AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus *Streptomyces*

Bradon R. McDonald,[a,b] Cameron R. Currie[a,b]

Department of Bacteriology, University of Wisconsin—Madison, Madison, Wisconsin, USA[a]; DOE Great Lakes Bioenergy Research Center, University of Wisconsin—Madison, Madison, Wisconsin, USA[b]

ABSTRACT Lateral gene transfer (LGT) profoundly shapes the evolution of bacterial lineages. LGT across disparate phylogenetic groups and genome content diversity between related organisms suggest a model of bacterial evolution that views LGT as rampant and promiscuous. It has even driven the argument that species concepts and tree-based phylogenetics cannot be applied to bacteria. Here, we show that acquisition and retention of genes through LGT are surprisingly rare in the ubiquitous and biomedically important bacterial genus *Streptomyces*. Using a molecular clock, we estimate that the *Streptomyces* bacteria are ~380 million years old, indicating that this bacterial genus is as ancient as land vertebrates. Calibrating LGT rate to this geologic time span, we find that on average only 10 genes per million years were acquired and subsequently maintained. Over that same time span, *Streptomyces* accumulated thousands of point mutations. By explicitly incorporating evolutionary timescale into our analyses, we provide a dramatically different view on the dynamics of LGT and its impact on bacterial evolution.

IMPORTANCE Tree-based phylogenetics and the use of species as units of diversity lie at the foundation of modern biology. In bacteria, these pillars of evolutionary theory have been called into question due to the observation of thousands of lateral gene transfer (LGT) events within and between lineages. Here, we show that acquisition and retention of genes through LGT are exceedingly rare in the bacterial genus *Streptomyces*, with merely one gene acquired in *Streptomyces* lineages every 100,000 years. These findings stand in contrast to the current assumption of rampant genetic exchange, which has become the dominant hypothesis used to explain bacterial diversity. Our results support a more nuanced understanding of genetic exchange, with LGT impacting evolution over short timescales but playing a significant role over long timescales. Deeper understanding of LGT provides new insight into the evolutionary history of life on Earth, as the vast majority of this history is microbial.

KEYWORDS antibiotics, evolutionary genomics, horizontal gene transfer, molecular clock, species concepts

Received 21 April 2017 **Accepted** 8 May 2017 **Published** 6 June 2017

Citation McDonald BR, Currie CR. 2017. Lateral gene transfer dynamics in the ancient bacterial genus *Streptomyces*. mBio 8:e00644-17. https://doi.org/10.1128/mBio.00644-17.

Editor Paul Keim, Northern Arizona University

Address correspondence to Cameron R. Currie, currie@bact.wisc.edu.

This article is a direct contribution from a Fellow of the American Academy of Microbiology. External solicited reviewers: Christopher Marx, University of Idaho; Patrick Abbot, Vanderbilt University.

The bacterial domain encompasses prodigious diversity generated over billions of years of evolution (1). Despite the critical role that bacteria play in shaping nearly every aspect of life on Earth, understanding the complex evolutionary processes that generate this diversity remains a challenge. In contrast to sexual reproduction in eukaryotes, bacteria were originally thought to undergo strict clonal cell division with little to no genetic exchange. The discovery of conjugative plasmids (2), and later transformation and transduction (3), suggested that genetic exchange plays a role in bacterial evolution. The subsequent linking of nonhomologous lateral gene transfer (LGT) to important bacterial phenotypes, such as antibiotic resistance (4) and virulence (5), resulted in the recognition of LGT as a driving force in bacterial evolution. With the advent of comparative genomics and the identification of significant gene content

differences between related bacteria (6), the prevailing view of rampant exchange of genes across bacteria emerged. Expanding on this view, some have argued that genetic exchange is so rampant that bacterial species do not exist as discrete entities (7) and their evolutionary histories fit a web of life model rather than a tree of life (8–11).

The disruptive impact of genetic exchange largely depends on several factors, including the degree to which barriers to LGT structure the exchanges between distantly related lineages. At larger phylogenetic scales, gene transfers have been shown to be more common within than between phyla, and some phyla exchange genes more frequently than others (12, 13). At the population level, both geographic distribution and sequence dissimilarity can lead to reduced rates of homologous recombination (14, 15). Analyses that span the species level to intermediate, genus-level diversity provide opportunities to investigate the combined effects of LGT and mutation across physiologically similar organisms that are diverging, either due to neutral processes or due to selective pressures in the different ecological niches that they occupy. This has the advantage of providing sufficient diversity to reliably detect LGT and identify trends that occur over evolutionary timescales, such as the loss of acquired genes that are selectively neutral or mildly deleterious (16–18). By including sufficient sampling of closely related organisms, it also reduces the amount of variation due to differences in core physiology between organisms in the data set. Focusing on a single genus that has broadly similar life history traits and is found in diverse environments enables easier detection of rapidly evolving diversity and ecologically relevant variation.

The ubiquitously distributed bacterial genus *Streptomyces* provides an excellent model for intermediate-scale analysis of genetic exchange and mutation. These diverse filamentous bacteria have been isolated from soil, marine, and host-associated environments, with ecological roles ranging from plant biomass degradation (19, 20) to defensive mutualisms with eukaryotes (21). Complex interactions between *Streptomyces* and other organisms are often characterized by the production of natural products (22), which have been mined for drug discovery for decades (23). Here, we utilize phylogenomic analyses combined with molecular clock dating to investigate the temporal scale of *Streptomyces* genomic and phylogenetic diversity, focusing on non-homologous LGT and point mutations.

