*lsa*BGC provides a comprehensive framework for evolutionary analysis of biosynthetic gene clusters within focal taxa

Rauf Salamzade[1], J.Z. Alex Cheong[1], Shelby Sandstrom[1], Mary Hannah Swaney[1], Nicole Lane Starr[1], Anne Marie Singh[2], and Lindsay R. Kalan[1,3]

[1]Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA
[2]Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA
[3]Department of Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA

Address for Correspondence: Lindsay Kalan, 1550 Linden Dr, 6155 Microbial Sciences Building, Madison, WI, USA, 53706. lkalan@wisc.edu

Abstract:

Bacterial secondary metabolites, encoded by biosynthetic gene clusters (BGCs), can underlie microbiome homeostasis and be developed into high-value commercialized products, which have historically been mined from a select group of taxa. Despite this, dedicated bioinformatics tools for comparative analysis of BGCs are limited. We have thus developed *lsa*BGC to aid exploration of microdiversity and evolutionary trends across BGCs in any bacterial taxa of interest through calculation of population genetic statistics and metagenomic exploration. As a proof of concept, we applied *lsa*BGC to four genera commonly found in skin microbiomes to reveal novel insights on both well and less studied taxa and their secondary metabolites. We find evidence of horizontal gene transfer (HGT) within multiple genera. The BGC encoding for the carotenoid staphyloxanthin, a virulence factor in *S. aureus*, is ubiquitous across the genus of *Staphylococcus* and exhibits signatures of inter-species HGT, including mobilization onto plasmids. We also identified a non-ribosomal peptide synthetase (NRPS) BGC flanked by transposable elements, encoding for a predicted siderophore, which is highly conserved across several divergent *Corynebacterium* species commonly isolated from skin such as *C. kefirresidentii*. From strain-resolution metagenomics, we further determined that the *C. kefirresidentii* species complex is the most prevalent clade of *Corynebacterium* in skin metagenomes and is found in the skin microbiome of every participant surveyed. While ubiquitous across skin, the number of publicly-available genomes for this species complex is currently limited. Using *lsa*BGC's framework for metagenomic mining, we identified 34,545 novel single nucleotide variants (SNVs) in predicted BGCs of the species complex, including several novel, non-synonymous SNVs within highly conserved sites of enzymes. Ultimately, we envision that application of *lsa*BGC can highlight important, yet poorly studied, secondary metabolites as well as aid in the discovery and design of more efficacious commercialized metabolites.

Introduction:

The secondary metabolome of bacteria has served as a valuable reservoir of natural products with great societal benefit[1]. Historically, many commercialized secondary metabolites, such as antibiotics, have been identified from particular taxonomic groups, such as the genus of *Streptomyces* within Actinomycetota[2,3]. Several studies have explored the secondary metabolome of these metabolically rich microbial taxa to develop a comprehensive understanding of their chemical diversity and unique traits[4–7]. Most of these studies begin by first identifying biosynthetic gene clusters within genomes using the popular antiSMASH software[8] and then group BGCs into gene cluster families (GCFs) using BiG-SCAPE[9]. Critically, this approach, in addition to other bioinformatics software for genomics annotation and analysis of BGCs[10–12], uses a set of key protein domains as their base to group BGCs predicted to encode similar metabolites. However, when working in the context of a single genus or species, it becomes more appropriate to use full protein sequences which enables the use of advanced algorithms for the identification of orthologs and paralogs[13] to reliably determine evolutionary relationships between BGCs.

Recent studies which explored the vertical descent and conservation of BGCs across the genus of *Salinospora* and the fungal species *Aspergillus flavus* have shown that intra-genus or intra-species evolution can result in chemical and regulatory diversification of the encoded metabolites[6,14]. Thus, improved understanding of the evolutionary emergence of BGCs could be informative towards elucidating the function of individual genes and encoded metabolites[9,15–17].

Due to difficulties in the cultivation of many bacterial taxa, several recent studies have begun to explore metagenomic datasets to unearth novel secondary metabolites[18–21]. While most of these endeavors have relied on performing metagenomic assembly followed by BGC identification using software such as antiSMASH, recently, two read-based approaches were described to reliably and sensitively search for key BGC domains in metagenomes[20,22]. This highlights a larger trend in which these approaches primarily aim to identify highly novel secondary metabolites and thus prioritize investigation of poorly studied or newly discovered taxa[23]. However, provided that most metabolites have been mined from select genera, mostly within the *Actinomycetia* class[2], there could be tremendous reward for efforts to advance our understanding of cataloged but uncharacterized BGCs from well studied taxa. This is especially important because many species within the metabolically rich *Streptomyces* only have limited representative genomes, thus methods to leverage metagenomic datasets and identify homologous instances of rare BGCs can lead to critical insight into their function and ecological distribution[4].

To aid the advancement of such taxa-specific analyses of BGCs, we introduce the comprehensive software suite *Lineage Specific Analysis* of BGCs (*lsa*BGC; https://github.com/Kalan-Lab/lsaBGC), which consists of several programs for comparative genetics of BGCs as well as metagenomic mining. We showcase the utility of *lsa*BGC in this study through application to four major genera which are commonly found in healthy human skin microbiomes. Several secondary metabolites which function as antibiotics[24–26], virulence

factors[27], or in microbe-host interactions[28,29] have been identified in bacteria from skin microbiomes, primarily from staphylococci. Human skin is also easily accessible to sampling and most species can be cultivated for follow up studies. Using *lsa*BGC, we reveal several new ecological and evolutionary insights for both well characterized and unknown predicted metabolites, including signatures of inter-species transfer for the virulence-associated staphylococcal carotenoid staphyloxanthin and a predicted siderophore found to be highly conserved among divergent *Corynebacterium* species. Additionally, through mining skin metagenomic datasets, we find novel coding variants within the functionally important ketide synthase (KS) domain of the polyketide synthase (PKS) responsible for mycolic acid biosynthesis, a defining cell wall component of *Corynebacterium* species implicated in immune recognition.

Results:

*lsa*BGC allows for sensitive and systematic identification of BGC homolog groups

The *lsa*BGC suite consists of seven core programs, three workflow programs, and several additional scripts for analyses (Figure 1, S1; Supplementary Methods and Results). These programs allow for clustering BGCs into GCFs, refining boundaries of BGCs, and sensitive, yet specific, rapid detection of GCF instances in draft-quality assemblies. The suite additionally includes dedicated tools for GCF visualization, evolutionary and population genetic analysis of individual genes within GCFs, and base-resolution identification of novel SNVs in complex metagenomes.

To demonstrate the capabilities of the *lsa*BGC suite we applied it to four genera which are stable and integral members of the human skin microbiome[30–34]. These included three genera in the Actinomycetota (formerly Actinobacteria) phylum: *Corynebacterium*, *Cutibacterium*, and *Micrococcus* as well as the Bacillota (formerly Firmicutes) genus *Staphylococcus*. *lsa*BGC was also independently applied to the species or species-complex from each genus most commonly represented in skin microbiomes. These were the *C. kefirresidentii* species-complex, *C. acnes*, *M. luteus*, and *S. epidermidis* (Table S1).

