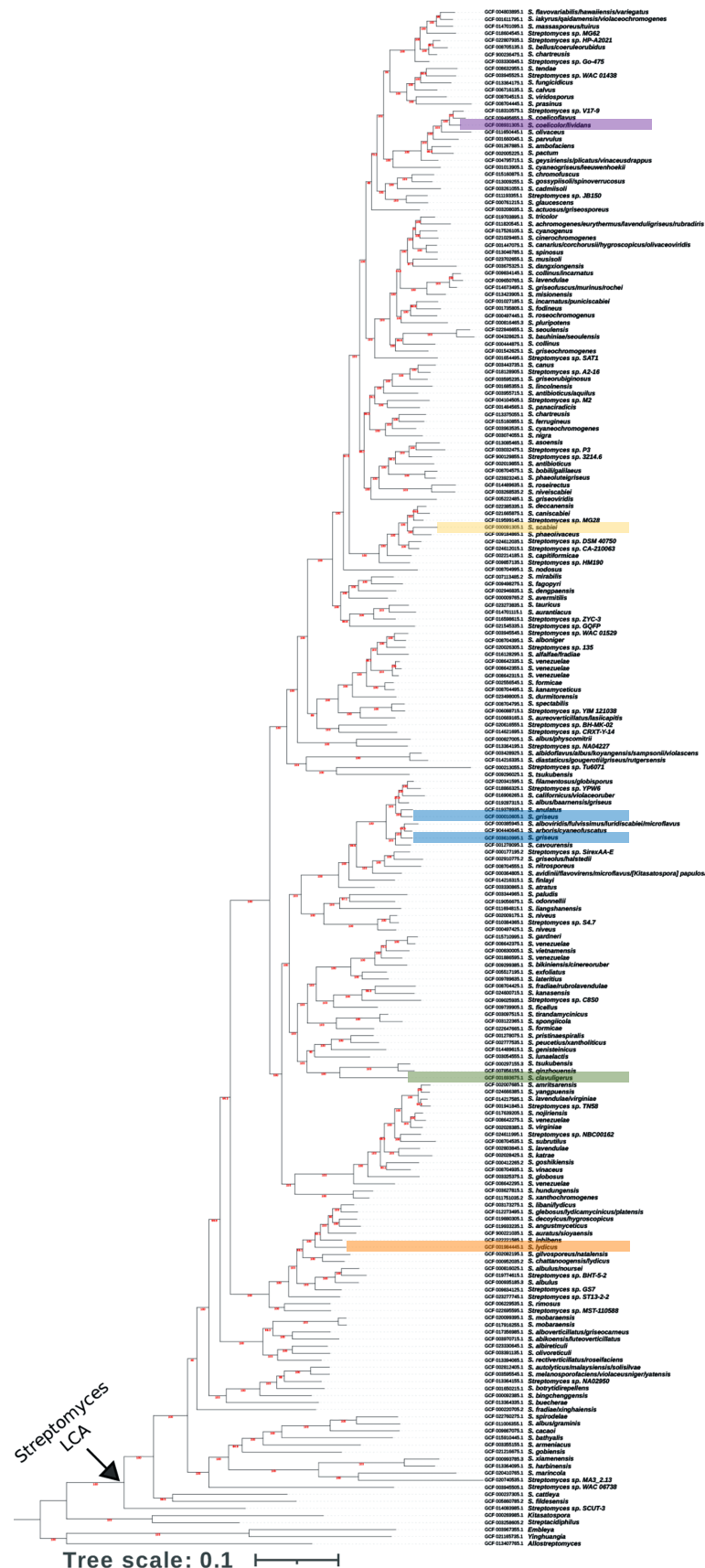


Figure 1.30: A core-genome phylogeny of 218 *Streptomyces* species, based on 318 core protein orthologous groups, highlights the placement of key pharmaceutical species - *S. griseus* (blue), *S. clavuligerus* (green), *S. coelicolor* (purple), *S. lydicus* (orange) and the plant pathogen *S. scabiei* (yellow). Figure adapted from Nikolaidis et al., 2023.

1.8 Thesis outline

In this thesis, I investigate the taxonomic relationships among all publicly available *Streptomyces* genomic sequences using several classification approaches to establish a reliable taxonomy of *Streptomyces* to improve classification of *Streptomyces* genomic sequences for pangenomic analyses. In the process, I aim also to answer the following questions:

1. What discriminatory power do these approaches offer?
2. Do these methods agree with each other, and with current taxonomic opinion?
3. What useful information about the taxonomic structure of *Streptomyces* do they reveal?

In chapter 2, I perform comprehensive classification of *Streptomyces* using 16S and whole-genome distance methods, to delineate the genomic landscape of this group of organisms, and to identify and interpret incongruences between the two approaches. In this study, I use all 48,981 publicly available full-length 16S rRNA *Streptomyces* sequences obtained between October 26, 2020, and November 10, 2020, as well as 2,276 *Streptomyces* genomes available as of July 8, 2021. Using 14,239 distinct full-length *Streptomyces* 16S rRNA sequences, I reconstruct the most comprehensive 16S phylogeny of *Streptomyces* known at the time of writing. I investigate the effect of clustering these 16S rRNA sequences into OTUs across a range of threshold identities, including previously recommended 97% and zOTU identity thresholds, to determine the theoretical taxonomic resolving power of 16S sequences. I also examine the distribution of individual

16S rRNA sequences across *Streptomyces* genomes to determine whether a mapping exists between 16S rRNA sequences and whole genome-derived species boundaries, with a specific goal to understand whether zOTUs are congruent with genome-based classifications.

In chapter 3, I survey all available *Streptomyces* genomes to update the canonical pubMLST *Streptomyces* scheme, incorporating 568 novel STs. Using this updated scheme, I investigate taxonomic relationships among all *Streptomyces* genomic sequences using MLST and whole-genome based methods, to delineate the extent of genomic diversity and investigate congruence between these two methods. Using graph-based analyses I examine the divisions within the genus based on shared common alleles and determine congruencies between MLSA, MLST and whole genome taxonomy. My findings support claims from other taxonomists that current *Streptomyces* classifications require further resolution.

In chapter 4, I focus on a comprehensive classification of *Streptomyces* using core-genome phylogeny (SCOGs) and a whole-genome distance method (ANI). I use whole genome approaches to achieve a highly resolved and robust taxonomy to improve input sets for pangenomic analyses, and resolve the taxonomic conflicts identified in previous chapters. I present a phylogeny reconstructed from all single-copy orthologs shared across complete publicly-available genomes currently assigned to *Streptomyces*. I discuss the integration of various taxonomic classification methods, including ANI analysis and SCOGs phylogeny, along with statistical approaches and graph algorithms, to delineate candidate genus boundaries within *Streptomyces*. This approach was designed to better understand the taxonomic structure of the genus and investigate the congruence between

SCOGs and ANI whole-genome methods.

In chapter 4 I will also expand on the investigation of SCOGs shared across *Streptomyces* to identify and investigate evidence of potential horizontal gene transfer and its effects on the taxonomic classification of this group.

In the final chapter (chapter 5), I summarise my findings from the previous chapters, and discuss future research directions.

16S rRNA phylogeny and clustering is not a reliable proxy for genome-based taxonomy in *Streptomyces*

2.1 Introduction

2.1.1 Motivation

As previously discussed in Chapter 1 section 1.1, the current range of available antibiotics are becoming increasingly less effective owing to the rapid emergence of AMR, which needs urgent attention (Murray et al., 2022a). Although *Streptomyces* produce over 80% of all clinically approved antibiotics and are a promising source of novel natural products, drug discovery using pangenomic and comparative genomic analyses could lead to false interpretations due to their contested taxonomy (Chevrette et al., 2019b) (general introduction section 1.2.4). The uncertainty in the classification of microorganisms can make drug discovery efforts based on genomics misleading. For example, if a new antibiotic-producing strain is misclassified or placed in the wrong group, researchers may fail to recognise its full potential for producing valuable drugs, overlook promising strains, or misdirect research efforts, all of which could hinder the discovery and development of new bioactive compounds.

The previous most comprehensive 16S phylogenetic study of the genus *Streptomyces* was carried out in 2012, although this gene is still routinely used for identification of *Streptomyces* isolates and for metabarcoding studies (Labeda et al., 2012). Hence, reconstruction of the 16S phylogeny for *Streptomyces* to consider all available sequences is overdue to better understand the suitability of this gene for assignment of taxonomy in *Streptomyces*.

2.1.2 Public 16S databases

Recent advances in sequencing technology have had a revolutionary impact on a data availability (Buermans & Dunnen, 2014), and enabled the production of large public sequence datasets for 16S sequences. The most widely used 16S rRNA databases are SILVA (<https://www.arb-silva.de>) (Yilmaz et al., 2014), Greengenes (<https://greengenes.secondgenome.com>) (DeSantis et al., 2006) and the Ribosomal Database Project (RDP) (<https://rdp.cme.msu.edu>) (Cole et al., 2014). Each of these databases acquires their 16S rRNA sequences from repositories of record such as Genbank and EMBL/DDBJ. Unfortunately, these databases are not frequently maintained. SILVA is the only 16S database updated since 2020. Greengenes was last updated in 2013, and RDP in 2016. Despite this, each remains in frequent use (Babis et al., 2024; McDonald et al., 2024).

These databases collect sequences from similar sources, but assign taxonomic annotations using different approaches. For instance, taxonomy annotation in SILVA and Greengenes is based on automated *de novo* phylogeny construction and by implementation of database-specific prediction approaches (McDonald et al., 2012; Yilmaz et al., 2014). In contrast, RDP uses a Naïve Bayesian Classifier (NBC) to assign taxonomy trained on Bergey’s taxonomy (Wang et al., 2007). The annotation error rate is a poten-

tial concern for these databases, as approximately 17% of annotations were incorrectly mapped in SILVA and Greengenes, and 10% in RDP. Annotation errors are known to include conflicting annotations up to phylum rank, despite 100% sequence similarity (Edgar, 2018a). Additionally, all databases implement sequence quality checks, but only Greengenes checks for and removes chimeric sequences (DeSantis et al., 2006).

2.1.3 Considerations for construction of a 16S phylogenetic tree

Acquiring sequences

The first step in constructing any phylogenetic tree is the acquisition of relevant sequences. A phylogenetic tree can be inferred from any set of sequences, but what sequences are included in the phylogeny determine whether the phylogeny is biologically meaningful (Kapli et al., 2020). Thus, careful consideration is essential to avoid calculation of meaningless phylogenies or challenges for interpretation, or both.

Most phylogenetic methods, such as those implemented in RAxML (Kozlov et al., 2019), produce a bifurcating tree (section 1.4.3). This structure implies a single, recognisable common ancestor for all sequences, which is important for ensuring a meaningful evolutionary interpretation. If the input sequence set fails to meet these criteria—for example, includes sequences from chicken collagen, bacterial beta-lactamase, and archaeal DNA polymerase—the fundamental assumption of building a phylogenetic tree is not met. Sequences that share a recognisable, recent common ancestor are termed homologous, and ensuring homology among selected sequences is essential for calculating meaningful phylogenies (Kapli et al., 2020).

Identifying homologous sequences isn't a one-size-fits-all approach; it depends on the question at hand. Once suitable criteria are established, the selection of appropriate

sequences can be carefully determined, including sourcing them from relevant databases (section 2.1.2) or experimental data.

As discussed in section 1.3.3, microbial diversity can be studied using varying amounts of biological information, ranging from a single gene to an entire genome. When investigating the evolutionary relationships of a collection of isolates and inferring a phylogenetic tree, we can identify shared sequences, such as the 16S rRNA gene, a subset of housekeeping genes commonly used in MLSA and MLST, or core sequences. However, relying on a single marker gene may not capture the necessary level of variation needed to accurately differentiate organisms of interest. In such cases, selecting multiple sequences can be more effective. These sequences are chosen from the organisms being studied to ensure that the observed variation is suitable for reconstructing their evolutionary history. By using a concatenated set of aligned sequences instead of a single gene, the resulting phylogenetic tree provides a more reliable representation of evolutionary relationships, reducing biases that could arise from the variability of individual genes, if these genes are not affected by horizontal gene transfer (HGT) or recombination.

When inferring phylogenies from sequences obtained from database(s), there are several crucial aspects to consider:

1. Legal considerations, as sequences may be protected by intellectual property.
2. Database maintenance to ensure that errors are corrected and that the database remains comprehensive and up-to-date with current knowledge, minimising bias in observations.

3. Quality checks to understand potential limitations and data quality issues (eg. contamination, misannotation, chimeras).

Quality control and sequence clustering

Using sequences from a public database without quality checks or implementing data curation processes should be avoided. Several factors can impact the accuracy of phylogenetic analyses:

1. Redundant sequences: These sequences, if included in phylogenetic analyses, not only consume computational resources unnecessarily but also do not bring new biological information (Zou et al., 2018). However, in the context of epidemiology, removing such samples may not always be appropriate, as they could still provide important information for outbreak tracking.
2. Chimeric sequences: These are artifacts formed due to PCR errors, where at least two biological sequences are incorrectly joined together (Gonzalez et al., 2005). Inclusion of chimeric sequences in phylogenetic analyses can result in misleading evolutionary estimates (Schloss et al., 2011). For instance, chimeric sequences can affect the positioning of other, genuine sequences, ultimately producing an incorrect overall topology.
3. Contaminated sequences with ambiguous nucleotides: Ambiguity bases represent uncertainty in base-calling. The presence of such bases can introduce inaccuracies and distortions in phylogenetic analysis by influencing branch length estimates, the choice of evolutionary model, and topology (Lemmon et al., 2009). Determining the appropriate threshold for discarding ambiguity data depends on the dataset

being used.

4. Partial or sequences of different lengths: Depending on the analysis being conducted, it may be necessary to exclude sequences with partial or missing information. For instance, as mentioned in section 1.3.3, some taxonomic studies rely on individual or combined variable regions for classification, while others utilize whole 16S sequences. If the goal is to perform taxonomic classification using whole 16S sequences, including partial sequences that lack certain variable regions may not be appropriate.

As noted in section 1.3.3, sequence clustering is a common approach for studying evolutionary history through 16S sequences. By clustering sequences at a specific identity threshold allows to represent biologically relevant groups (Edgar, 2018c), such as species, allowing for the use of a single representative sequence from each group. This method reduces redundancy, removing sequences that would otherwise consume computational resources without contributing new biological insights, making phylogenetic analyses more efficient (Zou et al., 2020).

Some bioinformatics tools for sequence clustering implement greedy algorithms that process input sequences in user specified order, or in order according to abundance or length (Edgar, 2010; Rognes et al., 2016). Since the order in which sequences are processed can influence the clustering outcome, different processing orders can affect the clustering outcome (eg. cluster composition or selection of representative sequence). These differences, in turn, might affect the resulting phylogenetic tree, potentially inferring different topologies and branch lengths.

Software tools commonly used for 16S analysis that implement *de novo* clustering with greedy algorithms include VSEARCH (Rognes et al., 2016), USEARCH (Edgar, 2010) and CD-HIT (Li et al., 2012) (Table 2.1). Both VSEARCH and CD-HIT are open source software, while USEARCH became open-source starting from version 12. All three tools offer chimera filtering, dereplication, and clustering of sequences. However, a key difference is that USEARCH and CD-HIT also support amino acid sequence analysis, a feature not available not available in VSEARCH. The key difference between these tools is the sorting algorithm for the input sequences. CD-HIT sorts input sequences by length, where USEARCH sorts the sequences by abundance. In contrast to this, VSEARCH enables the user to pre-sort sequences by length, abundance or user specified order.

Feature	usearch	vsearch	CD-HIT
chimera filtering	✓	✓	✓
clustering	✓	✓	✓
dereplication	✓	✓	✓
nucleotide sequences	✓	✓	✓
protein sequences	✓	✗	✓
open source	✗	✓	✓

Table 2.1: Commonly used sequence quality control tools.

Multiple sequence alignment

The next step in building evolutionary trees is the alignment of the cleaned (and potentially clustered) sequences to produce a multiple sequence alignment (MSA) (Figure 2.1). Aligned sequences form a discrete character matrix, where each row and column are the representation of sequences and nucleotide derived from the same position in a common ancestral sequence (homologous nucleotide), respectively (Chatzou et al., 2015). Although, the input sequences are not required to be of the same length, the output alignment matrix must consist of equal length sequences. Each alignment matrix consists of matches, mismatches and deletions. A mismatch in the alignment represents a possible substitution of the nucleotide and an indel, represented as a gap, is a possible insertion or deletion (Higgins et al., 2005). A match is a position in which the compared sequences have identical nucleotide or amino acid.

CATT--ATATTCTAAA	(Sequence 1)
x	
CATTAGATA--CTTAA	(Sequence 2)
x	
CAATAGATA--CTTAA	(Sequence 3)

Figure 2.1: Illustrative representation of multiple sequence alignment. Sequences are represented as rows, and homologous nucleotides are represented as columns. MSA can consist of matches (black vertical lines), mismatches (red cross) and deletions (green horizontal lines).

One of the first and most fundamental algorithms for comparing nucleotide and protein sequences are the Needleman-Wunsch and Smith-Waterman algorithms (Chao et al., 2022). The Needleman-Wunsch algorithm performs a global alignment, meaning that it aligns all bases of both sequences from end to end (Needleman & Wunsch, 1970). This method is particularly useful when the sequences being compared are of similar length and share a significant degree of similarity. In contrast, the Smith-Waterman algorithm is designed for local alignment, focusing on identifying regions of high similarity within two sequences (Smith, Waterman, et al., 1981). Rather than aligning the entire sequence, it finds the most similar regions, making it more suitable for cases where the sequences may vary in length or contain only partial similarities. Both algorithms rely on dynamic programming, but require significant time and memory resources. The challenge grows even more when trying to calculate pairwise alignments for thousands of sequences.

Modern sequence aligners use heuristic methods to reduce the time and computational costs associated with dynamic programming. One such example is MAFFT (Multiple Alignment using Fast Fourier Transform), which has become widely used for aligning large number of sequences (Katoh & Standley, 2013). MAFFT achieves computational efficiency by first calculating a distance matrix, which is based on the number of shared 6-tuples (substrings of length six) between sequences, without the need to calculate pairwise alignments. Using the distance matrix, MAFFT constructs a guide tree that captures the relationships between the sequences. The sequences are then progressively aligned, beginning with the most closely related pairs, followed by the inclusion of more distantly related sequences. After the initial alignment, the guide tree is re-estimated

based on the current alignment, enabling further refinement of the relationships between sequences. Finally, MAFFT performs a realignment using the updated guide tree, which improves the overall alignment accuracy and optimises the alignment score. In addition to MAFFT, other progressive alignment tools that adopt heuristic approaches include Clustal (Sievers & Higgins, 2018) and MUSCLE (Edgar, 2004).

Progressive alignment methods are widely used due to their advantages in speed and ability to produce optimal alignments. However, progressive alignments can sometimes introduce artefactual gaps, particularly in large datasets, as it may be challenging to achieve fully accurate alignment across all sequences (Golubchik et al., 2007). An excessive number of gaps can lead to unreliable and biologically meaningless alignments. Reference-based approaches can help avoid inappropriate gap insertion. Tools like Nextalign are reference-based aligners that calculate pairwise distances by comparing input sequences against a reference sequence, with only sites present in the reference included in the alignment (Hadfield et al., 2017). However, reference-based alignments may be less suitable for distantly related sequences, as they depend on the availability of a highly similar reference sequence for accurate comparisons.

Many available alignment algorithms align nucleotide sequences in a codon-unaware manner (Hadfield et al., 2017). This can pose challenges when using nucleotide sequences for coding regions because the function of a protein is primarily determined by its amino acid sequence, rather than by the underlying nucleotide sequence. Since interactions among different amino acids determine the structure and folding into complex 3D structures that carry out specific tasks, inappropriate placement of gaps and disrupting codons can shift the reading frame, leading to the loss of functional information

and sequence homology (Hall, 2005). In a MSA, each column represents a positional relationship among the sequences, but the algorithm behind the alignment is blind to biological context and only aims to maximise or minimise a scoring function. By enforcing codon awareness, biological relevance is imposed on the alignment process, ensuring that disruptions, such as frameshifts, are minimised, thereby preserving functional and evolutionary insights. Without this biological consideration, downstream analyses can be compromised, leading to incorrect bases being compared or affecting the choice of evolutionary models, which in turn may result in misleading phylogenies. Aligning protein sequences allows for construction of codon-aware alignments, preserving functional information and avoiding biases from codon-unaware alignments. In scenarios where nucleotide alignment is necessary for estimating phylogeny or other downstream analyses, the common practice is to convert protein alignments back into a DNA alignment by threading the nucleotide sequences onto a protein alignment.

In some cases, phylogenetic reconstruction involves the use of multiple genes. The common practice is to align the individual genes separately, back-translate and trim (see below) the alignment appropriately, and then concatenate the alignments (Kapli et al., 2020). The reason for aligning the sequences separately is to maximise or minimise the alignment score for each gene. Additionally, by calculating separate alignments, all information is preserved, ensuring accurate representation of each gene's evolutionary history and maintaining the sequence homology (Horton & Taylor, 2023; Long et al., 2014).

Alignment trimming

Filtering and trimming of a constructed MSA is required to remove unreliable or uninformative columns before phylogenetic tree estimation. Positions mostly represented by gaps within a MSA can introduce uncertainty into the reconstruction of phylogenetic trees, as they can accommodate various explanatory topologies. The phylogenetic tree serves as a model that explains the sequence data, and any factors that complicate this explanation—such as misalignment, chimeric sequences, gaps, and ambiguous bases—can reduce the confidence in the resulting phylogeny. Therefore, removing positions that are largely represented by gaps, along with poor-quality sequences (e.g., chimeric sequences and those with excessive ambiguity bases), can increase the accuracy and precision of phylogenetic reconstruction by increasing the evolutionary signal (Castresana, 2000; Talavera & Castresana, 2007).

Two commonly used bioinformatics tools that have been developed for alignment filtering and trimming are Gblocks (Talavera & Castresana, 2007) and trimAl (Capella-Gutiérrez et al., 2009). Both tools are conceptually similar and were designed to remove columns according to a user specified threshold for both gap score and similarity score component. However, trimAl differs from Gblocks in that it offers automated selection of those thresholds by a number of built-in heuristic approaches.

Evolutionary model estimations

During phylogenetic tree reconstruction, assumptions about how sequences evolve over time are made. It cannot be assumed that each nucleotide or amino acid have equal chances of change (Hall, 2018). Some regions may be more conserved due to functional

constraints, while others may vary more freely. Evolutionary models inform us about the likelihood of nucleotide replacement within a sequence, allowing for more accurate representations of evolutionary relationships.

In phylogenetic trees, branch lengths indicate the amount of genetic change between sequences in an alignment (Gregory, 2008). While it may seem intuitive to measure branch lengths by calculating the percentage of observed changes between an ancestor and its descendants, this method assumes that there are no additional unobserved changes. However, there could be changes that occurred in the past and are no longer apparent. Two scenarios might be overlooked with the calculation of observed changes:

1. Multiple changes occur at a site but only one is observed, for example, from G to A and then from A to C.
2. No changes are observed as the site mutated from G to A and then back to G.

Evolutionary models operate under the assumption that there may be more changes at a site than the ones observed (Hall, 2018). Various evolutionary models exist, and one of the biggest challenges we face is determining which model is most appropriate for a given dataset. However, several bioinformatics tools for estimating evolutionary models have been developed, such as ModelTest-NG (Darriba et al., 2019), DT-ModSel (Minin et al., 2003), and JModelTest (Darriba et al., 2012), which enable more robust and accurate estimation of the models.

The choice of evolutionary models is crucial, as it has been found to affect inferred phylogenies (Kuhner & Felsenstein, 1994; Lemmon & Moriarty, 2004). It can be influenced by factors such as the inclusion of contaminants like poor-quality sequences

and chimeras, missing information in the alignment (e.g., over representation of gaps), or the inclusion of unrelated sequences (Lemmon et al., 2009). This further highlights the importance of cleaning and filtration steps in the phylogenetic analysis process.

Estimation of phylogenetic trees

There are four main approaches for phylogenetic tree estimation: distance, parsimony, bayesian and likelihood (Hall, 2018). The most widely used distance method is neighbour joining (NJ) (Saitou & Nei, 1987). The principle of this method is to find the distance between each pair of sequences and group the most similar sequences together. The NJ method is particularly useful on large datasets due to its computational speed (Elias & Lagergren, 2009). However, NJ fails to account for the complexities of evolutionary dynamics—such as varying mutation rates—potentially leading to less accurate representations of true evolutionary relationships among sequences (Hall, 2018) as more sophisticated methods, like Maximum Likelihood or Bayesian inference, do. Given the advancements in computational efficiency, there is little justification for continuing to use NJ; ML methods have become sufficiently fast and robust to serve as the default choice in phylogenetic analysis.

Maximum parsimony (MP) seeks to find phylogenetic tree by minimising the total number of estimated evolutionary steps required to explain the sequence data. In MP, the assumption is that the simplest (i.e. most parsimonious) explanation that can explain the data is more preferred over more complex explanations (Fitch, 1971). The biggest benefit of using MP is the speed required to process large amount of data especially for sequences with low divergence (Takahashi & Nei, 2000). However, this method is associated with a high rate of inconsistencies, especially for data with a

constant rate of nucleotide substitution (Takezaki & Nei, 1994).

Perhaps the most commonly used modern method applied to estimate evolutionary events is Maximum likelihood (ML), which was introduced by Felsenstein in 1981 (Felsenstein, 1981). The ML method assumes that the evolution of each nucleotide site occurs independently according to a substitution model. The topology of the resulting phylogenetic network is obtained by finding the tree which maximises the probability of observing the data, for a given evolutionary model (Harrison & Langdale, 2006). This is achieved by estimating a number of trees, starting from a number of different initial topologies and modifying the trees until they converge on a topology that is most consistent with the input data. Although ML was considered to be slow and computationally expensive, this method is superior to NJ and MP and the power of modern personal computers makes it a good default choice under many circumstances.

One of the most commonly used tools for estimation of ML phylogenetic trees is RAxML-NG (Kozlov et al., 2019). RAxML-NG was developed by incorporating concepts of other widely used bioinformatic tools for the estimation of ML evolutionary trees like RAxML and ExaML. The efficiency of RAxML-NG was demonstrated to be increased by 10-60% and offers improved accuracy. RAxML-NG also predicts the most suitable evolutionary model for the given dataset and enables post-hoc analysis of phylogenetic trees including bootstrapping.

Another widely used method for estimating phylogenies is Bayesian Inference (BI) that uses Bayes' Theorem to compute the posterior probability of each possible tree (Didelot et al., 2018). Unlike ML that seeks a single most likely tree, BI it uses prior assumptions about evolutionary processes, and updates these assumptions by analysing

the data. In practice, the process begins with an initial tree, for which the probability is calculated (Hall, 2018). This tree is then subject to slight modifications, which may involve adjusting its topology or altering branch lengths. After these modifications, the likelihood score of the revised tree is recalculated. If the likelihood of the new configuration surpasses that of the previous tree, the revised tree is accepted as the new current state.

BI differs from the ML approach in its ability to consider the same tree multiple times during the search process, which can prove advantageous (Hall, 2018). To illustrate this, we consider a metaphorical landscape in which we seek the best tree, represented by the highest point on a hill. In the ML search, the algorithm may become trapped on a hill that, while tall, is not the highest peak available; therefore, it may overlook a much higher hill, representing a better tree. In contrast, BI's approach allows it to explore various options more freely, potentially leading to the identification of superior phylogenetic trees that might otherwise remain undiscovered. This capacity to revisit and refine trees enables BI to escape local maxima, consequently enhancing the likelihood of converging on a more accurate phylogenetic estimation.

Bootstrapping

Central to our confidence in any phylogenetic reconstruction is statistical analysis to evaluate the robustness of the estimated tree topology. The reliability of phylogenetic tree topology is often estimated by bootstrapping (Hall, 2018), a concept introduced to phylogenetics by Felsenstein in 1985 (Felsenstein, 1985). The basic idea behind (Felsenstein) bootstrapping is to measure sensitivity of the tree topology to resampling of columns from the initial alignment. This form of bootstrapping involves the generation of

many (typically 100-1000) alignments of the same size as the original input by resampling (with replacement) random columns from the original alignment, and inferring a new phylogeny. These phylogenetic trees are then subdivided into subsets at specific nodes (bipartitions) and compared to the original tree. If a bipartition in the new tree matches that of the original, it is considered a match and typically assigned a score of 1. The resulting bootstrap values - the sum of scores on each bipartition in the original tree - indicate how frequently a particular bipartition was observed out of the total number of replications. Higher bootstrap values suggest greater support for the corresponding bipartition in the phylogenetic tree, indicating the sensitivity of that bipartition to changes in the input dataset (Hall, 2018).

Felsenstein's bootstrap is arguably one of the most widely used methods for assessing the robustness of phylogenetic inferences. However, in the era of big data, its ability to accurately reflect statistical support for evolutionary relationships has been questioned (Lemoine et al., 2018). Studies on microbial diversity, especially with large datasets, have shown that Felsenstein's bootstrap tends to produce lower support values, particularly for deeper branches (Soltis & Soltis, 2003). This occurs because, as the number of sequences increases, so does the number of possible tree topologies (tree space). With more sequences, the bootstrapping process samples this tree space more broadly, leading to greater variation and potentially different topologies. Since Felsenstein's method only considers exact matches between tree bipartitions, this can result in less precise statistical support. An alternative approach, Transfer Bootstrap Expectation (TBE), offers a better estimation of phylogenetic support (Lemoine et al., 2018). Unlike Felsenstein's bootstrap, TBE takes into account not only exact matches but also similar topologies,

measured by transfer distance. This adjustment improves the accuracy of support values for evolutionary inferences, making it a more robust tool for large-scale phylogenetic analyses.

2.1.4 Best practices

The methodology of the analysis should be documented in such way that each step of the analysis can be replicated, and the results can be reproduced. It has been estimated that approximately only 40% of published phylogenetic trees are reproducible (Magee et al., 2014). Failure to reproduce results in analysis can prevent other scientists from verifying the findings of the study, identifying errors or biases, and may raise concerns regarding result accuracy and reliability.

Numerous recommendations are available to ensure the reproducibility of bioinformatics analyses, applicable not only to phylogenetic reconstruction but to all analyses (Sandve et al., 2013). These recommendations include clear reporting of:

1. Software used for specific analyses and tasks. When solving a problem, there might be multiple options to explore, each potentially yielding different results due to variations in algorithms.
2. Software versions. Algorithms to solve specific tasks might evolve in response to the available data, and the progress of understanding of the problem at hand. This evolution may lead to different algorithms being implemented across software versions or alterations to algorithms due to bug fixes within the code base.
3. Parameters, which can significantly affect data handling and processing. In phylogenetic reconstruction, it has been demonstrated that parameter selection

can impact final outcomes, particularly concerning evolutionary models or seed parameters (Shen et al., 2020).

4. Databases and sequence accessions. Many databases, including NCBI, are dynamic and their content may vary over time. The difference in the available data can lead to discrepancies in results and observations.
5. Computing resources, including processor types and CPU cores. Previous investigations have indicated that computational resources may influence the resulting phylogenies observed (Shen et al., 2020).
6. Sequence alignments and tree files. Providing these as a minimum standard is crucial for ensuring transparency and reproducibility in phylogenetic analyses. These files allow others to evaluate the robustness of the results, validate the methods used, and contribute to the principles of open science.

2.1.5 Aims and objectives

The overall aim of this chapter is to estimate *Streptomyces* diversity and assess the agreement between 16S and whole-genome distance methods, using all publicly available 46,981 full-length 16S rRNA sequences and the 2,276 *Streptomyces* genomic sequences available at the time of writing.

Specific objectives of this chapter are:

1. Given the significant increase in available genomic sequence data (Figure 1.6) and the development of advanced phylogenetic methods since 2012, when the most extensive 16S phylogeny for *Streptomyces* was constructed (Labeda et al., 2012),

I will reconstruct a more complete and updated phylogeny for this genus using sequences from SILVA, Greengenes, RDP and NCBI.

2. Approximately 17% of annotations in SILVA and Greengenes, and 10% in the RDP database, have been found to be incorrect (Edgar, 2018a). Moreover, the lack of active maintenance for Greengenes and RDP, along with inconsistencies in quality control procedures across these databases, highlights the need for validation (section 2.1.2). Therefore, I will assess the quality of genomic *Streptomyces* sequences and validate the taxonomic nomenclature assigned in the 16S databases used.
3. Since the threshold for 16S sequence clustering is currently under debate (section 1.3.3), I will investigate the accuracy and impact of sequence clustering across a spectrum of threshold identities ranging from 98% to 100%. (zero-radius Operational Taxonomic Units).
4. I will explore intragenomic heterogeneity of 16S rRNA sequences across *Streptomyces* genomes and investigate the underlying biology and whether it supports mapping between 16S and whole-genome taxonomy, using a combination of graph theory and distance methods.

The findings of this study will benefit taxonomists and researchers interested in studies that heavily rely on accurate taxonomy and, in particular, 16S-based classification methods. By conducting large-scale comparisons of 16S and whole-genome distance methods in *Streptomyces*, this work will provide greater clarity on the effectiveness of single-gene taxonomy classification, and will lead to better understanding of the

taxonomic structure of this pharmaceutically important genus.

2.2 Methodology

2.2.1 Data summary and availability

All analyses in this and other chapters were carried out on a MacBook Pro with 2 GHz Quad-Core Intel Core i5 and 32GB RAM, Python v3.9, RStudio v2023.06.1+524 with R v4.3.2 and default parameters, unless stated otherwise. All code, raw and supporting data used in this chapter are publicly available from GitHub (https://github.com/sipbs-compbiol/Kiepas_et_al_2024_16S) and Zenodo (<https://zenodo.org/records/10991761>). The 16S sequence data used in this study are also available from Greengenes v13.5 (https://gg-sg-web.s3-us-west-2.amazonaws.com/downloads/greengenes_database/gg_13_5/gg_13_5.fasta.gz), SILVA v138.1 (https://www.arb-silva.de/fileadmin/silva_databases/release_138_1/Exports/SILVA_138.1_SSURef_tax_silva.fasta.gz), RDP v11.5 (http://rdp.cme.msu.edu/download/current_Bacteria_unaligned.fa.gz), and NCBI under BioProject PRJNA33175. Additionally, the accessions for all 16S sequences used in this manuscript, as well as for *Streptomyces* genome accessions, are available in the Supplementary File 2, as content of the source databases may vary over time.

2.2.2 Acquisition of 16S rRNA *Streptomyces* sequences from major 16S rRNA databases

The flowchart provided in Figure 2.2 provides an overview of analysis steps and serves as a guide for which Supplementary Files were generated during reconstruction of the 16S phylogeny. 16S rRNA sequences were manually downloaded from SILVA

v138.1 (Yilmaz et al., 2014), RDP v11.5 (Cole et al., 2014), NCBI (New ribosomal RNA BLAST) PRJNA33175 (Sayers et al., 2021) and Greengenes v13.5 (DeSantis et al., 2006) as indicated in Table 2.2. Sequence accessions for the downloaded 16S databases are provided in **Supplementary file 2**.

Greengenes is the only one of these databases that stores sequence and taxonomy information in separate files. The taxonomy file (`gg_13.5_taxonomy.txt`, **Supplementary File 2**) was accessed on the 26th of October 2020 from https://greengenes.secondgenome.com/?prefix=downloads/greengenes_database/gg-13.5/. This taxonomic information was mapped to sequence by matching the sequence identifier (`gg_map_taxonomy.py`; **Supplementary File 2**).

Table 2.2: Summary description of 16S rRNA databases used in the study.

Database	Version	Acquisition date	No. of sequences
SILVA	138.1	26th Oct 2020	2 224 740
GreenGenes	13.5	26th Oct 2020	1 262 982
RDP	11.5	10th Nov 2020	3 196 041
NCBI	PRJNA33317	10th Nov 2020	20 801
			TOTAL: 6 704 552

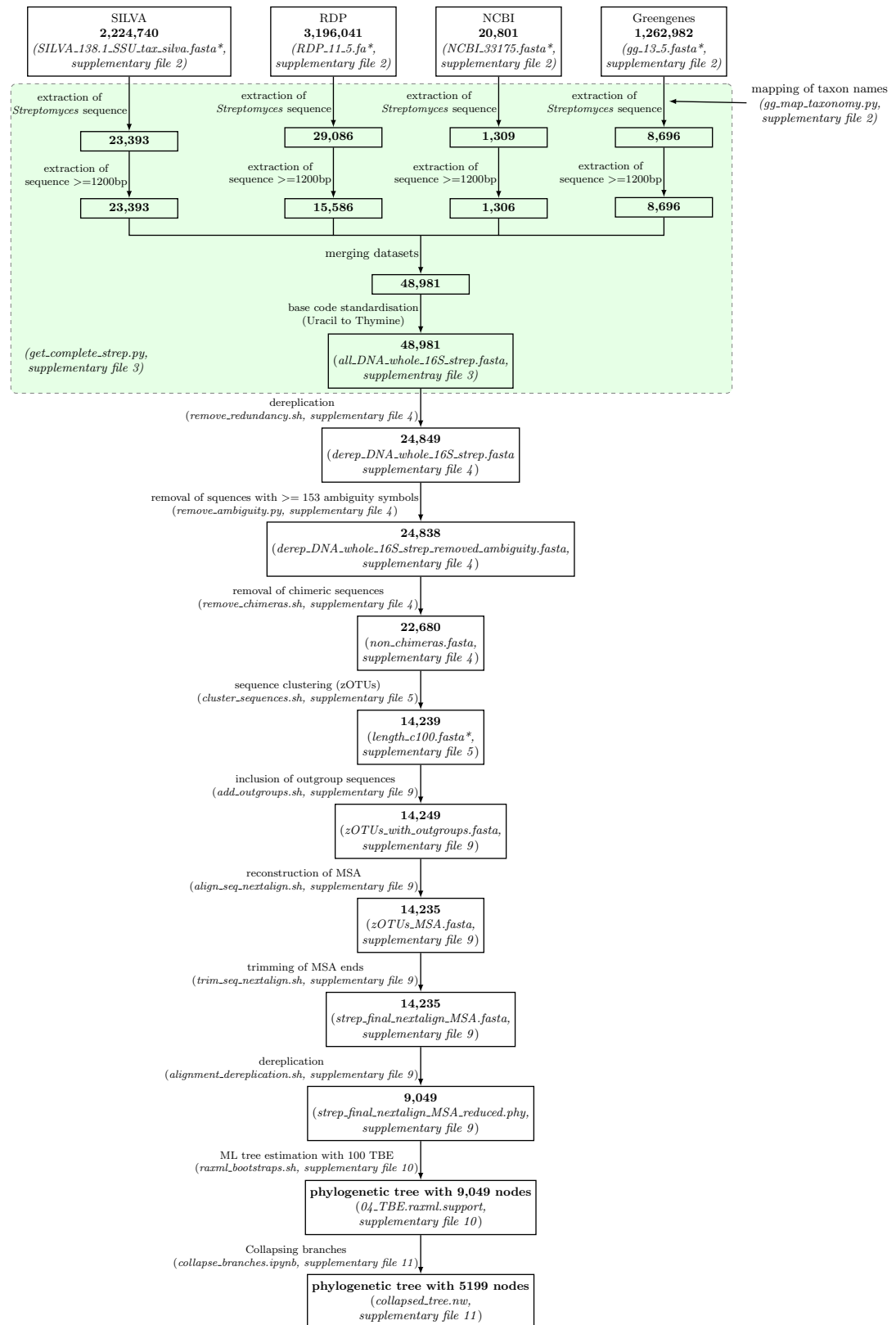


Figure 2.2: Schematic workflow for construction of the full-length 16S rRNA *Streptomyces* phylogeny. Each arrow represents a process and is annotated with the name of the script used and corresponding supplementary file. Output/data files, and the number of remaining sequences after each step, are indicated by rectangles. The green shading represents a single processing step of collecting and collating 16S database sequences.

2.2.3 Selection of full-length *Streptomyces* 16S rRNA sequences

As the downloaded databases consist of 16S rRNA sequences covering many genera, I filtered the downloaded databases to retain only sequences with the keyword *Streptomyces* in the taxonomy field. To obtain only full-length 16S sequence candidates, I further filtered the sequences by removing sequences shorter than 1200bp. The average length of a complete 16S gene is approximately 1550bp, and a 1200bp lower threshold on sequence length was chosen to capture all hypervariable regions, maximising information about sequence variation, and to filter out database sequences targeting only a subset of hypervariable regions. I refer to this >1200bp filtered set as “full-length” sequences.

The resulting 48,981 sequences were combined to create a local 16S database, and base coding was standardised to replace uracil with thymine ($U \rightarrow T$; `get_complete_strep_seq.py`, Supplementary File 3).

2.2.4 LPSN Nomenclature Validation

Standardised nomenclature data was downloaded from the List of Prokaryotic Names with Standing in Nomenclature (Parte, 2018) (LPSN; `LPSN_taxonomy.csv`, Supplementary File 6) on 16th February 2023. Species-level nomenclature previously assigned to the 48,981 full-length 16S rRNA *Streptomyces* sequences in the source database(s) was validated against this list (`get_NCBI_taxID_and_LPSN_status.py`; Supplementary File 6).

2.2.5 Elimination of nomenclature disagreements at higher taxonomic ranks

In some cases, the stated taxonomy at higher ranks of a sequence was a lineage not recognised to be ancestral to *Streptomyces* spp. The nomenclature at ranks up to Kingdom assigned to extracted full-length 16S rRNA *Streptomyces* sequences in the source database(s) was validated (`check_nomenclature_hierarchy.py`, Supplementary File 3) to identify and note, but not correct, this and similar cases of nomenclature hierarchy disagreement.

2.2.6 Removing redundant and ambiguous sequences

As the 16S rRNA databases used in this study might inherit sequences from similar major international sequence databases and may contain identical sequences without bringing the new biological information, I identified and removed 24,132 strictly identical redundant sequences identified using vsearch v2.15.1 (Rognes et al., 2016) (`--derep_fulllength, --sizeout; remove_redundancy.sh`, Supplementary File 4).

To avoid the accuracy of the analysis being affected by poor quality sequences with a high ambiguity bases count, the number of ambiguity bases present in each sequence were counted. Sequences with no ambiguity bases were treated as high-quality by default.

To identify an appropriate statistical threshold to identify outliers, the distribution of ambiguity bases present in each sequence was investigated (`remove_ambiguity` in Supplementary File 4). As shown in Figure 2.3, the number of ambiguity count per

sequence follows skewed distribution, and the variance (498.3) is larger than the mean (5.5), also known as overdispersion. Thus, a negative binomial statistical model was chosen to identify a threshold at which outlying sequences containing an unusually large number of ambiguity bases could be removed.

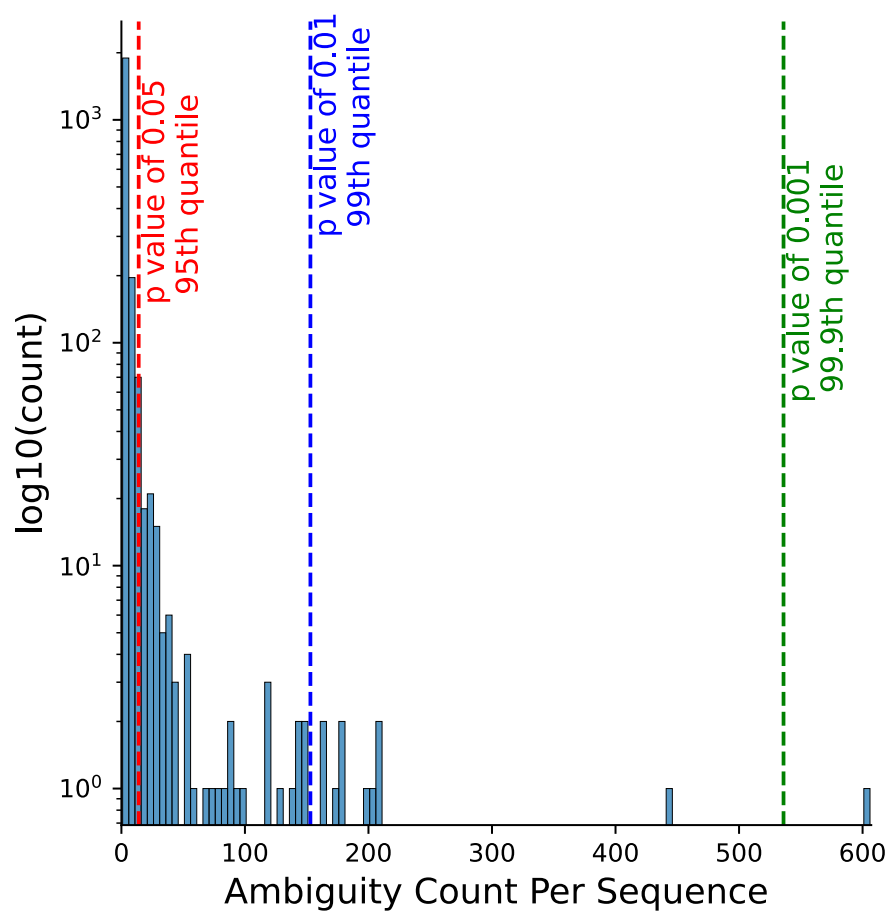


Figure 2.3: The number of ambiguity bases per sequence follows a negative binomial distribution, with observed variance (498.3) much larger than the mean (5.5). Outliers identified at the 95th quantile (p-value of 0.05), 99th quantile (p-value of 0.01), and 99.9th quantile (p-value of 0.001) based on this negative binomial distribution are highlighted in red, blue, and green, respectively.

A threshold value was calculated for Negative Binomial Distribution in R using `qnbinom()`, where μ was the mean count of ambiguity per sequences, and n was the run of positive outcomes treated as dispersion parameter, calculated with the following formulae:

$$\mu = \frac{\text{total count of ambiguity bases}}{\text{total number of sequences}}$$

$$n = \frac{\mu p}{1 - p}$$

where p is the probability calculated as follow:

$$p = \frac{\text{total count of ambiguity bases}}{\text{total count of all bases in all sequences}}$$

The `qnbinom()` function also requires an additional argument specifying the quantile for which we want to determine the threshold. In statistical analysis, a p-value quantifies how extreme an observed result is relative to a distribution of values. For instance, a p-value of 0.01 corresponds to the 99th quantile, indicating that any sequence with ambiguity counts exceeding this threshold falls within the top 1% of values. Such sequences are considered outliers, suggesting they are significantly different from the majority of the data.

The analysis showed that sequences with more than 14, more than 153, or more than 536 ambiguous bases could be outliers at the p-values of 0.05, 0.01 and 0.001, respectively. The minimum length of 16S rRNA sequences in this database is 1200bp.

Thus, retaining sequences with as many as 536 ambiguity bases would mean that these sequences lack almost half of the expected biological information compared to sequences with no ambiguity bases. This could negatively affect phylogenetic analysis and lead to inaccurate tree topology estimation. Therefore, excluding sequences at $P < 0.001$ was not considered.

Retaining sequences with as many as 153 ambiguity bases is comparable to losing one conserved region from the 16S sequence. Taking into consideration that this study is based on whole 16S rRNA sequences, I considered there would still be enough biological information to distinguish between different species if 153 bases were removed. However, only 100 sequences contained more than 14 and less than 153 ambiguity bases, and I considered that discarding this set of sequences would not substantially reduce the number of sequences in the analysis. Hence, all sequences with more than 153 ambiguity bases were discarded (at $P < 0.01$), which reduced the number of total 16S rRNA sequences to 24,838.

All databases used in this *in silico* study perform their own sequence quality checks. However, not all actively check for and remove chimeric sequences. Therefore, to reduce the negative effect chimeras could have on the phylogenetic analysis of 16S rRNA *Streptomyces* sequences, 2,158 chimeric sequences were identified and discarded using vsearch v2.15.1 (Rognes et al., 2016) (`--uchime_denovo; remove_chimeras.sh`, Supplementary File 4). These operations reduced the $\approx 49,000$ full-length 16S rRNA *Streptomyces* database to a total of 22,680 sequences (`non_chimeras.fasta`; Supplementary File 4).

2.2.7 Clustering of complete 16S rRNA *Streptomyces* sequences

The 22,680 retained full-length *Streptomyces* sequences were clustered using vsearch 2.15.1 (Rognes et al., 2016) at pairwise percentage sequence identity thresholds ranging from 98% to 100% in steps of 0.1% (`--cluster_fasta, --centroids;cluster_sequences.sh`, Supplementary File 5). The number of distinct taxonomic assignments in a cluster for each clustering threshold was determined using the NCBI reference taxonomy (Schoch et al., 2020) (downloaded on the 31st January 2023 from <https://ftp.ncbi.nih.gov/pub/taxonomy/taxdmp.zip>; `names.dmp`, Supplementary File 6). Each sequence was assigned NCBI taxID corresponding to the LPSN-validated nomenclature assigned to it in the source database(s). This process generated 14,239 zOTUs with 100% pairwise sequence identity according to the clustering threshold, although the input sequence set was non-redundant. Nomenclature and corresponding taxIDs for redundant sequences removed in methodology section 2.2.6 were assigned to the retained representative sequence for this analysis, as multiple different taxonomic classifications were identified for many redundant sequences (`cluster_composition_analysis.py`, Supplementary File 6).

2.2.8 Phylogenetic reconsuction

Representative sequences from all 14,239 16S rRNA zOTUs, and a further ten 16S rRNA outgroup sequences belonging to isolates from *Kitasatospora*, *Streptoalloteichus* and *Clavibacter* genera (`outgroup.fasta`, Supplementary File 9) were aligned using nextalign v0.1.4 (Hadfield et al., 2017) against the GCF_008931305.1 16S rRNA reference sequence (`S_coelicolor_A32.fasta`, Supplementary File 9) from *Streptomyces*

coelicolor A3(2).

Alignments were trimmed using trimAl v1.4 (Capella-Gutiérrez et al., 2009) (trim_alignments.sh, Supplementary File 9) and subsequently dereplicated using RaxML (alignment_dereplication.sh, Supplementary File 9). A Maximum-Likelihood tree was estimated using RaxML-NG v1.0.2 with 100 Transfer Bootstrap Expectation (TBE) replicates (--model GTR+F0; --seed 24875;raxml_bootstraps.sh; raxml_tbe.sh, Supplementary File 10) on the ARCHIE-West computing cluster with Intel XI(R) Silver 4216 CPU 2.10Hz, 32 cores and 187 GB RAM.

2.2.9 Assessment of unique 16S rRNA sequences from *Streptomyces* genomes

All 2,276 publicly available *Streptomyces* genome sequences (streptomyces_genomes.txt, Supplementary File 17) were downloaded from NCBI (Sayers et al., 2021) on July 8th, 2021 with ncbi-genome-download v0.3.3 (<https://github.com/kblin/ncbi-genome-download>; download_genomes.sh, Supplementary File 17). The flowchart in Figure 2.4 outlines the workflow processes and supplementary materials used for analysis of 16S rRNA sequences from *Streptomyces* genomes.

The assembly status of each genome was checked against the NCBI assembly report downloaded on 30th January 2023 from https://ftp.ncbi.nlm.nih.gov/genomes/refseq/assembly_summary_refseq_historical.txt (assembly_summary_refseq_historical.txt, check_genome_status.py, Supplementary File 17). I discarded 120 suppressed genomes from the analysis. Updated versions of five replaced genomes (streptomyces_replaced_genomes.txt, Supplementary File 17) were manually

downloaded from NCBI (Sayers et al., 2021) on 30th January 2023. A total of 6,692 16S rRNA sequences were extracted from the 2,156 *Streptomyces* genomes by matching the key word '16S' in the `gene_qualifiers` GenBank field (`extract_16S.py`, Supplementary File 18). I filtered the extracted sequences to a total of 4,227 by retaining only those from the 1,369 genomes (`retained_genomes.txt`, Supplementary File 18) that exclusively contain only full-length and ambiguity bases-free 16S rRNA sequences (`filter_16S_seq.py`, `filtered_16S_seq_from_strep_genomes.fasta`, Supplementary File 18). All 4,227 such sequences extracted from *Streptomyces* genomes were aligned using nextalign v0.1.4 (Hadfield et al., 2017) against the same GCF_008931305.1 16S rRNA reference sequence as used previously (`S_coelicolor_A32.fasta`, Supplementary File 9). The alignment was trimmed using trimAl v1.4 (Capella-Gutiérrez et al., 2009) (`trim_alignment.sh`, Supplementary File 19), and genomes sharing identical 16S rRNA sequences were clustered (`get_input_genomes_for_pyani.py`, Supplementary File 19) and used as input sequences to determine their taxonomic boundaries with ANI using pyANI v0.3 (Pritchard et al., 2015) (`pyani_analysis.sh`, Supplementary File 19). As mentioned in section 1.3.3, a proposed ANI threshold suggests that isolates with $\geq 95\%$ genome identity likely belong to the same species. However, genus boundaries remain debated, with some arguing that genomes sharing $< 50\%$ of their genetic material may be more similar to unrelated lineages, warranting separate classification (Pritchard et al., 2015). Based on these considerations, I adopted the 95% identity and 50% coverage thresholds to estimate species and genus boundaries, respectively.

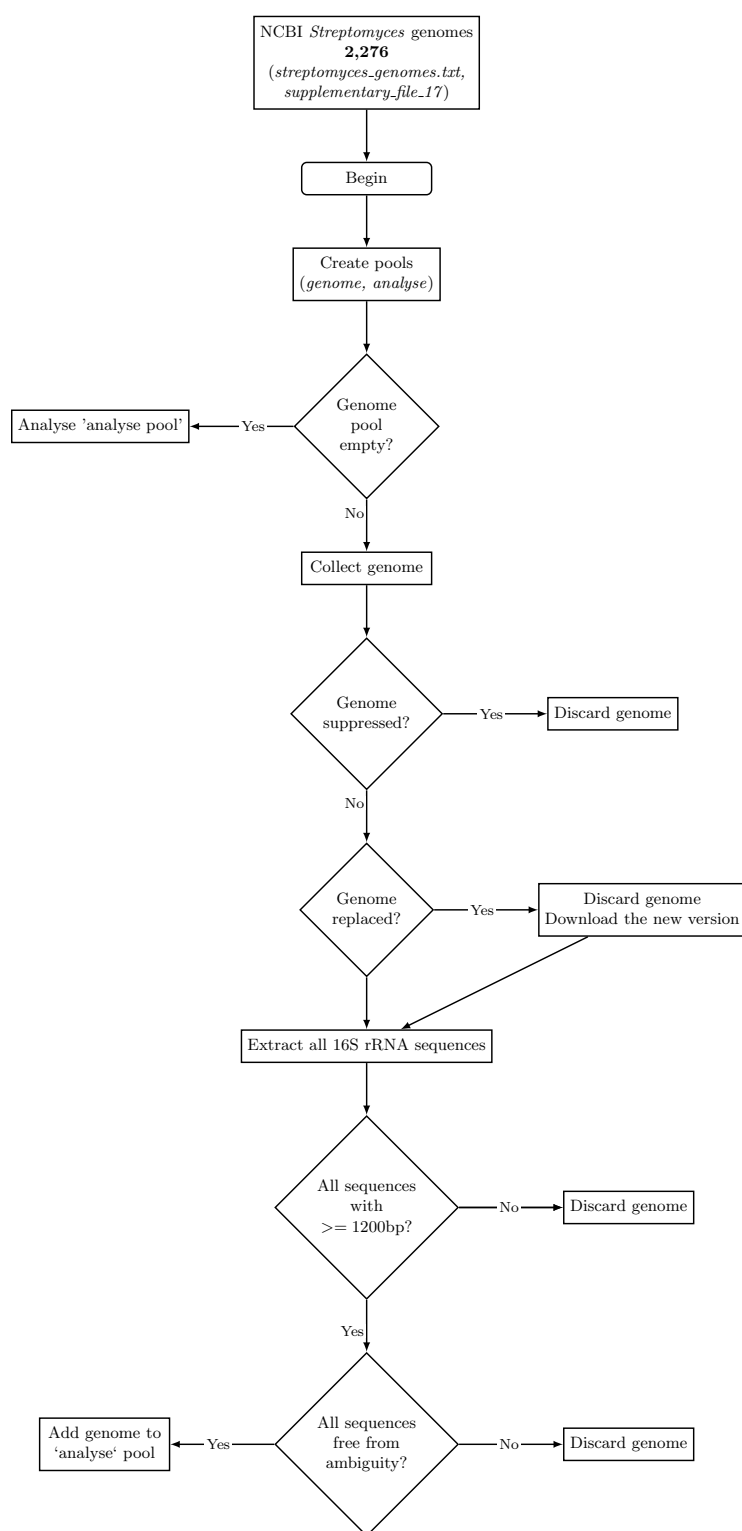


Figure 2.4: Schematic representation of the pipeline used to filter publicly available *Streptomyces* genomes.

2.2.10 Network analysis of genomes based on shared 16S rRNA sequences

I constructed a network representing individual genomes as nodes and assigning edges between genomes with weights corresponding to the number of shared identical 16S sequences, to represent the 1,369 *Streptomyces* genomes that contain only full-length and ambiguity bases-free 16S rRNA sequences. Edges corresponding to pairs of genomes with no 16S sequence in common were removed so that an edge in the graph implied at least one identical 16S sequence in common. The graph was constructed and processed using NetworkX (Hagberg et al., 2008), and visualised interactively with plotly v5.6.0 (<https://plotly.com/python/>; `genome_16S_NetworkX.ipynb`; Supplementary File 20). Node layout was calculated using Cytoscape v3.9.0 (Shannon et al., 2003) with Prefuse Forced Directed layout. ANIm analysis was performed using pyANI v0.3 (Pritchard et al., 2015) to determine taxonomic boundaries for all genomes found in the same subgraph, as above (`pyani_analysis.sh`, Supplementary File 20).

2.3 Results and Discussion

2.3.1 Public 16S *Streptomyces* sequence databases include records with low-quality or redundant sequence, or that have issues with taxonomic nomenclature

High-quality sequences are crucial for ensuring the accuracy and reliability of research and analysis, particularly in the fields of genomics, taxonomic classification, and applied microbiology. Public sequence databases provide remarkable opportunities for large

and comprehensive studies of evolutionary relationships and hold exceptional value that has shaped modern systematics. Using unintentionally inaccurate database records can potentially lead to false interpretations and flawed research outcomes.

To examine the robustness of phylogenetic relationships of the genus of *Streptomyces* based on 16S rRNA diversity, all available 16S rRNA sequences were downloaded from SILVA, GreenGenes, RDP and NCBI databases (Methodology Section 2.2.2). Knowing that these databases acquire sequences from similar authoritative database(s) it was possible that these contained redundant sequences. Inclusion of redundant sequences in phylogenetic analyses can waste computational resources and slow down analyses without providing additional biologically meaningful information and make it difficult to distinguish between noise and phylogenetic signal. Therefore, it was necessary to identify and exclude redundant sequences for phylogenetic reconstruction. Among these databases, there were 62,482 16S rRNA sequences belonging to the genus *Streptomyces*, of which only 48,981 (78.4%) were determined to be full-length sequences (methodology section 2.2.2). In total, I identified 24,849 non-redundant full-length 16S rRNA *Streptomyces* sequences (50.7% of all full-length *Streptomyces* sequences; 39.8% of all database *Streptomyces* sequences).

Prokaryotic taxonomic nomenclature is a pivotal mechanism for unambiguous communication about an organism's identity. Correct nomenclature helps avoid undesirable clinical, ecological, agricultural, and pharmaceutical consequences (Boykin, 2014; Janda, 2020). In medical diagnostics, misclassification of a bacterium under an incorrect genus or species name can result in the prescription of inappropriate antibiotics, or failure to give suitable treatment. For example, the recent reclassification of *Ochrobactrum spp.* as

Brucella spp. illustrates this issue (Hördt et al., 2020). *Ochrobactrum*, an opportunistic pathogen with low virulence, typically causes less severe infections that require minimal treatment. In contrast, *Brucella* is highly pathogenic and requires at least eight weeks of antibiotic therapy for both infected patients and those who have been in close contact (Moreno et al., 2023). This renaming has been the cause of misidentification of these pathogens preventing patients from receiving appropriate medical care (Park et al., 2024). In pangenomic and comparative genomics, assuming that taxonomic assignments are accurate can lead to significant issues. For instance, if two genomes are both labeled as *Streptomyces*, but one genome has been incorrectly classified and actually belongs to a closely related but distinct genus, this mislabeling can obscure the pangenomic analysis (as discussed in 1.2.4). Specifically, if the misclassified genome is a member of a different genus that is less closely related to *Streptomyces* than originally thought, the analysis may inaccurately identify a smaller number of core genes shared between the two genomes and a larger number of accessory genes. This misinterpretation can result in an overestimation of the pangenome’s openness and obscure effective drug discovery (general introduction section 1.2.4).

The databases considered in this work rely on a variety of taxonomic authorities: RDP uses Bergey’s Manual (Bergey, 2001; Kämpfer, 2020), SILVA uses LPSN and Bergey’s Manual (Parte, 2018); and NCBI combines nomenclature provided by the submitter with that in Greengenes, basing its nomenclature on that in the NCBI taxonomy (DeSantis et al., 2006; Schoch, n.d.). All of these schemes are good-faith efforts to follow the International Code of Nomenclature of Prokaryotes (ICNP) (Parker et al., 2015), but I found that records are in some cases inaccurate. LPSN is an

online database that catalogues the validly published names of prokaryotes in accordance with the Rules of ICNP (Parte, 2018). I validated taxonomic nomenclature assigned to each of the full-length sequences in the source database(s) against LPSN (Methodology Section 2.2.4) and find that only 14,859 (30.3%) of the 48,981 species names assigned to extracted full-length *Streptomyces* sequences were found within LPSN (eg. EU570732.1.1437 assigned *Streptomyces clavuligerus* species name). A further 1,400 (2.9%) were synonyms of valid names (eg. FJ486381.1.1443 assigned *Streptomyces variabilis* name, synonym of *Streptomyces griseoincarnatus*), but 28,333 (57.8%) sequences were labelled as unclassified *Streptomyces* (eg. KY921882.1.1268), 17 (0.03%) were misspelled (eg. AJ781349.1.1478 assigned *Streptomyces morookaense* species name instead of *Streptomyces morookaensis*), and no record in LPSN was found for 4,372 (8.9%) sequences (eg. EU273531.1.1483 assigned *Streptomyces verne* species name). The LPSN status of all 48,981 sequences used in this Chapter is provided in `full_length_strep_records_info.csv` (Supplementary File 6).

In the originating databases, sequences may be annotated with synonyms of taxon names at various ranks, and it is reasonable to expect that these taxon names should be consistent within the same lineage. For instance, a sequence assigned the species name *Streptomyces clavuligerus* should be assigned to Bacteria kingdom, *Actinomycetota* phylum, *Actinomycetes* class, *Streptomycetales* order, *Streptomycetaceae* family, *Streptomyces* genus taxonomic hierarchy. It would be incorrect for this sequence to be classified under a different taxonomic hierarchy such as Bacteria kingdom, *Firmicutes* phylum, *Bacilli* class, *Bacillales* order, *Bacillaceae* family, *Bacillus* genus. Investigation of nomenclature within the source databases at ranks from Kingdom to Genus

(Methodology Section 2.2.5) identifies sequences having the *Streptomyces* keyword in the taxonomy field that are assigned: (i) correctly to all ranks above *Streptomyces*, eg. AWQW01000120.100.1367 belongs to Bacteria kingdom, *Actinobacteriota* phylum, *Actinobacteria* class, *Streptomycetales* order, *Streptomycetaceae* family, *Streptomyces* genus and *Streptomyces niveus* species (note that nomenclature is fluid and, for example, at phylum level *Actinobacteriota* has been superseded by *Actinomycetota* (Oren & Garrity, 2021), but choose to reflect the assigned nomenclature in the database); (ii) to higher ranks not expected to contain the *Streptomyces* genus (eg. KY753270.1.1450 belongs to Bacteria kingdom, *Firmicutes* phylum, *Bacilli* class, *Bacillales* order, *Bacillaceae* family, *Bacillus* genus, and *Streptomyces pseudovenezuelae* species); (iii) to higher ranks that correctly include *Streptomyces* from kingdom to genus, but where the annotated nomenclature nevertheless disagrees on the parent genus name (eg. JN987181.1.1444 belongs to Bacteria kingdom, *Actinobacteriota* phylum, *Actinobacteria* class, *Streptomycetales* order, *Streptomycetaceae* family, *Streptomyces* genus and *Lactobacillus apodemi* species); and (iv) to ambiguous hierarchies where there is a complete lack of information about higher ranks (eg. NR_042095.1). This is consistent with previous observations of taxon names assigned to conflicting nomenclature at higher ranks in Greengenes and SILVA (Edgar, 2018a). All identified nomenclature at ranks from kingdom to genus are summarised in Figure 2.5.

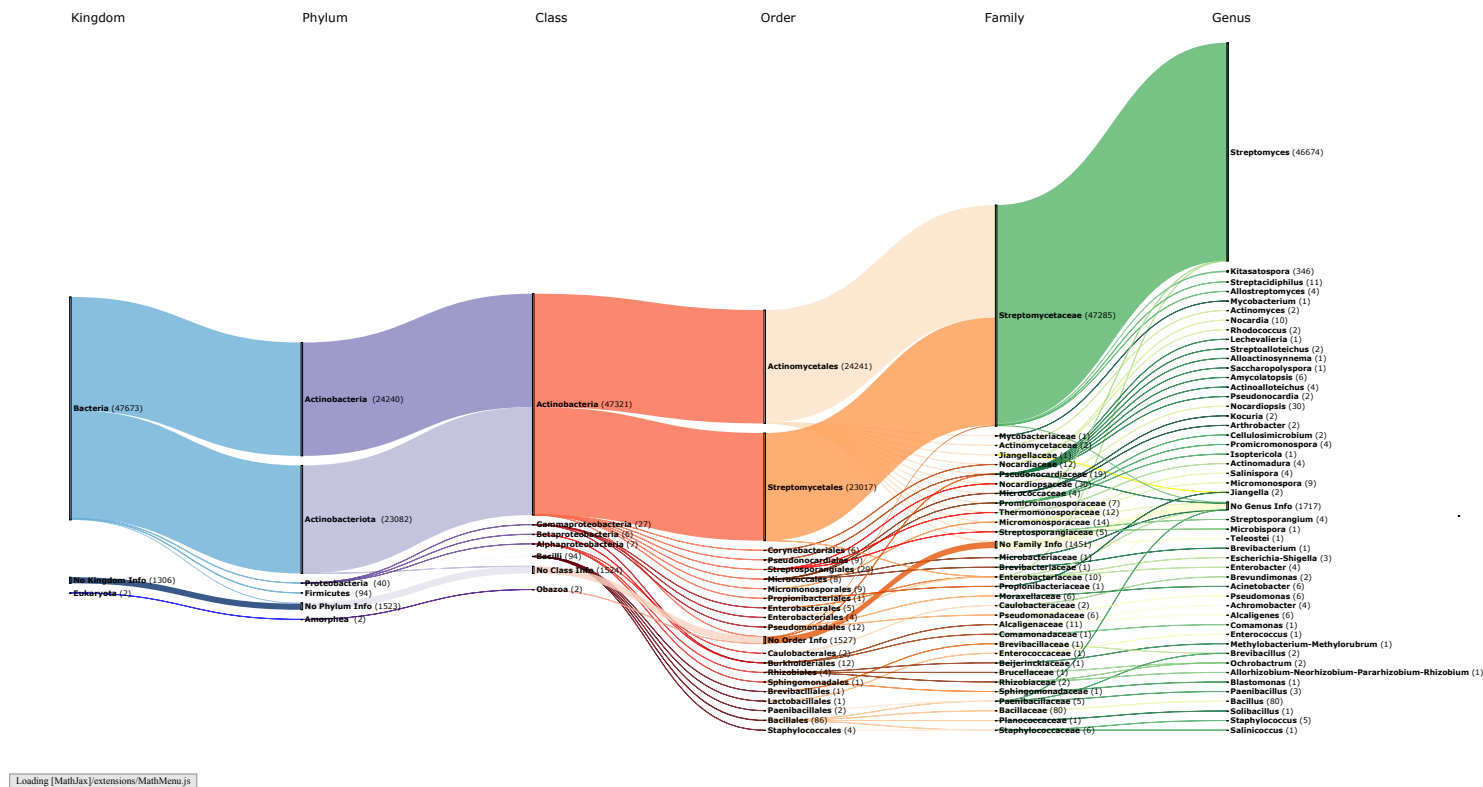


Figure 2.5: Sankey plot showing counts of taxonomic names in source databases, assigned at ranks from phylum to genus, to sequences identified with a key word ‘*Streptomyces*’ in the taxonomy field. Note that Actinobacteria and Actinobacteriota are synonyms in LPSN for the correct Phylum name Actinomycetota, but that Actinomycetales and Streptomycetales are not taxonomic synonyms for each other. Streptomycetales is synonymous in LPSN with the correct name Kitasatosporales; Actinomycetales is a distinct taxonomic Order. The parent order of the Family Streptomycetaceae in LPSN is Kitasatosporales.

Chimeric sequences, and sequences with a high proportion of ambiguity bases, can be disruptive in phylogenetic analyses, leading to model misspecification, incorrect branch length and topology estimation (Lemmon et al., 2009). Eleven sequences with more than 153 nucleotide ambiguity bases were discarded from the dataset prior to analysis; some sequences contained more than 600 ambiguity bases. As the average length of a 16S rRNA gene sequence is 1500bp long, these sequences were deemed to be of low quality.

During the data cleaning process, a further 2,158 potential chimeric sequences were also identified and removed from the dataset. Following the filtration and cleaning process, 22,680 full-length non-redundant high-quality sequences (46% of the initial dataset) were taken forward for further analyses and phylogenetic tree reconstruction. Despite significant and diligent long-term efforts by curators to remove poor quality sequences from the databases used in this analysis (DeSantis et al., 2006), identification and exclusion of poor quality reads was still required to avoid introducing identifiable sources of potential inaccuracy to the analysis (Methodology Section 2.2.6).

2.3.2 16S percentage sequence identity thresholds do not reliably delineate existing *Streptomyces* species assignments

The long-standing 16S rRNA clustering threshold for species separation of 97% sequence identity has been robustly questioned (Edgar, 2018c), and current best practice is to use zOTUs, or Amplicon Sequence Variants (ASVs) for taxonomic classification and clustering of 16S and other marker genes. To identify whether any clustering threshold adequately circumscribes *Streptomyces* taxa, I applied a range of threshold identities to

over 22,000 full-length *Streptomyces* 16S rRNA sequences, then determined the agreement between taxonomic species and cluster membership (Methodology Section 2.2.7). If zOTUs were assumed to be an accurate proxy for species, then with 14,239 known zOTUs one might conclude that there are the same number of candidate *Streptomyces* species. Currently at least 650 species are recognised within *Streptomyces*, and so many observed zOTUs might suggest either a high degree of cryptic species diversity, or that 20 or more distinct 16S rRNA sequences may be characteristic of the same *Streptomyces* species. Initial examination of the pharmaceutically important strains *Streptomyces griseus* (streptomycin producer), *Streptomyces lydicus* (natamycin, lydimycin and streptolydigin producer), *S. clavuligerus* (clavulanic acid and cephamycin C producer; Procópio et al., 2012) and the phytopathogen *Streptomyces scabiei* (Kers et al., 2005) shows that 16S rRNA sequences assigned these species names are split across multiple zOTUs.

Specifically, *S. clavuligerus* is split across 8 zOTUs, *S. griseus* is found in 145 zOTUs, *S. lydicus* in 16, and *S. scabiei* in 62. These observations were not unexpected, as bacteria have multiple 16S rRNA copies and each potentially varying in sequence (Větrovský & Baldrian, 2013), including members of the genus *Streptomyces* (Coenye & Vandamme, 2003). Additionally, these findings align with previous observations, such as those detailed in chapter 1 section 1.7, where it was noted that *S. caeculicus* contains five distinct copies of 16S rRNA sequences (Figure 1.29; Wink et al., 2017). The existence of multiple distinct 16S rRNA sequences corresponding to a single species implies that naïve 16S metabarcoding of communities assuming that zOTU diversity reflects species diversity may overestimate the number of species per sample, as well as result in conflicting branching orders.

As I raise the 16S sequence identity clustering threshold from 98% to 100%, clusters increase in number and tend to have fewer 16S sequence members (Figure 2.6). The number of unique taxa per cluster also tends to fall as the identity threshold approaches 100% (Figure 2.7). I identified 10,548 zOTUs containing two or more sequences. Of these, 8,326 (78.9%) included at least one sequence currently named as *Streptomyces sp.* whose classification is ambiguous, and 4,820 (45.7% of zOTUs) consisted only of such unclassified sequences. These sequences introduce ambiguity and disrupt estimation of the true taxonomic accuracy of 16S marker sequences.

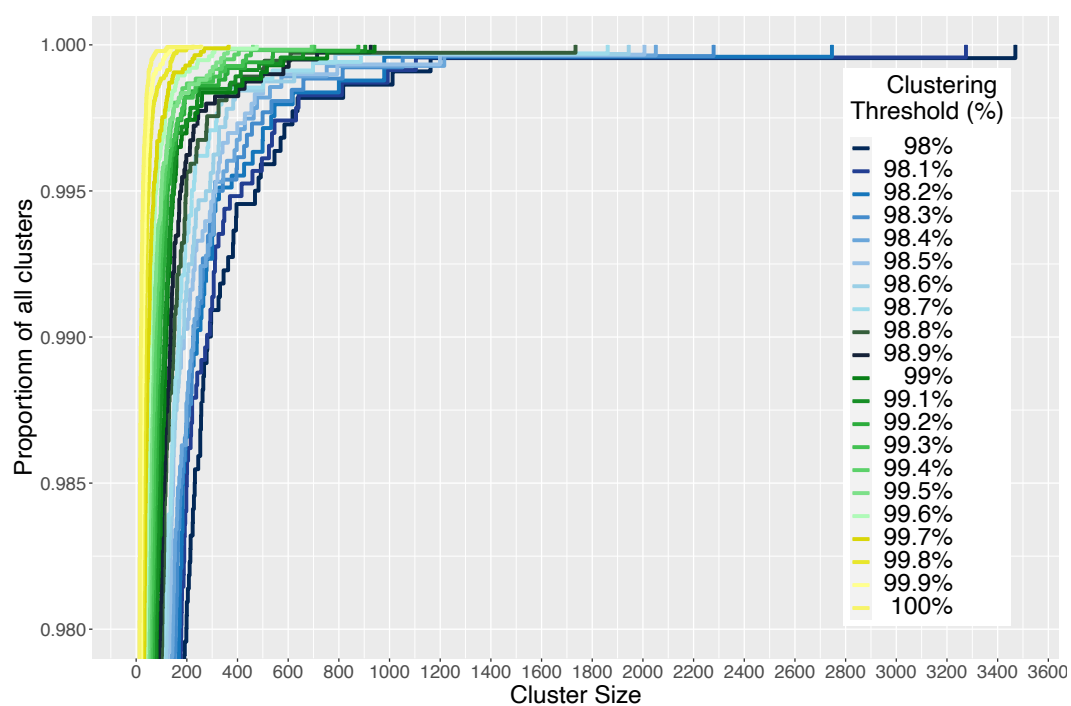


Figure 2.6: Cluster sizes. Empirical cumulative frequency plot showing cluster sizes generated for each clustering threshold.