RESULTS

Based on our pan-genomic analyses of 122 *Streptomyces* genomes, 80 publicly available genomes and 42 additional genomes sequenced for this study, we identified significant phylogenetic and genomic diversity. The 42 strains chosen for genome sequencing were selected to obtain genome coverage across the genus; we selected strains to fill in gaps based on the phylogenetic location of the publicly available genomes in a 16S rRNA gene phylogeny (see Fig. S1 in the supplemental material). Using this expanded genomic data set, multilocus phylogenies were generated using both a traditional multilocus approach based on 94 housekeeping genes (Fig. 1, Fig. S2) and an alternative gene tree consensus-based approach implemented in ASTRAL-II (24) (Fig. S3). Both phylogenies are largely congruent, with two major monophyletic clades of *Streptomyces* containing 88 genomes (here referred to as clade I and clade II) and a number of other *Streptomyces* lineages containing the remaining 34 genomes (Fig. 1; also Fig. S2). These clades did not match the genome distribution on the 16S rRNA gene phylogeny, likely due to the poor resolution of 16S at finer phylogenetic scales. Most *Streptomyces* isolates from marine environments are found in more ancient lineages, suggesting a possible marine origin for the genus. Further, many strains in clade I were isolated from insect-associated niches (Fig. S2). Support values were high across both phylogenies with the exception of internal nodes in clade II, where tree topology differed in poorly supported nodes between the two phylogenetic methods. Exploring gene content, we identified a total of 39,893 gene families across the genus. Of these, 1,048 were conserved in 95% of *Streptomyces* genomes in the data set (Fig. 2). Each of the two major clades contained ~200 gene families that were conserved in the

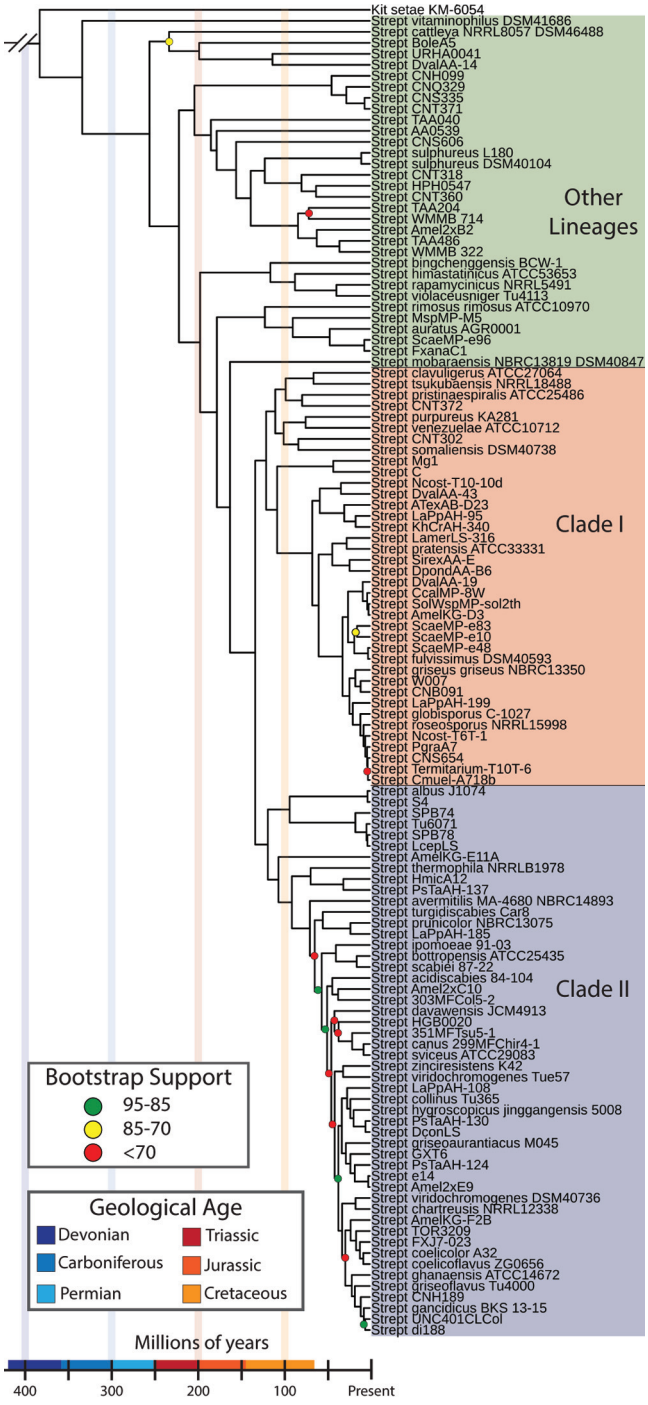

FIG 1 Genome-based phylogeny and molecular clock for the genus *Streptomyces*. TIGRFAM-based multilocus phylogeny of *Streptomyces* using 94 universally conserved housekeeping genes. Branch lengths indicate Reltime-estimated divergence times. Bootstrap values are shown by colored circles on all nodes with values of ≤95. *Streptomyces* and *Kitasatospora* are abbreviated as Strept and Kit, respectively.

respective group but not the others. Each *Streptomyces* clade also contained a large number of unique genes, with 3,558 unique genes in clade I and 8,140 in clade II. Of these, a small percentage (5% in clade I and 2.5% in clade II) were conserved across the group. Shared genome content was correlated with phylogenetic distance, as more closely related genomes shared a higher proportion of genes (Fig. 2B). Overall, the results of our gene content analysis, without incorporating divergence times, are

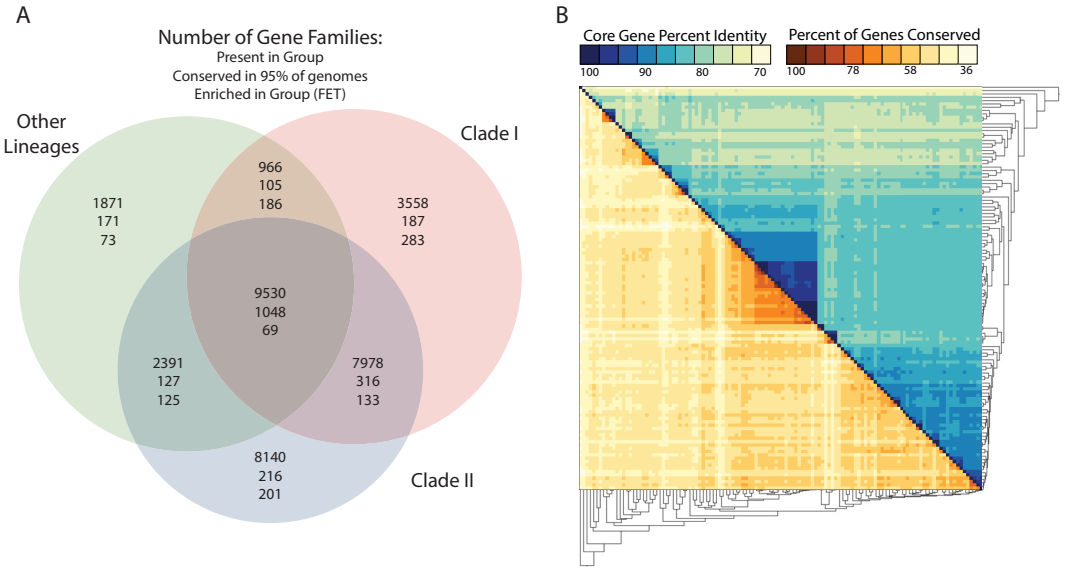

FIG 2 *Streptomyces* genome content conservation and divergence. (A) Proteinortho gene families present, conserved, and enriched in three groups of *Streptomyces* (clade I, clade II, and other lineages; based on Fig. 1). Gene family enrichment within subsets of *Streptomyces* was determined using Fisher's exact test. (B) Pairwise comparison of conserved Proteinortho gene family percentage and TIGRFAM core gene sequence percent identity.

consistent with the prevailing view of bacterial evolution: high gene content diversity driven by frequent LGT.