We used *lsa*BGC-AutoProcess to run gene-calling, BGC annotation, and *de novo* homolog group delineation with complete or chromosomally complete genomes for each genus or species. We then applied *lsa*BGC-Cluster to group homologous BGCs into GCFs and systematically searched for instances of GCFs using *lsa*BGC-AutoExpansion (Figure 1b; Table S2, S3; Supplementary Methods and Results). To assess the validity and sensitivity of our approach, two benchmarking tests using BGCs identified in *M. luteus* were performed (Figure 1c, S3; Supplementary Methods and Results) showing that the *lsa*BGC-AutoExpansion framework leads to greater sensitivity compared to running *de novo* antiSMASH on fragmented draft-quality assemblies. This increased sensitivity was due to more robust, yet reliable, detection of auxiliary genes located on the edges of scaffolds not containing core domains used by antiSMASH for BGC detection. In addition, *lsa*BGC-AutoExpansion is highly efficient and allowed us to search for 63 GCFs identified from high-quality *Staphylococcus* genomes in

14,978 draft-quality genomes in approximately 14 hours using 40 cores and less than 150 Gb of memory. We found a positive association between detection of complete BGC instances and assembly quality (Figure 1d; Supplementary Methods and Results).

_lsa_BGC enables reliable comparative genomics to explore evolutionary trends of BGCs

The increased sensitivity of *lsa*BGC to uniformly detect auxiliary components of BGCs in each genome enabled us to perform reliable comparative genetic analyses on them. We focused on two of the four genera with the greatest breadth of diversity, *Corynebacterium* and *Staphylococcus*, and selected a subset of 456 and 229 representative genomes from each genera, respectively (Table S1). To better understand which species in these genera were most relevant to skin health, we first profiled how often distinct species were found among healthy skin microbiomes using a large, recently generated metagenomics dataset[35,36] (Table S4). Skin-associated species (found in more than 10 of 270 metagenomes) were phylogenetically clustered for both genera, within the *S. epidermidis* clade (Figure S4a) and within the *C. kefirresidentii* species complex (Figure S4b; Supplementary Methods and Results).

Association of individual homolog groups from these GCFs revealed several were enriched within skin-associated staphylococci, but only two were enriched for skin-associated corynebacteria, both encoding hypothetical proteins (Table S5). Nearly half of the homolog groups enriched for skin-associated staphylococci (5 of 12; 41.7%) were related to GCF-10, predicted to encode for an NRP synthetase corresponding to pyrazinone biosynthesis, molecules reported to regulate virulence in *S. aureus*[27,37]. Staphylococcal pyrazinones were previously identified as being a self-contained BGC consisting of two genes, *pznA* and *pznB*, specific to species commonly associated with human skin, including *S. aureus*, *S. epidermidis*, and *S. capitis*[27,37]. Using *lsa*BGC, we discover novel variants of pyrazinones, which encapsulate seven distinct GCFs (Figure 2a), that belong to species not commonly found on human skin, such as *S. felis* (GCF-56), the dominant species of *Staphylococcus* on cat skin. While speciation had resulted in orthologous copies of *pznA* and *pznB* being highly diverged in sequence, it is surprising to find that the surrounding genomic context of the two genes is highly variable for all seven GCFs (Figure 2b), suggesting their ancestral mobilization or relocation within or between genomes.

Application of *lsa*BGC-Divergence followed by Bayesian shrinkage analysis was then used to systematically assess whether pyrazinones or other *Staphylococcus* and *Corynebacterium* GCFs exhibited indications of HGT between species, as suggested by high Beta-RD values (Figure S4cd; Table S6; Supplementary Methods and Results). Despite the small size of pyrazinones, we do not find support for the frequent mobilization of any of the seven GCFs encoding them across staphylococcal species. However, high Beta-RD posterior ranges were observed for two NRPS associated GCFs from *Corynebacterium*. One of these GCFs, predicted to encode for an NRPS-dependent siderophore[38], was found in several species commonly isolated from skin[35], including *C. kefirresidentii*. Using results from *lsa*BGC-PopGene for this GCF, which includes determination of a consensus gene order across BGC instances, we find that the NRPS is highly conserved in sequence across species and found flanked by

mobile genetic elements, as previously reported in *C. jeikeium*[38] (Figure 2cd). The taxonomic breadth and high sequence conservation of this GCF suggest that it has undergone recent HGT between distantly related species.

We next used *lsa*BGC-PopGene to identify genes under sweeping or balancing selection by comprehensively calculating Tajima's D for each homolog group within a GCF across all four genera, including *Micrococcus* and *Cutibacterium* (Table S7). We found that a hybrid terpene/T3PKS GCF, in *Staphylococcus* exhibits the greatest signal for intra-species sweeping selection and is one of four GCFs observed to be ancestral to the *S. epidermidis* clade (Figure S5; Supplementary Methods and Results). The terpene component of this hybrid GCF encodes for staphyloxanthin carotenoid production[39] and manual inspection determined that the T3PKS prediction actually corresponds to hydroxymethylglutaryl-CoA (HMG-CoA) synthase, which is involved in the mevalonate pathway for isoprenoid synthesis and is homologous to KS domains of polyketide synthases[40]. In addition to these two BGC predictions, the regions in between and flanking the staphyloxanthin encoding BGC and HMG-CoA synthase include several genes important to staphylococcal virulence, some of which were found in multiple copies in certain species (Figure S6ab). Further, while the five *crt* operon genes encoding for staphyloxanthin were found in multiple species within the *S. epidermidis/aureus* clade, they are particularly conserved in *S. aureus* as indicated by low Tajima's D values (Figure S6b).

The virulence-associated carotenoid staphyloxanthin is ubiquitous across the *Staphylococcus* genus

Although staphyloxanthin is well characterized as an *S. aureus* virulence factor[41], we were unable to identify a report of its distribution across the *Staphylococcus* genus, with only a single study, to our knowledge, suggesting its production in another species[42]. *lsa*BGC analysis revealed a total of five staphylococcal GCFs encoding for staphyloxanthin. Following GCF-3, the second most prevalent GCF encoding staphyloxanthin, GCF-6, had a high Beta-RD distribution (Figure S4c) and was widely distributed across divergent species in the genus (Figure 3a), as well as *Mammaliicoccus,* the closest phylogenetic neighbors of staphylococci[43]. Constructing a phylogeny based on a concatenated alignment of *crtM* and *crtN* revealed the appropriate delineation of GCF-3 and GCF-6 despite their co-occurrence in the genomes of particular species, such as *S. warneri* (Figure 3b, S7a). Additional analyses showed that GCF-6 has mobilized across species and can be found on plasmids (Figure 3bcd).

