nature microbiology

Supplementary information

https://doi.org/10.1038/s41564-026-02385-5

Complex multicellularity is linked with expanded specialized metabolite production in microorganisms

In the format provided by the authors and unedited

Supplementary Information for "*Complex multicellularity is linked with expanded specialized metabolite production in microbes*"

RA Salamzade, LR Kalan, and CR Currie 2026, Nature Microbiology

Contents*:*

Supplementary Notes

Systematic investigations of the relationship between bacterial lifestyle, genetic traits, and environmental distributions with BGC-ome size

Of the 2,442 characterized BGCs cataloged in the MIBiG database[23], 96.3% were from five bacterial phyla, *Actinomycetota*, *Bacillota*, *Cyanobacteriota*, *Myxococcota* and *Pseudomonadota*, and one fungal phylum, *Ascomycota*. To investigate why certain species encode for greater numbers of specialized metabolites, we assessed the relationship of BGC-ome size with other quantitative variables reflective of genomic, functional, and ecological traits (**Extended Data Fig. 3ab; Supplementary Table S3**). Using a dataset of 17,844 species-representative genomes across the aforementioned five bacterial phyla, we reestablished that BGC-ome size was correlated with known factors such as genome size[91–93].

The number of distinct carbohydrate utilization genes was broadly found to be the second most strongly associated factor with BGC-ome size. Phylogenetically-aware genome-wide association studies performed individually for each phylum revealed specific CAZyme families associated with increased BGC-ome size[94]. These analyses revealed 75 families in *Actinomycetota*, 80 families in *Bacillota*, 56 families in *Cyanobacteriota*, 1 family in *Myxococcota*, and 92 families in *Pseudomonadota* as significantly associated with increased biosynthetic capacity. Only two families were found as significantly associated in four or more phyla, glycosyl hydrolase 19 and glycosyl hydrolase 19 subfamily 2, which correspond to a family of chitinases (**Extended Data Fig. 3c)**.

We also observed that genome-wide frequencies for some amino acids, as well as the number of distinct oxidative phosphorylation genes and adherence genes are consistently and significantly associated with larger BGC-omes in each of the five bacterial phyla. In contrast, some factors, such as the frequencies of other amino acids, were negatively correlated with BGC-ome size. In particular, the frequency of leucine was observed to be positively correlated with BGC-ome size, whereas the frequency of isoleucine was negatively associated. This trend might occur because the biosynthesis of leucine and coenzyme A, a key precursor of polyketides, share an enzymatic step that is not required for isoleucine synthesis[95].

Genomic prediction of optimal growth temperatures further revealed that species with larger BGC-omes are mostly mesophilic (**Extended Data Fig. 3d**), which aligns with recent findings that mesophiles genomically encoded greater numbers of toxins[96] and thermophilic fungi exhibit lower biosynthetic potential[97]. Further, of the bacterial genera which have at least ten characterized BGCs, many have large BGC-omes, devote at least 300 kb in their genomes to specialized metabolism, and are environmentally associated (**Extended Data Fig. 3e**).

Evolutionary investigations to determine support for proposed phylogenetic model of *Actinomycetes*

In addition to our primary phylogeny for *Actinomycetota* based on single-copy core ortholog groups, we constructed several alternate phylogenetic models. For some of the alternate models, multicellular *Actinomycetes* were discreetly placed within a single clade; however, further analyses suggest that their placement is likely a technical artifact due to them sharing greater numbers of genes with each other than with morphologically simpler genera (**Extended Data Fig. 5**).

Next, we developed a statistical framework to measure the number of conserved sites along protein alignments for single-copy core ortholog groups that supported morphologically simple families from clade-2 sharing a more recent last common ancestor with the *Streptomycetaceae* family than with families from clade-1 (**Fig. 2cde**). This framework provided consensus support for our phylogenetic model in contrast to one where multicellular genera are aggregated within a single clade (**Fig. 2e; Extended Data Fig. 5**). Constructing a phylogeny of SsgB provided further support that the marker gene has largely evolved vertically, with instances from clade-1 and clade-2 genera mostly grouping separately (**Fig. 2f**).

We also searched for remnants of multicellularity-associated genetic markers, such as *ssgB* and *ramAB*[30,33,34,78,98,99] across the entire *Actinomycetota* phylum. Several morphologically simple *Acitnomycetes* genera in clade-2 were found to harbor orthologs of multicellularity-associated markers, most notably chaplins, but not found outside of the *Actinomycetes* class (**Fig. 2a**).

Finally, a recent study identified a positive correlation between the extent of gene loss and mutation rates[100]. Ancestral state reconstruction analysis for ortholog group copy counts showed that genera with the greatest counts of cumulative gene loss were also phylogenetically the most distant from the last common ancestor of *Actinomycetes* along single-copy ortholog group-based core genome, providing yet additional support for the phylogenetic model we propose (**Fig. 2g**).

Evolutionary analysis of ansamycin synthesizing BGCs and rifamycin resistance across *Actinomycetota*

Clinically used rifamycins are antimicrobials derived from specialized metabolites produced by *Amycolatopsis* and *Micromonospora,* two highly divergent genera from clade-1 *Actinomycetes.* The metabolite belongs to the family of naphthalenic ansamycins, which are also produced by some taxa belonging to clade-2 *Actinomycetes*, mainly *Streptomyces*. Ansamycin production requires 3-amino-5-hydroxybenzoic acid (AHBA) biosynthesis which are encoded by the genes *rifG-N* and *rifJ* in the rifamycin BGC from *Amycolatopsis* (**Extended Data Fig. 7a)**. While many AHBA biosynthesis genes are related to genes involved in shikimate biosynthesis, AHBA synthase, encoded by *rifK*, is proposed to have a distinct evolutionary origin[101] and underlie two distinct roles in AHBA biosynthesis[36]. It has also been used as a marker to identify new ansamycins in metagenomic data[102].

We constructed a phylogeny of proteins from *Actinomycetota* genomes that were homologous to RifK to chart the protein family's evolutionary landscape. A distinct clade corresponding to close RifK-like homologs was identified and found to be *Actinomycetes* specific (**Extended Data Fig. 7b**). While early diverging variants of this clade were primarily from clade-2 genomes, many sub-clades of the protein family featured proteins from both clade-1 and clade-2 genomes. Proteins for some of these sub-clades were further partitioned distinctly for clade-1 and clade-2 genomes in accordance with our established evolutionary model for the class. However, some RifK homologs featured high numbers of insertion-element (IS)-associated genes in their nearby genomic context (+/- 50 kb). For instance, investigation into the genomic proximity of a RifK homolog from a *Micromonospora* genome revealed the presence of polyketide synthase (PKS) genes homologous to those found in the rifamycin BGC, additional AHBA biosynthesis genes, and flanking transposases (**Extended Data Fig. 7c**). Both sides of the BGC feature multiple transposases with one side additionally featuring genes predicted to encode for a putative phage integrase and a reverse transcriptase. This genomic architecture strongly suggests the BGC was

acquired via lateral transfer. In addition, two phylogenetically related RifK-like homologs were also found on scaffolds predicted to be plasmids.

The subclade containing RifK is largely associated with naphthalenic ansamycins (**Extended Data Fig. 7b**). The distinct grouping of these RifK genes matches what has previously been reported from a limited analysis of RifK homologs from characterized ansamycin-associated BGCs[103]. Many RifK homologs in this subclade also feature multiple PKS genes in their nearby genomic flanking contexts. In addition, this subclade featured RifK homologs from both clade-1 and clade-2 *Actinomycetes* genomes with some degree of phylogenetic divergence between proteins from different clades. A smaller subclade within the clade featured proteins from both clade-1 and clade-2 and included a RifK homolog from the BGC underlying rubradarin biosynthesis, which has a substantially different genetic organization in comparison to other BGCs characterized to synthesize naphthalenic ansamycins. Thus, this subclade might have diverged ancestrally and be paralogous to other BGCs associated with naphthalenic ansamycins.

To further assess whether rifamycin-like naphthalenic ansamycin BGCs have evolved primarily through vertical evolution, we comprehensively searched for rifamycin homologs across *Actinomycetota* genomes using fai[104] (**Extended Data Fig. 7d**). Comparative analysis of non-fragmented and high-quality BGC instances alongside characterized BGCs was then performed to determine domain-resolution ortholog groups. A set of 33 ortholog groups was identified as single-copy and present in at least 90% of BGC instances investigated. These 33 domain-resolution ortholog groups were used to establish a multi-gene-based phylogeny (**Extended Data Fig. 7e**). Similar to the RifK-specific phylogeny, this high-resolution phylogeny also revealed a subclade with a clear partition between BGCs from clade-1 and clade-2 genomes. However, unlike the RifK-phylogeny, it further resolved that BGCs from *Amycolatopsis* and *Micromonospora* partitioned distinctly from each other. We examined the distribution of domain-resolution ortholog groups across BGCs and determined that BGCs from *Amycolatopsis* and *Micromonospora* shared more ortholog groups with each other than with BGCs from *Streptomyces* (**Extended Data Fig. 7f-g**). Because both *Amycolatopsis* and *Micromonospora* are genera belonging to distinct orders of clade-2, these observations suggest that rifamycin-like BGCs were ancestrally acquired by their last common ancestor and have since diverged primarily through vertical evolution. Manual investigation of sequence similarity and organization of characterized naphthalenic BGCs provided further support that the rifamycin BGC from *Amycolatopsis* was more similar to rifamorpholine and kanglemycin BGCs from other *Amycolatopsis* than to the rifamycin BGC from *Micromonospora* (**Extended Data Fig. 7h**).