Since mutation and LGT are dynamic processes that occur through time, identifying the rate of these events is critical for understanding their impact on the long-term evolution of a bacterial lineage. We estimated divergence times across the *Streptomyces* phylogeny using cyanobacterial fossils (25, 26), the estimated origin of life on Earth (27, 28), and the *Escherichia-Salmonella* divergence time (29) as calibration points for relaxed molecular clock analysis (Fig. S4; Table S1). Our analyses inferred that the genus *Streptomyces* diverged from *Kitasatospora* approximately 382 million years ago (mya) (confidence interval [CI], 250 to 514 mya), in the late Devonian period, and the two major clades diverged approximately 132 mya (CI, 87 to 177 mya), in the early to mid Cretaceous period. These divergence times also allow us to approximate the time span required for strains to diverge by 1% amino acid identity. Among 11 pairs of strains separated by 1% amino acid divergence in the core genes used for the phylogeny, the average divergence time is 8 $\pm$ 2.5 my.

Investigation of LGT dynamics in *Streptomyces* revealed both functional and phylogenetic biases in gene transfer events. In total, we inferred 320,263 genes laterally acquired by *Streptomyces* lineages using the gene tree reconciliation approach implemented in AnGST (30). Gene functional classes overrepresented in LGT events consisted of secondary metabolism and xenobiotic metabolism (Table S2). Core biological functions such as transcription and translation were underrepresented (Table S3). Combined with the molecular clock dating, we estimate that overall rates of detectable LGT events per node into clade I or clade II are 5.93 and 9.08 per my, respectively (Fig. 3A; also Table S4). Nodes in clade I and clade II were significantly more likely to receive a gene transferred from a member of their own clade than from another source ($P <$ 1e$-$5, Fisher's exact test), with rates of 3.82 and 7.07 per million years for clades I and II, respectively. The estimated transfer rate per node from clade I to clade II is 1.08 transfers per my, and that from clade II to clade I is 1.43 per my. Inferred transfers of genes from other *Streptomyces* lineages to a genome in one of the major clades are significantly less common, occurring at approximately once every 2 million years. Acquisition of genes from different actinobacterial genera occurred even less frequently, at once every 3 million years for clade II and once every 5 million years for clade I.

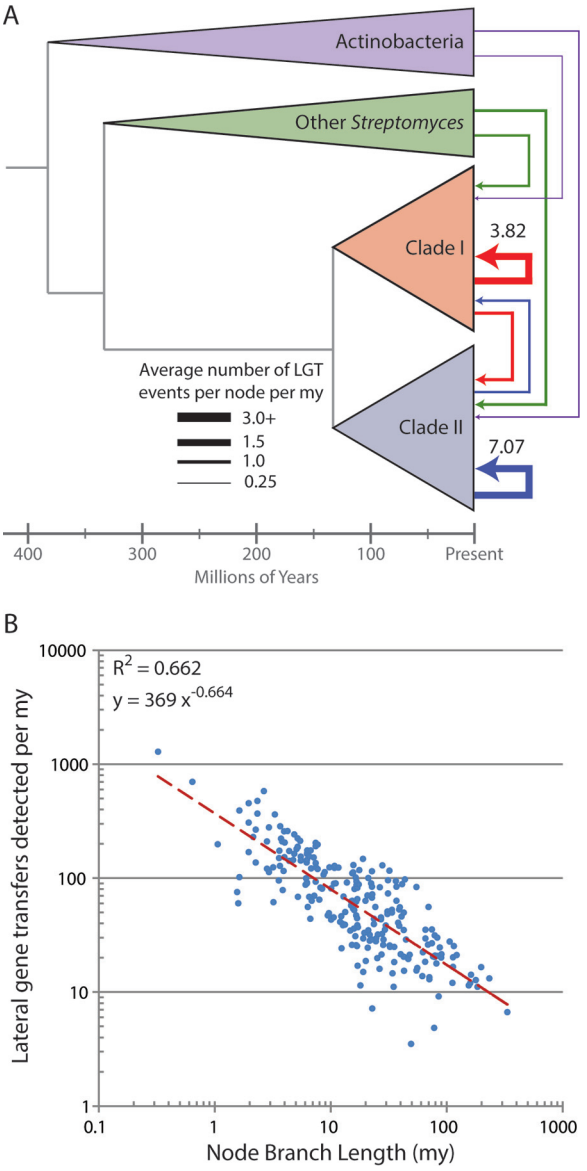

FIG 3 Rate of lateral gene transfer in *Streptomyces*. (A) Average rate of LGT across the *Streptomyces* phylogeny. Line thickness indicates the average number of detected LGT events per genome (including ancestral reconstructions) per million years from each source. Rates greater than 3 are labeled, and all rates appear in Table S2 in the supplemental material. (B) Detected rate of LGT on each branch of the phylogeny. Detected LGT rate is negatively correlated with branch length.

To investigate how the loss of neutral or deleterious genes acquired through LGT impacts estimates of gene transfer rates, we calculated transfer rate versus branch length across the *Streptomyces*. We found that the rate of detectable LGT events per million years is negatively correlated with branch length (Fig. 3B), and the rate can be approximated relative to branch length using a power law function ($\alpha = -0.664$ and $R^2 = 0.662$). KEGG (31, 32) genes with functions involved in replication and repair, translation, cell growth and death, and mobile elements make up a greater proportion of LGT events in short-branch-length nodes than in long-branch-length nodes (two-sided *t* test, *P* values of 0.009, 0.018, 0.019, and 0.027, respectively). This suggests that the lower number of detected LGT events involving core genes is due to stronger selection against transferred core genes, not a reduced number of actual transfers in these categories. It also suggests that mobile elements are gained and lost more rapidly on evolutionary timescales relative to other gene classes, which is consistent with expectations based on their biology.