To assess whether carriage of the *crt* operon translates to staphyloxanthin production in non-aureus species, we identified staphylococcal isolates cultivated from human skin with whole-genome sequences generated by our lab that encode a staphyloxanthin related GCF. We identified a rare instance of an *S. epidermidis* isolate with GCF-3 (only 1-3% of *S. epidermidis* genomes encode GCF-3), strain LK1136, and an *S. warneri* isolate with both GCF-3 and GCF-6, strain LK413. While we observe GCF-3 in all *S. warneri* genomes, the GCF is degraded encoding only the dehydrosqualene synthase gene, *crtM*, and the dehydrosqualene desaturase gene, *crtN* (Figure 3a; Table S8). However, roughly half of the available *S. warneri* genomes (47.5%) encode GCF-6 in full and thus have two copies of *crtM* and *crtN*. We used long-read

sequencing to generate near-complete genomes for *S. epidermidis* LK1136 and *S. warneri* LK413 and found that both strains encode GCF-3 within their chromosomes while GCF-6 is encoded on a plasmid within *S. warneri*. (Figure S7b; Supplementary Methods and Results).

To check for phenotypic effects in strains carrying different staphyloxanthin GCFs, we assessed pigmentation levels in strains of *S. aureus*, *S. epidermidis*, and *S. warneri* with and without full *crt* operons (Table S9). Cells of *S. epidermidis* encoding GCF-3 are pigmented compared to strains without the cluster (Figure 3e-f). Cells of *S. warneri* LK413 containing the complete *crt* operon (GCF-6) and additional copies of *crtMN* displayed the strongest golden pigmentation, confirmed as staphyloxanthin by spectrophotometry[39,44] (Figure 3f).

_lsa_BGC provides a framework for mining for base-resolution novelty in BGCs

We have demonstrated how the *lsa*BGC suite can be used to identify homologous BGC instances into GCFs and subsequently uncover their evolutionary trends and phylogenetic distribution. With increasing metagenomic datasets becoming available, often from complex microbiomes, we developed *lsa*BGC-DiscoVary to directly profile BGCs and rapidly identify novel variation within GCF related genes from metagenomes (Figure S2; Supplementary Methods and Results). The *lsa*BGC framework thus allows users to first identify all instances of GCFs in genomic assemblies to build a comprehensive database of known alleles for GCF genes and subsequently apply *lsa*BGC-DiscoVary to find novel SNVs within genes which are absent in the available assemblies. Further, because *lsa*BGC-DiscoVary extracts the subset of reads supporting the presence of novel SNVs, exhaustive methods to assess the validity of metagenomic SNV calling, which would be computationally intensive and impractical to perform for the full set of reads from a metagenome, are made accessible for further validation of novel SNVs.

We first assessed and found high concordance (>96% overlap) between novel SNVs reported by *lsa*BGC-DiscoVary compared to an assembly-based approach using whole-genome sequencing readsets for 132 *M. luteus* isolates to benchmark it's sensitivity and specificity (Figure S8; Table S10, S11; Supplementary Methods and Results). Next, to demonstrate *lsa*BGC-DiscoVary's application in metagenomic datasets, its primary intended usage, we investigated the microdiversity of BGCs from *C. acnes*, the most abundant species at sebaceous sites, within healthy skin metagenomes generated by our lab[35]. We found that cutimycin, an anti-Staphylococcal thiopeptide, exhibited genomic signatures of recent acquisition in *C. acnes* and likely has since been lost in most isolates besides those belonging to clade IB and clade III (Figure 4ab, S9, S10a; Supplementary Methods and Results). Extensive metagenomic analysis demonstrated the reliability of *lsa*BGC-DiscoVary to type clade-specific alleles of genes within the cutimycin BGC in metagenomic datasets which correlated with strain presence inferences (Figure 4cd, S10bc; Table S12). We used *lsa*BGC to further investigate novel SNVs detected within cutimycin genes and found that many were from the metagenomes of a single individual and belonged to clade III *C. acnes* (Figure 4efgh). Through observation that the majority of novel SNVs from this individual also corresponded to cytosine to thymine changes, we discovered that clade III *C. acnes* have elevated rates of

cytosine deamination, which coincides with the clade's carriage of a truncated uracil DNA glycosylase, the enzyme responsible for the repair of such deaminations (Figure 4ijk).

*lsa*BGC identifies novel SNVs within polyketide synthase catalytic domains

We determined the *C. kefirresidentii* species complex is the most broadly distributed *Corynebacterium* across our skin metagenomes (Figure S4b); however, only 22 genomic assemblies, representing five species, were available for the complex at the time of this analysis. This clade thus represents an ideal taxonomic group to examine novel BGC microdiversity using *lsa*BGC-DiscoVary. We ran *lsa*BGC-DiscoVary systematically on six GCFs identified in the species complex, largely encoding for unknown metabolites, and found 40,019 putatively novel SNV instances across 105 homolog groups and 109 metagenomic samples. Of these, 5,474 instances were filtered because reads supporting their presence aligned with a higher score to other regions in a concatenated database of all *Corynebacterium* genomes from GTDB as compared to the reference database used for *lsa*BGC-DiscoVary. While ubiquitous, *C. kefirresidentii* species complex members are present at very low abundances in some body-sites where metagenomic assembly would struggle to construct their genomes (Figure S11a). We performed sample-specific metagenomic assemblies[45] and found that at least 11,659 novel SNV instances detected by *lsa*BGC-DiscoVary would be missed by assembly-based investigation (Figure 5a; Supplementary Methods and Results).

To further investigate the novel SNVs detected, we focused on 68 homolog groups overlapping or in close proximity to protocore regions of BGCs and further filtered based on standardized coverage metrics, retaining 5,802 instances of 2,343 unique novel SNVs. Novel SNVs were more likely to correspond to synonymous changes if they were found in multiple metagenomic samples (Figure S11b). Further, metagenomic samples from the same subject at different body sites and from the same body sites across different subjects were more likely to share novel SNVs when compared to metagenomic samples from different body sites and different subjects (p<1E-8; two-sided Wilcoxon rank sum test) (Figure S11c). We additionally found that homolog groups for which *lsa*BGC-PopGene estimated lower Tajima's D values, suggesting the presence of sweeping selection, tended to have fewer novel SNV counts (Figure 5b).

Finally, we expected and observed that the ratio of synonymous to non-synonymous novel SNVs increases at deeply conserved sites along homolog groups (Figure 5c; Table S13). The homolog group with the greatest number of novel SNVs, and novel, non-synonymous SNVs within highly conserved regions, was the PKS gene responsible for synthesis of mycolic acid, a major component of the cell wall in most *Corynebacterium* species recently shown to have immunomodulatory effects on the skin[28,46,47]. We identified five non-synonymous SNVs along conserved sites of the PKS gene, including three within the ketoacyl synthase domain, which catalyzes condensation[48], and could alter reaction kinetics[49] or scaffold structure[50] (Figure 5d, S12).

