Title: ActDES – a Curated Actinobacterial Database for Evolutionary Studies

Jana K. Schniete[1,3], Nelly Selem-Mojica[2], Anna S. Birke[1], Pablo Cruz-Morales[2], Iain S. Hunter[1], Francisco Barona-Gómez[2], & Paul A. Hoskisson[1*]

Affiliations: [1]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street, Glasgow, G4 0RE, United Kingdom

[2]Evolution of Metabolic Diversity Laboratory, Langebio, Cinvestav-IPN, Libramiento Norte Carretera Leon Km 9.6, 36821, Irapuato, Guanajuato, México.

[3]Biology Department, Edge Hill University, St Helens Road, Ormskirk, Lancashire, L39 4QP, UK.

***Corresponding Author:** [1]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street, Glasgow, G4 0RE, United Kingdom, Tel.: +44 (0)141 548 2819; Fax: +44 (0)141 548 4124; E-mail: Paul.hoskisson@strath.ac.uk

Key words: *Streptomyces,* evolution, primary metabolism, specialised metabolism, antibiotics, natural product, biosynthetic gene cluster.

Repositories: All data, databases, files and scripts can be found available on Microbiology Society Figshare [link to be added by the Society], https://github.com/nselem/ActDES and https://hub-binder.mybinder.ovh/user/nselem-actdes-2p4esa8n/tree#notebooks

Abbreviations: BGC - Biosynthetic gene cluster; CCR – Carbon-catabolite repression; CDS – Coding sequence; HGT – Horizontal gene transfer; PTS - phosphoenolpyruvate-dependent phosphotransferase system; WGS - whole genome sequence

Abstract

Actinobacteria is a large and diverse phylum of bacteria that contains medically and ecologically relevant organisms. Many members are valuable sources of bioactive natural products and chemical precursors that are exploited in the clinic and made using the enzyme pathways encoded in their complex genomes. Whilst the number of sequenced genomes has increased rapidly in the last twenty years, the large size, complexity and high G+C content of many Actinobacterial genomes means that the sequences remain incomplete and consist of large numbers of contigs with poor annotation, which hinders large scale comparative genomic and evolutionary studies. To enable greater understanding and exploitation of Actinobacterial genomes, specialised genomic databases must be linked to high-quality genome sequences. Here we provide a curated database of 612 high-quality actinobacterial genomes from 80 genera, chosen to represent a broad phylogenetic group with equivalent genome re-annotation. Utilising this database will provide researchers with a framework for evolutionary and metabolic studies, to enable a foundation for genome and metabolic engineering, to facilitate discovery of novel bioactive therapeutics and studies on gene family evolution.

Significance as a bioresource to the community

The Actinobacteria are a large diverse phylum of bacteria, often with large, complex genomes with a high G+C content. Sequence databases have great variation in the quality of sequences, equivalence of annotation and phylogenetic representation, which makes it challenging to undertake evolutionary and phylogenetic studies. To address this, we have assembled a curated, taxa-specific, non-redundant database to aid detailed comparative analysis of Actinobacteria. ActDES constitutes a novel resource for the community of Actinobacterial researchers that will be useful primarily for two types of analyses: (i) comparative genomic studies - facilitated by reliable identification of orthologs across a set of defined phylogenetically-representative genomes and (ii) phylogenomic studies which will be improved by identification of gene subsets at specified taxonomic level. These analyses can then act as a springboard for the studies of the evolution of virulence genes, the evolution of metabolism and identification of targets for metabolic engineering.

Data summary

All genome sequences used in this study can be found in the NCBI taxonomy browser https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi and are summarised along with Accession numbers in Table S1

All other data are available on Microbiology Society Figshare [link to be added by the Society]

a. Perl script files available on GitHub https://github.com/nselem/ActDES including details of how to batch annotate genomes in RAST from the terminal https://github.com/nselem/myrast
b. **Supp. Table S1** List of genomes from NCBI (Actinobacteria database.xlsx) available on Microbiology Society Figshare https://doi.org/10.6084/m9.figshare.13143407.v1
c. CVS genome annotation files including the FASTA files of nucleotide and amino acids sequences (individual .cvs files) available on Microbiology Society Figshare https://doi.org/10.6084/m9.figshare.13143407.v1
d. BLAST nucleotide database (.fasta file) available on Microbiology Society Figshare https://doi.org/10.6084/m9.figshare.13143407.v1
e. BLAST protein database (.fasta file) available on Microbiology Society Figshare https://doi.org/10.6084/m9.figshare.13143407.v1
f. Supp. Table S2 Expansion table genus level (Expansion table.xlsx Tab Genus level) available on Microbiology Society Figshare [link to be added by the Society]
g. Supp. Table S2 Expansion table species level (Expansion table.xlsx Tab species level) available on Microbiology Society Figshare https://doi.org/10.6084/m9.figshare.13143407.v1
h. All GlcP and Glk data – blast hits from ActDES database, MUSCLE Alignment files and .nwk tree files can be found at available on Microbiology Society Figshare https://doi.org/10.6084/m9.figshare.13143407.v1
i. Interactive trees in Microreact for Glk tree https://microreact.org/project/w_KDfn1xA/5a178533 and associated files can be found at Microbiology Society Figshare https://doi.org/10.6084/m9.figshare.13143407.v1
j. Interactive trees in Microreact for GlcP tree https://microreact.org/project/VBUdiQ5_k/045c95e1 and associated files can be found at Microbiology Society Figshare https://doi.org/10.6084/m9.figshare.13143407.v1
k. Jupyter Notebook for exploring ActDES in MyBinder can be found https://github.com/nselem/ActDES