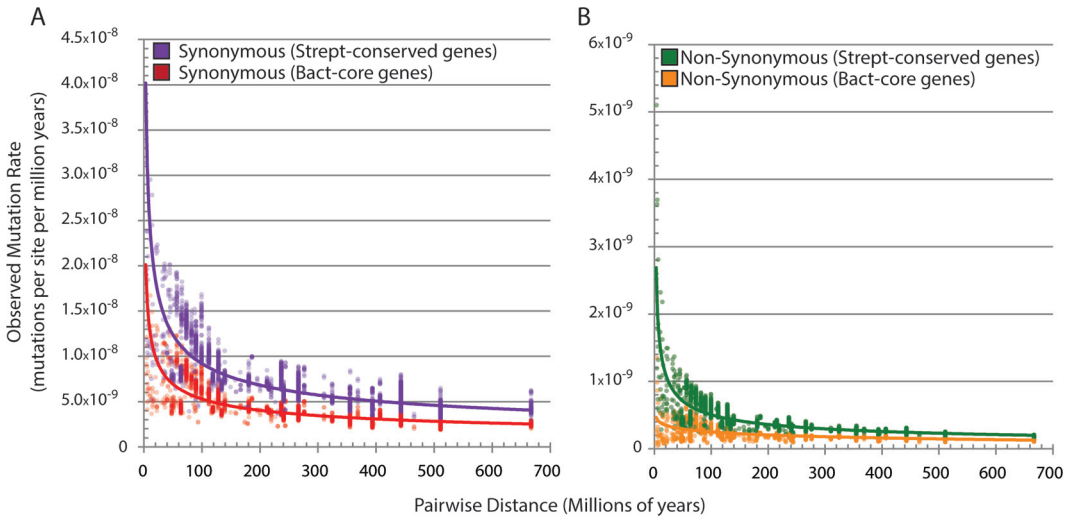

FIG 4 Rate of point mutations in TIGRFAM gene families. Bact-core genes consist of 94 housekeeping TIGRFAM gene families conserved across bacteria. *Strept*-conserved genes consist of 705 TIGRFAM gene families that are found in 95% of our *Streptomyces* genome data set. (A) The observed rate of synonymous point mutations varies by gene data set and pairwise distance. (B) The observed rate of nonsynonymous sites also varies by data set and pairwise distance, but to a lesser degree than that of synonymous sites.

We also investigated the relative contribution of point mutation versus LGT to *Streptomyces* diversity over time by calculating the rate of synonymous and nonsynonymous point mutations per million years in two different sets of conserved TIGRFAM (33) gene families: 705 families conserved across all *Streptomyces* (*Strept*-conserved) and the 94 universally conserved genes used to generate our phylogenies (Bact-core). Observed mutation rates differed between the two sets of genes that we analyzed, particularly for synonymous mutations (Fig. 4). The synonymous mutation rate was 1.5- to 2-fold higher in *Strept*-conserved genes than in the Bact-core genes. Similarly to the LGT rate, observed rates of both synonymous and nonsynonymous mutations are also influenced by the evolutionary distance between genome pairs; the synonymous mutation rate appears much higher in closely related genome pairs, while the difference in observed nonsynonymous mutation rates is less dramatic but still apparent. Using only pairwise comparisons of genomes separated by less than 100 my from the *Strept*-conserved data set, the estimated median rate of synonymous mutations is 1.62×10^{-8} per site per year and the median nonsynonymous mutation rate is 1.78×10^{-9} per site per year. Extrapolating the ratio of synonymous and nonsynonymous sites in the gene sequences that we analyzed to total coding sequence length, we estimate that a total of 13,714 synonymous and 10,429 nonsynonymous mutations accumulate in *Streptomyces* lineages per million years.

Given that natural product biosynthetic gene clusters (BGCs) were overrepresented in LGT events, and the known role that some of these play in producing small molecules that shape ecological interactions that are predicted to be under selection (34, 35), we examined the distribution and exchange of these secondary-metabolite-producing pathways. We identified a total of 4,945 natural product BGCs in *Streptomyces*; 1,759 clusters could be classified into 405 BGC families based on Pfam (36) domain content, while the domain structures of the other 3,186 were too dissimilar to match another *Streptomyces* cluster. We found that nearly all BGC families were nonrandomly distributed in *Streptomyces*, based on a branch-length permutation test: across the genus, 82.7% of BGC families were more phylogenetically restricted than expected by chance. This pattern also holds true at finer phylogenetic scales, as 79.2% and 76.1% of BGC families were more phylogenetically restricted than expected by chance within clade I and clade II, respectively. Interestingly, we also found very few cases of transfer and subsequent maintenance of complete BGC operons (Fig. 5). Analysis of BGC genes affected by LGT suggests that, at least over long evolutionary timescales, the vast

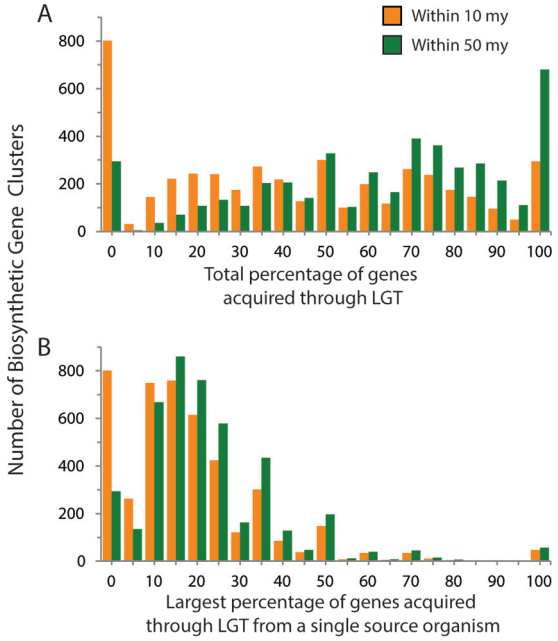

FIG 5 Sources of laterally transferred secondary metabolite biosynthesis genes. (A) Most biosynthetic gene clusters have acquired some genes through LGT within the last 50 my. (B) The vast majority of clusters appear to be a mix of genes from multiple sources, including being retained over millions of years within a lineage through vertical inheritance.

majority of BGCs appear to have been affected by LGT. Specifically, our findings imply that 93% of BGCs acquired at least one gene through LGT within the last 50 my. However, only 57 BGCs had been acquired intact from one source, while the other BGCs were composed of a mixture of genes from multiple sources, including vertically inherited genes. Of the BGCs entirely acquired from a single source, most have been acquired within the last 10 my.