In addition to AHBA synthase, another key enzyme in the biosynthesis of AHBA is aminoDAHP synthase (RifH), which is related to DAHP synthases involved in the shikimate biosynthesis pathway. Because the shikimate pathway is widespread across different bacteria[101], to understand the evolution of aminoDAHP synthases, all DAHP synthases were extracted from *Actinomycetota* genomes and used to construct a phylogeny including RifH homologs from characterized ansamycin(-like) BGCs included[23] (**Extended Data Fig. 7i**). The majority of inferred aminoDAHP synthases were found within a single clade of type II DAHP synthases and featured instances from both clade-1 and clade-2 genomes. Further assessment of the taxonomic prevalence of different types of DAHP synthases and suspected RifH homologs in inferred AHBA gene clusters revealed that 90.5% of *Actinomycetes* genomes feature type II DAHP synthases (**Extended Data Fig. 7j**). In contrast, only 35.3% of other *Actinomycetota* genomes belonging to different classes featured type II DAHP synthases. This analysis ultimately provided further support that AHBA and ansamycin biosynthesis are *Actinomycetes* innovations.

Because production of antimicrobials requires self-resistance, to further understand the emergence of ansamycins, which can perform such functions, we also investigated the distribution of rifamycin resistance across the *Actinomycetota* phylum. Although rifamycin resistance has been reported to be widespread across multiple bacterial phyla[37,105], the rifamycin resistance associated regulatory motif, rifamycin-associated element (RAE)[105], and HelR helicase homologs, shown to contribute to rifamycin resistance[37], were specific to or enriched in the *Actinomycetes* class (**Extended Data Fig. 7k-l**). Furthermore, these rifamycin resistance associated elements were found in both morphologically complex and simple taxa as well as both clade-1 and clade-2 genomes. Interestingly, the RAE was particularly prevalent within *Streptomycetaceae*.

Together, we think our investigations support the emergence of ansamycins within *Actinomycetes*, potentially within ancestral taxa more similar to extant representatives of clade-2 than clade-1 of the class. This is because: (i) early diverging variants of AHBA synthase were primarily from clade-2 genomes, (ii) *Streptomyces*, which is part of clade-2, appears to possess a more diverse repertoire of ansamycin BGCs relative to individual clade-1 genera, and (iii) RAE is particularly prevalent in clade-2 families. Although close homologs of RifK were identified within potential mobile genetic elements, signatures of vertical evolution are also present when investigating the domain architectures of BGCs for synthesis of rifamycin-like naphthalenic ansamycins. Thus, we think either: (i) the common ancestor of clade-1 *Actinomycetes* acquired the rifamycin-like BGC family from clade-2 or (ii) the last common ancestor of the full *Actinomycetes* class resembled large-genome taxa from clade-2 more than other extant members of the class.

Other scenarios that we cannot conclusively rule out are that *Amycolatopsis* and *Micromonospora* acquired a chaxamycin-like ansamycin BGC in independent transfer events that followed after they speciated. Donors for more recent transfer events could further either be from clade-2, e.g. *Streptomyces*, or another clade-1 genome. For the former scenario, similarities in the auxiliary gene content of ansamycin BGCs within genera and between more related genera could simply arise by convergence because the pangenomic repertoire of biosynthetically associated genes is more similar between them. However, we think convergent refinement of a basal ansamycin BGC to synthesize identical rifamycin compounds[106] is unlikely. In addition, for the latter scenario, while sequential intra-clade-1 transfer is a possibility, we think the high sequence divergence between ansamycin BGCs from *Amycolatopsis* and *Micromonospora* - approaching the level of sequence divergence they exhibit to naphthalenic ansamycin synthesizing BGCs from *Streptomyces* - indicates that if this was the evolutionary path taken, then the transfer likely occurred early in the evolutionary history of the genera.

Literature-based assessment of links between multicellularity and biosynthetic potential in non-fungal eukaryotes

Beyond fungi, a review of scientific literature suggests that other eukaryotic lineages that form more complex multicellular structures relative to morphologically simpler phylogenetic neighbors might be enriched biosynthetically. For instance, slime molds - a group of multicellular fungal-like protists[107] - are known to produce a wide array of specialized metabolites[58,108]. While often overlooked, animal genomes also encode BGCs[59,109,110]; however, their unicellular phylogenetic neighbors, choanoflagellates, are also noted to produce some secondary metabolites, such as sterols[111]. To our knowledge, further research is thus needed to resolve whether animals are enriched in biosynthetic traits relative to choanoflagellates.

Similarly, to our knowledge, such investigations are currently also needed for the plant kingdom, which are known to be rich in specialized metabolites and a major source of natural products[56]. However, non-vascular plants, such as the bryophyte *Calohypnum plumiforme*, have been reported to produce BGCs[112]. Recently, the biosynthetic potential of plants was systematically characterized using plantiSMASH by the authors of the bioinformatics toolkit[57,113], including charophytes, freshwater green algae which are phylogenetic neighbors of land plants. Re-assessment of this data revealed that early-diverging *Streptophyta* plants, which are generally physiologically simpler, had fewer BGCs compared to more morphologically advanced vascular land plants (p=1.85e-09; *Wilcoxon rank-sum statistic*=5.89; one-sided Wilcoxon rank-sum test). Future characterization of additional non-vascular green algae, including from the phylum *Chlorophyta*, should further resolve whether multicellularity is associated with biosynthetic expansions in plants.

Supplementary Methods

Construction of a bacterial and archaeal timetree and determination of evolutionary expansions in biosynthetic traits

High-resolution phylogenies were independently constructed for the five bacterial phyla *Actinomycetota*, *Bacillota*, *Cyanobacteriota*, *Myxococcota* and *Pseudomonadota* at the genus-resolution. For each phylogeny, outgroup genomes from alternate phyla were included for rooting. The best hit for 74 proteins which are largely universal and found in single-copy across all bacteria were identified across genomes using GToTree (v1.7.07) in "best-hit" mode with otherwise default parameters[66]. Some genera were dropped from investigation because they did not feature >50% of pre-computed single copy-core genes. Protein alignments were retained and filtered for sites where <10% of sequences featured gaps or ambiguous ("X") characters. Full sequences which had gaps or ambiguous characters for ≥10% sites were also subsequently filtered. Retained protein alignments were then used as input to IQ-TREE (v2.2.0.3) to construct a phylogeny using an edge-unlinked partition analysis with models selected using ModelFinder limited to the LG and WAG substitution matrices and with 1000 bootstraps requested (https://github.com/Kalan-Lab/Ancestral_BGC_Exansions_Study/blob/main/Supplementary_Figures.pdf)[67,114].

To assess the topological rigidity of the phylogenies constructed for the five phyla, we further filtered protein alignments to retain only 50% of the sites which are least impacted by compositional heterogeneity[115]. For each phylum, we first determined the median value of median pairwise GARP:FIMNKY ratios across protein alignments. This value was used as the GARP:FIMNKY ratio cutoff to discern between AT and GC rich genera. Afterwards, we computed ζ, the degree of compositional bias per site as described in Muñoz-Gómez et al. 2019 and filtered for 50% of the sites with the lowest score. These filtered protein alignments were then used as input to IQ-TREE to construct a phylogeny once again using an edge-unlinked partition analysis. The topological structure of phylogenies achieved with and without accounting for compositional heterogeneity for each of the five phyla were largely congruent.

To construct a reliable and comprehensive phylogeny of bacteria and archaea, we first selected representative genomes for 77 bacterial and archaeal orders with greater than 200 genomes in GTDB R214 and which had a "Complete" or "Chromosome" quality assembly with an N50 ≥100,000 from the major (most sequenced) genus to serve as a representative (**Supplementary Table S1**). The orders of *Vampirovibrionales* and *Cyanobacteriales* were automatically included to leverage valuable fossil information for construction of a timetree downstream. The orders UBA6257, which belongs to Candida Phyla Radiation[69,116] and *Fusobacteriales*, which have been shown to be highly impacted by horizontal gene transfer[117], were excluded to achieve clearer partitioning between bacteria and archaea and improve accuracy in phylogenomic modeling. GToTree was next used to identify the best-hit for a set of 16 universal ribosomal proteins[69]. The representative genome for the order of *Nitrososphaerales* was dropped for not featuring at least eight of the ribosomal proteins. The alignments for the ribosomal proteins were filtered for sites which featured gaps in fewer than 10% of genomes which resulted in retention of 1,095 amino acid sites along eight of the ribosomal proteins (L14, S10, S8, S6, L5, L2, S3 and S17). IQ-TREE was run using an edge-unlinked partition analysis and ModelFinder limited to the LG and WAG substitution matrices with a constraint tree manually constructed based on the phylum specific phylogenies constructed. Version v1.6.12 of IQ-TREE was used for this modeling instead of v2.2.0.3 to properly apply phylogenetic constraints. Targeted assessment of alternate phylogenetic placements of

Myxococcales was additionally performed to resolve whether it was more appropriate to clade the order closer to *Desulfobacterota* or *Pseudomonadota* based on aggregate bootstrap support. The final phylogeny was rooted using the minimal ancestor deviation (MAD) technique[118].