Introduction

The increase in availability of bacterial whole genome sequencing (WGS) provides large amounts of data for evolutionary and phylogenetic analysis. However, there is great variation in the quality, annotation and phylogenetic skew of the data available in large universal databases, meaning that evolutionary and phylogenetic studies can be challenging. To address this variation, curated, high-level, taxa-specific, non-redundant sub-databases need to be assembled to aid detailed analysis. Given that there is a direct correlation between phylogenetic distance and the discovery of novel function [1–3], it is imperative that any derived databases must be phylogenetically representative and non-redundant to enable insight into the evolution of genes, proteins and pathways within a given group of taxa [1].

The phylum *Actinobacteria* is a major taxon amongst the *Bacteria,* which includes phenotypically and morphologically diverse organisms found on every continent and in virtually every ecological niche [4]. They are particularly common in soils, yet within their ranks are potential human and animal pathogens such as *Corynebacterium, Mycobacterium*, *Nocardia* and *Tropheryma,* inhabitants of the gastrointestinal tract (*Bifidobacterium* and *Scardovia)* as well as plant commensals and pathogens such as *Frankia*, *Leifsonia* and *Clavibacter* [4, 5]. Perhaps the most notable trait of the phylum is the renowned ability to produce bioactive natural products such as antibiotics, anti-cancer agents and immuno-suppressive agents, with genera such as *Amycolatopsis, Micromonospora* and *Streptomyces* being particularly prominent [6]. As a result, computational 'mining' of Actinobacterial genomes has become an important part of the drug discovery pipeline, with increasing numbers of online resources and software devoted to identification of natural product biosynthetic gene clusters (BGCs)[7–9]. It is important to move beyond approaches that rely on similarity searches of known BGCs and to expand searches to identify hidden chemical diversity within the genomes [6, 7, 10–13].

A recent study of 830 Actinobacterial genomes found >11,000 BGCs comprising 4,122 chemical families, indicating that there is a vast diversity of strains and chemistry to exploit [14], yet within each of these strains there will be hidden diversity in the form of cryptic BGCs. To exploit this undiscovered diversity as the technology develops and databases expand, new biosynthetic logic will emerge, yet we know little of how natural selection shapes the evolution of BGCs and how biosynthetic precursors are supplied to gene products of BGCs from primary metabolism and to identify targets for metabolic engineering of industrially relevant strains. Such logic will expedite industrial strain improvement processes, enabling titre increases and development of novel molecules, as well as the engineering of strains to use more sustainable feedstocks.

To aid this process we have created an Actinobacterial metabolism database including functional annotations for enzymes from 612 species to enable phylum-wide interrogation of

142 gene expansion events that may indicate adaptive evolution, help shape metabolic robustness
143 for antibiotic production [15] or enable the identification of targets for metabolic engineering.
144 **Act**inobacterial **D**atabase for **E**volutionary **S**tudies (ActDES) provides a curated list of high-
145 quality, phylum specific genomes and data to help researchers navigate the redundancy and
146 inconsistency in sequence databases in a simplified format that enables researchers with little
147 taxonomic knowledge to develop testable evolutionary hypotheses. To demonstrate the utility
148 of ActDES, we have detailed its construction and used it to investigate the glucose
149 permease/glucokinase system phylogeny across the Actinobacteria.

Methods

We generated ActDES, a database for evolutionary analysis of Actinobacterial genomes, in two formats - a database for interrogation by BLASTn or BLASTp for phylogenetic analysis and a primary metabolic gene expansion table, which can be mined at different taxonomic levels (**Supp. Table S1 and S2**) for specific metabolic functions from primary metabolism. A schematic overview of generation of the dataset is shown in **Fig.1**.

The database was generated via the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) to identify Actinobacterial genome sequences. The quality of the genome sequences was filtered by the number of contigs (<100 contigs per 2Mb of genome sequence) and the genomes were downloaded from the NCBI WGS repository (https://www.ncbi.nlm.nih.gov/Traces/wgs/). These genomes were then dereplicated to ensure that the database comprised a wide taxonomic range of the phylum, resulting in 612 species from 80 genera within 13 suborders of the Actinobacteria (**Supp. Table S1**).