Discussion:

The *lsa*BGC suite packages several important, as well as novel, functionalities for large-scale evolutionary analysis of BGCs from any taxa in a structured framework. In particular we developed several algorithms to overcome challenges when working with draft-quality genomic assemblies. The suite's metagenomic mining functionalities further provide valuable insight into the presence and sequence diversity of a specific taxa's BGCs across diverse microbiomes. This is a particularly important feature for rare BGCs and taxa with a limited numbers of genomic assemblies available, such as taxa that are difficult to cultivate from complex metagenomes. A current limitation of the *lsa*BGC framework includes its basis for using chromosomally complete genomes to perform initial identification of BGCs and GCFs, which can miss rare BGCs only observed in draft genomes. To alleviate this limitation, future development of the suite will aim to allow direct incorporation of cataloged BGC and GCF instances from publicly available databases or software[12,51–53]. We may also consider the use of pan-genome alignment approaches to potentially increase the efficiency of *lsa*BGC for metagenomic identification of novel variants within BGC genes[54,55].

Here, we apply *lsa*BGC's functionalities to major taxa commonly found in skin microbiomes as a proof of concept to demonstrate its ability to uncover novel insights into the diversity and prevalence of both well-studied and uncharacterized secondary metabolites. For example, we uncovered the ubiquity of the carotenoid staphyloxanthin across the genus of *Staphylococcus,* a molecule conferring the signature golden pigmentation in *S. aureus*[56,57]. These findings raise the hypothesis that staphyloxanthin, which has been identified primarily as a virulence factor conferring oxidative stress resistance in the pathogen *S. aureus*[41,58], could similarly contribute to virulence in other staphylococcal species or serve additional unknown functions. Supporting this hypothesis, we observe that one staphyloxanthin encoding GCF appears ancestral to the skin-associated, multi-species clade which features both *S. aureus* and *S. epidermidis*. The observation that multiple copies of *crt* genes can be found within a single genome further suggests staphyloxanthin is of functional importance since evolutionary retention of paralogous genes has previously been shown to underlie ecological advantages[59,60]. Additionally, while this GCF is present in most skin-associated species, it is absent in approximately 95% of *S. epidermidis* genomes, suggesting large-scale loss within the species. It will thus be interesting to further explore how the loss of staphyloxanthin, if it contributes to oxidative stress resistance within *S. epidermidis*, coincides with the species' reputation as a beneficial symbiont of host skin[32,61,62].

Using the novel metagenomic mining functionality of *lsa*BGC, we identified novel SNVs distinguishing uncultured *C. acnes* and *Corynebacterium* strains from isolates with genomes available in public databases. With regards to *Corynebacterium*, we determined that some of the most common skin residing species[31,35] belong to the *C. kefirresidentii* species complex, which were widely distributed across skin metagenomes and highly abundant at moist body sites. Because this species complex is vastly underrepresented in public genomic databases and additional species and strains belonging to it likely remain to be discovered, application of *lsa*BGC-DiscoVary led to the identification of thousands of novel SNVs within the BGCs of the

clade. Critically, some of these variants could have functional implications as they lie within highly conserved regions of important domains for natural product biosynthesis.

The comprehensive nature of *lsa*BGC allows analyses to be applied to both public and proprietary microbial genomic databases. Besides accelerating fundamental evolutionary research of BGCs from diverse taxa, we envision *lsa*BGC will be particularly useful for secondary metabolite discovery and mechanism of action studies. Specifically, we believe the evolutionary insight gathered from *lsa*BGC can be used as prior information to guide algorithms in the emerging field of using artificial intelligence to identify new antibiotics[63,64]. Finally, an improved understanding of the relationship between intra-taxa genomic diversity and BGC content novelty at the base-resolution using *lsa*BGC highlights which sub-lineages or sub-clades bear the largest reservoir of untapped secondary metabolic potential.

Acknowledgments:

This work was supported by grants from the National Institutes of Health (NIAID U19AI142720 and NIGMS R35GM137828 [L.R.K]). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors gratefully acknowledge additional members of the Kalan lab, Dr. Cameron Currie, Dr. Reed Stubbendieck, Dr. John-Demian Sauer, Dr. Jason Kwan, Dr. Karthik Anantharaman, Dr. Caroline Grunenwald, Dr. Chase Clark, and Dr. Milton Drott for helpful discussions.

Materials and Methods:

Code availability

The *lsa*BGC software suite was developed in Python3 and R and is available on Github at: https://github.com/Kalan-Lab/lsaBGC. Algorithmic descriptions of programs can be found on the Github wiki and in the Supplementary Methods and Results document. A test dataset pertaining to the *C. kefirresidentii* species complex for running the seven core programs is provided on the Github repository, along with expected results.

An overview of the *lsa*BGC software suite

The first core program is *lsa*BGC-Cluster which clusters homologous and non-fragmented instances of BGCs, identified by antiSMASH from chromosomally complete genomic assemblies, into gene cluster families (GCFs), similar to BiG-SCAPE[9].Users can then generate a report to select for the most appropriate parameters for downstream analyses of interest. Because antiSMASH groups neighboring BGCs into hybrid gene clusters if they are located close to one another on the genome[7], *lsa*BGC-Refiner has the option to manually separate hybrid clusters based on user-defined boundary homolog groups. Homologous instances of GCFs across all available genomic assemblies for the taxa of interest can then be identified through *lsa*BGC-Expansion (Figure S1b). Similar to ClusterFinder[10], *lsa*BGC-Expansion uses profile Hidden Markov models (HMMs) to identify homologs of GCF associated

genes in assemblies followed by a classical HMM approach to define BGC regions and their boundaries. Unlike ClusterFinder, *lsa*BGC-Expansion: (i) uses all coding genes associated with a GCF instead of being restricted to core protein domains, (ii) is specifically designed to account for assembly fragmentation through conditional assessment of predicted regions, and (iii) features a "quick mode" option for fast detection of BGCs using Diamond[65].

The remaining four core programs are analytical and designed to be run on a single GCF for ease of parallelization: *lsa*BGC-See, *lsa*BGC-PopGene, *lsa*BGC-Divergence, and *lsa*BGC-DiscoVary. *lsa*BGC-See permits visualization of BGCs across a user-provided species tree or a BGC based phylogeny and can handle multiple fragmented BGC instances from the same genome, depicting them as neighboring leaves on phylogenies[66]. *lsa*BGC-PopGene can then infer evolutionary and population genetics statistics for each homolog group associated with a GCF. It generates codon based alignments for each homolog group which can then be used as input for *lsa*BGC-Divergence and *lsa*BGC-DiscoVary. *lsa*BGC-PopGene also determines the consensus order and direction of homolog groups for a GCF across all BGC instances, which provides intuitive assessment of its reports. *lsa*BGC-Divergence calculates a comprehensive statistic between pairs of homologous BGCs from the same GCF but different genomes, which we refer to as Beta-RD, that captures how similar the BGC sequences are relative to what would be expected based on genome-wide similarity. Finally, *lsa*BGC-DiscoVary is a multi-functional program which can mine raw sequencing readsets from metagenomes or single genomes, for GCF instances and then reliably identify whether they possess novel single nucleotide variants (SNVs) yet to be observed in available genomic assemblies for a taxa (Figure S2).