DISCUSSION

Using *Streptomyces* divergence times rather than sequence similarity provides a new perspective on the rate of LGT and its potential impact on bacterial evolution over small and large evolutionary timescales. Although we inferred over 300,000 gene transfer events across the *Streptomyces*, a single successful gene transfer every hundred thousand years per lineage is sufficient to generate the observed number of events given the huge time span encompassed by *Streptomyces* evolution. Further, gene transfers from distantly related *Streptomyces* or other *Actinobacteria* are orders of magnitude less frequent than transfers from closely related lineages, suggesting that distantly related bacterial lineages are typically genetically isolated from each other for tens to hundreds of thousands of years at a time. We also find a strong effect of evolutionary distance between sampled genomes on the inferred rate of LGT. Our estimated transfer rate decreases steadily with branch length according to a two-thirds power law, likely due to acquisition and subsequent loss of genes with neutral or deleterious fitness effects (16). This suggests our calculated average LGT rate primarily measures acquisition rates for genes that are retained over evolutionary time. A similar correlation between evolutionary distance and inferred LGT rate was also found in *Pseudomonas syringae*, using percent amino acid divergence as the measure of distance rather than time (37). Overall, these results stand in contrast to the prevailing view that successful LGT events are rampant over short evolutionary time spans. The inferred rate of LGT accumulation may be impacted by taxon sampling, accurate species tree reconstruction, and LGT detection methods. Also, our analysis focused solely on the actinobacterial genus *Streptomyces*; further research is needed to determine how LGT dynamics vary over

time in different lineages of bacteria. Nevertheless, given that our estimates of genomic diversity and total LGT events, using traditional pan-genome approaches, identify genome content patterns consistent with previous work across bacteria (e.g., ~300,000 LGT events), *Streptomyces* bacteria are unlikely to be outliers with respect to LGT rates.

We also show that approximately 23,000 point mutations accumulate per million years. Although each point mutation changes only a single base while an LGT event may involve the addition of thousands of base pairs, the phenotypic effect of any one mutation or LGT event is difficult to predict. A single point mutation in an active site may completely change an enzyme's function, while an acquired gene that is not expressed may have no phenotypic effect. Our data suggest that while positively selected LGT events provide a rare but important source of genetic diversity, the vast majority of events that generate sequence diversity in *Streptomyces* are likely point mutations. Similar to the evolutionary distance effect on inferred LGT rate, pairwise comparisons between more distantly related organisms lead to lower estimated mutation rates, although this effect is weaker for point mutations than LGT events. As a result, the neutral rate of LGT is more difficult to infer from older lineages than the neutral point mutation rate. Estimated point mutation rates are also lower in genes that are conserved across most bacteria than in genes conserved in all *Streptomyces* species. The effects of evolutionary distance and gene data set on inferred mutation rate are stronger in synonymous sites, suggesting widespread fitness effects of synonymous mutations in *Streptomyces* core genes, as has been shown in a variety of bacteria and eukaryotes (38–41). These results suggest that comparisons of evolutionary event rates between bacterial groups can be strengthened through normalizing rates based on evolutionary distance between samples within each group.

Genes involved in the biosynthesis of natural products, the compounds that are critical for modern medicine and for which *Streptomyces* is well known, are among the genes most frequently acquired through LGT. In contrast to previous work in the related actinobacterial genus *Salinispora* (35), we found that most biosynthesis clusters were composed of genes apparently acquired from multiple sources rather than a single full-operon transfer event. These differences could be explained by the *Salinispora* data set being comprised of much more closely related genomes than our *Streptomyces* data set. Many BGC transfer events in *Streptomyces* might in fact be transfers of full clusters, followed by gene shuffling with other clusters in the same genome over evolutionary time. This would result in the scattered distribution of transferred genes that we observed in *Streptomyces* BGCs, without requiring different transfer dynamics than those seen in *Salinispora*.

Applying the temporal framework to *Streptomyces*, our results show that the biomedically and ecologically important genus *Streptomyces* is truly ancient; *Streptomyces* bacteria as a group are approximately as old as tetrapods and 60 my older than seed plants. The two major *Streptomyces* clades are approximately as old as flowering plants and older than the divergence of *Salmonella* and *Escherichia*. Therefore, despite being taxonomically defined as a genus, bacteria within lineages like the *Streptomyces* should not be considered "closely related." Our analysis provides insight into intermediate-scale LGT dynamics, but our phylogenetic sampling does not enable us to investigate population-scale dynamics. Investigating processes that occur at this scale, including the short-term impacts of population size and drift on fixation and retention of genes acquired through LGT, will require intensive sampling of much more closely related bacterial strains, i.e., 99%-plus average nucleotide identity.

Our molecular clock analysis is consistent with several other molecular clock analyses performed across all bacteria (42) and *Actinomycetes* (43), using different methods and data sets. The majority of comparable nodes in our molecular clock analysis fall within the credibility intervals of previous rigorous analysis in bacteria performed by Battistuzzi et al. (42) (Table S1). Older nodes have larger confidence intervals and are more variable. This variability has limited effect on our analyses of *Streptomyces*, because only nodes that recently diverged, i.e., less than 400 my ago, were employed as reference points for the *Streptomyces* molecular clock that was used for downstream

analyses. Although our molecular clock analysis is consistent with previous work, uncertainty in molecular clock dating generally means that the absolute rates of LGT that we identified are estimates. However, the extremely low rate of LGT means that our general conclusions remain valid even if the *Streptomyces* lineage is significantly younger than our analysis indicates. Further, the relative number of LGT events versus point mutations is robust to uncertainty in the molecular clock, since the two rates are calculated using the same divergence times.