Each of these 612 genomes was reannotated using RAST. Default settings were used to ensure equivalence of annotation across the database and the annotation files of each genome were downloaded (Data File: cvs files). These annotation files were subsequently used to extract all protein and nucleotide sequences into two files. Each of these files was subsequently converted into BLAST databases (a protein database and a nucleotide database https://doi.org/10.6084/m9.figshare.12167724) to facilitate phylogenetic analysis. Sequences of interest can be aligned using MUSCLE [16] and phylogenetic trees constructed using a range of tree construction software such as QuickTree [17], IQ tree [18] or MrBayes [19]. Subsequent trees may be visualized in software such as FigTree v1.4.2 (http://tree.bio.ed.ac.uk/ software/figtree/).

The RAST annotation files were also used to extract the functional roles of each CDS per genome and the level of gene expansion was assessed for each genome by counting the number of genes per species per functional category (gene function annotation). The dataset was then curated manually for central carbon metabolism and amino acid biosynthesis pathways to create the gene expansion table (**Supp. Table S2**) with the organisms grouped according to their taxonomic position. The quality of the data was checked at each step for duplicates and inconsistencies and was curated manually to exclude faulty entries.

As the NCBI taxonomy browser database is overrepresented in *Streptomyces* genomes due to the number of species that have been sequenced relative to other Actinobacteria, this is also reflected in the ActDES database (288 *Streptomyces* genomes from a total of 612 genomes). However this was addressed in the expansion table (**Supp. Table S2**) by calculating the mean occurrence of each functional category within each genus and then calculating an overall mean for the phylum to compensate. The mean occurrence of each

functional category per genus plus the standard deviation was also calculated and this was used to analyse the occurrence of each functional gene category per species within **Supp. Table S2**. A gene function annotation with a gene copy number value above the mean plus the standard deviation for each genus, indicated that there had been a gene expansion event in that species and this was noted. The gene expansion table (**Supp. Table S2**) enables researchers to identify groups of genes of interest for subsequent phylogenetic and evolutionary analysis, which can be performed with confidence due to the highly curated nature of the data included in the database.

Results

The gene expansion table (**Supp. Table S2**) lists 612 species of 80 genera within the Actinobacteria with data that provides an extensive analysis at the phylum level, which is the starting point for detailed phylogenomic studies. Gene expansions were identified in separate datasets at the genus and species levels, along with details of the numbers of genes in each functional category per species and the average numbers of genes in each functional category per genus expanded within the genomes. These data can be used subsequently in phylogenomic analyses to identify targets for metabolic engineering and gene function studies. Identification of expanded gene families may also facilitate the recognition of novel natural product biosynthetic gene clusters, for which gene expansion events of primary metabolic genes have been classified to be associated within BGCs as biosynthetic enzymes or through provision of additional copies of antibiotic targets that may subsequently function as resistance mechanisms [6, 11, 20–24].

This database has found utility for studying primary metabolic gene expansions in *Streptomyces*. It enabled a detailed *in silico* analysis of the duplication event leading to the two pyruvate kinases in the genus of *Streptomyces* subsequently enabling the functional characterisation of the two isoenzymes to reveal how they contribute to metabolic robustness [15]. ActDES may also be useful for investigating the distribution of primary metabolic genes across the phylum to link phenotype to genotype and phylogenetic position. An initial RpoB phylogeny has been constructed previously using this database [15] which provided a robust universal phylogeny for comparison of individual protein trees [25].

To demonstrate the utility of ActDES, the glucose permease/glucokinase system of the Actinobacteria was investigated. The role of nutrient-sensing in regulation of antibiotic biosynthesis is well known [26] with the enzyme glucokinase (Glk) playing a central role in carbon-catabolite repression in *Streptomyces* [27]. In most bacteria, CCR is mediated by the phosphoenolpyruvate-dependent phosphotransferase system (PTS), yet in *Streptomyces,* glucose uptake is mediated by the Major-Facilitator Superfamily (MFS) transporter, glucose permease (GlcP), and there is evidence for direct interaction between Glk and GlcP which