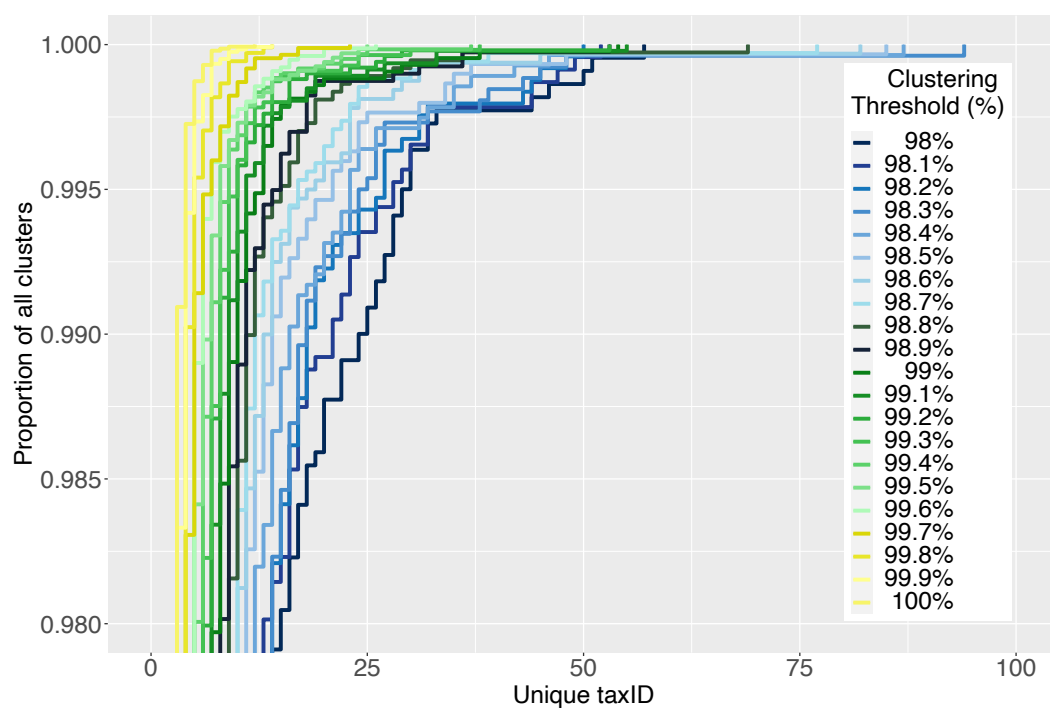


Figure 2.7: Cluster taxID abundance. Empirical cumulative frequency plot for unique taxIDs present at each clustering threshold.

Nevertheless, if 16S rRNA sequences do provide sufficient sequence diversity to distinguish between *Streptomyces* species, and zOTUs are a reliable proxy for taxonomic assignment at or below species-level in *Streptomyces*, then each zOTU cluster should contain only a single unambiguous species name. However, the analysis shows that 3,747 (26%) of zOTUs contain 16S sequences currently assigned to at least two distinct *Streptomyces species* (Figure 2.7). One zOTU notably includes sequences annotated as twelve distinct species (*S. coelicolor*, *S. albidoflavus*, *S. somaliensis*, *S. rutgersensis*, *S. paulus*, *S. limosus*, *S. griseochromogenes*, *S. sampsonii*, *S. resistomycificus*, *S. felleus*, *S. violascens* and *S. sp.*). I therefore find that a substantial fraction of full-length 16S zOTUs do not map exactly to a single *Streptomyces* species assignments, and in particular single 16S sequences frequently map to multiple distinct species.

2.3.3 A comprehensive *Streptomyces* 16S phylogeny

Despite numerous taxonomic inconsistencies when clustering analysis of sequences at 100% identity, zOTUs provide the best possible opportunity for reconstructing a 16S gene-based phylogeny. To estimate the evolutionary relationship amongst *Streptomyces* all 14,239 zOTU sequences were used to produce a multiple sequence alignment (MSA; Methodology Section 2.2.8). Ten outgroup sequences from related non-*Streptomyces* genera were added to the analysis to aid in root placement. The MSA was trimmed to 1,086 nucleotides and 9,049 sequences after redundant sequences were removed (no positions in the alignment were absolutely conserved), and a maximum-likelihood tree was calculated (Methodology Section 2.2.8). Clades containing a single species taxonomic assignment were collapsed to single leaf nodes to facilitate visualisation of

the phylogenetic tree for all 5,064 nodes shown in Figure 2.8 (full tree provided in newick format in Supplementary File 10; 04_TBE.raxml.support; collapsed newick file version provided in Supplementary File 11; collapsed_strep_tbe.new).

To my knowledge this is the largest, most comprehensive 16S rRNA phylogenetic reconstruction attempted for *Streptomyces* to date. No clade received TBE support greater than 60%, and only 18 clades have a TBE support above 50%; hence I do not consider the topology of this tree to be robust as presented. Figure 2.9 describes a phylogenetic tree where clades having TBE value of 50% or higher are marked.

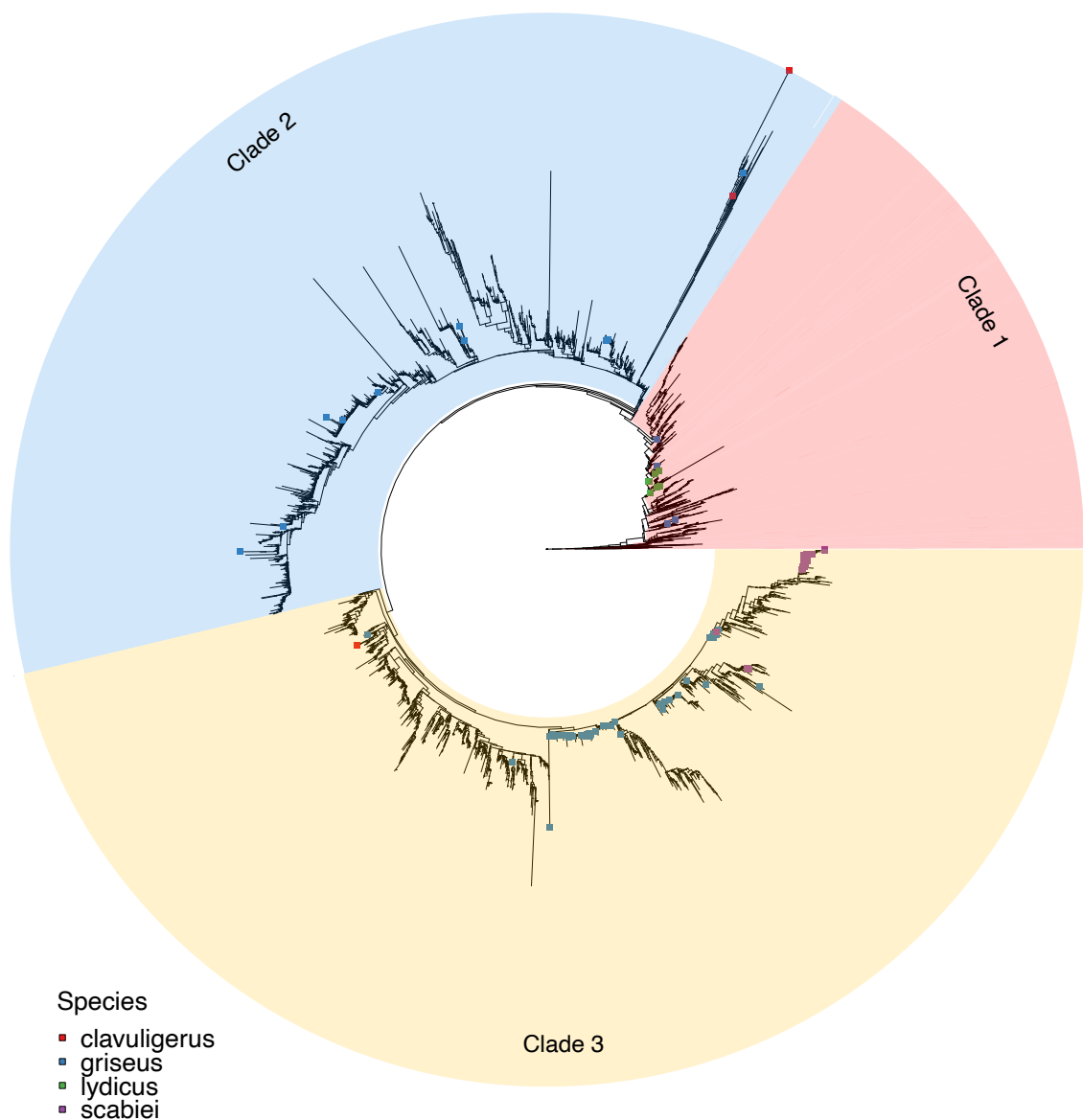


Figure 2.8: Clades containing a single assigned species-level taxon were collapsed to single leaf nodes. Three major clades are highlighted in distinct colours. Squares indicate distributions of 16S sequences assigned the same species names in the source database(s); *S. griseus* sequences are shown in blue, *S. clavuligerus* in red, *S. lydicus* in green, and *S. scabiei* in purple. Sequences with these species assignments tend not to be monophyletic, indicating incongruence between taxonomy and the 16S gene tree.

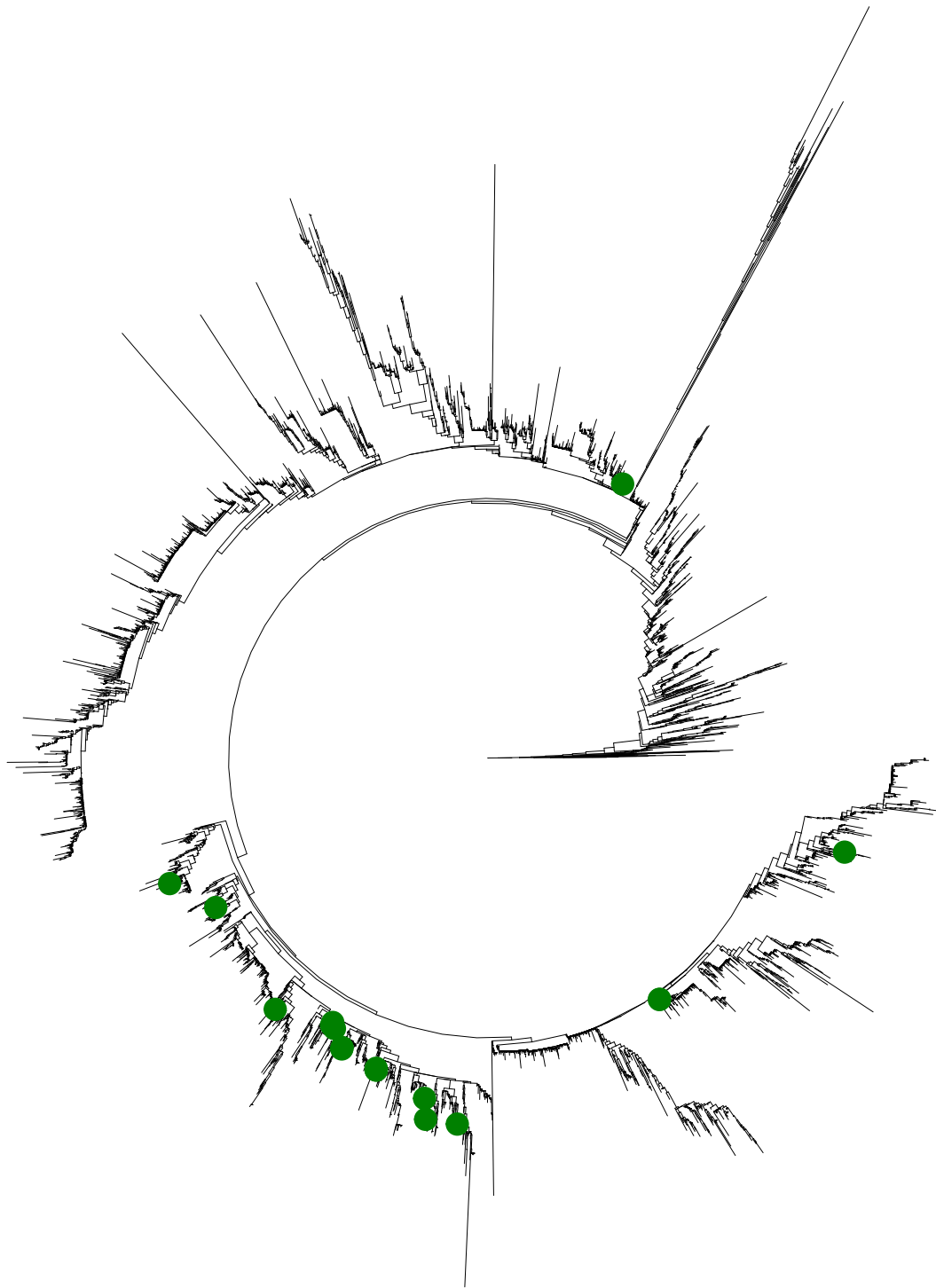


Figure 2.9: Maximum-likelihood tree of the genus *Streptomyces* showing branches with transfer bootstrap expectation support of $\geq 50\%$.

A notable feature of the ML tree is the presence of three clearly distinguishable major clades (Figure 2.8). If 16S sequence databases were simply consistent with the true *Streptomyces* species tree, we might expect to find that 16S sequences taxonomically assigned to the same species are broadly monophyletic (allowing for ancestral duplications, given that a large proportion of *Streptomyces* species contain multiple 16S loci). If the 16S tree departs from this distribution of taxonomic assignment, it might reflect an unreliability in classification of taxa using the existing database assignments. To examine this, I mapped pharmaceutically and agriculturally important isolates to the comprehensive 16S ML phylogeny in Figure 2.8. I find that some species annotations are consistently found within a single clade (*S. lydicus*, *S. albulus* and *S. venezuelae*, Figure 2.10), and some exhibit dispersion within major clades consistent with limited misannotation (*S. scabei*, *S. lavendulae* and *S. rimosus* (Figure 2.11). However, I find some taxon representatives to be distributed widely across the tree, in two or more major clades (*S. griseus*, *S. albus*: Figure 2.12). Other species assignments are represented by an insufficient number of sequences to clearly demonstrate a pattern (*S. clavuligerus* or *S. coelicolor*: 2.13).

Species

- albus
- lydicus
- venezuelae

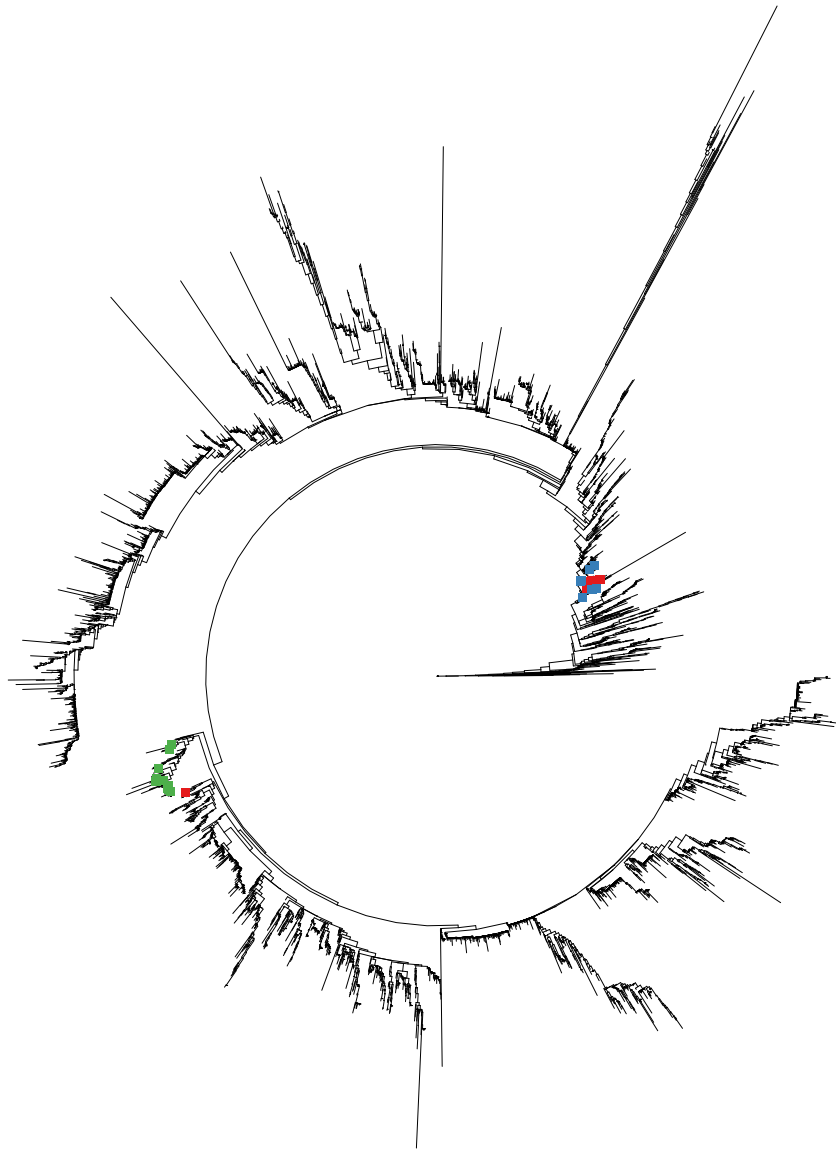


Figure 2.10: Maximum-likelihood tree of the genus *Streptomyces* showing distribution of *Streptomyces albus* (red), *Streptomyces lydicus* (blue) and *Streptomyces venezuelae* (green).

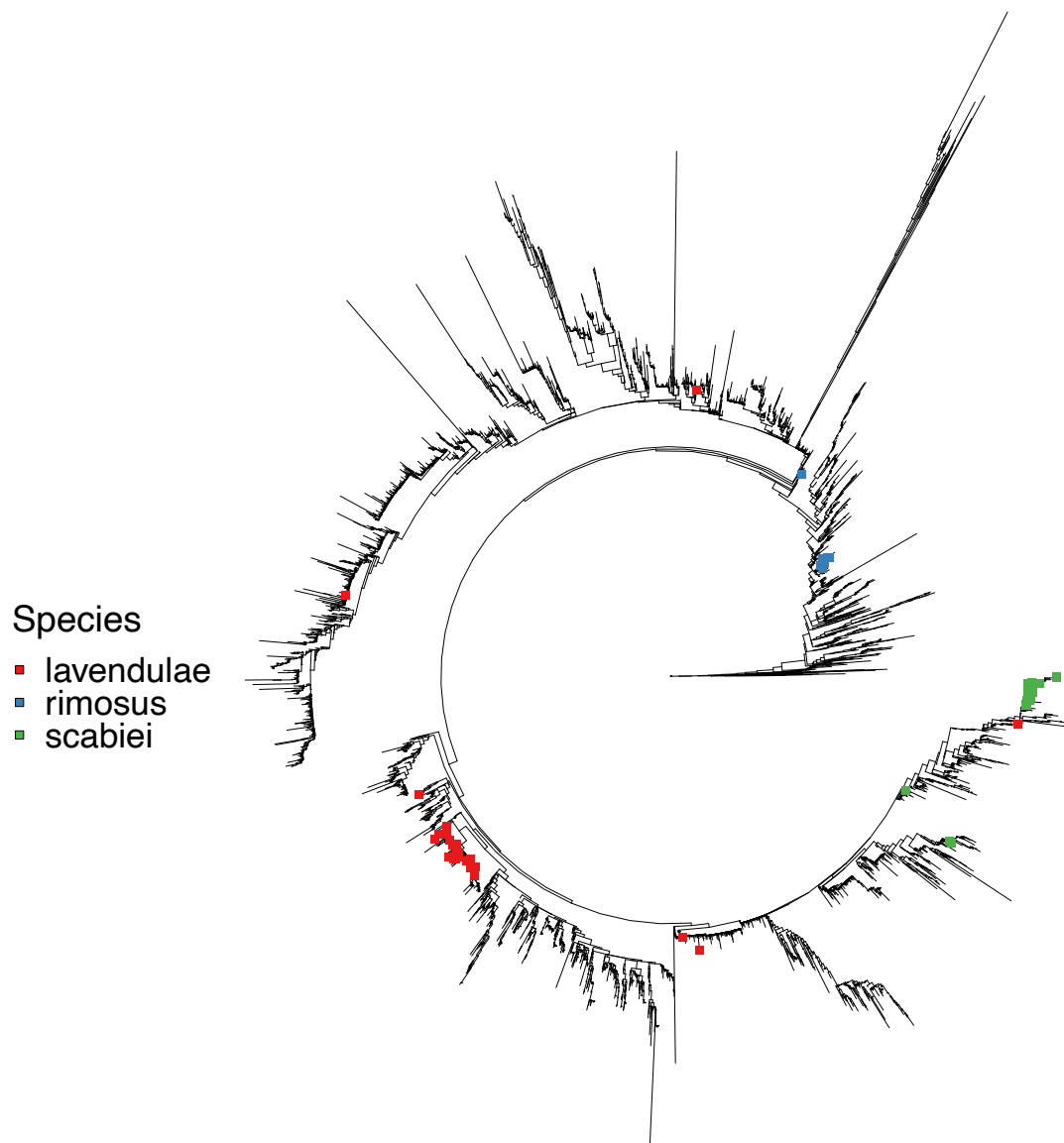


Figure 2.11: Maximum-likelihood tree of the genus *Streptomyces* showing distribution of *Streptomyces lavendulae* (red), *Streptomyces rimosus* (blue) and *Streptomyces scabiei* (green).

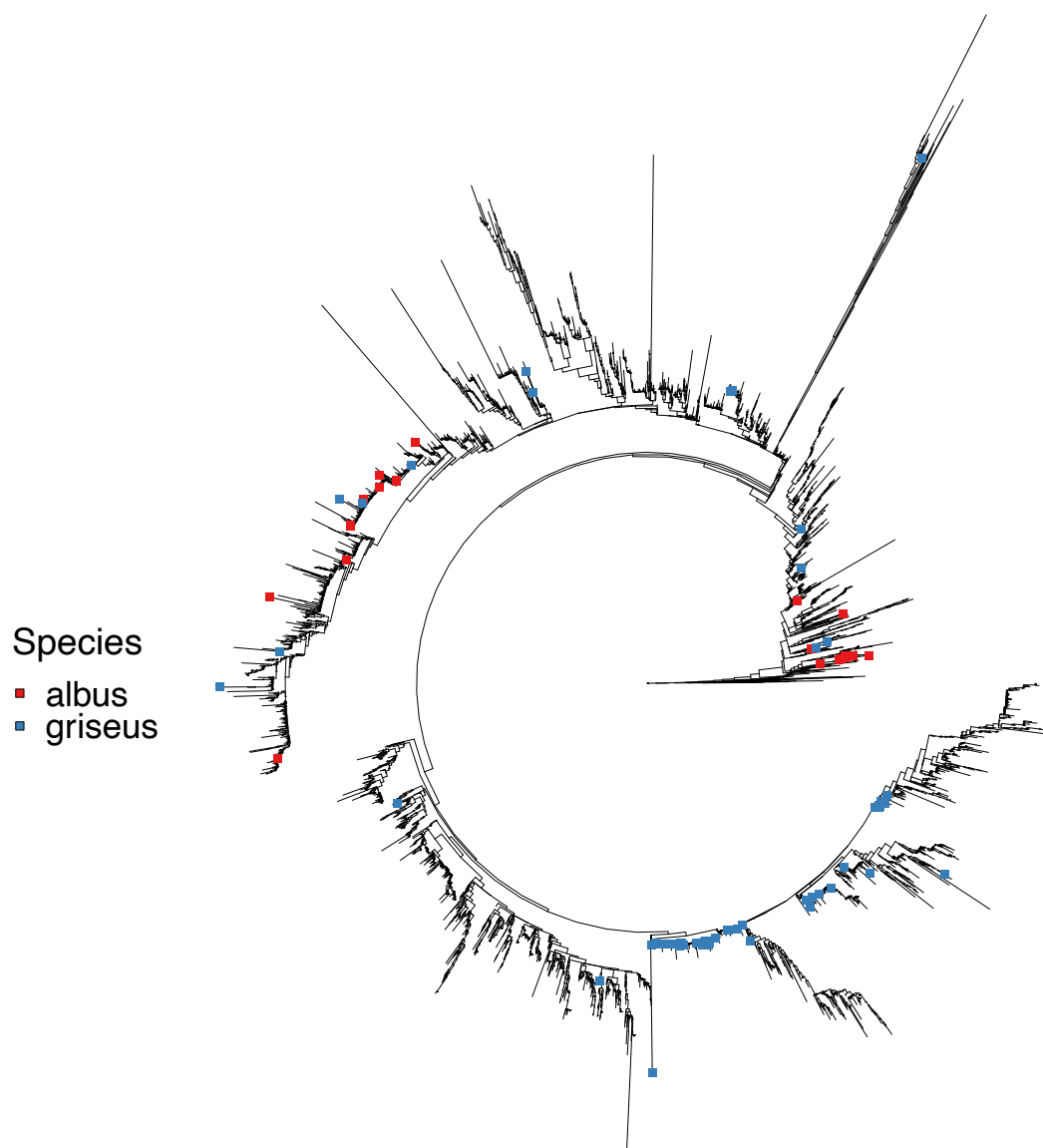


Figure 2.12: Maximum-likelihood tree of the genus *Streptomyces* showing distribution of *Streptomyces albus* (red) and *Streptomyces griseus* (blue).

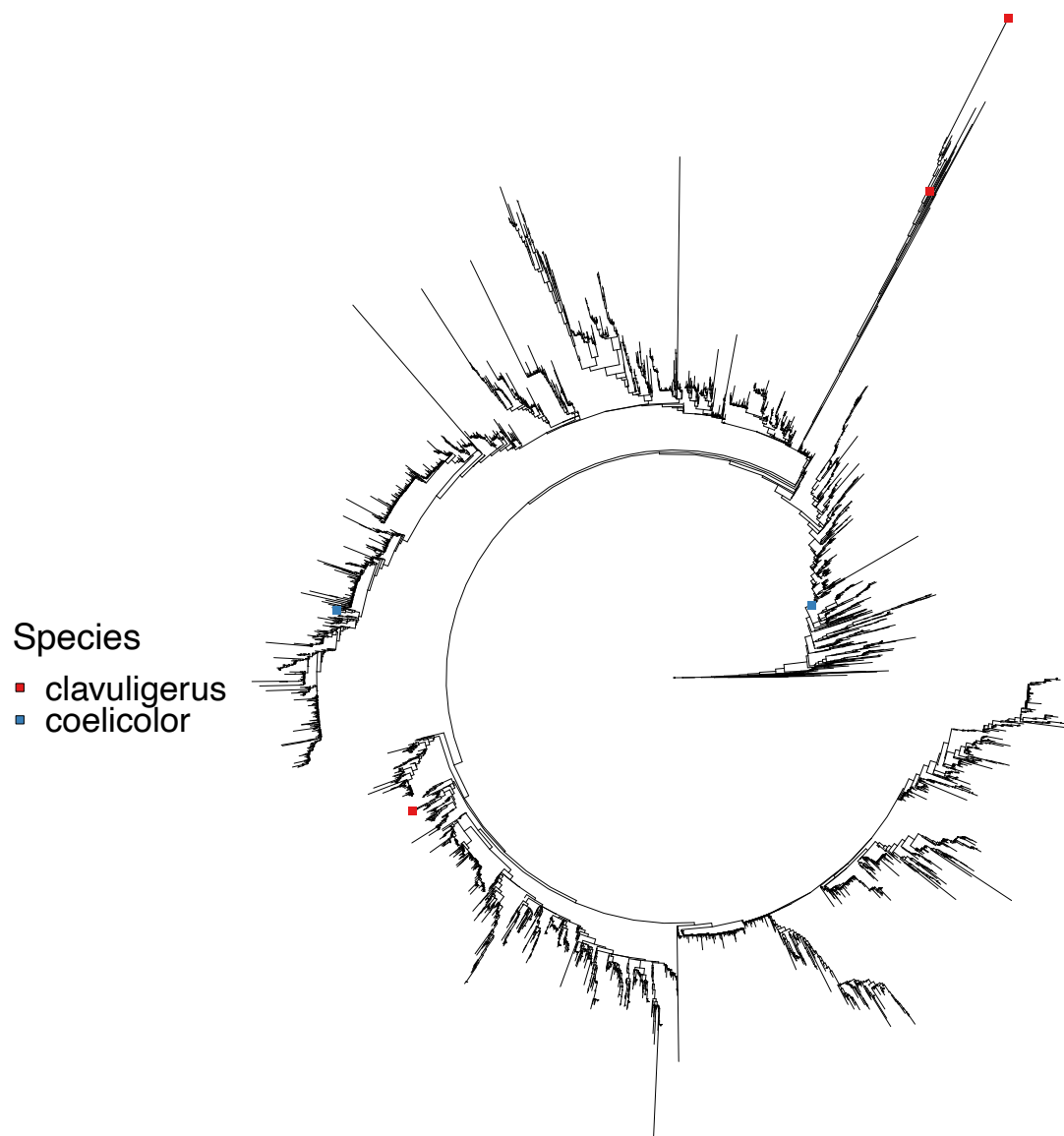


Figure 2.13: Maximum-likelihood tree of the genus *Streptomyces* showing distribution of *Streptomyces clavuligerus* (red) and *Streptomyces coelicolor* (blue).

Additionally, recent reclassifications within the Streptomycetaceae family led to the proposal of six novel genera (Madhaiyan et al., 2022). To determine whether the proposed novel genera resolve in my analysis, I checked their distribution across the 16S ML phylogeny. I find that members of *Wenjunlia* (Figure 2.14) are consistently placed on the comprehensive 16S ML phylogeny, whereas members of *Actinacidiphila* are scattered across the tree (Figure 2.15). Additionally, my analysis did not include any representative from *Peterkaempferia* genus, and the available sequences for *Mangrovactinospora* (Figure 2.16), *Phaeacidiphilus* (Figure 2.17) and *Streptantibioticus* (Figure 2.18) were insufficient to establish a clear pattern in their distribution. Overall I find evidence of taxonomic misassignment across the full scope of 16S sequences, consistent with observations of sequence mis-annotation previously estimated for SILVA and GreenGenes to be around 17% at ranks up to phylum, and a similar mis-annotation rate of 10% in RDP (Edgar, 2018a). It is possible that, in some cases, the apparent dispersion of a single taxon across the tree could be the result of limited sequence variation within the 16S rRNA and failure to obtain a robust phylogeny, given that most internal nodes have a TBE support value of lower than 50% (Figure 2.9). However, the three major clades do appear to be relatively robustly distinguished in the phylogeny, and where the same taxon has representatives in two or more of these clades I consider that this calls into question the corresponding assignment.

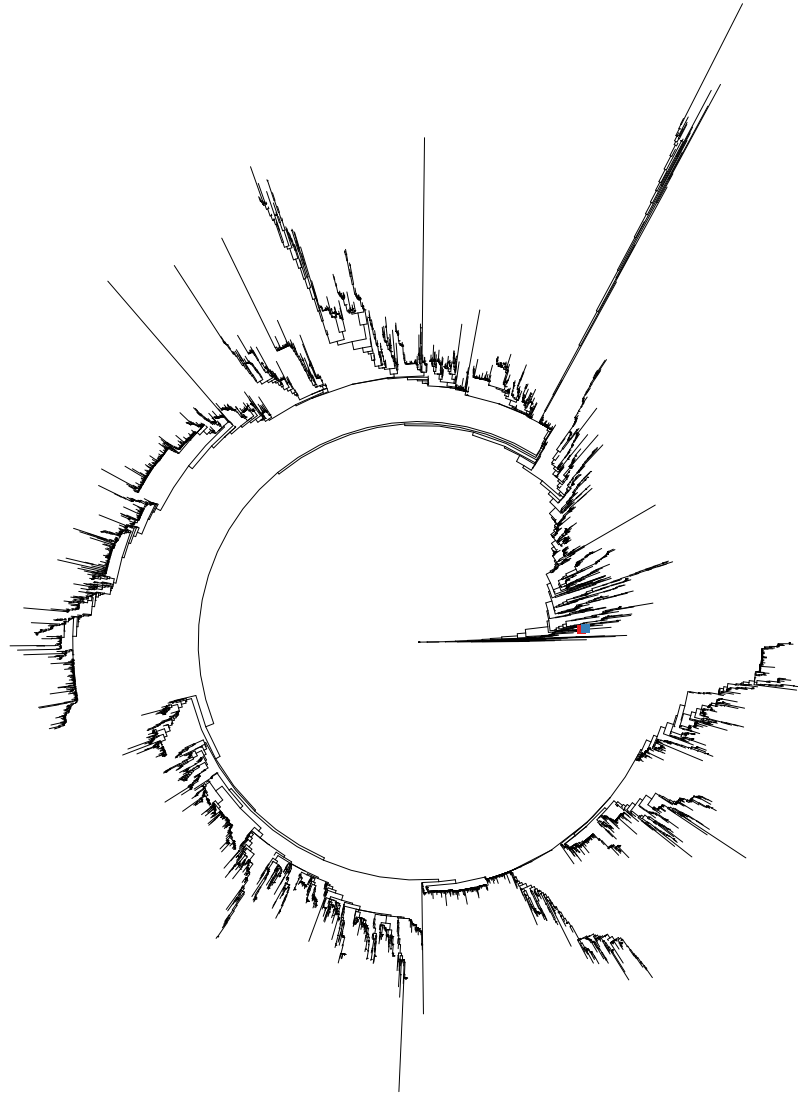


Figure 2.14: Distribution of members of the novel *Wenjunlia* genus on the ML tree.

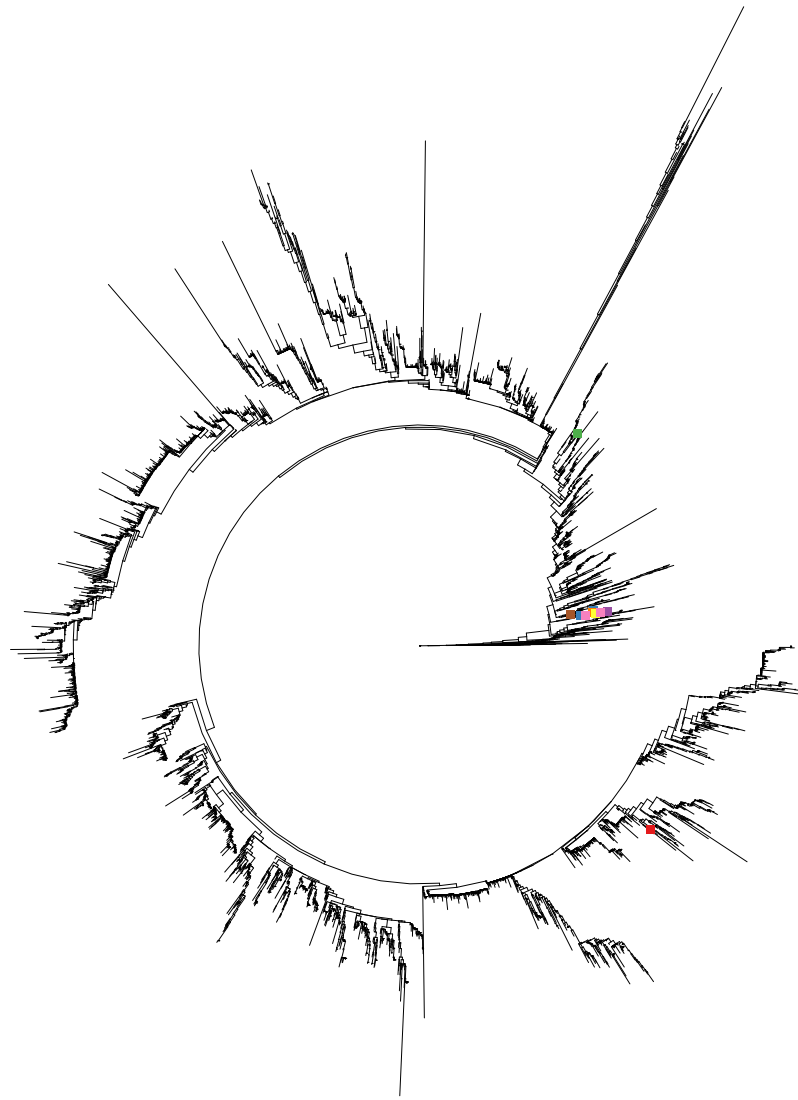


Figure 2.15: Distribution of members of the novel *Actinacidiphila* genus on the ML tree.

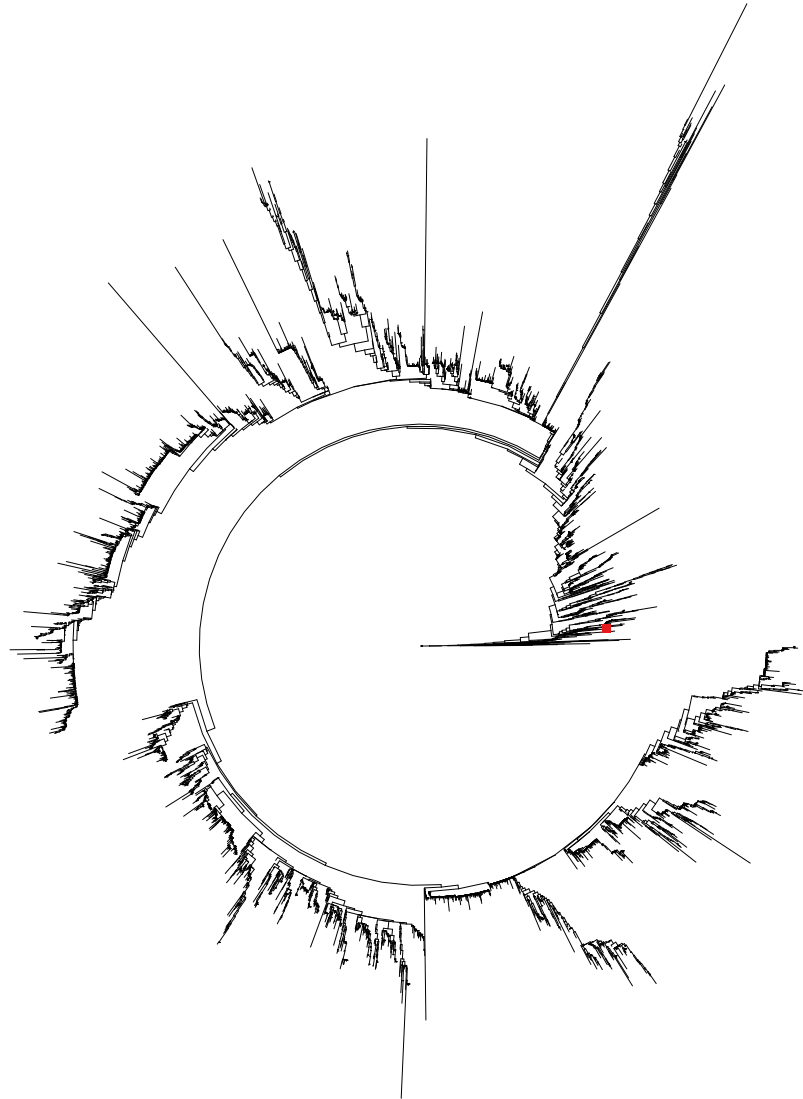


Figure 2.16: Distribution of members of the novel *Mangrovactinospora* genus on the ML tree.

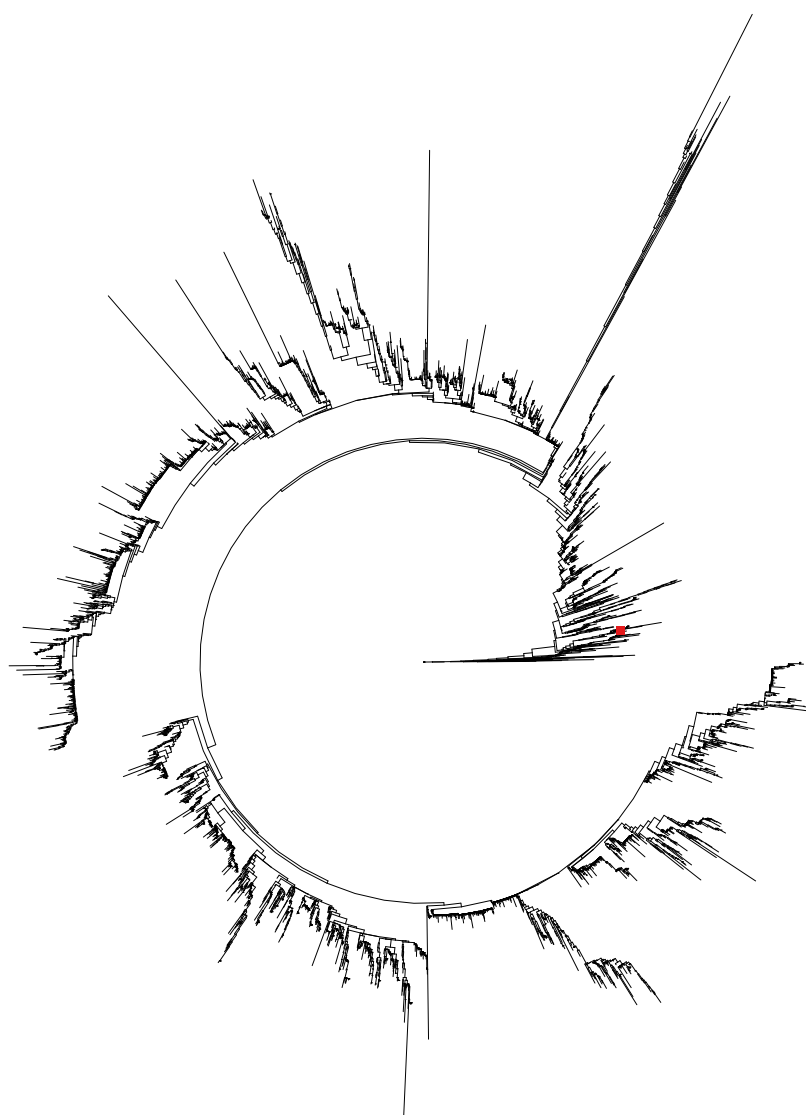


Figure 2.17: Distribution of members of the novel *Phaeacidiphilus* genus on the ML tree.

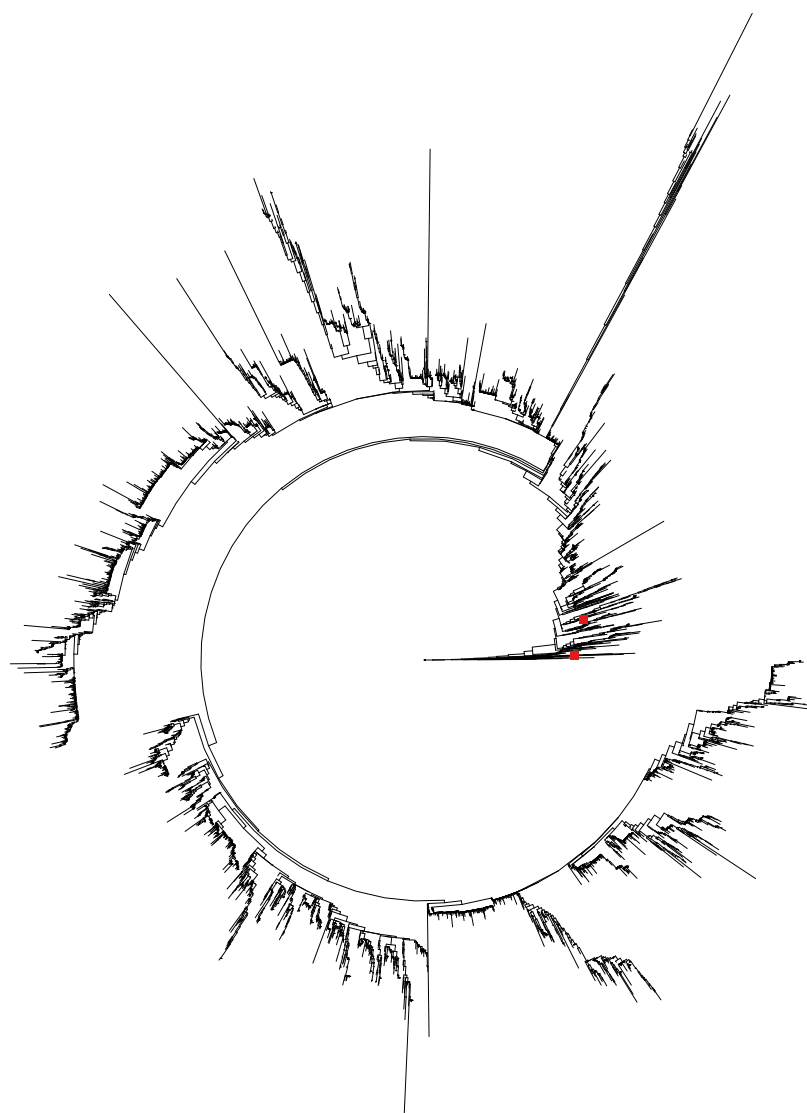


Figure 2.18: Distribution of members of the novel *Streptantibioticus* genus on the ML tree.

2.3.4 Whole-genome sequence classification indicates that distinct *Streptomyces* species can share identical full-length 16S sequences

No strain information linking to sequenced genomes could be recovered for 16S rRNA sequences from SILVA, Greengenes, RDP or NCBI new ribosomal RNA BLAST databases. Therefore, to map the 16S sequence landscape to the whole-genome classification landscape, I extracted 6,692 16S rRNA full-length sequences from 2,156 publicly available *Streptomyces* genomes (Methodology Section 2.2.9). Eighty-seven of the published assemblies lacked identifiable 16S rRNA sequences, and 700 genomes contained at least one 16S rRNA sequences with ambiguity bases or partial sequences that could lead to biased observations and overestimation of the intragenomic diversity of 16S rRNA sequences. For sequences with ambiguity bases, it is unclear whether these sequences match the remaining copies in the genome or represent distinct sequences, due to the uncertainty about ambiguous bases. Similarly, for partial sequences, even if the known portion aligns with existing full sequences, the missing segments could either match or differ from other sequences in the genome. Therefore, these genomes were excluded from the analysis, yielding a dataset comprising 4,227 16S rRNA sequences from 1,369 genomes containing only full-length, ambiguity base-free 16S rRNA sequences.

Streptomyces genomes most commonly contain six copies of 16S rRNA operons (Wezel et al., 1991). Across the 1,369 assemblies analysed I find that 16S rRNA sequence copy number varies from one to twelve copies per genome (Figure 2.19). It was found that 359 (26.2%) of the assemblies contained six copies, 144 (10.5%) had more than six copies, and 865 (63.1%) contained fewer than six copies of 16S rRNA. The genomes

of 375 (27.4%) assemblies were found to contain multiple non-identical 16S rRNA sequences. A single assembly (GCF_900199205.1) was found to contain eight distinct 16S rRNA sequences but the majority, 993 genomes (72.5%), possessed only a single 16S rRNA sequence variant, of which 811 contained only one detectable copy of 16S rRNA. Inconsistency in the number of 16S rRNA operons and their intragenomic heterogeneity could be due to a range of causes, but, in addition to the underlying biological variation, is likely due to inclusion of *Streptomyces* genomes assembled to different levels of completeness and quality (eg. contig, scaffold, complete and chromosome). Assemblies could also have been affected by the presence of duplication artifacts or collapsed 16S rRNA sequences resulting from sequencing errors, such as overassembly of short reads from distinct 16S rRNA loci into a single artifactual 16S sequence. It may be assumed that NCBI complete and chromosomal *Streptomyces* assemblies reflect the true genomic heterogeneity of 16S rRNA sequences. In this context, complete assemblies include both complete chromosome and plasmids, whereas chromosomal assemblies contain the entire chromosome but exclude plasmids. Given this, it would be expected that 69% of *Streptomyces* isolates would contain multiple distinct 16S rRNA sequences (Figure 2.20). This proportion is consistent with previous observations that *Streptomyces* strains may contain multiple distinct 16S copies, but twice that observed across all *Streptomyces* genome assemblies in NCBI (Khadayat et al., 2020).

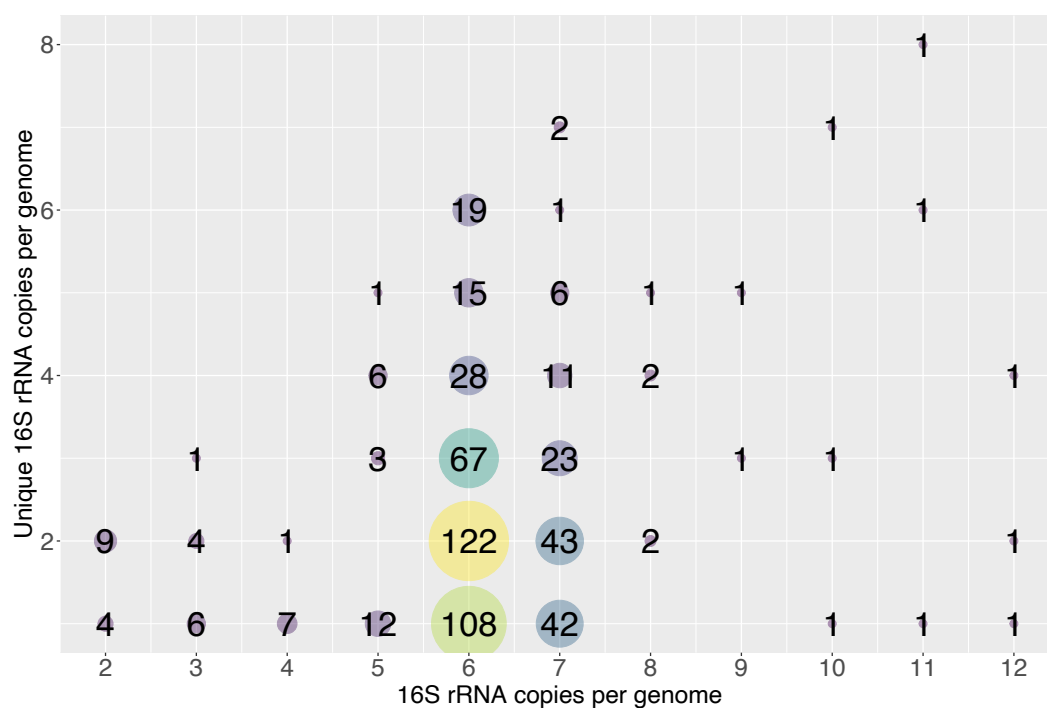


Figure 2.19: Intragenomic 16S rRNA heterogeneity within 1,369 *Streptomyces* genomes which exclusively contain only full-length and ambiguity base-free 16S rRNA sequences. A total of 811 genomes containing single 16S rRNA sequences are not shown.

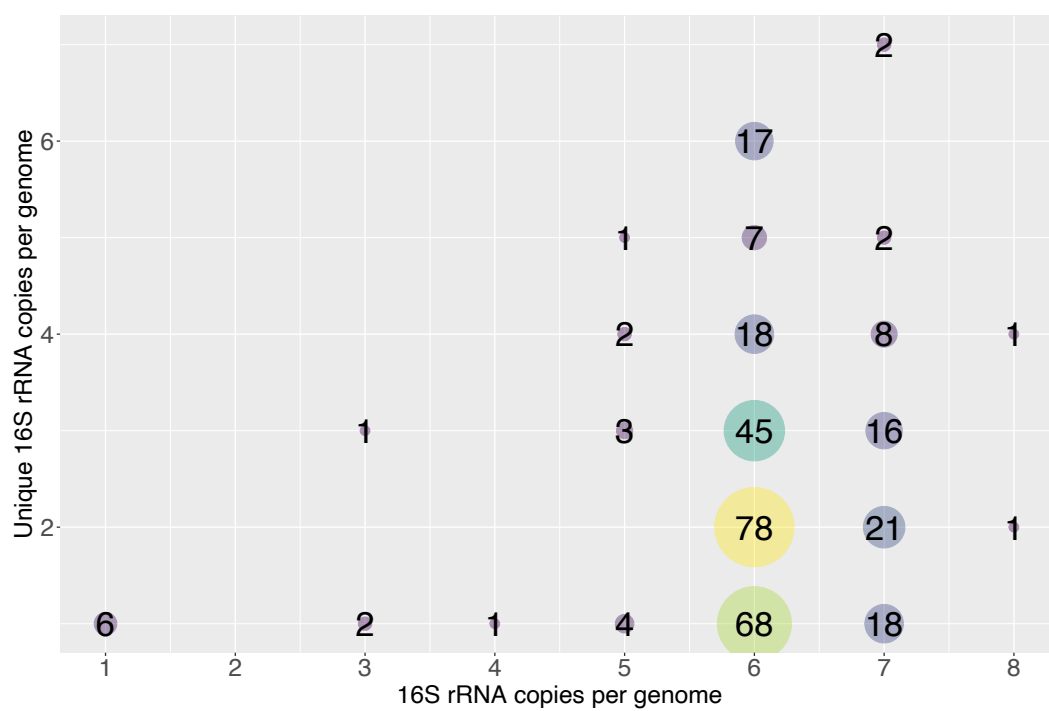


Figure 2.20: Distribution of 16S copies per genome with a distinction between unique and total copies for genomes at assembly level complete and chromosome.

To further investigate the potential for inaccurate *Streptomyces* taxonomic assignment when using 16S rRNA sequences, I examined the relationship between distinct genomic 16S sequences and the species assignments of their corresponding genomes. I examined the distribution of 16S sequences by the number of genomes they occur in, and the number of uniquely assigned species names in NCBI associated with those genomes (Figure 2.21). I find that a single 16S sequence variant may be represented in as many as 33 genomes, and be associated with a group of genomes assigned as many as six species names.

Previous whole-genome analyses of *Streptomyces* also observed that identical 16S rRNA sequences are present in strains assigned to different species (Komaki & Tamura, 2021). Some of the discrepancy may arise from differing approaches to, and knowledge of, taxonomic assignment over time that leads to, for example, the same strain being assigned to a different species depending on when the analysis was done. However, some of these observations may genuinely reflect common 16S sequence shared across sequence boundaries, as 16S substitution rates may be slow in relation to speciation events.

For some *Streptomyces* species multiple distinct 16S rRNA sequences can be found within the same genome (Figure 2.20), implying that there is a one-to-many mapping between *Streptomyces* species and 16S rRNA sequence. It follows that it is not always possible to cluster *Streptomyces* 16S sequence data without splitting a single organism into multiple zOTUs. Simple counts of *Streptomyces* zOTUs when metabarcoding with 16S may thus overestimate species numbers. It also follows that comprehensive

16S rRNA gene trees reflect gene histories, and may not directly recapitulate the corresponding accurate species trees (results and discussion section 2.3.3).

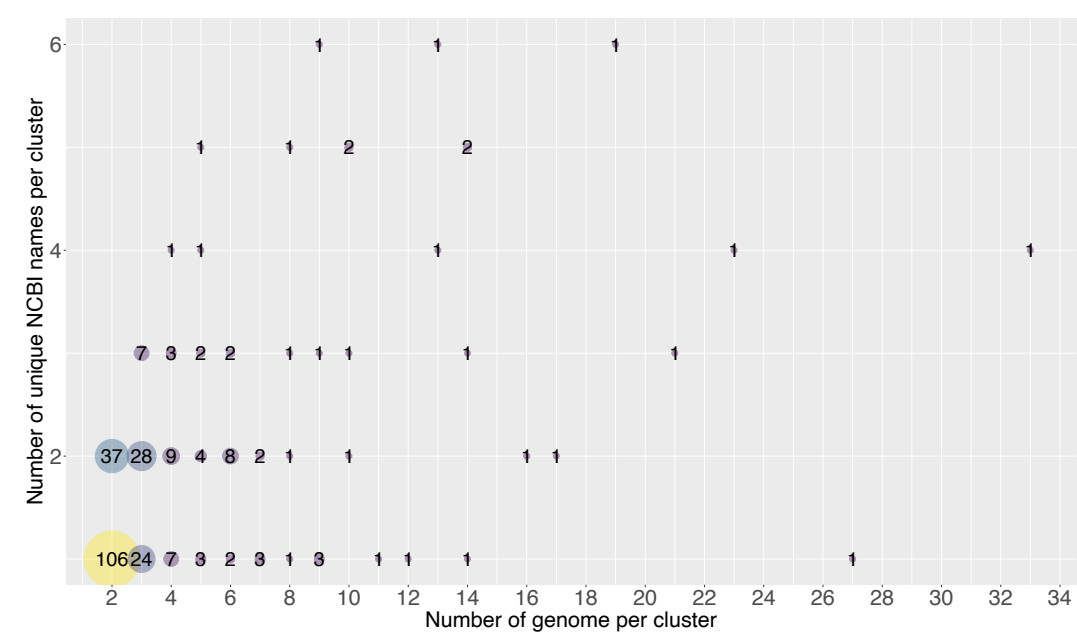


Figure 2.21: Genomes sharing identical 16S rRNA sequences are assigned different names in NCBI. A total of 1,030 singleton clusters are not shown.

I constructed network graphs from genome-derived 16S sequences to visually represent connections between *Streptomyces* genomes based on their common 16S sequences, and thereby interpret the relationship of this network to whole-genome similarity based species classifications (Methodology Section 2.2.9). I represented each of 1,369 *Streptomyces* genomes as a node in the graph, connecting two genomes with an edge if they shared an identical full-length 16S rRNA sequence. The network analysis resolved the 1,369 *Streptomyces* genomes into 709 connected components (Figure 2.22). The largest connected component united 47 genomes, but 527 (74.3%) genomes were singletons, sharing no 16S sequence with any other *Streptomyces* assembly.

If there was a direct mapping between 16S rRNA sequences and species, then connected components would be expected to form cliques (k-complete graphs) where each genome within a single connected component would be linked to every other genome in the same component by at least one edge. However, 22 connected components formed non-cliques (Figure 2.22), indicating that some genomes within a single connected component may not share any identical 16S sequence with some of the other genomes in the same component. If all members of a component belong to the same taxon, this would imply that two members of the same taxon might share no 16S sequence in common. Alternatively, some genomes from distinct taxa may share identical 16S rRNA sequences, perhaps resulting in multiple species being found within a single connected group of genomes. This might arise for a number of reasons, including inter-species recombination (Doroghazi & Buckley, 2010; Tidjani et al., 2019) or selective pressures from natural products that act upon the ribosome (Hansen et al., 2003).

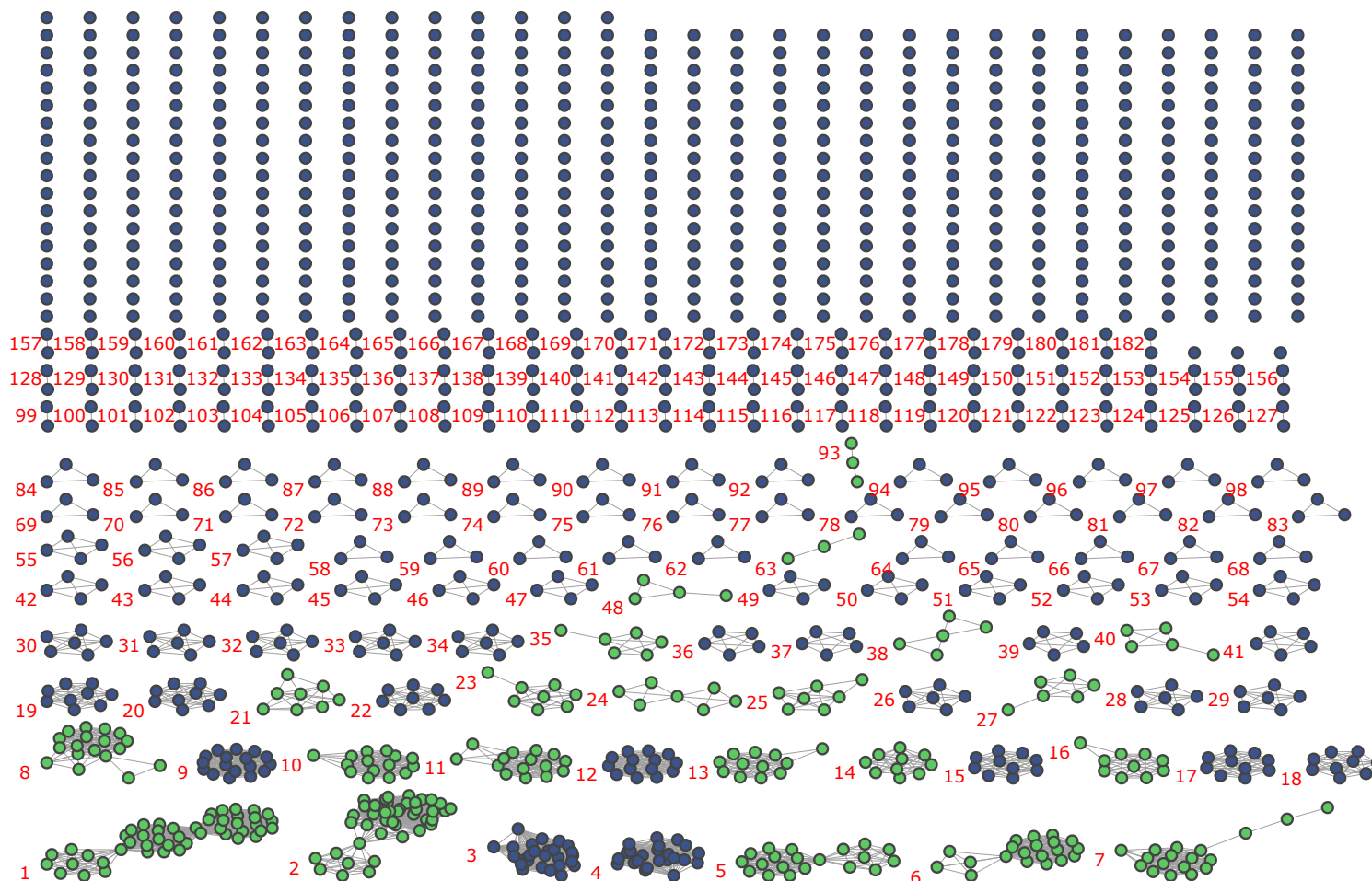


Figure 2.22: Network graph with 1369 genomes and 709 connected components. Each node represents a distinct genome assembly, each edge corresponds to at least one identical 16S sequence being shared between that pair of genomes. Blue connected components form cliques, in which every genome shares at least one identical 16S sequence with all other genomes in the same connected component. Green connected components do not have this property. Assigned connected component IDs are displayed in red. No IDs are shown for components consisting of a single node/genomes improve visualisation.

I performed ANI analysis on the genomes comprising each connected component to establish whether the subgraph corresponded to a single grouping of genomes at genus or species level. I defined genomes as belonging to the same candidate genus if they shared at least 50% genome coverage, and belonging to the same species if they shared at least 95% ANI (methodology section 2.2.9). These boundaries are approximations, but correspond to commonly-used heuristics (Richter & Rosselló-Móra, 2009).

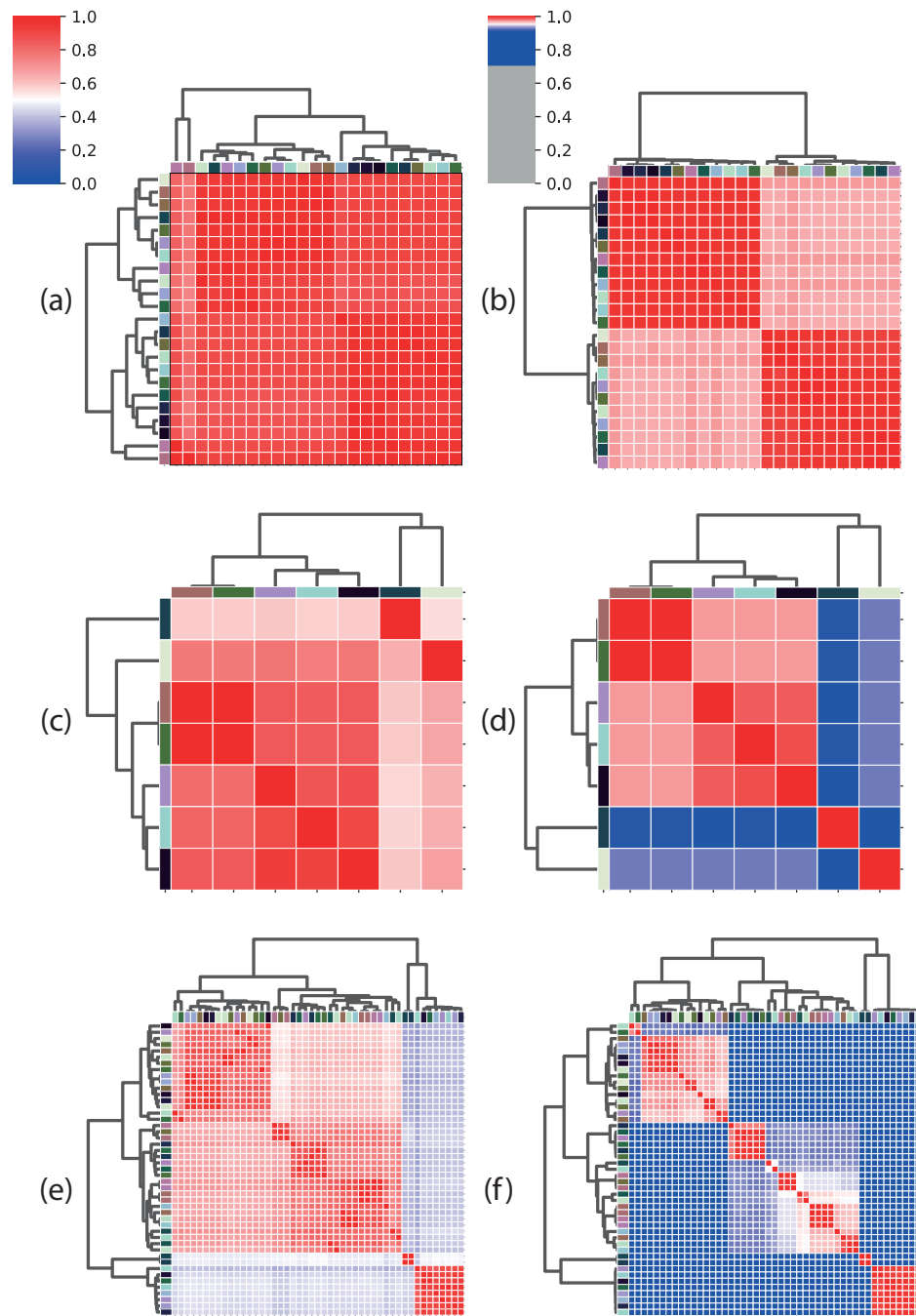


Figure 2.23: Heatmaps of ANIm coverage (left), and ANIm identity (right) for three example connected components from Figure 2.22. Heatmaps in the same row correspond to comparisons for the same connected component. The left column represents percentage genome coverage, the right column %ANI. Red cells in coverage plots correspond to genome coverage of 50% or above, interpreted as common membership of the same genus; blue cells correspond to coverage below 50% and imply distinct genus assignments. In ANI plots, red cells correspond to genome identity of 95% or above, interpreted as membership of the same species; blue cells represent imply distinct species. In some cases ANIm species classifications map onto components containing genomes from a single genus and species (a-b; connected component 4), to distinct species in the same genus (c-d; connected component 24), or to distinct genera (e-f; connected component 1).

Table 2.3: Summary statistics for taxonomic composition of subgraphs uniting at least two genome assemblies.

Category	Interpretation	Count	Percentage
$\geq 50\%$ coverage; $\geq 95\%$ identity	same species	154	84.62%
$\geq 50\%$ coverage; $< 95\%$ identity	same genus, different species	25	13.74%
$< 50\%$ coverage; 95% identity	different genera	3	1.65%

By mapping ANIm coverage to visually represent the distribution of unique candidate genera within each connected component, I find that 174 (98.35%) non-singleton subgraphs unite isolates that share at least 50% genome coverage, indicating likely membership of the same genus (Figure 2.24). However, three connected components (component 1, 32 and 23; Figure 2.24) appear to comprise assemblies from distinct candidate genera. Using our whole-genome comparison threshold to define genus, and although the prevalence of such groups is low, I find that full-length 16S sequences are not always sufficient to resolve *Streptomyces* at genus level. I find, with a similar analysis using %ANIm identity (Figure 2.25), that the majority (84%) of connected components likely represent a single species. However, 28 connected components (Figure 2.25) contain assemblies representing multiple species.

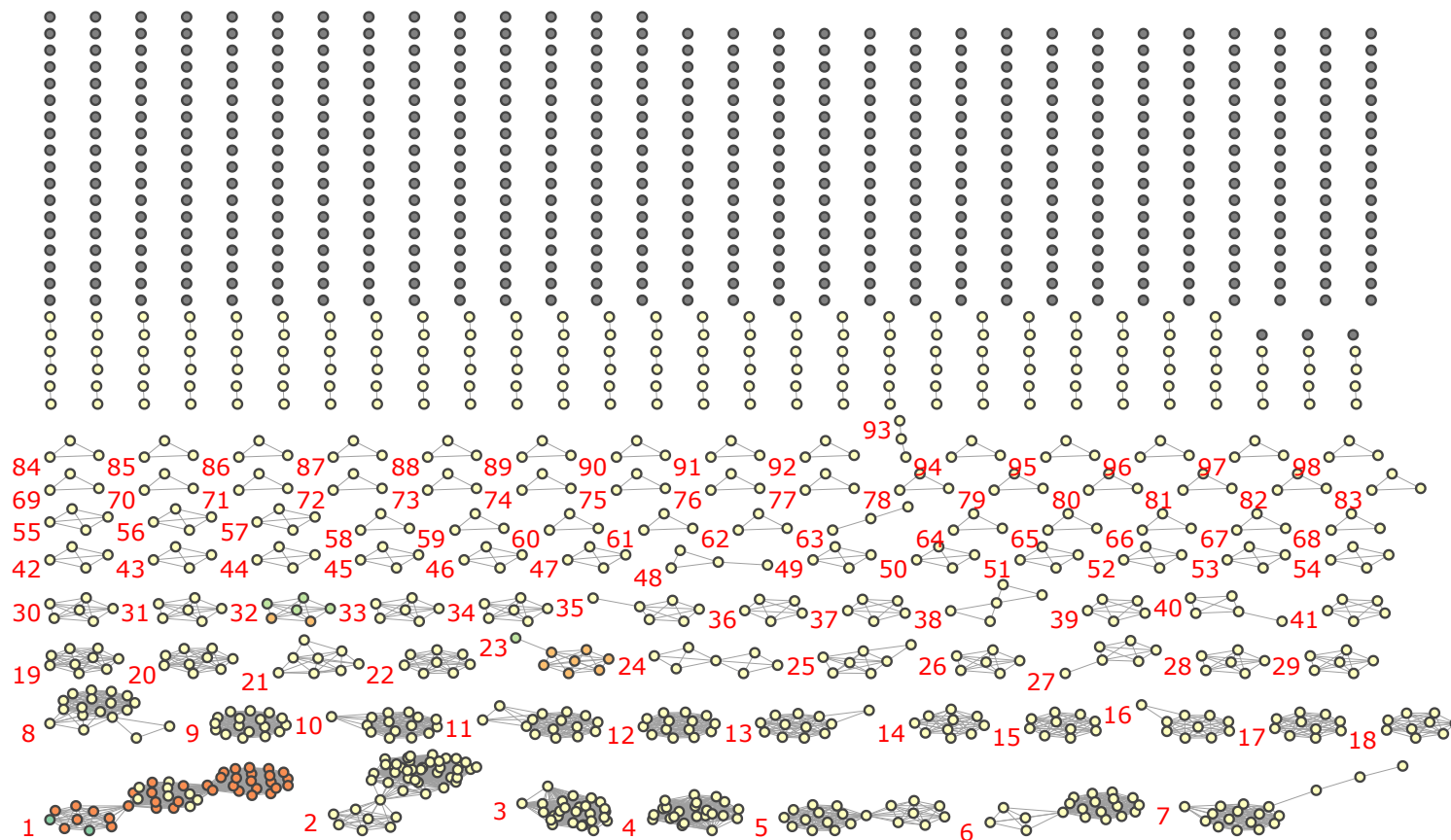


Figure 2.24: Network graph of genomes sharing common full-length 16S sequences showing number of unique genera within each connected component. Each candidate genus is represented as a single node colour within a connected component. Connected components 1, 23 and 32 connect assemblies from distinct candidate genus.

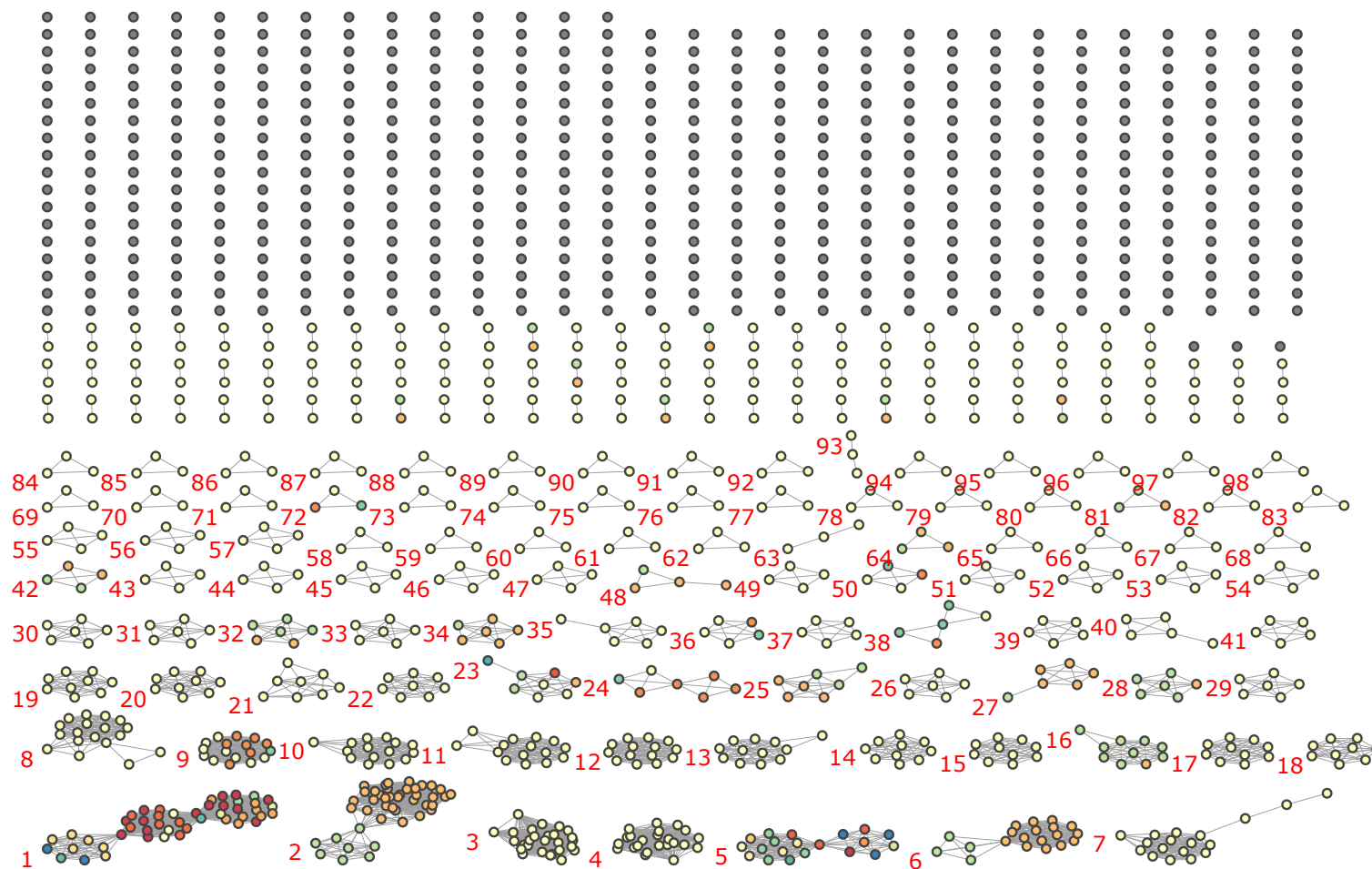


Figure 2.25: Network graph of genomes sharing common full-length 16S rRNA sequences showing number of unique species within each connected component. Each candidate species is represented as a single node colour within a connected component.