Usage of genomic and metagenomic datasets

To overcome issues with misclassification and contamination in NCBI, we used GTDB (release 202) [67], to query for reliable genomic assemblies and respective species designations for the genera of *Corynebacterium*, *Cutibacterium*, *Micrococcus*, and *Staphylococcus* (Table S1). For *Micrococcus*, we further included 132 additional genomes of the type species *M. luteus* which our lab had sequenced (BioProject ID PRJNA803478) but were not yet uploaded to NCBI to be included in the GTDB release used. These isolates were classified as *M. luteus* by GTDB-tk[68] and exhibited > 95% ANI to the NCBI representative genome for the species *M. luteus* NCTC 2665. A similar approach was used for gathering genome sets for species-level analyses performed (Table S1). For the *C. kefirresidentii* species complex, an additional five genomes from NCBI were gathered, classified using GTDB-tk[68], and incorporated into analysis if they belonged to one of the five species designated as belonging to the complex.

Because a single representative per species as provided by GTDB might not adequately capture the diversity of BGCs found across a genus, we instead employed a more granular approach to select for representative genomes for a genus. We performed dereplication using MASH[69] with an estimated ANI cutoff of 0.99 to pair similar genomes and then performed single-linkage clustering and selected representative genomes from each cluster based on assembly N50 (Table S1). This dereplication was performed using the script

popSizeAndSampleSelector.py provided in the *lsa*BGC suite. Complete or chromosome level assemblies among the representative genomes were run through lsaBGC-AutoProcess.py and lsaBGC-Cluster.py to identify the core set of GCFs per genus. GCFs determined from the core set of genomes were comprehensively identified in the remaining genomes using lsaBGC-AutoExpansion.py, including those which were not deemed as representatives from the MASH based analysis to showcase scalability (e.g. all 15,055 *Staphylococcus* and not just the 229 representative genomes of which 77 corresponded to complete/chromosome level assemblies). To expedite population genetics analysis however, we only ran the downstream functional analytics of *lsa*BGC using the representative genomes for each genus. For species-level analyses, all genomes available were used for such downstream analyses and to comprehensively profile known alleles for genes in BGCs, which was critical to enable metagenomic mining for base-resolution novelty previously unseen in assemblies available for a species. For population level analyses within species, strain delineations were performed using the popSizeAndSampleSelector.py script in *lsa*BGC with strain groups determined from single-linkage clustering of pairs of genomes found to exhibit $\geq$98% ANI and $\geq$80% genomic content similarity by FastANI[70].

For determining skin-associated species amongst the genera investigated in our study and to demonstrate mining functionalities with *lsa*BGC-DiscoVary, we used a paired-end metagenomics dataset sequenced on the Illumina NovaSeq where individual samples corresponded to microbiome surveys of one of eight body sites for one of 34 participants, collected in the state of Wisconsin (BioProject ID PRJNA763232)[35]. As previously described[35], adapter removal, quality filtering, human sequence decontamination, and tandem repeat removal were performed using fastp (v0.21.0)[71] and KneadData (v0.8.0) (https://huttenhower.sph.harvard.edu/kneaddata/).

Genomic assessment for staphyloxanthin carriage in the lab isolate collection

We had previously constructed draft genomic assemblies for 44 *Staphylococcus* isolates from skin (BioProject ID PRJNA803478; BioProject ID XXXX). Illumina sequencing was performed at the Microbial Genome Sequencing Center (MiGS). Briefly, assemblies were generated using Unicycler[72] with default settings after quality and adapter trimming using fastp[71]. The set of 44 *Staphylococcus* genomes were searched for carriage of staphyloxanthin encoding GCFs identified in our study using *lsa*BGC-AutoExpansion (Table S9).

Library preparation, sequencing and genomic assembly

Reconstruction of Illumina draft assemblies for 132 M. luteus: We assessed the performance of multiple programs in the *lsa*BGC suite using a set of genomic assemblies for 132 *M. luteus* isolates which were cultured from skin. For these isolates, assemblies were reconstructed from raw-reads (BioProject ID PRJNA803478) and were uploaded to NCBI's Genbank database (Table S1). Briefly, Illumina sequencing reads were processed using fastp (v. 0.20.1)[71] to trim for Poly-G tail artifacts and TrimGalore (v. 0.6.5)[73] was additionally used to detect and remove adapters. Afterwards, assembly was performed using Unicycler (v. 0.4.6)[72] with default settings.

Construction of Illumina+Nanopore hybrid assemblies for staphyloxanthin producing S. warneri *LK413 and* S. epidermidis *LK1136*: To learn the genomic location of staphyloxanthin encoding GCFs in *S. epidermidis* LK1136 and *S. warneri* LK413, we performed additional long-read sequencing using Oxford Nanopore Technologies (ONT). Library construction, using a PCR-free ligation method, and sequencing were performed at the Microbial Genome Sequencing Center (MiGS). Hybrid assemblies were constructed using a modified version of the Hybrid Assembly workflow in seQuoia[74]. Illumina reads, used for construction of the initial draft genomic assemblies for the isolates, were first reprocessed using fastp (v. 0.20.1)[71] to trim for Poly-G tail artifacts and TrimGalore (v. 0.6.5)[75] was subsequently used to detect and remove adapters. Similarly, ONT FASTQ reads, provided by MIGS after basecalling, were filtered for potential adapters using PoreChop (v. 0.2.3) (https://github.com/rrwick/Porechop) and filtered to retain only reads greater than 3 kb in length as well as subsampled to 300 mb (while being inclusive of all reads greater than 20 kb) using the script fastqfilter.py included in seQuoia. Afterwards, hybrid-assemblies were constructed using Unicycler (v. 0.4.6)[72] with default settings and additional Pilon (v. 1.23)[76] polishing was iteratively performed by aligning ONT reads[77] as well as the Illumina reads[78] to assemblies and refining them, allowing for a maximum of 10 iterations. Quality assessment of the two near-complete assemblies was performed using GAEMR (https://github.com/broadinstitute/GAEMR) (Supplementary Methods and Results).

Experimental validation of staphyloxanthin production in non-aureus *Staphylococcus*

Isolates were grown overnight at 37°C on tryptone soy agar plates and inoculated into tryptone soy broth at a concentration of $1x10^6$ CFU/mL. Liquid cultures were grown in shaking conditions at 37°C for 24 hours. 10 mL of the liquid culture was then pelleted and washed twice with phosphate-buffered saline before pigment extraction in 1 mL of methanol at 55°C for 15 mins. Spectral scans from 300 nm to 600 nm were taken using a plate reader (EPOCH2, Biotek, Winooski, VT). Absorbance readings were blanked with methanol and normalized to liquid culture cell density based on OD600. Spectra shown are representative of at least two biological replicates.

MASH distance neighbor-joining tree construction

Average nucleotide identity estimates between pairs of genomes were computed with MASH[69] and used to create a distance matrix. The distance matrix was then used to construct a neighbor-joining tree with the ape library in R[79,80].