may mediate CCR [28]. Understanding the nature and distribution of these enzymes will play a key role in developing industrial fermentations with glucose as major carbon source. Investigating the distribution of the glucose permease/glucokinase system across the phylum shows that GlcP and Glk have been the subject of gene expansion events in some members of the Streptomycetales, most notably the *Streptomyces,* with a patchy distribution of the Glk/GlcP system across the remainder of the phylum (**Table S2; Genus tab**). However, where the Glk/GlcP system is found, the number of expansion events observed is greater for Glk than for GlcP (**Fig. 2 A & B**). The phylogenetic trees (**Fig. 2A & 2B**) clearly show two clades for Glk and GlcP within the Streptomycetales (Interactive trees are available via Microreact [29]: Glk https://microreact.org/project/w_KDfn1xA/5a178533 and GlcP https://microreact.org/project/VBUdiQ5_k/045c95e1). However, these clades differ in the number of sequences, with the Glk clades being equal in number, suggesting that a duplication event has occurred within the Streptomycetales (**Fig. 2A**). The is consistent throughout the order, with the patterns largely the same as that observed for *S. coelicolor*. This species has two ROK-family ATP-dependent glucokinases, SCO2126 (*glkA*) and SCO6260, that share around 50% amino acid sequence identity and each is found in one of the distinct clades (permease-associated kinases and orphan kinases; **Fig. 2A**). Whilst SCO2126 is a GlcP-associated kinase, the gene encoding SCO6260 is located in an operon including genes encoding a putative carbohydrate ABC-transporter system, which has been reported previously [30]. SCO6260 appears to be the only glucokinase in the database that is associated with an ABC-transporter. This may suggest that expansion of the Glk gene family in Streptomycetales might have occurred to extend the number of CCR-mediating kinases in the genome, adding increased regulatory complexity to carbohydrate metabolism in this group of organisms that use CCR as a major regulator of specialised metabolism.

The two clades for GlcP within the Streptomycetales differ in size suggesting either gene duplication followed by gene loss, or an expansion through horizontal gene transfer (HGT) has occurred. A detailed examination of these clades by species (**Table S2; Species Tab**) shows the presence of both scenarios. There are duplicated enzymes located within the same clade (as observed in *S. coelicolor;* Group I) or additional copies of the permease which are located in a phylogenetically distinct clade, which lacks congruence with the RpoB tree [15] and remarkably consists entirely of sequences from the genus *Streptomyces* (Group II; **Fig. 2B**). This suggests that they may have been acquired via HGT. The expansive nature of the duplicated Glk enzymes compared to GlcP may be due to the role played in CCR by the GlkA enzymes [27] and the different transcriptional activities under glycolytic and gluconeogenic conditions [31], yet quite how these different Glk enzymes interact with the permease(s) under various conditions requires further experimental investigation to understand their exact

physiological role, and how this may be translated in to industrial strain improvement processes.

Discussion

Large scale WGS and phylogenomic analysis is increasingly used for identifying targets for genome and metabolic engineering, studies of metabolic capabilities, pathogen phylogenomics and evolutionary studies. These studies are often complicated by the large number of sequences in the databases, database redundancy and the poor quality of some genome sequence data. The development of the high-quality, curated ActDES database, reported here, enables phylum-wide taxonomic representation of the Actinobacteria coupled with quality-filtered genome data and equivalent annotation for each CDS.

The intended primary use of ActDES will be in the study of primary metabolism, but it is not limited. It can also inform the development and evolution of metabolism in strains that produce bioactive metabolites, given the high representation of genera renowned for their ability to produce natural products such as *Streptomyces* and *Micromonospora*. Due to a greater understanding of BGC evolution and genome organisation in Actinobacteria it is becoming increasingly clear that genes whose functions are in primary metabolism may actually contribute directly to the biosynthesis of specialised metabolites and, hence, the identification of duplicates may indicate the presence of cryptic BGCs [6, 11] or, when associated with precursor biosynthetic genes, provide the raw material for the enzymes across multiple BGCs [32–34].

ActDES may also find utility in evolutionary studies of expanded gene families across the Actinobacterial phylum that contribute to virulence such as the *mce* locus which is known to facilitate host survival in Mycobacteria [35], but also facilitates xenobiotic substrate uptake in *Rhodococcus* [36] and enables root colonization and survival in *Streptomyces* [37]. With phylum-wide taxonomic representation of established Actinobacterial animal and plant pathogens, the scope for evolutionary studies using these data is enormous.

Usage Notes

The CVS files of each genome contains the RAST annotation details in addition to the DNA and protein sequences for each annotated CDS (https://doi.org/10.6084/m9.figshare.12167880). The Genome list contains the RAST ID (which is equivalent to the name of the .cvs file) along with the NCBI ID (sequence ID; **Table S1**) plus the species name which are included in the dataset. Further details of annotating batches of genomes in RAST can be found here https://github.com/nselem/myrast.