This data are consistent with previous observations made by Chevrette et al., 2019b, that distinct *Streptomyces* species confirmed by ANI may share identical 16S rRNA sequences. Similar observations have also been reported in other bacterial genera, suggesting that 16S rRNA sequence identity alone may not be sufficient for species delineation (Bartoš et al., 2024). Moreover, my results also indicate that some networks of genomes which would be assigned as the same species using ANI do not form cliques linked by identical 16S sequences, and so it is possible that genomes assigned to the same *Streptomyces* species by whole-genome methods may not share any identical 16S sequences. Thus, in *Streptomyces*, there is a one-to-many mapping from *Streptomyces* species to 16S sequence, and a one-to-many mapping from 16S sequence to species. Taken together, these results demonstrate that use of 16S rRNA sequences in isolation for taxonomic classification of *Streptomyces* (as is often the case in 16S metabarcoding) can lead to misclassifications not just at the species level, as might be expected, but also at genus level.

To further delineate the relationship between 16S sequence variation and whole genome taxonomy, I measured relatedness within each 16S zOTU using ANIm (Methodology Section 2.2.9). I define a single zOTU genome cluster as a group of at least two genomes sharing an identical, full length and ambiguity base-free 16S rRNA sequence (Figure 2.22). I classified pairs of genomes as belonging to the same genus if they share at least 50% ANIm coverage, and belonging to the same species if they share at least 95% ANIm identity, as before. I show the pairwise comparison results as 1D scatter plots of pairwise genome coverage (Figure 2.26) and pairwise genome identity for each

zOTU (Figure 2.27). In the figure I subdivide zOTUs into groups corresponding to the number of distinct species currently assigned to the *Streptomyces* genomes containing that OTU (zOTUs used in this analysis contain between one and six distinct assigned *Streptomyces* species). I further overlay whole-genome comparison information by colouring comparisons differently if the participants correspond to distinct genera or species by our ANIm thresholds.

I find four (1.4%) clusters containing assemblies that share less than 50% genome coverage, some with as little as 32% coverage (Cluster 2, assigned six unique taxon names; Figure 2.26 – Red Box). I find that it is rare, but possible, for genomes from different candidate *Streptomyces* genera to share an identical 16S rRNA sequence. I also find that 36 (13%) of clusters include assemblies sharing less than 95% ANI (as low as 86% ANI in Cluster 2 with six unique taxon names; Figure 2.27 - red box). This observation is especially evident in, but not restricted to, clusters whose genomes have already been assigned distinct species names in NCBI. These results again show a one-to-many mapping between 16S sequence and *Streptomyces* genera and species as determined by whole-genome comparison, and that assignment of taxonomy based only on 16S rRNA sequences may be misleading. However frequent the potential for misclassification, our data confirm that 98.6% of clusters comprise only representatives of a single genus, and 86% representatives of a single species, as determined by whole-genome comparison. Our *Streptomyces* genome sample is large but not exhaustive, so this may truly reflect that 16S rRNA sequences are often unique to a single species or genus. However, it remains possible that some of these 16S sequences may also be found in as yet unsequenced (or unreleased) assemblies with a different taxonomic classification.

I note that some *Streptomyces* appear to be classified with more precision, potentially due to their industrial or medical importance: members of cluster 139, representing 27 genomes currently assigned to *Streptomyces clavuligerus*, share 96% coverage, and 100% identity. *S. clavuligerus* is an industrially important organism due to its ability to produce clavulanic acid (Liras & Martín, 2021). By contrast, individual members of 125 (45%) clusters seem to have been assigned incorrectly to distinct species, as all members of the cluster share at least 50% genome coverage and 95% identity. Overall, our data show that, while many 16S rRNA sequences do resolve to single species level, there is not a one-to-one mapping between 16S and whole-genome taxonomy and, in general, 16S does not provide sufficient resolution to discriminate between species. I also conclude, on the basis of these observations, that extensive revision of the genus *Streptomyces* is required, in line with recent work based on a smaller dataset of 456 strains, suggesting that there are at least six validly-describable genera within the current single genus *Streptomyces* (Madhaiyan et al., 2022).

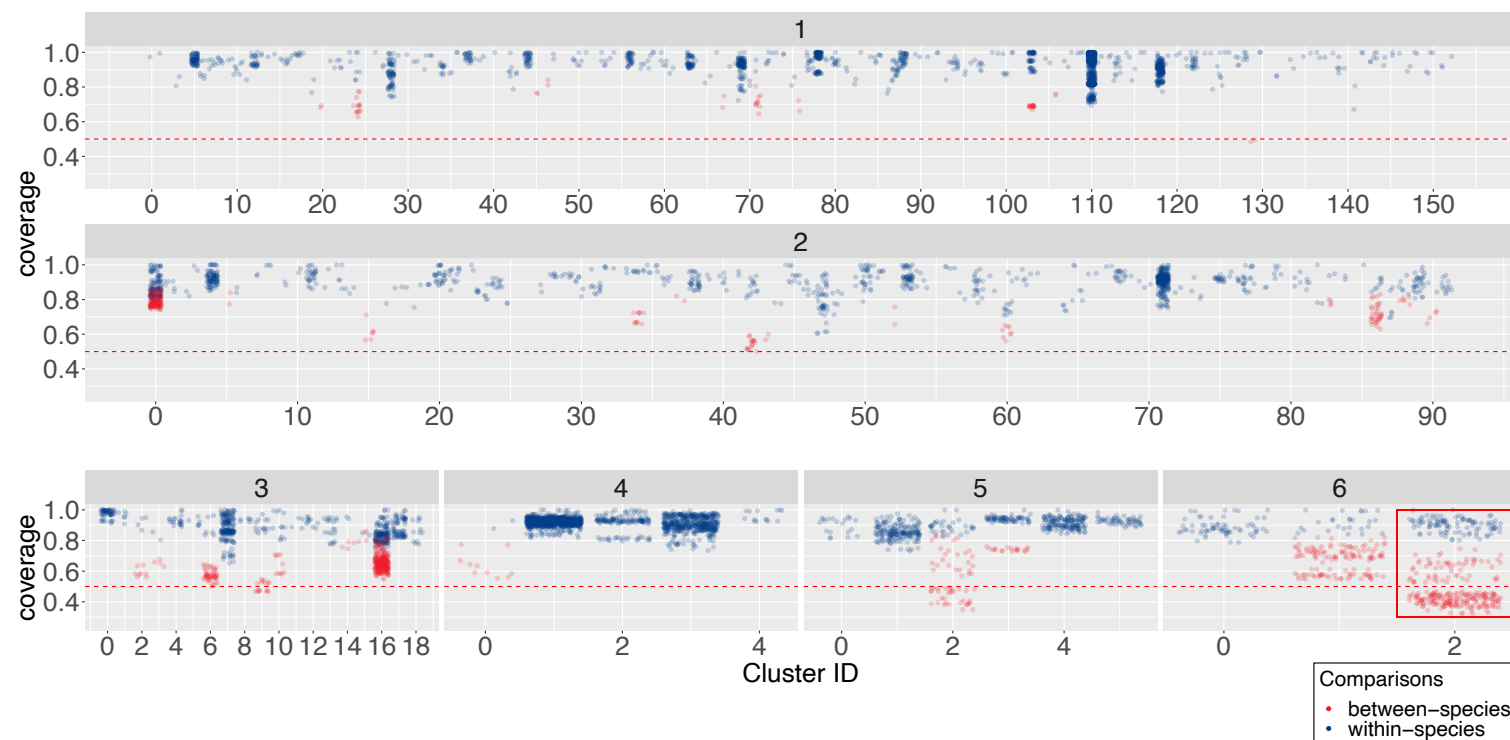


Figure 2.26: Scatterplots showing genome coverage for pairwise ANI comparisons for genomes sharing identical full-length and ambiguity base-free 16S sequences. The number of unique species names assigned per cluster is displayed at the top of each subgrouping, and the red horizontal line at 50% indicates the whole-genome genus threshold. Within-species pairwise comparisons ($\geq 95\%$ genome identity) are shown in blue, and between-species comparisons ($< 95\%$ genome identity) are shown in red. Cluster uniting genomes with the lowest genome coverage is outlined in the red box. (note that not all clusters below the 50% threshold are highlighted; this box is only used to indicate the cluster with the lowest genome coverage).

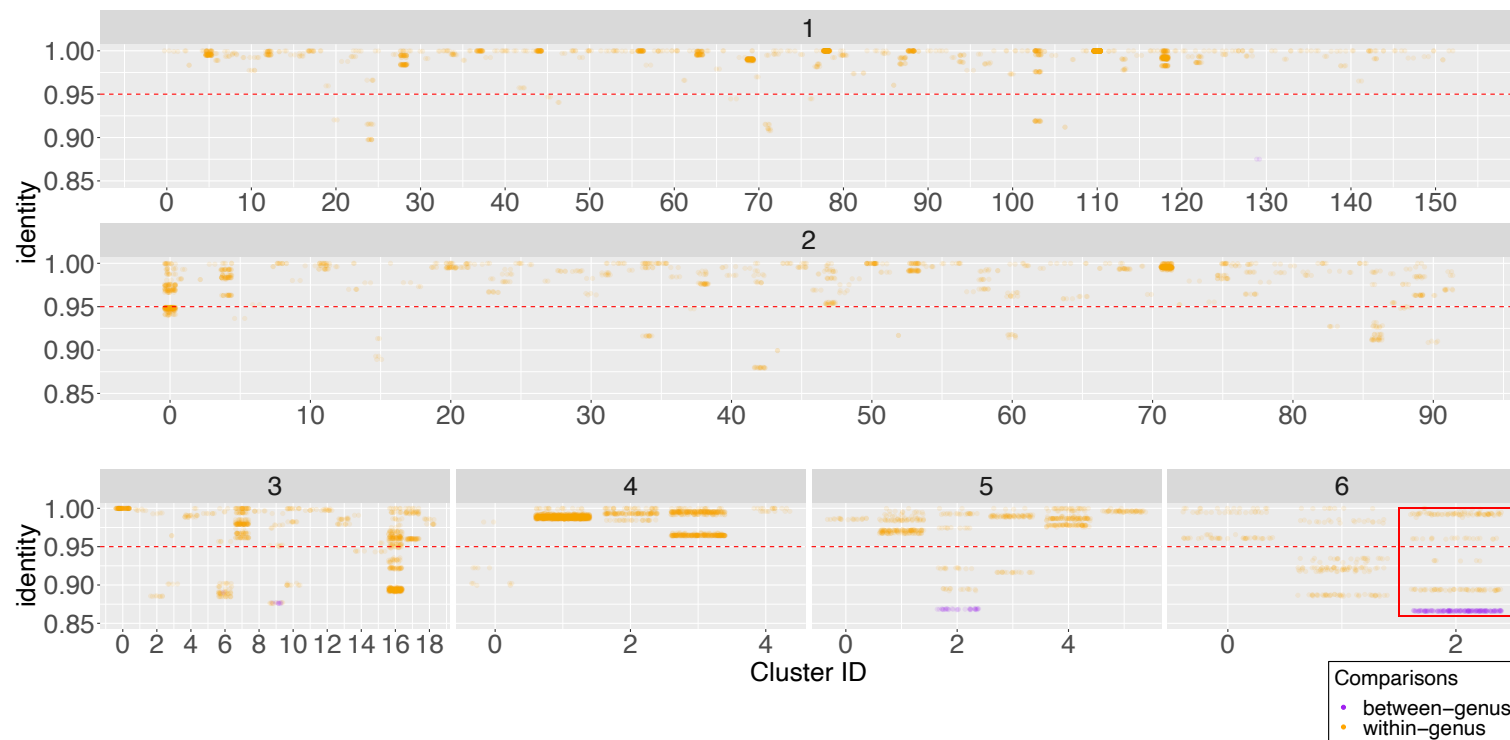


Figure 2.27: Scatterplots showing genome identity for pairwise ANI comparisons for genomes sharing identical full-length and ambiguity base free 16S sequences. The number of unique species names assigned per cluster are displayed at the top of each plot, and the red horizontal line at 95% indicates the whole-genome species threshold. Within-genus comparison ($\geq 50\%$ genome coverage) are shown in orange, and between-genus comparisons ($< 50\%$ genome coverage) are shown in purple. Cluster uniting genomes with the lowest genome identity is outlined in the red box.

Updated Multilocus Sequence Typing (MLST) scheme for *Streptomyces* reveals a complex taxonomic structure

3.1 Introduction

3.1.1 Motivation

In Chapter 2, I estimated the most comprehensive 16S phylogeny to date for members of the genus *Streptomyces* and compared intragenomic heterogeneity of 16S sequences from *Streptomyces* genomes using ANI. This comparison revealed that a single-gene tree based on 16S sequences is unlikely to provide a robust phylogeny and that there is not a one-to-one mapping between 16S sequences and species.

In this chapter, I consider MLST analysis in the context of publicly available *Streptomyces* genomes. MLST, as discussed in Chapter 1, remains actively used for exploring evolutionary relationships between taxa. The current canonical *Streptomyces* MLST scheme provided by pubMLST comprises six markers and 237 sequence types (STs) (Jolley et al., 2018). However, only three new STs have been reported since 2016, and the exact relationship between MLST and whole-genome-based taxonomy remains unexplored. Therefore, the central aim of this chapter is to update the canonical

Streptomyces scheme using all publicly available genomes and investigate taxonomic relationships among all *Streptomyces* genomic sequences using both MLST and whole-genome distance methods.

3.1.2 Aims and Objectives

The objectives of this chapter are as follows:

1. Since 2016 only three new STs have been added to the pubMLST *Streptomyces* database (Jolley et al., 2018), despite a substantial increase in the number of publicly available genomes in NCBI (Figure 1.12). Given this limited expansion, I will update the canonical pubMLST *Streptomyces* scheme by incorporating all publicly available *Streptomyces* genomic sequences at the time of this study.
2. Previous analyses based on 16S sequences suggest that the MLST scheme might produce a disjoint overall graph of ST profiles (Figure 3.20), potentially indicating useful categorisations for *Streptomyces*. To investigate this further, I will reconstruct a minimum spanning tree (MST) to investigate the relationships between ST profiles and their corresponding genomes, and then assess the taxonomic composition of the connected subgraphs using ANIm methods.
3. After identifying the composition of the subgraphs in the MST, I will examine the distribution of currently assigned species names in NCBI to see how they align with existing or proposed taxonomic categories. Additionally, I will conduct ANIm analysis on all genomes assigned the same name in NCBI to assess the accuracy of their classifications and identify any potential misclassification in *Streptomyces* genomes.

4. To explore whether sequencing additional *Streptomyces* genomes in the future might connect all STs, I will investigate whether the observed divisions in the MLST within the MST are due to insufficient sampling.
5. Since MLST profiles are categorical and do not explicitly consider phylogenetic information there may be a lack of congruence between these methods. Therefore, I will reconstruct a comprehensive MLSA phylogeny for members of the genus *Streptomyces* to investigate whether the MLST divisions form monophyletic clades on the MLSA tree.
6. Given that the allele choices for MLST were made historically, before routine availability of whole-genome sequences, their suitability for accurately representing *Streptomyces* evolutionary relationships may be in question. To address this, I will assess how the exclusion of each marker gene affects the representation of STs by evaluating changes in ST assignments for each genome assembly.
7. Since genera like *Kitasatospora* have been reclassified outwith *Streptomyces* genus multiple times (see section 1.7), I will investigate the representation of *Streptomyces* STs in sister genera to better understand their taxonomic relationships and potential overlaps.

3.2 Methodology

3.2.1 Data retrieval and availability

This chapter's code, all raw and supporting data can be accessed by the public through GitHub (https://github.com/kiepczi/Kiepas-et.al.2023_MLST).

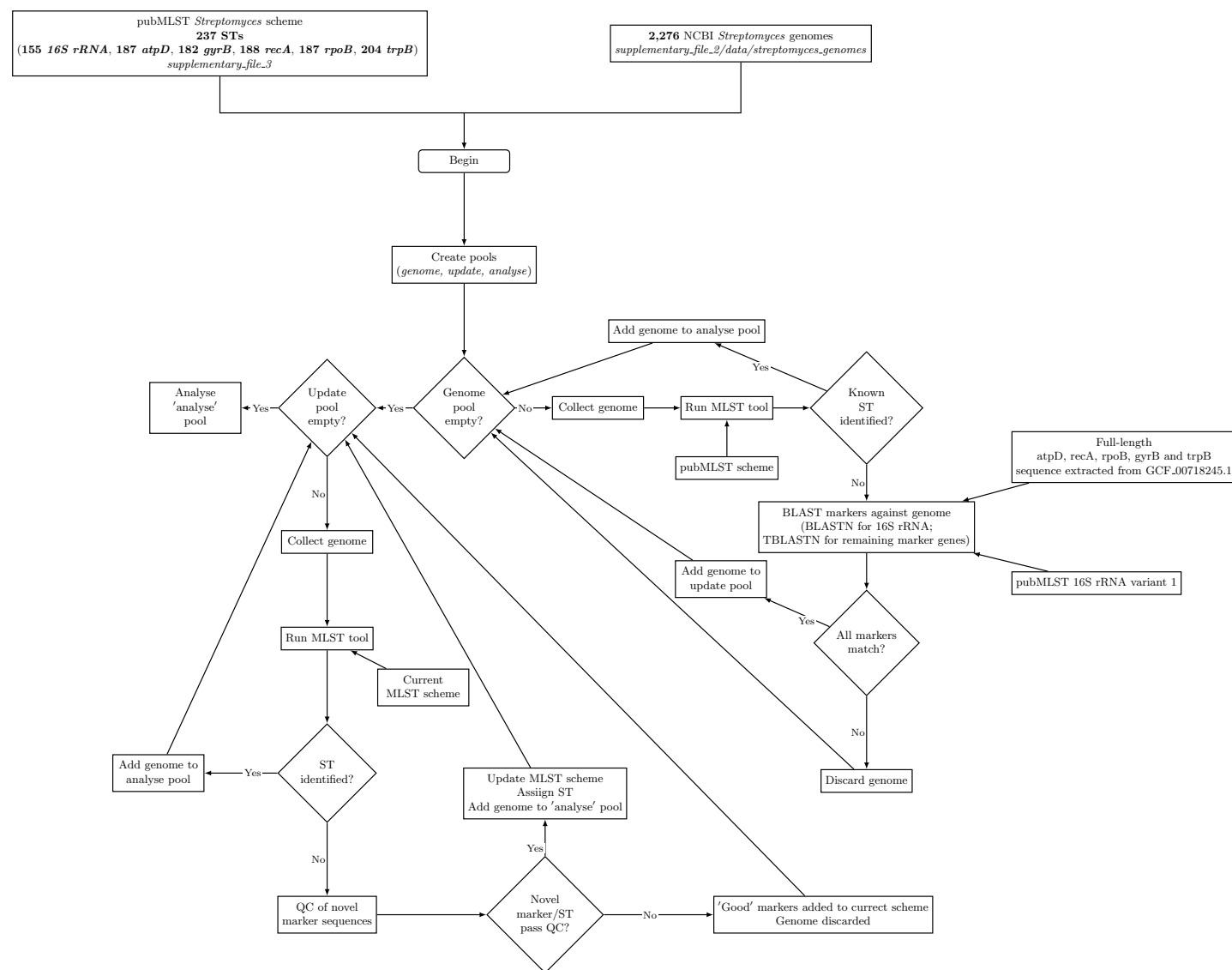
The canonical *Streptomyces* MLST scheme (`streptomyces_pubMLST.txt`; Supplementary File 3) with 237 STs, together with 155 16S rRNA, 187 *atpD*, 182 *gyrB*, 188 *recA*, 187 *rpoB* and 204 *trpB* allele sequences (`*.tfa` files; Supplementary File 3) was manually downloaded from pubMLST (Jolley et al., 2018; https://pubmlst.org/bigsdb?db=pubmlst_streptomyces_seqdef; accessed 24th August 2023).

All 2,276 publicly available *Streptomyces* genome sequences were downloaded from NCBI (Sayers et al., 2021) on July 8th, 2021 (as described in section 2.2.9). The flowchart in Figure 3.1 outlines the sequence of steps involved in updating and filtering of the MLST *Streptomyces* scheme.

3.2.2 Filtration of *Streptomyces* genomes

Genomes were included in the analysis if they possessed a known ST, or all six marker sequences were present. To identify genomes with known STs, I ran MLST v2.22.0 <https://github.com/tseemann/mlst> on all 2,276 *Streptomyces* genomes using the canonical *Streptomyces* scheme (`streptomyces_pubMLST.txt`; Supplementary File 4) and `--mincov 80` parameter using the `01_run_MLST.sh` bash script provided in Supplementary File 4.

For the remaining genomes, I used BLAST to identify genomes that met the criteria of having at least 80% identity and 80% coverage to a known allele for all six marker sequences; using `blastn v2.6.0+` for 16S rRNA sequences and `tblastn` (Camacho et al., 2009) `v2.6.0+` for the 5 remaining alleles (`03_blastn_and_tblastn.sh`; Supplementary File 5). The canonical pubMLST scheme characterises *Streptomyces* species based on full length 16S rRNA sequences, and internal fragments of the remaining marker genes. Therefore, to filter genomes, I used 16S rRNA variant 1 from the pubMLST database



(16S.fasta; Supplementary File 5) as a query for blastn analysis, and I extracted the remaining full length five marker sequences from GCF_00718245.1 (identified to belong to ST2; atpD.fasta, gyrB.fasta, recA.fasta, rpoB.fasta, trpB.fasta; Supplementary File 5) using 02_get_marker_seq_for_blast.py (Supplementary File 5), and used these as the tblastn query sequences for the other five markers. For the subject sequences, I created local BLAST databases comprising genomes that were not assigned a ST profile in the MLST analysis above, using the 01_get_genomes_for_blast.py Python script (Supplementary File 5). This reduced the total number of *Streptomyces* genomes under consideration to 1,938.

3.2.3 Identification of novel allele sequences and ST assignment

I used the 1,938 *Streptomyces* genomes (all_genomes_for_mlst_extention; Supplementary File 5) to identify novel alleles and update the canonical *Streptomyces* scheme as follows:

1. The canonical pubMLST scheme was taken as the initial (current) *Streptomyces* MLST scheme.
2. I used MLST v2.22.0 (<https://github.com/tseemann/mlst>) and the current *Streptomyces* scheme to predict new allele sequences in the currently unassigned (initially 1,938 genomes) genome set, using the same script and parameters as described in section 3.2.2.
3. All reported novel sequences were checked for the presence of ambiguous bases or partial marker sequences. I discarded novel sequences that contained any ambiguity bases, and I also checked all reported novel sequences against the

sequences already present in the scheme, and against all other novel reported sequences (`01_novel_seq_filtration.py`; Supplementary File 6). No novel sequences were found to be partial sequences of known or other novel marker sequences,.

4. I assigned unique allele numbers to all novel sequences (`01_novel_seq_filtration.py`; Supplemenetary File 6) and added these to the current scheme.
5. Steps 2-4 were repeated until no further novel allele sequences were found (six rounds were required). The MLST tool results and novel sequences identified after each run are provided in `supplementary_file_4/output/extension_scheme_round.*`, while the subsequent updated schemes are provided in `supplementary_file_6/output/schemes`.
6. I combined the allele numbers into profiles, and updated the current canonical pubMLST scheme by assigning a unique ST to each profile (`02_ST_assignment.py`; Supplementary File 6), which were submitted to the PubMLST database if:
 - (a) the profile had not been previously reported in the pubMLST database
 - (b) the genome did not contain multiple distinct copies of any marker genes

3.2.4 Scheme refinement

I checked the assembly status of all 2,276 downloaded genomes against the assembly report (downloaded on the 6th September 2023; `assembly_summary_refseq_historical.txt`; Supplementary File 7) from

https://ftp.ncbi.nlm.nih.gov/genomes/refseq/assembly_summary_refseq_historical.txt (01_check_assembly_status.py; Supplementary File 7). I removed suppressed and replaced genomes from the MLST scheme. Replacement genomes (streptomyces_replaced_genomes; Supplementary File 2) were downloaded manually from NCBI on the 1st of May 2023, and analysed as described in Methodology Section 3.2.3.

3.2.5 Visualisation of MLST scheme: Minimum Spanning Tree

I calculated a minimum spanning tree (MST; using Kruskal's algorithm) for the final ST profile set (methodology section 3.2.4) using NetworkX v2.6.3 (Hagberg et al., 2008) and a pairwise Hamming distance calculated as the count of differing allele numbers between two profiles (02_calculate_and_investigate_MST.ipynb; Supplementary File 16). Edges with a Hamming distance of six correspond to pairs of STs with no allele in common, and such edges were removed from the tree. I fixed the positioning of the nodes in each tree using Cytoscape (Shannon et al., 2003) v3.9.0 with Prefused Force-Directed layout, and visualised the MSTs using plotly v5.6.0 (<https://plotly.com/python/>; 01_visualise_MST_with_NetworkX.ipynb provided in Supplementary File 20).

3.2.6 Genome Quality Assessment

All 874 genomes with an identified ST were assessed for completeness and contamination using checkM v1.2.2 (Parks et al., 2015) to identify any poor-quality genomes that may have been included and could have influenced the results (01_genome_supplementary_data.py provided in Supplementary File 34.).

3.2.7 Influence of genome sampling on the connectivity of MST

If every naturally-occurring *Streptomyces* isolate had been sequenced, the MST would provide an accurate representation of the relationships among their genomes. However, if the currently sequenced *Streptomyces* genomes represent only a small fraction of the *Streptomyces* total genomic diversity, the MST might show many disconnected subgraphs. These subgraphs would not reflect true evolutionary separations but rather gaps in sequencing data (e.g., sampling bias), where sequencing more *Streptomyces* genomes in the future could potentially bridge these divides. To assess whether sampling depth is impacting the current MST, I analyse the effect of removing a subset of the sequenced genomes. If the graph becomes more disconnected after the removal, it would suggest that the current sequencing effort is insufficient to fully capture the genomic relationships within *Streptomyces*. Conversely, if the graph's structure remains largely unchanged, we might assume that the structure is a fair representation of the structure that would be obtained if it was possible to sequence all naturally-occurring *Streptomyces* spp. and tentatively infer that the existing MST is a reasonable approximation of the true genomic relationships, even as additional genomes are sequenced.

To assess the effect of sampling representation of currently sequenced *Streptomyces* isolates on the connectivity of the MST, I repeated the MST construction with random subsamples of the sequenced genomes. The sampled proportions ran from 10% to 90% of the available genomes (each repeated 100 times) progressing in 10% intervals. In each case, I calculated and normalised the distribution of sizes of connected components to the number of genomes being considered (02_calculate_and_investigate_MST.ipynb;

Supplementary File 16), to determine whether alternative sub-samplings of the same data result in a different set of MST topologies.

Separately, to investigate the possibility that observed changes in MST topologies can theoretically arise through sampling bias, I generated an artificial MLST scheme consisting of 5000 STs and six markers. This scheme was generated through a random process, where each unique sequence type was assigned an allelic profile composed of six marker genes. Each marker gene was randomly assigned a number between 1 and 2000. To ensure some level of connectivity between the profiles, the scheme was designed so that each allelic profile shared at least one marker in common with at least 100 other profiles. This approach prevented the creation of a completely disconnected set of 5000 STs, thereby introducing a degree of relatedness among them (`02_calculate_and_investigate_MST.ipynb`; Supplementary File 16). Subsequently, I analysed the changes in MST topologies by calculating the number of disjoint components and distribution of component sizes through random subsampling of the artificial scheme, following the same methodology as described above, to act as a form of control analysis under the assumption of a random distribution of sequence types.

3.2.8 Empirical non-parametric network test

The distribution of node degrees (i.e., the number of connections each node has in the network) was compared across 237 pubMLST STs, 568 novel STs, 150 non-GenBank-represented pubMLST STs, and 87 GenBank-represented pubMLST STs using an empirical non-parametric test (`03_empirical_network_test.ipynb`; Supplementary File 16). I randomly resampled nodes (STs) from the original MST graph one million

times, where the selected nodes had the same degree of connections as in the original graph. For the pubMLST analysis, I sampled 237 nodes/STs, and for the GenBank analysis, I sampled 150 nodes/STs in each iteration. I then calculated how often I observed the same number or more high-degree nodes (≥ 7 for pubMLST, and ≥ 6 for GenBank-represented STs). This provides an estimate of the probability of having a network with a similar or more extreme distribution of highly-connected nodes by chance alone.

3.2.9 ANI analysis

Three separate Average Nucleotide Identity (ANI) analyses using pyANI v0.3 (Pritchard et al., 2015) were performed to determine taxonomic boundaries for genomes: (i) assigned the same species name in the NCBI taxonomy, (ii) sharing identical STs, and (iii) within the same connected subgraph of STs in MST (Supplementary File 17). Similar to the boundaries adopted in chapter 2.2.9, I applied the 95% identity threshold to define species boundaries and the 50% genome coverage threshold to estimate genus boundaries. As discussed in Section 1.3.3, a proposed ANI threshold suggests that isolates with $\geq 95\%$ genome identity likely belong to the same species. However, genus boundaries remain debated, with some researchers arguing that genomes sharing $< 50\%$ of their genetic material may be more similar to unrelated lineages and should be classified separately (Pritchard et al., 2015). While these thresholds are approximations, they align with commonly used heuristics in microbial classification (Pritchard et al., 2015; Richter & Rosselló-Móra, 2009).

For these analyses, a 50% coverage threshold was applied to estimate genus boundaries, and a 95% identity threshold was used to estimate species boundaries.

In cases where genome boundaries were ambiguous—such as when genome A shared 50% coverage with genomes B and C, but genomes B and C only shared 49% of their genome by alignment length (as discussed in Chapter 2, Section 1.5.4)—I used NetworkX v2.6.3 (Hagberg et al., 2008) to resolve these issues. I approached this by modeling genomes as nodes and representing pairwise comparisons as edges, with the lowest genome coverage used for genus assignment and the average genome identity used for species assignment. Edges with the lowest genome coverage (for genus assignment) or genome identity (for species assignment) were iteratively removed until the cliques were clearly defined (`08_assign_genus_species_from_pyANI_analysis.ipynb`; **Supplementary File 17**).

3.2.10 MLSA Phylogenetic reconstruction from MLST markers

A phylogenetic reconstruction was made using the concatenated full-length sequences of all 6 marker sequences for each genome. To extract full-length sequences, I carried out tblastn v2.6.0+ (Camacho et al., 2009) analysis (**Supplementary File 18**) using reported alleles as query sequences and their corresponding genomes as subjects to determine their location on the genome. For each genome, I obtained 16S nucleotide sequences from the revised MLST scheme (`01_get_16S_seq.py`; **Supplementary File 10**) and for the remaining five marker genes the protein and nucleotide sequences were extracted from GenBank files using the determined location on the genome (`02_get_prot_and_nt.py`; **Supplementary File 19**). I found that 19 markers (four *gyrB*, six *recA*, four *rpoB*, and seven *trpB*) were labeled as pseudogenes in GenBank, meaning nucleotide sequences were available, but no corresponding protein sequences existed. Therefore, I aligned the protein sequences for alleles reported from non-pseudogenes using

MAFFT v7.520 (Kato & Standley, 2013) (`03_align_seq_mafft_no_pseudo_genes.sh`; Supplementary File 19) and the nucleotide sequences were back-threaded onto these alignments using T-coffee v12.00.7fb08c2 (Notredame et al., 2000) (`04_backthread.sh`; Supplementary File 19). I then aligned the nucleotide sequences reported from pseudogenes with the existing backthreaded alignments using MAFFT v7.520 (`05_add_nucleotide_seq_for_pseudo_genes.sh`). The alignments were then concatenated (`06_concatenate_alignments.py`; Supplementary File 19), trimmed using trimal v1.4.rev15 (Capella-Gutiérrez et al., 2009) with `-automated1` parameter (`07_trim_alignement.sh`; Supplementary File 19) and the best evolutionary models for each partition were determined using ModelTest-NG v0.1.7 (Darriba et al., 2019) (`09_evolutionary_model_test.sh`; Supplementary File 19). Maximum Likelihood (ML) phylogeny with 100 Transfer Bootstrap Expectation (TBE) values and `--seed 1655486274` parameter were reconstructed with RAxML-NG v1.1 (Kozlov et al., 2019) (`10_build_tree.sh`; Supplementary File 19) on the ARCHIE-West computing cluster with Intel XI(R) Silver 4216 CPU 2.10Hz, 32 cores and 187 GB RAM. Finally, the congruence between the inferred phylogenetic tree and MST was checked using `ete3` (Huerta-Cepas et al., 2016) python module (`11_check_congruence.ipynb`; Supplementary File 19), and visualized in R (`12_tree_vizualisation.R`; Supplementary File 19).

3.2.11 Sensitivity test

To evaluate the contribution of each marker gene to the classification of genomes into distinct STs, I excluded each marker singly in turn and assessed how its absence affected the classification. To evaluate the contribution of each marker gene to the classification of genomes into distinct STs, I excluded each marker singly in turn and assessed how

its absence affected the classification. For example, if two genomes share the same five markers—16S, *atpD*, *rpoB*, *gyrB*, and *recA*—but differ only by the *trpB* marker, removing *trpB* would result in both genomes being classified under the same ST. I examined how the removal of each marker gene impacted genome groupings, by identifying how often two genomes are classified differently under a six-marker scheme but as the same under a five-marker scheme (`01_sensitivity_test.py`; Supplementary File 9).

3.2.12 Representation of *Streptomyces* in sister genera

To determine whether any pubMLST STs applied to any genomes from outwith the *Streptomyces* genus, I downloaded all publicly available 66 *Kitasatospora*, 2 *Streptoalloteichus* and 15 *Streptacidiphilus* genomes from NCBI on September 7th, 2023 and ran MLST v2.22.0 using the revised updated scheme (methodology section 3.2.4; `revised_scheme` provided in Supplementary File 8) using the same methodology and parameters as used above. The genome accessions for *Kitasatospora*, *Streptoalloteichus* and *Streptacidiphilus* genomes are provided in `kitasatospora_genomes.txt`, `streptoalloteichus_genomes.txt` and `streptacidiphilus_genomes.txt` (Supplementary File 2), respectively.

3.3 Results

3.3.1 Updated scheme

I updated the current canonical pubMLST scheme using a set of 2,276 available *Streptomyces* genomes (methodology sections 3.2.2-3.2.4). I began by identifying a subset of 1,938 genomes containing all six marker sequences required for the analysis (methodology section 3.2.2). I then excluded 156 genomes from the 1,938 that had

been suppressed in NCBI, leaving a total of 1,782 genomes for analysis to update the scheme (methodology section 3.2.4). From these 1,782 genomes, 873 contained all marker sequences in a single variant copy, making it possible to assign 568 novel STs. The resulting updated MLST scheme for *Streptomyces* consists of 805 STs, and uses 787 (632 novel) 16S rRNA, 938 (751 novel) *atpD*, 716 (534 novel) *gyrB*, 1019 (831 novel) *recA*, 999 (812 novel) *rpoB* and 936 (732 novel) *trpB* distinct allele sequences. The novel STs and allele sequences were submitted to PubMLST on October 25th, 2023 (Figure 3.2), and are now publicly accessible through PubMLST <https://pubmlst.org/organisms/streptomyces-spp> and in the supplementary data (`revised_scheme`; Supplementary File 8). The examples of novel STs and genomes that corresponded to the extension of the canonical *Streptomyces* scheme are shown in Table 3.1, while the overall distribution of NCBI assigned species with novel and existing, or combination of STs is shown in Figure 3.3.

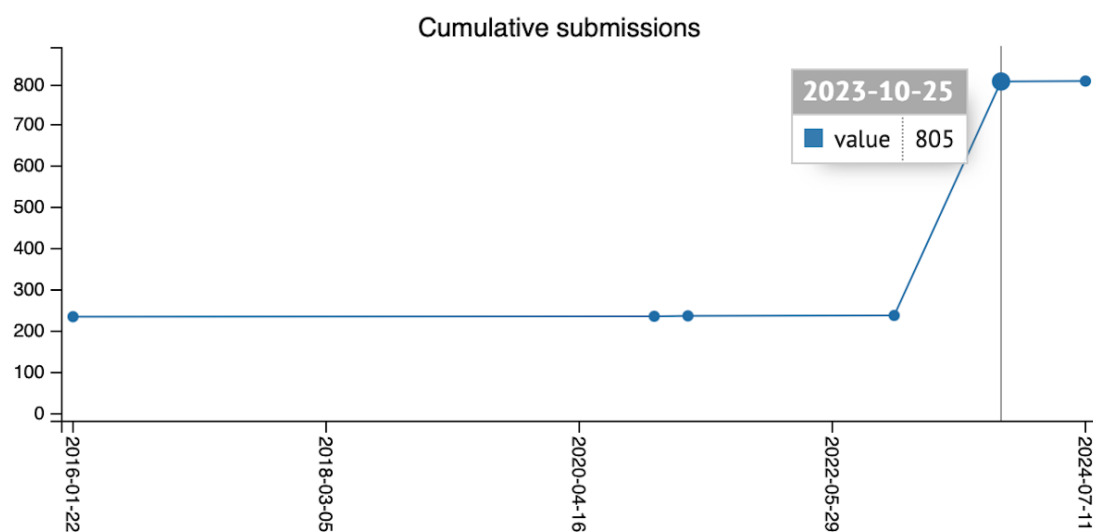


Figure 3.2: History of submission of STs for *Streptomyces* scheme in pubMLST. Spike in submissions on the 25th of October corresponds to the contribution made by the work presented in this chapter. Figure taken from https://pubmlst.org/bigssdb?db=pubmlst_streptomyces_seqdef&page=schemeInfo&scheme_id=1.

Table 3.1: Examples of novel STs with corresponding genome accessions and assigned NCBI names.

NCBI name	Accession	ST	16S	atpD	gyrB	recA	rpoB	trpB
<i>S. griseus</i>	GCF_000716515.1	269	46	58	496	214	689	605
<i>S. rimosus</i>	GCF_000718755.1	287	243	298	250	312	304	285
<i>S. scabiei</i>	GCF_001550245.1	327	248	165	161	317	309	290
<i>S. lincolensis</i>	GCF_003344445.1	402	291	357	286	378	370	806
<i>S. lydicus</i>	GCF_004125265.1	554	42	805	573	867	821	710

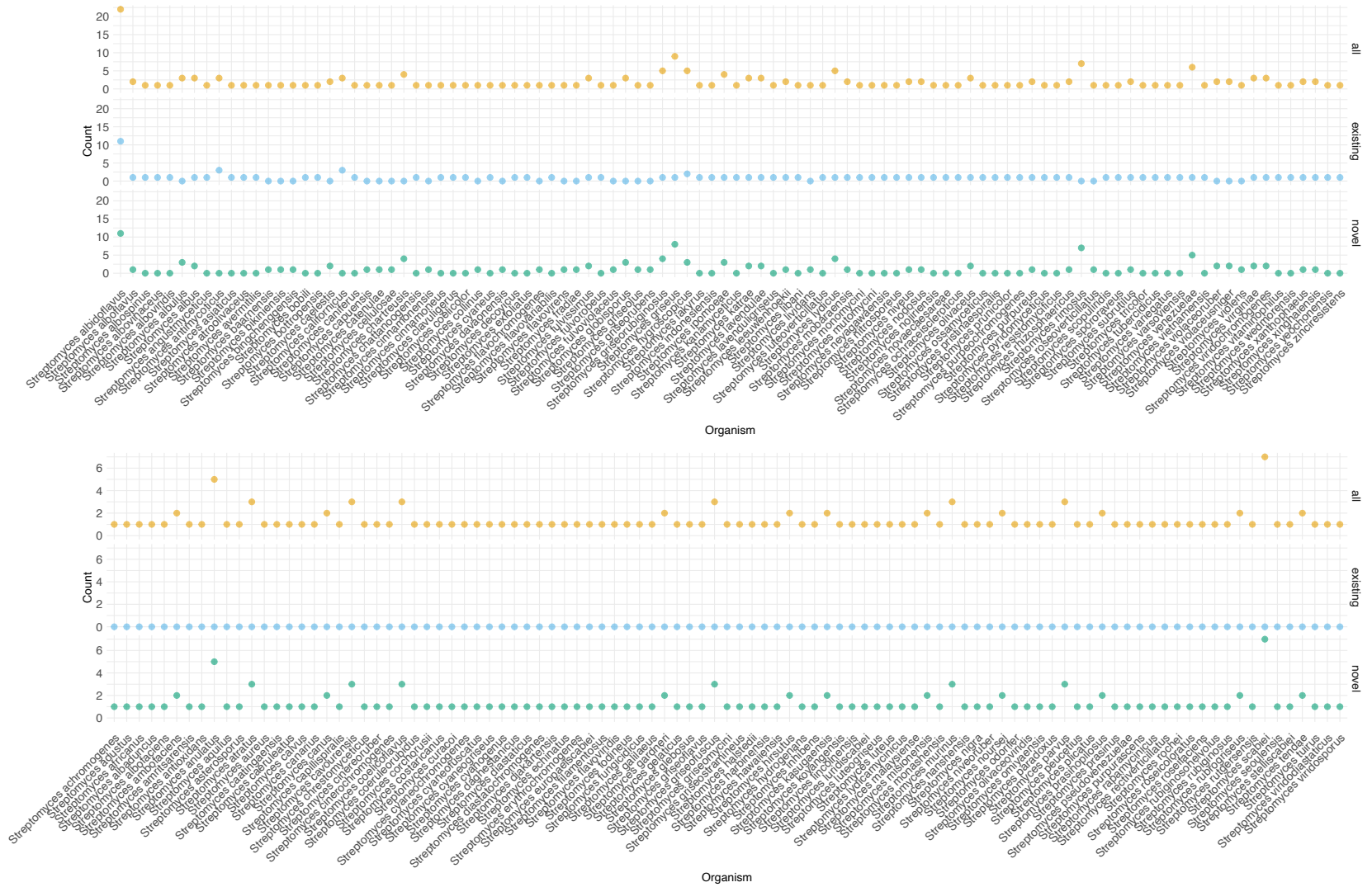


Figure 3.3: Distribution of NCBI-assigned species across STs. The top panel represents the total count of STs per species, including both novel and existing STs. The middle panel shows the distribution for existing STs, while the bottom panel highlights the distribution for novel STs. The graph is split into two parts to enhance visibility and clarity.

It might seem counterintuitive that there are more allele sequences than STs. However, the extension of the scheme was repeated six times, during which any new allele sequences that were identified were added to the scheme, even if the genomes did not have all six marker alleles present. This iterative process resulted in the identification of more novel alleles than sequence types. While 925 genomes had complete sets of all six alleles, others with missing alleles were still included in the analysis (as detailed in Table 3.2). Additionally, some genomes contained multiple non-identical copies of marker genes, contributing to the overall count; specifically, 77 genomes had more than one 16S sequence variant, and one genome (GCF_001550235.1) had two non-identical copies of *recA*. Examples of novel STs and genomes that corresponded to the extension of the canonical *Streptomyces* scheme are shown in Table 3.1, while the overall distribution of NCBI-assigned species with novel and existing STs, or combination of STs is shown in Figure 3.3.

Table 3.2: An overview of the distribution of missing allele copies across the 1,782 analysed *Streptomyces* genomes.

Number of missing markers	Number of genomes
0	925
1	366
2	230
3	73
4	72
5	86
6	30

I updated the current canonical scheme using publicly available genomes that had been assembled to different levels of quality (i.e. contig, scaffold, complete and chromosome). Of the 873 genomes identified with either novel or existing STs, 466 were assembled to contig level, 251 to scaffold level, 141 to complete level and 15 to chromosome level. A total of 462 genomes collectively assembled to scaffold and contig level were crucial in the assignment of novel STs, and no genomes assembled to chromosomal or complete level contained novel STs. Using incomplete assemblies to update the scheme has limitations. For example, assemblies at the scaffold or contig level may miss alleles, giving the false impression that an allele is absent, which prevents ST assignment and fails to capture the full diversity of *Streptomyces* sequences. Additionally, incomplete assemblies may overlook multiple non-identical allele copies, leading to inaccurate ST assignments that do not exist in nature. The `Genome_ST_info.csv` (Supplementary File 8) file provides extended information about each genome's assembly status level. Additionally, out of the 873 genomes identified with either novel or existing STs, 204 belonged to type strains. Among these, 86 type strain assemblies were associated with existing STs, while 118 contributed to the assignment of novel STs. A complete list detailing which genomes are type strains and which are not can be found in `Genome_ST_info.csv` (Supplementary File 8).

During allele identification (methodology section 3.2.3), six novel alleles were identified that contained ambiguity bases. These would prevent determination of the true sequence variation, and the assignment of a unique allele number. As a result, these sequences were not included in the amended scheme. Additionally, one genome

(GCF_001550235.1) had multiple non-identical copies of *recA* and 16S rRNA, and 76 other genomes were found to have multiple non-identical copies of 16S rRNA sequences.

Among the genomes with existing or novel STs, 259 genomes had multiple identical copies of at least one marker gene, this being most common for the 16S rRNA marker, in 255 genomes. The count of allele copies per genome are summarised in Table 3.3, while the extended information about each genomes total number of all and unique marker copies is provided in `Genomes_allele_count.csv` (Supplementary File 8).

Table 3.3: The count of genomes with single and multiple identical copies for each marker.

Marker	Number of Genomes with Multiple Identical Copies	Number of genomes with a single copy
<i>16S</i>	255	618
<i>rpoB</i>	3	870
<i>atpD</i>	1	872
<i>gyrB</i>	1	872
<i>recA</i>	1	872
<i>trpB</i>	1	872

3.3.2 Graph based analysis of STs

I represented the updated scheme as a minimum spanning tree, where nodes represent STs and edges indicate transitions between STs that share at least one common allele. The minimum spanning tree resolved the 805 STs into 278 connected components (Figure 3.4; methodology section 3.2.5). The largest component comprises 80 STs (9.94% of all STs; Figure 3.4 connected component 1), and 178 STs are singletons (63.03% of all disjoint graphs).

STs connect with varying numbers of other STs, with node degree from zero to 13 (Figure 3.6 and Figure 3.7). The number of STs with higher degree/levels of connectivity is relatively low. Nodes (STs) with high degree in a network are known as "hubs" and may be important members of the network. These hubs may play a critical role in maintaining the structural integrity of the network by linking otherwise unrelated STs. For example, if ST 1 shares a common allele with six other STs, but these six STs do not share any alleles among themselves, ST 1 effectively acts as a bridge, creating connections between otherwise disconnected STs. Additionally, nodes with high degree may also exhibit high betweenness centrality, meaning they frequently lie on the shortest paths between other STs. The distribution of connected components sizes is shown in Figure 3.5, while the distribution of connections (degrees) among STs is shown in Figure 3.6.

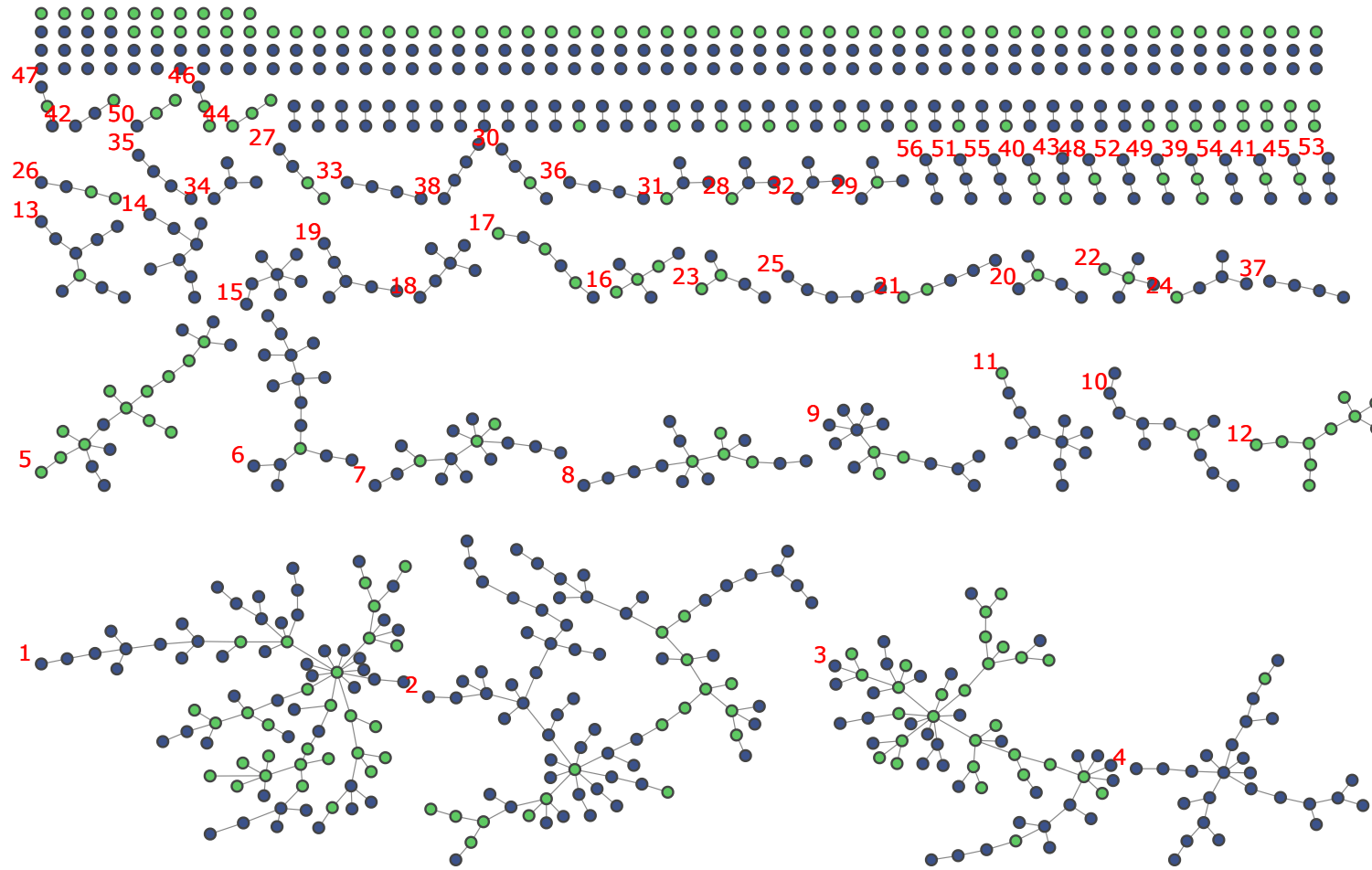


Figure 3.4: MST with 805 STs and 282 connected components describing all sequences *Streptomyces* genomes, and all STs from the pubMLST database. Each node represents a unique ST, and each edge corresponds to that pair of STs sharing at least one allele in common. The distribution of novel STs (blue) and pubMLST STs (green) on the MST. Assigned connected component IDs are displayed in red. No IDs are shown for components consisting of two or fewer STs to improve visualisation. Connected component 1 is the largest component, uniting 80 STs, while the largest connected component consisting only of pubMLST STs is connected component 12, and the largest connected component consisting only of novel STs is connected component 14.

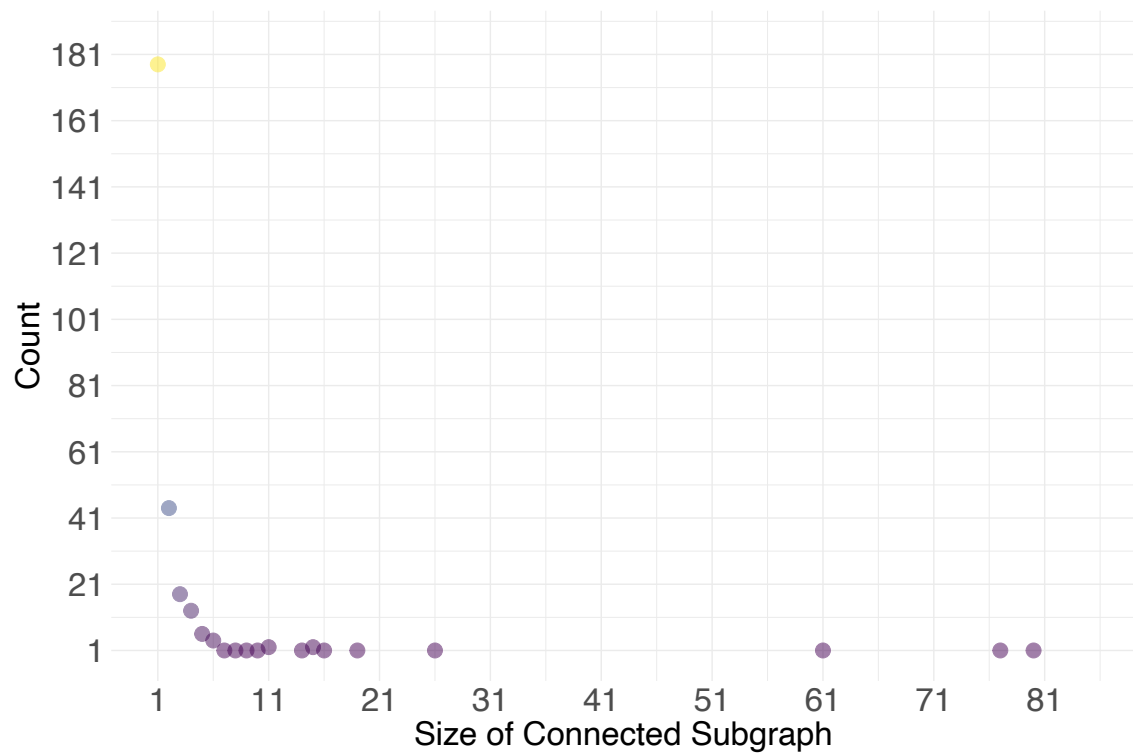


Figure 3.5: Distribution of connected component sizes showing that smaller connected components are more frequent, while larger connected components are less common.

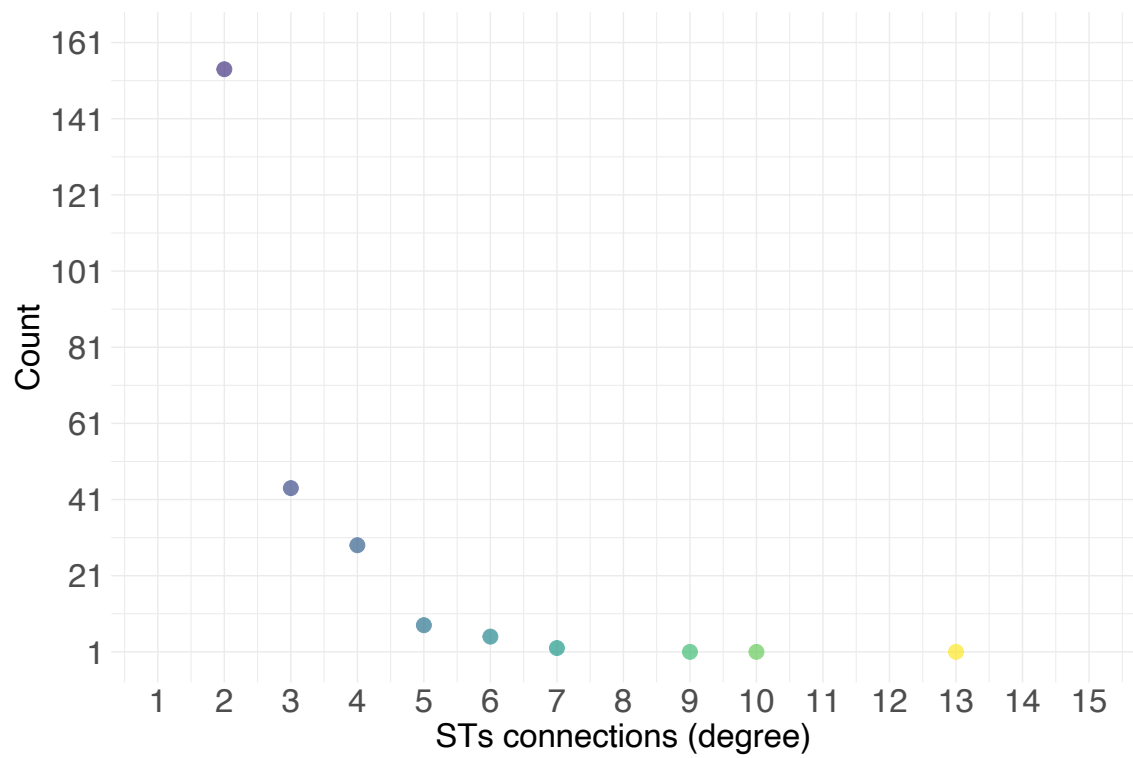


Figure 3.6: Distribution of connections (degrees) of STs in the MST representation of MLST scheme for *Streptomyces*.

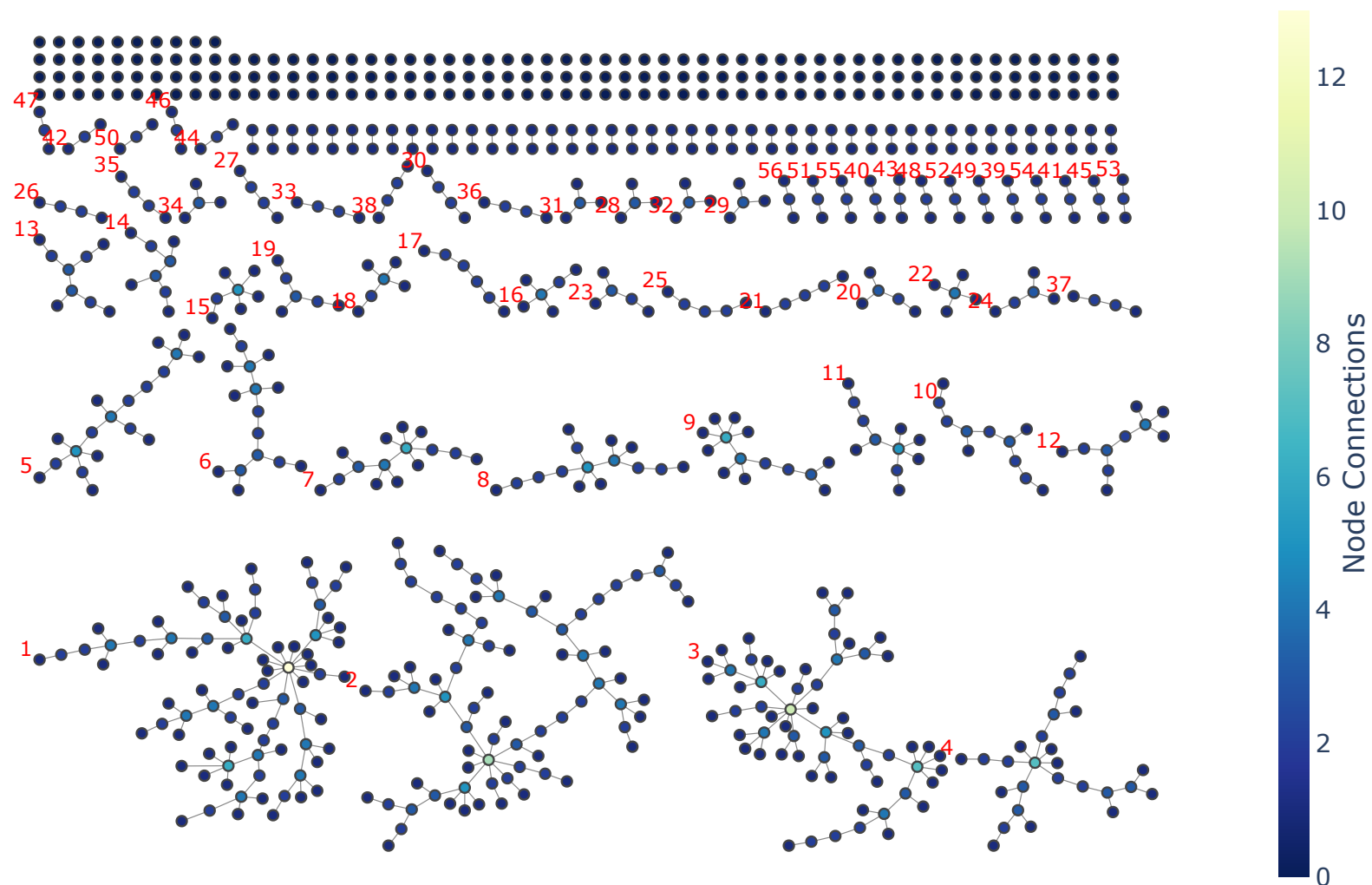


Figure 3.7: MST of the updated pubMLST *Streptomyces* scheme showing number of ST connections (node degree). Nodes are colored based on their degree of connectivity, with fewer connections in blue and more connections in yellow.

There are 120 connected components containing at least one pubMLST ST, 63 of which are not singletons. There are 69 components exclusively consisting of pubMLST STs, and only six of which are not singletons; the largest such subgraph unites 10 STs (Figure 3.4 connected component 12). A total of 209 connected components contain at least one novel ST, and 115 of these components are not singletons. There are 158 connected components exclusively consisting of novel STs, of which 43 are not singletons and the largest unites a total of 8 STs (Figure 3.4 connected component 14). These findings suggest three key patterns: (i) the 69 pubMLST-only subgraphs remain isolated, with no new connections formed; (ii) 51 mixed pubMLST/novel subgraphs have expanded out from known STs, where genomic data bridge gaps between established STs and new sequence variations; and (iii) 158 subgraphs, composed entirely of novel STs, represent distinct "islands" of new sequence variation, detached from the existing pubMLST framework.

150 of the 237 canonical pubMLST STs are not present among the sequenced genomes in NCBI; such STs account for 63.29% of all pubMLST STs and 18.63% of all STs in the updated *Streptomyces* scheme. The 150 STs not found in Genbank are distributed across a total of 71 connected components, of which 36 are not singletons (Figure 3.8). Additionally, there are 40 components exclusively consisting of non-GenBank represented STs, with four of them being non-singletons, and the largest one uniting 10 STs (Figure 3.8 connected component 12). Where there is a GenBank sequence for an ST, that provides genome-level confirmation of the ST. However, in the absence of a GenBank entry, there is no such confirmation, leaving open the possibility that these

STs may represent co-cultures, sequencing artifacts, or other uncertainties.

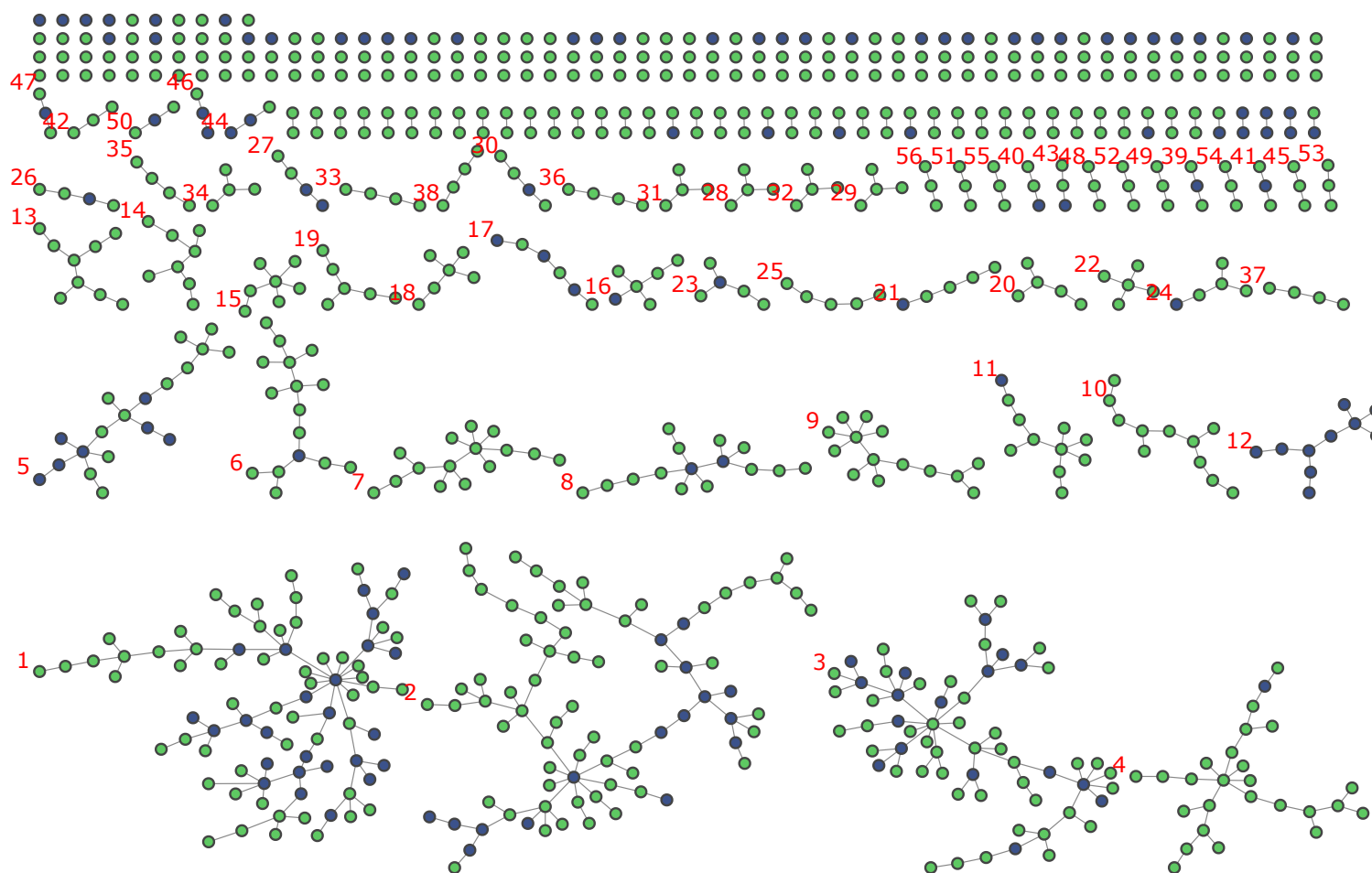


Figure 3.8: Minimum Spanning Tree of the updated pubMLST *Streptomyces* scheme showing STs with a representative in GenBank (green), and STs with no representative in GenBank (blue).

3.3.3 Connectivity of MST

The MST for *Streptomyces* comprises 278 components that are not connected to each other by a shared allele sequence. This might reflect the true distribution of connectivity for all naturally-occurring *Streptomyces* - a result of their divergence into several groups that cannot, or do not get the opportunity to, recombine. Alternatively, if we were to sequence all *Streptomyces* genomes and repeat the analysis, we might find that they are all in fact connected into a single component, and that the graph in Figure 3.7 is an artefact of sampling. To determine whether the sampling of *Streptomyces* genomes is responsible for the distribution of component sizes in the MST, I investigated changes in MST topology as a result of repeated MST construction using random subsamples of known STs ranging from 10% to 90% of the available genomes (Methodology section 3.2.7). I found that the number of disjoint components increases exponentially with the number of sampled genomes (Figure 3.9), i.e. addition of new genomes brings more diversity to the group, rather than connecting existing components. Additionally, the distribution of the relative sizes of connected components (as a proportion of the total number of genomes in the analysis), was observed to be the same at each sampling depth (Figure 3.10). Taken together these results suggest that additional sequencing of the natural variation of *Streptomyces* spp. will uncover further variation, but that the sampled distribution of related sequences is representative and extensive sequencing will not "fill in gaps" in the tree.

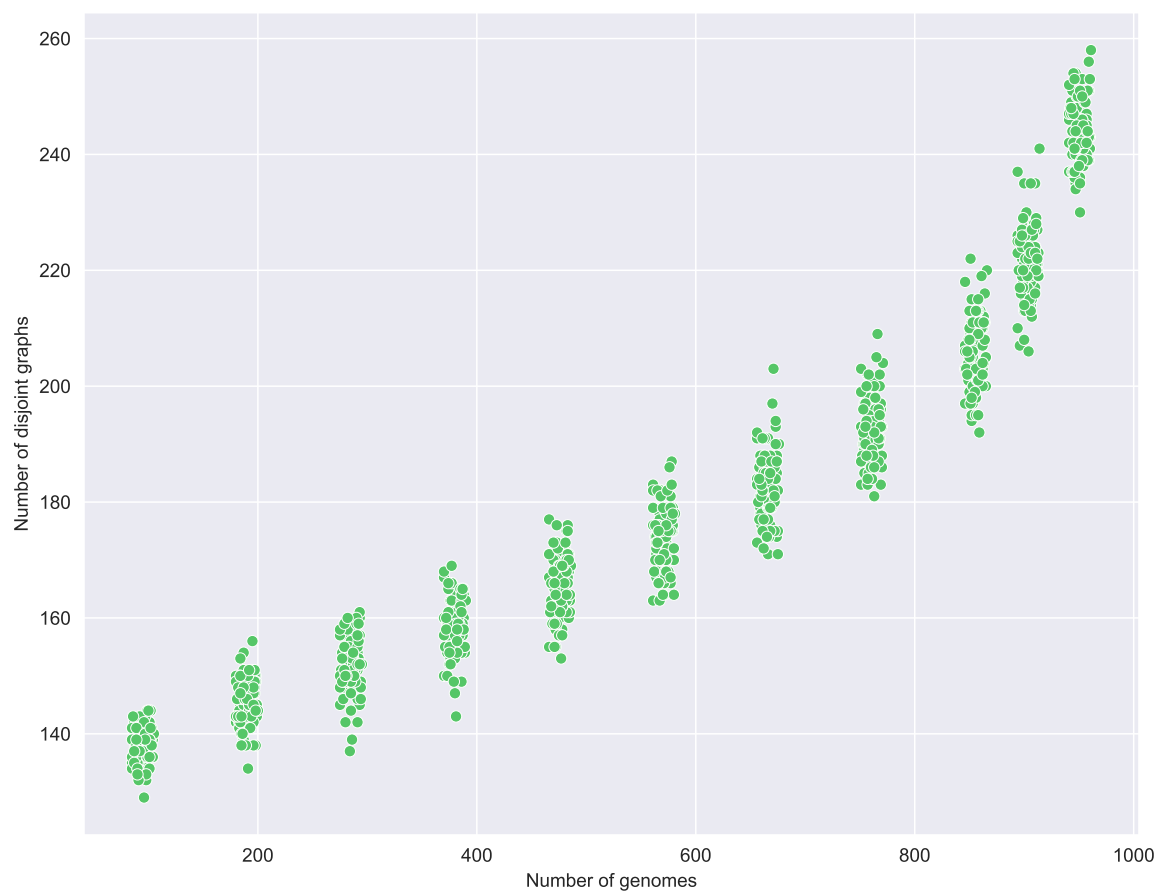


Figure 3.9: Scatter plot showing the relationship between the number of randomly subsampled genomes (10-90% of the original dataset) and number of disjoint graphs. Addition of genomes generates more disjoint components, rather than uniting existing subgraphs.

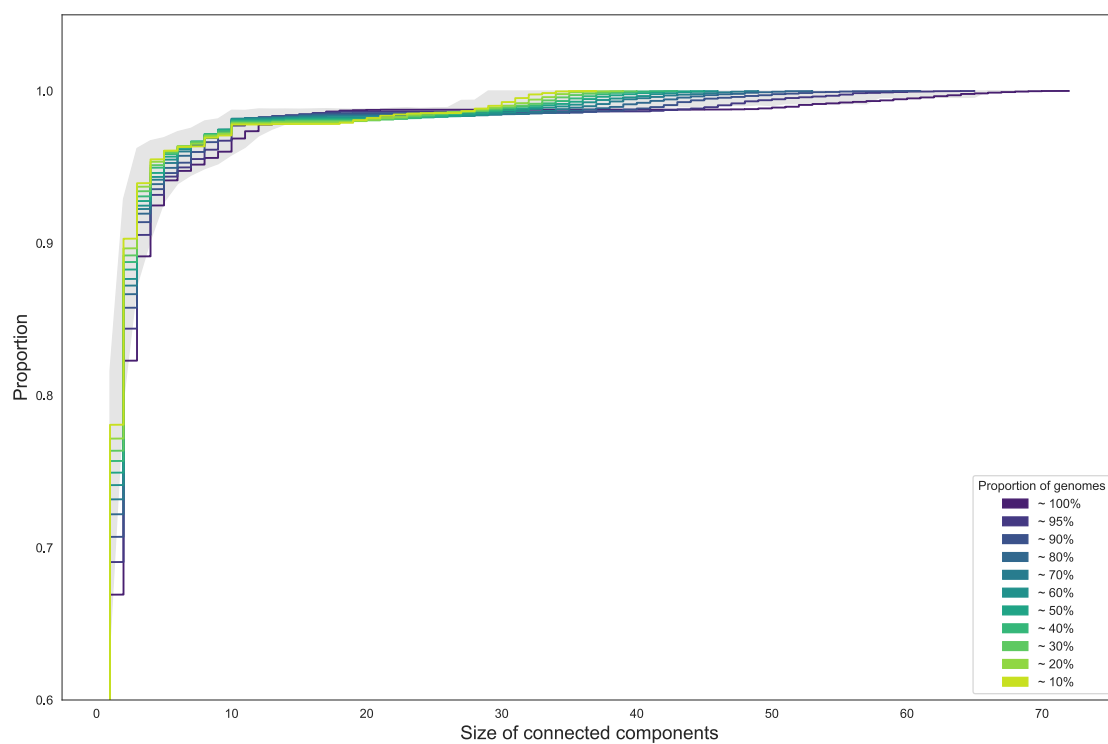


Figure 3.10: The distribution of relative connected component sizes as the number of randomly sampled genomes varies from 10-90% of the original dataset size. The distribution of (relative) subgraph sizes is essentially independent of subsampling depth.

As the genome data provided by the six genome markers do not fully capture the extent of diversity present in natural *Streptomyces* population, and the results obtained from current sequencing efforts may be an artefact of sampling bias, I repeated my analysis on a simulated dataset. I investigated the effect of subsampling on the distribution of component sizes using an artificial scheme consisting of 5000 STs and six marker genes. In this scheme, when constructing a MST using all 5000 artificial STs and allowing changes of up to five alleles, the result is a single disjoint graph. With complete sampling, all genomes are be connected into a single component. This represents a simulated biological reality corresponding to one possible outcome of very extensive sequencing: all "gaps" in the current MST are gradually filled in to form one connected tree of organisms.