Ribosomal and BGC phylogeny construction

Ribosomal phylogeny construction: Ribosomal phylogenies for each species were constructed based on a concatenated alignment of 16 ribosomal proteins[81,82]. HMMER (v3) was used to search for profile HMMs of each ribosomal protein in the predicted proteomes of samples of interest belonging to a particular taxonomic group. Afterwards, the best matching hits for each sample based on alignment E-value were identified and used to construct individual alignments

for each ribosomal protein with the L-INS-I method in MAFFT[83]. Protein alignments were then converted to codon alignments using PAL2NAL[84] and subsequently concatenated together. Samples which featured ≥ 25% gaps in the concatenated alignment were filtered after which core SNV sites were extracted and used for constructing a maximum likelihood phylogeny with RAxML[85], specifying a GTRCAT model and performing 100 bootstraps. Where needed, ribosomal phylogenies were subset to showcase relationships between select samples using PareTree (v1.0.2)[86].

Staphyloxanthin phylogeny construction: A maximum likelihood phylogeny was constructed from a concatenated protein alignment of CrtM and CrtN amongst representative *Staphylococcus* genomes to understand the relationship between different GCFs encoding for staphyloxanthin. Proteins designated as HG0001876 (CrtM) and HG0001904 (CrtN) were gathered for each staphyloxanthin encoding GCF found in each sample and aligned using the L-INS-I method in MAFFT[83]. To account for assembly fragmentation resulting in multiple proteins from a sample being assigned to the same homolog group for a specific GCF, a consensus sequence was determined and multi-allele sites were replaced as gaps. A concatenated alignment of such consensus sequences was subsequently constructed and used for phylogeny construction with RAxML using a PROTCAT model with JM selected as the best amino-acid replacement matrix through the automatic maximum likelihood model selector[85]. The phylogeny of CrtMN sequences exhibited high correspondence to GCF delineations with the exception of a single *S. lugdunensis* BGC which is likely misclassified as GCF-17 and should be GCF-3. Upon further examination, it was identified that the reason for the misclassification was likely due to the absence of mevalonate pathway genes, classified as a type III polyketide synthase (T3PKS), within *S. lugdunensis* which are found downstream of the *crt* operon in all other species with GCF-3 (Figure S7b).

Pyrazinone phylogeny construction: A maximum likelihood phylogeny was constructed from a concatenated protein alignment of PznA and PznB amongst representative *Staphylococcus* genomes using the same methods as described for phylogeny construction from CrtMN proteins.

Designation of species belonging to *Mammaliicoccus*

The *Staphylococcus_A* genus designation in GTDB R202 was largely concordant with the recent reclassification of five species from *Staphylococcus* to *Mammaliicoccus*[43]. In the latest GTDB R207 release, this genus has now been reclassified to *Mammaliicoccus*. The recent update to taxonomic names also includes the reclassification of species *S. pasteuri_A* to *M. fleurettii*. Notably, we do not regard *S. schleiferi* as part of *Mammaliicoccus* because it is categorized as *Staphylococcus* by GTDB (both R202 and R207) and groups within the *Staphylococcus* genus in our ribosomal protein based maximum-likelihood phylogeny. Similar to our assessment, the reclassification of this species to *Mammaliicoccus* was recently questioned[87].

Ubiquity assessment of skin residing species and strains of *Staphylococcus* and *Corynebacterium*

We searched for the presence of representative *Staphylococcus* and *Corynebacterium* strains in 270 skin metagenomes from 34 participants across 8 body sites using StrainGST[36]. To construct the database of distinct representative genomes for each genus, the protocol described in the StrainGE documentation was followed. *S. aureus* genomes 16405 and 278 were dropped from the analysis because at the time of downloading them, the plasmid and chromosome were switched. Briefly, genomes were downloaded from NCBI using ncbi-genome-download[88] and dereplication was performed using built-in StrainGE functionalities. As some representative genomes were not part of the GTDB release R202, GTDB-tk[68] was used to annotate such genomes and perform species classification. We considered a species as skin-associated if it was found in at least 10 of the 270 skin metagenomes, regardless of the abundance (Table S4). Certain skin-associated species from each genus were also identified in the 19 negative control metagenomic samples, which included air-swab, bench, desk, and door-handle samplings (Table S4). Because we expect heavy overlap in species content between these environmental samples and skin microbiome samples, we still considered such species as skin-associated if they were found in 10 or more skin metagenomes.

We found the majority of skin-associated *Staphylococcus* belonged to the *S. epidermidis/aureus* clade, as has previously been noted[40]. Similarly, most of the skin-associated *Corynebacterium* belonged to *C. kefirresidentii* and it's close neighboring species (including *C. kefirresidentii*, *C. aurimucosum* type E, *C.* sp. 900539985, *C. tuberculostearicum*, *C. tuberculostearicum* type C), which we denote in this study as the *C. kefirresidentii* species complex. The minimal ANI separating genomes belonging to this species complex was determined to be 88.59% using FastANI[70]. To assess whether any of the 24 new *Corynebacterium* species reported by Saheb Kashaf et al. 2022[34] belonged to this species complex, FastANI was used to look at their ANI similarity to genomes from the clade and regard them as additional members if they exhibited 88.59% ANI or greater. Of note, many additional *Corynebacterium* species have been routinely isolated from the skin[35] which we did not regard as skin-associated based on our criteria in this study. In particular, the species of *C. amycolatum* (detected in 4 metagenomes), *C. kroppenstedtii* (detected in 4 metagenomes), and *C. simulans* (detected in 5 metagenomes) are commonly found on skin.

To estimate the relative abundance of *Cutibacterium* (see section "Identification of *Cutibacterium* Strains within Metagenomes and Determination of Near-Fixed Sites Differentiating Cutimycin Biosynthesis Alleles between Clade I and Clade III *C. acnes*") and *Corynebacterium* strains within metagenomic samples we normalized the estimated relative abundances for strains (*rapct*) by the estimated relative abundance of the genus (*pan%*) as reported by StrainGST[36].

Identification of homolog groups associated with skin residing species of *Staphylococcus* and *Corynebacterium*

We developed a script called crawlingFisher.py, provided in the *lsa*BGC suite, which tests for enrichment or depletion of homolog group presence amongst genomes in focal clades

using Fisher's exact test. Provided with a phylogenetic tree and matrix specifying the presence of homolog groups across genomes, the program automatically performs such tests for each homolog group at each inner-node of the phylogeny. Multiple testings correction was performed comprehensively at the end using Benjamini-Hochberg false discovery rate and cases were considered statistically significant if the adjusted p-value was less than 0.05. Results were further filtered to retain only homolog groups which were present in at least 80% of genomes under the focal node and at most 20% of other genomes or, alternatively, at most 20% of genomes under the focal node and at least 80% of other genomes. Finally, only results pertaining to focal nodes where at least 80% of genomes were classified as skin-associated (see section "Ubiquity Assessment of Skin Residing Species and Strains of *Staphylococcus* and *Corynebacterium*") were reported (Table S5).

Assessing the divergence of BGCs relative to their background genomic contexts

BGC divergence does not always track with background genomic divergence and speciation[14]. The program *lsa*BGC-Divergence allows for easy assessment of how BGC divergence compares to whole-genome divergence, primarily through computing a statistic we refer to as Beta-Relative-Divergence (Beta-RD) between pairs of genomes for a single GCF. The Beta-RD statistic is simply the average amino or nucleotide identity between homologous instances of proteins for a GCF observed in two genomes normalized by the average amino or nucleotide identity observed between the full genomes. *lsa*BGC allows users to specify whether amino or nucleotide divergence should be used and automated workflows (see section Automated Workflows for Running *lsa*BGC) can be used to compute pairwise AAI or ANI between genomes using CompareM[89] or FastANI[70], respectively.