The primary metabolism expansion tables (**Supp. Table S2**) are organised by metabolic pathway along the top row with the Enzyme Commission (EC) number and functional

297 annotation, with the first column being the taxonomic assignment. The genus table shows the
298 average number of genes of the annotated function. Highlighted cells reflect gene expansion
299 events, i.e. those genes that are present in higher number than the overall mean across the
300 database plus the standard deviation.
301 It is suggested that the gene expansion table (**Supp. Table S2**) is searched in the first instance
302 (either by species or genus of interest or by a specific enzymatic function). This can be carried
303 out by a simple text search. This will then allow the identification of a query sequence from a
304 species or gene of interest (either nucleotide or amino acid sequence) which can then be
305 searched against the curated BLAST database allowing a detailed phylogenetic analysis of a
306 gene/protein of interest by using standard alignment and tree building software tools. These
307 data can also be used in detailed evolutionary analysis of selection, mutation rates etc. We
308 have set up a Jupyter Notebook through the MyBinder (https://mybinder.org/) project here to
309 enable ease of use of the code https://github.com/nselem/ActDES with a tutorial to enable use
310 of the database https://github.com/nselem/ActDES.
311
312

Funding Information

This work was funded through PhD studentships from the Scottish University Life Science Alliance (SULSA) to JKS and by the Industrial Biotechnology Innovation Centre (IBioIC) with GlaxoSmithKline funded PhD studentship to ASB.

Author Contributions

Conceptualization: JKS, NS-M, PAH, FB-G

Data curation: JKS & NS-M

Formal analysis: JKS, N.S-M, PC-M, & ASB

Funding acquisition: PAH, FB-G

Methodology: JKS, NS-M, PC-M, FB-G & PAH

Project administration: PAH & FB-G

Supervision: PAH, ISH & F.B-G

Writing – original draft: JKS, ASB & PAH

Writing – review and editing: JKS, N.S-M, PC-M, ASB, ISH, FB-G & PAH

Acknowledgements

We thank the Scottish Universities Life Science Alliance (SULSA) for BioSkape PhD funding to JKS, and a Mac Robertson Travelling Scholarship awarded to JKS to visit the laboratory of FB-G., an Industrial Biotechnology Innovation Centre (IBioIC) and GlaxoSmithKline funded PhD studentship to ASB, and a NERC (grant NE/M001415/1), BBSRC (grants BB/N023544/1 and BB/T001038/1) and BBSRC/NPRONET (grant NPRONET POC045) to PAH. PAH would also like to acknowledge the support of the Royal Academy of Engineering for the award of a Research Chair in Engineering Biology of Antibiotic Production. Work in the laboratory of FB-G. was funded by CONACyT, Mexico Metabolic Robustness in Streptomyces, and Langebio institutional funds to support PC-M. as a postdoctoral fellow and Royal Society Newton Advanced Fellowship (NAF\R2\18063).

Conflicts of interest

The authors declare that there are no conflicts of interest

Ethical statement

No ethical approval was required.

References

1. **Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, *et al.*** 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat Biotechnol* 2017;35:676–683.

2. **Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, *et al.*** A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 2009;462:1056–60.

3. **Kunin V, Cases I, Enright AJ, Lorenzo V de, Ouzounis CA**. Myriads of protein families, and still counting. *Genome Biol* 2003;4:401.

4. **Goodfellow M**. Bergey's Manual of Systematics of Archaea and Bacteria. 2015; Springer.

5. **Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, *et al.*** Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum. *Microbiology and Molecular Biology Reviews* 2007;71:495–548.

6. **Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A, *et al.*** Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat Prod Rep*. Epub ahead of print 2019. DOI: 10.1039/c9np00048h.

7. **Adamek M, Alanjary M, Ziemert N**. Applied evolution: phylogeny-based approaches in natural products research. *Nat Prod Rep* 2019;36:1295–1312.

8. **Ziemert N, Alanjary M, Weber T**. The evolution of genome mining in microbes – a review. *Nat Prod Rep* 2016;33:988–1005.

9. **Medema MH, Fischbach MA**. Computational approaches to natural product discovery. *Nat Chem Biol* 2015;11:639–648.

10. **Adamek M, Alanjary M, Sales-Ortells H, Goodfellow M, Bull AT, *et al.*** Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in *Amycolatopsis* species. *BMC Genomics* 2018;19:426.

11. **Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, *et al.*** Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biol Evol* 2016;8:1906–1916.

12. **Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, *et al.*** A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 2019;1–9.

13. **Sélem-Mojica N, Aguilar C, Gutiérrez-García K, Martínez-Guerrero CE, Barona-Gómez F**. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb Genom*;5. Epub ahead of print 2019. DOI: 10.1099/mgen.0.000260.

14. **Doroghazi JR, Metcalf WW**. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* 2013;14:611.

15. **Schniete JK, Cruz-Morales P, Selem-Mojica N, Fernández-Martínez LT, Hunter IS, *et al.*** Expanding Primary Metabolism Helps Generate the Metabolic Robustness To Facilitate Antibiotic Biosynthesis in *Streptomyces*. *Mbio* 2018;9:e02283-17.

16. **Edgar RC**. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;5:113.

17. **Howe K, Bateman A, Durbin R**. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 2002;18:1546–1547.

18. **Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, *et al.*** IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;37:1530–1534.

19. **Ronquist F, Teslenko M, Mark P van der, Ayres DL, Darling A, *et al.*** MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biol* 2012;61:539–42.