Repeating the subsampling analysis that was applied to the observed ST data above, I found that the number of subgraphs decreased as more genomes were "sequenced" and more STs were added to the scheme (Figure 3.11). This relationship is contrary to that observed in Figure 3.9. As the number of "sequenced" genomes increased there there was an accompanying change to the distribution of the relative size of connected components (Figure 3.12). Again, this was not the relationship observed for the actual *Streptomyces* ST data (Figure 3.10). These simulations, and the scale-free distribution of subgraph sizes (Figures 3.9, 3.10) suggest that the subgraphs observed may truly represent distinct populations, and not artifacts of subsampling from a larger, incompletely-sampled *Streptomyces* population that could be represented as a single, large connected graph.

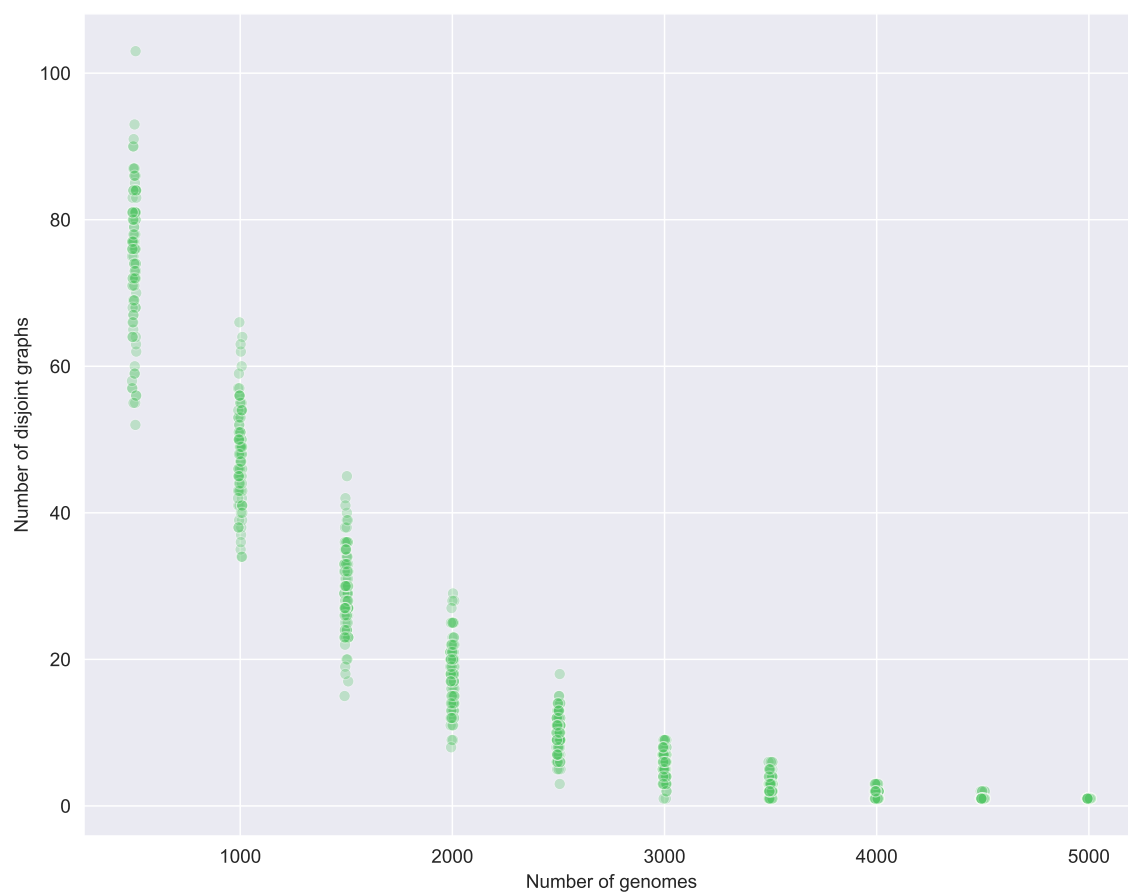


Figure 3.11: Relationship between the number of randomly sampled genomes (10-90% of the original dataset) and number of observed disjoint subgraphs, for the artificial scheme.

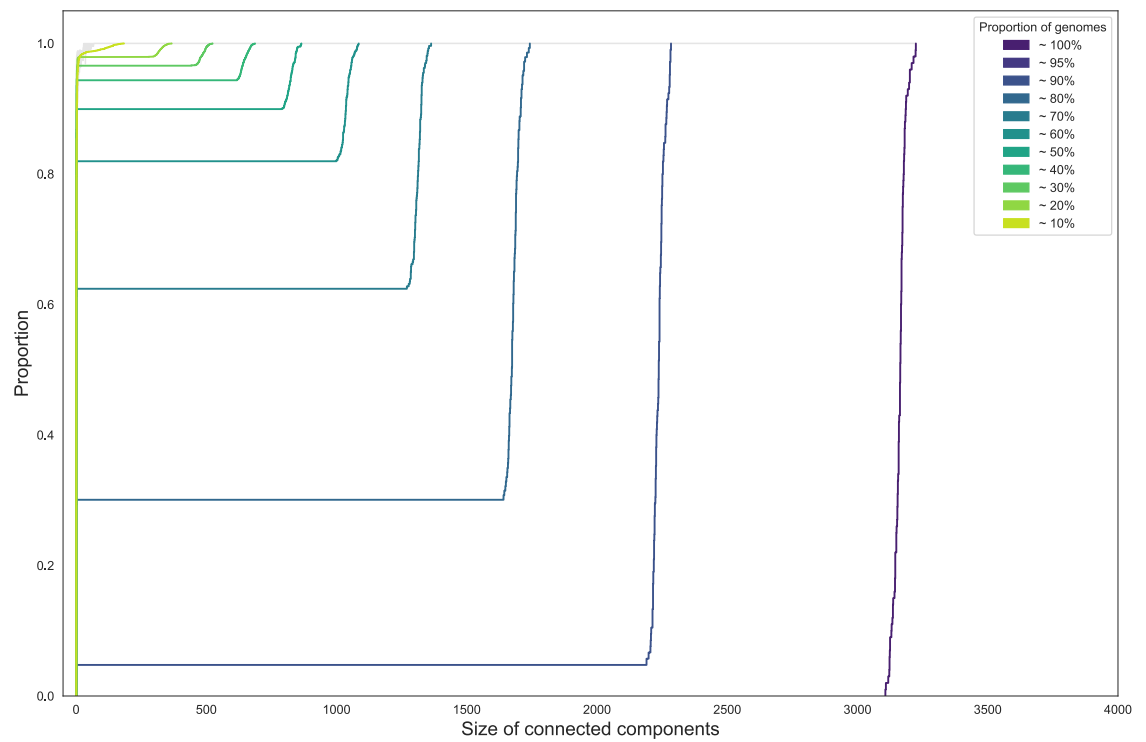


Figure 3.12: Distribution of relative connected component sizes (as a proportion of total number of genome) when randomly sub sampling 10-90% genomes from the artificial scheme.

3.3.4 Empirical test

I observed that pubMLST STs appear to be more "central" in the MST subgraphs (Figure 3.4) in the sense of pubMLST nodes tending to have a higher degree than nodes representing novel STs. Nodes representing pubMLST STs are seen to connect with as many as 13 distinct STs (ST8 located in connected component 1, i.e. a node degree of 13), while the novel STs connect with only up to seven other STs (ST342 located in connected component 4, i.e. a node degree of 7) (Table 3.4).

Using an empirical non-parametric network test (methodology section 3.2.8) I reproduced the same network structure but randomly reassigned the same numbers of "pubMLST" and "novel" labels to each node, I estimated the probability of observing the same or higher count of high-degree pubMLST nodes (≥ 7). It is possible that shuffling the labels does not affect the overall distribution of nodes; thus, pubMLST nodes would generally retain a high degree. This would imply that the high degree of pubMLST nodes is an expected outcome when numerous pubMLST and novel sequence types (STs) are present within a network. Alternatively, the current distribution of node degrees among pubMLST and novel STs might be unlikely to occur by chance, suggesting that shuffling labels would rarely yield a similar distribution. This would imply that shuffling the labels would rarely yield a similar distribution, suggesting that there exists some biological or artefactual significance that positions pubMLST nodes in "central" locations within the network. I find the latter scenario is more likely; under this model, derived from the same underlying graph structure, we would expect to see fewer high-degree pubMLST nodes ($p=0.028$). This result implies that the observed

high degree of pubMLST nodes is unlikely to arise as a coincidence and suggests that additional factors, beyond randomness, might be influencing the central placement of pubMLST nodes.

Table 3.4: Summary of node degrees for novel and pubMLST STs.

Degree	Count of pubMLST nodes	Count of novel nodes
0	63	115
1	79	303
2	47	107
3	18	26
4	17	12
5	5	3
6	4	1
7	1	1
9	1	0
10	1	0
11	0	0
12	0	0
13	1	0

Likewise, the distribution of the STs with no genome sequence in GenBank ("non-GenBank") is such that they tend to be central in each subgraph, and with high degree (linking to a higher number of genomes/STs, Figure 3.8). In total, I identified four (0.62%) STs represented by genomes in Genbank to have a degree ≥ 6 . Six (4%) non-Genbank STs have at least degree 6 (Table 3.5).

By a similar empirical non-parametric network test to that above (described in Methodology Section 3.2.8), shuffling "GenBank" and "not GenBank" labels on the existing MST graph, I estimated the probability of observing the same or higher number of non-GenBank nodes to have degree ≥ 6 . I estimated that for a random subset of 150 nodes extracted from the original graph, it is unlikely that at least six would have a degree ≥ 6 ($P=0.04196$). This suggests that STs without representative genomes and pubMLST STs may act as bridges connecting well-defined STs (eg. sequenced STs with representative genomes).

Table 3.5: Summary of degree node connections represented and not represented in Genbank.

Degree	Count of nodes represented in Genbank	Count of nodes not represented in Genbank
0	142	36
1	334	48
2	120	34
3	33	11
4	17	12
5	5	3
6	2	3
7	1	1
9	0	0
10	1	0
11	0	0
12	0	0
13	0	1

3.3.5 Comparing MLST divisions and whole-genome sequence classification landscapes

ANIm analysis of genomes sharing identical STs

A total of 552 STs were represented by a single genome and 103 STs by between two and 27 genomes. The five most highly-represented STs in Genbank are summarised in Table 3.6. Examples of genomes representing the same STs were found to: i) have a consistent assigned taxonomy (eg. ST167), ii) be represented by multiple names assigned in NCBI (eg. ST249, ST2, ST168); or iii) currently lack any assigned species names (ST241). I found a total of 11 STs corresponding to genomes assigned conflicting species names at NCBI (Figure 3.13).

To establish whether this observation stems from conflicts in nomenclature arising from misclassification or whether the current set of markers lack discriminatory power to differentiate distinct species, I established taxonomic boundaries using ANI comparisons between genomes sharing identical STs (Methodology Section 3.2.9). Remarkably, although genomes sharing identical STs may have been assigned different species designations in NCBI, the genome coverage does not fall below 69.4% (Figure 3.14), and the lowest pairwise average nucleotide identity recorded was 98.9%, indicating that a single ST is likely always to correspond to a single species (Figure 3.15). In cases where discrepancies arise, it may be possible to identify genome sequences of type strains in the database, which could help pinpoint misclassified strains and correct errors in the existing literature. For instance, despite being assigned to *S. rimosus*, *S. capuensis*, and *Streptomyces sp.*, the ST 249 was found to include three genomes (GCF_008704655.1,

GCF_000717285.1 and GCF_000331185.2) that were type strains, all of which were consistently assigned the *S. rimosus* species designation. Thus, it remains likely that the misannotated genomes should be reannotated as *S. rimosus* in this case.

Table 3.6: Five larges STs with representatives in GenBank and their corresponding taxonomic assignments.

ST	Genome Count	Dominant taxonomic assignment	Additional taxonomic assignment
167	27	<i>Streptomyces clavuligerus</i> (27)	NA
249	22	<i>Streptomyces rimosus</i> (18)	<i>Streptomyces capensis</i> (2) <i>Streptomyces sp.</i> (2)
2	10	<i>Streptomyces californicus</i> (7)	<i>Streptomyces purpeochromogenes</i> (2) <i>Streptomyces sp.</i> (1)
168	9	<i>Streptomyces coelicolor</i> (8)	<i>Streptomyces sp.</i> (1)
241	8	<i>Streptomyces sp.</i> (8)	NA

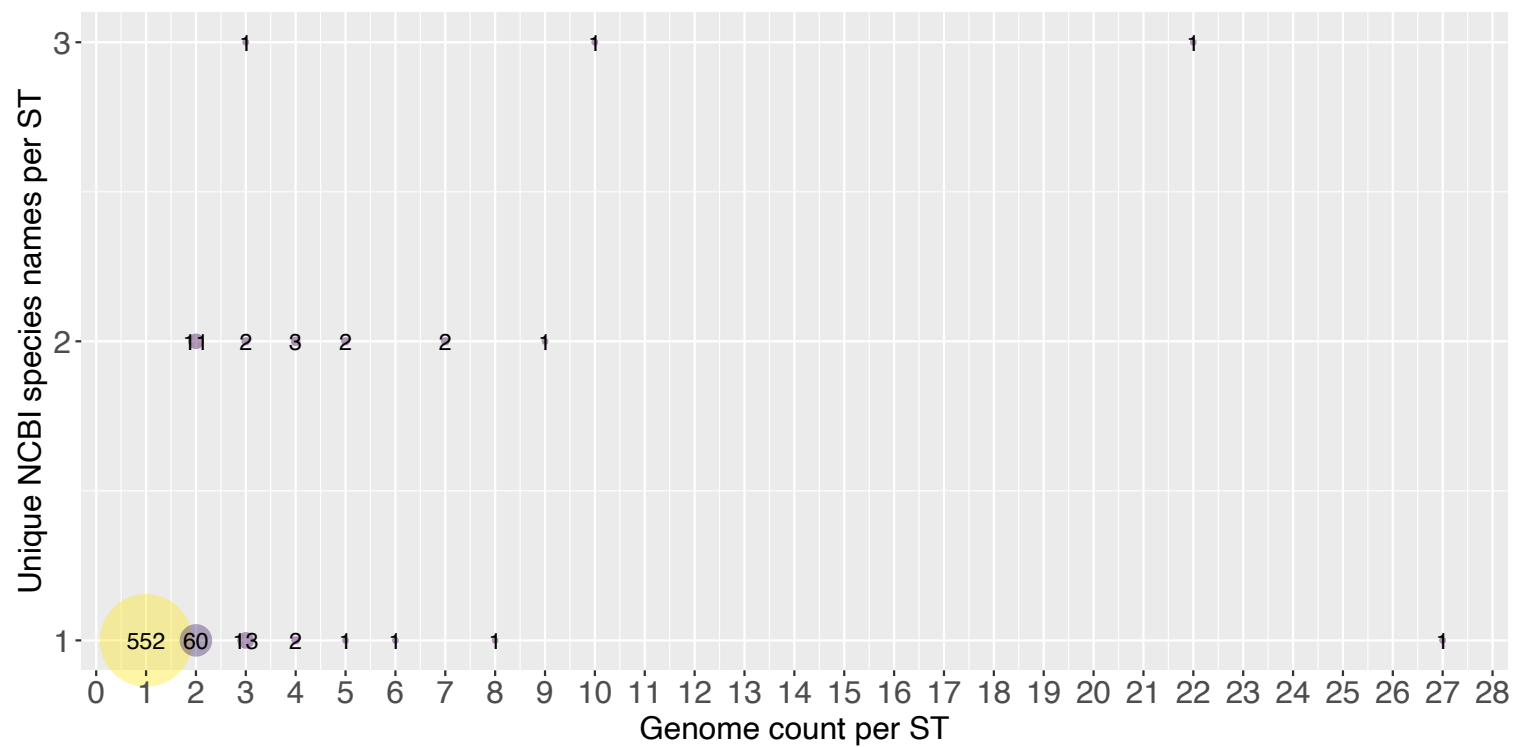


Figure 3.13: Genomes sharing identical STs can be assigned different taxonomic names in NCBI. The y-axis represents the number of unique NCBI species names, excluding *Streptomyces* sp., that are shared among genomes with identical STs. The x-axis indicates the number of genomes corresponding to each ST.

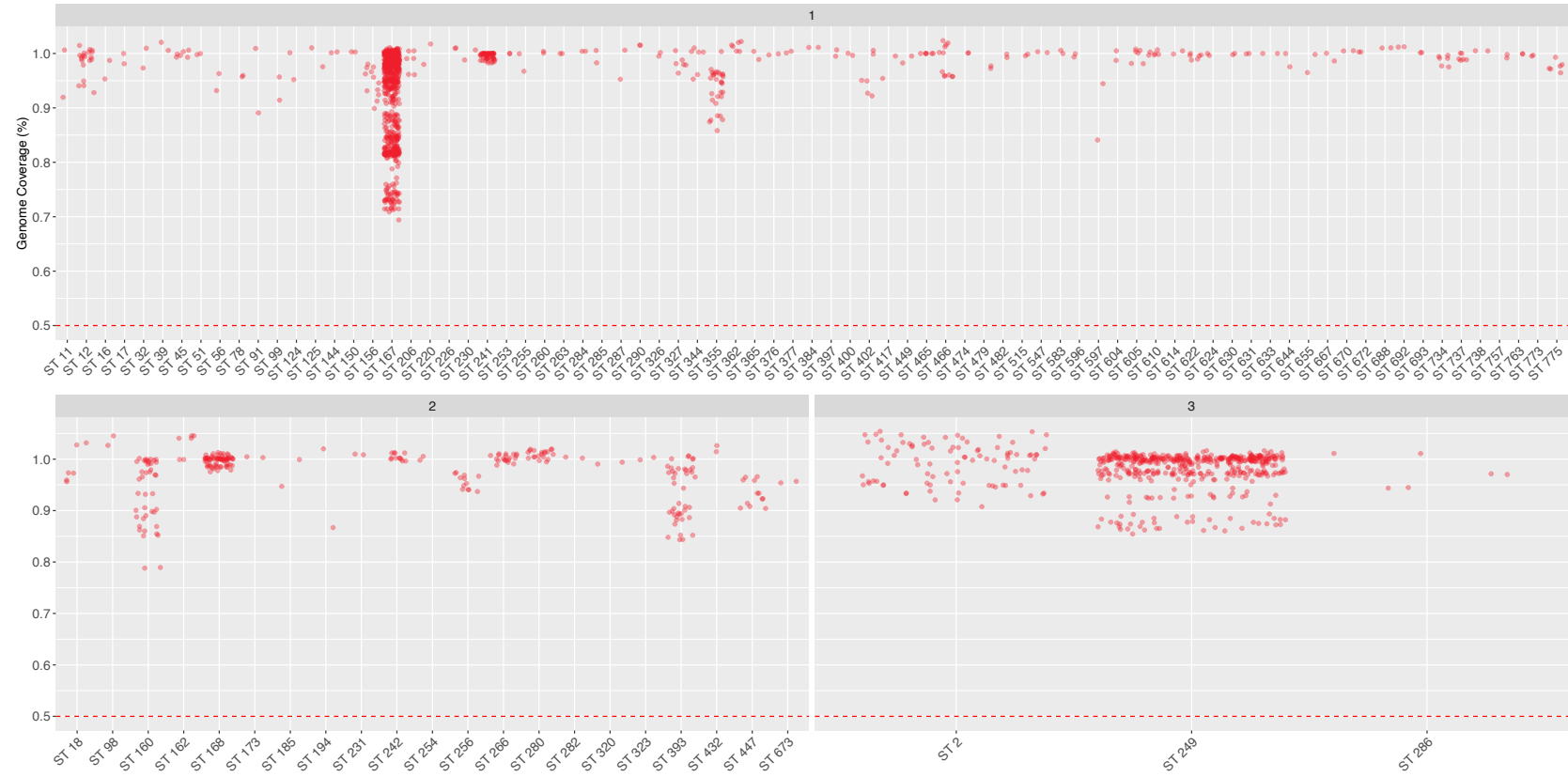


Figure 3.14: Scatter plots of genome coverage for pairwise ANIm comparisons for genomes sharing identical STs. The number of unique species names among genomes sharing identical STs is displayed at the top of each plot. The red horizontal line indicates the whole-genome genus threshold (50%).

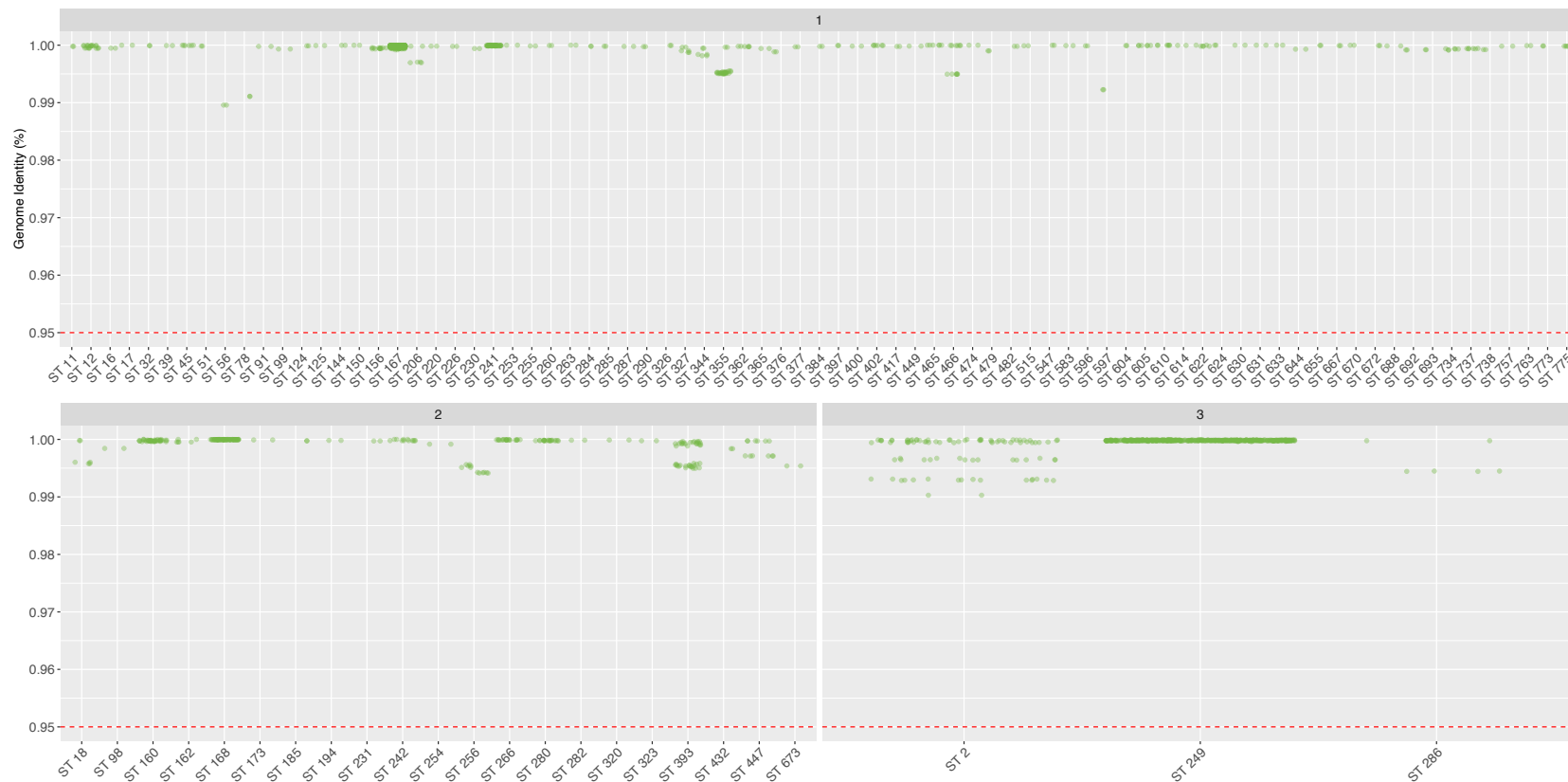


Figure 3.15: Scatter plots of genome identity for pairwise ANIm comparisons for genomes sharing identical STs. The number of unique species names among genomes sharing identical STs is displayed at the top of each plot. The red horizontal line indicates the whole-genome species threshold (95%).

ANIm analysis of connected components

I investigated the taxonomic composition of each MLST connected component subgraph using ANIm analysis (Methodology Section 3.2.9). I found cases where ANIm species classification aligns with MLST divisions, bringing together single species, where all genomes within a single connected component share a minimum of 50% genome coverage (Figure 3.16A) and 95% average nucleotide identity (Figure 3.16B) (connected component 13 in the MST). However, some subgraphs connect genomes belonging to the same genus but distinct species, where all genomes share at least 50% genome coverage (Figure 3.16C), but not all share $\geq 95\%$ ANI in a pairwise comparison (Figure 3.16D) (connected component 10 in the MST). I also identified cases where a single connected component unites genomes that could be considered distinct genera in other groups of bacteria; in these cases, genomes share less than 50% genus (Figure 3.16E) and less than 95% genome identity (Figure 3.16F) (connected component 2 in the MST). Thus, while a single ST is likely to map to a single species, a connected subgraph linking STs by at least one common allele may link members of more than one *Streptomyces*, or genus-level grouping. The counts of subgraphs falling into each category are summarised in Table 3.7. Heatmaps for all remaining ANIm pairwise comparisons can be found in the `supplementary_file_17/output/pyani_heatmaps_connected_components` folder, and scatterplots summarising these pairwise ANIm comparisons are shown in Figure 3.17 (for genome coverage) and Figure 3.18 (for genome identity).

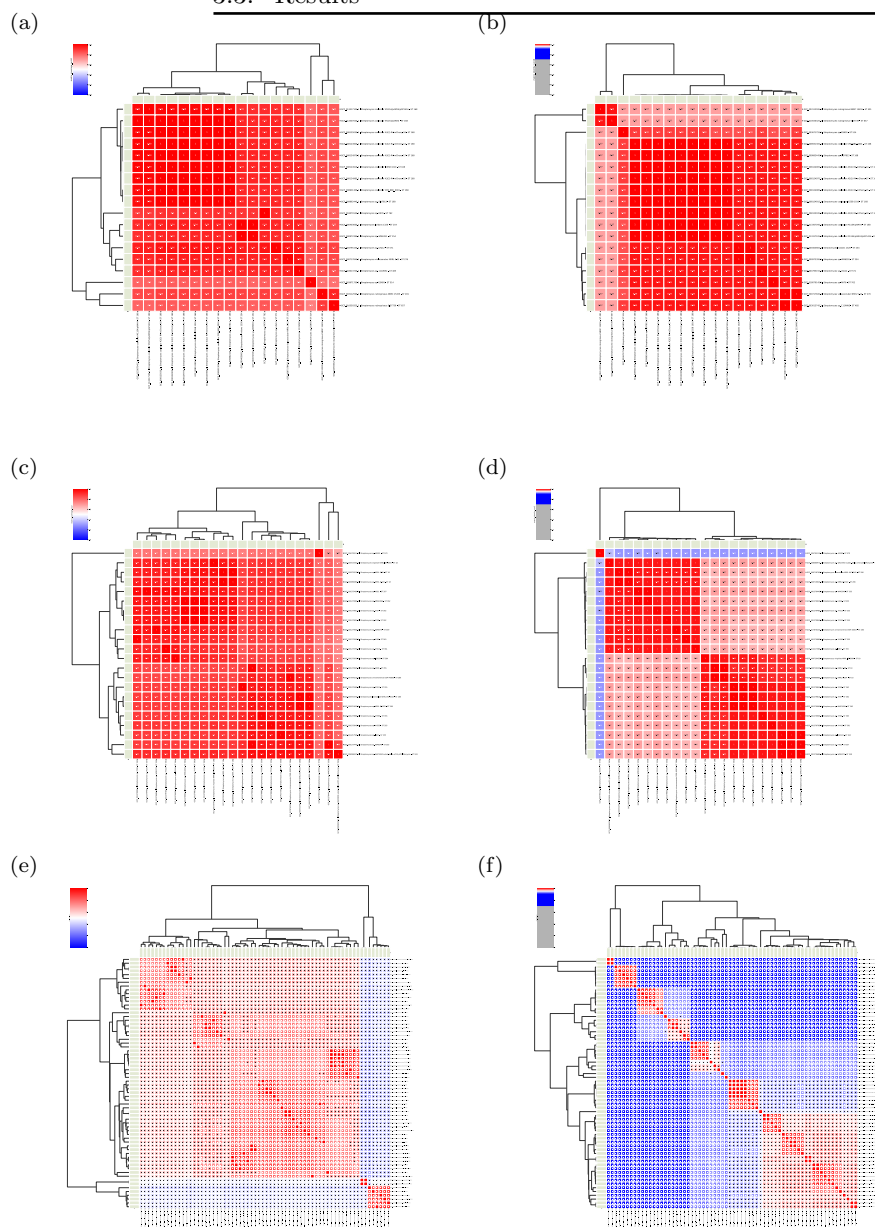


Figure 3.16: ANIm genome coverage (left) and ANIm identity (right) analysis of genomes found in the same group of connected STs. The heatmaps in each row represent comparisons between members of the same connected component. In the genome coverage plots (left), red cells indicate genome coverage of $\geq 50\%$, which suggests common membership of the same genus, whereas blue cells indicate coverage below 50%, implying presence of distinct candidate genus. In the genome identity plots (right), the red cells indicate genome identity of $\geq 95\%$, which can be interpreted as membership of the same species, whereas blue cells ($< 95\%$ genome identity) imply distinct species.

Table 3.7: Summary of the taxonomic composition of subgraphs uniting at least two genome assemblies.

Category	Interpretation	Count	Percentage
$\geq 50\%$ coverage; $\geq 95\%$ identity	same species	78	78%
$\geq 50\%$ coverage; $< 95\%$ identity	same genus, different species	14	14%
$< 50\%$ coverage; $< 50\%$ identity	different genera	8	8%

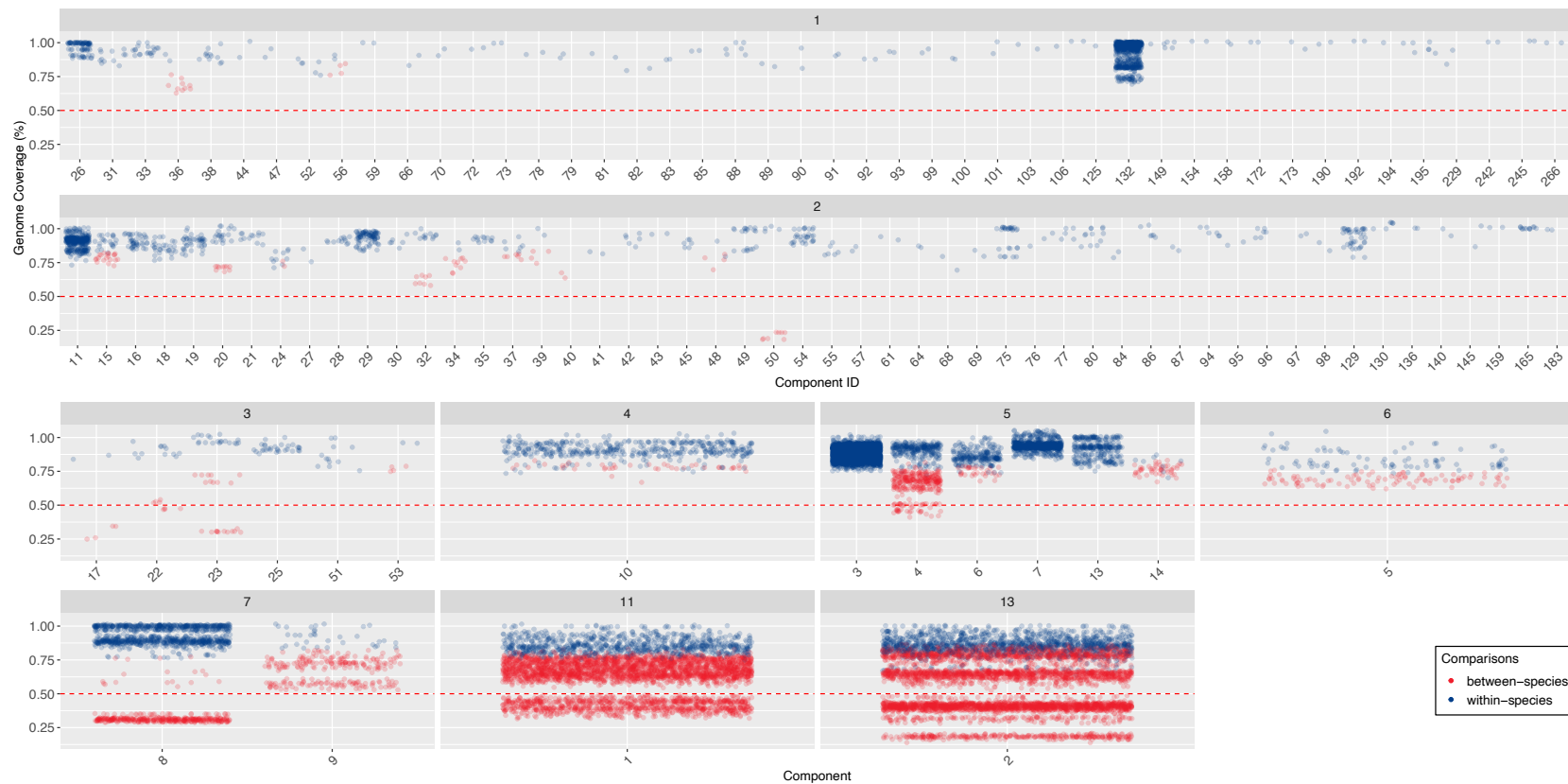


Figure 3.17: ANIm genome coverage analysis of genomes found in the same group of connected STs. The number of unique species names found in each connected component is shown at the top of each plot. Within species comparisons ($\geq 95\%$ genome identity) are shown in blue, whereas between species comparisons ($< 95\%$ genome identity) are shown in red. The red horizontal line indicates the whole-genome genus threshold (50%).

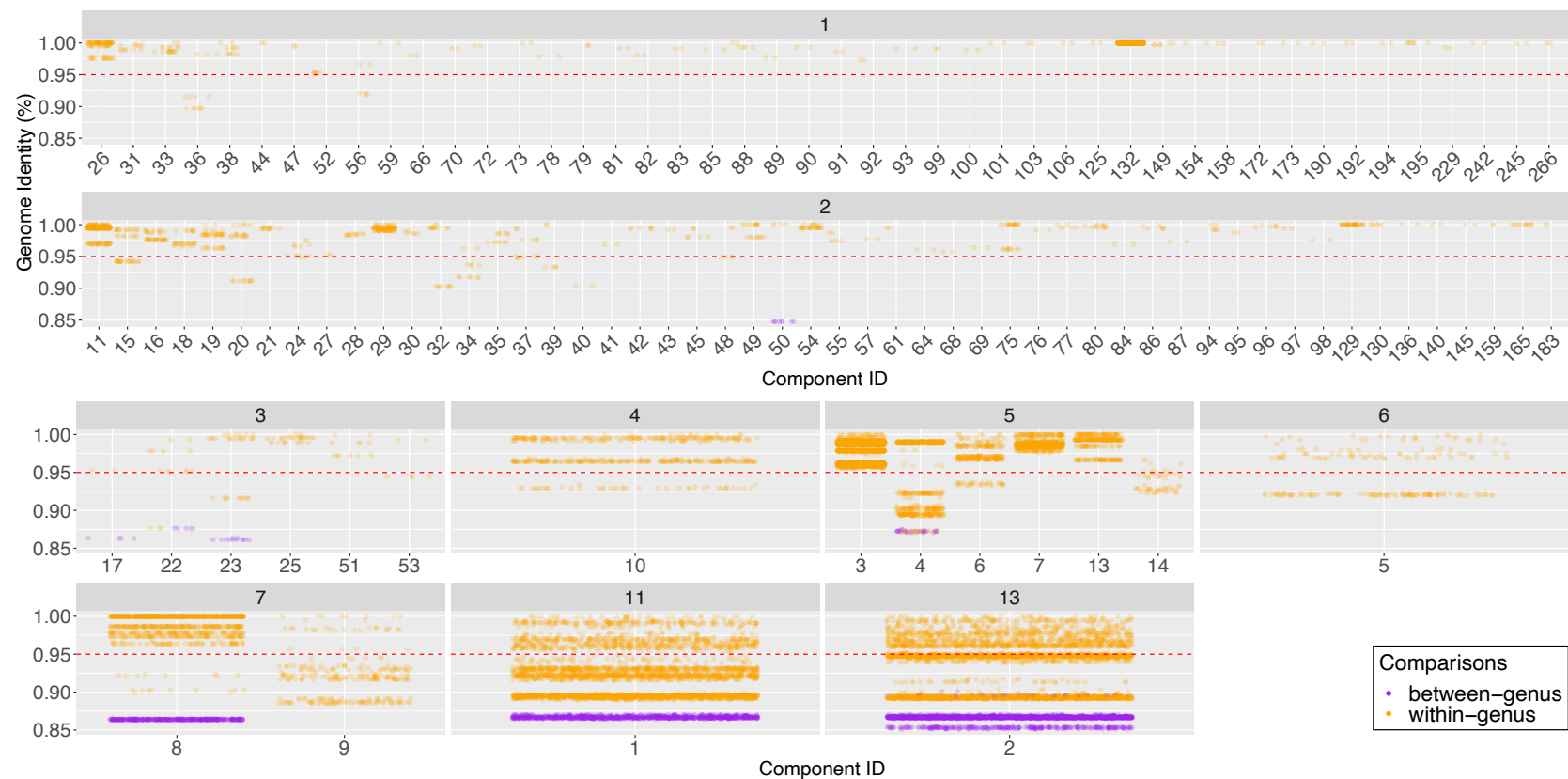


Figure 3.18: Genome identity scatter plots for all connected components comprising at least two sequenced genomes. The numbers at the top of each plots correspond to the number of unique species names found in each connected component. Within genus comparisons ($\geq 50\%$ genome identity) are shown in orange, whereas between genus comparisons ($< 50\%$ genome identity) are shown in purple. The red horizontal line indicates the whole-genome species threshold (95%).

The results of ANIm coverage analysis were mapped onto MST to visually illustrate the distribution of unique genera within each connected component (Figure 3.19). I identify 92 (92%) non-singleton connected components uniting only isolates that share at least 50% of their genomes by alignment length, indicating their likely membership of the same genus-level grouping. However, I also observed that eight (8%) connected components contain more than one distinct genus-level group; the second largest connected component (Figure 3.19 connected component 2) consists of as many as four such groupings.

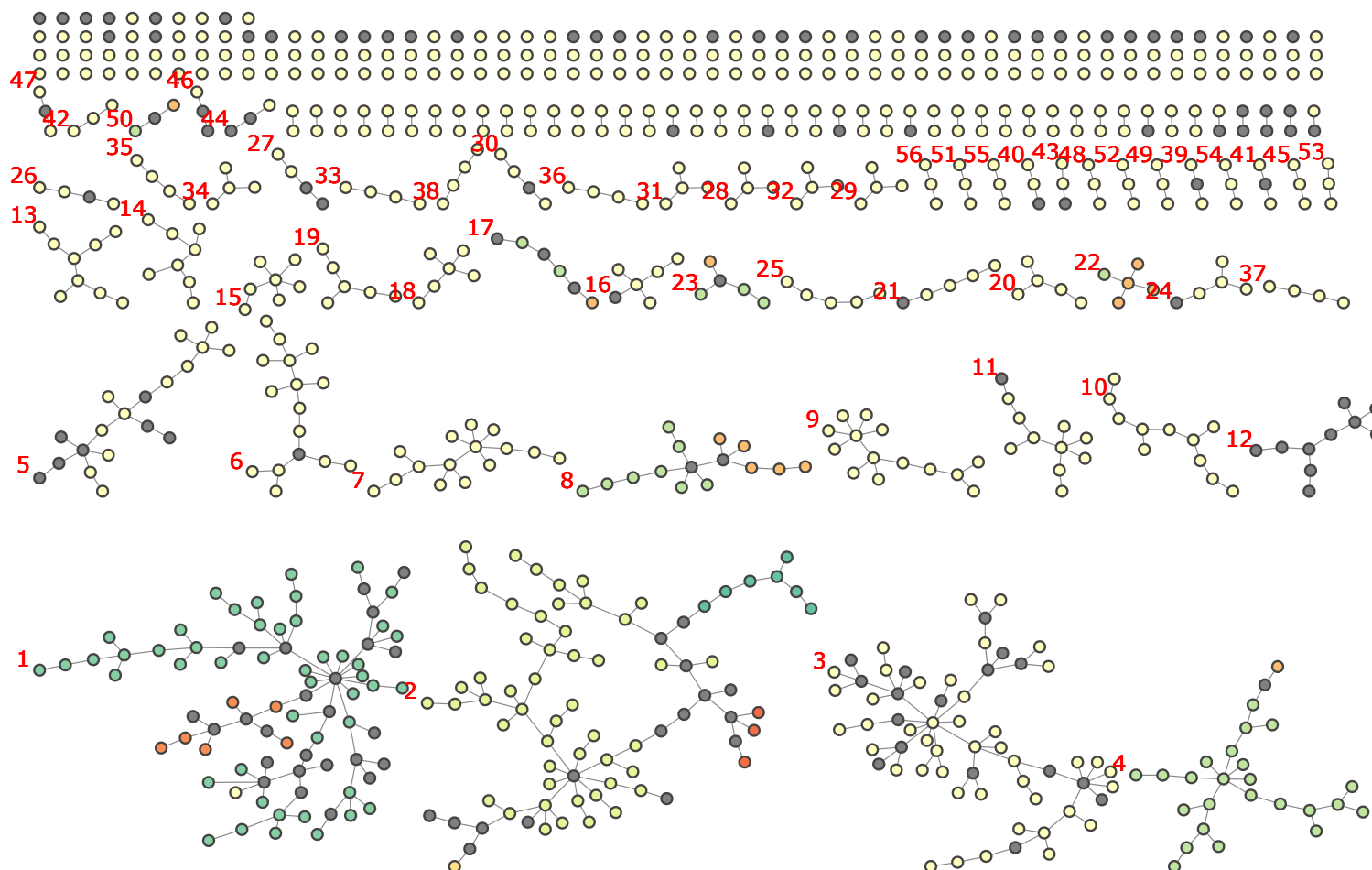


Figure 3.19: Minimum Spanning Tree of the updated pubMLST *Streptomyces* scheme showing a count of unique candidate genera per connected component. Each candidate genus is represented as a single node colour within a connected component. STs lacking a representative genome in NCBI are shown as grey nodes.

I also visually represented the %ANIm identity on the MST to illustrate the distribution of unique species within each connected component (Figure 3.20). This analysis revealed that 78 (78%) of the non-singleton connected components unite only a single candidate species, in which all genomes within the component share a genome identity of 95% or over. However, I also identified 22 (22%) connected components that unite distinct candidate species. The largest number of different candidate species within a single component was eleven (Figure 3.20 connected component 1). A notable feature of this MST is the presence of "hub" nodes—pubMLST STs that lack genomic representation in GenBank—which appear to serve as linkers between groups of organisms from different species and even genera.

Assignment of taxonomic status within each connected component using %ANIm identified 103 (35% of all species) instances where a single *Streptomyces* species was represented by multiple STs (Figure 3.21), while the remaining 192 (65%) species were represented by a single ST. This suggests that, while each ST may map to a single species, in general a single *Streptomyces* species can map to multiple STs.

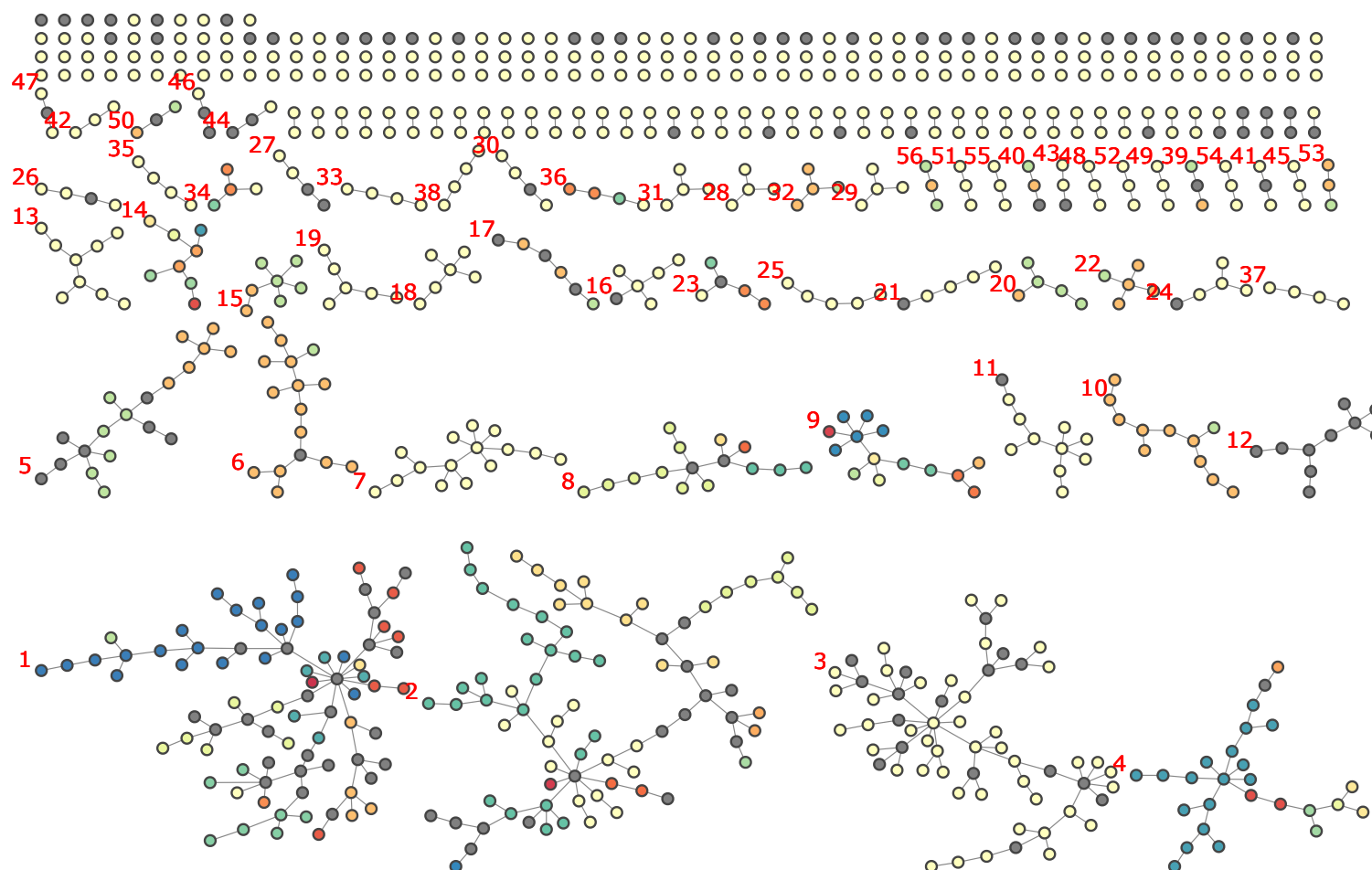


Figure 3.20: Minimum Spanning Tree of the updated pubMLST *Streptomyces* scheme showing a count of unique candidate species per connected component. Each candidate genus is represented as a single node colour within each connected component. STs lacking a representative genome in NCBI are shown as grey nodes.

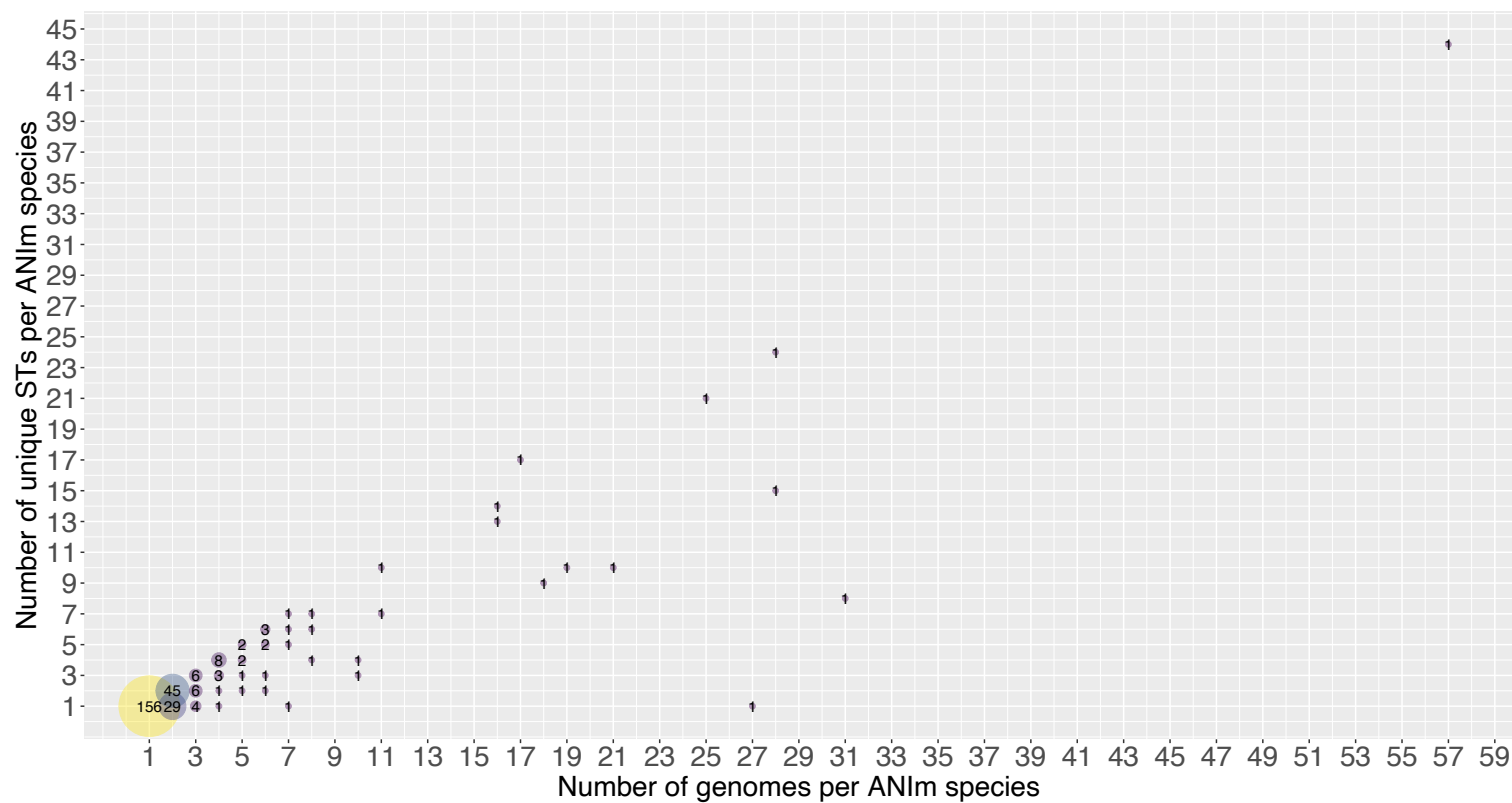


Figure 3.21: Multiple STs are used to describe single *Streptomyces* species ($\geq 50\%$ genome coverage; $\geq 95\%$ genome identity). The y-axis represents the number of unique STs per each species, while the x-axis shows the number of genomes representing each species.

Finally, I investigated the diversity of assigned nomenclature for ANI-classified species ($\geq 50\%$ genome coverage; $\geq 95\%$ genome identity) in each connected component. I found that in 222 species (of which 156 were represented by a single genome), all candidate species according to the ANIm cutoff threshold of $\geq 50\%$ genome coverage and $\geq 95\%$ genome identity had been assigned the same name. However I also found 73 incongruencies between the whole-genome species assignments and current nomenclature. In one particular case, a single species determined by whole-genome comparison was found to have been assigned five distinct names in NCBI (Figure 3.22 and Figure 3.23).

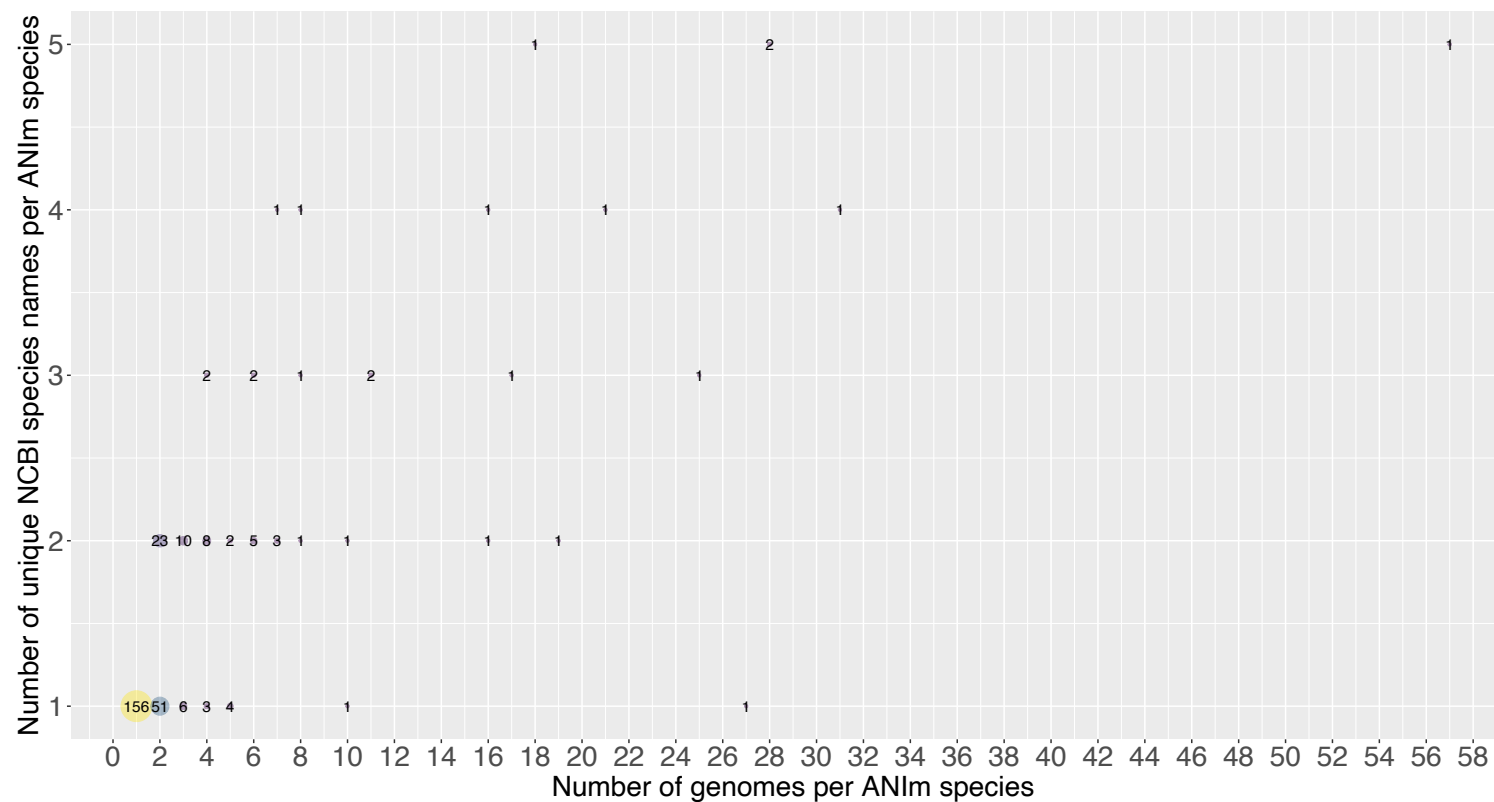


Figure 3.22: Multiple NCBI names are used to describe single *Streptomyces* species indicated by whole-genome circumscription ($\geq 50\%$ genome coverage; $\geq 95\%$ genome identity). The y-axis represents the number of unique names assigned in NCBI to each species, while the x-axis shows the number of genomes representing each species.

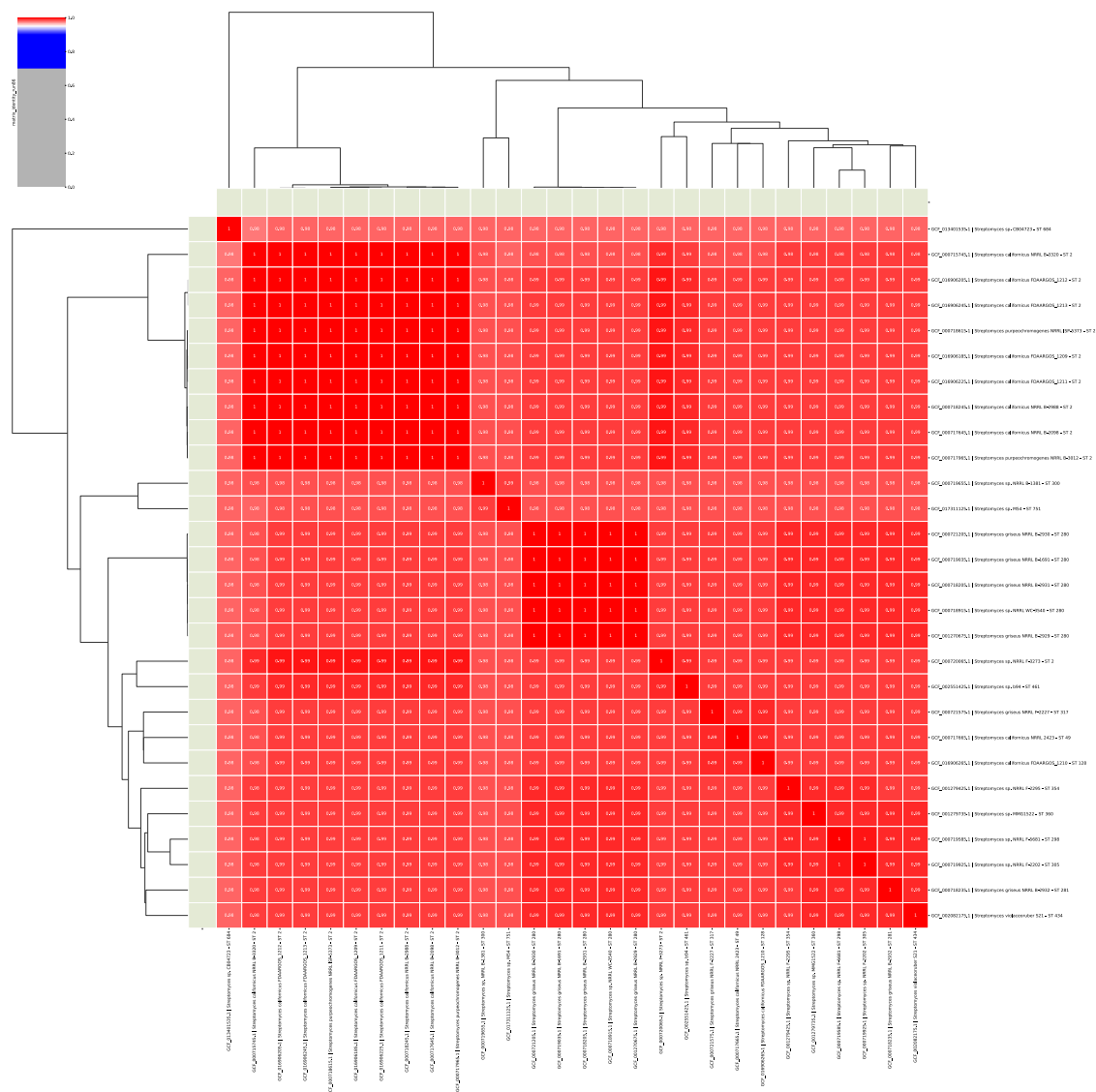


Figure 3.23: ANIm genome identity for an exemplar case where species designations in NCBI do not match, despite sharing $\geq 95\%$ genome identity. No values fall below the 95% ANIm threshold, yet the genomes share different names, including *Streptomyces californicus*, *Streptomyces purpeochromogenes*, *Streptomyces griseus*, *Streptomyces violaceoruber*, and *Streptomyces* sp.

3.3.6 ANIm analysis of genomes assigned the same species designations in NCBI

I found that among the 80 distinct Streptomyces species designations assigned to genomes (other than *Streptomyces sp.*) that were represented by at least two genomes: i) 29 species are represented by a single ST (eg. *Streptomyces coelicolor*, *Streptomyces clavuligerus*); ii) 31 species designations are present in the same connected component but are represented by multiple STs (eg. *Streptomyces scabiei*); and iii) 20 appear in multiple disjoint groups of STs, which do not share a single allele marker with each other (eg. *Streptomyces rimosus*, *Streptomyces griseus*, *Streptomyces olivaceus*). The distribution of this third group would be consistent with either misannotation or unusual allele diversity.

To determine the factors that contributed to inconsistencies between nomenclature and MLST divisions, ANI analysis was applied to genomes sharing identical species designations at NCBI (Methodology 3.2.9). The results are shown as scatter plots showing genome coverage (Figure 3.24) and genome identity (Figure 3.25). I found that all isolates sharing the same species designations in NCBI appearing in a single connected component share no less than 69.4% (*S. clavuligerus*) genome coverage, implying their membership of the same genus (Figure 3.24). It is also apparent that these isolates usually represent the same species as the majority of clusters share $\geq 95\%$ identity (Figure 3.25), with a single exception for *Streptomyces misionensis* genomes sharing 94.9% identity - though I note that the 95% identity threshold is not a strict cutoff, and these genomes almost certainly should be considered the same species. These data also

show that ten taxonomic species names (*Streptomyces artratus*, *Streptomyces globisporus*, *Streptomyces hygrosopicus*, *Streptomyces lydicus*, *Streptomyces niveus*, *Streptomyces olivaceus*, *Streptomyces subbrutillus*, *Streptomyces violaceusniger*, *Streptomyces rimosus*, *Streptomyces virginiae*) were split across multiple connected components despite sharing $\geq 50\%$ genome by alignment length (Figure 3.24) implying they represent the same genus-level grouping. Of these, genomes corresponding to *S. niveus*, *S. olivaceus* and *S. violaceusniger* shared more than 95% genome identity with other genomes of the same assignment but in distinct subgraphs, suggesting they are likely the same species despite their sharing no MLST alleles in common. These observations are unlikely to be due to poor genome quality, as assessed using checkM (ethodology section 3.2.6). Genome completeness ranged from 100% to 99.93% with contamination up to 1.58% for *S. violaceusniger* and from 100% to 99.53% completeness with contamination below 0.07% for *S. olivaceus*. However, the quality of *S. niveus* may have influenced these findings, as its completeness ranged from 95.2% to 94.61%, with contamination no higher than 0.89%. I also found that 10% (10) of all clustered assemblies, based on the same shared name in NCBI, share less than 50% of their total genome length, with *Streptomyces albus* (GCF_000719865.1; completeness 100% and contamination 0.53%) sharing genome coverage of 13.1% with the other sequenced isolates of that name (Figure 3.24 - *Streptomyces albus* was also found in 2 distinct connected components).

Similarly, I identified 17 cases of species-level misannotations, where genomes assigned the same species designations in NCBI share less than 95% genome identity, with 85% being the lowest identity seen, once again for *Streptomyces albus* (Figure 3.25).

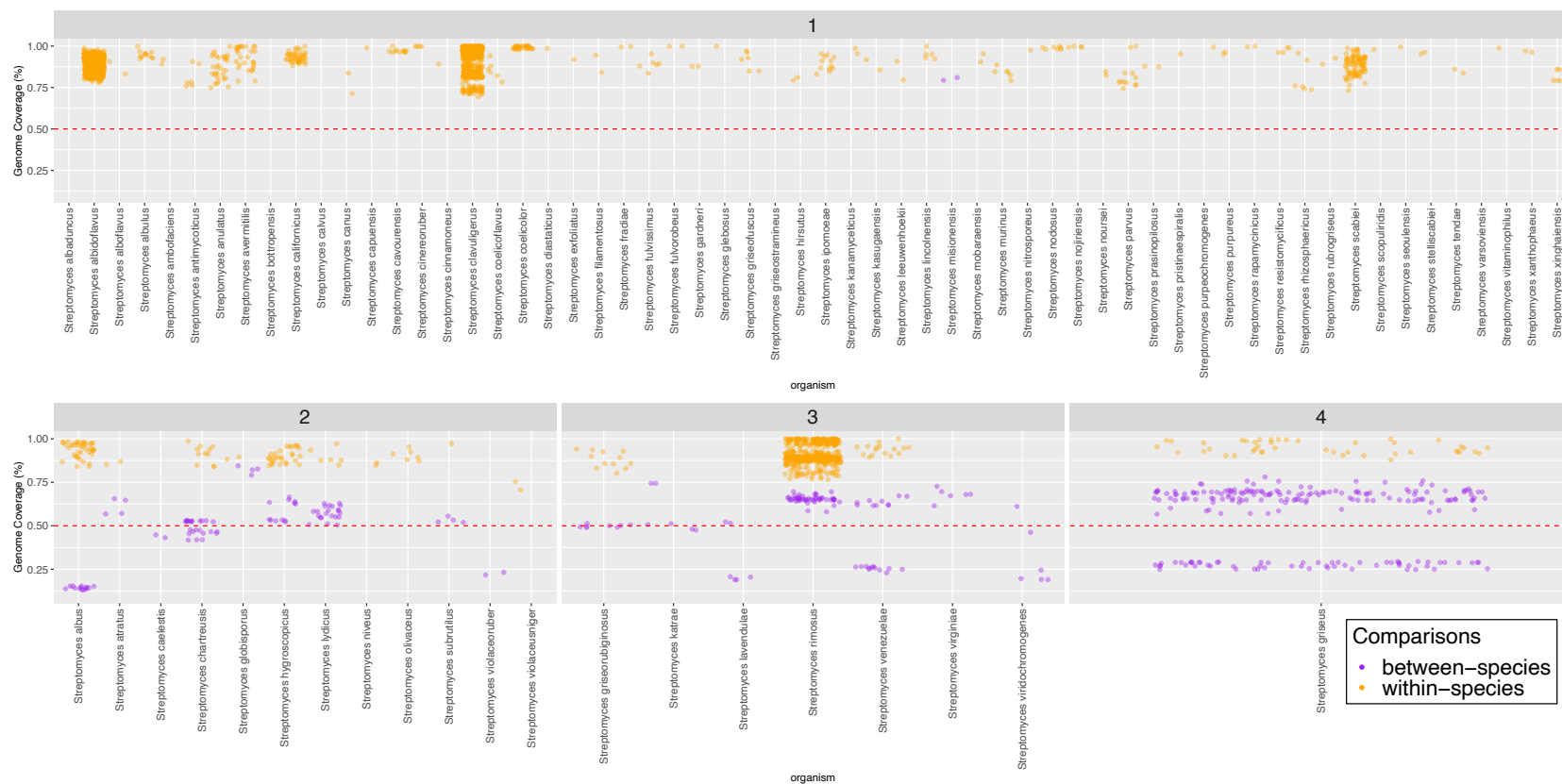


Figure 3.24: ANIm coverage analysis of genomes sharing the same name other than *Streptomyces sp.* in NCBI. The numbers at the top of each plot correspond to the count of connected components where the names are found, and the red horizontal line at 50% indicates the whole-genome genus threshold. Within species comparisons ($\geq 95\%$ genome identity) are shown in orange, and between species comparisons ($< 95\%$ genome identity) are shown in purple.

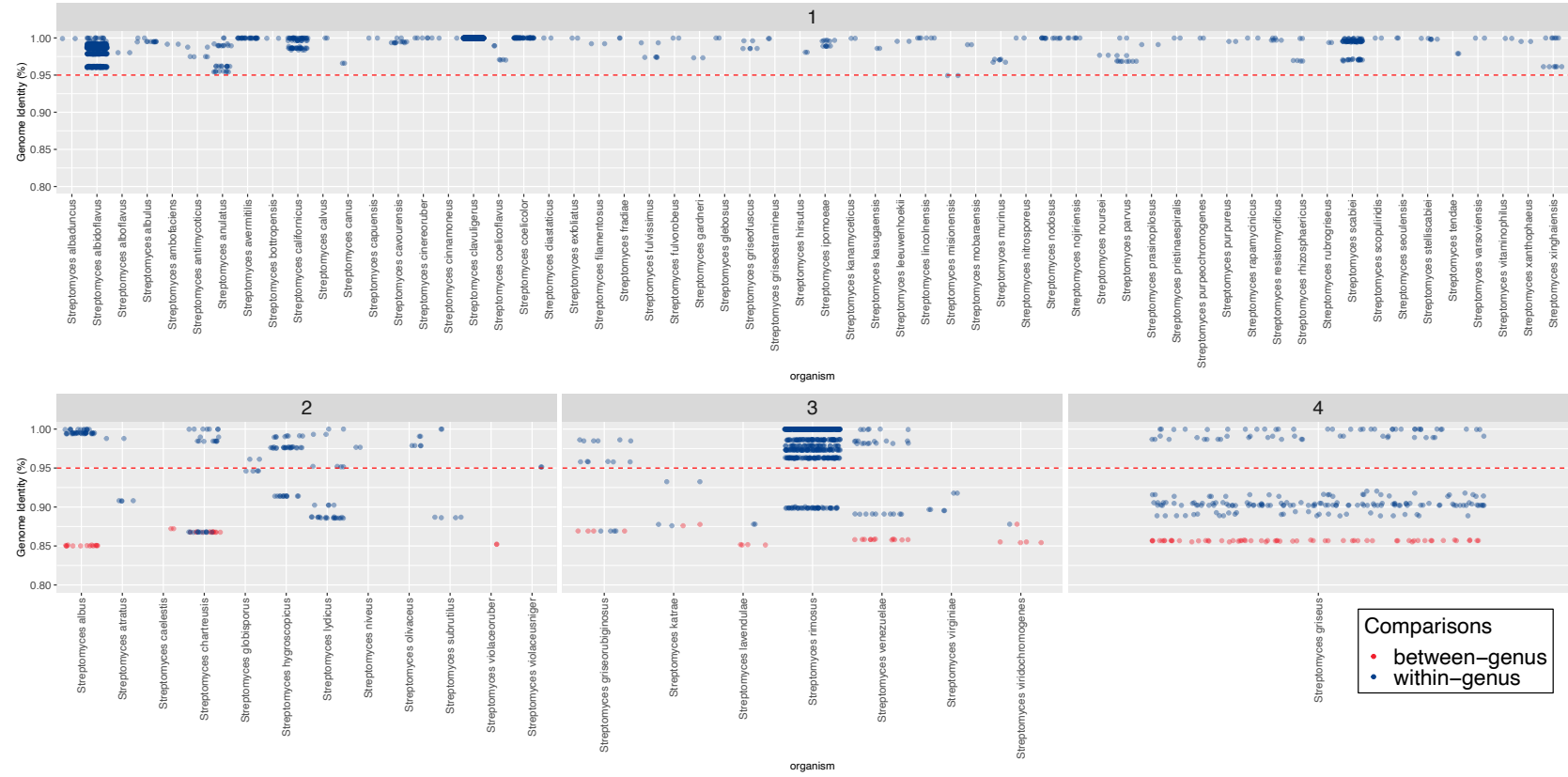


Figure 3.25: ANIm identity analysis of genomes sharing the same name other than *Streptomyces* sp. in NCBI. The numbers at the top of each plot correspond to the count of connected components where the names are found, and the red horizontal line at 95% indicates the whole-genome identity threshold. Within genus comparisons ($\geq 50\%$ genome identity) are shown in blue, and between genus comparisons ($< 50\%$ genome identity) are shown in red.

3.3.7 MLSA Phylogeny

The evolutionary relationships amongst all 873 *Streptomyces* with known STs were estimated using concatenated full length marker sequences, as an MLSA-derived phylogeny. A maximum-likelihood tree was inferred from the concatenated six full-length marker sequences. The complete alignment was trimmed to 10,368 nucleotides and consisted of 684 sequences after collapsing redundant sequences (Methodology Section 3.2.10). The MLSA tree has a total of 1366 internal nodes, of which 913 (66.8%) had Transfer Bootstrap Expectation (TBE) values of 100%, and 41 internal nodes with TBE values below 50% (3%), with 8% being the lowest TBE values observed (Figure 3.26). This suggests that the MLSA tree topology is quite robust.

I attempted to determine whether the subgraphs of the MST correspond exactly to divisions between clades in the MLSA tree. To accomplish this, I examined whether the 116 connected components of the MST that unite at least two genomes form monophyletic groups within the MLSA tree (Methodology Section 3.2.10). It should be noted that I am comparing connected components represented by at least two genomes, and not just the STs. This distinction is important because 20 singleton connected components were represented by multiple genomes. Additionally, 40 connected components exclusively consisted of non-Genbank represented STs, and could not be included in this analysis. I found that across 116 MST connected components uniting at least two genomes, 59 (50.9%) formed monophyletic groups on the MLSA tree. The largest group consists of as many as 57 genomes (Figure 3.26 - green), and all such monophyletic groups consist of a single genus and species ($\geq 50\%$ coverage; $\geq 95\%$ identity). However, I identified 57

(49.1%) cases where the MST divisions were found to be incongruent with the MLSA tree. Given the high TBE values supporting their placement splitting around the MLSA tree, it is likely that the discrepancies between the MLSA clades and MST components may be attributed to the lack of genomic representation for certain pubMLST STs in GenBank. This absence could be artificially linking otherwise distinct groups, leading to misleading interpretations of their evolutionary relationships.