To better understand how distributions of Beta-RD differ across GCFs, in this study, we applied a Bayesian hierarchical model to perform shrinking analysis and appropriately account for differences in the number of genomes GCFs are found (Figure S4cd; Table S6). For computational efficiency, a maximum of 500 Beta-RD observations were used for each GCF. Posterior predictions for Beta-RD were made using rstan (Stan Development Team 2021) with the following hierarchical model:

$$\overline{\mu} \sim Norm(\mu_0, \sigma_0^2)$$
$$\tau \sim half - t_7(1)$$
$$\mu_1 ... \mu_J \sim Norm(\overline{\mu}, \tau)$$
$$\sigma \sim half - t_7(1)$$
$$y_j \sim Norm(\mu_j, \sigma^2)$$

where for each GCF j, the posterior Beta-RD distribution is sampled as y_j using Markov chain Monte Carlo with four chains, each featuring 2,000 iterations, including a burnin of 1,000 iterations. A hierarchical prior was specified for $_j$, which represents a GCF-specific baseline. The values of $_0$ and σ_0 were set to the respective mean and standard deviation of raw Beta-RD values across all GCFs. R and STAN code used for this analysis are available within the scripts subdirectory of the *lsa*BGC software package.

Scrutinization of putative novel variants for *lsa*BGC-DiscoVary application to cutimycin in *C. acnes* and the comprehensive set of GCFs for the *C. kefirresidentii* species complex

By default, lsaBGC-DiscoVary.py only considers reads aligning at 95% identity to reference gene sequences and avoids reporting novel variants on genes if either annotation or coverage (relative to other genes in the BGC context) suggests they might be MGEs. The program also automatically extracts the subset of reads supporting the existence of novel variants, which can then be further screened with high stringency filters using otherwise computationally expensive methods. This can involve mapping reads supporting putative novel SNVs to a representative database of whole-genomes to observe for higher quality alignments or checking whether reads are classified as a particular taxonomy by Kraken2 (v2.0.8-beta)[90]. As described previously[35], we used a customized database for Kraken2 analysis constructed from chromosomally complete bacterial, viral, archaeal, fungal, protozoan genomes as well as the human genome available on RefSeq, where plasmid sequences were distinguished to alleviate potential misclassifications of reads to a particular taxonomy.

For reads supporting putative novel variants in the cutimycin BGC of *C. acnes*, we required them to be classified as *Cutibacterium* or *Priopionibacterium* by Kraken2[90] and to not map with a higher alignment score to a comprehensive database of all *Cutibacterium* genomes than was observed for the alignment to the reference gene sequence in lsaBGC-DiscoVary.py (Figure S2c). We generally do not recommend screening variant-supporting reads for an appropriate taxonomic match with Kraken2 as it can generally lead to false negatives (on average only 92.8% of read-pairs per sample were taxonomically classified and misclassification can also occur; Figure S12), but performed this here to provide further assurance that novel variants discovered are reliable. Standardized coverage metrics within lsaBGC-DiscoVary.py's novel SNV report were further used to conservatively filter sites and variants if they exhibited site or allele-specific coverage values greater than two median absolute deviations from the median coverage for the entire BGC.

For reads supporting putative novel variants within GCFs of the *C. kefirresidentii* species complex, we required them to not align with a higher alignment score to a comprehensive database of all *Corynebacterium* genomes. We did not require reads to be classified by Kraken2 as *Corynebacterium* because genomes from the species complex are likely to be poorly represented in the Kraken2 database. We did however further validate that reads supporting the existence of variants at conserved sites in the mycolic acid biosynthesis gene mapped best to the *C. kefirresidentii* species and confirmed our suspicion that they could be misclassified by Kraken2 given the representation in the underlying database (Figure S12). This *ad hoc* analysis was performed using paired-end information when mapping in Bowtie2 and accounting for concordant read pair alignments. While many reads supporting the existence of novel SNVs in the biosynthesis gene were classified as *C. acnes* by Kraken2, we found that none of the reads concordantly aligned to the comprehensive *Cutibacterium* genomes Bowtie 2 database whereas 88% concordantly aligned to the comprehensive *Corynebacterium* genomes Bowtie 2 database.

Note, the reasoning behind mapping SNV supporting reads to comprehensive genomic databases to purge false positive SNV calls, is that if a read truly aligns best to the BGC-associated homolog group where it was used to call a SNV, then it should at most generate a mapping to this comprehensive database of equal mapping quality. We align paired-end reads individually as was performed in lsaBGC-DiscoVary.py to have mapping scores be comparable with those from the original alignment. Furthermore, because our initial database within lsaBGC-DiscoVary.py consisted of allelic representatives of homolog groups, we realized that reads mapping to the start and end of genes could result in a lower mapping score than if a larger genomic context was provided. To ensure we do not remove such reads mapping to gene edges, scrutunizeNovelSNVSupport.py, the program we developed to perform Kraken 2 and genomic database alignment based filtering of novel SNV reports from lsaBGC-DiscoVary.py, also compares the reference sequences when it finds a better Bowtie 2 mapping for SNV-supporting reads in genomic databases. If the reference sequence in the genomic database encompasses the reference sequence in the original alignment to the lsaBGC-DiscoVary.py homolog group database and the read was previously noted to map to a gene edge, then reads are still retained and SNV read support is not decremented. Five retained and supportive reads are required for an SNV to be featured in the filtered report produced by the program.

_lsa_BGC-DiscoVary based analyses of BGCs fromo the _C. kefirresidentii_ species complex:

For the following analyses, we used 5,802 instances of 2,343 unique novel SNVs found in 68 homolog groups which were core or adjacent to core biosynthesis machinery of BGCs from the *C. kefirresidentii* species complex. These novel SNVs had site and allele-specific coverage values less than two median absolute deviations from the median coverage for the entire BGC.

Comparing SNVs shared across body sites and subjects: To comprehensively assess whether novel SNVs were shared more frequently for microbiomes of the same body site or from the same individual, we performed a pairwise, multi-iteration resampling analysis. For each pair of metagenomes, where each sample had at least 30 novel SNVs, we calculated the average Jaccard similarity across 1,000 simulations in which we randomly drew sets of 30 SNVs from each metagenomic sample to control for sequencing depth. Pairs of metagenomes were classified as one of three categories: “different body site, different participant”, “different body site, same participant”, and “same body site, different participant” (Figure S11c). A two-sided Wilcoxon rank sum test for differences between the three distributions of average Jaccard similarities revealed that there was a statistically significant difference between “different body site, same participant’ and “same body site, different participant” (p=6.01E-3). Of greater relevance, there were statistically strong significant differences between the “different body site, different participant” distribution and the two other distributions: “different body site, same participant” and “same body site, different participant” (p=3.67E-09 and p=1.39E-116, respectively). Thus, pairs of metagenomes were more likely to share novel SNVs if they were either from the same body site or same participant compared to if they were from different participants and different body sites. Two metagenomic samples for participant S002 were not used because they corresponded to a resampling of two of their body sites and we did not have

enough of such samples to pursue a temporal analysis. We assessed this testing using different considerations/cutoffs for: (i) the number of novel SNVs needed for a metagenome to be considered in the analysis and (ii) whether or not to account for singleton SNVs. In all cases, statistical testing for differences between distributions of pairwise metagenome categories yielded similar conclusions.