20. **Tang X, Li J, Millán-Aguiñaga N, Zhang JJ, O'Neill EC, *et al.*** Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. *Acs Chem Biol* 2015;10:2841–9.

21. **Schmidt KL, Peterson ND, Kustusch RJ, Wissel MC, Graham B, *et al.*** A Predicted ABC Transporter, FtsEX, Is Needed for Cell Division in *Escherichia coli*. *J Bacteriol* 2004;186:785–793.

22. **Steffensky M, Mühlenweg A, Wang Z-X, Li S-M, Heide L**. Identification of the Novobiocin Biosynthetic Gene Cluster of *Streptomyces spheroides* NCIB 11891. *Antimicrob Agents Ch* 2000;44:1214–1222.

23. **Kling A, Lukat P, Almeida DV, Bauer A, Fontaine E, *et al.*** Targeting DnaN for tuberculosis therapy using novel griselimycins. *Science* 2015;348:1106–1112.

24. **Peterson RM, Huang T, Rudolf JD, Smanski MJ, Shen B**. Mechanisms of Self-Resistance in the Platensimycin- and Platencin-Producing *Streptomyces platensis* MA7327 and MA7339 Strains. *Chem Biol* 2014;21:389–397.

25. **Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, *et al.*** Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. *Appl Environ Microb* 2007;73:278–288.

26. **Fernández-Martínez LT, Hoskisson PA**. Expanding, integrating, sensing and responding: the role of primary metabolism in specialised metabolite production. *Curr Opin Microbiol* 2019;51:16–21.

27. **Gubbens J, Janus M, Florea BI, Overkleeft HS, Wezel GP**. Identification of glucose kinase-dependent and -independent pathways for carbon control of primary metabolism, development and antibiotic production in *Streptomyces coelicolor* by quantitative proteomics. *Molecular Microbiology* 2012;86:1490–1507.

28. **Wezel GP van, König M, Mahr K, Nothaft H, Thomae AW, *et al.*** A New Piece of an Old Jigsaw: Glucose Kinase Is Activated Posttranslationally in a Glucose Transport-Dependent Manner in *Streptomyces coelicolor* A3(2). *J Mol Microb Biotech* 2006;12:67–74.

29. **Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, *et al.*** Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom*;2. Epub ahead of print 2016. DOI: 10.1099/mgen.0.000093.

30. **Bertram R, Schlicht M, Mahr K, Nothaft… H**. In silico and transcriptional analysis of carbohydrate uptake systems of Streptomyces coelicolor A3 (2). *Journal of Bacteriology* 2004;186:1362–1373.

31. **Schniete JK, Reumerman R, Kerr L, Tucker NP, Hunter IS, *et al.*** Differential transcription of expanded gene families in central carbon metabolism of *Streptomyces coelicolor* A3(2). *Access Microbiol*. Epub ahead of print 2020. DOI: 10.1099/acmi.0.000122.

32. **Chan YA, Podevels AM, Kevany BM, Thomas MG**. Biosynthesis of polyketide synthase extender units. *Nat Prod Rep* 2009;26:90–114.

33. **Pfeifer BA, Khosla C**. Biosynthesis of Polyketides in Heterologous Hosts. *Microbiol Mol Biol R* 2001;65:106–118.

34. **Zhang G, Li Y, Fang L, Pfeifer BA**. Tailoring pathway modularity in the biosynthesis of erythromycin analogs heterologously engineered in *E. coli*. *Science Advances* 2015;1:e1500077e1500077.

35. **Arruda S, Bomfim G, Knights R, Huima-Byron T, Riley L**. Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. *Science* 1993;261:1454–1457.

36. **Mohn WW, Geize R van der, Stewart GR, Okamoto S, Liu J, *et al.*** The Actinobacterial *mce4* Locus Encodes a Steroid Transporter. *Journal of Biological Chemistry* 2008;283:35368–35374.

37. **Clark LC, Seipke RF, Prieto P, Willemse J, Wezel GP van, *et al.*** Mammalian cell entry genes in *Streptomyces* may provide clues to the evolution of bacterial virulence. *Scientific Reports* 2013;3.

38. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, *et al.*** The RAST Server: Rapid Annotations using Subsystems Technology. *Bmc Genomics* 2008;9:75.