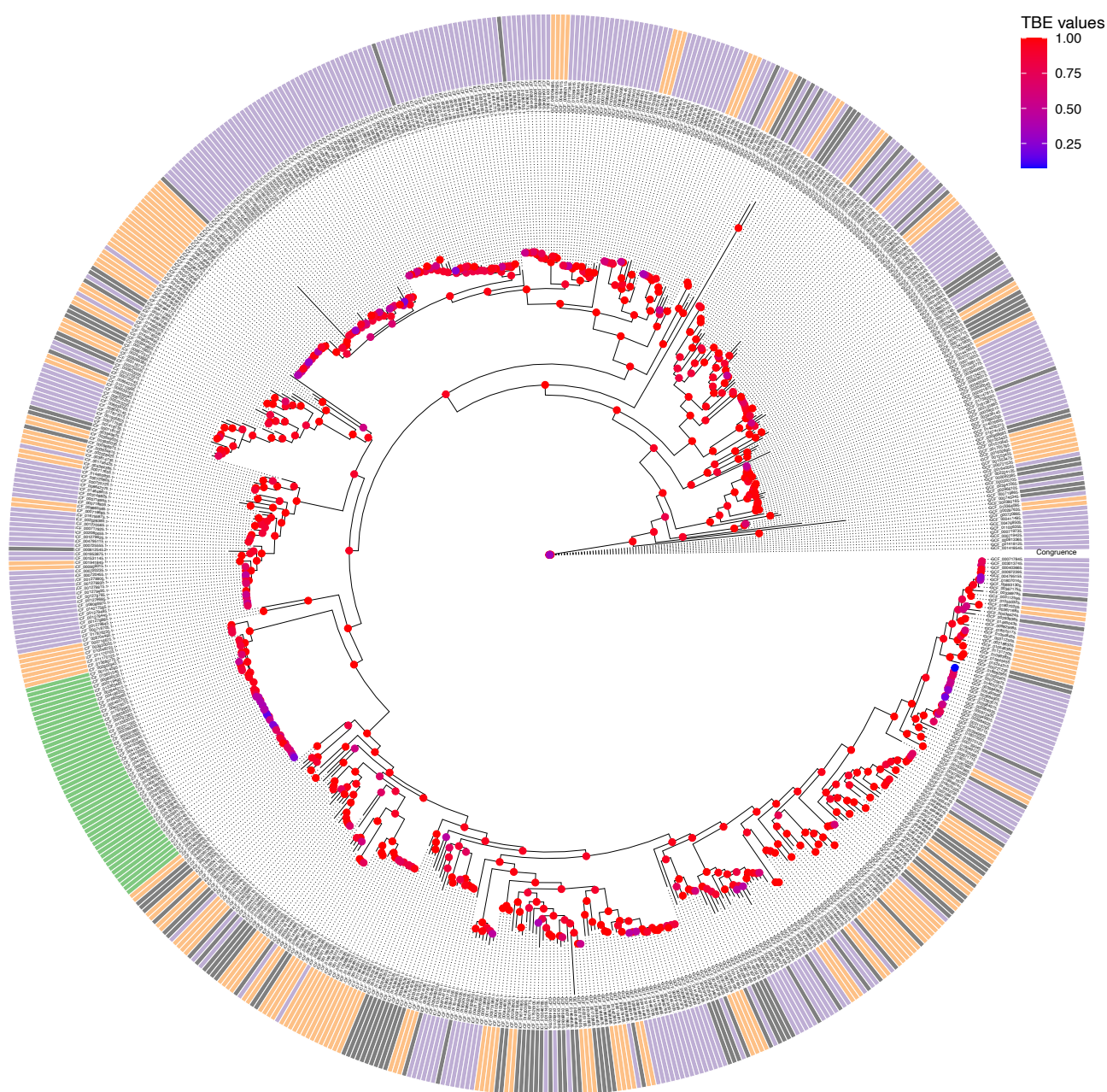


Figure 3.26: Midpoint rooted ML tree of six concatenated full-length allele sequences with mapped MST congruence. MST connected components which group as a single tree partition are shown in **orange**, except the biggest connected component forming a single monophyletic group which is shown in **green**. Genomes which are members of connected components that do not form a single partition are shown in **purple**, and **grey** correspond to connected components comprising of a single ST and represented by a single genome. The branch lengths of the tree were altered for visualisation purposes as some were obstructing a clear view of the clades.

3.3.8 Sensitivity test

I attempted to estimate the influence of each marker gene in the MLST scheme on classification - essentially, the sensitivity of classification to each individual marker sequence. This sensitivity test determined the number of times each pair of genomes was classified to the same, or to a different, ST when the marker sequence was removed from the scheme (Methodology Section 3.2.11). This allowed identification of assemblies that were previously assigned distinct STs, and would now be represented by the same ST after the marker is excluded. The analysis revealed that the classification of genomes was most affected by the removal of the *trpB* marker, resulting in the collapse of a total of 30 STs (Figure 3.31), followed by *recA* collapsing 13 unique STs (Figure 3.29), and 16S (Figure 3.27) and *rpoB* (Figure 3.32) collapsing 4 STs each, and *gyrB* (Figure 3.30) collapsing 3 unique STs. *atpD* showed the lowest sensitivity, collapsing only 2 STs (Figure 3.28 and Table 3.8). This finding suggests that *atpD* may not be essential for robust classification of *Streptomyces* genomes and raises the possibility of refining the MLST scheme by either excluding or replacing *atpD* marker with more informative gene.

Table 3.8: Summary of sensitivity test showing number of unique STs represented in GenBank after the exclusion of each marker.

Marker	Number of unique STs represented in GenBank after exclusion
16S	651
rpoB	651
atpD	653
gyrB	652
recA	642
trpB	625

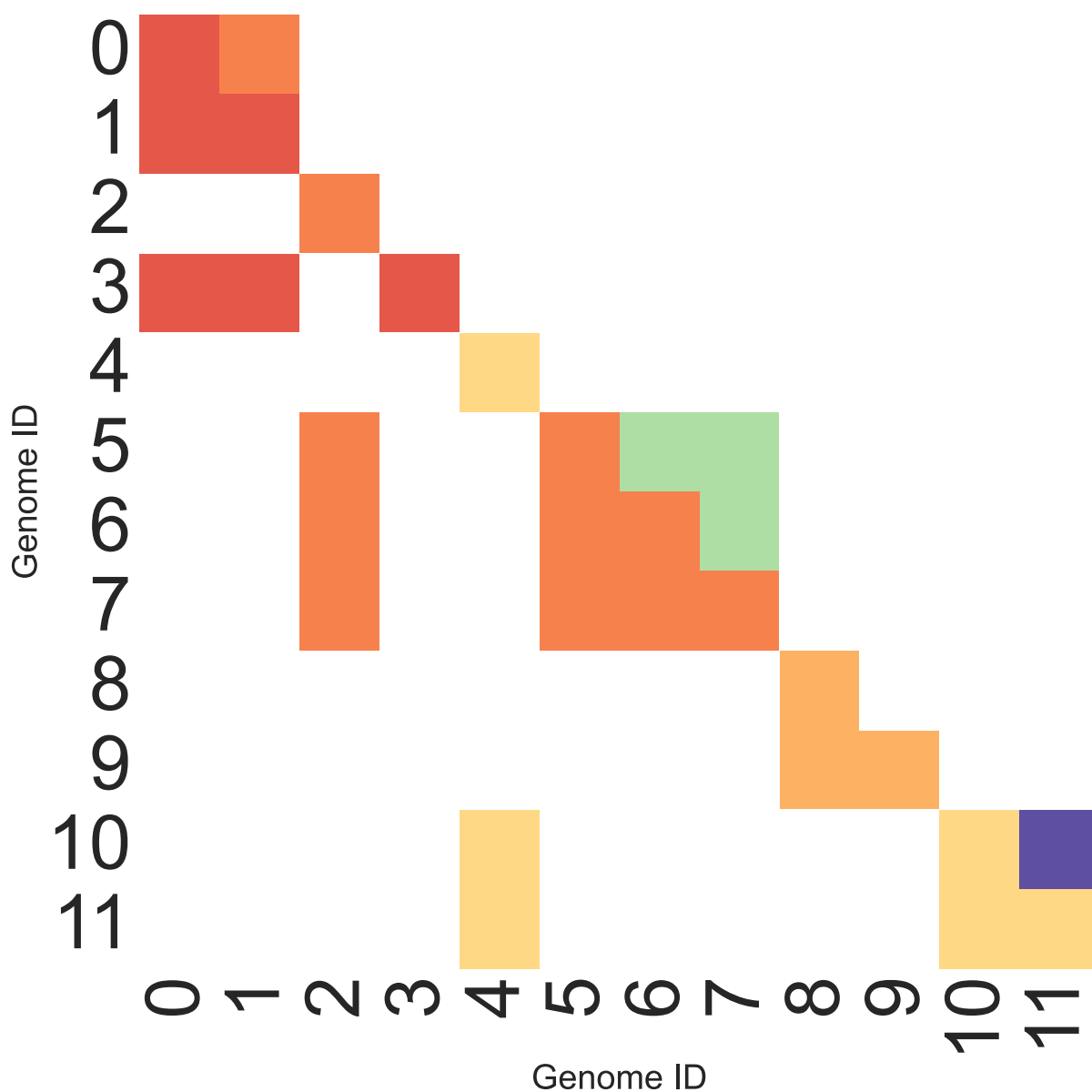


Figure 3.27: Heatmap of the sensitivity test results for *16S* marker gene. Each color in the heatmap represents distinct ST, with genomes sharing the same STs shown in the same color. The upper triangular region represents genomes sharing the same STs before the marker was excluded, and the lower triangular region represented genomes which are represented by the same STs after the exclusion of the *16S* gene from the scheme.

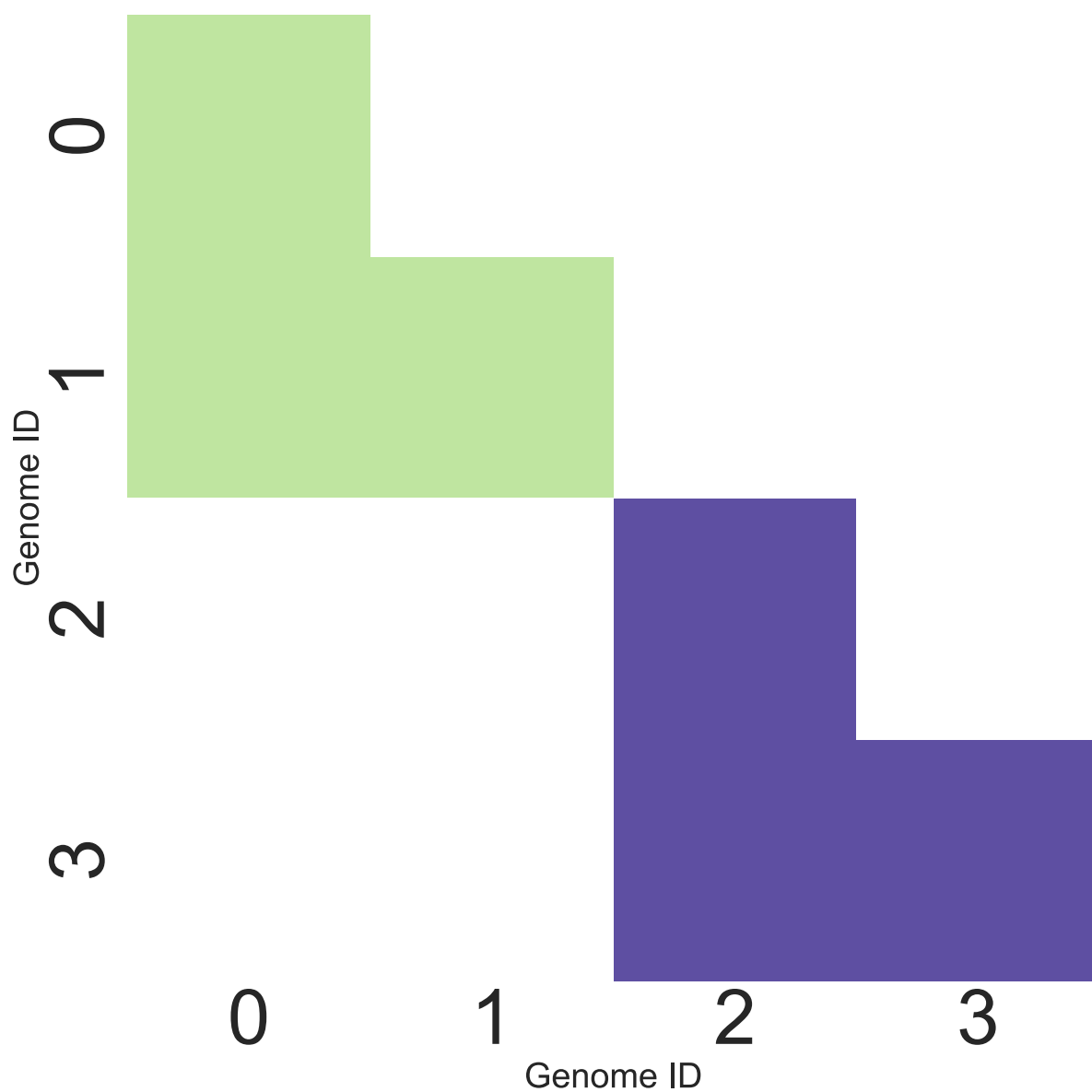


Figure 3.28: Heatmap of the sensitivity test results for *atpD* marker gene. Each color in the heatmap represents distinct ST, with genomes sharing the same STs shown in the same color. The upper triangular region represents genomes sharing the same STs before the marker was excluded, and the lower triangular region represented genomes which are represented by the same STs after the exclusion of the *atpD* gene from the scheme.

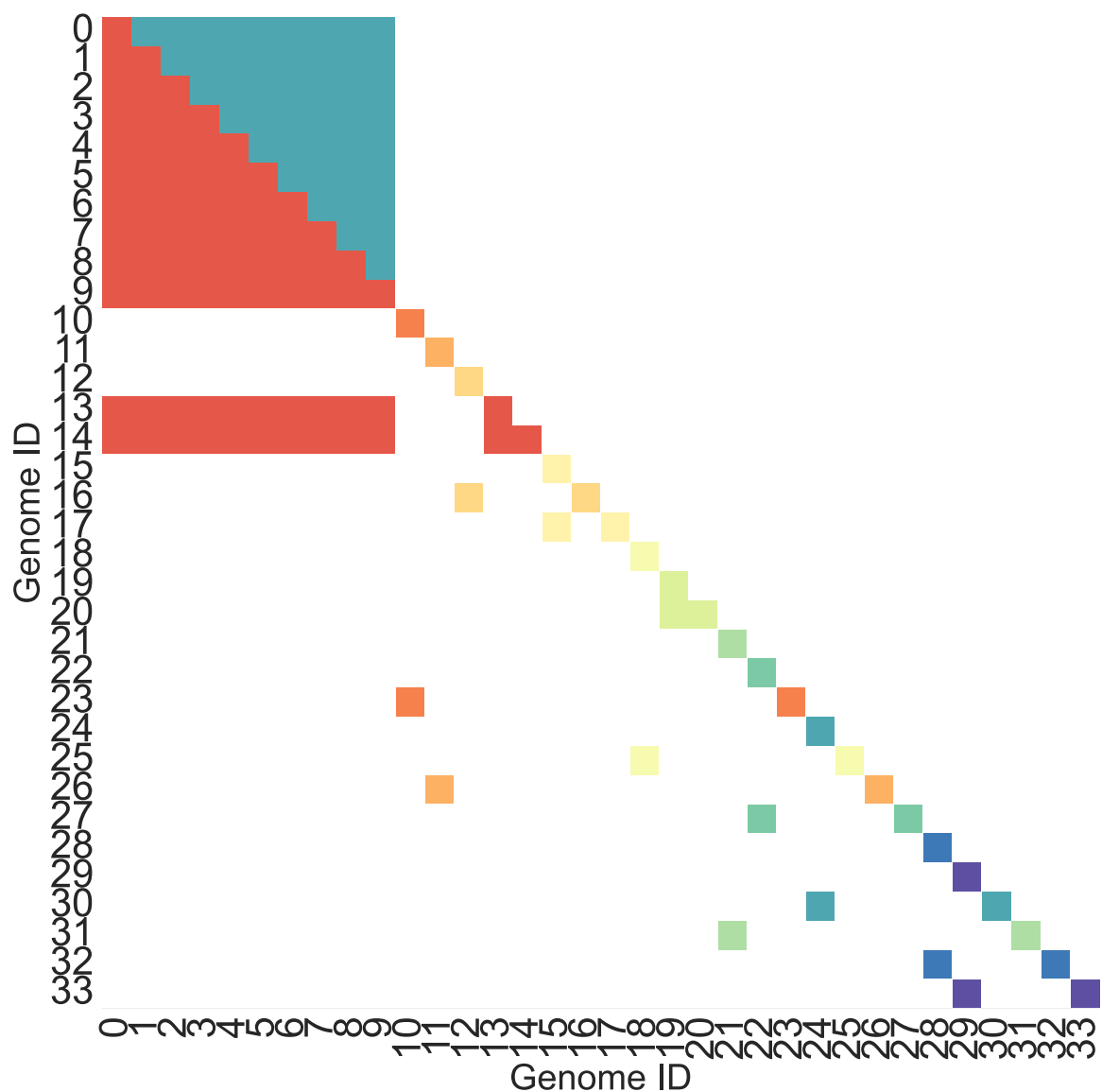


Figure 3.29: Heatmap of the sensitivity test results for *recA* marker gene. Each color in the heatmap represents distinct ST, with genomes sharing the same STs shown in the same color. The upper triangular region represents genomes sharing the same STs before the marker was excluded, and the lower triangular region represented genomes which are represented by the same STs after the exclusion of the *recA* gene from the scheme.

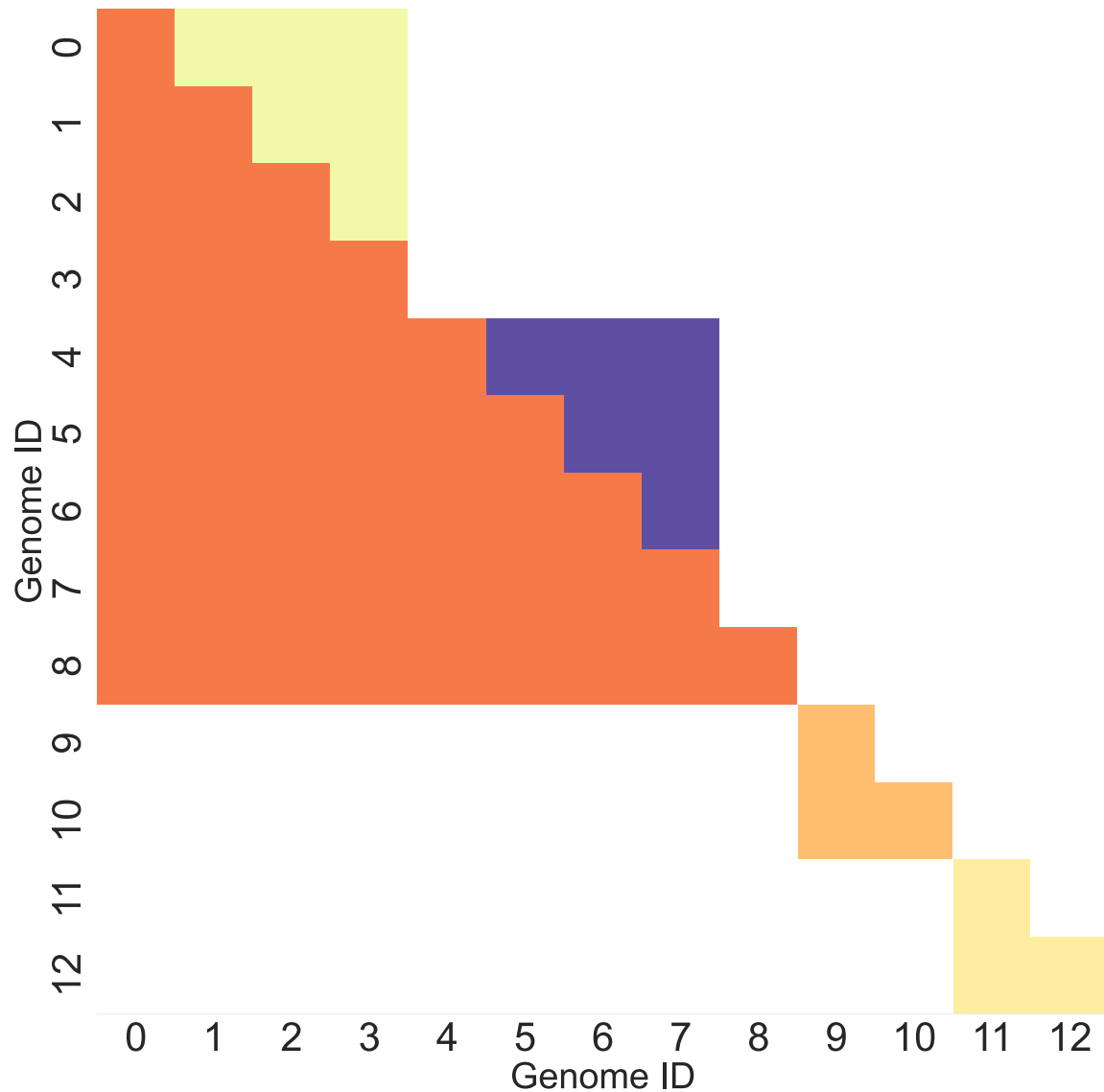


Figure 3.30: Heatmap of the sensitivity test results for *gyrB* marker gene. Each color in the heatmap represents distinct ST, with genomes sharing the same STs shown in the same color. The upper triangular region represents genomes sharing the same STs before the marker was excluded, and the lower triangular region represented genomes which are represented by the same STs after the exclusion of the *gyrB* gene from the scheme.

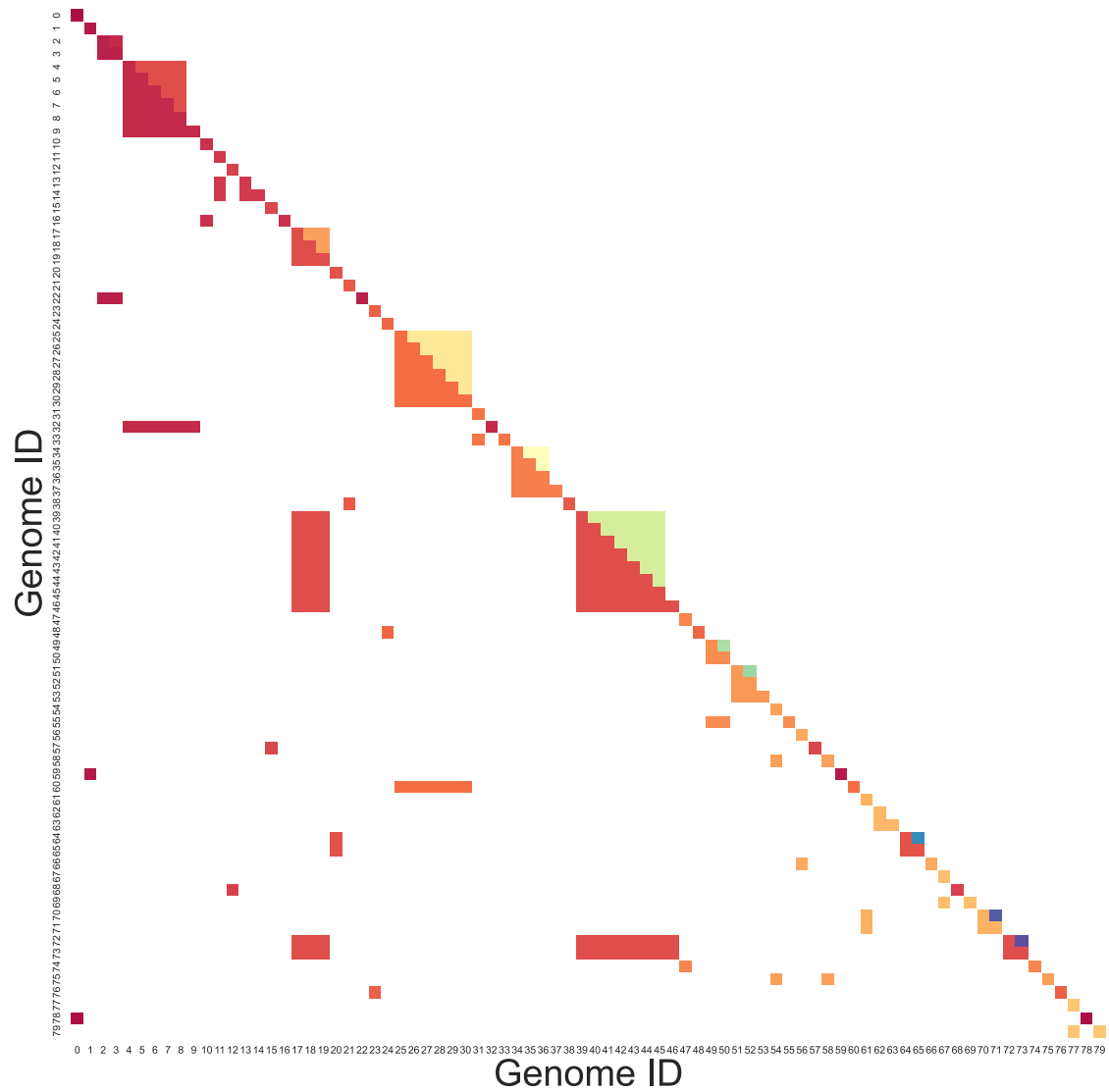


Figure 3.31: Heatmap of the sensitivity test results for *trpB* marker gene. Each color in the heatmap represents distinct ST, with genomes sharing the same STs shown in the same color. The upper triangular region represents genomes sharing the same STs before the marker was excluded, and the lower triangular region represented genomes which are represented by the same STs after the exclusion of the *trpB* gene from the scheme.

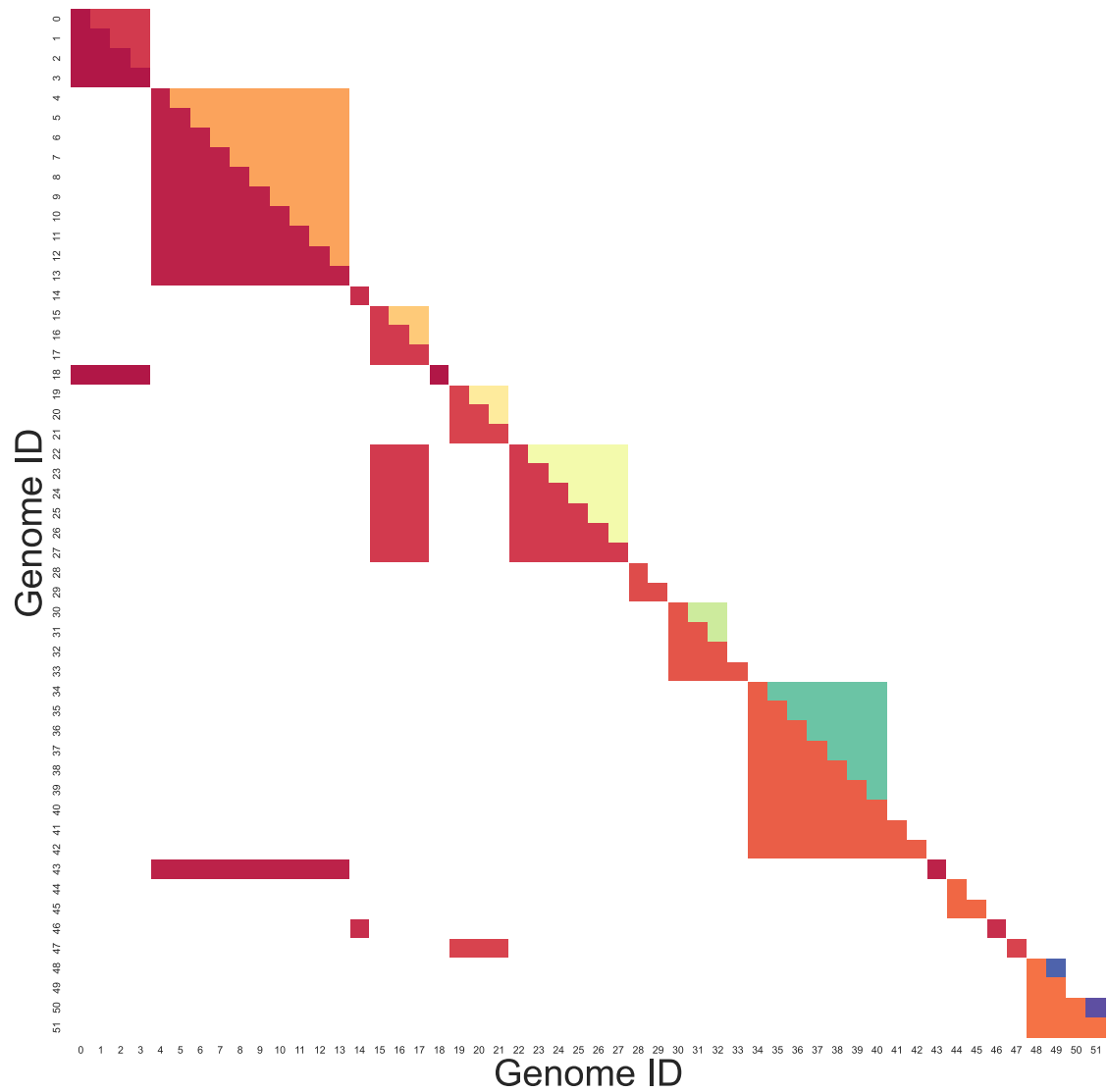


Figure 3.32: Heatmap of the sensitivity test results for *rpoB* marker gene. Each color in the heatmap represents distinct ST, with genomes sharing the same STs shown in the same color. The upper triangular region represents genomes sharing the same STs before the marker was excluded, and the lower triangular region represented genomes which are represented by the same STs after the exclusion of the *rpoB* gene from the scheme.

3.3.9 Sister genera

Using the approach described in methodology section 3.2.12, I found that only ST159, originally defined in the pubMLST *Streptomyces* scheme, is assigned to *Kitasatospora aureofaciens* (GCF_002082605.1, GCF_000978515.1 and GCF_000719175.1; `kitasatospora_MLST_output.txt;supplementary_file_4/output/sister_genera`). If the taxonomic assignments of these genomes are correct, it suggests that pubMLST STs not represented in GenBank as *Streptomyces* could potentially be reclassified outside of this genus.

In addition, I was able to identify a further five genomes currently assigned to the genus *Kitasatospora* in NCBI (GCF_026340775.1, GCF_026340263.1, GCF_024172095.1, GCF_001188955.3 and GCF_00716885.1) using the current set of markers in the *Streptomyces* scheme, as they possess all six loci. No STs for *Streptacidiphilus* and *Stoalloteichus* genera were represented under the *Streptomyces* pubMLST scheme, as none possessed all six loci. This suggests that without appropriate phenotypic checks and relying solely on MLST classifications for taxa identification, there is a risk of misclassifications.

3.4 Discussion

3.4.1 The canonical pubMLST Scheme is incomplete

After recognising the limitations of a single 16S marker for taxonomic classification and discrimination, as discussed in Chapter 2, a primary focus of this chapter was to investigate the congruence of MLST with other classification approaches, in the context

of *Streptomyces*. Prior to examining whether MLST could provide insights into the genomic diversity and evolutionary relationships within this pharmaceutically important genus, it became necessary to update the scheme. Only three new STs were added to the canonical scheme in the eight years prior to this study (Jolley et al., 2018), despite remarkable advancements in sequencing technology. As might expected, the canonical scheme was incomplete with respect to the current set of sequenced genomes, and I updated the scheme by identifying 568 novel STs and a total of 4,292 allele sequences (result section 3.3.1). This reflects a limitation of molecular typing schemes in public databases in that they are currently reliant on community annotation for timely updates. In general, the addition of more sequences to an MLST scheme is likely to lead to the discovery of more novel alleles. This exemplifies a different issue arising from the cumulative nature of the typing scheme, that if an organism within the *Streptomyces* genus possesses markers absent from the canonical scheme and significantly divergent from those represented, then without appropriate phenotypic checks, there's a risk of misclassification. Such an organism could be assigned to a different genus or potentially lead to incorrect proposal of a novel genus. Continuous automated or community maintenance of molecular typing schemes is necessary to avoid these issues.

One might argue that an MLST marker gene can be considered reliable and suitable for molecular typing, if it meets the following criteria:

1. Allele sequences have to be free from ambiguity, as a core principle of MLST is that a change in at least one base results in a new allele.
2. Each organism would contain a single copy of each marker allele to avoid ambiguous

interpretations.

3. Each marker would be present in all organisms of the target taxa enabling the study of all relevant isolates
4. Each marker would exhibit sufficient genetic diversity to enhance resolution to a specific taxonomic level, without overestimating or underestimating its diversity.

During the novel allele identification step for *Streptomyces* (methodology section 3.2.3), six novel allele sequences contained ambiguity bases and were not incorporated to the new scheme (result section 3.3.1). Such bases can arise from sequencing errors or the presence of mixed populations. While this poses a limitation, it is usually a technical challenge rather than a biological barrier. Furthermore, only six out of 1938 genomes were affected by the presence of ambiguity bases, indicating that the issue is not significant in scale for *Streptomyces*.

However, I found that the presence of multiple non-identical copies of marker genes could pose a larger challenge for MLST in *Streptomyces*. A total of 77 genomes could not be assigned STs due to the presence of multiple non-identical copies.

One genome (GCF_001550235.1; completeness 95.29%; contamination 6.48%) two non-identical copies of the *recA* gene, an unusual finding in *Streptomyces*, with only a single reported case of in *Streptomyces rimosus* (Mikoc et al., 2000). While it is possible that this genome contains two non-identical copies of *recA*, it could also be due to the genome's relatively low quality, as indicated by the checkM (methodology section 3.2.6). In addition, 77 genomes had multiple non-identical copies of the 16S gene. This result was expected, as approximately 69% of *Streptomyces* isolates with genomes assembled to

the chromosomal or complete level were found to possess multiple non-identical copies of the 16S rRNA gene, as discussed in Chapter 2. This suggests that the 16S rRNA gene is unsuitable as an MLST marker for the *Streptomyces* genus, as its variability prevents the scheme from encompassing all *Streptomyces* isolates.

Further limitations arise from the use of draft assemblies with missing 16S rRNA copies, which may be used to report novel STs. The problem arises due to potential variation between the reported copies and those which are absent from the assembly but present in the genome, which might result in the integration of non-existent STs into the scheme. For instance, a draft genome may contain only one (possibly chimeric) copy of the 16S rRNA gene, whereas the corresponding complete genome might possess six copies, all potentially non-identical, and perhaps different to that in the draft genome. Such circumstances could lead to the assignment of an ST that holds no biological relevance. I used a total of 462 genomes, collectively assembled to scaffold and contig level, to assign novel STs (result section 3.3.1), but there remains a possibility of missing copies of the 16S rRNA gene. Therefore, I cannot guarantee that these STs are reliable, as the missing copies of the 16S rRNA gene could differ from the ones that are currently present in the assemblies.

Another factor that might affect future reliability of an MLST scheme is the incorporation of genomes with multiple identical copies of a marker. During the extension step (methodology section 3.2.3), I identified 259 genomes that possess multiple identical copies of at least one marker gene (Table 3.3). While genomes with identical copies of marker genes can be assigned a ST now, future complications are a concern. Gene family expansions through duplication and horizontal gene transfer, followed by diversification,

are common in *Streptomyces* (Schniete et al., 2018b). This process generates new genetic material for natural selection to act upon. If one copy of a duplicated marker gene undergoes mutation while the other remains unchanged, or if both copies mutate but in different ways, this could lead to divergence between the two copies. At that point, the genome would contain non-identical versions of what was previously a single allele. This divergence could complicate and prevent the assignment of an ST in the future, as the current MLST scheme assumes that each marker gene represents a single allele per genome. Expanded gene copies can follow different evolutionary paths: they may be lost, preserved if they contribute to the organism’s fitness, or neofunctionalised into genes with altered functions (Birchler & Yang, 2022).

The markers chosen for an MLST scheme are typically differentially variable, and can show differing variabilities in different subgroups of the target clade. This leads to the question: how sensitive is the MLST scheme classification to variation in each marker. By carrying out a sensitivity test that dropped each marker from the scheme in turn, I identified assemblies that were assigned distinct STs in the complete scheme, but which would now be represented by the same ST after a marker was excluded (methodology section 3.2.11). If removing a single marker has little to no impact on the clustering of STs, it suggests that the marker may not be essential for distinguishing between STs, potentially indicating redundancy. On the other hand, if excluding the marker causes most STs to cluster together—resulting in the creation of a new ST every time the marker is included—it could indicate that the marker is introducing unnecessary complexity and making the scheme more difficult to manage. The sensitivity test revealed that *trpB* marker exhibited the most sensitivity by collapsing a total of 30

STs, but the it still did not affect the scheme in a significant scale. This indicates that markers with lower sensitivity could potentially be replaced with alternative markers to enhance the scheme's resolution in future applications.

3.4.2 The current set of MLST markers can lead to genus-level misclassifications

Traditionally, MLST schemes were designed to classify bacterial isolates by amplifying and sequencing specific genes through PCR, without requiring whole-genome sequencing. For instance, MLST schemes for *Brachyspira* spp. targeted genes such as *adh*, *est*, *gdh*, *glpK*, and *pgm*, allowing for bacterial typing without the necessity of sequencing entire genomes (Råsbäck et al., 2007). However, with the advent of whole-genome sequencing and the growing availability of genomic data, some MLST schemes have evolved, now being developed entirely based on sequenced genomes, such as that of *Enterococcus faecium*.

During the MLST scheme extension step (methodology section 3.2.3), I found that 150 of the 237 canonical pubMLST STs were not present among the sequenced genomes in NCBI. Each pubMLST scheme is typically linked to the relevant publications, providing users with detailed information about the development process, as in the case of *Brachyspira* spp. scheme which is linked to the Råsbäck et al., 2007 publication. This allows users to understand the steps involved in its creation, such as whether STs were derived from PCR amplification of isolates or associated with sequenced genomes.

However, there is no linked publication detailing the creation of the *Streptomyces* scheme, making it difficult to verify the origins of individual profiles. It might be possible that *Streptomyces* scheme was developed through PCR amplification of "isolates", raising

the possibility that some STs originated from co-cultures, or suffered from sequencing errors, potentially leading to chimeric sequences. This could explain why certain STs in pubMLST, which lack representation in GenBank, appear central in the network (result section 3.3.4 and Figure 3.8). If these STs arose from co-cultures, they might contain genetic material from multiple species, making them appear as key connectors that link diverse and potentially distinct STs, some of which may represent different lineages. This could result in these STs appearing central due to their artificial composite nature, even though they do not accurately reflect a single genetic lineage. Additionally, these isolates may have been sequenced only once, never sequenced again, or even lost over time, which could possibly explain the lack of representation of these STs in the NCBI database.

It is also possible that assemblies used to obtain the STs were suppressed or replaced due to changes in genome quality criteria, but it seems unlikely that this happened for as many as 150 genomes. Replaced genomes may differ in the sequences currently used for molecular typing, meaning different STs should be proposed to represent the isolate, or removed where appropriate.

However, there is currently no obvious mechanism to correct the pubMLST schemes in the case of retracted, suppressed or replaced genomes removing evidence for a ST (Jolley et al., 2018). In the case of suppressed genomes, the STs could be assigned to artefactual allelic profiles leading to false interpretation of novel taxa. Resequencing of type species from culture collections could be valuable in resolving any such issues.

At the time of conducting this study, 156 out of the 1,938 *Streptomyces* genomes used to update the current canonical *Streptomyces* scheme were suppressed by NCBI

since their initial download on the 9th July 2021 (Methodology section 3.2.4). Of these, 47 genomes were crucial in the assignment of novel STs, so these STs were removed from the updated scheme. Additionally, eight genomes were replaced, of which five were crucial in the assignment of novel STs. Updated versions of the replaced genomes were downloaded from NCBI on the 6th of September 2023, and I again ran MLST v2.22.0 with the updated scheme to ensure that they have the same ST as previously determined. Two genomes (GCF_012184365.1 and GCF_002879675.2) had two non-identical copies of a 16S rRNA marker gene, and corresponding STs were removed from the scheme, as no other genomes in NCBI represented the proposed allelic profile. These results demonstrate the importance of implementing automated measures to consistently revise schemes to resolve STs that may be affected by genome sequencing issues.

Another reason why some pubMLST STs are not supported by GenBank could be that MLST was unknowingly performed on non-axenic cultures, or STs have been unwittingly classified outwith the *Streptomyces* genus. If an organism from outside the *Streptomyces* genus possesses all the markers used in the *Streptomyces* MLST scheme, and these isolates were classified solely based on ST assignment to genomic data without phenotypic verification, it is possible that STs from isolates from distinct lineages could have been falsely included in the scheme. Using the approach described in methodology section 3.2.12, I found that, although it does not occur frequently, it is possible for non-Genbank represented STs to be classified outside of the *Streptomyces* genus. Specifically, ST159, which is now assigned to *Kitasatospora aureofaciens* (result section 3.3.9). This suggests that some *Streptomyces* STs might in fact correspond to organisms beyond the genus. In addition, further genomes currently assigned to the

genus *Kitasatospora* could now be classified using the current set of markers used in *Streptomyces* scheme, as they possess all six loci. The taxonomic status of the members of the genus *Kitasatospora* is controversial, as several isolates were transferred from and to the *Streptomyces* genus on several occasions (Kim et al., 2003; Wellington et al., 1992a; Zhang et al., 1997b). This highlights the difficulties in the family *Streptomycetaceae* (*Kitasatospora* and *Streptomyces*) given their common phenotypic and genomic features (Li et al., 2021). As sequencing technology advances, making genomic diversity-based computational assignments more feasible and likely to become the standard, the results indicate that relying solely on the current set of markers for identifying *Streptomyces* may be inadequate. These markers do not effectively differentiate organisms according to the currently accepted taxonomic boundaries within *Streptomycetaceae* and could lead to misclassifications.

3.4.3 Graph based analysis of STs subdivides *Streptomyces* into 278 distinct groups and which likely represent biologically-meaningful divisions

Kruskal's MST algorithm, applied using Hamming distance for matching alleles between STs, resolves the 805 STs into 278 subgraphs (connected components) (Figure 3.4). This suggests 278 distinct groups of *Streptomyces*, of varying size, in which no member ST of one subgraph shares any marker allele with members of any of the other subgraphs. This kind of distribution could reflect the true nature of diversity among *Streptomyces*, or be an artefact of sampling from a much larger, fully-connected *Streptomyces* population. It raises the question of whether the sampling of *Streptomyces* genomes affects the distribution of subgraph sizes (methodology section 3.2.7).

I reasoned that if the current set of sequenced *Streptomyces* genomes represented a monophyletic, closely-related group, I might expect to be able to traverse the entire set of STs, stepping from one ST to another by changing at most five alleles simultaneously. If, however, there were natural barriers to recombination, such as boundaries between taxonomic orders, lifestyle, niche occupancy, or other mode of isolation this might result in some STs not being reachable from all other *Streptomyces* genomes. Conversely, horizontal gene transfer including one or more markers might make the distance between two STs appear smaller than the true phylogenetic distance and "shortcut" some connections. A number of observations might be expected to follow from these possibilities. For example, the number of connected components might:

1. fall as the number of genomes increases - suggesting that at some point all components might eventually be joined into a single large group, or a small number of groups
2. increase to some limiting number as more genomes are added, indicating an asymptotic distribution of subgraph sizes. This might reflect natural barriers to recombination across *Streptomyces*, suggesting the discrete groups might have biological meaning
3. increase indefinitely as new genomes are introduced - suggesting that there might be much more more diversity available within *Streptomyces* than is currently represented in the genome data

My examination of how genome sampling affects the connectivity of the MST (methodology section 3.2.7) indicated that, in *Streptomyces*, the number of disjoint

components continues to increase (Figure 3.9). Also, there is no notable change in the distribution of relative sizes of connected components (Figure 3.10). These findings imply that the connected components are likely to remain disjoint even with the addition of more genomes and, where the inclusion of new genomes adds more diversity, it does so within the current distribution of disjoint groups. This result is unlikely to arise due to sampling issues, as simulation of a similar artificial scheme comprising 5000 sequence types (STs) representing a fully-connected MLST graph based on six marker genes indicated that increasing the number of STs led to a decrease in the number of connected components (Figure 3.11) and a reduction in their relative sizes (Figure 3.12). This suggests that the "true" graph of *Streptomyces* MLST sequence diversity is a set of disjoint subgraphs, not a fully connected graph.

3.4.4 MLST subgraphs do not generally correspond to *Streptomyces* taxa

As mentioned in section 3.4.2, there is currently a lack of published information regarding the *Streptomyces* pubMLST scheme, and it is unclear whether the scheme is able to provide subspecies resolution. If the scheme were intended to achieve this level of detail, we would expect to observe a one-to-one relationship where each ST represents a distinct species ($\geq 50\%$ coverage; $\geq 95\%$ identity), with potentially multiple STs associated with a single species. To establish whether this is the case for the current pubMLST *Streptomyces* scheme, I established taxonomic boundaries using ANI between genomes sharing identical STs (methodology section 3.2.9). I found that genomes sharing all six alleles in common are likely to represent the same genus and species, by whole-genome classification. This is supported by the observation that the lowest

genome coverage was 69.4% (Figure 3.14), and the lowest average nucleotide identity did not drop below 98.9% (Figure 3.15). These findings suggests that the current set of markers is unlikely to falsely describe distinct species under a single ST.

I also found that 103 species determined with ANI were represented by multiple STs, with one species being represented by as many as 48 distinct STs (Figure 3.21). This implies that the current scheme does provide subspecies resolution at least for some Streptomyces. This observation that multiple STs can exist for a single species could be a result of sequence diversity of the markers themselves being shaped by the selective pressures exerted by bioactive natural products that target specific loci used for MLST. This explanation is particularly intriguing when considering *rpoB*, *gyrB* and *recA* markers, which are frequently the target of naturally occurring antimicrobial compounds (Amusengeri et al., 2022; Chopra et al., 2012; Pavlopoulou, 2018). Similarly to 16S rRNA (given that many natural compounds target the ribosome (Hansen et al., 2003)) the selective forces acting on these loci (Chevrette et al., 2019a) can drive diversity in marker sequences, but also perhaps making them prone to horizontal transfer between distinct lineages - which would be a hindrance to their use in classification.

The division of *Streptomyces* into 278 components that share no marker allele with each other could imply a set of natural divisions between groups of isolates (discussion section 3.4.3). I applied ANI analysis to each connected component comprising at least two isolates (methodology section 3.2.9) to determine whether this set of natural divisions corresponds to a separation at genus or species level. If the current sets of markers are congruent with a taxonomic rank such as genus or species (or both), we might expect to find a single genus or species present in the same connected subgraph

of STs. I found that MLST subgraphs can unite genomes having various taxonomic relationships amongst their members. Specifically, 92% of non-singleton components unite a single genus ($\geq 50\%$ coverage), and 78% uniting only genomes from a single species (with approximately $\geq 50\%$ coverage and $\geq 95\%$ identity). However, 8% of connected components encompass more than one distinct candidate genus, with some components bridging as many as four distinct genus-level groups sharing as little as 13.81% genome coverage (Result Section 3.3.5). I also observed diversity at the species level within many connected components, with 22 components uniting multiple candidate species. Some subgraphs united as many as eleven distinct species, despite pairwise comparisons between some members sharing as little as 84.75% identity (result section 3.3.5). This suggests that the pubMLST STs absent from GenBank could originate from co-cultures or mixed populations, thereby linking genomes that are less related than implied by the MST. This is consistent with the observation that STs without representative genomes in GenBank exhibit a higher degree of connectivity (Table 3.5), as well as the empirical evidence suggesting that these STs are likely bridging groups of related STs (results section 3.3.4). Another possible explanation is that distantly-related organisms belonging to different species and even genera may still share a conserved marker forcing them to appear in the same connected component. This could be due to specific allele variants being transferred between distinct lineages if they enhance fitness under selective pressure. It could also be that inclusion of markers whose lineage does not reflect taxonomic divisions can also “force” dissimilar isolates together. In Chapter 2, I also found that distinct isolates at species and genus level confirmed with ANI (eg. $\geq 50\%$ coverage and $\geq 95\%$ identity) may share identical 16S rRNA sequences (results

and discussion section 2.3.4). In cases like this, individual markers can make isolates with distant common ancestors appear more related than they truly are.

3.4.5 Inconsistencies between MLST divisions and NCBI nomenclature

After confirming that all genomes represented by any single ST belonged to the same species according to ANIm (with $\geq 50\%$ coverage and $\geq 95\%$ identity) (results section 3.3.5), I examined their current taxonomic assignments. Given the high degree of genomic similarity among these genomes, one would expect consistent taxonomic names. I found that among genomes sharing the same ST and having multiple entries in NCBI, 79 cases (76.7%) had consistent taxonomic assignments. However, there were 24 instances where genomes with the same ST were assigned conflicting names, with up to three different species names despite confirming that they are, in fact, the same species based on the $\geq 50\%$ coverage and $\geq 95\%$ identity threshold. For example, ST 249 was represented by genomes labeled as *Streptomyces rimosus*, *S. capuensis*, and *Streptomyces sp.* These discrepancies may arise from different taxonomic assignment methods used during the sequencing period, leading to the same strain being classified under different species names depending on when the analysis was performed.

An interesting observation was made for ST167, the most represented ST in NCBI with 27 genomes, all consistently identified as *Streptomyces clavuligerus*. This consistency could be due to the industrial importance of *S. clavuligerus* for the production of clavulanic acid, prompting greater care in its taxonomic naming (or extensive resequencing of related isolates). A similar pattern was observed for ST168, represented nine times in NCBI, where eight out of nine genomes were consistently named *S. coelicolor*, a model organism for the genus (Bentley et al., 2002), with only one genome listed as

Streptomyces sp.

To further investigate the incongruities between whole-genome taxonomy, MLST divisions and nomenclature, I explored the distribution of genomes currently assigned the same names in NCBI across the MST (result section 3.3.6). I found that:

- 36.25% of the 80 distinct streptomycete names, excluding *Streptomyces* sp., exhibit identical sets of all six alleles, thereby being represented by the same ST.
- 38.75% of streptomycete names exist within the same connected group of STs but possess multiple allelic profiles.
- 25% of the names are distributed among disconnected groups of STs, ie. they do not share any allele sequences in common.

ANIm analysis was used to determine whether the inconsistencies between nomenclature and their distribution across the MST for these genomes were the result of misassigned nomenclature. Where the current set of markers does not capture the true diversity of *Streptomyces*, it would be expected to see the same taxa ($\geq 50\%$ genome coverage for genus; $\geq 95\%$ genome identity for species) being split across disconnected groups of STs. If, however, the inconsistencies were the result of taxonomic misassignment at genus level it would be expected that isolates with the same name would share less than 50% of their genome by alignment length (or if the taxonomic misassignments occurred at species level, the genome identity of $< 95\%$) (result section 3.3.6).

As discussed in section 3.4.4, all isolates sharing the same name in NCBI and belonging to the same connected group of STs likely represent the same species. However, I observed that ten names, despite not sharing any markers in common and therefore

being split across multiple connected components, were likely to represent the same genus. This inference is supported by their genome coverage remaining above the adapted 50% threshold. Among these names, three (*S. niveus*, *S. olivaceus*, and *S. violaceusniger*) were likely to be the same species, as they shared 95% or more of their genome identity (Figure 3.24 and 3.25). These confirm that it is possible for members of the same genus and even species to not share any marker alleles with each other. This highlights that the genomic diversity provided by the combination of markers does not contain sufficient resolution to capture the true diversity of *Streptomyces*. It is also possible that genomes that could potentially link these STs have not yet been sequenced, which may create a false impression of greater divergence than truly exists.

In my analysis, I have also uncovered nomenclature inconsistencies. I identified seventeen instances where genomes, despite sharing the same name in NCBI, possessed relatively low pairwise genome identity, reaching as low as 85%. Similarly, ten clusters, also sharing identical names in NCBI, are unlikely to represent the same genus, given their genome coverage drops as low as 13.1% (Figure 3.24 and 3.25). This implies that significant reclassification within the *Streptomyces* genus is not only needed but long overdue to accurately reflect the genetic diversity and relationships among these organisms.

To further examine the nomenclature issues within *Streptomyces* genus, I investigated the nomenclature diversity for each identified species ($\geq 50\%$ genome coverage; $\geq 95\%$ genome identity) per connected component (Result section 3.3.5; Figure 3.22). To test if there is agreement between the current taxonomic classification and nomenclature, isolates belonging to the same species should share the same name. Among the 295

identified species, 156 were represented by a single genome, leaving no opportunity for confusion. Among the remaining 139 species with multiple representations in the NCBI, only 66 (47.5%) showed agreement between whole-genome distance methods and nomenclature, while the remaining 73 (52.5%) did not. As discussed in section 3.4.4, I also identified cases where genomes, despite sharing identical allelic profiles, 50% genome coverage, and 95% genome identity, were assigned conflicting names at NCBI (result section 3.3.5, Figure 3.14 and Figure 3.14). Taken together, these findings indicate that the nomenclatural adjustments in the public record of Streptomycete genomes may be needed to avoid undesirable clinical, ecological, agricultural, and pharmaceutical consequences (Boykin, 2014; Janda, 2020).

3.4.6 Clades in MLSA phylogeny largely do not correspond with MLST subgraphs

I constructed a phylogenetic tree from all six MLST allele sequences from each genome (result section 3.3.7) to explore the congruence between MLST subgraphs and phylogenetic clades, which were inferred from concatenated six full-length allele marker sequences. This involved examining whether the 116 connected components of the MST, which unite at least two genomes, formed monophyletic groups within the MLSA tree, an approach that, to my knowledge, has not been previously applied in other taxa.. While I found that in 50.9% of cases, the MLST subgraphs formed monophyletic groups, with some comprising as many as 57 genomes, there were instances, accounting for 49.1% of cases, where incongruities between these MLST subgraphs and the MLSA tree occurred. Given that most clades in the inferred phylogeny are well-resolved, with 66.8% of all clades having a TBE value of 100%, and only 3% of clades in the MLSA tree having a

TBE value below 50% (result section 3.3.7, Figure 3.26), the topology of the MLSA tree can be considered robust, and supports the split of the MST connected components across the MLSA tree. The lack of one-to-one map between the MLSA clades and MST connected components could arise, as the result of non-Genbank represented STs that might inappropriately link groups of genomes that are not as related as the MST would imply. This is consistent with the finding that STs lacking a representative genome in GenBank exhibit a higher degree of node connections (Table 3.5) and the empirical test that indicates that STs lacking a representative genome in GenBank are likely uniting groups of STs (result section 3.3.4). Another possibility is that some markers might be prone to horizontal gene transfer. As mentioned in discussion section 3.4.5, certain markers like *rpoB*, *gyrB*, and *recA* are frequently targeted by naturally occurring antimicrobial compounds, which may contribute to these discrepancies (Amusengeri et al., 2022; Chopra et al., 2012; Pavlopoulou, 2018). Genomically distinct taxa that share the same ecological niche may be exposed to selective pressures exerted by bioactive natural products, which can drive gene transfer as a survival strategy. This is especially likely if one organism possesses a marker variant that grants resistance to the bioactive compounds present, thereby allowing it to survive better under those conditions.

Genomic insights into *Streptomyces* phylogeny, taxonomy and structure

4.1 Introduction

4.1.1 Motivation

Reclassification of *Streptomyces* to better understand their evolutionary relationships has been extensively explored in previous chapters. Both the 16S (Chapter 2) and MLST (Chapter 3) classification methods have shortcomings when for identification of useful operational taxonomic units for pangenomic and comparative analyses. These methods were found to be incongruent with whole-genome distance methods, leading to ambiguous interpretations for divisions of *Streptomyces*. As outlined in Chapter 1, whole-genome classification approaches have been developed that provide more accurate and reliable taxonomic assignments, including isolates belonging to the phylum *Actinobacteria* (Nouioui et al., 2018). This, and my previous observations, were a motivating factor for investigation of the taxonomic structure of *Streptomyces* using whole-genome classification methods. The main questions I attempt to answer in this chapter are:

1. What resolution does the current whole-genome classification approach offer?
2. Are whole-genome phylogenies congruent with whole-genome distance methods?
3. What useful information about *Streptomyces* can whole-genome taxonomy methods reveal?
4. Can whole-genome taxonomy methods help in identifying useful taxonomic units for pangenomic analysis by including closely related isolates with a consistent definition of "closely related"?

4.1.2 Aims and Objectives

The objectives of this chapter are as follows:

1. I will select a representative sample of *Streptomyces* genomes. This sample will aim to capture the full diversity of *Streptomyces* genomes used in this thesis while excluding poor quality genomes. To achieve this, I will assess the quality of *Streptomyces* genomes and select a set of representative genomes based on prior MLST analyses.
2. To address the limitations of previous reclassification efforts that were found incongruent with whole-genome distance methods and failed to align with natural divisions within the *Streptomyces* genus, I will identify the core genome (genes shared across all 295 representative *Streptomyces* genomes) of *Streptomyces* using graph-based methods and use single-copy orthologues to establish evolutionary relationships within the genus.

3. Given the current lack of consensus in interpreting genus boundaries through whole-genome coverage and identity, I will attempt to establish genus boundaries for *Streptomyces* using ANI, statistical methods, and graph theory.
4. Having established the genus boundaries, I will evaluate the congruence between whole-genome distance methods and whole-genome phylogenies.
5. Previous studies have observed that genes located in the core of *Streptomyces* (i.e., those shared across all investigated genomes) tend to be situated towards the center of the chromosome (Bury-Moné et al., 2023). I will investigate whether this pattern holds true for my set of *Streptomyces* genomes by examining the chromosomal locations of SCOGs.
6. Given that *Streptomyces* are known to experience frequent recombination and high levels of HGT (Zhou et al., 2012), I will investigate the distribution of SCOGs nucleotide variants within on the SCOG phylogenetic tree. This analysis will aim to identify potential evidence of HGT by examining patterns of nucleotide variation and clustering within the tree.

4.2 Methodology

4.2.1 Data retrieval and availability

All raw data, supporting data, and code used in this chapter are publicly available on GitHub (https://github.com/kiepczi/Kiepas-et_al_2024_SCOG).

4.2.2 Taxonomic sampling

In Chapter 3, I studied the taxonomic structure of *Streptomyces* based on MLST analysis. I identified 873 publicly available *Streptomyces* genomes containing all six MLST markers currently used in the pubMLST *Streptomyces* scheme (https://pubmlst.org/bigsdb?db=pubmlst_streptomyces_seqdef) (Jolley et al., 2018) in a single copy variant. Although the MLST divisions of *Streptomyces* do not obviously correspond directly to any particular natural taxonomic division, other than that each ST appears to be found in only one species (Section 3.3.5), this is a convenient classification for reducing the number of genomes for more extensive whole-genome study. I used this classification to identify a representative subset of 295 *Streptomyces* genomes (`representative_genomes.csv`; `supplementary_file_2`) to create a representative sample of *Streptomyces* diversity of a convenient size for computational analysis. For each ANI species found in a connected group of STs (Figure 3.20) a single representative genome was selected. In all cases, the assembly with highest genome completeness and lowest genome contamination, as identified by checkM v1.2.2 (Parks et al., 2015) (`01_genome_quality_assessment.sh`; `supplementary_file_2`) was chosen. The term species is used here to refer to assemblies sharing a minimum genome coverage of 50% and genome identity of 95%, as determined by ANIm analysis with pyANI v0.3 (Pritchard et al., 2015).

4.2.3 Nomenclature status

The NCBI taxonomy assigned to all 873 genomes at the species level was validated against the List of Prokaryotic names with Standing in Nomenclature (LPSN),

downloaded on February 16, 2023 (Supplementary File 13).

4.2.4 Identification of Orthogroups

Orthogroups shared by all 295 representative genomes were identified using Orthofinder v2.5.4 (Emms & Kelly, 2019) (`02_run_orthofinder.sh`; `supplementary_file_4`). Genes from orthogroups common to all 295 (100%) genomes regardless of their duplication status were categorised as core-genes, while the set of genes that were shared across all 295 representative genomes and have remained in a single copy (without duplications) were considered as Single Copy Orthologues (SCOGs). This strict definition of core genes was chosen to quantify the genomic similarities within *Streptomyces* by capturing only universally conserved genes, minimising the impact of lineage-specific gene loss. Genes from orthogroups represented in 280 to 294 genomes ($95\% \leq \text{genomes} < 99.9\%$) were classified as soft-core genes, which were also considered in a more specific context to highlight genes with a high level of conservation, but still allowing for minor variability across genomes. Finally, genes found in orthogroups represented in 44 to 279 genomes ($15\% \leq \text{genomes} < 95\%$) were classified as shell genes, and those in fewer than 43 genomes ($< 15\%$) were labeled as cloud genes.

4.2.5 Single Copy Orthologue phylogeny

SCOGs protein and nucleotide sequences were extracted from GenBank files by matching the appropriate protein IDs (`01_extract_protein_and_nucleotide_sequences.py`; `supplementary_file_5`). During the extraction it was brought to my attention that a total of 14 SCOG sequences appeared in multiple copies in some genomes. It was confirmed manually that these corresponded to Identical Protein Groups (IPGs) with identical protein and nucleotide sequences. Therefore, a single representative for each

IPG was chosen for further analyses. The extended information of such sequences is available in `IPG.csv` in `supplementary_file_5\output\additional_information`.

A phylogenetic reconstruction was estimated for all 295 representative genomes using concatenated sequences of all 137 SCOGs identified in Section 4.2.4. SCOG protein sequences were aligned separately using MAFFT v.7407 (Katoh & Standley, 2013) (`02_align_SCOGs.sh`; `supplementary_file_5`), and their corresponding nucleotide sequences were backthreaded onto the protein alignments using T-coffee v12.007bf08c2 (Notredame et al., 2000) (`03_backthread.sh`; `supplementary_file_5`). The nucleotide alignments were then concatenated (`04_concatenate_alignments.py`; `supplementary_file_5`) and gaps were removed using trimAl v.1.4.rev15 (Capella-Gutiérrez et al., 2009) (`05_remove_gaps.sh`; `supplementary_file_5`). To account for varying evolutionary processes across different SCOGs, ModelTest-NG v.0.1.7 (Darriba et al., 2019) (`07_evolutionary_model_test.sh`; `supplementary_file_5`) was used to predict the most appropriate evolutionary model for each alignment partition, selecting distinct models for each SCOG. A Maximum-Likelihood tree with 100 Transfer Bootstrap Expectation (TBE) replicates was inferred using RAxML-NG v.1.0.3 (Kozlov et al., 2019) on the ARCHIE-West computer cluster with Intel(R) Xeon(R) Gold 6138 CPU 2.00GHz, 40 cores and 188 GB RAM.

4.2.6 Testing congruence of ANIm taxonomic boundaries and SCOG phylogeny and identification of genus boundaries

ANIm analysis using pyANI v0.3 (Pritchard et al., 2015) (`04_run_anim.sh`; `supplementary_file_3`) was performed to determine taxonomic boundaries for all 295 representative genomes. I primarily wished to test the monophyly of *Streptomyces*

subgroups inferred based on ANIm analysis by mapping their distribution on the SCOG phylogeny estimated in section 4.2.6. To do this, I constructed a network using NetworkX v2.3.6 (Hagberg et al., 2008) representing genomes as nodes, and assigning edges between genomes with weights corresponding to the lowest genome coverage and average genome identity of each pairwise comparisons. Given the current lack of consensus on interpreting genus boundaries using whole-genome coverage, I tested a range of genome coverage thresholds from 40% to 80%, in steps of 0.1%. Edges falling below a specified threshold were removed, and the resulting graph components were individually examined. If non-cliques were formed in a subgraph, I removed edges with the lowest genome coverage until cliques (k-complete graphs) were formed. The subsequent subgroups generated for each starting genome coverage threshold were then examined to determine which (if any) formed monophyletic groups on the SCOG phylogenetic tree, using the ete3 (Huerta-Cepas et al., 2016) Python module (`01_assign_genus_IDs.ipynb`; `supplementary_file_9`).

To estimate robust genus boundaries, I predicted likely genome coverage and genome identity thresholds corresponding to a discontinuity in pairwise genome similarities by fitting piecewise linear regression to these values for all pairwise comparisons obtained from ANIm analysis. The number of optimum segments for the piecewise linear regression was determined based on the Bayesian Information Criterion (BIC) with piecewise-regression (aka segmented regression) (Pilgrim, 2021) using a Python module run across segment values ranging from one to ten (`07_piecewise_linear_regression.ipynb`; `supplementary_file_3`). I used the identified thresholds to determine an optimal number of subgroups in the genus *Streptomyces*. Members of each subgroup were identified

though network analysis involving removal of edges corresponding to pairs of genomes with genome coverage of $<45.9\%$ or genome identity $<86.8\%$ (or both), and by resolving non-cliques as described above (`01_assign_genus_IDs.ipynb`; `supplementary_file_9`).

4.2.7 SCOG location on the chromosome

Placement of SCOGs required fully assembled genomes, as partial genomes lack the necessary continuity to map gene position reliably across the chromosome. Therefore, to determine the location of all SCOGs on the chromosome, I initially identified a subset of 63 complete genomes (`streptomyces_complete_genomes.txt`; `supplementary_file_7\output`) from the previous set of 295 representative genomes (`01_get_general_information_genomes_and_SCOGs.py`; `supplementary_file_7`). I considered a genome to be "complete" if it was assembled to the complete or chromosomal level according to NCBI.

For each complete genome, I first checked the location of the *oriC* gene by determining the location of the first base of the *dnaA* gene (`02_check_oriC_location.py`; `supplementary_file_7`). In *Streptomyces*, the *oriC* is typically located near the midpoint of the linear chromosome (Bentley et al., 2002). However, three of the complete genomes (GCF_009834125.1, GCF_000147815.2 and GCF_018128905.1) had *oriC* located near the telomere of the chromosome. Due to the lack of complete assembly information regarding the chromosome structure (eg. linear or circular), these genomes were excluded from further analysis. The remaining 60 complete *Streptomyces* genomes were then reorientated based on the *oriC* region (`01_reorientation_of_genomes_oriC.py`; `supplementary_file_14`). In this step, the chromosomes were adjusted so that the *dnaA* gene was positioned on the positive strand. Following this reorientation of the chromo-

somes, I determined the locations of the predicted SCOGs (`03_get_SCOGs_location.py`; `supplementary_file_7`) with their positions calculated as a percentage along the chromosome from one end (left) to the other (right).

4.2.8 Distribution of SCOGs on the phylogeny

For each individual SCOG nucleotide variant, I checked its distribution on the SCOG phylogeny, identifying genomes that shared identical nucleotide sequences and assessing whether they formed monophyletic groups on the phylogenetic tree using the `ete3` (Huerta-Cepas et al., 2010) Python module. Additionally, I checked if the non-monophyletic SCOGs variants were scattered across multiple candidate *Streptomyces* subgroups identified in section 4.2.6 (`01_check_HGT_with_ete.ipynb`; `supplementary_file_10`).

4.3 Results

4.3.1 Representative set of genomes

The 873 *Streptomyces* genomes carrying all six MLST markers were found to be assembled to different levels of completeness. 251 genomes assembled to scaffold level, 466 to contig level, 141 to complete and 15 to chromosomal level. The genome sizes of the 873 *Streptomyces* assemblies ranged from 5.77Mb (GCF_000297635.1) to 13.13Mb (GCF_016741935.1) with an average size of 8.43Mb per genome. Detailed information about each genome's ST, completeness level and genome size is provided in `all_genomes_additional_data`; `supplementary_file_2\output`.

The number of predicted genes ranged from 5,216 to 10,983 total genes, and from 5,153 to 10,906 total protein coding sequences. The average number of genes per genome

was 7,534 with an average of 7,450 total protein coding sequences per genome. The number of plasmids per genome varied from 0 to 5. As is typical, all 873 *Streptomyces* genomes used in this analysis have a high G+C% content, ranging from 67.9% to 73.9% with an average of 71.9% per genome. The taxonomic nomenclature assigned to each genome in the NCBI database was validated against the List of Prokaryotic Names with Standing in Nomenclature (LPSN), a database that catalogues the validly published names of prokaryotes in accordance with the rules of the International Code of Nomenclature of Prokaryotes (ICNP) (section 4.2.3). This validation revealed that only 401 (45.93%) genomes were assigned a validly published name, while an additional 33 (3.78%) genomes were assigned synonyms of valid names. A total of 417 (47.77%) genomes were labeled as *Streptomyces sp.*, and 22 (2.52%) genomes had no corresponding record in the LPSN database.

I assessed the quality of these 873 *Streptomyces* genomes, finding that the completeness ranged from 85.85% to 100%, and contamination ranged from 0% to 17.67% (`quality.txt`; `supplementary_file_2`). To create a representative sample of *Streptomyces* that covers the entire diversity of *Streptomyces* genomes used in this thesis, without compromising the research outcomes by inclusion of poor quality genomes, I selected the best-quality (i.e. most complete, least contamination) genome assemblies from each species (as determined by ANIm with $\geq 50\%$ coverage; $\geq 95\%$ identity) found in each subgraph of the MST in figure 3.20 for further analysis (Section 4.2.2). Following this filtration process, 295 representative genomes were taken forward for further whole-genome analyses with completeness ranging from 93.99% to 100% and contamination ranging from 0% to 4.21%. The `all_genomes_additional_data.csv`

(`supplementary_file_2`) file provides extended information about all 873 genomes, and corresponding extended information about the representative genomes is provided in `representative_genomes.csv` (`supplementary_file_2`).

4.3.2 General features of *Streptomyces* pangenome

To obtain an overview of *Streptomyces* genome contents, orthogroups for all 295 representative genomes were calculated using Orthofinder (methodology section 4.2.4). This analysis identified 28,900 groups containing a total of 2,120,506 genes; 19,663 genes remained unassigned. The number of orthogroup genes per assembly ranging from 4,934 to 10,219, with the average orthogroup count being 7,188 per genome. The core genome (at least one gene in each of the 295 genomes) contains 463 orthogroups (222,300 genes), of which 137 (40,415 genes) are single-copy orthogroups (SCOGs) represented by only one gene in each assembly (Table 4.1). The majority of orthogroups, 21,346 in total, were categorised as cloud orthogroups (present in fewer than 43 genomes), with a total of 204,610 genes (Table 4.1).

Table 4.1: Summary statistics of *Streptomyces* pangenome.

Type	Number of genomes	% of genomes	Gene count	Orthogroup count
Core	295	100%	222,300	463
SCOGs	295	100%	40,415	137
soft-core	$280 \leq \text{genomes} \leq 294$	$95\% \leq \text{genomes} \leq 99.9\%$	760,193	1,980
shell	$44 \leq \text{genomes} \leq 279$	$15\% \leq \text{genomes} \leq 95\%$	933,403	5,111
cloud	>44	$>15\%$	204,610	21,346

4.3.3 Single Gene Copy Orthologue Phylogeny

For phylogenetic tree estimation, I constructed a codon aware multiple sequence alignment (MSA) of all 137 SCOGs sequence groups. Nucleotide sequences were backthreaded onto the protein sequence alignments, then concatenated and partitioned as described in methodology section 4.2.5. The final trimmed MSA consists of 97,464 nucleotide bases (`no_gaps_concatenated_sco.fasta`; `supplementary_file_5/output/alignments/concatenated`). I estimated a maximum-likelihood (ML) tree with 100 transfer bootstrap expectation values (TBE), fitting individual substitution models to each SCOG (Section 4.2.5). To my knowledge this is the most comprehensive whole genome-based phylogeny of the entire genus *Streptomyces* attempted to date. The tree contains 588 internal nodes (Figure 4.1; provided in newick format in `supplementary_file_5/output/tree/04_tbe.raxml.support`), and 587 nodes have $\text{TBE} \geq 70\%$ (538 nodes with 100% TBE), while only a single node with considerably low TBE value of 38% was observed, indicating a robust tree topology. The tree suggests the existence of three major clades of *Streptomyces*, consistent with the phylogenetic structure observed in the 16S rRNA analysis presented in chapter 2 (Figure 4.1).

I examined the distribution of genomes currently assigned at NCBI to *S. griseus* (six genomes, leaves shown in red), *S. clavuligerus* (one genome, leaf shown in green) and *S. rimosus* (three genomes, leaves shown in blue) on the SCO phylogenetic tree (Figure 4.1), to determine whether the core genome phylogeny places these species into monophyletic groups, which would be expected if the species assignments were correct.

Although neither species forms a monophyletic groups in the SCO phylogeny, I found that all *S. rimosus* genomes were closely related, forming a tightly clustered clade on the SCOG tree. Specifically, genomes GCF_000721045.1 and GCF_000721045.1 are represented as sister leaves, although the GCF_003865155.1 genome is slightly more divergent from the other two, yet still closely related. This finding confirms previous observations discussed in chapter 3, where I identified two species and a single genus across all 24 available *S. rimosus* genomes, distributed across three distinct components in the MST tree (Figure 3.24 and Figure 3.25).

In chapter 3, I also identified six candidate species and two candidate genera across all 19 available *S. griseus* genomes, with the larger genus consisting of five species (Figure 3.24 and Figure 3.25). Five of the six *S. griseus* genomes appear in the same clade, while the remaining genome is more distantly related, confirming these previous observations. For *S. clavuligerus*, the placement of which was also questioned in earlier chapters (Figure 2.13), no interpretation is possible as only a single representative genome was chosen for this species (Figure 3.24 and Figure 3.25). The mapping of *S. clavuligerus* was carried out to illustrate its placement within the SCOG tree, thereby providing a visual context for its evolutionary relationships among other species within the *Streptomyces*.

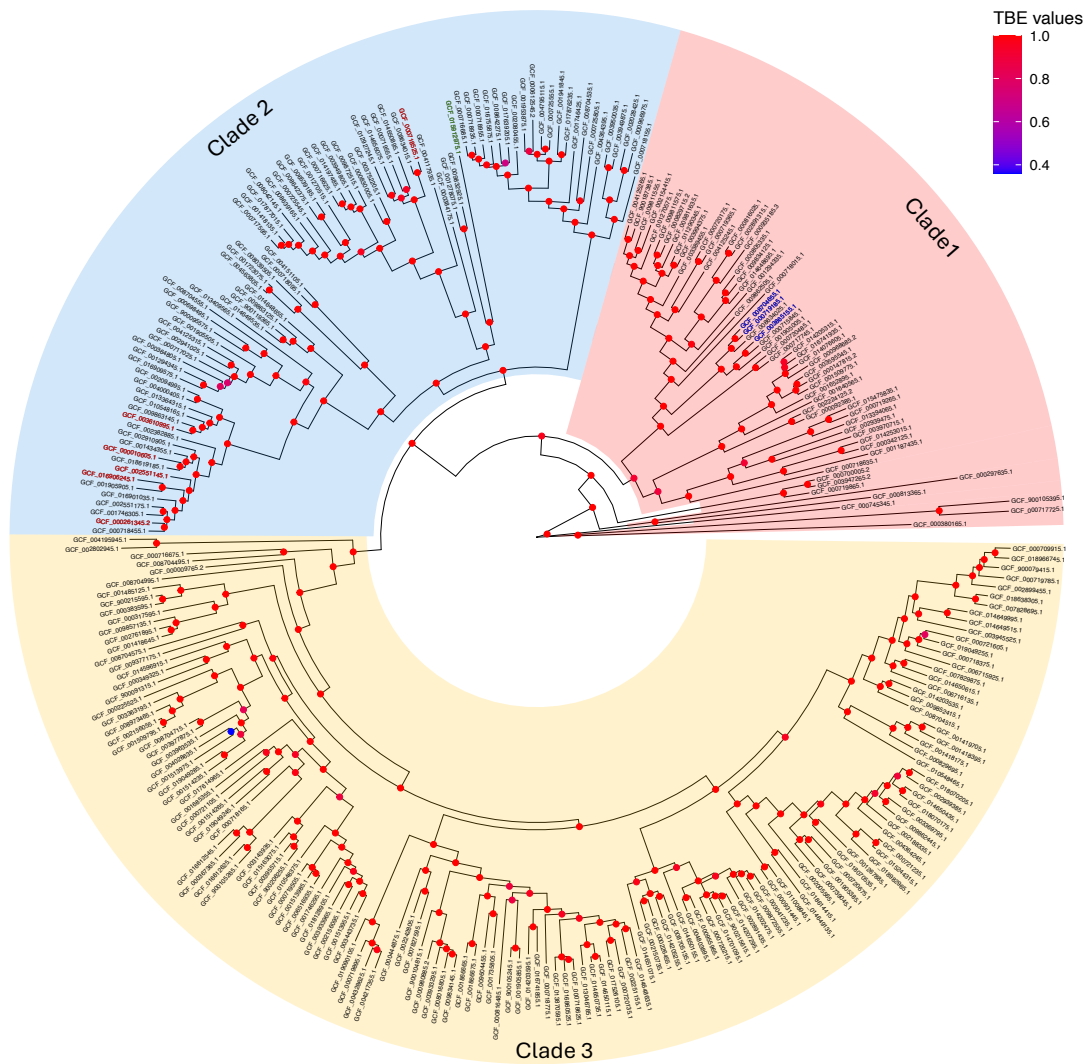


Figure 4.1: Maximum-likelihood tree of concatenated 117 SCOGs sequences with the distribution of genomes currently assigned *S. griseus* (red), *S. rimosus* (blue) and *S. clavuligerus* (green). The ML tree is rooted at midpoint with TBE values shown for each branch and ythree major clades (Clade 1 - Clade 3) being highlighted in distinct colours.