Bacterial and archaeal timetree construction: A concatenated alignment of the universal proteins was also created and formatted into Phylip format. CODEML (v4.10.0) was first used to estimate the Hessian matrix, gradient vector, and maximum-likelihood estimates of branch lengths. The LG amino acid substitution model rate matrix, an alpha value of 0.5 for gamma rates, and 5 discrete gamma categories were used to run CODEML. Afterwards, the *rst2* output file was renamed to *in.BV* and used as input for three replicate runs of MCMCTree to sample the posterior distribution for node ages. An approximate likelihood approach, an independent rate model and a uniform prior on phylogeny node ages were used to run MCMCTree. The parameters for the gamma-Dirichlet prior for the mean evolutionary rate were specified as α_μ=2, β_μ=40, α=1, based on expectations from past studies[70,119] and the parameters for the gamma-Dirichlet prior for the relaxed-clock model's rate dispersion parameter were set to α_μ=1, β_μ=10, α=1. The Monte Carlo Markov Chain (MCMC) was run with a burn-in of 100,000 and a sample size of 20,000 for 1,000 iterations. Results were assessed both visually and using effective sample size to confirm convergence in dating estimates[120,121]. All non-leaf nodes had an ESS of at least 4,141. TreeViewer[122] and iToL[123] were used to visualize results. For the iToL-based visualization in **Fig. 1c**, the standard deviation of the posterior age distribution was displayed as the width of branches.

Assessment of BGC-ome size differences across bacterial orders: To account for intra-order variation in BGC-ome sizes, genus-representative genomes were selected for each of the 74 orders (**Supplementary Table S1**). Selection of genus-representative genomes was based primarily on assembly completeness category and secondarily on N50 as a measure of completeness. Representative genomes also needed to have a minimal N50 of 100,000. BGC and CAZy annotations of the genus-level representative genomes across the 74 orders were performed using the same approach we took for annotating species-level representative genomes for the systematic investigation of the five bacterial phyla (*described below*). An Ornstein-Uhlenbeck processes-based phylogenetic lasso approach, implemented as the function "estimate_shift_configuration" in the *l1ou*[74] R library, was used to determine evolutionary shifts in the median BGC-ome size across the phylogeny (**Fig. 1c**).

Assessment of the number of distinct metabolites for each order in the phylogeny depicted in **Fig. 1c** were gathered from the Natural Products Atlas (using database files from 06/2023). The genus associated with each metabolite was mapped to orders based on the categorization of the genus in GTDB R214. Note, there is potential that some genera designations associated with metabolites might be inaccurate or have a different taxonomic classification if a genome was available for the source sample and classified using GTDB-Tk. The number of distinct metabolic nodes, which group metabolites more broadly compared to metabolic clusters, were calculated for each order.

Assessment of intra-order biosynthetic diversity: To assess the extent of biosynthetic diversity within orders, we annotated domains from Pfam-A (release 36.0) for key enzymes of BGCs using PyHMMER (v0.10.4). These key biosynthetic proteins were identified based on the keyword "biosynthetic (rule-based-clusters" in the "gene_functions" qualifier of CDS features in antiSMASH BGC region GenBank files and are important for rule-based detection of BGCs by antiSMASH. Domain annotations were sorted by score and sequentially regarded as present if they did not exhibit an overlap greater than

≥10% of its alignment length to regions matching other domains at higher scores. The carriage of different Pfam domains across bacterial and archaeal orders was encoded as a matrix and provided as input to psaps alongside our previously established maximum-likelihood phylogeny of the 74 microbial orders based on ribosomal proteins. The psaps analysis was run with default settings and used to assess the discovery of new Pfam domains from biosynthetic enzymes as a function of ribosomal phylogenetic divergence. For more information on psaps, please see the section: "*Phylogenetically standardized assessment of genome fluidity and pangenome size associated metrics".*

Genome selection and annotation for systematic investigations of genetic traits and ecological trends correlated with bacterial biosynthetic potential

To account for intra-genus variability in genomic content, representative genomes were selected for individual species (**Supplementary Table S3**). The process of representative genome selection for genera and species was primarily based on the category of assembly completeness and secondarily based on assembly N50. All genus-representative genomes were used to establish phylogenies; however, only genomes with an N50 ≥100,000 were used for functional annotation to alleviate false negative detection of genetic traits[124,125] and reduce partial detection of BGCs[126,127]. For some genera used for phylogenetic inference, no species representatives were selected for functional annotation. In certain instances, the most preferential genomic assembly for a species or genera was no longer available for download on NCBI (largely due to authors of assemblies making them private retroactively). For these cases, the next best available genomic assembly was selected using iterative assessment. Assemblies were downloaded using ncbi-genome-download (v0.3.3)[128] and gene-calling was performed using pyrodigal (v3.0.0)[129] and prepTG (v1.3.9) to create GenBank files with CDS features[104].

Functional annotations: BGC annotation was performed using antiSMASH (v7.0.0)[20] with largely default settings, prodigal[130] selected as the gene finding tool and GenBank files from prepTG processing as input. The BGC-ome size was computed as the summed length of all BGC regions identified whereas the NRPS+PKS-ome size was the sum of select BGC regions which feature NRPS or PKS associated types and at least one protocore region confidently predicted as a BGC (not necessarily NRPS or PKS). Plasmids and temperate phages were annotated in genomes using geNomad (v1.11.0)[131]. HMMs for carbohydrate utilization and transport genes from the dbCAN database (version 07262023)[88,132] were downloaded and searched for in predicted proteomes using pyhmmer[133] (v0.10.4). As recommended by the dbCAN curators, an E-value <1e-18 and a coverage >0.35 were required of alignments. Independent DIAMOND blastp alignment (v2.0.15)[134] searches in very-sensitive mode were used to identify homologs of transposons and insertion sequences from ISFinder[135] (downloaded from https://github.com/thanhleviet/ISfinder-sequences on Dec 24th, 2023) and adherence associated proteins from VFDB (downloaded on Nov 17th, 2023)[136]. Prior to searches in target species-representative genomes, sequences for the two datasets were independently collapsed for redundancy using CD-HIT (v4.8.1)[137] at 95% identity and 90% bi-directional coverage ("-c 0.95 -aS 0.90 -aL 0.90"). Top hits from the databases were identified for proteins from representative genomes based on bitscore provided that alignments exhibited ≥30% identity and ≥70% bi-directional coverage. To account for functional redundancy, the number of distinct records for carbohydrate utilization and adherence genes identified within each genome was calculated. For example, if multiple instances of a specific glycosyl hydrolase family were annotated in a genome (e.g. GH15), the count for the genome would only be

incremented by one. Each protein was allowed to represent multiple dbCAN/CAZy profile HMMs. The total count of proteins matching any CAZy profile per genome was also calculated. Insertion element counts were calculated per genome as the number of proteins shown to exhibit homology to sequences in the ISFinder-associated database. Genes involved in oxidative phosphorylation were identified using KofamScan[138] (v1.3.0) with KOfam profile HMMs associated with the KEGG pathway map00190. The unique count of KOfams identified in each target genome was assessed to overcome complications with functional redundancy when correlating the measure with BGC-ome size. The predicted optimal growth temperature was calculated based on the proportion of specific amino acids using the formula described by Zeldovich et al.[139].

Phylum-specific associations of CAZyme families with BGC-ome size: Pyseer (v1.3.11) was used to perform GWAS analyses to associate specific CAZyme families and sub-families with BGC-ome size for each phylum. Specifically, a linear-mixed-model was applied and accounted for similarities between genomes based on phylogenies constructed using GToTree[66].

Phylum-specific annotations of markers associated with multicellularity and sporulation: From comparative genomic analysis between genus representatives of the *Actinomycetota* phylum (see section: "In depth phylogenomic and evolutionary investigations of the *Actinomycetota* phylum") we identified an ortholog group that corresponded to SsgB based on sequence similarity to a query protein from *Streptomyces venezuelae* (SCO1541). A profile HMM was constructed for the ortholog group using HMMER (v3.3.2) and a reflexive search for the proteins used to construct the profile suggested that an appropriate score threshold for sensitive detection of the profile was 44.5. This was used to identify additional instances of the ortholog group amongst all species-representative *Actinomycetota* genomes. Social lifestyle-associated KOfam HMMs for *Myxococcota* were gathered from the study by Murphy et al. 2021[29] and searched across all species-representative genomes from the phylum using KofamScan. Genes associated with sporulation formation in *Myxococcota*, previously determined using proteomics by Dahl et al. 2007[76], as well as multicellularity and akinete formation in *Cyanobacteriota*, as described by Boden et al. 2025[31] and Singh et al. 2020[77], respectively, were searched using a strategy similar to what was described for identifying SsgB orthologs across *Actinomycetota*. For each phylum, OrthoFinder (v2.5.4) was run using default settings on a limited set of genus-representative genomes. Profile HMMs were then constructed for coarse ortholog groups corresponding to proteins from the three studies and hmmsearch was used to comprehensively search all species-representative genomes. Score thresholds were determined by reflexive searches of profile HMMs on the sequences used to construct them.