Our results provide new insight into the paradox that, despite widespread LGT events, bacteria seemingly form "natural groups with coherent properties" (37, 44). The argument that LGT leads to significant divergence from the paradigm of vertical transmission of DNA from parent to daughter cell in bacteria emerged from a combination of high pan-genomic diversity and examples of gene acquisitions from distantly related organisms (45). However, we find strong evidence of phylogenetic biases in *Streptomyces* LGT, even in the most frequently transferred gene classes such as secondary metabolism. We also find evidence that many transferred genes may be selectively neutral or deleterious, leading to rapid turnover of acquired genes. An evolutionary model incorporating high neutral turnover of genes acquired through LGT is also supported by a positive correlation between genome fluidity and effective population size; bacterial lineages with higher synonymous nucleotide diversity also exhibit higher gene content diversity (46). These results suggest that LGT dynamics can be highly structured by phylogenetic (47) and physiological (48, 49) factors even within a genus, which limit its disruptive effect on bacterial evolution. Our incorporation of an absolute timescale reveals that the actual rate of successful transfer is many orders of magnitude lower than estimated bacterial generation times in nature (50, 51), and we show that thousands of point mutations may accumulate over that same timescale. Viewed from this perspective, the high absolute number of transfers detected in large bacterial genomic data sets may not be the result of high transfer rates but of long evolutionary time spans separating sampled strains. Because even bacteria in the same genus can be separated by tens to hundreds of millions of years, placing LGT in a temporal context also potentially explains the variable gene content patterns observed among closely related genomes without the need to invoke rampant LGT over short evolutionary time spans. Together, our results support a model of rare positively selected LGT events over millions of years driving gene content diversity in bacteria, with vertically inherited point mutations and homologous recombination dominating bacterial evolution over short evolutionary time spans and finer phylogenetic levels.

MATERIALS AND METHODS

Genome annotation. In order to ensure consistent annotations across all genomes, protein-coding genes were predicted *de novo* using Prodigal (52). These were annotated using whole-protein HMMer3 (53) models generated from KEGG (31, 32) database gene families and TIGRFAM 13.0. They were also annotated using domain HMMer models from Pfam 27.0 and antiSMASH 2.0 (54). The TIGRFAM noise score cutoff and antiSMASH score cutoffs were used to remove false-positive hits, while KEGG hits were removed as false positives if their E value was greater than 1e–5, or if the difference in length between the HMMer model consensus and the protein was greater than 50%. rRNA genes were identified by performing BLAST v2.2.25 (55) analyses using sequences from the SILVA database (56). Proteins were also classified into families *de novo* based on sequence homology using Proteinortho v2 (57) with default parameters.

Genome-based phylogenetics. The *Streptomyces/Actinobacteria* multilocus phylogeny was generated using TIGRFAM annotated proteins. The 94 full TIGRFAM proteins in the "core bacterial protein" set (GenProp0799) were used as the molecular data set. The protein sequences with the top HMMer bitscore for each protein family in each genome were aligned using MAFFT (58). These protein alignments were then converted to codon alignments and were concatenated. Recombinant regions were identified using BratNEXTGEN (59) and masked to remove the potential confounding influence of homologous recombination on species tree topology. This removal did not have a significant influence on the final topology. RAxML-7.2.6 (60) was used to generate the phylogeny using the GTRgamma substitution model and 100 rapid bootstraps on the final, recombination-free alignment. The phylogeny of all bacteria used for molecular clock calibration was generated using a similar workflow, except that protein sequences were used to generate the phylogeny using the PROTGAMMABLOSUM62 substitution model in RAxML. Since genomes in this phylogeny were generally very distantly related to each other, no correction for homologous recombination was performed. The gene tree-based phylogeny was generated using ASTRAL-II. One hundred bootstrap alignments were generated for each of the core TIGRFAM families

using RAxML. Phylogenies for each of these alignments were generated with FastTree 2.0 (61). These phylogenies were then used as the input data for ASTRAL-II.

16S rRNA gene phylogeny. *Streptomyces* 16S rRNA gene sequences were obtained from RefSeq (62), along with 16S sequences extracted from the genomic data set and three cyanobacterial 16S sequences that were used as outgroups. These were aligned using MAFFT and then hand curated and trimmed to remove low-quality 16S sequences. The curated set of sequences was realigned and used to generate a phylogeny with FastTree.

Molecular clock analyses. Reltime (63) was used to approximate divergence times for two different phylogenies: the bacterial tree of life and the actinobacterial phylogeny containing the full set of *Streptomyces* genomes. The all-bacterium phylogeny and protein alignment described above were used as the input for Reltime. The algorithm was set to use "Many Clocks" and gamma-distributed rates with invariant sites. Approximate time intervals for the evolution of *Cyanobacteria* (2,500 to 3,500 million years ago) (25, 26), the divergence of *Salmonella* and *Escherichia* (50 to 150 million years ago) (29), and the origin of bacteria (3,500 to 3,800 million years ago) (27, 28) were used to calibrate the molecular clock found in Fig. S4 in the supplemental material. Using a single calibration point can correctly infer the divergence dates of the others with less than 20% error. The confidence intervals for the origins of *Actinobacteria* and *Streptomyces* and the divergence of *Streptomyces* clade I and clade II were then used to calibrate a second molecular clock analysis of the *Streptomyces*/*Actinobacteria* phylogeny found in Fig. 1 and S2.

Lateral gene transfer analysis. For each protein or domain database, the protein sequences for all gene families with more than 3 genes were aligned using MAFFT and then converted to codon alignments. Ten bootstrapped alignments were generated using RAxML for each gene family, and FastTree was used to generate a phylogeny for each bootstrapped alignment. These 10 bootstrap trees were then used as the gene tree inputs for each gene family in AnGST. Reconciliation event costs used were selected based on the criteria suggested in the original AnGST publication (30). Their weights for LGT and gene loss events were selected to minimize the mean change in genome size between parent and daughter nodes in the phylogeny. We analyzed LGT/loss weights of 10:1, 8:1, 5:1, and 3:1. Mean genome size changes for these weights were 4,344.12, 3,133.85, 1,159.95, and 702.71, respectively. Therefore, we used the 3:1 weights for the AnGST parameter settings. The *Streptomyces*/*Actinobacteria* molecular clock tree was provided as the species tree, and AnGST was run in ultrametric mode to avoid biologically improbable LGT events from extant genomes to deep ancestral nodes. The rate of LGT between or within subsets of *Streptomyces* was calculated as the number of genes acquired by genomes in the analyzed clade divided by the age of the last common ancestor of the clade in millions of years. LGT events affecting secondary metabolite clusters that occurred within a given time interval were identified by identifying genes that first appeared in a *Streptomyces* lineage within that interval. There is not a statistically significant difference in the inferred percentage of genes transferred into draft genomes versus complete genomes ($P = 0.21$, Mann-Whitney U), indicating that using predominantly draft genomes does not generate significant detection bias in our LGT analysis. The percentage of inferred gene losses is somewhat higher and more variable in draft genomes, a mean of 15.30% $\pm$ 8.91%, versus 10.41% $\pm$ 4.81% for complete genomes ($P = 0.06$, Mann-Whitney U). This is likely an artifact of variable-quality draft genome assemblies.