4.3.4 ANIm reveals significant genomic diversity within *Streptomyces*

To further investigate genomic diversity within the *Streptomyces* genus, ANIm analysis was conducted on the same set of 295 representative *Streptomyces* genomes (see methodology section 4.2.6). Pairwise ANIm comparisons (Figure 4.2) reveals important trends that highlight the relationships between identity and genome coverage. First, the relationship between identity and genome coverage is generally monotonic, meaning that as identity increases, so does coverage. However, noticeable variation exists within this trend. Two distinct patterns can be observed, differing in the slope of the relationship. The first pattern, which extends to approximately (88, 55) on the graph, likely represents between-genus comparisons. In this case, there is a rapid increase in shared homologous sequences without a corresponding increase in identity. In contrast, the second pattern shows an increase in identity with little additional homologous sequence, indicating possible between-species comparisons.

ANIm analysis also reveals significant genomic dissimilarity among the isolates. 95.1% of pairwise comparisons involve isolates that share less than 50% of their genome based on alignment length (Figure 4.2). The genomes showing the greatest level of divergence, with only 4.1% genome coverage, were GCF_000367365.1 (*Streptomyces prunicolor*) and GCF_000380165.1 (*Streptomyces vitaminophilus*).

Complementing the alignment coverage analysis, the genome identity analysis also highlighted significant genomic diversity within *Streptomyces*. As expected, given the aim of selecting one genome per species, genome identity ranged from 98.7% to 83.7%, with 99.9% of pairwise comparisons falling below the proposed species boundary of 95%

(Figure 4.2). The lowest genome identity, 83.7%, was observed between GCF_000380165.1 (*Streptomyces vitaminophilus*) and GCF_016612625.1 (*Streptomyces* sp. MBT62).

It is important to note that since the initial data download from NCBI on the 8th July 2021 (methodology section 2.2.9), GCF_000380165.1 has been reclassified outside the *Streptomyces* genus and is now known as *Wenjunlia vitaminophila*. Despite this reclassification, there are still instances of genomes that remain within the *Streptomyces* classification despite sharing as little as 5% of their genome by alignment length. Examples include GCF_000367365.1 (*Streptomyces prunicolor*) and GCF_000297635.1 (*Streptomyces* sp. AA0539), as well as GCF_000718165.1 (*Streptomyces fulvoviolaceus*) and GCF_900105395.1 (*Streptomyces* sp. TLI_053). In total 1.6% of pairwise comparisons are between genomes share less than 10% of homologous genome sequence by alignment length. This suggests a much higher-than-expected frequency of misclassification at the genus level within *Streptomyces* than is indicated in the literature. Although ongoing reclassification efforts are addressing these issues, discrepancies persist.

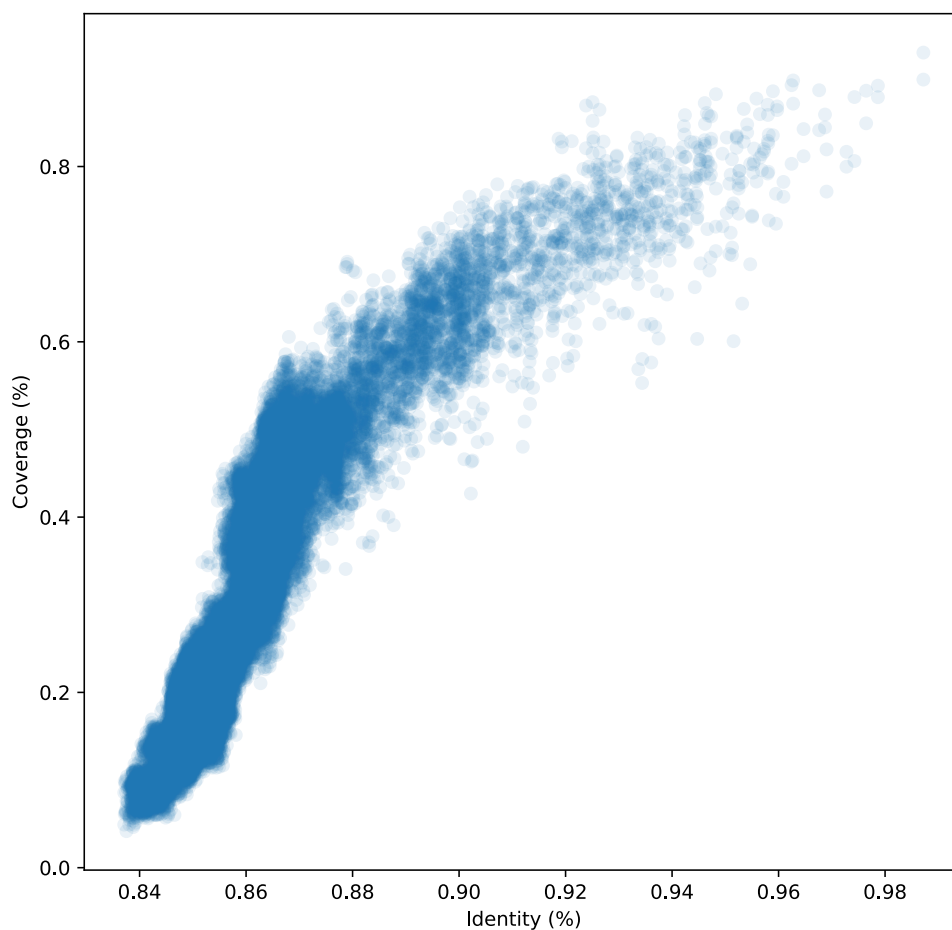


Figure 4.2: Pairwise ANIm comparisons of 295 representative *Streptomyces* genomes, with the X-axis representing genome identity (%) and the Y-axis representing genome coverage (%). Each point corresponds to a pairwise comparison between two genomes.

Due to the large size of genome coverage plot, the plot cannot be embedded directly in this document without causing issues. To ensure accessibility and maintain the integrity of this thesis, the plot is available in the online repository in `matrix_coverage_run1.pdf; supplementary_file_3\output\pyani_plots`. The genome coverage plot suggests the existence of many clusters of genomes that are more closely related to each other than to genomes outside their clusters. However, accurately determining the composition of these clusters from the plot itself is challenging, particularly since some clusters consist of genomes that do not all mutually share 50% genome coverage. This observation underscores the complexity of genomic relationships within *Streptomyces* and motivates for further in-depth investigation to better understand genomic relationships within the genus.

4.3.5 Identification of genus boundaries for *Streptomyces*

The observation of low genome coverage shared between some *Streptomyces* genomes (result section 4.3.4) raises the question of how many distinct genus-level groups (i.e. candidate novel genera) might actually exist within the *Streptomyces* group. Although there is a proposed threshold for ANI, where isolates sharing 95% or more of their genome identity are likely to represent the same species, there is no consensus on interpreting genus boundaries based on whole genome coverage or identity. Thus, I aimed to develop a method to establish clear boundaries for differentiating between genera. For this, I first clustered *Streptomyces* at genome coverage thresholds ranging from 40% to 85% in steps of 0.1%, and checked their distribution on the SCOG phylogeny to see whether they formed monophyletic groups, and whether these methods were congruent with each

other (methodology section 4.2.6). I found that genome coverage thresholds providing clear separation (i.e. that form exclusively monophyletic groups on the SCOG tree) are identifiable between 83.1% to 85% of genome coverage (Figure 4.3B). However, this consistency would result in dividing *Streptomyces* into a large number (284 to 291) of cluster, with the majority consisting of a single representative (Figure 4.3A). Thus, this is not a useful measure for identification of genus-level groupings.

Additionally, I identified thresholds that separate *Streptomyces* genomes into fewer clusters, although some thresholds produced a single non-monophyletic group on the SCOG phylogeny. Specifically, thresholds between 40% to 43.1% and 43.8% to 48.4% resulted in fewer clusters, with 30 to 34 clusters observed for the 40% to 43.1% range, and 36 to 45 clusters in the 43.8% to 48.4% range. The variability in these results offered a broad range of possible candidate genus boundaries. However, further analysis was necessary to determine thresholds that minimised the formation of non-monophyletic groups while also avoiding excessive separation of *Streptomyces* into numerous clusters and singletons, ensuring the preservation of true relationships at the genus level.



Figure 4.3: Assessment of genome coverage thresholds on the clustering of *Streptomyces* and their congruence with the SCOG phylogeny. A) Plots showing number of clusters and singletons formed at each genome coverage threshold. B) Plots showing counts of monophyletic and non-monophyletic clades at each threshold.

To address this challenge, I again used ANIm results, as the combination of genome coverage and genome identity to which I applied piecewise linear regression. This allowed the data to be segmented and for me to fit a separate linear regression model to each segment (Figure 4.4 and Figure 4.5). These segments yield breakpoints where the linear relationships in the data appears to change. In this case such breakpoints might be interpreted as candidate thresholds for defining boundaries for genus-level separation of *Streptomyces*. For piecewise linear regressions with more than three segments, there will be multiple breakpoints. Each of these breakpoints could potentially be considered as a candidate for defining a genus boundary separation. One could focus on the breakpoint that most distinctly separates the data into shallow and steep gradient regimes. This approach may offer more insightful separation by highlighting differences in the relationships between the genomes being compared. These distinctions could reflect different biological or genomic interactions, such as between-genus or within-genus comparisons. As more closely related genera tend to share greater genetic material (Goris et al., 2007), the overall sequence identity is likely to remain low due to the extended time for divergence, allowing for the accumulation of mutations and variations (Hershberg, 2015). For between-genus comparisons, we might expect a trend of a higher increase in shared homology, but not necessarily for sequence identity. Conversely, in between-species comparisons, we might expect to see a significant increase in genomic identity, as these species have had less time to diverge from their common ancestor. However, because closely related species exhibit higher recombination rates and different species may possess distinct genes for adaptation, the overall increase in genomic identity

might be relatively modest (Bury-Moné et al., 2023; McDonald & Currie, 2017a).

To determine the appropriate number of segments for this analysis, I tested a range of different models, varying from 1 to 10 segments, and calculated the Bayesian Information Criterion (BIC) values. The table 4.2 shows the BIC, convergence status, and Residual Sum of Squares (RSS) for models with varying numbers of segments. While models with up to six segments converged successfully, attempts to fit seven or more breakpoints resulted in failure to converge. Convergence failure suggests that the model is unable to find a stable solution with additional breakpoints, likely because the increased complexity makes it difficult to optimise the fit. This indicates overfitting or a lack of sufficient data to support such fine segmentation, reinforcing that adding more segments (beyond six segments) does not improve the model's reliability.

Table 4.2: Summary statistics for identification of optimal number of segments for piecewise linear regression.

No. of Segments	BIC	converged	Residual Sum of Squares
1	-5.0245e+05	True	264.29
2	-5.5316e+05	True	147.24
3	-5.6657e+05	True	126.12
4	-5.6928e+05	True	122.2
5	-5.6968e+05	True	121.61
6	-5.6989e+05	True	121.29
7	-	False	NA
8	-	False	NA
9	-	False	NA
10	-	False	NA

Although the model with six segments exhibits the lowest BIC value (Table 4.2) and thus provides the best statistical fit, interpreting these six segments is challenging (Figure 4.4). Notably, while the BIC continues to decrease with additional segments beyond four, the gains in fit become marginal. Specifically, the difference in BIC between models with four and six segments is minimal, and the reduction in RSS is also relatively small. However, fitting four segments, yields a much clearer separation, particularly evident at the breakpoint of 55.6% genome coverage and 88.1% identity (Figure 4.5). This supports earlier observations discussed in section 4.3.4, which highlight two distinct patterns based on the relationship's slope. The first trend, extending to around (88, 55) on the graph, likely corresponds to between-genus comparisons, where there is a rapid increase in shared homologous sequences but no significant rise in sequence identity. In contrast, the second trend shows an increase in identity with minimal additional homologous sequence being shared, suggesting between-species comparisons. Therefore, selecting four segments offers a practical balance between interpretability and model fit, avoiding the overfitting risks associated with more complex models. Additional segments might result in segments that are less meaningful and harder to interpret, while three breakpoints provide a more robust and comprehensible model.

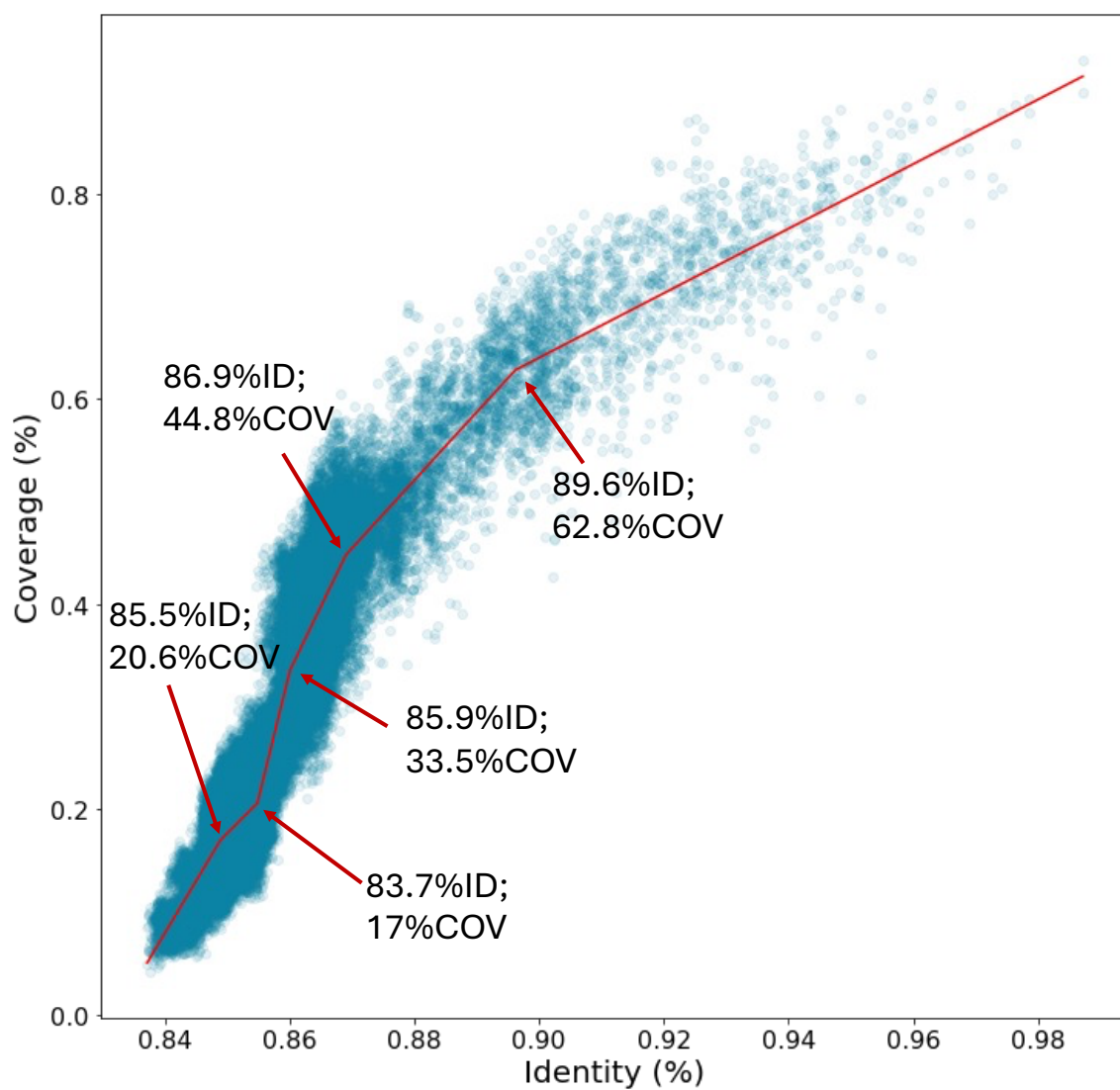


Figure 4.4: Piecewise linear regression with six segments for ANIm comparisons among 295 representative *Streptomyces* genomes does not yield clear boundary separation.

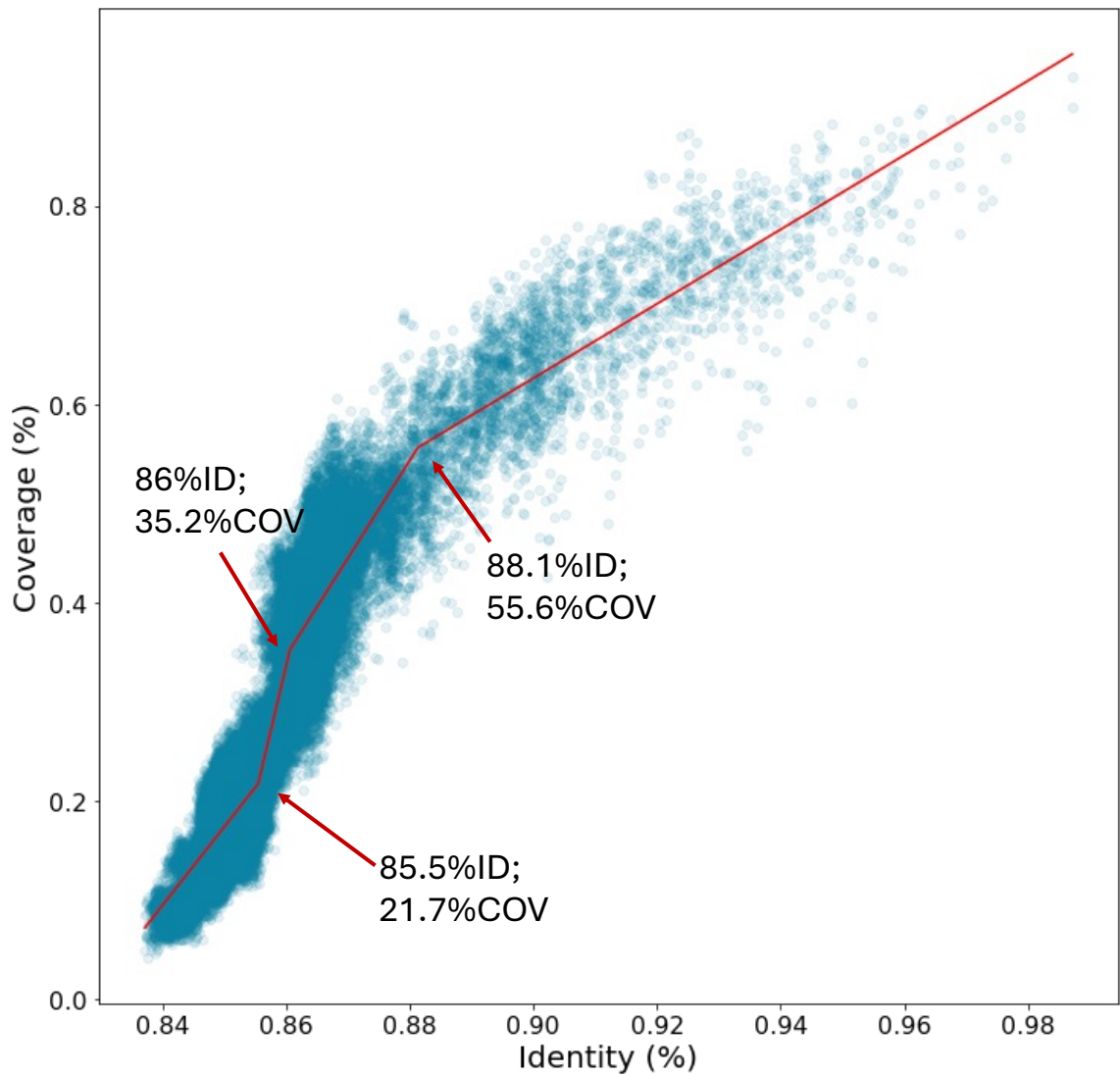


Figure 4.5: Piecewise linear regression with four segments for ANIm comparisons among 295 representative *Streptomyces* genomes reveals a clear boundary at 88.1% genome coverage and 55.6% genome identity. Up to this point, identity increases without a significant gain in homologous sequences, indicating between-genus comparisons. Beyond this threshold, there is a rise in shared homologous sequences without a corresponding increase in identity, suggesting within-genus comparisons.

Following the analysis of piecewise linear regression, a probable genus boundary was identified at 55.6% genome coverage and 88.1% genome identity. The next step was to delineate distinct subgroups within the genus *Streptomyces* and their corresponding members. This was accomplished using a combination of ANIm and graph theory (Section 4.2.6). Using the ANI results, I constructed a complete network, where nodes represented individual genomes and edges reflected the minimum genome coverage and average nucleotide identity between genome pairs. This was necessary because pyANI compares genomes bidirectionally—e.g., genome A vs. genome B and genome B vs. genome A—resulting in slightly different values for each direction. Edges with genome coverage below 55.6% and genome identity below 88.1% were removed, as these thresholds were found to separate different subgroups. In cases where non-cliques were formed—genomes that did not form cohesive groups and some falling below the genus boundary, resulting in ambiguous relationships—I removed edges with the lowest genome coverage until distinct cliques emerged, clarifying their relationships and allowing for a clearer understanding of subgroup divisions within *Streptomyces*. This analysis revealed 79 distinct groups within *Streptomyces*, 36 of which were represented by a single genome, and the largest genus consisting of as many as 20 genomes (genus 20). Each candidate genus was assigned a random number between 1 and 79 to ensure consistency for future analysis and to facilitate easier referencing in the text and thesis when needed. Once the different *Streptomyces* subgroups and their members were identified, it prompted the question of how these groups are distributed on the SCOG tree and whether they form monophyletic clades. Thus, I examined the distribution of the identified genera on

the SCOG phylogenetic tree. This analysis revealed that the majority of the proposed genera and their members form monophyletic groups, with the exception of genus 17 (highlighted in blue), which forms paraphyletic relationship with a singleton genus 66. Additionally, genus 16 (highlighted red) forms a paraphyletic relationship with genera 42,56,57,45 and 58. (Figure 4.6).

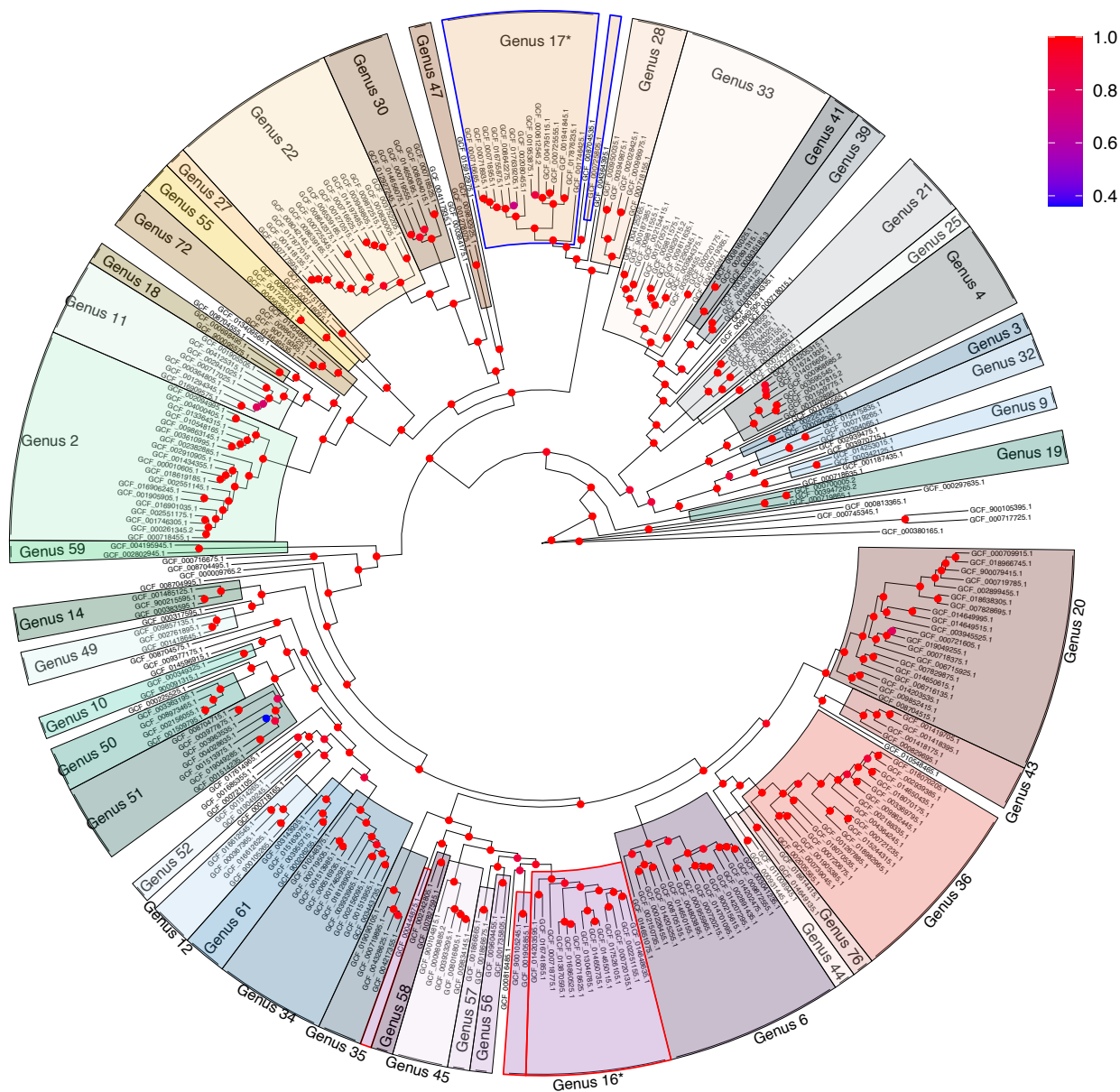


Figure 4.6: Maximum-likelihood tree of concatenated 137 SCOG sequences, rooted at the midpoint, with 79 labeled distinct groups, each representing a separate candidate genus within the broader *Streptomyces* lineage. These groups were identified by clustering the genomes based on genus boundaries of 45.8% genome coverage and 88.8% genome identity, as determined through a combination of ANIm analysis and piecewise linear regression (methodology section 4.2.6). Non-monophyletic genus 17 is highlighted in blue, while non-monophyletic genus 16 is highlighted in red.

4.3.6 Location of SCOGs on chromosome

63 out of 295 genomes met the criteria of having complete or chromosomal level assemblies, and only these were considered for the analysis. Before the location of SCOGs on the chromosomes could be determined, it was necessary to reorient genomes so that, for all genomes, *oriC* was on the positive strand. This reorientation was essential to ensure consistency in determining and comparing the location of SCOGs across all genomes. After checking the location of *oriC*, I found that in 60 out of 63 cases, the *oriC*, as expected, was located near the middle of the linear chromosome (Figure 4.7). In the remaining three cases, the *oriC* was found at the telomere of the chromosome. If the chromosomes were circular, the position of the *oriC* could be repositioned at the 50% mark because, in circular chromosomes, any point can be conceptually used as the origin or midpoint due to their continuous, unbroken nature. However, at the time of the study, there were no associated publications for these three genomes, making it impossible to confirm whether they indeed possessed circular chromosomes. Thus, to avoid bias and possible misinterpretation, the assemblies of GCF_009834125.1, GCF_000147815.2 and GCF_018128905.1 were excluded from the analysis.

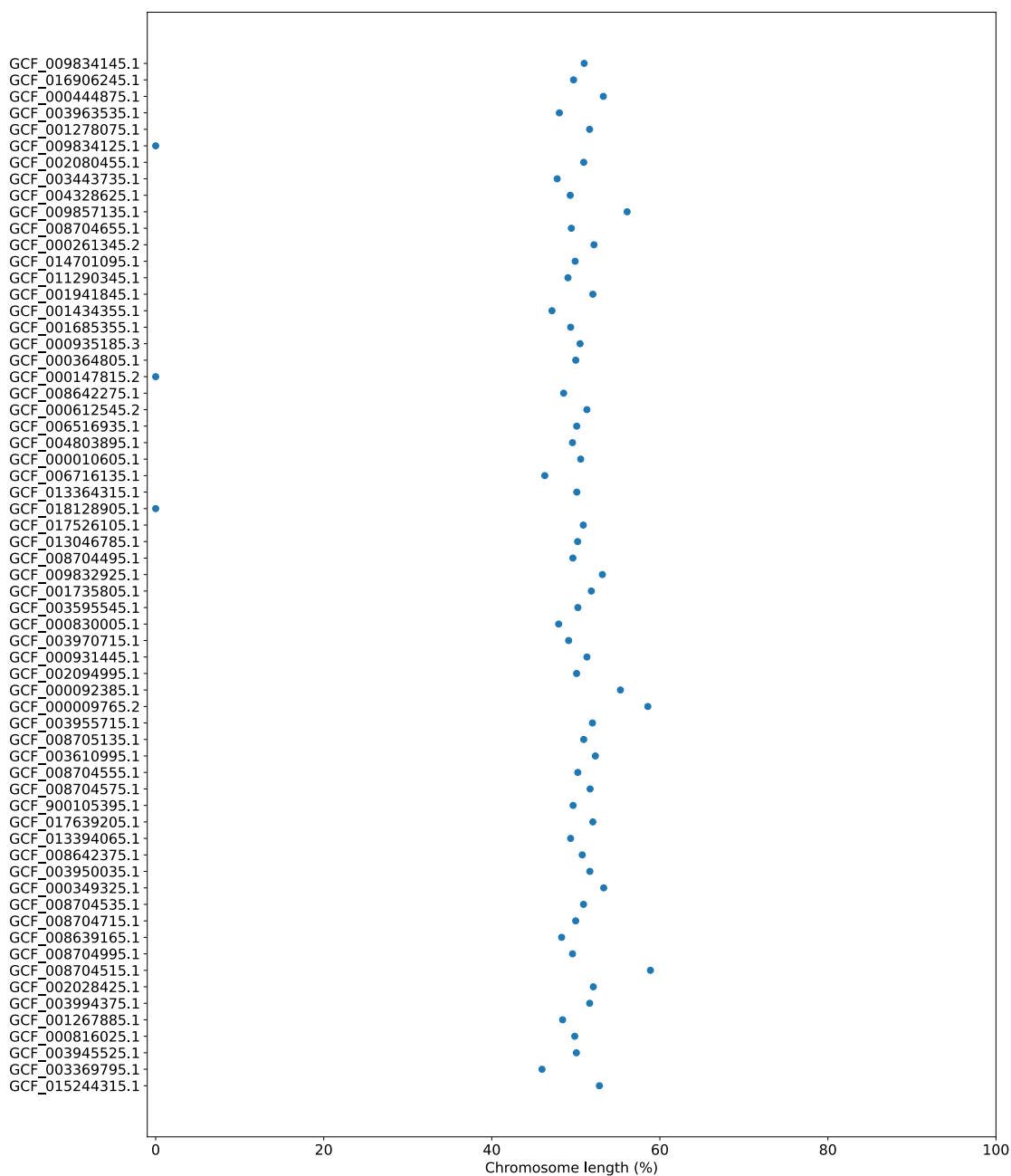


Figure 4.7: OriC location for genomes assembled to complete or chromosomal level in NCBI.

After analysing the chromosomal locations of SCOGs, I identified three distinct patterns (Figure 4.8):

1. **Core Exclusive:** This group comprised 76 SCOGs (55.5% of all SCOGs), consistently distributed between the 25th and 75th percentile of chromosome length (i.e. in the central 50% of the chromosome) across all 60 investigated genomes (Figure 4.8 green). The consistent central positioning of these SCOGs suggests they are situated in a region with lower recombination rates, likely providing greater genetic stability (Bentley et al., 2002).
2. **Predominantly Core:** 36 SCOGs (27%) were $\geq 50\%$ of the investigated SCOGs variants were situated within the central core region of the chromosome, with sporadic occurrences slightly beyond the adapted core boundary of 25%-50% (Figure 4.8 purple). The core region of the chromosome in *Streptomyces* was found to be slightly shifted to the left based on data from *Streptomyces coelicolor* A3(2), a situation observed when only a small number of genomes were available (Bentley et al., 2002). It remains possible that the core regions in other *Streptomyces* genomes might vary, potentially shifting towards either end of the chromosome. Therefore, SCOGs in this predominantly core category might be less prone to recombination due to their general positioning within a relatively stable core region.
3. **Predominantly Non-Core:** This group consisted of 25 SCOGs (17.5%), mainly located outside the core region. Few genomes containing these genes within

the core chromosome (Figure 4.8 red). The positioning of these SCOGs in the chromosomal arms, areas typically associated with higher recombination rates and genetic instability, suggests they are likely to be more prone to recombination (Widenbrant et al., 2007).

The majority of genes, whether located on chromosome arms or within the central core, are essential for an organism's survival rather than niche adaptation (Figure 4.8). This is because they encode fundamental cellular components, including structural ribosomal proteins (e.g., 50S L31, 50S L17), cell division genes (e.g., SepF, ribonucleases), transcriptional regulators (e.g., WhiB family), and DNA replication and repair proteins (e.g., RNA polymerase subunit beta, RecN).

The majority of genes, regardless of their location on the chromosome (e.g., within the arms or the central part of the chromosome), are essential for an organism's survival rather than niche adaptation (Figure 4.8). This is because they mostly encode fundamental cellular components, including structural ribosomal proteins (e.g., 50S L31, 50S L17), cell division genes (e.g., SepF, ribonucleases), transcriptional regulators (e.g., WhiB family), and DNA replication and repair proteins (e.g., RNA polymerase subunit beta, RecN). Additionally, I identified proteins with uncharacterised functions, such as Domain of Unknown Function (DUF) proteins and hypothetical proteins, which were localised in both the chromosome arms and the core. Moreover, proteins involved in the synthesis of structural ribosomal components (50S and 30S) were largely co-located (Figure 4.8; red).



4.3.7 Conservation of SCOGs nucleotide variants across the phylogenetic tree

I analysed the nucleotide variants of 137 SCOGs found across the 295 *Streptomyces* genomes. I found that 85 SCOGs (62%) were represented by 295 unique nucleotide variants, implying each variant was specific to each individual genome. The remaining 52 SCOGs (38%) had at least one nucleotide sequence variant that was common to two or more *Streptomyces* genomes, with some SCOGs having as few as 29 nucleotide sequence variants (Table 4.3).

Table 4.3: Lists 52 SCOGs with at least one nucleotide variant shared by two or more *Streptomyces* genomes. It includes the total number of unique nucleotide sequence variants, the corresponding protein products (as annotated in the sequence records), and the number of conserved nucleotide sequence variants categorized as either monophyletic or non-monophyletic in the SCOG tree.

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 <i>Streptomyces</i> genomes	Total Mono-phyletic	Total Non-Monophyletic
OG0002076	30S ribosomal protein S18	166	50	17	33
OG0002077	arD family transcriptional regulator	285	9	7	2
OG0002078	50S ribosomal protein L36	29	18	3	15
OG0002079	50S ribosomal protein L22	256	30	18	12
OG0002080	50S ribosomal protein L29	214	44	16	28
OG0002081	RNA-binding protein	281	12	10	2
OG0002082	50S ribosomal protein L30	190	48	10	38
OG0002083	bifunctional nuclease family protein	292	3	2	1
OG0002086	50S ribosomal protein L32	107	40	7	33
OG0002089	SDR family NAD(P)-dependent oxidoreductase	294	1	1	0
OG0002090	hypothetical protein	294	1	1	0
OG0002097	DUF3071 domain-containing protein	294	1	1	0
OG0002099	insulinase family protein	294	1	1	0
OG0002100	transcriptional repressor NrdR	294	1	1	0

Continued on next page

Table 4.3 – Continued from previous page

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 Streptomyces genomes	Total Monophyletic	Total Non-Monophyletic
OG0002102	4-hydroxy-tetrahydrodipicolinate reductase	294	1	1	0
OG0002103	RNA pseudouridine(55) synthase TruB	294	1	1	0
OG0002104	ribosome maturation factor RimP	294	1	1	0
OG0002105	ribosome maturation factor RimM	294	1	1	0
OG0002107	acylphosphatase	294	1	1	0
OG0002108	bifunctional DNA-formamidopyrimidine glycosylase	294	1	1	0
OG0002109	DUF177 domain-containing protein	294	1	1	0
OG0002110	pantetheine-phosphate adenylyltransferase	294	1	1	0
OG0002111	proline dehydrogenase family protein	294	1	1	0
OG0002115	cob(I)yrinic acid a,c-diamide adenosyltransferase	294	1	1	0
OG0002116	F0F1 ATP synthase subunit epsilon	279	15	13	2
OG0002118	F0F1 ATP synthase subunit delta	279	15	13	2
OG0002119	peptide chain release factor 1	279	15	13	2
OG0002120	50S ribosomal protein L31	249	32	15	17

Continued on next page

Table 4.3 – Continued from previous page

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 Streptomyces genomes	Total Mono-phyletic	Total Non-Monophyletic
OG0002121	hreonine synthase	249	32	15	17
OG0002122	peptide deformylase	249	32	15	17
OG0002125	mycoredoxin	287	7	5	2
OG0002128	O-methyltransferase	287	7	5	2
OG0002129	succinyl-diaminopimelate desuccinylase	287	7	5	2
OG0002130	WhiB family transcriptional regulator	294	1	1	0
OG0002132	cytidine deaminase	294	1	1	0
OG0002134	succinate dehydrogenase hydrophobic membrane anchor subunit	293	1	1	0
OG0002135	typtophan-tRNA ligase	293	1	1	0
OG0002136	malate dehydrogenase	293	1	1	0
OG0002139	hypothetical protein	293	1	1	0
OG0002144	D-tyrosyl-tRNA(Tyr) deacylase	293	1	1	0
OG0002145	folate-binding protein YgfZ	293	1	1	0
OG0002146	FABP family protein	293	1	1	0
OG0002147	phosphate ABC transporter ATP-binding protein	293	1	1	0

Continued on next page

Table 4.3 – Continued from previous page

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 Streptomyces genomes	Total Mono-phyletic	Total Non-Monophyletic
OG0002150	hypothetical protein	293	1	1	0
OG0002152	acil phosphoribosyltransferase	293	1	1	0
OG0002153	hypothetical protein	293	1	1	0
OG0002154	SEC-C domain-containing protein	293	1	1	0
OG0002155	ABC transporter permease	293	1	1	0
OG0002156	hypothetical protein	293	2	2	0
OG0002161	M18 family aminopeptidase	293	2	2	0
OG0002162	SseB family protein	293	2	2	0
OG0002163	molecular chaperone DnaJ	293	2	2	0
OG0002164	helix-turn-helix domain-containing protein	293	2	2	0
OG0002166	DUF5063 domain-containing protein	293	2	2	0
OG0002167	DUF4177 domain-containing protein	217	35	16	19
OG0002168	MarP family serine protease	217	35	16	19
OG0002169	phage holin family protein	217	35	16	19
OG0002170	ATP-binding protein	217	35	16	19
OG0002171	hypoxanthine phosphoribosyltransferase	292	3	2	1
OG0002172	DUF3180 domain-containing protein	292	3	2	1

Continued on next page

Table 4.3 – Continued from previous page

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 Streptomyces genomes	Total Monophyletic	Total Non-Monophyletic
OG0002173	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase	292	3	2	1
OG0002174	dihydroneopterin aldolase	292	3	3	0
OG0002176	Ppx/GppA family phosphatase	292	3	3	0
OG0002180	NADH-quinone oxidoreductase subunit Nuol	292	3	3	0
OG0002181	NADH-quinone oxidoreductase subunit M	292	3	3	0
OG0002184	YajQ family cyclic di-GMP-binding protein	292	3	3	0
OG0002185	50S ribosomal protein L11	267	22	10	12
OG0002186	50S ribosomal protein L10	275	16	9	7
OG0002188	50S ribosomal protein L3	281	11	6	5
OG0002189	50S ribosomal protein L4	287	8	5	3
OG0002190	50S ribosomal protein L23	255	27	15	12
OG0002191	30S ribosomal protein S19	182	42	20	22
OG0002192	30S ribosomal protein S3	292	3	2	1
OG0002193	50S ribosomal protein L16	250	27	11	16
OG0002194	30S ribosomal protein S17	254	29	13	16
OG0002195	50S ribosomal protein L24	254	30	17	13

Continued on next page

Table 4.3 – Continued from previous page

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 Streptomyces genomes	Total Mono-phyletic	Total Non-Monophyletic
OG0002196	50S ribosomal protein L5	272	16	13	3
OG0002197	30S ribosomal protein S8	249	34	20	14
OG0002198	50S ribosomal protein L6	282	12	8	4
OG0002199	50S ribosomal protein L18	279	13	7	6
OG0002200	50S ribosomal protein L15	283	11	9	2
OG0002201	30S ribosomal protein S13	270	20	14	6
OG0002202	50S ribosomal protein L17	289	5	4	1
OG0002203	chaperonin GroEL	289	5	4	1
OG0002205	IMP dehydrogenase	289	5	4	1
OG0002206	5-(carboxyamino)imidazole ribonucleotide mutase	289	5	4	1
OG0002210	ABC transporter permease	289	5	4	1
OG0002211	SsrA-binding protein SmpB	294	1	1	0
OG0002213	icotinate phospho-ribosyltransferase	294	1	1	0
OG0002214	M67 family metalloproteinase	294	1	1	0
OG0002215	MBL fold metallo-hydrolase	294	1	1	0
OG0002216	ribonuclease PH	294	1	1	0
OG0002217	co-chaperone GroES	292	3	3	0

Continued on next page

Table 4.3 – Continued from previous page

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 Streptomyces genomes	Total Monophyletic	Total Non-Monophyletic
OG0002221	icotinate-nucleotide adenyltransferase	293	2	2	0
OG0002222	PhoH family protein	293	2	2	0
OG0002223	RNA maturation RNase YbeY	293	2	2	0
OG0002225	RNA dihydrouridine synthase DusB	293	2	2	0
OG0002226	hypothetical protein	293	2	2	0
OG0002227	DUF3145 domain-containing protein	293	2	2	0
OG0002229	bifunctional DNA primase/polymerase	293	2	2	0
OG0002231	DUF3043 domain-containing protein	294	1	1	0
OG0002232	YggT family protein	258	31	22	9
OG0002236	ABC transporter permease	258	31	22	9
OG0002238	imidazoleglycerol-phosphate dehydratase HisB	258	31	22	9
OG0002239	PriA	258	31	22	9
OG0002240	PAC2 family protein	258	31	22	9
OG0002242	metal-sulfur cluster assembly factor	293	2	1	1
OG0002243	yl-CoA hydratase/isomerase family protein	293	2	1	1

Continued on next page

Table 4.3 – Continued from previous page

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 Streptomyces genomes	Total Mono-phyletic	Total Non-Monophyletic
OG0002246	SMC-Scp complex subunit ScpB	293	2	1	1
OG0002247	cell division protein SepF	294	1	0	1
OG0002250	biquitin-like protein Pup	291	4	1	3
OG0002251	Pup-protein ligase	291	4	1	3
OG0002252	WYL domain-containing protein	291	4	1	3
OG0002253	hypothetical protein	291	4	1	3
OG0002255	RNA methyltransferase	291	4	1	3
OG0002257	protein translocase subunit SecF	291	4	1	3
OG0002259	6,7-dimethyl-8-ribityllumazine synthase	291	4	1	3
OG0002260	phosphoribosyl-ATP diphosphatase	285	10	7	3
OG0002261	hypothetical protein	294	1	1	0
OG0002263	acil-DNA glycosylase	294	1	1	0
OG0002268	DNA repair protein RecN	294	1	1	0
OG0002270	DEAD/DEAH box helicase	294	1	1	0
OG0002272	Lrp/AsnC family transcriptional regulator	294	1	1	0
OG0002274	dephospho-CoA kinase	294	1	1	0

Continued on next page

Table 4.3 – Continued from previous page

Orthogroup ID	Protein Product	Total nt Sequence Variants	Common to ≥ 2 Streptomyces genomes	Total Mono-phyletic	Total Non-Monophyletic
OG0002277	DUF2029 domain-containing protein	294	1	1	0
OG0002278	DNA-directed RNA polymerase subunit beta	294	1	1	0
OG0002279	glutaredoxin family protein	294	1	1	0
OG0002280	peptide chain release factor N(5)-glutamine methyltransferase	294	1	1	0
OG0002281	PAS domain-containing sensor histidine kinase	294	1	1	0
OG0002282	hypothetical protein	231	43	25	18
OG0002284	DP-alcohol phosphatidyl-transferase family protein	231	43	25	18
OG0002285	dopeptidase La	231	43	25	18
OG0002286	GuaB1 family IMP dehydrogenase-related protein	231	43	25	18
OG0002287	hypothetical protein	231	43	25	18
OG0002290	preprotein translocase subunit SecG	280	14	10	4
OG0002291	FadR family transcriptional regulator	280	14	10	4
OG0002292	A pyrophosphatase	280	14	10	4

I examined the distribution of the 52 SCOG's nucleotide sequence variants that were common to more than two *Streptomyces* genomes, and I found that 51 of these SCOGs contained at least one monophyletic nucleotide variant, indicating a common ancestry among the organisms (Table 4.4, Figure 4.9). Additionally, I found that 38 SCOGs have at least one nucleotide sequence variant distributed in a non-monophyletic manner (Figure 4.10 and Table 4.5). Notably, all 38 of these SCOGs had at least one repeated nucleotide sequence variants that, while displaying non-monophyletic patterns on the SCOG tree, were confined within the same *Streptomyces* subgroup (with $\geq 45.8\%$ genome coverage and $\geq 88.8\%$ genome identity) (Table 4.6 and Figure 4.11). However, 16 SCOGs were found to have nucleotide sequence variants distributed in a non-monophyletic pattern, scattered across distinct *Streptomyces* subgroups (with genome coverage $> 45.8\%$ and genome identity $> 88.8\%$), suggesting a broader distribution beyond closely related taxa (Figure 4.12 and Table 4.7).

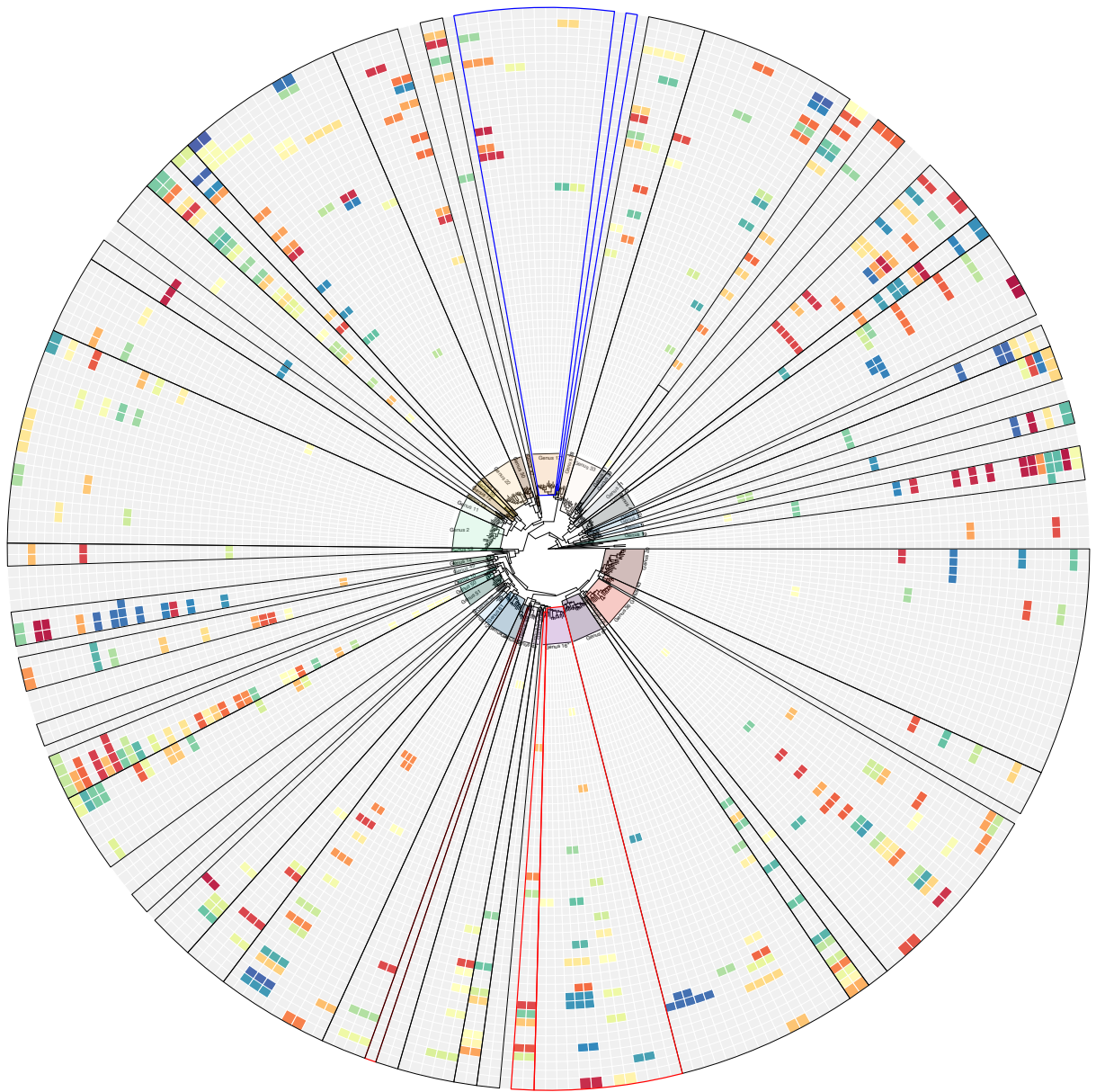


Figure 4.9: Variants of SCOG's nucleotide sequences, identified in two or more *Streptomyces* species, consistently form monophyletic clades in the core genome phylogenetic tree. Each individual ring represents a single SCOG, with the order of the ring (from inner to outer) and their corresponding protein products as annotated on sequence record is provided in Table 4.5. The same color within a ring indicates identical nucleotide variants. Variants that either form non-monophyletic clades or are unique to a single taxon on the core genome tree are shown in grey.

Table 4.4: List of 51 SCOGs in which "repeated" nucleotide sequence variants are shared across distinct *Streptomyces* genomes and form monophyletic clades. The corresponding protein products are annotated as per the sequence records. The order of the SCOGs aligns with the arrangement of the rings in Figure 4.9, from inner to outer.

Orthogroup ID	Protein Product	Total Mono-phyletic
OG0002089	SDR family NAD(P)-dependent oxidoreductase	1
OG0002104	ribosome maturation factor RimP	1
OG0002110	pantetheine-phosphate adenylyltransferase	1
OG0002130	WhiB family transcriptional regulator	1
OG0002134	succinate dehydrogenase hydrophobic membrane anchor subunit	1
OG0002211	SsrA-binding protein SmpB	1
OG0002214	M67 family metalloproteinase	1
OG0002231	DUF3043 domain-containing protein	1
OG0002242	metal-sulfur cluster assembly factor	1
OG0002250	biquitin-like protein Pup	1
OG0002261	hypothetical protein	1
OG0002083	bifunctional nuclease family protein	2
OG0002156	hypothetical protein	2
OG0002171	hypoxanthine phosphoribosyltransferase	2
OG0002192	30S ribosomal protein S3	2
OG0002221	icotinate-nucleotide adenylyltransferase	2
OG0002226	hypothetical protein	2
OG0002078	50S ribosomal protein L36	3
OG0002174	dihydroneopterin aldolase	3
OG0002217	co-chaperone GroES	3
OG0002202	50S ribosomal protein L17	4
OG0002125	mycoredoxin	5
OG0002189	50S ribosomal protein L4	5

Continued on next page

Table 4.4 – Continued from previous page

Orthogroup ID	Protein Product	Total Mono-phyletic
OG0002188	50S ribosomal protein L3	6
OG0002077	arD family transcriptional regulator	7
OG0002086	50S ribosomal protein L32	7
OG0002199	50S ribosomal protein L18	7
OG0002260	phosphoribosyl-ATP diphosphatase	7
OG0002198	50S ribosomal protein L6	8
OG0002186	50S ribosomal protein L10	9
OG0002200	50S ribosomal protein L15	9
OG0002081	RNA-binding protein	10
OG0002082	50S ribosomal protein L30	10
OG0002185	50S ribosomal protein L11	10
OG0002290	preprotein translocase subunit SecG	10
OG0002193	50S ribosomal protein L16	11
OG0002116	F0F1 ATP synthase subunit epsilon	13
OG0002194	30S ribosomal protein S17	13
OG0002196	50S ribosomal protein L5	13
OG0002201	30S ribosomal protein S13	14
OG0002120	50S ribosomal protein L31	15
OG0002190	50S ribosomal protein L23	15
OG0002080	50S ribosomal protein L29	16
OG0002167	DUF4177 domain-containing protein	16
OG0002076	30S ribosomal protein S18	17
OG0002195	50S ribosomal protein L24	17
OG0002079	50S ribosomal protein L22	18
OG0002191	30S ribosomal protein S19	20
OG0002197	30S ribosomal protein S8	20
OG0002232	YggT family protein	22

Continued on next page

Table 4.4 – Continued from previous page

Orthogroup ID	Protein Product	Total Mono-phyletic
OG0002282	hypothetical protein	25

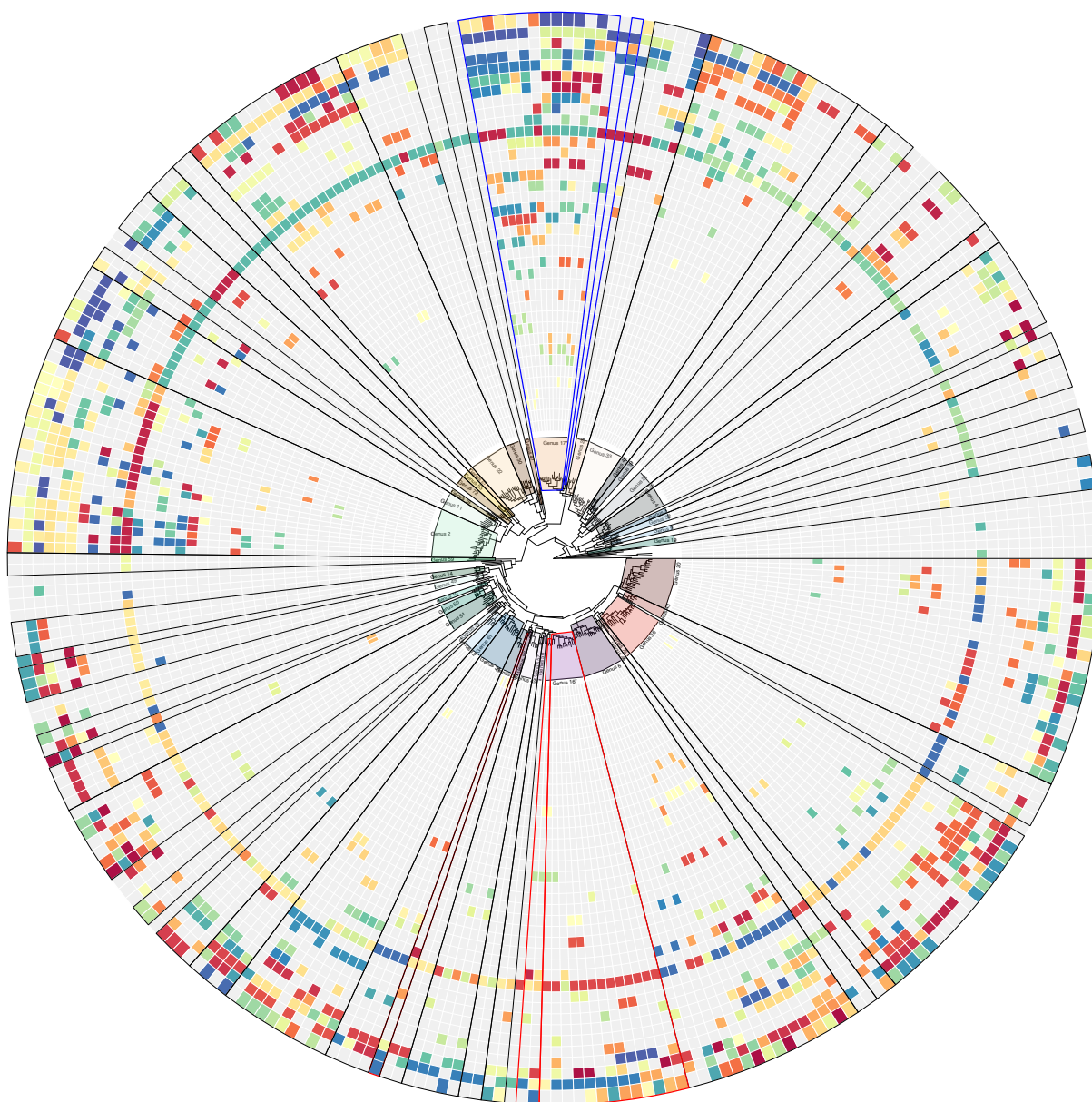


Figure 4.10: SCOG's repeated nucleotide sequence variants do not form monophyletic clades on the core genome tree. Each individual ring represents a single SCOG, with the order of the ring (from inner to outer) and their corresponding protein products as annotated on sequence record is provided in Table 4.5. The same color within a ring indicates identical nucleotide variants. Variants that either form monophyletic clades or are unique to a single taxon on the core genome tree are shown in grey.

Table 4.5: List of 38 SCOGs in which at least one repeated nucleotide variants are shared across distinct *Streptomyces* genomes and form non-monophyletic clades with their corresponding protein products as annotated in the sequence records. The order of the SCOGs corresponds to the order (from inner to outer) of the rings in the Figure 4.10.

Orthogroup ID	Protein Product	Total Non-Monophyletic
OG0002083	bifunctional nuclease family protein	1
OG0002171	hypoxanthine phosphoribosyltransferase	1
OG0002192	30S ribosomal protein S3	1
OG0002202	50S ribosomal protein L17	1
OG0002242	metal-sulfur cluster assembly factor	1
OG0002247	cell division protein SepF	1
OG0002077	arD family transcriptional regulator	2
OG0002081	RNA-binding protein	2
OG0002116	F0F1 ATP synthase subunit epsilon	2
OG0002125	mycoredoxin	2
OG0002200	50S ribosomal protein L15	2
OG0002189	50S ribosomal protein L4	3
OG0002196	50S ribosomal protein L5	3
OG0002250	biquitin-like protein Pup	3
OG0002260	phosphoribosyl-ATP diphosphatase	3
OG0002198	50S ribosomal protein L6	4
OG0002290	preprotein translocase subunit SecE	4
OG0002188	50S ribosomal protein L3	5
OG0002199	50S ribosomal protein L18	6
OG0002201	30S ribosomal protein S13	6
OG0002186	50S ribosomal protein L10	7
OG0002232	YggT family protein	9
OG0002079	50S ribosomal protein L22	12
OG0002185	50S ribosomal protein L11	12
OG0002190	50S ribosomal protein L23	12
OG0002195	50S ribosomal protein L24	13
OG0002197	30S ribosomal protein S8	14
OG0002078	50S ribosomal protein L36	15
OG0002193	50S ribosomal protein L16	16
OG0002194	30S ribosomal protein S17	16
OG0002120	50S ribosomal protein L31	17
OG0002282	hypothetical protein	18
OG0002167	DUF4177 domain-containing protein	19
OG0002191	30S ribosomal protein S19	22
OG0002080	50S ribosomal protein L29	28
OG0002076	30S ribosomal protein S18	33
OG0002086	50S ribosomal protein L32	33
OG0002082	50S ribosomal protein L30	38

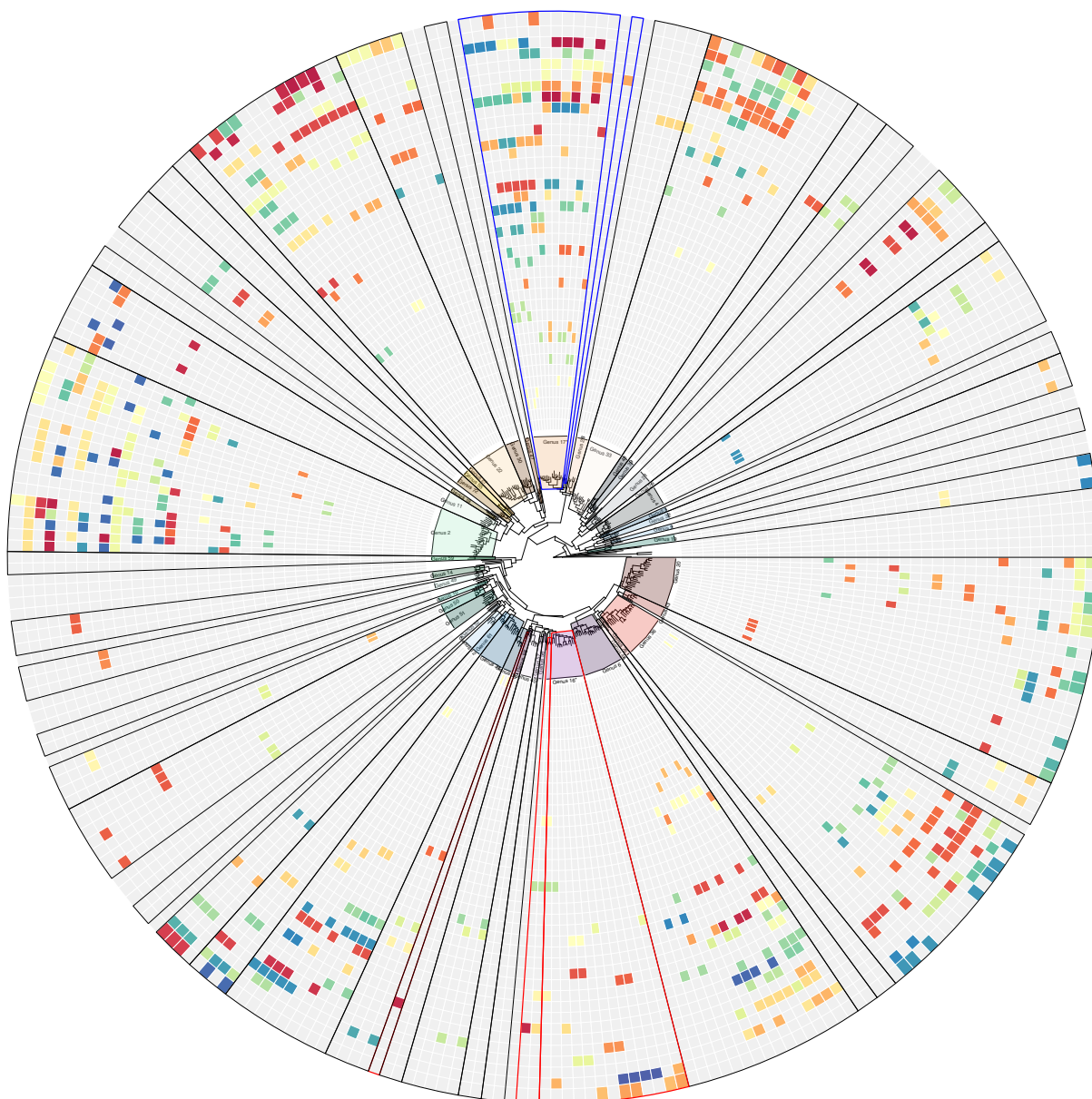


Figure 4.11: Repeated nucleotide sequence variants of SCOG do not form monophyletic clades in the core genome tree, but are confined within the same *Streptomyces* subgroups (with genome coverage of $\geq 45.8\%$ and genome identity of $\geq 88.8\%$). Each individual ring represents a single SCOG, with the order of the ring (from inner to outer) and their corresponding protein products as annotated on sequence record is provided in Table 4.6. The same color within a ring indicates identical nucleotide variants. Variants that either form monophyletic clades or are unique to a single taxon on the core genome tree are show in grey.

Table 4.6: List of 38 SCOGs, each containing at least one repeated nucleotide variant shared within the same *Streptomyces* subgroup and forming non-monophyletic clades. Corresponding protein products, as annotated in the sequence records, are also included. The order of the SCOGs follows the arrangement of rings (from inner to outer) in Figure 4.11.

Orthogroup ID	Protein Product	Total Non-Monophyletic
OG0002083	bifunctional nuclease family protein	1
OG0002171	hypoxanthine phosphoribosyltransferase	1
OG0002192	30S ribosomal protein S3	1
OG0002202	50S ribosomal protein L17	1
OG0002242	metal-sulfur cluster assembly factor	1
OG0002247	cell division protein SepF	1
OG0002077	arD family transcriptional regulator	2
OG0002081	RNA-binding protein	2
OG0002116	F0F1 ATP synthase subunit epsilon	2
OG0002125	mycoredoxin	2
OG0002200	50S ribosomal protein L15	2
OG0002189	50S ribosomal protein L4	3
OG0002196	50S ribosomal protein L5	3
OG0002250	biquitin-like protein Pup	3
OG0002260	phosphoribosyl-ATP diphosphatase	3
OG0002198	50S ribosomal protein L6	4
OG0002290	preprotein translocase subunit SecE	4
OG0002188	50S ribosomal protein L3	5
OG0002199	50S ribosomal protein L18	6
OG0002201	30S ribosomal protein S13	6
OG0002186	50S ribosomal protein L10	7
OG0002232	YggT family protein	9
OG0002079	50S ribosomal protein L22	12
OG0002185	50S ribosomal protein L11	12
OG0002190	50S ribosomal protein L23	12
OG0002195	50S ribosomal protein L24	13
OG0002197	30S ribosomal protein S8	14
OG0002078	50S ribosomal protein L36	15
OG0002193	50S ribosomal protein L16	16
OG0002194	30S ribosomal protein S17	16
OG0002120	50S ribosomal protein L31	17
OG0002282	hypothetical protein	18
OG0002167	DUF4177 domain-containing protein	19
OG0002191	30S ribosomal protein S19	22
OG0002080	50S ribosomal protein L29	28
OG0002076	30S ribosomal protein S18	33
OG0002086	50S ribosomal protein L32	33
OG0002082	50S ribosomal protein L30	38

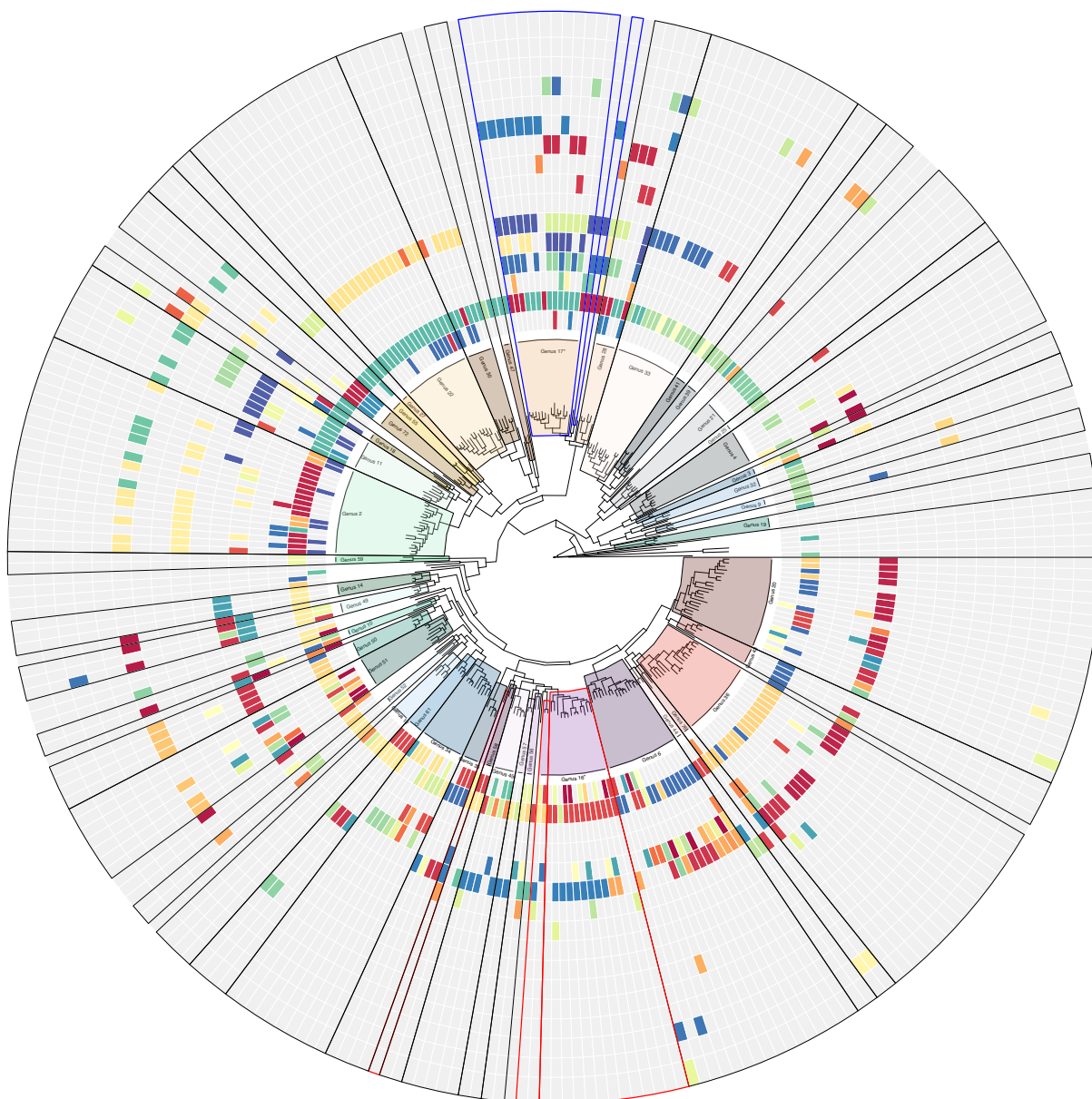


Figure 4.12: Some SCOG's repeated nucleotide sequence variants do not form monophyletic clades on the core genome tree and are shared across genomically distinct *Streptomyces* subgroups (>45.8% genome coverage; >88.8% genome identity). Each individual ring represents a single SCOG, with the order of the ring (from inner to outer) and their corresponding protein products as annotated on sequence record is provided in Table 4.7. The same color within a ring indicates identical nucleotide variants. Variants that either form monophyletic clades or are unique to a single taxon on the core genome tree are show in grey.

Table 4.7: List of 16 SCOGs in which repeated nucleotide variants are shared across distinct *Streptomyces* genomes and form non-monophyletic clades with their corresponding protein products as annotated in the sequence records. The order of the SCOGs corresponds to the order (from inner to outer) of the rings in the Figure 4.12.

Orthogroup ID	Protein Product	Total Non-Monophyletic
OG0002076	30S ribosomal protein S18	18
OG0002078	50S ribosomal protein L36	13
OG0002079	50S ribosomal protein L22	6
OG0002080	50S ribosomal protein L29	9
OG0002082	50S ribosomal protein L30	15
OG0002086	50S ribosomal protein L32	20
OG0002120	50S ribosomal protein L31	4
OG0002167	DUF4177 domain-containing protein	6
OG0002186	50S ribosomal protein L10	1
OG0002190	50S ribosomal protein L23	2
OG0002191	30S ribosomal protein S19	7
OG0002193	50S ribosomal protein L16	1
OG0002194	30S ribosomal protein S17	4
OG0002195	50S ribosomal protein L24	2
OG0002197	30S ribosomal protein S8	1
OG0002282	hypothetical protein	1

4.4 Discussion

4.4.1 Representative set of high-quality *Streptomyces* genomes reveals a small core genome

Studying prokaryotic pangenomes to understand adaptive mechanisms through their accessory genomes, or evolutionary patterns *via* the core genome, relies heavily on high-quality genomic sequences and accurate taxonomic placement. Pangenomic analyses are often affected by errors resulting from misannotations, contaminations, as well as fragmented and incomplete assemblies (Tonkin-Hill et al., 2023). These errors can underestimate the core genome, while simultaneously causing overestimation of the accessory gene size and the extent to which the taxa is considered “open”. This occurs when annotation software fails to detect genes or when genes are fragmented due to breaks in the assembly, leading to the exclusion of these genes from the core genome estimate (Tonkin-Hill et al., 2020).

To obtain a better overview of the *Streptomyces* genomes used in this study, I first examined the overall characteristics of *Streptomyces* genomes. The results in section 4.3.1 align with previous observations, which reported that *Streptomyces* genomes typically range from 6 to 15 Mb, with an average size of around 8 Mb. Additionally, the GC content of these genomes averages between 72% and 75%, and they generally contain approximately 7,000 protein-coding sequences (Bury-Moné et al., 2023; Lee et al., 2020; Subramaniam et al., 2020; Volff & Altenbuchner, 1998).

Given that only 17.9% of the *Streptomyces* analysed (section 4.3.1) were assembled to the complete or chromosomal level and the overall quality of the genomes was uncertain,

it was crucial to assess genome quality to avoid pitfalls associated with analysis and comparison of poor-quality sequences. My analysis revealed that some *Streptomyces* genomes in NCBI have contamination levels as high as 17.87% and completeness as low as 85.85%. Contamination of some genomes was expected, given the identification of over 2 million contaminated entries in GenBank (Steinegger & Salzberg, 2020) and estimates that around 10% of NCBI's prokaryotic genomes are contaminated (Astashyn et al., 2024). Human DNA contamination in bacterial genomes, particularly within small contigs of draft assemblies, is a common issue (Breitwieser et al., 2019), synthetic sequences like vectors, adapters, and primers are also recognised sources of contamination (Steinegger & Salzberg, 2020).

To address the concerns that arise from inclusion of poor quality genomes, I expanded on the analyses from chapter 3 to select a representative set of high-quality genomes. For each ANI species found in a connected group of STs in the MST tree (Figure 3.20), I selected a single species representative, prioritising genomes with the lowest contamination and highest completeness when multiple candidates were available (methodology section 4.2.2). This selection process resulted in a curated set of 295 genomes, each with a minimum completeness of 93.99% and contamination no higher than 4.21%.