Assessment of ecological distribution: The SandPiper (v0.2.0; https://sandpiper.qut.edu.au/)[82] database was assessed to determine the distribution of GTDB taxa across public metagenomes. SandPiper's classification of metagenomes as either environmental or eukaryotic-host associated based on a gradient boosting approach was also used for specific analyses. For investigations of well-represented genera in the MIBiG database, if multiple partitions of a genus existed in GTDB, metagenomic presence for the canonical genus was assessed across all of the individual partitions in aggregate.

In depth phylogenomic and evolutionary investigations of the *Actinomycetota* phylum

Annotation of BGCs, CAZymes and other functional traits: Distributions for BGC-ome size, genome size, distinct CAZy enzyme family counts, and distinct oxidative phosphorylation enzyme counts were calculated for each genus based on data gathered from species-representative genomes (**Supplementary Table S3)**. Consensus sequences were determined for orthologs using primarily MUSCLE in super5 mode (v5.1) using functionalities in HMMER (v3.3.2). For a small subset of ortholog groups, MAFFT with default options (v7.505)[140] was used to perform alignments because MUSCLE experienced difficulties. Consensus sequences for ortholog groups were used to simplify functional annotation with EggNOG mapper (webserver v2.1.12) with *de novo* Pfam domain alignment requested (**Supplementary Table S5**)[141].

Alternate phylogenomic and related investigations: To further assess the topology proposed in **Fig. 2ab** for the phylum of *Actinomycetota*, evolutionary model A, we applied alternate approaches for constructing a phylogeny of the phylum. First, we used IQ-TREE using ModelFinder, limited to substitution matrices WAG and LG, to generate individual phylogenies for the 34 single-copy core genes which were then provided as inputs to construct a consensus network using SplitsTree4CE (v4.19.2)[142] (**Extended Data Fig. 5a**). SplitsTree4CE consensus network construction was run using TreeSizeWeightedMean for edge weights, a threshold percent of 5.0 and a high-dimension filter applied. *Rubrobracter_D* was used to root the network. We also used a larger set of 84 near-SCC genes found in $\geq$ 80% of genomes and constructed a phylogeny similar to what was done for the strict-SCC set (**Extended Data Fig. 5d**). This phylogeny was in support of evolutionary model B, where *Streptomycetaceae* grouped with the majority of other mycelium forming *Actinomycetes* families from clade-1. To understand the discrepancy with our prior phylogenetic models, we further assessed the similarity in gene content between genomes. We developed hamtree, a program which accepts OrthoFinder results, computes the Hamming distance for shared ortholog groups between genomes and then uses these distances to construct a neighbor-joining tree. In the neighbor-joining tree constructed by hamtree, *Streptomycetaceae* and mycelium-forming clade-1 genomes were similarly grouped together in a monophyletic clade, suggesting the maximum likelihood phylogeny based on near-SCC ortholog groups might be affected by gene conservation in these taxa and large-scale gene loss in non-mycelium forming taxa (**Extended Data Fig. 5e**). Finally, we constructed protein alignments using MUSCLE for 186 core genes - ortholog groups found in all the genomes regardless of being single-copy per genome or not. Two core ortholog groups were dropped because alignments could not be properly established without the removal of sequences. Next, individual ortholog group phylogenies were created once more using IQ-TREE using ModelFinder limited to substitution matrices LG and WAG which were then provided as input for STAG (downloaded on Sep 12, 2023)[143] to determine a consensus species phylogeny (**Extended Data Fig. 5f**). The STAG based phylogeny was concordant with evolutionary model A for the phylum.

D.stat calculation: We developed the D.stat metric to assess genetic support between two phylogenetic models for the evolution of *Actinomycetota* based on a core protein alignment (**Fig. 2b-e**). The structure for this test was derived from Patterson's D-statistic, used to detect signals of relatively recent introgression across a focal genome amongst related eukaryotic species[144,145]. In our adaptation of the statistic, we similarly adopt a four-leaf phylogenetic structure; however, we use a concatenated protein alignment instead of a whole-genome nucleotide alignment. The D.stat was computed for multiple combinations of four genomes which featured (i) one genome belonging to clade-2 that was not part of

the family *Streptomycetaceae*, (ii) one genome from the family *Streptomycetaea*, (iii) one genome from clade-1 and (iv) an *Acidimicrobiia* genome. The use of a protein alignment is to account for the extreme evolutionary distances in the taxonomic relationships being investigated. Only biallelic sites along the concatenated protein alignment from 34 single-copy-core genes which featured the same residue in both *Acidimicrobiia* genomes were considered. Conserved residues in the *Acidimicrobiia* outgroup were treated as ancestral alleles (*A*). The alternate allele was regarded as a derived allele (*B*). Under evolutionary model A, where clade-2 includes *Streptomycetaceae* as an early diverging sister of the remaining families, we would expect there to be a greater number of *AABA* and *BBAA* sites in comparison to *ABAA* and *BABA* sites (**Fig. 2bc**). As an example, an *AABA* site is one where the genome belonging to a family from clade-1 has a derived alternate allele, while the rest of the genomes - smaller genomes from clade-2 and the *Streptomycetaceae* genome - feature the ancestral allele observed in the outgroup *Acidimicrobiia*. In contrast, under evolutionary model B, where *Streptomycetaceae* groups with the majority of the clade-1 families capable of mycelium formation, we would expect the occurrences of *AABA+BBAA* and *ABAA+BABA* sites to be equal (**Extended Data Fig. 4bc**). Positive values for D.stat, supportive of the model A topology structure, were largely observed regardless of whether we used the single-copy core protein alignment (feature 2,833 sites which are conserved amongst the outgroup *Acidimicrobiia*) or the near single copy core protein alignment (featuring 5,209 sites which are conserved amongst the outgroup *Acidimicrobiia*). Viewing D.stat distributions stratified by individual single-copy-core ortholog groups, revealed that some ortholog groups have distributions around 0 or have negative values, indicating either complex scenarios of horizontal transfer, the presence of convergent evolution, or *Streptomycetaceae*-specific differentiation from ancestral alleles common to the rest of *Actinomycetes*. Finally, we want to note, that positive values of D.stat could also arise if the last common ancestor of the entire *Actinomycetes* class resembled large-genome clade-2 taxa, such as *Streptomyces*, at the conserved sites assessed more than other extant members of the class.

Detection of additional multicellularity-associated genetic markers and construction of ssgB phylogeny: The genomic carriage of multicellularity-associated genetic markers, SsgB[30,78,79], SepX[99], RamAB[33,34,98] and ChpABCDEFGH[33,98], are shown based on the presence of corresponding ortholog groups from OrthoFinder analysis (**Fig. 2a; Supplementary Table S5**). Literature reports of mycelium formation or complex clustering of cells was also manually gathered, primarily from Bergey's Manual of Systematics of Archaea and Bacteria[32] (**Supplementary Table S4**). A maximum-likelihood phylogeny for the ortholog group corresponding to SsgB was constructed using IQ-TREE with ModelFinder limited to substitution matrices LG and WAG and based on a protein alignment generated using MUSCLE (v5.1) in super5 mode which was trimmed using trimAl (v1.4.1) in strict mode and filtered of a sequence which featured gaps or ambiguous characters for ≥25% sites. The program AnGST (committed Aug 8th, 2016)[146] was used to root the gene tree based on the species phylogeny (**Fig. 2f**).

Relation between phylogenetic distance of single-copy core and gene-loss rates: The count of ortholog groups, determined using OrthoFinder, at ancestral nodes of the *Actinomycetes* phylogeny was inferred using Count (downloaded version dated Apr 2, 2010 from https://www.iro.umontreal.ca/~csuros/gene_content/count.html)[147] with the asymmetric Wagner model (**Fig. 2g**). Because gene loss has been shown to be more common than gene duplications or gains in bacteria[148–150], the gene gain-to-loss cost ratio parameter was set to two. Results from Count were parsed to estimate the cumulative gene-loss events since the LCA of *Actinomycetes* for each extant genus. The

ete3 library[151] was further used to determine the phylogenetic distance between the LCA of *Actinomycetes* and each extant genus (leaves of the phylogeny).

Comprehensive assessment of the ecological distributions, resistance genes and characterized BGCs of representative families from Actinomycetota*:* The 133 *Actinomycetota* genomes used for comparative genomics featured genera across 31 families. The ete3 library was used to collapse the SCC phylogeny of the phylum down to the family level (**Extended Data Fig. 6a-e**). To understand the habitat distribution of the 31 families, we assessed the number of metagenomes from each broad environment category they were found in at ≥0.01% relative abundance using SandPiper[82]. Metagenomes which corresponded to common metadata labels (e.g. groundwater metagenome) were manually binned into select categories (e.g. aquatic; non-marine). To determine the distribution of characterized BGCs from MIBiGv3.1[23] across the *Actinomycetota* phylogeny, we extracted core biosynthesis proteins from their GenBank records and searched all 33,804 genomes belonging to the phylum in GTDB R214 using DIAMOND BLASTp with an E-value threshold of 1e-3 in fast mode. Characterized BGCs were regarded as present in a genome if ≥50% of core biosynthesis proteins were found, where each identified core biosynthesis protein had a match with ≥95% identity and ≥70% coverage. To sensitively assess antibiotic resistance traits across the phylum, which features many species for which resistance mechanisms might be currently poorly understood, we used the Core Resfams HMM set (v1.2)[152] with HMMER and applied recommended gathering thresholds. Mean counts of each Resfam per genome were calculated for each family.