Mutation rate estimation. The concatenated codon alignment used to generate the *Streptomyces* phylogeny, along with concatenated codon alignments of all genes conserved in 95% of the *Streptomyces* genomes, was used to calculate point mutation rates. The number of synonymous and nonsynonymous mutations and sites was identified by the codeML package in PAML (64), and molecular clock divergence dates were used to calculate mutation rates per million years. We estimated the total number of synonymous and nonsynonymous sites across all coding regions by multiplying the number of sites in the TIGRFAM protein coding sequence by the average total length of protein coding sequence. Total number of mutations per million years was calculated based on the mutation rate and the estimated total number of sites per genome.

Natural product biosynthesis cluster families. Natural product biosynthetic gene clusters were predicted using the ClusterFinder algorithm and Pfam annotations (65). Genes that were found on the end of predicted clusters and had less than an 0.8 probability of being part of a cluster were removed. After this trimming, clusters with fewer than 3 genes and clusters lacking genes with an antiSMASH domain hit were removed. These curated clusters then were grouped into families using the modified Lin similarity metric (65, 66) with a Jaccard weight of 0.36, GK-gamma weight of 0.64, and overall similarity threshold of 0.7. We then processed the matches with the MCL algorithm, which was run with default parameters, to generate final cluster families. We used a subtree permutation approach to identify cluster families that were phylogenetically restricted, as in the work of Cafaro et al. (67). For each cluster family, we generated subtrees from the multilocus phylogeny containing only the genomes that possessed the cluster. We then compared the total branch length of this subtree to the branch-length distribution of 1,000 subtrees containing the same number of taxa, randomly sampled from the multilocus phylogeny. Cluster families were identified as phylogenetically restricted if their subtree total branch length was significantly less than the distribution mean by two-sided *t* test, with a *P* value cutoff of 1e–5.

Accession number(s). Genome sequence data are available from NCBI (http://www.ncbi.nlm.nih.gov). Accession numbers for all genomes are listed in Table S5.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.00644-17.

FIG S1, TIF file, 0.7 MB.
FIG S2, TIF file, 1 MB.
FIG S3, TIF file, 2.3 MB.
FIG S4, TIF file, 0.9 MB.
TABLE S1, DOCX file, 0.04 MB.
TABLE S2, DOCX file, 0.04 MB.
TABLE S3, DOCX file, 0.05 MB.
TABLE S4, DOCX file, 0.03 MB.
TABLE S5, DOCX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank F. O. Aylward, G. Suen, J. L. Klassen, G. R. Lewin, and A. J. Book for discussions and comments on this work.

B.R.M. was supported by the National Institutes of Health (National Research Service award T32 GM007215) and the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494). Funding for C.R.C. was provided by the National Science Foundation (DEB-0747002 and MCB-0702025) and the National Institutes of Health (U19 AI109673). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

We declare no competing financial interests.

B.R.M. and C.R.C. designed analyses and wrote the manuscript. B.R.M. performed analyses.

REFERENCES

1. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling AE, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437. https://doi.org/10.1038/nature12352.
2. Lederberg J, Tatum EL. 1946. Gene recombination in *Escherichia coli*. Nature 158:558. https://doi.org/10.1038/158558a0.
3. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721. https://doi.org/10.1038/nrmicro1234.
4. Davies J. 1994. Inactivation of antibiotics and the dissemination of resistance genes. Science 264:375–382. https://doi.org/10.1126/science.8153624.
5. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304. https://doi.org/10.1038/35012500.
6. Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HLT, Donnenberg MS, Blattner FR. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A 99:17020–17024. https://doi.org/10.1073/pnas.252529799.
7. Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau MER, Nesbø CL, Case RJ, Doolittle WF. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 37:283–328. https://doi.org/10.1146/annurev.genet.37.050503.084247.
8. Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226–2238. https://doi.org/10.1093/oxfordjournals.molbev.a004046.
9. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482. https://doi.org/10.1038/nrg3962.
10. Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol 3:e316. https://doi.org/10.1371/journal.pbio.0030316.
11. Retchless AC, Lawrence JG. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci U S A 107:11453–11458. https://doi.org/10.1073/pnas.1001291107.
12. Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A 102:14332–14337. https://doi.org/10.1073/pnas.0504068102.
13. Andam CP, Gogarten JP. 2011. Biased gene transfer in microbial evolution. Nat Rev Microbiol 9:543–555. https://doi.org/10.1038/nrmicro2593.
14. Whitaker RJ, Grogan DW, Taylor JW. 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science 301:976–978. https://doi.org/10.1126/science.1086909.
15. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ. 2012. Patterns of gene flow define species of thermophilic Archaea. PLoS Biol 10:e1001265. https://doi.org/10.1371/journal.pbio.1001265.
16. Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3:e130. https://doi.org/10.1371/journal.pbio.0030130.
17. Kuo CH, Ochman H. 2009. The fate of new bacterial genes. FEMS Microbiol Rev 33:38–43. https://doi.org/10.1111/j.1574-6976.2008.00140.x.
18. Van Passel MWJ, Marri PR, Ochman H. 2008. The emergence and fate of horizontally acquired genes in *Escherichia coli*. PLoS Comput Biol 4:e1000059. https://doi.org/10.1371/journal.pcbi.1000059.
19. Book AJ, Lewin GR, McDonald BR, Takasuka TE, Wendt-Pienkowski E, Doering DT, Suh S, Raffa KF, Fox BG, Currie CR. 2016. Evolution of high cellulolytic activity in symbiotic *Streptomyces* through selection of expanded gene content and coordinated gene expression. PLoS Biol 14:e1002475. https://doi.org/10.1371/journal.pbio.1002475.
20. Chater KF, Biró S, Lee KJ, Palmer T, Schrempf H. 2010. The complex extracellular biology of *Streptomyces*. FEMS Microbiol Rev 34:171–198. https://doi.org/10.1111/j.1574-6976.2009.00206.x.
21. Kroiss J, Kaltenpoth M, Schneider B, Schwinger MG, Hertweck C, Maddula RK, Strohm E, Svatos A. 2010. Symbiotic streptomycetes provide antibiotic combination prophylaxis for wasp offspring. Nat Chem Biol 6:261–263. https://doi.org/10.1038/nchembio.331.
22. Kelsic ED, Zhao J, Vetsigian K, Kishony R. 2015. Counteraction of antibiotic production and degradation stabilizes microbial communities. Nature 521:516–519. https://doi.org/10.1038/nature14485.
23. Hopwood DA. 2007. *Streptomyces* in nature and medicine. Oxford University Press, New York, NY.
24. Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52. https://doi.org/10.1093/bioinformatics/btv234.