This set of high-quality genomes was used to provide a clearer overview of *Streptomyces* genome content (results section 4.3.2). The presence of a large accessory genome, comprising 91.54% of all orthogroups (with no soft-core orthogroups) (Table 4.1), aligns with previous observations that the *Streptomyces* pangenome (considering *Streptomyces* as a genus spanning these 295 genomes) is open, as new genes continue to be added with

the sequencing of additional genomes (Caicedo-Montoya et al., 2021; Otani et al., 2022). However, previous studies have found that approximately one-seventh (600–1,018 genes) of *Streptomyces* genomes are shared across all members (Bury-Moné et al., 2023), and larger sets of single-copy orthologues are shared among *Streptomyces* species (575 and 453 orthologues) (Kusuma et al., 2021; Martín-Sánchez et al., 2019). If *Streptomyces* represented a single group of closely related organisms, we would expect a larger number of SCOGs to be shared. However, ANIm analysis (see section 4.3.4) indicates significant diversity within *Streptomyces*. For instance, some *Streptomyces* genomes share only 4.1% of their genome homology based on alignment length. This finding is consistent with the relatively low proportion of core and single-copy orthologous genes shared among *Streptomyces* genomes observed here. This large genomic diversity within *Streptomyces* suggests that this relatively small core genome and extensive accessory genome may create a somewhat misleading view of *Streptomyces* pangenome. Pangenome analyses are typically performed on closely related strains, where the core genome consists of genes essential for fundamental biological processes and provides insights into evolutionary relationships, while accessory genes primarily reflect niche adaptations. However, when applied to highly diverse groups like *Streptomyces*, the vast genomic divergence leads to an artificially small core genome and an bigger accessory genome. This makes it challenging to define a meaningful core genome size and accurately interpret evolutionary relationships. To gain a more accurate understanding, *Streptomyces* should be categorised into finer, more granular groups that better align with genus-level distinctions used in other bacterial taxa. Repeating the analysis using these refined groupings would offer a clearer picture of the genomic diversity within this highly complex and diverse

taxa.

4.4.2 A highly resolved core genome phylogeny supports conclusions of widespread misclassification in *Streptomyces*

The reclassification of *Streptomyces* to better understand their evolutionary relationships has been extensively explored in previous chapters. However, both 16S rRNA (chapter 2) and MLST (chapter 3) methods proved inadequate for resolving phylogenetic conflicts and were found to be sometimes unreliable proxies for taxonomy. Recent advancements in sequencing technology have significantly increased the availability of whole-genome sequences (see section 1.2.3), leading to more frequent adoption of genome-based taxonomy (see section 1.3.3). The initial aim of this chapter was to infer a core-genome phylogeny for *Streptomyces* to evaluate its resolution and determine if it aligns with the misclassifications identified in chapter 3 section 3.3.6. As presented in section 4.4.1 I identified 137 SCOGs present in all 295 genomes, which I consider as the candidate core genome for the genus *Streptomyces*. I constructed a core-genome tree by aligning and concatenating the coding sequences of these 137 SCGO genes and used RAxML to calculate a maximum likelihood tree with a distinct evolutionary model to fit each gene (methodology section 4.2.5). This approach led to the most comprehensive whole-genome-based phylogeny of *Streptomyces* attempted to date, with a well-supported topology (Figure 4.1). All splits, except one, had a transfer bootstrap expectation (TBE) value of 70% or higher, with the majority showing a TBE value of 100%, accounting for 91.5% of all internal nodes, indicating overall topological stability (section 4.3.3). These findings align with previous studies demonstrating that concatenated SCOG data provide sufficient information to reconstruct reliable and

highly resolved phylogenies (Keller & Ankenbrand, 2021), including for *Streptomyces*, though on a more limited scale (Kim et al., 2015). The node with the lower TBE value might be attributed to high similarity of the SCOG sequences among taxa that share very recent common ancestors. This issue could potentially be resolved in future work by examining SCOGs from a common ancestor closer to the leaf nodes, and identifying a larger set of SCOGs to improve resolution. Given that the topology of the remaining parts of the tree is well-defined, it would be feasible to reintegrate the more accurately resolved subtree back into the overall species tree.

The overall structure of the SCOGs phylogenetic tree (Figure 4.1) reflects the three-group framework suggested by the 16S phylogeny from chapter 2 section 2.3.3 (Figure 2.8). The ancestral group (clade 1) is the smallest, comprising 60 *Streptomyces* genomes. In contrast, more recent divergences have resulted in two larger clades: clade 2, comprising 87 genomes, and clade 3, the largest, with 148 genomes. The presence of these three main clades in a core-genome phylogeny suggests the divergence of three distinct evolutionary lineages from a common ancestor. In prokaryotes, genome expansion frequently results from gene duplication and recombination events, enabling bacteria to acquire new genes that foster functional innovation and ecological adaptation (Nikolaidis et al., 2023; Zhou et al., 2012). This genome expansion may have played a key role in the divergence of *Streptomyces* into three distinct lineages, allowing different taxa to exploit a variety of ecological niches and exposing different them to unique selective pressures. This hypothesis aligns with the frequent observation of genome expansions in *Streptomyces*, with some expansions reaching up to 1.03 Mb due to gene duplications in the chromosomal arms (Nikolaidis et al., 2023). These pressures promote

diversification within their core genomes, with specific adaptations favored in certain environments, ultimately leading to the formation of distinct lineages. These selective pressures promote diversification within their core genomes, with specific adaptations favored in certain environments, ultimately leading to the formation of distinct lineages. One way to identify genes under selective pressure is through comparative genomic analyses, which offer a powerful approach to studying these evolutionary processes. Such analyses allow the identification of genes under positive selection using dN/dS ratios (non-synonymous/synonymous substitution rates) and reveal gene presence/absence variations and expansion/contraction patterns across lineages.

As discussed in chapter 3 section 3.3.6, several genus- and species-level misclassifications were identified among genomes with the same species names in NCBI. This raises the question of how these misclassified genomes, as well as those correctly classified, are distributed across the SCOG tree. If both SCOG phylogeny and ANI analyses are congruent, we would expect genomes with low ANI coverage to be present in distinct clades, while genomes identified as closely related through ANI analyses would cluster in the same clade. However, if the distribution of genomes on the SCOG tree does not align with the ANI-based taxonomy—such as observing unrelated genomes grouped in a single clade or related genomes scattered across the phylogenetic tree—it would indicate a lack of congruence between ANI and core-genome phylogeny. This discrepancy would require a closer examination of the reliability of each method for taxonomic classification and determine which, if any, provides the most accurate classification.

Through the investigation of the distribution of genomes currently assigned to *S. griseus*, *S. clavuligerus*, and *S. rimosus* on the SCO phylogenetic tree (Figure 4.1), it

appears that the first explanation is more likely, and that assigned taxonomy is generally consistent with the tree topology. Specifically, not all genomes currently assigned to *S. griseus* cluster within the same clade. For example, genome GCF_000718525.1, which shares only 29% of its genome with other *S. griseus* genomes (see appendix Figure A.1) based on alignment length, is separated from the remaining genomes (Figure 4.1; red leaves). In contrast, genomes currently identified as *S. rimosus* in NCBI, which were previously confirmed by ANI analysis to share no less than 58% genome coverage (see appendix Figure A.2), appeared to be closely related on the tree. This implies that both methods can be complementary in the detection and reclassification of misannotated genomes. Notably, ANI is currently used by NCBI to improve the accuracy of prokaryotic genome assignments (Ciufo et al., 2018). This confirms that the continued use of ANI at NCBI is beneficial and should be maintained. Additionally, using both methods together can offer a more comprehensive understanding of the evolutionary relationships between organisms.

4.4.3 New genus boundary threshold reveals 79 distinct *Streptomyces* groups

In chapters 2 and 3, I identified significant genomic diversity among publicly available *Streptomyces* genomes, despite their current classification within the same genus in NCBI. My analysis revealed that some *Streptomyces* strains share only a small portion of their genomes by alignment length, even when they share identical 16S rRNA sequences (section 2.3.4), belong to the same MLST subgraphs (section 3.3.5), or are assigned the same species name in NCBI (section 3.3.6). This suggests that marker-based approaches to classification are not capturing sufficient genomic variation to accurately classify

Streptomyces.

In this chapter, I investigated the genomic diversity among a representative set of 295 *Streptomyces* genomes (methodology section 4.2.6), which were carefully selected to capture the entire diversity of *Streptomyces* currently available in NCBI while avoiding potential misinterpretations that could arise from poor-quality genomes. None of the genomes were re-annotated, as *Streptomyces* species have multiple start codons (Burger et al., 1998), making it impossible to determine the correct one solely from the sequence. The true start codon would require laboratory techniques, such as ribosome profiling, to accurately identify the translation initiation site. To explore the genomic diversity within this set of 295 *Streptomyces* genomes, I used Average Nucleotide Identity (ANI) to investigate the genomic similarity among them. The analysis revealed that for the majority of *Streptomyces* genomes (95.1%), less than 50% of their genome sequences were shared by alignment length (result section 4.3.4; Figure 4.2). I found that genome coverage was very low in some pairwise comparisons, such as between GCF_000367365.1 (*Streptomyces prunicolor*) and GCF_000380165.1 (*Streptomyces vitaminophilus*), with only 4.1% coverage observed. Although GCF_000380165.1 has recently been reclassified as *Wenjunlia vitaminophila* — a novel genus that was previously placed within *Streptomyces* (Madhaiyan et al., 2022)—this reclassification likely reflects the observed genomic dissimilarities. Despite this reclassification, some genomes within the *Streptomyces* classification still showed low genomic similarity, with as little as 5% shared alignment length, yet they remain assigned to *Streptomyces* genus (result section 4.3.4). These findings, along with recent proposals of six novel genera within the *Streptomyces* genus (Madhaiyan et al., 2022), raise questions about the potential number of genera that

could reasonably be assigned within this group.

The current lack of a precise definition for bacterial genera under the International Code of Nomenclature of Prokaryotes (ICNP) (Parker et al., 2015), combined with the absence of a robust quantitative definition (threshold levels for genus definition based on alignment identity and coverage) (Pritchard et al., 2015), creates uncertainty in bacterial classification. In this thesis, I do not aim to redefine the concept of genus, but rather to propose criteria for what a genus could represent in the context of Streptomycetes. Specifically, I propose that genera should consist of groups of organisms that are more closely related to one another than to any organisms outside the group, reflecting a consistent evolutionary relationship between all members of the genus. This reasoning aligns with previously published perspectives, which argue that if two genomes share less than 50% of their genomic material, it is likely that each genome shares more than 50% of its material with different, unrelated organisms (Pritchard et al., 2015). However, certain cases may complicate this approach. For instance, if genome A shares more than 50% of its genomic content with both genomes B and C, but genomes B and C share less than 50% with each other, this could lead to ambiguous interpretations of genus boundaries. In such cases, simply categorising them under the same genus may be problematic.

Therefore, I propose that a genus should comprise organisms that includes consistent and quantifiable genomic relationships across all members. Specifically, I suggest that by evaluating ANI comparisons of genome coverage and identity, all members of a genus should share more than a predetermined percentage of their genomic content with one another.

In my initial analysis I explored a range of genome coverage thresholds (40% to 85%) to identify thresholds at which *Streptomyces* genomes form distinct and coherent groups (k-complete graphs) that could be considered candidate genera (methodology section 4.2.6). At the upper threshold range (83.1% to 85% coverage), the results show a clear separation into monophyletic groups (Figure 4.3; result section 4.3.5). However, these thresholds lead to excessive separation, with 284 to 291 clusters, with the majority consisting of single representatives (singletons). This suggests that such high coverage thresholds, while promoting phylogenetic monophyly, may not correspond to the current genus concept, and almost certainly overestimate the number of genera, over-splitting *Streptomyces*. At lower genome coverage thresholds (43.8% to 48.4%), fewer clusters were observed (36 to 45 clusters), but resulted in phylogenetic inconsistency: a non-monophyletic clade on the phylogenetic tree. Despite this, it is likely that no single genome coverage threshold can define genus boundaries for *Streptomyces*. Genomic classification likely involves a range of coverage values due to evolutionary processes such as genome expansion and the acquisition or loss of genetic material, which are common in bacteria, including *Streptomyces* (Caicedo-Montoya et al., 2021; Nikolaidis et al., 2023). For instance, genome expansion can lower the reported shared genomic content between species, even when their core genomes remain the same. For example, if two *Streptomyces* species originally share 75% of their genes but one undergoes genome expansion (e.g., through duplications or horizontal gene transfer), the shared content may drop to 60%, not due to a loss of core similarity, but because of the increased genome size. Additionally, the presence or absence of plasmids—which vary across *Streptomyces* species (Caicedo-Montoya et al., 2021)—can also affect genome

comparisons, as differences in the number of plasmids present across taxa can affect the proportion of shared genomic content. Given these factors, it is likely that genus boundaries in *Streptomyces* are defined by a range of genome coverage values rather than a single threshold, reflecting the genomic complexity within the genus.

To estimate a threshold for genus boundary separation in *Streptomyces* for future analyses and to statistically and mathematically validate and refine these thresholds, I applied piecewise linear regression to pairwise comparisons of genome coverage and genome identity (see methodology section 4.2.6). Although the six-segment model has the lowest BIC value and provides the best statistical fit, it introduces complexity in interpreting the breakpoints (result section 4.3.5; Table 4.2). There is no clear visual separation between the different slopes, making it difficult to distinguish meaningful trends (see Figure 4.4). In contrast, the four-segment model, which suggests thresholds of 55.6% genome coverage and 88.1% genome identity, offers a more interpretable and practical approach (Figure 4.5). This model appears to show less overfitting and presents a clear separation between two distinct trends: one characterised by a high increase in genome identity with minimal additional genome coverage, likely corresponding to species-level comparisons, and the other by a rapid increase in genome coverage with a low increase in genome identity, suggesting genus-level comparisons Figure 4.5. This clearer differentiation makes the four-segment model a more robust and realistic framework for defining genus boundaries in *Streptomyces*. Additionally, there was a minimal change in BIC and RSS values between the four-segment and six-segment models, further supporting the conclusion that the four-segment model with breakpoints at 55.6% genome coverage and 88.1% genome identity is the better choice (Table 4.2).

Furthermore, this proposed threshold of 55.6% genome coverage aligns more closely with earlier suggestions of approximately 50% genome coverage for genus-level separation and ensures a more robust and practical classification.

I mapped the *Streptomyces* subgroups identified at $\geq 55.6\%$ genome coverage and $\geq 88.1\%$ genome identity onto the core-genome phylogeny (Figure 4.6). The resulting groups varied in size, with some containing only a single representative. It is possible that for these smaller subgroups, additional members are yet to be sequenced.

The non-monophyletic group (proposed genus 16) is situated in clade 3, where it is interrupted by proposed genera 42, 56, 57, 45, and 58. This disruption may suggest that these groups were once part of proposed genus 16 but have since accumulated enough changes to become more distantly related. Similarly, proposed genus 17, located in clade 2, is interrupted by singleton genus 66, which could indicate that they were also once a single genus that later diverged. These disruptions in monophyly could also be artifacts arising from incomplete genome data. Some genomes used in this study are not complete, and the missing information may have contributed to the observed non-monophyly. Alternative scenarios are possible, and analyses with more complete genome sequences may help resolve these ambiguities in the future.

4.4.4 Single-Copy Orthologues in *Streptomyces* predominantly reside in the core chromosomal region with extensions into the arms

In this study, I characterised the organisation of all 137 SCOGs across 60 complete *Streptomyces* genomes for which the chromosomal structure was well-resolved (methodology section 4.2.7). While the initial dataset included 295 *Streptomyces* genomes, only 60 met the criteria for analysis. The genomes selected for analysis had assemblies that were

either complete or at least at the chromosomal level, ensuring the necessary continuity to accurately map gene positions across the chromosome. Additionally, the *oriC* region in these genomes was located approximately at the 50% mark of the chromosome length. Given that previous research has shown that in *Streptomyces* *oriC* is centrally located with bidirectional replication (Bentley et al., 2002; Bury-Moné et al., 2023), genomes where *oriC* was positioned on the chromosome arms were excluded to avoid interpretational bias.

The tripartite structure of the *Streptomyces* chromosome, consisting of a central core and two arms, was identified by Bentley et al. in 2002 through the study of *Streptomyces coelicolor* A3(2) (Figure 4.13). Bentley’s research identified that the core region of the chromosome is slightly off-center, spanning from approximately 1.5 Mb to 6.4 Mb. The right arm was found to be somewhat shorter, at about 2.3 Mb, while the left arm extends roughly 1.5 Mb. More recent sequencing of a greater number of assemblies of *Streptomyces* allows us to update this assessment.

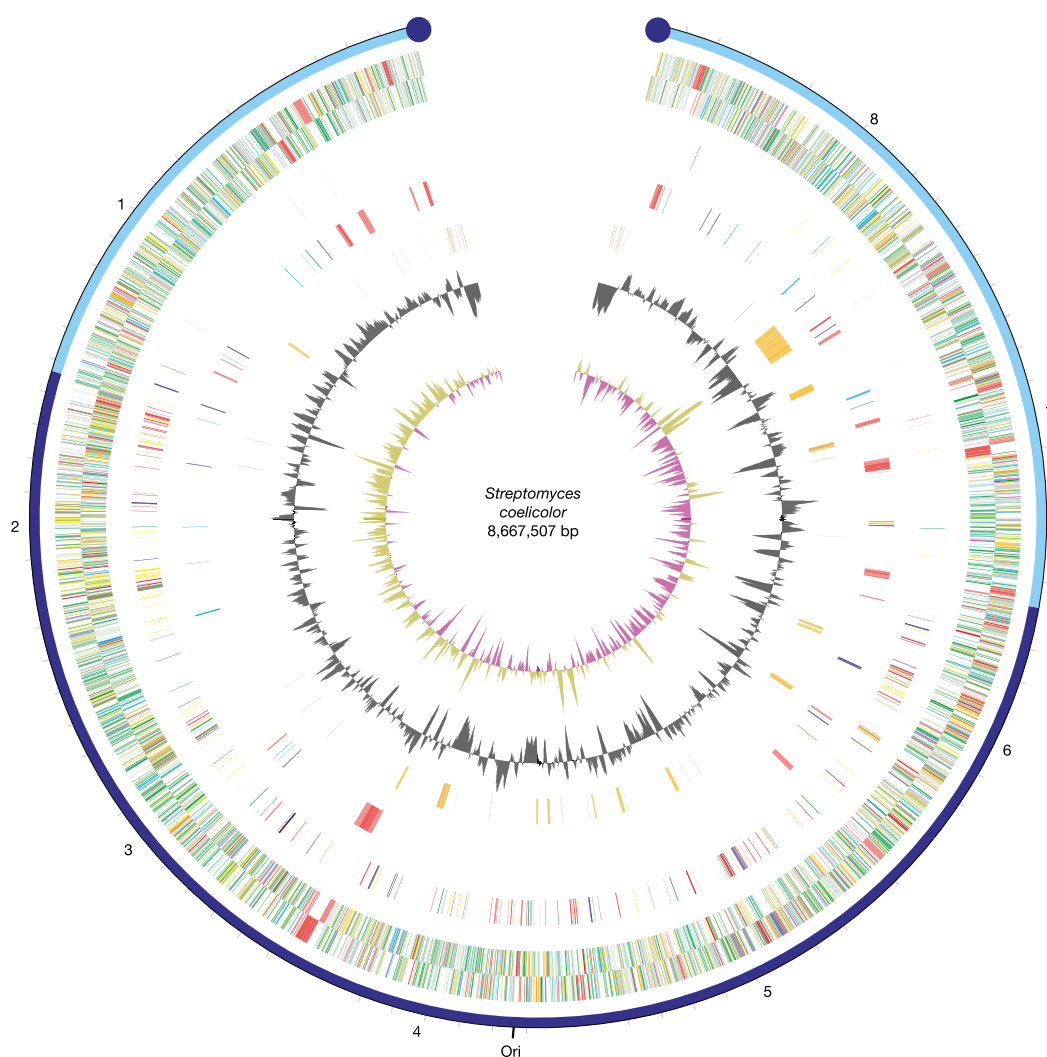


Figure 4.13: Tripartite structure of the *Streptomyces coelicolor* A3(2) chromosome. The outermost ring shows the length of the chromosome in megabases, with labeled regions: the core (dark blue) and the arms (light blue). Moving inward: Rings 2 and 3: Represent genes identified on the reverse and forward strands, respectively. Genes are color-coded based on their functions: black: energy metabolism; red: information transfer and secondary metabolism; dark green: surface-associated functions; cyan: degradation of large molecules; magenta: degradation of small molecules; yellow: central/intermediary metabolism; pale blue: regulators; orange: conserved hypothetical proteins; brown: pseudogenes; pale green: unknown functions; grey: miscellaneous functions. Ring 4: shows selected essential genes involved in processes like cell division, DNA replication, transcription, translation, and amino acid biosynthesis, using the same color scheme as Rings 2 and 3. Ring 5: shows selected contingency genes: red: secondary metabolism; pale blue: coenzymes; dark blue: conservation; green: gas vesicle proteins. Ring 6: shows mobile genetic element such as transposases in brown and putative laterally acquired genes in orange with >1 shown in khaki and <1 shown in purple. The diagram also marks the origin of replication (Ori) and terminal protein. Figure adapted from Bury-Moné et al., 2023.

Given that this tripartite structure and their boundaries were established when only a limited number of *Streptomyces* genomes were available and was primarily based on the structure of *Streptomyces coelicolor*, it remains possible that the core regions of other *Streptomyces* genomes might be more variably positioned, potentially shifting to and from either end of the chromosome. Therefore, I considered the core part of the chromosome to be located between the 25% and 75% points of the chromosome length, and regions outside this threshold to be the terminal arms (result section 4.3.6). Although, these boundaries are approximations, they closely correspond to previously proposed cut off of the tripartite structure proposed by Bentley and aimed to capture a more general organisation of *Streptomyces* chromosome while accounting for potential variations.

As previously reported, *Streptomyces* genomes exhibit genetic compartmentalisation, with core genes clustered in the central region and less conserved genes, including those related to antibiotic production and niche adaptation, located in the terminal regions (arms) (Bury-Moné et al., 2023; Lioy et al., 2021). Observations of large deletions near the chromosome arms, accounting for roughly 33% of the genome size, under stressful conditions such as UV light and heat, as well as spontaneous deletions, suggest that these regions are particularly prone to recombination in *Streptomyces* genomes (Widenbrant et al., 2007). Further analysis of genome shuffling, focusing on the type and position of genes, revealed that core genes are typically situated in the stable central backbone of the chromosome suggesting a protective effect of the core backbone against genetic instability (Lorenzi et al., 2021).

In this study, I demonstrated that, in my set of *Streptomyces*, more than half (55.5%) of SCOGs (shared across all *Streptomyces* genomes) are located in the central part of the chromosome, consistent with previous observations (results section 4.3.6 and Figure 4.8 green). Additionally, I identified 36 SCOGs (27%) that $\geq 50\%$ of investigated SCOGs were situated within this core region, with only occasional instances where they slightly extend beyond the defined boundaries (Figure 4.8 purple). However, given that the boundaries adopted in this study are approximations, and recognising that slight shifts in the core region have been observed in different *Streptomyces* genomes, it is likely that these genes should be considered to reside within the core region of the chromosome.

SCOGs identified as core-exclusive and predominantly core-located are likely strategically positioned within the stable central region of the chromosome to ensure their preservation and functionality across various *Streptomyces* taxa. The location of core genes in this area likely reflects evolutionary pressure to protect essential genetic elements from the higher rates of recombination and instability typically observed in chromosome arms (Lorenzi et al., 2021). This positioning may be crucial for maintaining the integrity of fundamental cellular processes and physically-interacting protein complexes. This strategic placement is particularly important given that many of these SCOGs include structural ribosomal proteins, DNA replication and repair components, and transcriptional regulators (results section 4.3.6), all of which are likely essential for the survival.

Additionally, I found that proteins involved in the synthesis of structural ribosomal components (50S and 30S) were largely co-located and are found in the core region of the chromosome. This suggests that their co-location is likely an evolutionary strategy

to minimize the risk of disruption, even if recombination events occur within this region. It may also help ensure that all essential genes required for ribosome assembly are transferred together as a functional unit. By clustering these essential ribosomal proteins together in the core, *Streptomyces* ensures the integrity and reliability of ribosome assembly, which is critical for efficient protein synthesis and, ultimately, the bacterium's survival and adaptability in various environments.

Nevertheless, I found that 25 SCOGs (17.5%) were mainly located on one or other arm of the chromosome (red). Since the terminal arms of *Streptomyces* chromosomes are known to be unstable (Lorenzi et al., 2021), this suggests that these genes may be prone to recombination. The presence of SCOGs in the chromosome arms could be a strategic adaptation, allowing these genes to benefit from the adaptive advantages of rapid evolution. This positioning might enhance the emergence of novel traits and *Streptomyces* ability to adapt to dynamic environments and/or competition.

Noting that *Streptomyces* originated around 380 million years ago (McDonald & Currie, 2017b), it is possible that these SCOGs were acquired relatively recently. Their presence across all *Streptomyces* species might indicate that they confer certain benefits, even if they have not yet been incorporated into the core region of the chromosome. Alternatively, these genes might have been beneficial in the past but are now being gradually displaced from the core region by genes that offer more substantial advantages. This shift could reflect changing evolutionary pressures faced by *Streptomyces*. Yet another possible explanation is that the placement of some SCOGs at the arms of the chromosome serves as a protective mechanism against challenges to the population. While SCOGs are generally expected to reside in the core genome due to their essential

functions, which benefit from the stability of that region (Bentley et al., 2002), this conventional view may overlook the potential strategic value of positioning certain genes at the arms of the chromosome. The arms of the chromosome, though more prone to recombination, might offer a dynamic environment where these SCOGs can evolve more rapidly in response to external pressures (Bury-Moné et al., 2023), acting as a reservoir for adaptive potential. This scenario is supported by the observation that none of these SCOGs appear to be randomly distributed. Distinct patterns emerge in their distribution on the chromosome: SCOGs are either consistently found in the core, while others are always positioned at the arms. This spatial separation could be evolutionarily advantageous. By localising certain SCOGs in the arms, a "pool" of highly recombining genes may be maintained, fostering increased genetic diversity across the population. Such diversity could prove crucial for the species' long-term survival, allowing the population to rapidly adapt to fluctuating environmental conditions.

4.4.5 Phylogenetic distribution of Single-Copy Orthologue nucleotide variants reveals non-monophyletic patterns in *Streptomyces*

Initially, bacteria were believed to reproduce exclusively through clonal cell division, with minimal or no genetic exchange (Smith et al., 1993). It is now widely recognised that HGT shapes the evolution of bacterial lineages (Arnold et al., 2022). Unlike vertical inheritance, which passes genes from parent to offspring, HGT enables the acquisition of new genes from unrelated organisms which drive genetic diversity, rapid evolution and adaptations in bacterial populations. It is well reported that genes that provide bacteria with advantageous traits such as antibiotic resistance, virulence factors and BGCs can often spread through populations via HGT (Barlow, 2009). There are three mechanisms

of HGT in bacteria: conjugation (transfer of DNA through direct cell-to-cell contact, often via plasmids), transformation (uptake of free DNA from the environment), and transduction (transfer of DNA by bacteriophages) (Thomas & Nielsen, 2005).

The exchange of genetic material via HGT has also been observed in *Streptomyces*, predominantly through conjugation, although smaller-scale occurrences of other mechanisms (transformation and transduction) have been noted as well (Morino & Takahashi, 1997; Roelants et al., 1976; Stuttard, 1979). Furthermore, the presence of genes involved in DNA recombination and transposase activity within the core genome of *Streptomyces* suggests that HGT may play a role in shaping the core genome (Zhou et al., 2012). Incidences of HGT within the central chromosomal regions of *Streptomyces* have also been reported (Choufa et al., 2022).

During my investigation of SCOGs' locations on the chromosome, I discovered that some SCOGs were positioned outside the core chromosomal region, suggesting potential genomic instability (discussion section 4.4.4). This observation, along with reports of HGT in *Streptomyces*, raised the question of whether any nucleotide variants in SCOGs were acquired through mechanisms other than vertical inheritance (Choufa et al., 2022). To explore this, I examined the distribution of each SCOG nucleotide sequence variant across the phylogenetic tree (methodology section 4.2.8), observing several distinct patterns in their distribution.

I found that the majority of SCOGs (62%) were represented by a unique nucleotide variant in each representative *Streptomyces* genome (result section 4.3.7). This suggests that no two genomes share the exact same nucleotide sequence for within these SCOGs, indicating that while the gene is conserved in terms of its presence and likely its function,

its DNA sequence varies between genomes. The unique variants may reflect speciation events, where orthologues evolved slightly differently in each lineage over time.

I also identified a set of 52 SCOGs where at least one nucleotide sequence variant appeared more than once, suggesting that some *Streptomyces* share the same nucleotide sequence (Table 4.3). I observed several distinct patterns in the distribution of these 'repeated' nucleotide sequence variants across the phylogenetic tree. Such as, I observed that 13 (25%) out of these 52 SCOGs exclusively consisted of repeated nucleotide variants, which were shared among closely related *Streptomyces* species (with $\geq 45.8\%$ genome coverage and $\geq 88.1\%$ genome identity) and exhibited a monophyletic pattern on the phylogenetic tree (Figure 4.10). Additionally, there were 38 additional SCOGs where at least one repeated nucleotide variant was found to be monophyletic, though not all variants showed this pattern.

But I also identified 38 (52%) SCOGs where at least one repeated nucleotide sequence variant was shared among closely related *Streptomyces* subgroups (with $\geq 45.8\%$ genome coverage and $\geq 88.1\%$ genome identity), where these were not organised in a monophyletic pattern. Additionally, I found 16 (30.8%) cases where repeated nucleotide sequences in SCOGs were shared across distantly related *Streptomyces* subgroups (with $> 45.8\%$ genome coverage and $> 88.1\%$ genome identity), appearing in a non-monophyletic arrangement (Figure 4.12 and Table 4.5). This could be the result of certain lineages acquiring mutations that provided some selective advantage, leading to variations in the nucleotide sequences. For instance, if a group of *Streptomyces* initially shared the same nucleotide sequence for a SCOG gene from a common ancestor, some lineage might have diverged due to selective pressures or adaptations, resulting in different

nucleotide sequences. Meanwhile, other lineages may have retained the original form of the nucleotide sequence. Additionally, it is possible that two *Streptomyces* genomes independently acquired the same mutation at different points in their evolutionary history due to similar selective pressures. Such convergent mutations could cause these genomes to exhibit slightly different sequences compared to others, despite originating from distinct evolutionary events. It is also possible that strong selective pressures may have driven the exchange of genes crucial for the organism's survival, potentially spreading these genes across populations through HGT. Previous studies have shown that antibiotic-induced selective pressure can promote specific genetic variations in certain groups. For instance, mutations in the 50S ribosomal gene *rplB*, which are associated with resistance to the azithromycin protein synthesis inhibitor, have been found to be susceptible to HGT (Manoharan-Basil et al., 2021). Given that a wide range of antibiotics target ribosomal subunits, particularly the 50S and 30S subunits (Champney, 2020), and considering that the majority of these SCOGs with scattered patterns on the phylogenetic tree encode structural ribosomal proteins (50S and 30S), it is plausible that HGT might influence genes involved in the function and assembly of these subunits. Thus, HGT could extend to what I currently consider core genes of *Streptomyces*.

Conclusions and Recommendations

5.1 Conclusions

Members of the genus *Streptomyces* are known for their ability to synthesise a diverse range of bioactive compounds and enzymes with applications in pharmacy, biotechnology and agriculture (Procópio et al., 2012) (chapter 1 section 1.6). Enumeration of bioactive compounds encoded in the genomes of and produced by *Streptomyces* suggest they still have great potential as reservoirs of new natural products that could be explored to help mitigate the effects of antimicrobial resistance. Comparative genomics has contributed to the identification of biosynthetic gene clusters with potential pharmaceutical or biotechnological applications from closely related genomics sequences (chapter 1 section 1.2.3 and section 1.2.4). However, the lack of a robust taxonomic classification of the genus *Streptomyces* (Mispelaere et al., 2024), complicates the application of these methods (chapter 1 section 1.7). Resolving their taxonomy would benefit comparative genomics-based searches for novel bioactive compounds with potential pharmaceutical applications.

Despite several approaches having been proposed for taxonomic classification, and a codified set of rules for how to compose names for bacteria, there is currently no

universally accepted method. This is largely due to freedom of taxonomic opinion and the dynamic nature of taxonomy, which continually evolves in response to the growing volume of biological and sequence data, advances in technology, and our improved understanding of biology (chapter 1 section 1.3). Commonly used sequence-based methods for taxonomic classification include single-gene, multi-gene, and whole-genome phylogenies, as well as techniques like MLST and ANI. However, these approaches frequently yield conflicting results, reflecting the varying levels of genomic data used to delineate species.

The work presented in this thesis offers a thorough investigation of taxonomic relationships among all publicly available *Streptomyces* genomic sequences. I used several classification approaches, to establish more accurate taxonomic classifications to provide improved input sets for pangenomic analyses, and improve our understanding of evolutionary relationships within this taxonomically complex genus.

The 16S rRNA gene is a widely-used phylogenetic marker for studying microbial diversity due to its universal presence in bacteria and its relatively high rate of sequence variation (Clarridge, 2004; Woese & Fox, 1977). Traditionally, 16S rRNA sequences were grouped into operational taxonomic units (OTUs) at a 97% similarity threshold, later revised to 98.7%, but it is now more common to use zero-radius OTUs (zOTUs), where each unique sequence is treated as a distinct taxonomic unit (Edgar, 2018c). Historically, *Streptomyces* species have been classified based on 16S rRNA differences, with one of the most comprehensive 16S phylogenies published in 2012 (Labeda et al., 2012). With the availability of new *Streptomyces* genomic and 16S data, in chapter 2 I updated the *Streptomyces* phylogenetic tree using all current genomic data, and

assessed the congruence between 16S rRNA-based and whole-genome classifications.

I downloaded 16S rRNA sequences from Greengenes, RDP, SILVA, and NCBI, starting with an initial dataset of 48,981 full-length *Streptomyces* sequences, which I refined to 14,239 high-quality, non-redundant sequences after excluding poor quality sequences. I constructed the most comprehensive 16S rRNA sequence-based phylogenetic tree of *Streptomyces* to date, using the Maximum Likelihood method (section 2.3.3). Despite a relatively low level of bootstrap confidence, I found a clear partitioning of *Streptomyces* into three major clades. I investigated the distribution of pharmaceutically and agriculturally important *Streptomyces* species, where I found that some taxonomic assignments placed the same species name at multiple points in the topology, inconsistent with a common lineage, and occasionally placed the same names in more than one major clade. By clustering zOTUs, I observed that unique 16S sequences are often associated with more than one current database taxonomic assignment, and that unique 16S sequences can often be associated with more than one *Streptomyces* species, or even candidate genus, when whole-genome approaches are used for classification (section 2.3.2 and section 2.3.4). By surveying complete *Streptomyces* genomes, I identified the distribution of 16S rRNA sequences in sequenced isolates, finding numerous cases where a single genome contains several distinct 16S sequences (section 2.3.4). The results presented in chapter 2 demonstrate that there is a one-to-many relationship between 16S rRNA sequence and *Streptomyces* species, and a one-to-many relationship between *Streptomyces* species and 16S rRNA sequence. Without specific corrections, 16S metabarcoding will overestimate both the number of *Streptomyces* species and their relative abundances.

Given that 16S rRNA phylogeny and clustering are not reliable proxies for genome-based taxonomy in *Streptomyces*, this raises the question of whether approaches that rely on more sequence data from each genome could offer a more accurate classification. In chapter 3, I chose to use MLST, which uses more sequence-data for taxon delineation and is still actively used for exploring evolutionary relationships between taxa (Maiden et al., 1998b; Wu et al., 2023a). The pre-existing canonical *Streptomyces* MLST scheme provided by pubMLST comprised six markers and 237 STs, with only three new STs reported since 2016, despite a significant increase in available whole-genome sequences.

I surveyed all 2,276 publicly available *Streptomyces* assemblies as a means of identifying existing and novel STs. I identified 568 novel STs (section 3.4.1), which have now been incorporated into the public pubMLST resource. However, I observed that the existing set of markers does not allow for typing of all *Streptomyces* isolates. This limitation stems from the strict requirement for single-copy marker sequences, even though *Streptomyces* often carries multiple non-identical copies of the 16S gene, or from instances where some genomes are missing at least one allele. I found that almost two-thirds of pubMLST STs currently lack a corresponding genome in NCBI. By applying the *Streptomyces* scheme to sister genera, I found that non-Genbank STs could be associated with taxa outside the *Streptomyces* genus (section 3.4.2). Additionally, I identified cases where the current set of all markers might be present in genera other than *Streptomyces*, potentially leading to misclassification at the genus level if not supplemented by additional phenotypic or genomic verification.

Graph-based analysis of the updated scheme subdivides *Streptomyces* into 278 distinct groups, and I found that sequencing more *Streptomyces* genomes is unlikely

to unify unconnected components (section 3.3.3). This implies that there is natural structure within *Streptomyces* reflected in ST profiles. I used ANI to determine species- and genus-level boundaries within and between MLST groups. Although a single ST can represent single species (with $\geq 50\%$ coverage and $\geq 95\%$ genome identity), I found that a single connected group of STs can unite multiple candidate taxa at different levels ($\geq 50\%$ genome coverage for genus and $\geq 95\%$ genome identity for species), and that members of the same taxon may be scattered across unconnected ST groups. Through further surveying of *Streptomyces* genomes, I also identified taxonomic misassignment within the genus, where individual species (with $\geq 50\%$ genome coverage and $\geq 95\%$ genome identity) are assigned conflicting names, or distinct candidate species or genera are assigned the same name in NCBI.

The findings presented in chapter 3 underscore that extensive reclassification of the *Streptomyces* genus is needed. Additionally, current MLST subgraphs do not align with whole-genome based taxonomy and, as a result, are unlikely to improve the efficiency and accuracy of pangenomic analyses. Therefore in chapter 4, I explored alternative taxonomic classifications based on whole-genome sequence data.

In chapter 4, to address potential issues related to poor-quality assemblies, I built upon the analyses from chapter 3 by selecting a single high-quality representative per species from each MLST-connected component, that I took forward for a more complete phylogenetic analysis (section 4.3.1). These representatives were chosen based on the lowest genome contamination and the highest genome completeness. Using 137 single-copy orthologues obtained through whole-genome comparisons across this high-quality 295 representative assemblies I was able to infer a highly resolved Maximum Likelihood

phylogeny and a robust topology, that recapitulates the three group structure (section 4.3.3). I also observed that the core-genome tree suggests the same misclassifications identified in chapter 2. Specifically, assemblies confirmed by ANI as representing distinct species do not cluster within the same clades, while species with conflicting names are found within the same clades (section 4.4.2). However, I did find evidence of horizontal gene transfer affecting even highly conserved core genes (section 4.4.5). This requires further investigation to determine whether some core genes have been acquired by HGT and to assess any potential impact on the core-genome phylogeny presented in this thesis.

In chapter 4, I attempted to estimate a more quantitative candidate genus boundary for *Streptomyces*, revealing 79 distinct groups, with all but one being monophyletic in the core-genome phylogeny (section 4.4.3). The full list of members for each genus and species is provided in **Supplementary Data 12** in Kiepas_et_al_2023_SCOG. The work demonstrated that some genomes have been assigned genus designations as *Streptomyces* despite sharing as little as 4% of the genome content (section 4.3.4). This finding is consistent with previous proposals advocating for the reassignment of certain *Streptomyces* genomes to novel genera, such as the reclassification of GCF_000380165.1 into *Wenjunlia*. However, the 79 distinct groups identified in this study suggest that the scope of necessary taxonomic reassessment extends far beyond the cases currently being addressed. These findings suggest that the current designations might not be reliable for analyses that require accurate taxonomic classification, such as pangenomic analyses aimed at identifying novel compounds with potential commercial or pharmaceutical applications.

The subgroups identified in this thesis provide a more meaningful framework for calculating pangenomes, helping the community avoid the issues that arise from assuming that existing taxonomic assignments are correct—such as the inclusion of several distinct genus-level groups under a single *Streptomyces* genus. By refining these classifications, researchers can prevent the misallocation of resources on inappropriate groupings and achieve more accurate genomic analyses.

The distribution of SCOGs across the phylogeny challenges the traditional view that essential (core) genes, often used to trace evolutionary history, are stable and representative of the entire genome’s evolution (result section 4.4.5). Instead, their presence across distantly related groups and the occurrence of some SCOGs at the chromosome arms suggest that recombination may play a more significant role in shaping the genome over time.

While respecting the principle of taxonomic freedom—the right of scientists to independently classify organisms based on their interpretations of data—the findings in this thesis strongly suggest that the community should prioritise whole-genome-based taxonomy over single-gene or multi-gene (MLST) approaches. Although single-gene and MLST methods do work in some cases, their limitations—particularly their inability to capture the full complexity of evolutionary relationships—are significant. In contrast, whole-genome sequencing provides a more comprehensive and accurate framework for species delineation and higher taxonomic ranks, resulting in more reliable and consistent classifications. This recommendation is further supported by recent advancements in sequencing technology, which have made whole-genome data far more accessible and affordable, particularly in high-income countries, though costs may still be a barrier in

low- and middle-income countries

5.2 To rename or not to rename - this is the question

In this thesis, I identified several instances of misclassification within the genus *Streptomyces* and identified several genomically distinct subgroups. The key question is whether these should be reassigned to different or novel species, and even distinct genera. While there is considerable taxonomic freedom and some may have differing opinions, it is crucial to consider the practical implications of nomenclature changes. Reclassifications should be guided by community judgment to ensure that they reflect both scientific accuracy and practical considerations.

Reclassifying organisms based solely on genomic data has led to controversial decisions that could have lead to adverse clinical outcomes. For example, renaming *Ochrobactrum* spp. (occasional opportunistic pathogens) to *Brucella* spp. (highly infectious, notifiable pathogens) has raised concerns due to the significant difference in aetiology, diagnosis, treatment and prophylaxis for the two organisms (Moreno et al., 2023). A similarly situation has also arisen in the *Mycobacterium* genus (Meehan et al., 2021). These cases demonstrate that taxonomic reassignments can cause considerable clinical confusion and negatively impact patient care, highlighting the need for a cautious approach.

Achieving complete consensus on taxonomic classifications within *Streptomyces* and other complex bacterial lineages may be unlikely due to the wide range of opinions in the field. Advances in genome sequencing are enabling more precise taxonomic assignments and frequent renaming of strains and taxa, which can disrupt continuity

with historical literature and established practices. Nevertheless, these advancements also offer potential solutions through whole-genome strain classification methods, such as LINgroups and genomeRxiv (Mazloom et al., 2022; Tian et al., 2020). These methods provide a neutral, genome distance-based framework for resolving alternative taxonomic classifications. While prokaryote classification inevitably involves some degree of value judgment, and complete agreement may remain elusive even with perfect sequence data and extensive phenotyping, future developments may improve our ability to bridge differing taxonomic perspectives.

5.3 Future work

5.3.1 Is horizontal gene transfer occurring in the core genome of *Streptomyces*?

Considerably more work will need to be done to determine whether HGT truly occurs - in *Streptomyces* - for single copy orthologues. One approach to investigate HGT could involve comparing the compositional structures of genomic sequences, such as GC content. Genomes have distinct GC content patterns shaped by a mix of environmental and genetic influences (Hildebrand et al., 2010). HGT genes can be spotted because their unique characteristics (such as GC content) contrast with those of the host genome, potentially signaling their foreign origin. This method was previously used to identify genes in *Escherichia coli* and *Salmonella* that were acquired through means other than vertical inheritance (Lawrence & Ochman, 1997). However, this method has limitations. While this method can be useful, it primarily identifies recent HGT events. Ancient events may not be detectable with this approach due to potential changes

in compositional patterns over time so that they adjust to their new genome pattern (gene amelioration) (Daubin et al., 2003; Lawrence & Ochman, 1997). Additionally, this method may fail to detect HGT when the transferred genes share a similar GC content with the recipient genome. For example, Lawrence and Ochman found that 425 horizontally transferred genes in *E. coli* could not be identified because their GC content was not atypical.

However, greater confidence in detecting evidence of HGT could potentially be achieved through a phylogeny-based approach, which compares the phylogeny of the target gene with a well-resolved species tree to identify unusual gene distribution patterns across the tree (David & Alm, 2011). For instance, Kim et al. implemented this approach and uncovered potential HGT events in 548 gene families within the *Chlamydiae* phylum. A number of bioinformatic tools based on explicit phylogenetic approach for HGT prediction have been developed including Ranger-DTL (Bansal et al., 2018) that could prove useful in providing evidence to support the detection of HGT events in single-copy orthologues of *Streptomyces*.

5.3.2 Explore the impact of HGT of single-copy orthologues in the context of *Streptomyces*

HGT can affect the topology of phylogenetic trees by making taxa appear more closely related than they truly are when horizontally acquired genes are included (Philippe & Douady, 2003). In my analysis of *Streptomyces*, if HGT is identified in the single-copy orthologues (SCOGs) used for phylogenetic inference (section 4.4.2), it will be important to assess how these genes affect the tree's topology. This can be done by excluding SCOGs with strong HGT evidence and recalculating the tree. The new

tree can then be compared to the original using the Robinson-Foulds (RF) distance, a metric that quantifies differences between trees by counting the number of differing splits (bipartitions) (Robinson & Foulds, 1981).

5.3.3 Exploring the bioactive potential of *Streptomyces*

The exploration of the *Streptomyces* bioactive compound repertoire could be explored *in silico*. As outlined in section 1.2.3 and section 1.2.4, both comparative genomics and pangenomics have proven crucial for identifying novel bioactive compounds with potential pharmaceutical and biotechnological applications, especially through the analysis of closely related genomic sequences. Since these approaches depend on robust taxonomic classification — which I have established as discussed in section 4.4.2 — a logical next step is to calculate the pangenomes of key species and clades. This could include the 79 distinct groups of *Streptomyces* identified in section 4.3.5, as well as genomes found to represent the same species in section 3.3.5. Calculating these pangenomes will enable us to assess the genomic diversity and openness of each subgroup or species, allowing us to delve into their accessory genes. This analysis can reveal genes with specialised functions that hold promise for clinical or agricultural applications. Furthermore, comparing pangenomes can help uncover BGCs that are not universally present across strains but may be involved in the production of unique or rare secondary metabolites.

Additionally, conducting genome mining analyses using tools like antiSMASH across all representative genomes could aid in identifying and enumerating BGCs encoded in *Streptomyces* genomes. This approach has already led to the discovery of humimycins, a novel class of antibiotics demonstrated to be effective against methicillin-resistant *Staphy-*

lococcus aureus in murine models (Chu et al., 2016). The discovery of unique or rare BGCs through this approach could lead to novel bioactive compounds with previously unexplored pharmaceutical properties, opening new avenues for drug development.

Additionally, by mapping these biosynthetic gene clusters onto the *Streptomyces* SCOGs phylogeny could help us to evaluate the evolutionary patterns and distribution of these BGCs, revealing clusters that are either lineage-specific or acquired through horizontal gene transfer. The discovery of unique or rare BGCs through this approach could lead to novel bioactive compounds with previously unexplored pharmaceutical properties, opening new avenues for drug development.

CHAPTER 8

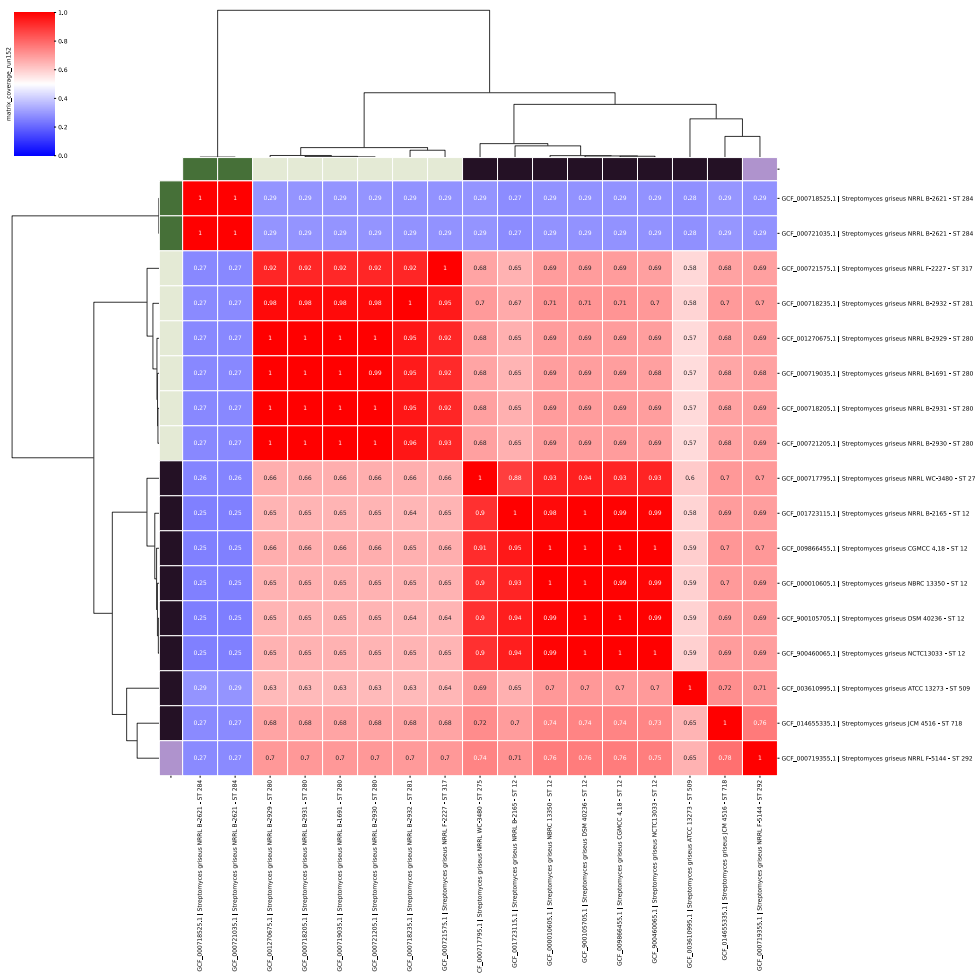


Figure A.1: pyANI genome coverage of genomes currently named as *S. griseus* in NCBI.

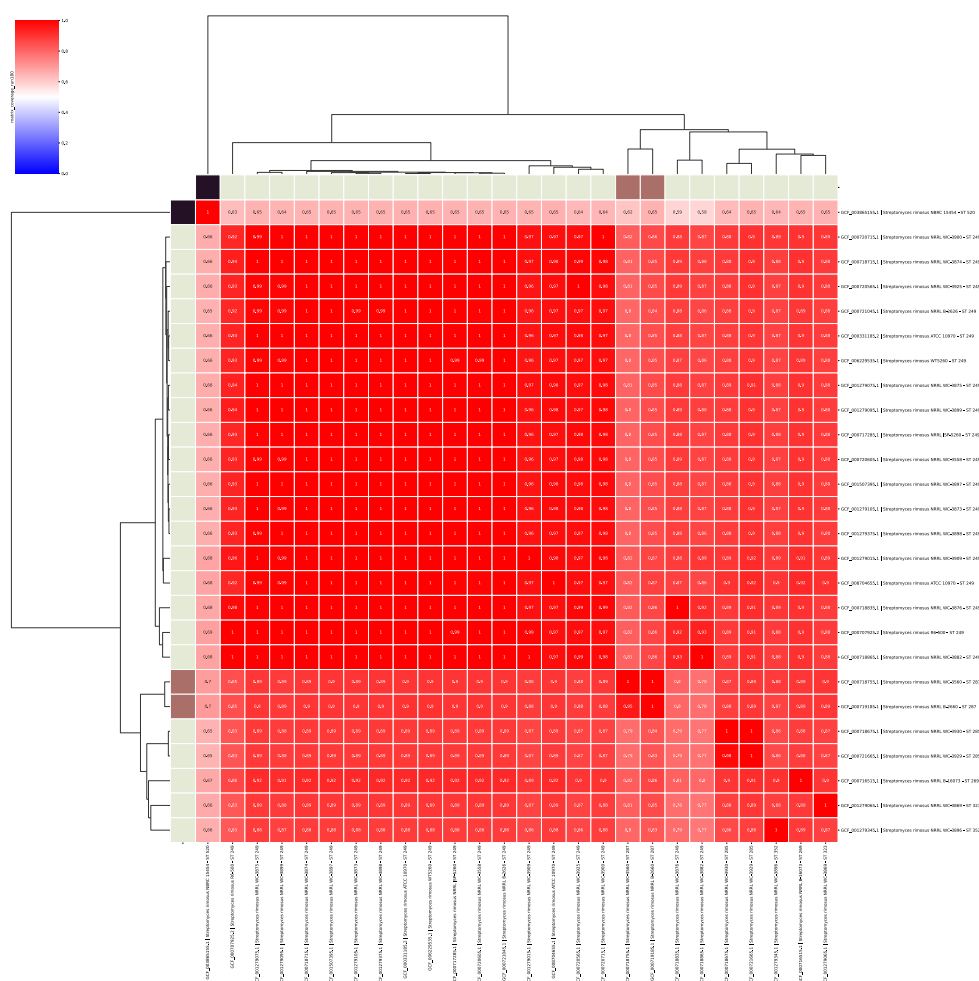


Figure A.2: pyANI genome coverage of genomes currently named as *S. rimosus* in NCBI.

Bibliography

- Aanensen, D. M., & Spratt, B. G. (2005). The multilocus sequence typing network: mlst.net. *Nucleic Acids Research*, 33, 728–733. <https://doi.org/10.1093/nar/gki415>
- Acaban, M. B., Sarı, A., Demirbağ, H. O., Ersöz, G., & Aktaş, S. (2024). The effects of topical mafenide acetate application on skin graft survival in bacterial contaminated wounds. *Burns*, 50(2), 433–443.
- Akaberi, M., Sahebkar, A., & Emami, S. A. (2021). Turmeric and Curcumin: From Traditional to Modern Medicine. *Advances in experimental medicine and biology*, 1291, 15–39. https://doi.org/10.1007/978-3-030-56153-6_2
- Alam, M. T., Merlo, M. E., Takano, E., & Breitling, R. (2010). Genome-based phylogenetic analysis of *Streptomyces* and its relatives. *Molecular Phylogenetics and Evolution*, 54(3), 763–772. <https://doi.org/10.1016/j.ympev.2009.11.019>
- Alhomoud, F., Aljamea, Z., Almahasnah, R., Alkhalifah, K., Basalelah, L., & Alhomoud, F. K. (2017). Self-medication and self-prescription with antibiotics in the middle east—do they really happen? a systematic review of the prevalence, possible reasons, and outcomes. *International journal of infectious diseases*, 57, 3–12.
- Alibu, V. P., Richter, C., Voncken, F., Marti, G., Shahi, S., Renggli, C. K., Seebeck, T., Brun, R., & Clayton, C. (2006). The role of *Trypanosoma brucei* mrpa

- in melarsoprol susceptibility. *Molecular and biochemical parasitology*, 146(1), 38–44.
- Allcock, S., Young, E. H., Holmes, M., Gurdasani, D., Dougan, G., Sandhu, M. S., Solomon, L., & Török, M. E. (2017). Antimicrobial resistance in human populations: challenges and opportunities. *Global Health, Epidemiology and Genomics*, 2, e4. <https://doi.org/10.1017/gheg.2017.4>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2)
- Amusengeri, A., Khan, A., & Bishop, T. (2022). The Structural Basis of *Mycobacterium tuberculosis* RpoB Drug-Resistant Clinical Mutations on Rifampicin Drug Binding. *Molecules*, 27(3), 885. <https://doi.org/10.3390/molecules27030885>
- Anderson, A. S., & Wellington, E. (2001). The taxonomy of *Streptomyces* and related genera. *International journal of systematic and evolutionary microbiology*, 51(3), 797–814.
- Aoki, H., SAKAI, H.-I., KOHSAKA, M., KONOMI, T., HOSODA, J., KUBOCHI, Y., IGUCHI, E., & IMANAKA, H. (1976). Nocardicin a, a new monocyclic β -lactam antibiotic i. discovery, isolation and characterization. *The Journal of antibiotics*, 29(5), 492–500.
- Arnold, B. J., Huang, I.-T., & Hanage, W. P. (2022). Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, 20(4), 206–218.
- Astashyn, A., Tvedte, E. S., Sweeney, D., Sapojnikov, V., Bouk, N., Joukov, V., Mozes, E., Strobe, P. K., Sylla, P. M., Wagner, L., et al. (2024). Rapid and sensitive

- detection of genome contamination at scale with fcs-gx. *Genome biology*, 25(1), 60.
- Babis, W., Jastrzebski, J. P., & Ciesielski, S. (2024). Fine-tuning of dada2 parameters for multiregional metabarcoding analysis of 16s rna genes from activated sludge and comparison of taxonomy classification power and taxonomy databases. *International Journal of Molecular Sciences*, 25(6), 3508.
- Baldwin, A., Loughlin, M., Caubilla-Barron, J., Kucerova, E., Manning, G., Dowson, C., & Forsythe, S. (2009). Multilocus sequence typing of *Cronobacter sakazakii* and *Cronobacter malonaticus* reveals stable clonal structures with clinical significance which do not correlate with biotypes. *BMC Microbiology*, 9(1), 223–223. <https://doi.org/10.1186/1471-2180-9-223>
- Baltz, R. H. (2021). Genome mining for drug discovery: progress at the front end. *Journal of Industrial Microbiology & Biotechnology*, 48(9-10), kuab044. <https://doi.org/10.1093/jimb/kuab044>
- Bansal, M. S., Kellis, M., Kordi, M., & Kundu, S. (2018). Ranger-dtl 2.0: Rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18), 3214–3216.
- Baptiste, E., van Iersel, L., Janke, A., Kelchner, S., Kelk, S., McInerney, J. O., Morrison, D. A., Nakhleh, L., Steel, M., Stougie, L., et al. (2013). Networks: Expanding evolutionary thinking. *Trends in Genetics*, 29(8), 439–441.
- Barberán, A., Bates, S. T., Casamayor, E. O., & Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal*, 6(2), 343–351.

- Barco, R., Garrity, G., Scott, J., Amend, J., Nealson, K., & Emerson, D. (2020). A genus definition for bacteria and archaea based on a standard genome relatedness index. *MBio*, 11(1), 10–1128.
- Barka, E. A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Meier-Kolthoff, J. P., Klenk, H.-P., Clément, C., Ouhdouch, Y., & Wezel, G. P. v. (2016). Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiology and Molecular Biology Reviews*, 80(1), 1–43. <https://doi.org/10.1128/mnbr.00019-15>
- Barlow, M. (2009). What antimicrobial resistance has taught us about horizontal gene transfer. *Horizontal Gene Transfer: Genomes in Flux*, 397–411.
- Bartoš, O., Chmel, M., & Swierczková, I. (2024). The overlooked evolutionary dynamics of 16s rrna revises its role as the “gold standard” for bacterial species identification. *Scientific Reports*, 14(1), 9067.
- Beiko, R. G., Harlow, T. J., & Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*, 102(40), 14332–14337.
- Bentley, S. D., Chater, K. F., Cerdño-Tárraga, A.-M., Challis, G. L., Thomson, N., James, K. D., Harris, D. E., Quail, M. A., Kieser, H, Harper, D., et al. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* a3 (2). *nature*, 417(6885), 141–147.
- Bergey, D. H. (2001). *BERGEY'S MANUAL OF Systematic Bacteriology* (P. W. B. Whitman & J. Wiley, Eds.; First Edition).

- Beytur, A., Yakupogullari, Y., Oguz, F., Otlu, B., & Kaysadu, H. (2015). Oral amoxicillin-clavulanic acid treatment in urinary tract infections caused by extended-spectrum beta-lactamase-producing organisms. *Jundishapur journal of microbiology*, 8(1).
- Biffi, G., Boretti, G., Di Marco, A., & Pennella, P. (1954). Metabolic behavior and chlortetracycline production by *Streptomyces aureofaciens* in liquid culture. *Applied Microbiology*, 2(5), 288–293.
- Birchler, J. A., & Yang, H. (2022). The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *The Plant Cell*, 34(7), 2466–2474. <https://doi.org/10.1093/plcell/koac076>
- Blomme, E. A., & Will, Y. (2016). Toxicology strategies for drug discovery: Present and future. *Chemical research in toxicology*, 29(4), 473–504.
- Bolourian, A., & Mojtahedi, Z. (2018). Immunosuppressants produced by *Streptomyces*: evolution, hygiene hypothesis, tumour rapalog resistance and probiotics. *Environmental Microbiology Reports*, 10(2), 123–126. <https://doi.org/10.1111/1758-2229.12617>
- Bosshard, P., Zbinden, R., Abels, S., Boddingtonhaus, B., Altwegg, M., & Bottger, E. (2006). 16s rrna gene sequencing versus the api 20 ne system and the vitek 2 id-gnb card for identification of nonfermenting gram-negative bacteria in the clinical laboratory. *Journal of clinical microbiology*, 44(4), 1359–1366.
- Bouteau, F., Grésillon, E., Chartier, D., Arbelet-Bonnin, D., Kawano, T., Baluška, F., Mancuso, S., Calvo, P., & Laurenti, P. (2021). Our sisters the plants? notes

- from phylogenetics and botany on plant kinship blindness. *Plant Signaling & Behavior*, 16(12), 2004769. <https://doi.org/10.1080/15592324.2021.2004769>
- Boykin, L. M. (2014). *Bemisia tabaci* nomenclature: lessons learned: *bemisia tabaci* nomenclature. *Pest Management Science*, 70(10), 1454–1459. <https://doi.org/10.1002/ps.3709>
- Breitwieser, F. P., Perte, M., Zimin, A. V., & Salzberg, S. L. (2019). Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome research*, 29(6), 954–960.
- Bridgeman, M. B., & Abazia, D. T. (2017). Medicinal Cannabis: History, Pharmacology, And Implications for the Acute Care Setting. *Pharmacy and therapeutics*, 42(3), 180–188.
- Brook, K., Bennett, J., & Desai, S. P. (2017). The Chemical History of Morphine: An 8000-year Journey, from Resin to de-novo Synthesis. *Journal of Anesthesia History*, 3(2), 50–55. <https://doi.org/10.1016/j.janh.2017.02.001>
- Brown, C. T., & Irber, L. (2016). Sourmash: A library for minhash sketching of dna. *Journal of open source software*, 1(5), 27.
- Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., & Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLOS ONE*, 15(2), e0228899. <https://doi.org/10.1371/journal.pone.0228899>
- Buermans, H., & Den Dunnen, J. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10), 1932–1941.

- Buermans, H., & Dunnen, J. d. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10), 1932–1941. <https://doi.org/10.1016/j.bbadis.2014.06.015>
- Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, 6(1), 190007. <https://doi.org/10.1038/sdata.2019.7>
- Burger, A., Brandt, B., Süssstrunk, U., Thompson, C. J., & Wohlleben, W. (1998). Analysis of a *Streptomyces coelicolor* a3 (2) locus containing the nucleoside diphosphate kinase (ndk) and folylpolyglutamate synthetase (folc) genes. *FEMS microbiology letters*, 159(2), 283–291.
- Bury-Moné, S., Thibessard, A., Lioy, V. S., & Leblond, P. (2023). Dynamics of the *Streptomyces* chromosome: Chance and necessity. *Trends in Genetics*.
- Cacciapuoti, B., CICERONI, L., & BARBINI, D. A. (1991). Fatty Acid Profiles, a Chemotaxonomic Key for the Classification of Strains of the Family *Leptospiraceae*. *International Journal of Systematic and Evolutionary Microbiology*, 41(2), 295–300. <https://doi.org/10.1099/00207713-41-2-295>
- Caicedo-Montoya, C., Manzo-Ruiz, M., & Ríos-Esteva, R. (2021). Pan-genome of the genus *Streptomyces* and prioritization of biosynthetic gene clusters with potential to produce antibiotic compounds. *Frontiers in microbiology*, 12, 677558.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1), 421. <https://doi.org/10.1186/1471-2105-10-421>

- Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), 1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>
- Carroll, L. M., Larralde, M., Fleck, J. S., Ponnudurai, R., Milanese, A., Cappio, E., & Zeller, G. (2021). Accurate de novo identification of biosynthetic gene clusters with GECCO. *bioRxiv*, 2021.05.03.442509. <https://doi.org/10.1101/2021.05.03.442509>
- Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, 17(4), 540–552. <https://doi.org/10.1093/oxfordjournals.molbev.a026334>
- Centers for Disease Control and Prevention (CDC), National Center for Emerging Zoonotic and Infectious Diseases (U.S.), National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (U.S.), & National Center for Immunization and Respiratory Diseases (U.S.) (2013). Antibiotic resistance threats in the united states, 2013. <https://stacks.cdc.gov/view/cdc/20705>
- Chakraborty, S., Gruber, T., Barry III, C. E., Boshoff, H. I., & Rhee, K. Y. (2013). Para-aminosalicylic acid acts as an alternative substrate of folate metabolism in *Mycobacterium tuberculosis*. *Science*, 339(6115), 88–91.
- Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, 69(2), 330–339. <https://doi.org/10.1016/j.mimet.2007.02.005>

- Champney, W. S. (2020). Antibiotics targeting bacterial ribosomal subunit biogenesis. *Journal of Antimicrobial Chemotherapy*, 75(4), 787–806.
- Chao, J., Tang, F., & Xu, L. (2022). Developments in algorithms for sequence alignment: A review. *Biomolecules*, 12(4), 546.
- Chatzou, M., Magis, C., Chang, J.-M., Kemena, C., Bussotti, G., Erb, I., & Notredame, C. (2015). Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics*, 17(6), 1009–1023. <https://doi.org/10.1093/bib/bbv099>
- Chevrette, M. G., Gutiérrez-García, K., Selem-Mojica, N., Aguilar-Martínez, C., Yañez-Olvera, A., Ramos-Aboites, H. E., Hoskisson, P. A., & Barona-Gómez, F. (2019a). Evolutionary dynamics of natural product biosynthesis in bacteria. *Natural Product Reports*, 37(4), 566–599. <https://doi.org/10.1039/c9np00048h>
- Chevrette, M. G., Carlos-Shanley, C., Louie, K. B., Bowen, B. P., Northen, T. R., & Currie, C. R. (2019b). Taxonomic and Metabolic Incongruence in the Ancient Genus *Streptomyces*. *Frontiers in Microbiology*, 10, 2170. <https://doi.org/10.3389/fmicb.2019.02170>
- Chevrette, M. G., Carlson, C. M., Ortega, H. E., Thomas, C., Ananiev, G. E., Barns, K. J., Book, A. J., Cagnazzo, J., Carlos, C., Flanigan, W., Grubbs, K. J., Horn, H. A., Hoffmann, F. M., Klassen, J. L., Knack, J. J., Lewin, G. R., McDonald, B. R., Muller, L., Melo, W. G. P., ... Currie, C. R. (2019c). The antimicrobial potential of *Streptomyces* from insect microbiomes. *Nature Communications*, 10(1), 516. <https://doi.org/10.1038/s41467-019-08438-0>

- Chokshi, A., Sifri, Z., Cennimo, D., & Horng, H. (2019). Global Contributors to Antibiotic Resistance. *Journal of Global Infectious Diseases*, 11(1), 36–42. https://doi.org/10.4103/jgid.jgid_110_18
- Cholo, M. C., Steel, H. C., Fourie, P. B., Germishuizen, W. A., & Anderson, R. (2012). Clofazimine: Current status and future prospects. *Journal of antimicrobial chemotherapy*, 67(2), 290–298.
- Chopra, S., Matsuyama, K., Tran, T., Malerich, J. P., Wan, B., Franzblau, S. G., Lun, S., Guo, H., Maiga, M. C., Bishai, W. R., & Madrid, P. B. (2012). Evaluation of gyrase B as a drug target in *Mycobacterium tuberculosis*. *Journal of Antimicrobial Chemotherapy*, 67(2), 415–421. <https://doi.org/10.1093/jac/dkr449>
- Choufa, C., Tidjani, A.-R., Gauthier, A., Harb, M., Lao, J., Leblond-Bourget, N., Vos, M., Leblond, P., & Bontemps, C. (2022). Prevalence and mobility of integrative and conjugative elements within a *Streptomyces* natural population. *Frontiers in Microbiology*, 13, 970179.
- Chu, J., Vila-Farres, X., Inoyama, D., Ternei, M., Cohen, L. J., Gordon, E. A., Reddy, B. V. B., Charlop-Powers, Z., Zebroski, H. A., Gallardo-Macias, R., Jaskowski, M., Satish, S., Park, S., Perlin, D. S., Freundlich, J. S., & Brady, S. F. (2016). Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nature chemical biology*, 12(12), 1004–1006. <https://doi.org/10.1038/nchembio.2207>
- Chung, M., Munro, J. B., Tettelin, H., & Hotopp, J. C. D. (2018). Using Core Genome Alignments To Assign Bacterial Species. *mSystems*, 3(6), e00236–18. <https://doi.org/10.1128/msystems.00236-18>

- Ciccarelli, F. D., Doerks, T., Mering, C. v., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, *311*(5765), 1283–1287. <https://doi.org/10.1126/science.1123061>
- Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., Brover, S., Schoch, C. L., Kimchi, A., & DiCuccio, M. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the ncbi. *International journal of systematic and evolutionary microbiology*, *68*(7), 2386–2392.
- Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*, *17*(4), 840–62. <https://doi.org/10.1128/cmr.17.4.840-862.2004>
- Coates, A. R., Halls, G., & Hu, Y. (2011). Novel classes of antibiotics or more of the same? *British Journal of Pharmacology*, *163*(1), 184–194. <https://doi.org/10.1111/j.1476-5381.2011.01250.x>
- Coenye, T., & Vandamme, P. (2003). Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters*, *228*(1), 45–49. [https://doi.org/10.1016/s0378-1097\(03\)00717-1](https://doi.org/10.1016/s0378-1097(03)00717-1)
- Cohn, F. (1872). Untersuchungen über bakterien. beitrage zur biologie der pflanzen. 1 heft 2.
- Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*, D633–D642. <https://doi.org/10.1093/nar/gkt1244>

- Coll, F., Gouliouris, T., Blane, B., Yeats, C. A., Raven, K. E., Ludden, C., Khokhar, F. A., Wilson, H. J., Roberts, L. W., Harrison, E. M., et al. (2024). Antibiotic resistance determination using *Enterococcus faecium* whole-genome sequences: A diagnostic accuracy study using genotypic and phenotypic data. *The Lancet Microbe*, 5(2), e151–e163.
- Costa, S. S., Guimarães, L. C., Silva, A., Soares, S. C., & Baraúna, R. A. (2020a). First steps in the analysis of prokaryotic pan-genomes. *Bioinformatics and Biology Insights*, 14, 1177932220938064.
- Costa, S. S., Guimarães, L. C., Silva, A., Soares, S. C., & Baraúna, R. A. (2020b). First Steps in the Analysis of Prokaryotic Pan-Genomes. *Bioinformatics and Biology Insights*, 14. <https://doi.org/10.1177/1177932220938064>
- Coyne, J. A., Elwyn, S., Kim, S. Y., & Llopart, A. (2004). Genetic studies of two sister species in the drosophila melanogaster subgroup, d. yakuba and d. santomea. *Genetics Research*, 84(1), 11–26.
- Curbete, M. M., & Salgado, H. R. N. (2016). A critical review of the properties of fusidic acid and analytical methods for its determination. *Critical reviews in analytical chemistry*, 46(4), 352–360.
- Darken, M. A., Berenson, H., Shirk, R. J., & Sjolander, N. O. (1960). Production of tetracycline by *Streptomyces aureofaciens* in synthetic media. *Applied microbiology*, 8(1), 46–51.
- Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8), 772. <https://doi.org/10.1038/nmeth.2109>

- Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2019). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution*, 37(1), 291–294. <https://doi.org/10.1093/molbev/msz189>
- Daubin, V., & Szöllősi, G. J. (2016). Horizontal gene transfer and the history of life. *Cold Spring Harbor perspectives in biology*, 8(4), a018036.
- Daubin, V., Lerat, E., & Perrière, G. (2003). The source of laterally transferred genes in bacterial genomes. *Genome biology*, 4, 1–12.
- David, L. A., & Alm, E. J. (2011). Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, 469(7328), 93–96.
- Davin-Regli, A., Lavigne, J.-P., & Pagès, J.-M. (2019). *Enterobacter spp.*: Update on taxonomy, clinical aspects, and emerging antimicrobial resistance. *Clinical microbiology reviews*, 32(4), 10–1128.
- De Been, M., Pinholt, M., Top, J., Bletz, S., Mellmann, A., Van Schaik, W., Brouwer, E., Rogers, M., Kraat, Y., Bonten, M., et al. (2015). Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. *Journal of clinical microbiology*, 53(12), 3788–3797.
- De Felice, B., Spicer, L. J., & Caloni, F. (2023). Enniatin b1: Emerging mycotoxin and emerging issues. *Toxins*, 15(6), 383.
- De Simeis, D., & Serra, S. (2021). *Actinomycetes*: A never-ending source of bioactive compounds—an overview on antibiotics production. *Antibiotics*, 10(5), 483.
- Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature biotechnology*, 34(5), 518–524.

- DeBarber, A. E., Mdluli, K., Bosman, M., Bekker, L.-G., & Barry 3rd, C. E. (2000). Ethionamide activation and sensitivity in multidrug-resistant *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, 97(17), 9677–9682.
- DeSalle, R., & Riley, M. (2020). Should networks supplant tree building? *Microorganisms*, 8(8), 1179.
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7). <https://doi.org/10.1128/aem.03006-05>
- Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R., & Wilson, D. J. (2018). Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic acids research*, 46(22), e134–e134.
- Diestel, R. (2000). *Graph theory*. Springer-Verlag.
- Díez-Aguilar, M., & Cantón, R. (2019). New microbiological aspects of fosfomycin. *Revista Española de Quimioterapia*, 32(1), 8.
- Domingo-Sananes, M. R., & McInerney, J. O. (2021). Mechanisms that shape microbial pangenomes. *Trends in Microbiology*, 29(6), 493–503.
- Doroghazi, J. R., & Buckley, D. H. (2010). Widespread homologous recombination within and between *Streptomyces* species. *The ISME Journal*, 4(9), 1136–1143. <https://doi.org/10.1038/ismej.2010.45>

- Drews, G. (2000). The roots of microbiology and the influence of Ferdinand Cohn on microbiology of the 19th century. *FEMS Microbiology Reviews*, 24(3). <https://doi.org/10.1111/j.1574-6976.2000.tb00540.x>
- Drinkovic, D., Fuller, E. R., Shore, K. P., Holland, D. J., & Ellis-Pegler, R. (2001). Clindamycin treatment of *Staphylococcus aureus* expressing inducible clindamycin resistance. *Journal of Antimicrobial Chemotherapy*, 48(2), 315–316.
- Du, L., Liu, R.-H., Ying, L., & Zhao, G.-R. (2012). An efficient intergeneric conjugation of dna from *Escherichia coli* to mycelia of the lincomycin-producer *Streptomyces lincolnensis*. *International journal of molecular sciences*, 13(4), 4797–4806.
- Ebach, M. C., Williams, D. M., & Morrone, J. J. (2006). Paraphyly is bad taxonomy. *Taxon*, 55(4), 831–832. <https://doi.org/10.2307/25065678>
- Edgar, R. (2018a). Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ*, 6, e5030. <https://doi.org/10.7717/peerj.5030>
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>
- Edgar, R. C. (2018b). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, 6, 4652. <https://doi.org/10.7717/peerj.4652>

- Edgar, R. C. (2018c). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, *34*(14), 2371–2375. <https://doi.org/10.1093/bioinformatics/bty113>
- Ehrenberg, C. (1834). Dritter beitrage zur erkenntniss grosser organisation in der richtung des kleinsten raumes. *Berlin: Konigl. Akad. d. Wiss.*, 1833, 145–336.
- Elias, I., & Lagergren, J. (2009). Fast neighbor joining. *Theoretical Computer Science*, *410*(21-23), 1993–2000. <https://doi.org/10.1016/j.tcs.2008.12.040>
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. <https://doi.org/10.1186/s13059-019-1832-y>
- Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, *4*(12), 1111–1119. <https://doi.org/10.1111/2041-210x.12114>
- Escherich, T. (1885). Die darmbakterien des neugeborenen und säuglings. *Fortschritte der Medicin*, *3*(16 und 17), 515–554.
- Evans, D. R., Griffith, M. P., Sundermann, A. J., Shutt, K. A., Saul, M. I., Mustapha, M. M., Marsh, J. W., Cooper, V. S., Harrison, L. H., & Tyne, D. V. (2020). Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. *eLife*, *9*, 53886. <https://doi.org/10.7554/elife.53886>
- Feigin, V. L., Krishnamurthi, R. V., Parmar, P., Norrving, B., Mensah, G. A., Bennett, D. A., Barker-Collo, S., Moran, A. E., Sacco, R. L., Truelsen, T., et al. (2015).