Assessing the phylogenetic distribution and contexts of BGC associated ortholog groups: Key biosynthetic ortholog groups were identified if at least one member protein was found within an antiSMASH BGC region and marked as “rule-based-clusters" for the qualifier gene-function of CDS features. Of these, 366 ortholog groups were identified as present within BGC contexts in genomes from both major clades of *Actinomycetes* and prioritized for further investigation (**Extended Data Fig. 6f-g**). To more specifically associate ortholog groups with certain types of BGCs, the antiSMASH designated type of only the corresponding protoclusters it was found to overlap with were considered. This was important because BGC regions might include several independent protoclusters in BGC-rich taxa that are simply called as belonging to the same region due to genomic proximity[153].

Assessing the evolutionary relationship between ansamycin encoding BGCs and the origins of AHBA biosynthesis: A comprehensive HMM-based search for RifK (K16016) was performed across all 33,804 *Actinomycetota* genomes from GTDB R214[75] using KoFamScan[138]. Afterwards 1,122 sequences for matching hits and close homologs (depicting E-value ≤1e-100) were gathered. The program hmmsearch from HMMER3[80] was also used to search for homologs of RifK more broadly across the bacterial domain using NCBI's bacterial RefSeq database[154] (downloaded on Nov. 4, 2021). Based on those results, three RifK homologs from separate *Bacillota* genera were included with our set of RifK homologs from *Actinomycetota* and 21 RifK homologs from MIBiG[23] BGCs for phylogenetic analysis. Sequences were aligned using MUSCLE super5[155] and trimmed in strict mode using trimal[156]. The phylogeny was constructed using FastTree 2[157] on the trimmed protein alignment of RifK and midpoint rooted for visualization (**Extended Data Fig. 7b**).

Afterwards, we extracted 50 kb of upstream and downstream flanking contexts for RifK homologs in *Actinomycetota* genomes using fai (v1.6.13) in single-query mode. Homologs from which such context

could be extracted in full were assessed for the presence of insertion sequence (IS) elements and polyketide synthases (PKSs). IS-associated protein sequences were downloaded from https://github.com/thanhleviet/ISfinder-sequences, sourced from the ISFinder database[135]. PKS protein sequences were extracted from the *Amycolatopsis* rifamycin BGC in MIBiG. These were used as queries with DIAMOND blastp[134] to search for homologs in the proteins found in the contexts of RifK homologs with an E-value threshold of 1e-20. Predictions of PKS homologs were further required to be ≥1,000 aa in length. Genomes with RifK homologs were also annotated for temperate phages and plasmids using geNomad to determine whether RifK homologs were ever found in such MGE contexts.

A coarse and sensitive search for homologous BGCs to rifamycin (MIBiG entry BGC0000136) across all 33,804 *Actinomycetota* genomes from GTDB R214 was performed using fai (v1.3.11)[104]. fai was run requesting that homologous regions to the query BGC in target genomes possess matches to ≥50% of core proteins for biosynthesis with E-value thresholds of 1e-10 and loosened criteria for the percentage of total query genes needing a match (**Extended Data Fig. 7d**). A subset of 60 homologous regions identified (those that featured ≥25% of the total genes from the query BGC at an average amino acid identity (AAI) of ≥60%) alongside 11 reference BGCs from MIBiG were further compared and investigated using zol (v1.6.13). To account for vast divergences in sequence between the BGC instances in consideration, zol was run in "--domain-mode" to determine domain-resolution ortholog groups. This led to the identification of 33 near single-copy core domain-resolution ortholog groups conserved in ≥90% of considered BGCs. Trimmed protein alignments for these ortholog groups were gathered from zol output and provided as input to IQ-TREE to construct a multi-gene phylogeny using the same settings that we used for constructing the *Actinomycetota* species phylogeny (**Extended Data Fig. 7e**).

A subset of 39 naphthalenic BGCs which formed a distinct phylogenetic clade that included both rifamycin BGCs from *Amycolatopsis* and *Micromonospora* were further investigated for indications of vertical evolution. This included 15 BGCs from *Streptomyces*, 15 BGCs from *Amycolatopsis*, and 7 BGCs from *Micromonospora*. The Jaccard index was used to measure shared domain-resolution ortholog groups (from zol analysis) between BGCs belonging to the same and to different genera. Characterized BGCs from this set of 39 BGCs, in addition to the BGC for synthesizing rubradirin, were further compared and visualized using clinker[158].

The protein sequence for RifH, an aminoDAHP synthase, from the rifamycin BGC in *Amycolatopsis mediterranei* was used as a query to extract 10 homologs from 27 ansamycin(-like) BGCs from MIBiGv3.1[23]. Instances of 3-deoxy-D-arabinoheptulosonate 7-phosphate (DAHP) synthases were then identified across the full set of *Actinomycetota* genomes in GTDB R214 using KofamScan with the three KEGG ortholog (KO) profile HMMs corresponding to the enzyme. Classification of type I and II DAHP synthases[159] was performed by using hmmsearch without the "--domtblout" option to annotate TIGRfam protein families[160] TIGR00034.1 (aroFGH; type I) and TIGR01358.1 (DAHP_synth_II; type II) as well the Pfam domain PF18152 (DAHP_snth_FXD), where the sequences were assigned to the best matching domain based on E-value. The aggregate set of DAHP synthase proteins was collapsed to remove redundancy using CD-HIT[137] with parameters: “-c 0.80 -aS 0.90 -aL 0.90” and afterwards supplemented with the RifH homologs from MIBiG BGCs. Next, MUSCLE super5[155] was used to perform multiple sequence alignment, trimal[156] was applied in strict mode to perform site-specific filtering and sequences with 125 or more gaps (threshold selected based on histogram of gap counts) were discarded prior to phylogeny construction of the remaining 2,923 sequences using FastTree 2[157] and visualization using iTol[123].

As part of a supplementary analysis, instances of characterized BGCs for the synthesis of rifamycin (BGC0000136 and BGC0000137) and chaxamycin (BGC0001785) were identified in isolate genomes belonging to *Amycolatopsis* (GCA_000282715.1), *Micromonospora* (GCA_000384275.1), and *Streptomyces* (GCA_001013905.1). OrthoFinder was run on predicted proteins from these genomes and the output were provided as input to paiof (v1.0.2b) to determine the percent identity of core one-to-one orthologs between genomes.

Assessing the distribution and conservation of rifamycin-resistance markers: The rifamycin associated element (RAE)[105] was searched for across all 33,804 *Actinomycetota* genomes in GTDB R214 using BLASTn (v2.5.0)[161] with the task mode set to blastn-short and restrictions set to perform only ungapped alignments with the following four 19-mers provided as queries: "GGGGCTTGCGGCAAGGCCC", "GGGCCTTGCGGCAAGGCCC", "GGGGCTTGCGGCAAGCCCC", "GGGCCTTGCGGCAAGCCCC". Only exact hits which featured a sequence identity and coverage of 100% were considered (**Extended Dat Fig. 7k**; **Supplementary Table S6**).

The query sequence for HelR, a recently identified and widespread helicase contributing to rifamycin resistance[37], was obtained from the CARD database (ARO:3007208; *Streptomyces venezuelae*)[162]. DIAMOND BLASTp was used to search for homologous proteins across the full set of *Actinomycetota* genomes from GTDB R214 in very-sensitive mode with an E-value threshold of 1e-3. Alignments were loosely filtered for bi-directional coverage of ≥50% and sequence identity ≥30% to retain remote homologs and the corresponding target proteins were extracted. Only hits to proteins from genomes regarded as species-level representatives for the *Actinomycetota* phylum, previously selected for systematic investigations of variables associated with BGC-ome size, were considered. Multiple sequence alignment of the HelR homologous sequences was performed using MUSCLE super5 mode[155], trimmed using trimal[156] in strict mode and used to create an approximate maximum-likelihood tree with FastTree2 (v2.1.11)[157] (**Extended Dat Fig. 7l**).

Phylogenomic investigations of biosynthetic potential across the fungal kingdom

Timetree construction: A timetree for fungi was constructed by pruning the comprehensive phylogeny of 316 genomes for a subset of 212 fungal genomes featuring ≥55 Universal Fungal Core Genes (UFCGs) universal gene markers per assembly in addition to the outgroup genome, *Capsaspora owczarzaki* (GCF_000151315.2) using ete3. The stringent requirement for the number of UFCG markers per genome was to construct a concatenated protein alignment to provide to MCMCTree. Across all 213 genomes, 22 UFCGs were used to establish the strict core genome, where each gene contained sites that were ambiguous or gaps in at most 10% of genomes. The phylogeny was rooted by the *Capsaspora* outgroup genome and updated to encode priors for calibrations for MCMCTree analysis (**Supplementary Table S8**). MCMCTree analysis was performed using an identical methodology to what was previously described for constructing a timetree of bacterial and archaeal order representatives. To achieve convergence, the Monte Carlo Markov Chain (MCMC) was run with a burn-in of 20,000 and a sample size of 40,000 for 100 iterations in triplicate. All non-leaf nodes were confirmed to possess an effective sample size of at least 772 for the run used for dating estimates[120].