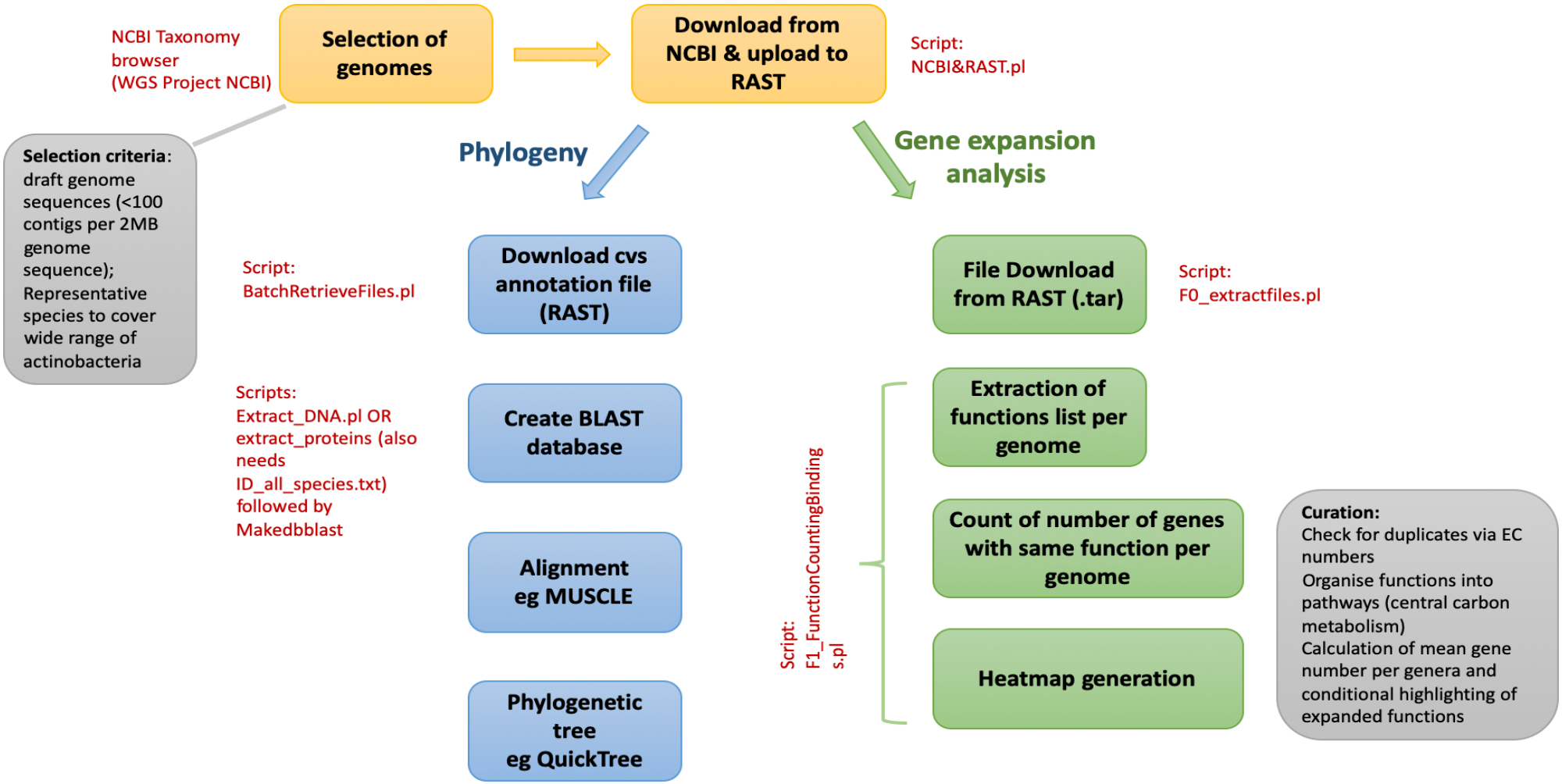
39. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.

Figures

Fig. 1. Schematic workflow for the creation of the ActDES database. Genomes were selected from NCBI Taxonomy browser and uploaded for annotation to RAST[38]. The annotated genomes were the processed for two different analyses. Firstly, the functional roles were downloaded and for each functional role the numbers of occurrences per genome were counted in order to obtain an expansion table (**Supp. Table S2**) by comparing the mean of each genus to the overall mean of all genera. Secondly, the genomes were used to extract all nucleotide and protein sequences in FASTA format which could then be queried by sequence using BLAST [39]. The hits were aligned in MUSCLE [16] and after refinement the alignment was used to construct phylogenetic trees in Quicktree [17].