- Update on the global burden of ischemic and hemorrhagic stroke in 1990-2013: The gbd 2013 study. *Neuroepidemiology*, 45(3), 161–176.
- Feigin, V. L., Stark, B. A., Johnson, C. O., Roth, G. A., Bisignano, C., Abady, G. G., Abbasifard, M., Abbasi-Kangevari, M., Abd-Allah, F., Abedi, V., et al. (2021). Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Neurology*, 20(10), 795–820. [https://doi.org/10.1016/s1474-4422\(21\)00252-0](https://doi.org/10.1016/s1474-4422(21)00252-0)
- Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., & Spratt, B. G. (2004). eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data. *Journal of Bacteriology*, 186(5), 1518–1530. <https://doi.org/10.1128/jb.186.5.1518-1530.2004>
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), 368–376. <https://doi.org/10.1007/bf01734359>
- Felsenstein, J. (1985). CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution*, 39(4), 783–791. <https://doi.org/10.1111/j.1558-5646.1985.tb00420.x>
- Fidock, D. A., Rosenthal, P. J., Croft, S. L., Brun, R., & Nwaka, S. (2004). Antimalarial drug discovery: Efficacy models for compound screening. *Nature reviews Drug discovery*, 3(6), 509–520.
- Figueras, M. J., Beaz-Hidalgo, R., Hossain, M. J., & Liles, M. R. (2014). Taxonomic Affiliation of New Genomes Should Be Verified Using Average Nucleotide Identity

- and Multilocus Phylogenetic Analysis. *Genome Announcements*, 2(6), e00927–14. <https://doi.org/10.1128/genomea.00927-14>
- Finlay, A. C., Hobby, G., Hochstein, F., Lees, T., Lenert, T., Means, J., P'an, S., Regna, P., Routien, J., Sobin, B., et al. (1951). Viomycin, a new antibiotic active against mycobacteria. *American review of tuberculosis*, 63(1), 1–3.
- Fischbach, M. A., Walsh, C. T., & Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. *Proceedings of the National Academy of Sciences*, 105(12), 4601–4608.
- Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. *Systematic Zoology*, 19(2), 99–113. <https://doi.org/10.2307/2412448>
- Fitch, W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology*, 20(4), 406. <https://doi.org/10.2307/2412116>
- Fitzmaurice, C., Dicker, D., Pain, A., Hamavid, H., Moradi-Lakeh, M., MacIntyre, M. F., Allen, C., Hansen, G., Woodbrook, R., Wolfe, C., et al. (2015). The global burden of cancer 2013. *JAMA oncology*, 1(4), 505–527.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *science*, 269(5223), 496–512.
- Fleming, A. (1980). On the Antibacterial Action of Cultures of a *Penicillium*, with Special Reference to Their Use in the Isolation of *B. influenzae*. *Clinical Infectious Diseases*, 2(1), 129–139. <https://doi.org/10.1093/clinids/2.1.129>

- Fofana, D., George, E. O., & Bowman, D. (2021). Combining assumptions and graphical network into gene expression data analysis. *Journal of Statistical Distributions and Applications*, 8(1), 9. <https://doi.org/10.1186/s40488-021-00126-z>
- Founou, L. L., Founou, R. C., & Essack, S. Y. (2016). Antibiotic resistance in the food chain: A developing country-perspective. *Frontiers in microbiology*, 7, 232834.
- Fox, G. E., Wisotzkey, J. D., & JR., P. J. (1992). How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. *International Journal of Systematic and Evolutionary Microbiology*, 42(1), 166–170. <https://doi.org/10.1099/00207713-42-1-166>
- Galvidis, I., Lapa, G., & Burkin, M. (2015). Group determination of 14-membered macrolide antibiotics and azithromycin using antibodies against common epitopes. *Analytical biochemistry*, 468, 75–82.
- Gao, Z., Jiang, C., Zhang, J., Jiang, X., Li, L., Zhao, P., Yang, H., Huang, Y., & Li, J. (2023). Hierarchical graph learning for protein–protein interaction. *Nature Communications*, 14(1), 1093. <https://doi.org/10.1038/s41467-023-36736-1>
- Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Peer, Y. V. d., Vandamme, P., Thompson, F. L., & Swings, J. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, 3(9). <https://doi.org/10.1038/nrmicro1236>
- Glaeser, S. P., & Kämpfer, P. (2015). Multilocus sequence analysis (mlsa) in prokaryotic taxonomy. *Systematic and applied microbiology*, 38(4), 237–245.
- Godoy, D., Randle, G., Simpson, A. J., Aanensen, D. M., Pitt, T. L., Kinoshita, R., & Spratt, B. G. (2003). Multilocus Sequence Typing and Evolutionary Relation-

- ships among the Causative Agents of Melioidosis and Glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *Journal of Clinical Microbiology*, 41(5), 2068–2079. <https://doi.org/10.1128/jcm.41.5.2068-2079.2003>
- Goldman, A. D., & Landweber, L. F. (2016). What Is a Genome? *PLoS Genetics*, 12(7), e1006181. <https://doi.org/10.1371/journal.pgen.1006181>
- Golubchik, T., Wise, M. J., Easteal, S., & Jermin, L. S. (2007). Mind the gaps: Evidence of bias in estimates of multiple sequence alignments. *Molecular biology and evolution*, 24(11), 2433–2442.
- Gonzalez, J. M., Zimmermann, J., & Saiz-Jimenez, C. (2005). Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics*, 21(3), 333–337. <https://doi.org/10.1093/bioinformatics/bti008>
- Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). Dna–dna hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology*, 57(1), 81–91.
- Gregory, T. R. (2008). Understanding Evolutionary Trees. *Evolution: Education and Outreach*, 1(2), 121–137. <https://doi.org/10.1007/s12052-008-0035-x>
- Griffith, F. (1928). The significance of pneumococcal types. *Epidemiology & Infection*, 27(2), 113–159.
- Guo, Y., Zheng, W., Rong, X., & Huang, Y. (2008). A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *International Journal of Systematic and Evolutionary Microbiology*, 58(1), 149–159. <https://doi.org/10.1099/ijs.0.65224-0>

- Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2017). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics (Oxford, England)*, *34*(23), 4121–4123. <https://doi.org/10.1093/bioinformatics/bty407>
- Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123.
- Haeckel, E. (1866). Generelle morphologie der organismen: Allgemeine grundzuge der organischen formen-wissenschaft, mechanisch begründet durch die von charles darwin reformierte descendenz-theorie. band 1: Allgemeine anatomie. band 2: Allgemeine entwicklungsgeschichte. *de Gruyter*.
- Hagberg, A., Swart, P. J., & Schult, D. A. (2008). *Exploring network structure, dynamics, and function using networkx* (tech. rep.). Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).
- Hall, B. (2018). *Phylogenetic trees made easy. a how-to manual*. Sinauer Associates.
- Hall, B. G. (2005). Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences. *Molecular Biology and Evolution*, *22*(3), 792–802. <https://doi.org/10.1093/molbev/msi066>
- Hall, B. D., & Spiegelman, S. (1961). Sequence complementarity of t2-dna and t2-specific rna. *Proceedings of the National Academy of Sciences*, *47*(2), 137–146.
- Hansen, J. L., Moore, P. B., & Steitz, T. A. (2003). Structures of Five Antibiotics Bound at the Peptidyl Transferase Center of the Large Ribosomal Subunit.

- Journal of Molecular Biology*, 330(5), 1061–1075. [https://doi.org/10.1016/s0022-2836\(03\)00668-5](https://doi.org/10.1016/s0022-2836(03)00668-5)
- Harling-Lee, J. D., Gorzynski, J., Yebra, G., Angus, T., Fitzgerald, J. R., & Freeman, T. C. (2022). A graph-based approach for the visualisation and analysis of bacterial pangenomes. *BMC Bioinformatics*, 23(1), 416. <https://doi.org/10.1186/s12859-022-04898-2>
- Harrison, C. J., & Langdale, J. A. (2006). A step by step guide to phylogeny reconstruction. *The Plant Journal*, 45(4), 561–572. <https://doi.org/10.1111/j.1365-313x.2005.02611.x>
- Hashemian, S. M. R., Farhadi, T., & Ganjparvar, M. (2018). Linezolid: A review of its properties, function, and use in critical care. *Drug design, development and therapy*, 1759–1767.
- Hennig, W. (1965). Phylogenetic systematics. *Annual review of entomology*, 10(1), 97–116.
- Hernández-González, I. L., Moreno-Hagelsieb, G., & Olmedo-Álvarez, G. (2018). Environmentally-driven gene content convergence and the *Bacillus* phylogeny. *BMC Evolutionary Biology*, 18(1), 148. <https://doi.org/10.1186/s12862-018-1261-7>
- Hershberg, R. (2015). Mutation—the engine of evolution: Studying mutation and its role in the evolution of bacteria. *Cold Spring Harbor perspectives in biology*, 7(9), a018077.
- Higgins, D. G., Blackshields, G., & Wallace, I. M. (2005). Mind the gaps: Progress in progressive alignment. *Proceedings of the National Academy of Sciences of the*

- United States of America*, 102(30), 10411–10412. <https://doi.org/10.1073/pnas.0504801102>
- Hildebrand, F., Meyer, A., & Eyre-Walker, A. (2010). Evidence of selection upon genomic gc-content in bacteria. *PLoS genetics*, 6(9), e1001107.
- Hördt, A., López, M. G., Meier-Kolthoff, J. P., Schleuning, M., Weinhold, L.-M., Tindall, B. J., Gronow, S., Kyrpides, N. C., Woyke, T., & Göker, M. (2020). Analysis of 1,000+ type-strain genomes substantially improves taxonomic classification of alphaproteobacteria. *Frontiers in microbiology*, 11, 468.
- Horesh, G., Taylor-Brown, A., McGimpsey, S., Lassalle, F., Corander, J., Heinz, E., & Thomson, N. R. (2021). Different evolutionary trends form the twilight zone of the bacterial pan-genome. *Microbial genomics*, 7(9), 000670.
- Horton, J. S., & Taylor, T. B. (2023). Mutation bias and adaptation in bacteria. *Microbiology*, 169(11), 001404. <https://doi.org/10.1099/mic.0.001404>
- Hou, P., Nowak, V. V., Taylor, C. J., Calcott, M. J., Knight, A., & Owen, J. G. (2023). A genomic survey of the natural product biosynthetic potential of actinomycetes isolated from new zealand lichens. *Msystems*, 8(2), e01030–22.
- Huerta-Cepas, J., Dopazo, J., & Gabaldón, T. (2010). Ete: A python environment for tree exploration. *BMC bioinformatics*, 11, 1–7.
- Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6), 1635–1638. <https://doi.org/10.1093/molbev/msw046>

- Hughes, D., & Andersson, D. I. (2017). Environmental and genetic modulation of the phenotypic expression of antibiotic resistance. *FEMS microbiology reviews*, 41(3), 374–391.
- Hughes, W. T. (1998). Use of dapsone in the prevention and treatment of *Pneumocystis carinii* pneumonia: A review. *Clinical infectious diseases*, 27(1), 191–204.
- Hutchings, M., Truman, A., & Wilkinson, B. (2019). Antibiotics: past, present and future. *Current Opinion in Microbiology*, 51. <https://doi.org/10.1016/j.mib.2019.10.008>
- Hörandl, E., & Stuessy, T. F. (2010). Paraphyletic groups as natural units of biological classification. *Taxon*, 59(6), 1641–1653. <https://doi.org/10.1002/tax.596001>
- Inoue, K., Kabeya, H., Hagiya, K., Kosoy, M. Y., Une, Y., Yoshikawa, Y., & Maruyama, S. (2011). Multi-locus sequence analysis reveals host specific association between *Bartonella washoensis* and squirrels. *Veterinary Microbiology*, 148(1), 60–65. <https://doi.org/10.1016/j.vetmic.2010.08.007>
- Jacobson, M. J., Lin, G., Whittam, T. S., & Johnson, E. A. (2008). Phylogenetic analysis of *Clostridium botulinum* type A by multi-locus sequence typing. *Microbiology*, 154(8), 2408–2415. <https://doi.org/10.1099/mic.0.2008/016915-0>
- Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2017). High-throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *bioRxiv*, 225342. <https://doi.org/10.1101/225342>
- Janda, J. M. (2020). Proposed nomenclature or classification changes for Bacteria of medical importance: Taxonomic update 5. *Diagnostic Microbiology and Infectious Disease*, 97(3), 115047. <https://doi.org/10.1016/j.diagmicrobio.2020.115047>

- Jenke-Kodama, H., & Dittmann, E. (2009). Evolution of metabolic diversity: Insights from microbial polyketide synthases. *Phytochemistry*, 70(15-16), 1858–1866.
- Jiang, H., Xu, X., Fang, Y., Ogunyemi, S. O., Ahmed, T., Li, X., Yang, Y., Yan, C., Chen, J., & Li, B. (2023). Metabarcoding reveals response of rice rhizosphere bacterial community to rice bacterial leaf blight. *Microbiological Research*, 270, 127344.
- Jin, Y., Zhou, J., Zhou, J., Hu, M., Zhang, Q., Kong, N., Ren, H., Liang, L., & Yue, J. (2020). Genome-based classification of *Burkholderia cepacia* complex provides new insight into its taxonomic status. *Biology Direct*, 15(1), 6. <https://doi.org/10.1186/s13062-020-0258-5>
- Johnson, B. A., Anker, H., & Meleney, F. L. (1945). Bacitracin: A new antibiotic produced by a member of the *B. subtilis* group. *Science*, 102(2650), 376–377.
- Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1), 5029. <https://doi.org/10.1038/s41467-019-13036-1>
- Jolley, K. A., & Maiden, M. C. (2014). Using MLST to study bacterial variation: prospects in the genomic era. *Future Microbiology*, 9(5), 623–630. <https://doi.org/10.2217/fmb.14.24>
- Jolley, K. A., Bray, J. E., & Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications.

- Wellcome Open Research*, 3, 124. <https://doi.org/10.12688/wellcomeopenres.14826.1>
- Jovel, J., Dieleman, L. A., Kao, D., Mason, A. L., & Wine, E. (2018). The human gut microbiome in health and disease. *Metagenomics*, 197–213.
- Kaltenpoth, M., Göttler, W., Herzner, G., & Strohm, E. (2005). Symbiotic bacteria protect wasp larvae from fungal infestation. *Current Biology*, 15(5), 475–479.
- Kaltenpoth, M., Goettler, W., Dale, C., Stubblefield, J. W., Herzner, G., Roeser-Mueller, K., & Strohm, E. (2006). ‘*Candidatus streptomyces philanthi*’, an endosymbiotic streptomycete in the antennae of *Philanthus* digger wasps. *International Journal of Systematic and Evolutionary Microbiology*, 56(6), 1403–1411.
- Kamarudheen, N., & Rao, B. (2018). An Overview of Protease Inhibitors from Actinobacteria. *Research Journal of Biotechnology*, 13(1), 115–122.
- Kaplan, T. (2014). The Role of Horizontal Gene Transfer in Antibiotic Resistance. *Eukaryon*, 10, 80–81.
- Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7), 428–444. <https://doi.org/10.1038/s41576-020-0233-0>
- Kapustina, Medžiūnė, J., Alzbutas, G., Rokaitis, I., Matjošaitis, K., Mackevičius, G., Žeimytė, S., Karpus, L., & Lubys, A. (2021). High-resolution microbiome analysis enabled by linking of 16S rRNA gene sequences with adjacent genomic contexts. *Microbial Genomics*, 7(9), 000624. <https://doi.org/10.1099/mgen.0.000624>

- Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Katz, L., & Baltz, R. H. (2016). Natural product discovery: past, present, and future. *Journal of Industrial Microbiology & Biotechnology*, 43(2-3), 155–176. <https://doi.org/10.1007/s10295-015-1723-5>
- Keller, A., & Ankenbrand, M. J. Inferring core genome phylogenies for bacteria. In: *Bacterial pangenomics: Methods and protocols*. Springer, 2021, pp. 59–68.
- Kers, J. A., Cameron, K. D., Joshi, M. V., Bukhalid, R. A., Morello, J. E., Wach, M. J., Gibson, D. M., & Loria, R. (2005). A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. *Molecular Microbiology*, 55(4), 1025–1033. <https://doi.org/10.1111/j.1365-2958.2004.04461.x>
- Khadayat, K., Sherpa, D. D., Malla, K. P., Shrestha, S., Rana, N., Marasini, B. P., Khanal, S., Rayamajhee, B., Bhattarai, B. R., & Parajuli, N. (2020). Molecular Identification and Antimicrobial Potential of *Streptomyces* Species from Nepalese Soil. *International Journal of Microbiology*, 2020, 1–8. <https://doi.org/10.1155/2020/8817467>
- Khaldi, N., Seifuddin, F. T., Turner, G., Haft, D., Nierman, W. C., Wolfe, K. H., & Fedorova, N. D. (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology*, 47(9), 736–741. <https://doi.org/10.1016/j.fgb.2010.06.003>

- Kiepas, A. B., Hoskisson, P. A., & Pritchard, L. (2023). 16s rna phylogeny and clustering is not a reliable proxy for genome-based taxonomy in *Streptomyces*. *bioRxiv*, 2023–08.
- Kiepas, A. B., Hoskisson, P. A., & Pritchard, L. (2024). 16s rna phylogeny and clustering is not a reliable proxy for genome-based taxonomy in *Streptomyces*. *Microbial Genomics*, 10(9), 001287.
- Kim, H., Kwak, W., Yoon, S. H., Kang, D.-K., & Kim, H. (2018). Horizontal gene transfer of *Chlamydia*: Novel insights from tree reconciliation. *PLoS One*, 13(4), e0195139.
- Kim, H.-J., Lee, J. S., Kwak, N., Cho, J., Lee, C.-H., Han, S. K., & Yim, J.-J. (2019). Role of ethambutol and rifampicin in the treatment of *Mycobacterium avium* complex pulmonary disease. *BMC Pulmonary Medicine*, 19, 1–10.
- Kim, J.-N., Kim, Y., Jeong, Y., Roe, J.-H., Kim, B.-G., & Cho, B.-K. (2015). Comparative genomics reveals the core and accessory genomes of *Streptomyces* species. *Journal of microbiology and biotechnology*, 25(10), 1599–1605.
- Kim, S. B., Lonsdale, J., Seong, C.-N., & Goodfellow, M. (2003). *Streptacidiphilus* gen. nov., acidophilic actinomycetes with wall chemotype i and emendation of the family *streptomycetaceae* (waksman and henrici (1943)al) emend. rainey et al. 1997. *Antonie van Leeuwenhoek*, 83(2), 107–116. <https://doi.org/10.1023/a:1023397724023>
- Komaki, H. (2021). Reclassification of 15 *Streptomyces* species as synonyms of *Streptomyces albogriseolus*, *Streptomyces althioticus*, *Streptomyces anthocyanicus*, *Streptomyces calvus*, *Streptomyces griseoincarnatus*, *Streptomyces mutabilis*,

- Streptomyces pilosus* or *Streptomyces rochei*. *International Journal of Systematic and Evolutionary Microbiology*, 71(3), 004718.
- Komaki, H., & Tamura, T. (2021). Reclassification of *Streptomyces cinnamonensis* as a later heterotypic synonym of *Streptomyces virginiae*. *International Journal of Systematic and Evolutionary Microbiology*, 71(5). <https://doi.org/10.1099/ijsem.0.004813>
- Komaki, H., Sakurai, K., Hosoyama, A., Kimura, A., Igarashi, Y., & Tamura, T. (2018). Diversity of nonribosomal peptide synthetase and polyketide synthase gene clusters among taxonomically close *Streptomyces* strains. *Scientific reports*, 8(1), 6888.
- Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, 39(1), 309–38. <https://doi.org/10.1146/annurev.genet.39.073003.114725>
- Koskella, B., & Brockhurst, M. A. (2014). Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiology Reviews*, 38(5), 916–931. <https://doi.org/10.1111/1574-6976.12072>
- Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21), 4453–4455. <https://doi.org/10.1093/bioinformatics/btz305>
- Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1), 48–50.

- Kuhner, M. K., & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3), 459–68. <https://doi.org/10.1093/oxfordjournals.molbev.a040126>
- Kumagai, T., Koyama, Y., Oda, K., Noda, M., Matoba, Y., & Sugiyama, M. (2010). Molecular cloning and heterologous expression of a biosynthetic gene cluster for the antitubercular agent d-cycloserine produced by *Streptomyces lavendulae*. *Antimicrobial agents and chemotherapy*, 54(3), 1132–1139.
- Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2), R12–R12. <https://doi.org/10.1186/gb-2004-5-2-r12>
- Kusuma, A. B., Nouioui, I., & Goodfellow, M. (2021). Genome-based classification of the *Streptomyces violaceusniger* clade and description of *Streptomyces sabulosicollis* sp. nov. from an indonesian sand dune. *Antonie Van Leeuwenhoek*, 114, 859–873.
- Kämpfer, P. (2012). Systematics of prokaryotes: the state of the art. *Antonie van Leeuwenhoek*, 101(1), 3–11. <https://doi.org/10.1007/s10482-011-9660-4>
- Kämpfer, P. (2020). Bergey’s Manual of Systematics of Archaea and Bacteria, 1–414. <https://doi.org/10.1002/9781118960608.gbm00191>
- Labeda, D. P., Goodfellow, M., Brown, R., Ward, A. C., Lanoot, B., Vannanneyt, M., Swings, J., Kim, S.-B., Liu, Z., Chun, J., Tamura, T., Oguchi, A., Kikuchi, T., Kikuchi, H., Nishii, T., Tsuji, K., Yamaguchi, Y., Tase, A., Takahashi, M., ... Hatano, K. (2012). Phylogenetic study of the species within the family *Streptomycetaceae*. *Antonie van Leeuwenhoek*, 101(1), 73–104. <https://doi.org/10.1007/s10482-011-9656-0>

- Labeda, D. P. (2011). Multilocus sequence analysis of phytopathogenic species of the genus *Streptomyces*. *International Journal of Systematic and Evolutionary Microbiology*, 61(10), 2525–2531. <https://doi.org/10.1099/ijss.0.028514-0>
- Lambert, D., & Loria, R. (1989). *Streptomyces acidiscabies* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 39(4), 393–396.
- Lan, R., & Reeves, P. R. (1996). Gene transfer is a major factor in bacterial evolution. *Molecular biology and evolution*, 13(1), 47–55.
- Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., et al. (2015a). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, 15, 141–161.
- Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., & Ussery, D. W. (2015b). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, 15(2), 141–161. <https://doi.org/10.1007/s10142-015-0433-4>
- Larsen, J., Enright, M. C., Godoy, D., Spratt, B. G., Larsen, A. R., & Skov, R. L. (2012). Multilocus sequence typing scheme for *Staphylococcus aureus*: revision of the gmk locus. *Journal of clinical microbiology*, 50(7), 2538–9. <https://doi.org/10.1128/jcm.00290-12>
- Lawrence, J. G., & Ochman, H. (1997). Amelioration of bacterial genomes: Rates of change and exchange. *Journal of molecular evolution*, 44, 383–397.
- Lawrence, J. G., & Ochman, H. (2002). Reconciling the many faces of lateral gene transfer. *Trends in Microbiology*, 10(1), 1–4.

- Le, K. D., Yu, N. H., Park, A. R., Park, D.-J., Kim, C.-J., & Kim, J.-C. (2022). *Streptomyces* sp. an090126 as a biocontrol agent against bacterial and fungal plant diseases. *Microorganisms*, 10(4), 791.
- Lee, A. S., De Lencastre, H., Garau, J., Kluytmans, J., Malhotra-Kumar, S., Peschel, A., & Harbarth, S. (2018). Methicillin-resistant *Staphylococcus aureus*. *Nature reviews Disease primers*, 4(1), 1–23.
- Lee, I., Kim, Y. O., Park, S.-C., & Chun, J. (2016a). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology*, 66(2), 1100–1103. <https://doi.org/10.1099/ijsem.0.000760>
- Lee, J.-H., Kim, Y.-G., Lee, K., Kim, C.-J., Park, D.-J., Ju, Y., Lee, J.-C., Wood, T. K., & Lee, J. (2016b). *Streptomyces*-derived actinomycin d inhibits biofilm formation by *Staphylococcus aureus* and its hemolytic activity. *Biofouling*, 32(1), 45–56.
- Lee, N., Kim, W., Hwang, S., Lee, Y., Cho, S., Palsson, B., & Cho, B.-K. (2020). Thirty complete *Streptomyces* genome sequences for mining novel secondary metabolite biosynthetic gene clusters. *Scientific data*, 7(1), 55.
- Lemmon, A. R., & Moriarty, E. C. (2004). The Importance of Proper Model Assumption in Bayesian Phylogenetics. *Systematic Biology*, 53(2), 265–277. <https://doi.org/10.1080/10635150490423520>
- Lemmon, A. R., Brown, J. M., Stanger-Hall, K., & Lemmon, E. M. (2009). The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference. *Systematic Biology*, 58(1), 130–145. <https://doi.org/10.1093/sysbio/syp017>

- Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44(1), 99–121.
- Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., & Gascuel, O. (2018). Renewing felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, 556(7702), 452–456.
- Lerat, E., Daubin, V., & Moran, N. A. (2003). From gene trees to organismal phylogeny in prokaryotes: The case of the γ -proteobacteria. *PLoS biology*, 1(1), e19.
- Lewis-Rogers, N., Bendall, M. L., & Crandall, K. A. (2009). Phylogenetic Relationships and Molecular Adaptation Dynamics of Human Rhinoviruses. *Molecular Biology and Evolution*, 26(5), 969–981. <https://doi.org/10.1093/molbev/msp009>
- Li, J. W.-H., & Vederas, J. C. (2009). Drug Discovery and Natural Products: End of an Era or an Endless Frontier? *Science*, 325(5937), 161–165. <https://doi.org/10.1126/science.1168243>
- Li, T., & Yin, Y. (2022). Critical assessment of pan-genomic analysis of metagenome-assembled genomes. *Briefings in Bioinformatics*, 23(6), bbac413. <https://doi.org/10.1093/bib/bbac413>
- Li, W., Fu, L., Niu, B., Wu, S., & Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in Bioinformatics*, 13(6), 656–668. <https://doi.org/10.1093/bib/bbs035>
- Li, Y., Wang, M., Sun, Z.-Z., & Xie, B.-B. (2021). Comparative Genomic Insights Into the Taxonomic Classification, Diversity, and Secondary Metabolic Poten-

- tials of *Kitasatospora*, a Genus Closely Related to *Streptomyces*. *Frontiers in Microbiology*, 12, 683814. <https://doi.org/10.3389/fmicb.2021.683814>
- Lin, G. N., Zhang, C., & Xu, D. (2011). Polytohy identification in microbial phylogenetic reconstruction. *BMC systems biology*, 5, 1–11.
- Lin, X., & Kück, U. (2022). Cephalosporins as key lead generation beta-lactam antibiotics. *Applied Microbiology and Biotechnology*, 106(24), 8007–8020.
- Linnaeus, C. (1759). *Systema naturae*.
- Lioy, V. S., Lorenzi, J.-N., Najah, S., Poinsignon, T., Leh, H., Saulnier, C., Aigle, B., Lautru, S., Thibessard, A., Lespinet, O., et al. (2021). Dynamics of the compartmentalized *Streptomyces* chromosome during metabolic differentiation. *Nature communications*, 12(1), 5221.
- Lippi, D., & Gotuzzo, E. (2014). The greatest steps towards the discovery of *Vibrio cholerae*. *Clinical Microbiology and Infection*, 20(3), 191–195. <https://doi.org/10.1111/1469-0691.12390>
- Liras, P., & Martín, J. F. (2021). *Streptomyces clavuligerus*: the omics era. *Journal of Industrial Microbiology & Biotechnology*, 48(9-10), kuab072. <https://doi.org/10.1093/jimb/kuab072>
- Livingstone, P. G., Morphew, R. M., & Whitworth, D. E. (2018). Genome Sequencing and Pan-Genome Analysis of 23 *Coralloccoccus spp.* Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. *Frontiers in Microbiology*, 9, 3187. <https://doi.org/10.3389/fmicb.2018.03187>

- Löfmark, S., Edlund, C., & Nord, C. E. (2010). Metronidazole is still the drug of choice for treatment of anaerobic infections. *Clinical infectious diseases*, 50(Supplement_1), S16–S23.
- Loman, N. J., & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology*, 13(12), 787–794.
- Long, H., Sung, W., Miller, S. F., Ackerman, M. S., Doak, T. G., & Lynch, M. (2014). Mutation Rate, Spectrum, Topology, and Context-Dependency in the DNA Mismatch Repair-Deficient *Pseudomonas fluorescens* ATCC948. *Genome Biology and Evolution*, 7(1), 262–271. <https://doi.org/10.1093/gbe/evu284>
- Lorenzi, J.-N., Lespinet, O., Leblond, P., & Thibessard, A. (2021). Subtelomeres are fast-evolving regions of the *Streptomyces* linear chromosome. *Microbial genomics*, 7(6), 000525.
- Loria, R., Kers, J., & Joshi, M. (2006). Evolution of plant pathogenicity in *Streptomyces*. *Annu. Rev. Phytopathol.*, 44(1), 469–487.
- López-Aladid, R., Fernández-Barat, L., Alcaraz-Serrano, V., Bueno-Freire, L., Vázquez, N., Pastor-Ibáñez, R., Palomeque, A., Oscanoa, P., & Torres, A. (2023). Determining the most accurate 16S rRNA hypervariable region for taxonomic identification from respiratory samples. *Scientific Reports*, 13(1), 3974. <https://doi.org/10.1038/s41598-023-30764-z>
- Madhaiyan, M., Saravanan, V. S., See-Too, W.-S., Volpiano, C. G., Sant’Anna, F. H., Mota, F. F. d., Sutcliffe, I., Sangal, V., Passaglia, L. M. P., & Rosado, A. S. (2022). Genomic and phylogenomic insights into the family *Streptomycetaceae*

- lead to the proposal of six novel genera. *International Journal of Systematic and Evolutionary Microbiology*, 72(10). <https://doi.org/10.1099/ijsem.0.005570>
- Magee, A. F., May, M. R., & Moore, B. R. (2014). The Dawn of Open Access to Phylogenetic Data. *PLoS ONE*, 9(10), e110268. <https://doi.org/10.1371/journal.pone.0110268>
- Mahajan, R. (2013). Bedaquiline: First fda-approved tuberculosis drug in 40 years.
- Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., & Spratt, B. G. (1998a). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6), 3140–3145. <https://doi.org/10.1073/pnas.95.6.3140>
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., et al. (1998b). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6), 3140–3145.
- Maiti, P. K., Das, S., Sahoo, P., & Mandal, S. (2020). *Streptomyces* sp sm01 isolated from indian soil produces a novel antibiotic picolinamycin effective against multi drug resistant bacterial strains. *Scientific Reports*, 10(1), 10092. <https://doi.org/10.1038/s41598-020-66984-w>
- Manoharan-Basil, S. S., Laumen, J. G. E., Van Dijck, C., De Block, T., De Baetselier, I., & Kenyon, C. (2021). Evidence of horizontal gene transfer of 50s ribosomal

- genes rplb, rpld, and rply in *Neisseria gonorrhoeae*. *Frontiers in Microbiology*, 12. <https://doi.org/10.3389/fmicb.2021.683901>
- Manyi-Loh, C., Mamphweli, S., Meyer, E., & Okoh, A. (2018). Antibiotic Use in Agriculture and Its Consequential Resistance in Environmental Sources: Potential Public Health Implications. *Molecules : A Journal of Synthetic Chemistry and Natural Product Chemistry*, 23(4), 795. <https://doi.org/10.3390/molecules23040795>
- Marschall, J., Carpenter, C. R., Fowler, S., & Trautner, B. W. (2013). Antibiotic prophylaxis for urinary tract infections after removal of urinary catheter: Meta-analysis. *Bmj*, 346.
- Martín-Sánchez, L., Singh, K. S., Avalos, M., van Wezel, G. P., Dickschat, J. S., & Garbeva, P. (2019). Phylogenomic analyses and distribution of terpene synthases among *Streptomyces*. *Beilstein journal of organic chemistry*, 15(1), 1181–1193.
- Martínez-Núñez, M. A., & López, V. E. L. y. (2016). Nonribosomal peptides synthetases and their applications in industry. *Sustainable Chemical Processes*, 4(1), 13. <https://doi.org/10.1186/s40508-016-0057-6>
- Mast, Y., & Wohlleben, W. (2014). Streptogramins—two are better than one! *International Journal of Medical Microbiology*, 304(1), 44–50.
- Mazloom, R., Pritchard, L., Brown, C. T., Vinatzer, B. A., & Heath, L. S. Lingroups as a principled approach to compare and integrate multiple bacterial taxonomies. In: In *Proceedings of the 13th acm international conference on bioinformatics, computational biology and health informatics*. 2022, 1–7.
- McDonald, B. R., & Currie, C. R. (2017a). Lateral gene transfer dynamics in the ancient bacterial genus streptomyces. *MBio*, 8(3), 10–1128.

- McDonald, B. R., & Currie, C. R. (2017b). Lateral gene transfer dynamics in the ancient bacterial genus *Streptomyces*. *mBio*, 8(3), 10.1128/mbio.00644–17. <https://doi.org/10.1128/mbio.00644-17>
- McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6(3), 610–618. <https://doi.org/10.1038/ismej.2011.139>
- McDonald, D., Jiang, Y., Balaban, M., Cantrell, K., Zhu, Q., Gonzalez, A., Morton, J. T., Nicolaou, G., Parks, D. H., Karst, S. M., et al. (2024). Greengenes2 unifies microbial data in a single reference tree. *Nature biotechnology*, 42(5), 715–718.
- McInerney, J. O., McNally, A., & O’Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature Microbiology*, 2(4), 17040. <https://doi.org/10.1038/nmicrobiol.2017.40>
- Medema, M. H., Blin, K., Cimermanic, P., Jager, V. d., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., & Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, 39(Web Server issue), W339–W346. <https://doi.org/10.1093/nar/gkr466>
- Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., De Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., et al. (2015). Minimum information about a biosynthetic gene cluster. *Nature chemical biology*, 11(9), 625–631.

- Meehan, C. J., Barco, R. A., Loh, Y.-H. E., Cogneau, S., & Rigouts, L. (2021). Reconstituting the genus *Mycobacterium*. *International journal of systematic and evolutionary microbiology*, 71(9), 004922.
- Miao, V., Coeffet-LeGal, M.-F., Brian, P., Brost, R., Penn, J., Whiting, A., Martin, S., Ford, R., Parr, I., Bouchard, M., et al. (2005). Daptomycin biosynthesis in *Streptomyces roseosporus*: Cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology*, 151(5), 1507–1523.
- Michael, C. A., Dominey-Howes, D., & Labbate, M. (2014). The antimicrobial resistance crisis: Causes, consequences, and management. *Frontiers in public health*, 2, 110657.
- Mikoc, A., Ahel, I., & Gamulin, V. (2000). Construction and characterization of a *Streptomyces rimosus* recA mutant: the RecA-deficient strain remains viable. *Molecular Genetics and Genomics*, 264(3), 227–232. <https://doi.org/10.1007/s004380000284>
- Minin, V., Abdo, Z., Joyce, P., & Sullivan, J. (2003). Performance-Based Selection of Likelihood Models for Phylogeny Estimation. *Systematic Biology*, 52(5), 674–683. <https://doi.org/10.1080/10635150390235494>
- Mispelaere, M., De Rop, A.-S., Hermans, C., De Maeseneire, S. L., Soetaert, W. K., De Mol, M. L., & Hulpiau, P. (2024). Whole genome-based comparative analysis of the genus *Streptomyces* reveals many misclassifications. *Applied Microbiology and Biotechnology*, 108(1), 1–12.

- Miyoshi-Akiyama, T., Hayakawa, K., Ohmagari, N., Shimojima, M., & Kirikae, T. (2013). Multilocus Sequence Typing (MLST) for Characterization of *Enterobacter cloacae*. *PLoS ONE*, 8(6), e66358. <https://doi.org/10.1371/journal.pone.0066358>
- Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., & Prunotto, M. (2017). Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug Discovery*, 16(8), 531–543. <https://doi.org/10.1038/nrd.2017.111>
- Mohite, O. S., Lloyd, C. J., Monk, J. M., Weber, T., & Palsson, B. O. (2022). Pangenome analysis of *Enterobacteria* reveals richness of secondary metabolite gene clusters and their associated gene sets. *Synthetic and Systems Biotechnology*, 7(3), 900–910. <https://doi.org/10.1016/j.synbio.2022.04.011>
- Moreno, E., Middlebrook, E. A., Altamirano-Silva, P., Al Dahouk, S., Araj, G. F., Arce-Gorvel, V., Arenas-Gamboa, Á., Ariza, J., Barquero-Calvo, E., Battelli, G., et al. (2023). If you're not confused, you're not paying attention: *Ochrobactrum* is not *Brucella*. *Journal of clinical microbiology*, 61(8), e00438–23.
- Morino, T., & Takahashi, H. (1997). Transduction of a plasmid with an inserted r4 phage dna fragment in *Streptomyces lividans*. *Bioscience, biotechnology, and biochemistry*, 61(6), 1047–1048.
- Mosher, R. H., Camp, D. J., Yang, K., Brown, M. P., Shaw, W. V., & Vining, L. C. (1995). Inactivation of chloramphenicol by o-phosphorylation: A novel resistance mechanism in *Streptomyces venezuelae* isp5230, a chloramphenicol producer. *Journal of Biological Chemistry*, 270(45), 27000–27006.

- Munoz-Davila, M. J. (2014). Role of old antibiotics in the era of antibiotic resistance. highlighted nitrofurantoin for the treatment of lower urinary tract infections. *Antibiotics*, 3(1), 39–48.
- Murray, C., Ikuta, K. S., Sharara, F., Swetschinski, L., Aguilar, G. R., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Hamadani, B. H. K., Kumaran, E. A. P., McManigal, B., ... Naghavi, M. (2022a). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *Lancet*, 399(10325), 629–655.
- Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Aguilar, G. R., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., et al. (2022b). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *The lancet*, 399(10325), 629–655.
- Najjar, P. A., & Smink, D. S. (2015). Prophylactic Antibiotics and Prevention of Surgical Site Infections. *Surgical Clinics of North America*, 95(2), 269–283. <https://doi.org/10.1016/j.suc.2014.11.006>
- Nang, S. C., Han, M.-L., Yu, H. H., Wang, J., Torres, V. V. L., Dai, C., Velkov, T., Harper, M., & Li, J. (2019). Polymyxin resistance in *Klebsiella pneumoniae*: multifaceted mechanisms utilized in the presence and absence of the plasmid-encoded phosphoethanolamine transferase gene mcr-1. *Journal of Antimicrobial Chemotherapy*, 74(11), 3190–3198. <https://doi.org/10.1093/jac/dkz314>

- Navon-Venezia, S., Kondratyeva, K., & Carattoli, A. (2017). *Klebsiella pneumoniae*: A major worldwide source and shuttle for antibiotic resistance. *FEMS microbiology reviews*, *41*(3), 252–275.
- Ndovie, W., Havránek, J., Leconte, J., Koszucki, J., Chindelevitch, L., Adriaenssens, E. M., & Mostowy, R. J. (2025). Exploration of the genetic landscape of bacterial dsDNA viruses reveals an anti gap amid extensive mosaicism. *mSystems*, e01661–24.
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443–453.
- Nemeth, J., Oesch, G., & Kuster, S. P. (2015). Bacteriostatic versus bactericidal antibiotics for patients with serious bacterial infections: Systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy*, *70*(2), 382–395.
- Nickrent, D. L., Blarer, A., Qiu, Y.-L., Vidal-Russell, R., & Anderson, F. E. (2004). Phylogenetic inference in rafflesiales: The influence of rate heterogeneity and horizontal gene transfer. *BMC Evolutionary Biology*, *4*, 1–17.
- Nikolaidis, M., Hesketh, A., Frangou, N., Mossialos, D., Van De Peer, Y., Oliver, S., & Amoutzias, G. (2023). A panoramic view of the genomic landscape of the genus.
- Normark, B. H., & Normark, S. (2002). Evolution and spread of antibiotic resistance. *Journal of internal medicine*, *252*(2), 91–106.
- Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, *302*(1), 205–17. <https://doi.org/10.1006/jmbi.2000.4042>

- Nouioui, I., Carro, L., García-López, M., Meier-Kolthoff, J. P., Woyke, T., Kyrpides, N. C., Pukall, R., Klenk, H.-P., Goodfellow, M., & Göker, M. (2018). Genome-Based Taxonomic Classification of the Phylum Actinobacteria. *Frontiers in Microbiology*, 9, 2007. <https://doi.org/10.3389/fmicb.2018.02007>
- Ochman, H., Lerat, E., & Daubin, V. (2005). Examining bacterial species under the specter of gene transfer and exchange. *Proceedings of the National Academy of Sciences*, 102(suppl.1), 6595–6599.
- Ojha, K. K., Mishra, S., & Singh, V. K. Chapter 5 - computational molecular phylogeny: Concepts and applications (D. B. Singh & R. K. Pathak, Eds.). In: *Bioinformatics* (D. B. Singh & R. K. Pathak, Eds.). Ed. by Singh, D. B., & Pathak, R. K. Academic Press, 2022, pp. 67–89. ISBN: 978-0-323-89775-4. <https://doi.org/https://doi.org/10.1016/B978-0-323-89775-4.00025-0>.
- Omura, S., TAKAHASHI, Y., IWAI, Y., & TANAKA, H. (1982). *Kitasatosporia*, a new genus of the order *Actinomycetales*. *The Journal of antibiotics*, 35(8), 1013–1019.
- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using minhash. *Genome biology*, 17, 1–14.
- O'Neill, J. (2014). Antimicrobial resistance: Tackling a crisis for the health and wealth of nations. *Review on Antimicrobial Resistance*.
- Oren, A., & Garrity, G. M. (2021). Valid publication of the names of forty-two phyla of prokaryotes. *International journal of systematic and evolutionary microbiology*, 71(10). <https://doi.org/10.1099/ijsem.0.005056>

- Otani, H., Udworthy, D. W., & Mouncey, N. J. (2022). Comparative and pangenomic analysis of the genus *Streptomyces*. *Scientific reports*, 12(1), 18909.
- Park, J. H., Kim, T. S., Park, H., & Kang, C. K. (2024). Delay in the diagnosis of *Brucella abortus* bacteremia in a nonendemic country: A case report. *BMC Infectious Diseases*, 24(1), 489.
- Parker, C. T., Tindall, B. J., & Garrity, G. M. (2015). International Code of Nomenclature of Prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 69(1A), S1–S111. <https://doi.org/10.1099/ijsem.0.000778>
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7). <https://doi.org/10.1101/gr.186072.114>
- Parte, A. C. (2018). LPSN – List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. *International Journal of Systematic and Evolutionary Microbiology*, 68(6). <https://doi.org/10.1099/ijsem.0.002786>
- Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., & Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, 4(1), 10. <https://doi.org/10.1186/1756-0381-4-10>
- Pavlopoulou, A. (2018). RecA a universal drug target in pathogenic bacteria. *Frontiers in Bioscience*, 23(1), 36–42. <https://doi.org/10.2741/4580>
- Petkovic, H., Cullum, J., Hranueli, D., Hunter, I. S., Perić-Concha, N., Pigac, J., Thamchaipenet, A., Vujaklija, D., & Long, P. F. (2006). Genetics of *Strepto-*

- myces rimosus*, the oxytetracycline producer. *Microbiology and molecular biology reviews*, 70(3), 704–728.
- Philippe, H., & Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Current opinion in microbiology*, 6(5), 498–505.
- Pilgrim, C. (2021). piecewise-regression (aka segmented regression) in Python. *Journal of Open Source Software*, 6(68), 3859. <https://doi.org/10.21105/joss.03859>
- Pilo, S., Valenci, G. Z., Rubinstein, M., Pichadze, L., Scharf, Y., Dveyrin, Z., Rorman, E., & Nissan, I. (2021). High-resolution multilocus sequence typing for *Chlamydia trachomatis*: improved results for clinical samples with low amounts of *C. trachomatis* DNA. *BMC Microbiology*, 21(1), 28. <https://doi.org/10.1186/s12866-020-02077-y>
- Pinheiro, A., Pinheiro, H. P., & Sen, P. K. The use of hamming distance in bioinformatics. In: In *Handbook of statistics*. Vol. 28. Elsevier, 2012, pp. 129–162.
- Porooshat, D. (2019). Antimicrobial Resistance: Implications and Costs. *Infection and Drug Resistance*, 12, 3903–3910. <https://doi.org/10.2147/idr.s234610>
- Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6), 1389–1401.
- Prinzi, A. M., & Moore, N. M. (2023). Change of Plans: Overview of Bacterial Taxonomy, Recent Changes of Medical Importance, and Potential Areas of Impact. *Open Forum Infectious Diseases*, 10(7), 269. <https://doi.org/10.1093/ofid/ofad269>
- Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., & Toth, I. K. (2015). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobac-

- terial plant pathogens. *Analytical Methods*, 8(1), 12–24. <https://doi.org/10.1039/c5ay02550h>
- Procópio, R. E. d. L., Silva, I. R. d., Martins, M. K., Azevedo, J. L. d., & Araújo, J. M. d. (2012). Antibiotics produced by *Streptomyces*. *Brazilian Journal of Infectious Diseases*, 16, 466–471.
- Procópio, R. E. d. L., Silva, I. R. d., Martins, M. K., Azevedo, J. L. d., & Araújo, J. M. d. (2012). Antibiotics produced by *Streptomyces*. *The Brazilian Journal of Infectious Diseases*, 16(5), 466–471. <https://doi.org/10.1016/j.bjid.2012.08.014>
- Qin, Q.-L., Xie, B.-B., Zhang, X.-Y., Chen, X.-L., Zhou, B.-C., Zhou, J., Oren, A., & Zhang, Y.-Z. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of bacteriology*, 196(12), 2210–2215.
- Qin, Z., Munnoch, J. T., Devine, R., Holmes, N. A., Seipke, R. F., Wilkinson, K. A., Wilkinson, B., & Hutchings, M. I. (2017). Formicamycins, antibacterial polyketides produced by *Streptomyces formicae* isolated from African Tetraponera plant-ants. *Chemical Science*, 8(4). <https://doi.org/10.1039/c6sc04265a>
- Råsbäck, T., Johansson, K.-E., Jansson, D., Fellström, C., Alikhani, M., La, T., Dunn, D., & Hampson, D. (2007). Development of a multilocus sequence typing scheme for intestinal spirochaetes within the genus *Brachyspira*. *Microbiology*, 153(12), 4074–4087.
- Rhoads, A., & Au, K. F. (2015). Pacbio sequencing and its applications. *Genomics, Proteomics and Bioinformatics*, 13(5), 278–289.
- Ribeiro-Gonçalves, B., Francisco, A. P., Vaz, C., Ramirez, M., & Carriço, J. A. (2016). PHYLOViZ Online: web-based tool for visualization, phylogenetic inference,

- analysis and sharing of minimum spanning trees. *Nucleic Acids Research*, 44(1), 246–251. <https://doi.org/10.1093/nar/gkw359>
- Rice, L. B. (2008). Federal funding for the study of antimicrobial resistance in nosocomial pathogens: No escape.
- Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*, 106(45), 19126–19131.
- Richter, M., Rosselló-Móra, R., Glöckner, F. O., & Peplies, J. (2016). JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics*, 32(6), 929–931. <https://doi.org/10.1093/bioinformatics/btv681>
- Ridley, M. (2004). Evolution. mark ridley. blackwell publishing. 2003. 751 pages. isbn 1 4051 0345 0. price£ 27.50. *Genetics Research*, 83(1), 65–66.
- Riedlinger, J., Schrey, S. D., Tarkka, M. T., Hampp, R., Kapur, M., & Fiedler, H.-P. (2006). Auxofuran, a novel metabolite that stimulates the growth of fly agaric, is produced by the mycorrhiza helper bacterium *Streptomyces* strain ach 505. *Applied and environmental microbiology*, 72(5), 3550–3557.
- Riera, E., Cabot, G., Mulet, X., García-Castillo, M., del Campo, R., Juan, C., Cantón, R., & Oliver, A. (2011). *Pseudomonas aeruginosa* carbapenem resistance mechanisms in spain: Impact on the activity of imipenem, meropenem and doripenem. *Journal of antimicrobial chemotherapy*, 66(9), 2022–2027.
- Riley, M. A., & Lizotte-Waniewski, M. (2009). Population genomics and the bacterial species concept. *Horizontal gene transfer: Genomes in Flux*, 367–377.

- Rimbara, E., Mori, S., Matsui, M., Suzuki, S., Wachino, J.-i., Kawamura, Y., Shen, Z., Fox, J. G., & Shibayama, K. (2012). Molecular Epidemiologic Analysis and Antimicrobial Resistance of *Helicobacter cinaedi* Isolated from Seven Hospitals in Japan. *Journal of Clinical Microbiology*, 50(8), 2553–2560. <https://doi.org/10.1128/jcm.06810-11>
- Risdian, C., Landwehr, W., Rohde, M., Schumann, P., Hahnke, R. L., Spröer, C., Bunk, B., Kämpfer, P., Schupp, P. J., & Wink, J. (2021). *Streptomyces bathyalis* sp. nov., an actinobacterium isolated from the sponge in a deep sea. *Antonie Van Leeuwenhoek*, 114, 425–435.
- Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2), 131–147.
- Rodríguez, M., Núñez, L. E., Brana, A. F., Méndez, C., Salas, J. A., & Blanco, G. (2011). Mutational analysis of the thienamycin biosynthetic gene cluster from *Streptomyces cattleya*. *Antimicrobial agents and chemotherapy*, 55(4), 1638–1649.
- Roelants, P., Konvalinkova, V., Mergeay, M., Lurquin, P., & Ledoux, L. (1976). Dna uptake by *Streptomyces* species. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis*, 442(1), 117–122.
- Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4, 2584. <https://doi.org/10.7717/peerj.2584>
- Rokas, A., & Carroll, S. B. (2006). Bushes in the tree of life. *PLoS biology*, 4(11), e352.
- Rong, X., & Huang, Y. (2010a). Taxonomic evaluation of the *Streptomyces griseus* clade using multilocus sequence analysis and dna–dna hybridization, with proposal to

- combine 29 species and three subspecies as 11 genomic species. *International journal of systematic and evolutionary microbiology*, 60(3), 696–703.
- Rong, X., & Huang, Y. (2010b). Taxonomic evaluation of the *Streptomyces griseus* clade using multilocus sequence analysis and DNA–DNA hybridization, with proposal to combine 29 species and three subspecies as 11 genomic species. *International Journal of Systematic and Evolutionary Microbiology*, 60(3), 696–703. <https://doi.org/10.1099/ijs.0.012419-0>
- Rosselló-Móra, R., & Amann, R. (2015). Past and future species definitions for Bacteria and Archaea. *Systematic and Applied Microbiology*, 38(4), 209–216. <https://doi.org/10.1016/j.syapm.2015.02.001>
- Rouli, L., Merhej, V., Fournier, P.-E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, 7, 72–85. <https://doi.org/10.1016/j.nmni.2015.06.005>
- Saati-Santamaría, Z., Peral-Aranega, E., Velázquez, E., Rivas, R., & García-Fraile, P. (2021). Phylogenomic Analyses of the Genus *Pseudomonas* Lead to the Rearrangement of Several Species and the Definition of New Genera. *Biology*, 10(8), 782. <https://doi.org/10.3390/biology10080782>
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4), 406–25. <https://doi.org/10.1093/oxfordjournals.molbev.a040454>
- Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10), e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>

- Santos, A., van Aerle, R., Barrientos, L., & Martinez-Urtaza, J. (2020). Computational methods for 16s metabarcoding studies using nanopore sequencing data. *Computational and Structural Biotechnology Journal*, 18, 296–305.
- Saunders, N. J., Boonmee, P., Peden, J. F., & Jarvis, S. A. (2005). Inter-species horizontal transfer resulting in core-genome and niche-adaptive variation within *Helicobacter pylori*. *BMC genomics*, 6, 1–16.
- Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., ... Sherry, S. T. (2021). Database resources of the national center for biotechnology information. *Nucleic acids research*, 50(D1), 20–26. <https://doi.org/10.1093/nar/gkab1112>
- Schatz, A., Bugle, E., & Waksman, S. A. (1944). Streptomycin, a Substance Exhibiting Antibiotic Activity Against gram-Positive and gram-Negative Bacteria. *Proceedings of the Society for Experimental Biology and Medicine*, 55(1), 66–69. <https://doi.org/10.3181/00379727-55-14461>
- Schleifer, K. H. (2009). Classification of Bacteria and Archaea: Past, present and future. *Systematic and Applied Microbiology*, 32(8), 533–542. <https://doi.org/10.1016/j.syapm.2009.09.002>
- Schleifer, K. H., & Kandler, O. (1972). Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriological reviews*, 36(4), 407–477.

- Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS ONE*, 6(12), e27310. <https://doi.org/10.1371/journal.pone.0027310>
- Schniete, J. K., Cruz-Morales, P., Selem-Mojica, N., Fernández-Martínez, L. T., Hunter, I. S., Barona-Gómez, F., & Hoskisson, P. A. (2018a). Expanding primary metabolism helps generate the metabolic robustness to facilitate antibiotic biosynthesis in *Streptomyces*. *MBio*, 9(1), 10–1128.
- Schniete, J. K., Cruz-Morales, P., Selem-Mojica, N., Fernández-Martínez, L. T., Hunter, I. S., Barona-Gómez, F., & Hoskisson, P. A. (2018b). Expanding Primary Metabolism Helps Generate the Metabolic Robustness To Facilitate Antibiotic Biosynthesis in *Streptomyces*. *mBio*, 9(1), e02283–17. <https://doi.org/10.1128/mbio.02283-17>
- Schoch, C. (n.d.). NCBI Taxonomy. Retrieved July 6, 2023, from <https://www.ncbi.nlm.nih.gov/books/NBK53758/>
- Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database : the journal of biological databases and curation*, 2020. <https://doi.org/10.1093/database/baaa062>
- Scully, B., Parry, M., Neu, H., & Mandell, W. (1986). Oral ciprofloxacin therapy of infections due to *Pseudomonas aeruginosa*. *The Lancet*, 327(8485), 819–822.

- Seipke, R. F., Kaltenpoth, M., & Hutchings, M. I. (2012). *Streptomyces* as symbionts: an emerging and widespread theme? *FEMS Microbiology Reviews*, 36(4), 862–876. <https://doi.org/10.1111/j.1574-6976.2011.00313.x>
- Servedio, R., Rubinfeld, R., Backurs, A., & Indyk, P. (2015). Edit Distance Cannot Be Computed in Strongly Subquadratic Time (unless SETH is false). *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, 51–58. <https://doi.org/10.1145/2746539.2746612>
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11), 2498–2504. <https://doi.org/10.1101/gr.1239303>
- Shen, X.-X., Li, Y., Hittinger, C. T., Chen, X.-x., & Rokas, A. (2020). An investigation of irreproducibility in maximum likelihood phylogenetic inference. *Nature Communications*, 11(1), 6096. <https://doi.org/10.1038/s41467-020-20005-6>
- Shi, J., Zeng, Y. J., Zhang, B., Shao, F. L., Chen, Y. C., Xu, X., Sun, Y., Xu, Q., Tan, R. X., & Ge, H. M. (2019). Comparative genome mining and heterologous expression of an orphan nrps gene cluster direct the production of ashimides. *Chemical Science*, 10(10), 3042–3048.
- Shi, T., & Falkowski, P. G. (2008). Genome evolution in cyanobacteria: The stable core and the variable shell. *Proceedings of the National Academy of Sciences*, 105(7), 2510–2515.

- Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, 27(1), 135–145. <https://doi.org/10.1002/pro.3290>
- Simeis, D. D., & Serra, S. (2021). *Actinomycetes*: a never-ending source of bioactive compounds—an overview on antibiotics production. *Antibiotics*, 10(5), 483. <https://doi.org/10.3390/antibiotics10050483>
- Singh, N., Singh, V., Rai, S. N., Mishra, V., Vamanu, E., & Singh, M. P. (2022). Deciphering the gut microbiome in neurodegenerative diseases and metagenomic approaches for characterization of gut microbes. *Biomedicine & Pharmacotherapy*, 156, 113958.
- Singh, S. B., & Barrett, J. F. (2006a). Empirical antibacterial drug discovery—Foundation in natural products. *Biochemical Pharmacology*, 71(7), 1006–1015. <https://doi.org/10.1016/j.bcp.2005.12.016>
- Singh, S. B., & Barrett, J. F. (2006b). Empirical antibacterial drug discovery—foundation in natural products. *Biochemical pharmacology*, 71(7), 1006–1015.
- Sivakala, K. K., Gutiérrez-García, K., Jose, P. A., Thinesh, T., Anandham, R., Barona-Gómez, F., & Sivakumar, N. (2021). Desert environments facilitate unique evolution of biosynthetic potential in *Streptomyces*. *Molecules*, 26(3), 588.
- Sivalingam, P., Hong, K., Pote, J., & Prabakar, K. (2019). Extreme Environment *Streptomyces*: Potential Sources for New Antibacterial and Anticancer Drug Leads? *International Journal of Microbiology*, 2019, 5283948. <https://doi.org/10.1155/2019/5283948>

- Skinnider, M. A., Merwin, N. J., Johnston, C. W., & Magarvey, N. A. (2017). PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Research*, 45, 320. <https://doi.org/10.1093/nar/gkx320>
- Smilack, J. D. Trimethoprim-sulfamethoxazole. In: In *Mayo clinic proceedings*. 74. (7). Elsevier. 1999, 730–734.
- Smith, J. M., Smith, N. H., O'Rourke, M., & Spratt, B. G. (1993). How clonal are bacteria? *Proceedings of the National Academy of Sciences*, 90(10), 4384–4388.
- Smith, T. F., Waterman, M. S., et al. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1), 195–197.
- Soltis, P. S., & Soltis, D. E. (2003). Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, 256–267.
- Spilker, T., Vandamme, P., & LiPuma, J. J. (2012). A Multilocus Sequence Typing Scheme Implies Population Structure and Reveals Several Putative Novel *Achromobacter* Species. *Journal of Clinical Microbiology*, 50(9). <https://doi.org/10.1128/jcm.00814-12>
- Srivastav, R., & Suneja, G. (2019). Recent advances in microbial genome sequencing. *Microbial Genomics in Sustainable Agroecosystems: Volume 2*, 131–144.
- Stackebrandt, E. (2006). Taxonomic parameters revisited: Tarnished gold standards. *Microbial Today*, 33, 152.
- Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, 44(4), 846–849. <https://doi.org/10.1099/00207713-44-4-846>

- Steemers, F. J., & Gunderson, K. L. (2005). Illumina, inc. *Pharmacogenomics*, 6(7), 777–782.
- Steinegger, M., & Salzberg, S. L. (2020). Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in genbank. *Genome biology*, 21, 1–12.
- Stuttard, C. (1979). Transduction of auxotrophic markers in a chloramphenicol-producing strain of *Streptomyces*. *Microbiology*, 110(2), 479–482.
- Subramaniam, G., Thakur, V., Saxena, R. K., Vadlamudi, S., Purohit, S., Kumar, V., Rathore, A., Chitikineni, A., & Varshney, R. K. (2020). Complete genome sequence of sixteen plant growth promoting *Streptomyces* strains. *Scientific Reports*, 10(1), 10294. <https://doi.org/10.1038/s41598-020-67153-9>
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3), 209–249.
- Sutherland, R., Boon, R., Griffin, K., Masters, P., Slocombe, B., & White, A. (1985). Antibacterial activity of mupirocin (pseudomonic acid), a new antibiotic for topical use. *Antimicrobial agents and chemotherapy*, 27(4), 495–498.
- Swinney, D. C. (2020). Phenotypic Drug Discovery. *Drug Discovery*, 1–19. <https://doi.org/10.1039/9781839160721-00001>
- Takahashi, K., Terai, Y., Nishida, M., & Okada, N. (2001). Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in lake tanganyika

as revealed by analysis of the insertion of retroposons. *Molecular biology and evolution*, 18(11), 2057–2066.

Takahashi, K., & Nei, M. (2000). Efficiencies of Fast Algorithms of Phylogenetic Inference Under the Criteria of Maximum Parsimony, Minimum Evolution, and Maximum Likelihood When a Large Number of Sequences Are Used. *Molecular Biology and Evolution*, 17(8), 1251–1258. <https://doi.org/10.1093/oxfordjournals.molbev.a026408>

Takezaki, N., & Nei, M. (1994). Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. *Journal of Molecular Evolution*, 39(2), 210–218. <https://doi.org/10.1007/bf00163810>

Talavera, G., & Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology*, 56(4), 564–577. <https://doi.org/10.1080/10635150701472164>

Terefework, Z., Nick, G., Suomalainen, S., Paulin, L., & Lindström, K. (1998). Phylogeny of *Rhizobium galegae* with respect to other rhizobia and agrobacteria. *International Journal of Systematic and Evolutionary Microbiology*, 48(2), 349–356.

Tettelin, H., Massignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Ros, I. M. y., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., ... Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for

- the microbial “pan-genome”. *Proceedings of the National Academy of Sciences*, 102(39), 13950–13955. <https://doi.org/10.1073/pnas.0506758102>
- Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology*, 3(9), 711–721.
- Thomas, J. C., Vargas, M. R., Miragaia, M., Peacock, S. J., Archer, G. L., & Enright, M. C. (2006). Improved multilocus sequence typing scheme for *Staphylococcus epidermidis*. *Journal of clinical microbiology*, 45(2), 616–9. <https://doi.org/10.1128/jcm.01934-06>
- Thompson, C. C., Chimetto, L., Edwards, R. A., Swings, J., Stackebrandt, E., & Thompson, F. L. (2013). Microbial genomic taxonomy. *BMC Genomics*, 14(1), 913–913. <https://doi.org/10.1186/1471-2164-14-913>
- Tian, L., Huang, C., Mazloom, R., Heath, L. S., & Vinatzer, B. A. (2020). Linbase: A web server for genome-based identification of prokaryotes as members of crowdsourced taxa. *Nucleic Acids Research*, 48(1), 529–537.
- Tidjani, A.-R., Lorenzi, J.-N., Toussaint, M., Dijk, E. v., Naquin, D., Lespinet, O., Bontemps, C., & Leblond, P. (2019). Massive Gene Flux Drives Genome Diversity between Sympatric *Streptomyces* Conspecifics. *mBio*, 10(5), e01533–19. <https://doi.org/10.1128/mbio.01533-19>
- Timmins, G. S., & Deretic, V. (2006). Mechanisms of action of isoniazid. *Molecular microbiology*, 62(5), 1220–1227.
- Toghueo, R. M. K., & Boyom, F. F. (2020). Endophytic *Penicillium* species and their agricultural, biotechnological, and pharmaceutical applications. *3 Biotech*, 10(3), 107.

- Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., et al. (2020). Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome biology*, *21*, 1–21.
- Tonkin-Hill, G., Corander, J., & Parkhill, J. (2023). Challenges in prokaryote pangenomics. *Microbial Genomics*, *9*(5), 001021.
- Tsang, A. K. L., Lee, H. H., Yiu, S.-M., Lau, S. K. P., & Woo, P. C. Y. (2017). Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny. *Scientific Reports*, *7*(1), 4536. <https://doi.org/10.1038/s41598-017-04707-4>
- Tyc, O., Song, C., Dickschat, J. S., Vos, M., & Garbeva, P. (2017). The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends in Microbiology*, *25*(4). <https://doi.org/10.1016/j.tim.2016.12.002>
- Uddin, T. M., Chakraborty, A. J., Khusro, A., Zidan, B. R. M., Mitra, S., Emran, T. B., Dhama, K., Ripon, M. K. H., Gajdács, M., Sahibzada, M. U. K., et al. (2021). Antibiotic resistance in microbes: History, mechanisms, therapeutic strategies and future prospects. *Journal of infection and public health*, *14*(12), 1750–1766.
- Upadhyay, U., & Vishwa, P. C. V. (2014). Growth promoters and novel feed additives improving poultry production and health, bioactive principles and beneficial applications: The trends and advances-a review. *Int. J. Pharmacol*, *10*(3), 129–159.

- Urwin, R., & Maiden, M. C. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology*, 11(10), 479–487. <https://doi.org/10.1016/j.tim.2003.08.006>
- Valent, P., Groner, B., Schumacher, U., Superti-Furga, G., Busslinger, M., Kralovics, R., Zielinski, C., Penninger, J. M., Kerjaschki, D., Stingl, G., et al. (2016). Paul ehrlich (1854-1915) and his contributions to the foundation and birth of translational medicine. *Journal of innate immunity*, 8(2), 111–120.
- Vandamme, P., & Peeters, C. (2014). Time to revisit polyphasic taxonomy. *Antonie Van Leeuwenhoek*, 106(1), 57–65.
- Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research*, 43(14), 6761–6771. <https://doi.org/10.1093/nar/gkv657>
- Vasilchenko, A. S., Julian, W. T., Lapchinskaya, O. A., Katrukha, G. S., Sadykova, V. S., & Rogozhin, E. A. (2020). A Novel Peptide Antibiotic Produced by *Streptomyces roseoflavus* Strain INA-Ac-5812 With Directed Activity Against gram-Positive Bacteria. *Frontiers in Microbiology*, 11, 556063. <https://doi.org/10.3389/fmicb.2020.556063>
- Velasco, J. D. (2009). When monophyly is not enough: exclusivity as the key to defining a phylogenetic species concept. *Biology & Philosophy*, 24(4), 473–486. <https://doi.org/10.1007/s10539-009-9151-4>

- Venkateswarlu, G, Murali Krishna, P., & Venkateswar Rao, L. (1999). Production of rifamycin using *Amycolatopsis mediterranei* (mtcc14). *Bioprocess Engineering*, 20, 27–30.
- Ventola, C. L. (2015). The Antibiotic Resistance Crisis. *P&T*, 40(4), 277–283.
- Verma, M., Lal, D., Kaur, J., Saxena, A., Kaur, J., Anand, S., & Lal, R. (2013). Phylogenetic analyses of phylum Actinobacteria based on whole genome sequences. *Research in Microbiology*, 164(7), 718–728. <https://doi.org/10.1016/j.resmic.2013.04.002>
- Veyrier, F., Pletzer, D., Turenne, C., & Behr, M. A. (2009). Phylogenetic detection of horizontal gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*. *BMC evolutionary biology*, 9, 1–14.
- Volff, J.-N., & Altenbuchner, J. (1998). Genetic instability of the *Streptomyces* chromosome. *Molecular microbiology*, 27(2), 239–246.
- Větrovský, T., & Baldrian, P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE*, 8(2), e57923. <https://doi.org/10.1371/journal.pone.0057923>
- Waksman, S. A., & Henrici, A. T. (1943). The Nomenclature and Classification of the Actinomycetes. *J Bacteriol.*, 46(4), 337–41. <https://doi.org/10.1128/jb.46.4.337-341.1943>
- Waksman, S. A., & Woodruff, H. B. (1941). *Actinomycetes antibioticus*, a new soil organism antagonistic to pathogenic and non-pathogenic bacteria. *J Bacteriol*, 42(2), 231–49. <https://doi.org/10.1128/jb.42.2.231-249.1941>

- Waksman, S. A., Reilly, H. C., & Johnstone, D. B. (1946). Isolation of streptomycin-producing strains of *Streptomyces griseus*. *Journal of bacteriology*, 52(3), 393–397.
- Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy †. *Applied and Environmental Microbiology*, 73(16), 5261–5267. <https://doi.org/10.1128/aem.00062-07>
- Wang, W., Yu, L., Hao, W., Zhang, F., Jiang, M., Zhao, S., & Wang, F. (2021). Multi-Locus Sequence Typing and Drug Resistance Analysis of Swine Origin *Escherichia coli* in Shandong of China and Its Potential Risk on Public Health. *Frontiers in Public Health*, 9, 780700. <https://doi.org/10.3389/fpubh.2021.780700>
- Wellington, E. M. H., Stackebrandt, E., Sanders, D., Wolstrup, J., & Jorgensen, N. O. G. (1992a). Taxonomic Status of *Kitasatosporia*, and Proposed Unification with *Streptomyces* on the Basis of Phenotypic and 16S rRNA Analysis and Emendation of *Streptomyces* Waksman and Henrici 1943, 339AL. *International Journal of Systematic and Evolutionary Microbiology*, 42(1), 156–160. <https://doi.org/10.1099/00207713-42-1-156>
- Wellington, E., Stackebrandt, E., Sanders, D., Wolstrup, J., & Jorgensen, N. (1992b). Taxonomic status of *Kitasatosporia*, and proposed unification with *Streptomyces* on the basis of phenotypic and 16s rRNA analysis and emendation of streptomyces waksman and henrici 1943, 339al. *International Journal of Systematic and Evolutionary Microbiology*, 42(1), 156–160.

- Wezel, G. P. v., Vijgenboom, E., & Bosch, L. (1991). A comparative study of the ribosomal RNA operons of *Streptomyces coelicolor* A3(2) and sequence analysis of rrnA. *Nucleic Acids Research*, 19(16), 4399–4403. <https://doi.org/10.1093/nar/19.16.4399>
- Whitman, W. B. (2015). Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Systematic and Applied Microbiology*, 38(4), 217–222. <https://doi.org/10.1016/j.syapm.2015.02.003>
- Widenbrant, E. M., Tsai, H.-H., Chen, C. W., & Kao, C. M. (2007). *Streptomyces coelicolor* undergoes spontaneous chromosomal end replacement. *Journal of bacteriology*, 189(24), 9117–9121.
- Williams, D. H., Stone, M. J., Hauck, P. R., & Rahman, S. K. (1989). Why are secondary metabolites (natural products) biosynthesized? *Journal of natural products*, 52(6), 1189–1208.
- Wink, J., Schumann, P., Atasayar, E., Klenk, H.-P., Zaburannyi, N., Westermann, M., Martin, K., Glaeser, S. P., & Kämpfer, P. (2017). ‘*Streptomyces caelicus*’, an antibiotic-producing species of the genus *Streptomyces*, and *Streptomyces canchipurensis* li et al. 2015 are later heterotypic synonyms of *Streptomyces muensis* ningthoujam et al. 2014. *International Journal of Systematic and Evolutionary Microbiology*, 67(3), 548–556.
- Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *PNAS*, 74(11), 5088–5090. <https://doi.org/10.1073/pnas.74.11.5088>

- Woodford, N., & Ellington, M. J. (2007). The emergence of antibiotic resistance by mutation. *Clinical Microbiology and Infection*, 13(1), 5–18. <https://doi.org/10.1111/j.1469-0691.2006.01492.x>
- WorldBankGroup. (2016). *By 2050, drug-resistant infections could cause global economic damage on par with 2008 financial crisis* [Accessed: 3rd January 2024]. <https://www.worldbank.org/en/news/press-release/2016/09/18/by-2050-drug-resistant-infections-could-cause-global-economic-damage-on-par-with-2008-financial-crisis>
- Wu, C.-F., Chen, S.-H., Chou, C.-C., Wang, C.-M., Huang, S.-W., & Kuo, H.-C. (2023a). Serotype and multilocus sequence typing of *Streptococcus suis* from diseased pigs in taiwan. *Scientific Reports*, 13(1), 8263.
- Wu, H.-J., Xiao, Z.-G., Lv, X.-J., Huang, H.-T., Liao, C., Hui, C.-Y., Xu, Y., & Li, H.-F. (2023b). Drug-resistant *Acinetobacter baumannii*: From molecular mechanisms to potential therapeutics. *Experimental and therapeutic medicine*, 25(5), 1–10.
- Wuisan, Z. G., Kresna, I. D. M., Böhringer, N., Lewis, K., & Schäberle, T. F. (2021). Optimization of heterologous darobactin a expression and identification of the minimal biosynthetic gene cluster. *Metabolic Engineering*, 66, 123–136.
- Xu, L., Huang, H., Wei, W., Zhong, Y., Tang, B., Yuan, H., Zhu, L., Huang, W., Ge, M., Yang, S., et al. (2014). Complete genome sequence and comparative genomic analyses of the vancomycin-producing *Amycolatopsis orientalis*. *BMC genomics*, 15, 1–18.

- Yang, L. P., & Keam, S. J. (2008). Retapamulin: A review of its use in the management of impetigo and other uncomplicated superficial skin infections. *Drugs*, 68, 855–873.
- Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Priesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., & Glöckner, F. O. (2014). The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research*, 42(1), 643–648. <https://doi.org/10.1093/nar/gkt1209>
- Yoon, S.-H., Ha, S.-m., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek*, 110(10), 1281–1286. <https://doi.org/10.1007/s10482-017-0844-4>
- Yuan, W. M., & Crawford, D. L. (1995). Characterization of *Streptomyces lydicus* wyec108 as a potential biocontrol agent against fungal root and seed rots. *Applied and environmental microbiology*, 61(8), 3119–3128.
- Zavascki, A. P., Goldani, L. Z., Li, J., & Nation, R. L. (2007). Polymyxin b for the treatment of multidrug-resistant pathogens: A critical review. *Journal of antimicrobial chemotherapy*, 60(6), 1206–1215.
- Zhang, Z., Wang, Y., & Ruan, J. (1997a). A proposal to revive the genus *Kitasatospora* (omura, takahashi, iwai, and tanaka 1982). *International Journal of Systematic and Evolutionary Microbiology*, 47(4), 1048–1054.
- Zhang, Z., Wang, Y., & Ruan, J. (1997b). A Proposal To Revive the Genus *Kitasatospora* (Omura, Takahashi, Iwai, and Tanaka 1982). *International Journal of Systematic and Evolutionary Microbiology*, 47(4), 1048–1054. <https://doi.org/10.1099/00207713-47-4-1048>

- Zhou, Z., Gu, J., Li, Y.-Q., & Wang, Y. Genome plasticity and systems evolution in *Streptomyces*. In: In *Bmc bioinformatics*. 13. Springer. 2012, 1–17.
- Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K. L., & Jensen, P. R. (2014). Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proceedings of the National Academy of Sciences*, 111(12), E1130–E1139.
- Zou, Q., Lin, G., Jiang, X., Liu, X., & Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. *Briefings in Bioinformatics*. <https://doi.org/10.1093/bib/bby090>
- Zou, Q., Lin, G., Jiang, X., Liu, X., & Zeng, X. (2020). Sequence clustering in bioinformatics: An empirical study. *Briefings in bioinformatics*, 21(1), 1–10.