Functional annotations: BGCs were predicted using antiSMASH[20] with the taxon parameter set to fungi and GFF files describing CDS coordinates provided via the genefinding-gff3 argument. One

predicted BGC region was discarded because it was suspected of corresponding to either extremely recent horizontal gene transfer from bacteria or contamination. Briefly, key biosynthetic proteins were aligned to bacterial RefSeq which highlighted 29 BGC regions as exhibiting ≥80% amino acid identity and ≥80% query coverage to proteins found in bacterial genomes from RefSeq. BGC regions were further assessed using Whokaryote (v1.1.2) to determine if their gene structuring is likely eukaryotic or bacterial, revealing only one BGC region as a potential contaminant that was disregarded from downstream analyses. BGC regions were regarded as (i) NRPS-like if they featured a protocluster categorized as "thioamide-NRP", "NRPS", "NRPS-like", or "NRP-metallophore", (ii) as PKS-like if they featured a protocluster categorized as "hgLE-KS", "PKS-like", "prodigiosin" "T1PKS", "T2PKS", "T3PKS", "transAT-PKS", or "transAT-PKS-like" and (iii) as terpene if they featured a protocluster categorized as "terpene". The BGC-ome and strict NRPS-PKS-ome sizes of fungal genomes were computed similar to how they were for bacterial genomes. HMM profiles from the dbCAN database[132] were searched for in predicted proteomes using PyHMMER[133]. Hits were regarded if they featured an E-value <1e-17 and a coverage >0.45, as recommended for annotation of fungal genomes by the maintainers of dbCAN. The eggNOG-mapper server[141] was used to perform comprehensive functional annotation of proteins belonging to select ortholog groups or hierarchical ortholog groups highlighted by the *Pezizomycotina* comparative genomic or GWAS analyses (*see below for details on those analyses*). Functional descriptions from eggNOG-mapper were manually assessed to infer functional categories.

To predict proteins associated with heterokaryon incompatibility (HI), we first gathered sequences from UniProt[163] and NCBI Protein[164] databases for a set of 47 characterized proteins shown to be involved in self-recognition in one of three ascomycetes species: *Neurospora crassa*, *Podospora anserina*, and *Cryphonectria parasitica*[63,165–168], and investigated their domain architectures. Domains from Pfam-A were annotated along these 47 proteins using PyHMMER[133] with gathering thresholds. Annotated domains were sorted by score and greedily retained if they did not exhibit overlap ≥10% of the alignment length to regions already matching to other domains at higher scores. Domain architectures were finally determined for each protein by sequentially appending them in order while avoiding tandem-repeating instances of the same domain to alleviate the influence of alignment-related technical artifacts (**Extended Data Fig. 9a**). Note, some proteins characterized to participate in HI had no domains predicted.

Profile HMMs for 24 domains found in HI proteins were extracted from Pfam-A using hmmfetch in HMMER3[80] to create a custom database that can be used to annotate HI-associated domain architectures in any fungal proteome. The domains included were: PF06985.15 (HET), PF07217.15 (Het-C), PF14479.10 (HeLo), PF11558.12 (HET-s_218-289), PF05729.16 (NACHT), PF09273.15 (Rubis-subs-bind), PF00856.32 (SET), PF01753.22 (zf-MYND), PF00082.26 (Peptidase_S8), PF04419.18 (SERF-like_N), PF12734.11 (CYSTM), PF07727.18 (RVT_2), PF00400.36 (WD40), PF03477.20 (ATP-cone), PF02867.19 (Ribonuc_red_lgC), PF00317.25 (Ribonuc_red_lgN), PF00931.26 (NB-ARC), PF03881.18 (Fructosamin_kin), PF06244.16 (Ccdc124), PF01734.26 (Patatin), PF00505.23 (HMG_box), PF17043.9 (MAT1-1-2), PF08718.15 (GLTP), and PF13374.10 (TPR_10). This subset of Pfam-A profile HMMs, together with a multi-FASTA of the 47 HI-related proteins, and an overview of their domain architectures are provided on this study's Github repo (https://github.com/Kalan-Lab/Ancestral_BGC_Exansions_Study/tree/main/Fungal_Heterokaryon_Incompatability_Proteins_and_pHMMs). We next used hmmsearch with the profile HMM database and gathering thresholds to comprehensively identify proteins putatively associated with HI in the 312 fungal genomes investigated in this study. Proteins with at least one hit by a domain were extracted and more comprehensively profiled using the full Pfam-A[169] database. Afterwards, we determined domain architectures using an equivalent

approach to what was used for the characterized proteins. Finally, proteins were regarded as HI-associated if their architectures matched the architecture of one of the characterized proteins.

Ultimately, as we detected in the GWAS analysis (*please see following section*), the HET domain was determined as highly enriched in the BGC-rich *Pezizomycotina* clade (**Fig. 4a**; **Extended Data Fig. 9a**). While this domain is only associated with heterokaryon incompatibility to our knowledge[165], we constructed a phylogeny of the domain to further assess whether characterized proteins in HI were aggregated in a specific phylogenetic subclade (**Extended Data Fig. 9b**). We extracted the domain from proteins with it using the Biopython library[170]. Domains extracted from characterized proteins were aligned using MUSCLE align[155] and afterwards HET domain sequences from the 312 fungal genomes were incorporated into the resulting multiple sequence alignment (MSA) using MAFFT --add with the "--keeplength" option[140]. This approach was taken to achieve a higher quality and less sparse alignment than performing comprehensive alignment using MUSCLE super5. The resulting MSA was trimmed using trimal[156] in strict mode prior to applying FastTree 2[157] to infer an approximate maximum-likelihood phylogeny. To assess whether any regions of the phylogeny were highly dissimilar to HET from HI characterized proteins, DIAMOND blastp (v2.0.15) was used to align HET domains extracted from the 312 fungal genomes to characterized HI proteins and the E-value associated with the best alignment (highest bitscore) was determined for each query.

Orthology inference, comparative genomics of Pezizomycotina and biosynthetic potential GWAS in Dikarya: OrthoFinder[81] (v2.5.4) was used to infer ortholog groups between fungal genomes.

Comparative genomics between *Agaricomycetes* and BGC-rich *Pezizomycotina* with the complementary set of all other dikaryon genomes was performed using coarse ortholog groups determined by OrthoFinder and a permutation-based approach. Briefly, 100,000 permutations with label shuffling were used to calculate empirical P-values in the mean copy count of ortholog groups between the two genome sets[171]. Only ortholog groups which were found in >70% of one group but <30% of the other group were investigated. A total number of 124 ortholog groups were identified as significantly different in conservation or copy-count between the two groups (**Supplementary Table S11**).

The genome-wide association study (GWAS) to identify coarse ortholog groups associated with increased biosynthetic potential across the full set of 316 genomes was performed using a linear-mixed model in pyseer[94] (v1.3.11). The BGC-ome size of genomes in Mbp units was treated as a continuous phenotype, an ortholog group presence or absence matrix was provided as the genotypes, and the phylogeny constructed for the kingdom was converted into a kinship matrix using a script from the pyseer package to allow for population structure correction. After running pyseer, the count_patterns.py script within the pyseer package was used to determine an appropriate cutoff, accounting for multiple testing, for P-values adjusted for population structure. A total of 672 ortholog groups were identified as significantly associated with BGC-ome size (**Supplementary Table S12**).

Annotations for both ortholog group sets determined by the dikaryon comparative genomics analysis and GWAS were performed using a combination of approaches (**Supplementary Tables S11-S12**). Because many genes highlighted by these analyses appeared related to carbohydrate utilization or virulence towards plants, we primarily assessed the distribution of best matches to dbCAN / CAZy[88,132] profiles across protein members for each ortholog group. Protein sequences categorized as select ortholog groups were further individually annotated using the eggNOG-mapper webserver[141] with *de novo* searching for Pfam domains requested. The annotation with the largest alignment score was selected as the best annotation for each ortholog group; however, intra-ortholog group variability in annotations could

exist and thus these annotations should be interpreted carefully. In addition to using dbCAN and eggNOG-mapper, additional functional predictions were performed for the set of 124 ortholog groups identified via comparative genomics using sequence alignment to the curated Swiss-Prot database[172], via structural similarity using Phyre2[173] and based on available literature.

Typing of PK synthases, NRP synthetases and key terpene biosynthesis genes: Protein sequences featuring either condensation domains from NRPS-like BGCs or ketoacyl-synthase domains from PKS-like BGCs were extracted and further typed using synthaser[86] (v1.1.22) (**Fig. 3b**). Synthaser results were processed and used to determine the proportion of genomes for fungal taxonomic clades which featured different categories of type I PK synthases (non-reducing, partially reducing, or highly reducing), verify NRP synthetases and identify hybrid NRPS-PKS proteins. The proportion of genomes which featured suspected NRP synthases, possessing BGCs predicted as NRPS-like (antiSMASH types: NRPS, NRPS-like, NRP-metallophore and thioamide-NRP) with a condensation domain but not regarded as an NRPS by synthaser was individually determined as well. Similarly, the proportion of genomes which lacked a synthaser verified PKS but featured a PKS-like BGC with a ketoacyl-synthase domain was also determined. NRPS-PKS and PKS-NRPS hybrid proteins were regarded as the same category.