25. Brocks JJ, Logan GA, Buick R, Summons RE. 1999. Archean molecular fossils and the early rise of eukaryotes. Science 285:1033–1036. https://doi.org/10.1126/science.285.5430.1033.
26. Garvin J, Buick R, Anbar AD, Arnold GL, Kaufman AJ. 2009. Isotopic evidence for an aerobic nitrogen cycle in the latest Archean. Science 323:1045–1048. https://doi.org/10.1126/science.1165675.
27. Mojzsis SJ, Arrhenius G, McKeegan KD, Harrison TM, Nutman AP, Friend CR. 1996. Evidence for life on Earth before 3,800 million years ago. Nature 384:55–59. https://doi.org/10.1038/384055a0.
28. Rosing MT. 1999. 13C-depleted carbon microparticles in 3700-Ma seafloor sedimentary rocks from west Greenland. Science 283:674–676.
29. Ochman H, Wilson AC. 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol 26:74–86. https://doi.org/10.1007/BF02111283.
30. David LA, Alm EJ. 2011. Rapid evolutionary innovation during an Archaean genetic expansion. Nature 469:93–96. https://doi.org/10.1038/nature09649.
31. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182–W185. https://doi.org/10.1093/nar/gkm321.
32. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42:D199–D205. https://doi.org/10.1093/nar/gkt1076.
33. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2013. TIGRFAMs and genome properties in 2013. Nucleic Acids Res 41:D387–D395. https://doi.org/10.1093/nar/gks1234.
34. Sit CS, Ruzzini AC, Van Arnam EB, Ramadhar TR, Currie CR, Clardy J. 2015. Variable genetic architectures produce virtually identical molecules in bacterial symbionts of fungus-growing ants. Proc Natl Acad Sci USA 112:13150–13154. https://doi.org/10.1073/pnas.1515348112.
35. Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR. 2014. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. Proc Natl Acad Sci U S A 111:E1130–E1139. https://doi.org/10.1073/pnas.1324161111.
36. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. https://doi.org/10.1093/nar/gkt1223.
37. Nowell RW, Green S, Laue BE, Sharp PM. 2014. The extent of genome flux and its role in the differentiation of bacterial lineages. Genome Biol Evol 6:1514–1529. https://doi.org/10.1093/gbe/evu123.
38. Agashe D, Sane M, Phalnikar K, Diwan GD, Habibullah A, Martinez-Gomez NC, Sahasrabuddhe V, Polachek W, Wang J, Chubiz LM, Marx CJ. 2016. Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium. Mol Biol Evol 33:1542–1553. https://doi.org/10.1093/molbev/msw035.
39. Knöppel A, Näsvall J, Andersson DI. 2016. Compensating the fitness costs of synonymous mutations. Mol Biol Evol 33:1461–1477. https://doi.org/10.1093/molbev/msw028.
40. Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136:927–935.
41. Bailey SF, Hinz A, Kassen R. 2014. Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. Nat Commun 5:4076. https://doi.org/10.1038/ncomms5076.
42. Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol Biol 4:44. https://doi.org/10.1186/1471-2148-4-44.
43. Embley TM, Stackebrandt E. 1994. The molecular phylogeny and systematics of the Actinomycetes. Annu Rev Microbiol 48:257–289. https://doi.org/10.1146/annurev.mi.48.100194.001353.
44. Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6:431–440. https://doi.org/10.1038/nrmicro1872.
45. Cordero OX, Hogeweg P. 2009. The impact of long-distance horizontal gene transfer on prokaryotic genome size. Proc Natl Acad Sci U S A 106:21748–21753. https://doi.org/10.1073/pnas.0907584106.
46. Andreani NA, Hesse E, Vos M. 2017. Prokaryote genome fluidity is dependent on effective population size. ISME J https://doi.org/10.1038/ismej.2017.36.
47. Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol 28:1057–1074. https://doi.org/10.1093/molbev/msq297.
48. Lercher MJ, Pál C. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 25:559–567. https://doi.org/10.1093/molbev/msm283.
49. Taoka M, Yamauchi Y, Shinkawa T, Kaji H, Motohashi W, Nakayama H, Takahashi N, Isobe T. 2004. Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. Mol Cell Proteomics 3:780–787. https://doi.org/10.1074/mcp.M400030-MCP200.
50. Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A 95:9413–9417. https://doi.org/10.1073/pnas.95.16.9413.
51. Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. Proc Natl Acad Sci U S A 96:12638–12643. https://doi.org/10.1073/pnas.96.22.12638.
52. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
53. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
54. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, Weber T. 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res 41:W204–W212. https://doi.org/10.1093/nar/gkt449.
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
56. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219.
57. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124. https://doi.org/10.1186/1471-2105-12-124.
58. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
59. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 40:e6. https://doi.org/10.1093/nar/gkr928.
60. Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690. https://doi.org/10.1093/bioinformatics/btl446.
61. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
62. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 42:D553–D559. https://doi.org/10.1093/nar/gkt1274.
63. Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. Proc Natl Acad Sci U S A 109:19333–19338. https://doi.org/10.1073/pnas.1213199109.
64. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088.
65. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158:412–421. https://doi.org/10.1016/j.cell.2014.06.034.
66. Lin K, Zhu L, Zhang DY. 2006. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics 22:2081–2086. https://doi.org/10.1093/bioinformatics/btl366.
67. Cafaro MJ, Poulsen M, Little AEF, Price SL, Gerardo NM, Wong B, Stuart AE, Larget B, Abbot P, Currie CR. 2011. Specificity in the symbiotic association between fungus-growing ants and protective *Pseudonocardia* bacteria. Proc Biol Sci 278:1814–1822. https://doi.org/10.1098/rspb.2010.2118.