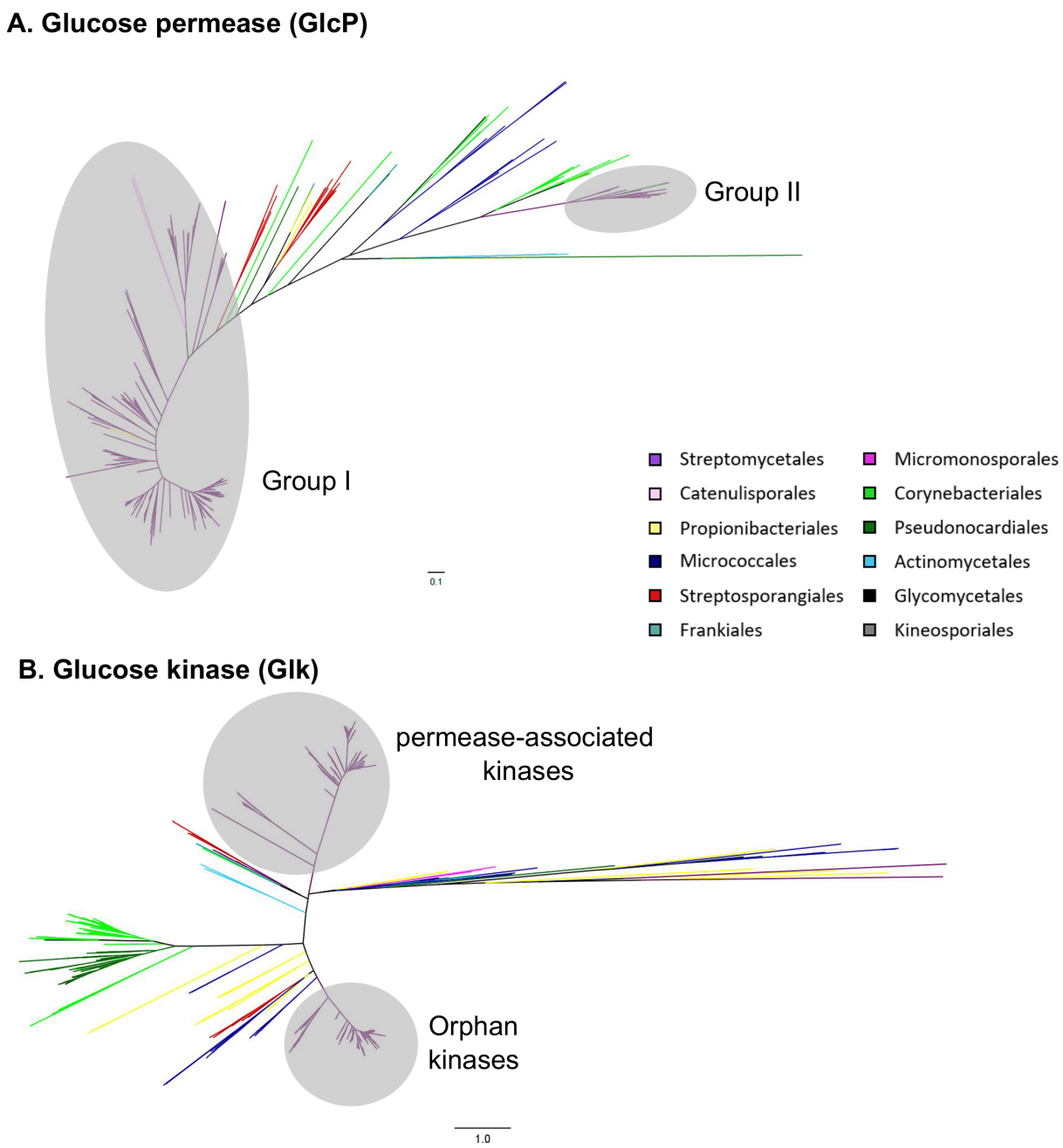
Fig. 2A) Actinobacterial-wide phylogenetic tree of glucose permeases (GlcP) **B)** Actinobacterial-wide phylogenetic tree of Glucokinases (Glk). Trees are colour-coded according to the NCBI Taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi). Interactive trees are also available via Microreact [29]: GlcP https://microreact.org/project/VBUdiQ5_k/045c95e1 and Glk https://microreact.org/project/w_KDfn1xA/5a178533.

Fig. 1


NCBI Taxonomy browser (WGS Project NCBI)
Selection of genomes
Download from NCBI & upload to RAST
Script: NCBI&RAST.pl
Selection criteria: draft genome sequences (<100 contigs per 2MB genome sequence); Representative species to cover wide range of actinobacteria
Phylogeny
Gene expansion analysis
Script: BatchRetrieveFiles.pl
Download cvs annotation file (RAST)
Scripts: Extract_DNA.pl OR extract_proteins (also needs ID_all_species.txt) followed by Makedbblast
Create BLAST database
Alignment eg MUSCLE
Phylogenetic tree eg QuickTree
File Download from RAST (.tar)
Script: F0_extractfiles.pl
Extraction of functions list per genome
Count of number of genes with same function per genome
Heatmap generation
Script: F1_FunctionCountingBindings.pl
Curation:
Check for duplicates via EC numbers
Organise functions into pathways (central carbon metabolism)
Calculation of mean gene number per genera and conditional highlighting of expanded functions

467

Fig. 2


A. Glucose permease (GlcP)
Group II
Group I
0.1
Streptomycetales
Micromonosporales
Catenulisporales
Corynebacteriales
Propionibacteriales
Pseudonocardiales
Micrococcales
Actinomycetales
Streptosporangiales
Glycomycetales
Frankiales
Kineosporiales
B. Glucose kinase (Glk)
permease-associated kinases
Orphan kinases
1.0

468