To similarly assess conservation of key terpene biosynthesis genes, we first identified six coarse ortholog groups from OrthoFinder analysis which were present in greater than 25% of fungal genomes and found as a key terpene biosynthesis protein in at least one genome by antiSMASH. After annotation of all proteins belonging to these six ortholog groups using eggNOG-mapper[141], we identified five common Pfam domains across them: PF03936.20 (Terpene_synth_C), PF00348.21 (polyprenyl_synt), PF00494.23 (SQS_PSY), PF13243.10 (SQHop_cyclase_C) and PF13249.10 (SQHop_cyclase_N). Profile HMMs for these domains were gathered from Pfam-A and systematically searched for in all proteins from across all genomes using PyHMMER[133] with gathering cutoffs applied at the domain resolution. For each terpene associated domain, the total proportion of genomes which feature the domain was reported regardless of whether hits were found within BGC contexts or not.

Phylogenetic investigations of PKS ketoacyl-synthase domains and NRPS condensation domains from zoosporic and early diverging fungi: As described in the previous section, PK synthase and NRP synthetase sequences from NRPS-like and type I PKS-like BGCs featuring either condensation or ketoacyl synthase domains, respectively, were identified and processed through synthaser[86]. Synthaser results were processed and ketoacyl synthase domains were extracted from verified polyketide synthases (inclusive of hybrid NRPS-PKS proteins) belonging to genus-representative genomes from early diverging lineages of fungi (fungi not belonging to *Dikarya*, *Mucoromycota, or Zoopagomycota*). Similarly, condensation domains of non-ribosomal peptide synthetases were also extracted from early diverging fungal genomes (also inclusive of hybrid NRPS-PKS proteins). The two sets of sequences were complemented with PKS ketosynthase domains and NRPS condensation domains from reference fungal BGCs cataloged in MIBiG (v3.1), which are primarily from *Ascomycota* and were similarly annotated for respective domains using synthaser. The protein sequences for the two categories of domains were then aligned using super5 mode in MUSCLE[155], trimmed with trimAl[156] in strict mode, and provided as input to FastTree 2[157] to infer phylogenies. Domain classes for sequences were determined using NaPDoS2[87].

Phylogenetically standardized assessment of genome fluidity and pangenome size associated metrics: While genome fluidity is an informative metric which has been used to understand bacterial pangenomes[174–176], it does not account for the core genome phylogenetic diversity represented within each taxonomic group making comparisons between different groups challenging. We thus developed psaps (https://github.com/raufs/psaps; v1.0.0) to assess informative metrics for pangenome expansion differences between taxonomic groups while accounting for phylogenetic diversity. Coarse ortholog groups determined by OrthoFinder and the core genome phylogeny of the fungal kingdom were provided as input to psaps. Briefly, by default, psaps calculates the summed branch lengths for subclades in the global phylogeny provided, by first pruning the global phylogeny to retain only leaves belonging to the focal subclade. It also computes the total number of distinct ortholog groups, auxiliary ortholog groups and the proportion of ortholog groups which are auxiliary for each subclade. The reason auxiliary ortholog groups are of interest is because core ortholog groups are likely to be present in the last common ancestor of the focal subclade and thus not informative for pangenome expansion assessment. The set of auxiliary ortholog groups per subclade is the total number of ortholog groups minus the core ortholog groups, where, by default, core ortholog groups are those found in at least 80% of genomes.

The psaps program also features an option to perform pairwise analyses between genomes independently for each subclade. This allows for the generation of a distribution of the various statistics for each subclade. In this mode the branch sum equates to simply the distance between pairs of genomes. If pairwise analysis is requested, the genomic fluidity metric[174] and the standardized genomic fluidity metric are also calculated. The genomic fluidity metric for a pair of genomes belonging to the same subclade corresponds to the sum of the number of non-shared ortholog groups observed divided by the sum of the total number of ortholog groups observed in each genome. The genome fluidity is typically calculated for an entire microbial population or species by averaging values across multiple pairs of genomes; however, psaps reports values for individual pairs of genomes to allow for more resolute assessment of the distribution of values. The standardized genomic fluidity simply divides the genomic fluidity metric by the branch distance between the two genomes being assessed. This results in inflating the genomic fluidity metric when pairs of genomes are more similar to each other in the core genome phylogeny and deflating the metric when pairs of genomes are phylogenetically more diverged.

For the investigations of genome fluidity and BGC-ome size in **Fig. 4e** and **Extended Data Fig. 10**, a few genomes which had taxonomic classifications in NCBI that were discordant with expectations based on our or Amses et al. 2022 phylogenomic analyses were disregarded. Genomes removed for these analyses included: GCA_025526915.1 (*Entophlyctis*), GCA_025602905.1 (*Phlyctochytrium*), GCA_027604625.1 (*Blyttiomyces*). These genomes were still included in other analyses since their broader taxonomic classification (e.g. membership in the *Chytridiomycota* phylum) was reliable.

Aspergillus *investigations of the relation between HI counts with BGC-ome and genome sizes*: Representative genomes for 78 *Aspergillus* species with CDS predictions were downloaded from the NCBI Genomes database (**Supplementary Table S10**). They were annotated for BGCs using antiSMASH, CAZymes using PyHMMER with dbCAN profile HMMs, and HI-associated proteins using HMMER with HET and NACHT profile HMMs with analogous parameter configurations used for processing and analysis of the fungi-wide dataset of 316 genomes. A core genome phylogeny for the 78 species was next established using protein alignments of UFCG marker genes and OrthoFinder was applied to infer ortholog groups between the species, also using equivalent settings to what was performed for the fungi-wide dataset. Orthology predictions together with the core genome phylogeny

and visually determined clade groupings for *Aspergillus* species, factoring in whether species exhibited similar counts of HI proteins, were provided as input for psaps to assess the relationship between the core genome phylogenetic breadth and number of auxiliary ortholog groups across different clades.

Inference on horizontal gene transfer in shaping BGC evolution

In addition to vertical evolution through speciation, BGCs and their individual genes can sometimes transfer horizontally between organisms. While HGT is likely to have influenced bacterial evolution for eons, for this study, we primarily focused on assessing and measuring the extend of recent HGT for different taxonomic groups. This is because signatures of recent HGT are more straightforward to detect at scale and confident inference of ancient HGT events often requires manual investigation and computationally costly phylogenetic analysis.

Development and application of codoff for codon-usage-based assessment of recent HGT: We had previously developed an auxiliary program, compareBGCtoGenomeCodonUsage.py, as part of our software suite *Isa*BGC[127] to assess whether BGCs exhibited divergent codon usage profiles to their genomic contexts. Divergent codon usage profiles between focal BGCs and background genomes can indicate recent horizontal gene transfer if codon usage between the donor and recipient organisms vary. However, such signatures of HGT are expected to fade over time as the transferred sequence gradually evolves to assimilate to its new context and increase translational efficiency. In addition, codon usage discordance for a focal region to its genomic context can arise from atypical expression of gene clusters and not involve HGT. This can be the case for gene clusters which are very highly expressed, such as ribosomal operons, or gene clusters which are infrequently expressed under highly specific conditions.

More recently, we have reimplemented compareBGCtoGenomeCodonUsage.py as an independent software package called codoff and made several improvements. Both its predecessor and codoff rely on simulations to infer how the codon usage divergence of the focal BGC (or any gene cluster or genomic region) to its background genome compares to similarly sized regions across the genome. In the most recent version of codoff however, we now perform these simulations by directly sampling regions in the genome rather than creating mock gene clusters composed of random genes from across the genome. This upgrade to preserve genomic structure and organization leads to more realistic simulations and alleviates technical artifacts of the previous approach related to BGC length. The second major update we made in codoff is the reporting of a new metric to assess whether a BGC was unusually discordant in its codon usage to its genomic context profile relative to regions of similar size assessed in simulations. In the current version of codoff, we report a "Discordance Percentile" for each focal BGC investigated with codoff. This metric is calculated as the percentage of simulations where regions of similar size had codon usage profiles as discordant or more discordant than the focal BGC to the background. Lower values for discordance percentiles thus indicate that the focal BGC of interest features an unusually divergent codon usage profile to the genome-wide codon usage profile. Additional details on codoff can be found on its Github repo (https://github.com/Kalan-Lab/codoff) and the associated Wiki - including assessment that simulations work as intended in bacterial and fungal genomes and a large-scale application to investigate codon usage discordance for BGCs across complete *Streptomyces* chromosomes.

We systematically applied codoff (v1.2.3) on BGC predictions for the set of 4,732 genus-representative genomes across 74 bacterial and archaeal orders. The criteria for running codoff to

completion were met by 21,843 of 21,848 bacterial BGCs. Boxplots in Tukey's style were used to visualize distributions of BGC codon usage discordance percentiles for each taxonomic order (**Extended Data Fig. 2ab**). Similarly, we systematically applied codoff on BGCs from the primary set of 312 fungal genomes investigated in this study (**Extended Data Fig. 2d**). Of the 7,215 fungal BGCs assessed, codoff produced results for 7,176 which met the criteria for analysis.

Additional detection of MGEs in bacteria and fungi: The software geNomad[131] was used to annotate temperate phages and plasmids in the bacterial genomes of 17,844 species-representatives from *Actinomycetota*, *Bacillota*, *Cyanobacteriota*, *Myxococcota* and *Pseudomonadota.* Afterwards, coordinates of BGCs were assessed for being located on phages or plasmids (**Extended Data Fig. 2c**). The presence of the tyrosine site-specific recombinase domain (DUF3435, PF11917.12), a key enzyme for the mobilization of Starship MGEs that can horizontally transfer BGCs in fungi[44], was assessed across the 312 fungal genomes using hmmsearch with the gathering threshold. The number of proteins with the domain per genome was reported for individual fungal taxonomic clades (**Extended Data Fig. 2e**).

ADDITIONAL REFERENCES

91. Belknap, K. C., Park, C. J., Barth, B. M. & Andam, C. P. Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. *Sci. Rep.* **10**, 2003 (2020).
92. Doroghazi, J. R. & Metcalf, W. W. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* **14**, 611 (2013).
93. Nett, M., Ikeda, H. & Moore, B. S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat. Prod. Rep.* **26**, 1362–1384 (2009).
94. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. Pyseer: A comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310–4312 (2018).
95. Amorim Franco, T. M. & Blanchard, J. S. Bacterial Branched-Chain Amino Acid Biosynthesis: Structures, Mechanisms, and Drugability. *Biochemistry* **56**, 5849–5865 (2017).
96. Danov, A. *et al.* Toxinome-the bacterial protein toxin database. *MBio* **15**, e0191123 (2024).
97. Steindorff, A. S. *et al.* Comparative genomic analysis of thermophilic fungi reveals convergent evolutionary adaptations and gene losses. *Commun. Biol.* **7**, 1124 (2024).
98. Elliot, M. A., Buttner, M. J. & Nodwell, J. R. Multicellular Development in*Streptomyces*. in *Myxobacteria* 419–438 (ASM Press, Washington, DC, USA, 2014).
99. Bush, M. J., Gallagher, K. A., Chandra, G., Findlay, K. C. & Schlimpert, S. Hyphal compartmentalization and sporulation in Streptomyces require the conserved cell division protein SepX. *Nat. Commun.* **13**, 71 (2022).
100. Bourguignon, T. *et al.* Increased Mutation Rate Is Linked to Genome Reduction in Prokaryotes. *Curr. Biol.* **30**, 3848-3855.e4 (2020).
101. Shende, V. V., Bauman, K. D. & Moore, B. S. The shikimate pathway: gateway to metabolic diversity. *Nat. Prod. Rep.* **41**, 604–648 (2024).
102. Peek, J. *et al.* Rifamycin congeners kanglemycins are active against rifampicin-resistant bacteria via a distinct mechanism. *Nat. Commun.* **9**, 4147 (2018).
103. Kang, Q., Shen, Y. & Bai, L. Biosynthesis of 3,5-AHBA-derived natural products. *Nat. Prod. Rep.* **29**, 243–263 (2012).
104. Salamzade, R. *et al.* Zol and fai: Large-scale targeted detection and evolutionary investigation of

gene clusters. *Nucleic Acids Res.* **53**, (2025).
105. Spanogiannopoulos, P., Waglechner, N., Koteva, K. & Wright, G. D. A rifamycin inactivating phosphotransferase family shared by environmental and pathogenic bacteria. *Proceedings of the National Academy of Sciences* **111**, 7102–7107 (2014).
106. Kim, T. K., Hewavitharana, A. K., Shaw, P. N. & Fuerst, J. A. Discovery of a new source of rifamycin antibiotics in marine sponge actinobacteria by phylogenetic prediction. *Appl. Environ. Microbiol.* **72**, 2118–2125 (2006).
107. Baldauf, S. L. & Doolittle, W. F. Origin and evolution of the slime molds (Mycetozoa). *Proc. Natl. Acad. Sci. U. S. A.* **94**, 12007–12012 (1997).
108. Dembitsky, V. M., Rezanka, T., Spízek, J. & Hanus, L. O. Secondary metabolites of slime molds (myxomycetes). *Phytochemistry* **66**, 747–769 (2005).
109. Butcher, R. A., Fujita, M., Schroeder, F. C. & Clardy, J. Small-molecule pheromones that control dauer development in Caenorhabditis elegans. *Nat. Chem. Biol.* **3**, 420–422 (2007).
110. Lin, Z., Li, F., Krug, P. J. & Schmidt, E. W. The polyketide to fatty acid transition in the evolution of animal lipid metabolism. *Nat. Commun.* **15**, 236 (2024).
111. Kodner, R. B., Summons, R. E., Pearson, A., King, N. & Knoll, A. H. Sterols in a unicellular relative of the metazoans. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9897–9902 (2008).
112. Mao, L. *et al.* Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12472–12480 (2020).
113. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
114. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**, 1–7 (2002).
115. Muñoz-Gómez, S. A. *et al.* An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. *Elife* **8**, (2019).
116. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
117. Mira, A., Pushker, R., Legault, B. A., Moreira, D. & Rodríguez-Valera, F. Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics. *BMC Evol. Biol.* **4**, 50 (2004).
118. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* **1**, 193 (2017).
119. dos Reis, M. *et al.* Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Curr. Biol.* **25**, 2939–2950 (2015).
120. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
121. Magee, A., Karcher, M., Matsen, F. A., 4th & Minin, V. M. How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lens of Monte Carlo error. *Bayesian Anal.* **19**, 565–593 (2024).
122. Patricia, C. B. G. *TreeViewer - Cross-Platform Software to Draw Phylogenetic Trees*. doi:10.5281/zenodo.10044388.
123. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

124. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6**, 90 (2014).
125. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genom.* **3**, e000131 (2017).
126. Klassen, J. L. & Currie, C. R. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* **13**, 14 (2012).
127. Salamzade, R. *et al.* Evolutionary investigations of the biosynthetic diversity in the skin microbiome using lsaBGC. *Microb Genom* **9**, (2023).
128. Blin, K. *Ncbi-Genome-Download*. (Zenodo, 2023). doi:10.5281/ZENODO.8192486.
129. Larralde, M. Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *J. Open Source Softw.* **7**, 4296 (2022).
130. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
131. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).
132. Zheng, J. *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res.* **51**, W115–W121 (2023).
133. Larralde, M. & Zeller, G. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* **39**, (2023).
134. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).
135. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32-6 (2006).
136. Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47**, D687–D692 (2019).
137. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
138. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2019).
139. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **3**, e5 (2007).
140. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
141. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
142. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
143. Emms, D. M. & Kelly, S. STAG: Species Tree Inference from All Genes. *bioRxiv* 267914 (2018) doi:10.1101/267914.
144. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
145. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
146. David, L. A. & Alm, E. J. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011).

147. Csurös, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
148. Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**, 17–25 (2002).
149. Bolotin, E. & Hershberg, R. Gene loss dominates as a source of genetic variation within clonal pathogenic bacterial species. *Genome Biol. Evol.* **7**, 2173–2187 (2015).
150. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596 (2001).
151. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
152. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).
153. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
154. Li, W. *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* **49**, D1020–D1028 (2021).
155. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
156. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
157. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
158. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
159. Walker, G. E., Dunbar, B., Hunter, I. S., Nimmo, H. G. & Coggins, J. R. Evidence for a novel class of microbial 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase in Streptomyces coelicolor A3(2), Streptomyces rimosus and Neurospora crassa. *Microbiology* **142 (Pt 8)**, 1973–1982 (1996).
160. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
161. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
162. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699 (2023).
163. UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
164. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res.* **53**, D20–D29 (2025).
165. Zhao, J. *et al.* Identification of allorecognition loci in Neurospora crassa by genomics and evolutionary approaches. *Mol. Biol. Evol.* **32**, 2417–2432 (2015).
166. Glass, N. L. & Dementhon, K. Non-self recognition and programmed cell death in filamentous fungi. *Curr. Opin. Microbiol.* **9**, 553–558 (2006).
167. Daskalov, A., Heller, J., Herzog, S., Fleißner, A. & Glass, N. L. Molecular mechanisms regulating cell fusion and heterokaryon formation in filamentous fungi. in *The Fungal Kingdom* 215–229 (ASM Press, Washington, DC, USA, 2017).

168. Zhang, D.-X., Spiering, M. J., Dawe, A. L. & Nuss, D. L. Vegetative incompatibility loci with dedicated roles in allorecognition restrict mycovirus transmission in chestnut blight fungus. *Genetics* **197**, 701–714 (2014).
169. Paysan-Lafosse, T. *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res.* **53**, D523–D534 (2025).
170. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
171. Salamzade, R., Swaney, M. H. & Kalan, L. R. Comparative genomic and metagenomic investigations of the Corynebacterium tuberculostearicum species complex reveals potential mechanisms underlying associations to skin health and disease. *Microbiol. Spectr.* e0357822 (2022).
172. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
173. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
174. Kislyuk, A. O., Haegeman, B., Bergman, N. H. & Weitz, J. S. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**, 32 (2011).
175. Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11**, 1719–1721 (2017).
176. Dewar, A. E., Hao, C., Belcher, L. J., Ghoul, M. & West, S. A. Bacterial lifestyle shapes pangenomes. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2320170121 (2024).