Assessing a relationship between normalized SNV rates and Tajima's D for homolog groups: We assessed how the normalized count of novel SNVs observed across all metagenomic samples for the 68 homolog groups compared with respective Tajima's D estimate computed from representative alleles found in genomic assemblies. For this analysis, we only considered homolog groups with at least 1 novel SNV detected in the metagenomes and at least four sequences available from genomic assemblies, the minimum number needed to compute the Tajima's D statistic. To normalize novel SNV counts per homolog group, we multiplied the raw number of novel SNVs observed for a homolog group by the number of representative sequences present in genomic assemblies and divided by the number of metagenomic samples deemed to possess the homolog group. The rationale behind this normalization was that more novel SNVs would be expected for homolog groups that occur in more metagenomes and similarly fewer novel SNVs would be expected if a greater number of sequences are already known from available genomic assemblies. We found that homolog groups with Tajima's D < 0 generally had lower normalized novel SNV rates ($\leq$10), while homolog groups with Tajima's D > 0 generally had higher novel SNV rates (>10) (p=1.63E-3, Fisher's exact test). We observed similar trends when using non-normalized novel SNV counts. Notably, the homolog groups with the lowest Tajima's D and novel SNV counts were from GCF-5 and GCF-6, which are specific to a single species, *C. kefirresidentii*.

Assessing trends for novel SNV synonymous to non-synonymous rate with ubiquity and site conservation: Using reports from lsaBGC-DiscoVary.py, we performed a comprehensive assessment of whether novel SNVs identified on the 68 homolog groups nearby or overlapping with protocore regions of BGCs in *C. kefirresidentii* corresponded to synonymous or non-synonymous substitutions. As described in other sections, SNV alleles were compared to corresponding alleles observed at the site on the reference gene sequences which SNVs were called upon. We aimed to check whether the rate of synonymous to non-synonymous novel SNVs showed a relationship to: (i) how common SNVs were across metagenomic samples and (ii) conservation levels across protein sequences.

Novel SNV synonymous to non-synonymous rate increases with SNV ubiquity amongst metagenomic samples: The occurrence of novel SNVs, a novel allele observed at a specific site in the multiple sequence codon alignment of a particular homolog group, was tabulated across metagenomic samples. While novel SNVs found in a single metagenome were only twice as likely to correspond to a synonymous change as opposed to a non-synonymous change, more prevalent novel SNVs, found in ten or more samples, were eight times more likely to correspond to a synonymous change (Figure S11b).

Novel SNV synonymous to non-synonymous rate increases at conserved sites in protein sequences: We searched RefSeq's bacterial NR database for remote homologs of the 68 homolog groups of interest. The top 20 homologs which belonged to classified bacterial species outside of the *Corynebacterium* genus or *Corynebacteriales* family were identified using HMMER3 with profile-HMMs gathered from *lsa*BGC-Expansion for each homolog group and an E-value threshold of 1E-20 (Table S13). The sequences of these representative homologs were added to protein alignments for homolog groups in *lsa*BGC-PopGene results using the '--add' option in MAFFT[83]. Afterwards, we used the alignments for scoring protein sequence conservation with Jensen-Shannon divergence[91]. Conservation scores were used to determine conservation percentiles of sites along protein alignments for individual homolog groups. Novel SNVs found in the top five to ten percentile of conserved sites for a homolog group were approximately five to seven times more likely to represent a synonymous change as opposed to a non-synonymous change (Figure 5c). In contrast, the bottom 20% of conserved sites (the least conserved sites), were almost equally as likely to represent non-synonymous substitutions as synonymous substitutions. Further, the number of novel SNVs, either synonymous or non-synonymous observed at conserved sites, amongst the top 20% of conserved sites was 649, and lower than the number of novel SNVs found in the 20-40 (n=739), 40-60 (n=855), 60-80 (n=806), and 80-100 (n=697) percentile ranges of conserved sites. We suspect that the number of novel SNVs identified decreases past the top 40-60 percentile range because alignments to such regions become more challenging.

References:

1. Pham, J. V. *et al.* A Review of the Microbial Production of Bioactive Natural Products and Biologics. *Front. Microbiol.* **10**, 1404 (2019).
2. Watve, M. G., Tickoo, R., Jog, M. M. & Bhole, B. D. How many antibiotics are produced by the genus Streptomyces? *Arch. Microbiol.* **176**, 386–390 (2001).
3. de Lima Procópio, R. E., da Silva, I. R., Martins, M. K., de Azevedo, J. L. & de Araújo, J. M. Antibiotics produced by Streptomyces. *Braz. J. Infect. Dis.* **16**, 466–471 (2012).
4. Chevrette, M. G. *et al.* The antimicrobial potential of Streptomyces from insect microbiomes. *Nat. Commun.* **10**, 516 (2019).
5. Waterworth, S. C. *et al.* Horizontal Gene Transfer to a Defensive Symbiont with a Reduced Genome in a Multipartite Beetle Microbiome. *MBio* **11**, (2020).
6. Drott, M. T. *et al.* Microevolution in the pansecondary metabolome of Aspergillus flavus and its potential macroevolutionary implications for filamentous fungi. *Proc. Natl. Acad. Sci. U.*

S. A. **118**, (2021).

7. Steinke, K., Mohite, O. S., Weber, T. & Kovács, Á. T. Phylogenetic Distribution of Secondary Metabolites in the Bacillus subtilis Species Complex. *mSystems* **6**, (2021).
8. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
9. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
10. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
11. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110 (2019).
12. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience* **10**, (2021).
13. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
14. Chase, A. B., Sweeney, D., Muskat, M. N., Guillén-Matus, D. G. & Jensen, P. R. Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. *MBio* **12**, e0270021 (2021).
15. Felnagle, E. A., Rondon, M. R., Berti, A. D., Crosby, H. A. & Thomas, M. G. Identification of the biosynthetic gene cluster and an additional gene for resistance to the antituberculosis drug capreomycin. *Appl. Environ. Microbiol.* **73**, 4162–4170 (2007).
16. Del Carratore, F. *et al.* Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. *Commun Biol* **2**, 83 (2019).
17. Crits-Christoph, A., Bhattacharya, N., Olm, M. R., Song, Y. S. & Banfield, J. F. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore

activity. *Genome Res.* (2020) doi:10.1101/gr.268169.120.

18. Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**, 440–444 (2018).
19. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
20. Pereira-Flores, E. *et al.* Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultrafast techniques. *bioRxiv* 2021.01.20.427441 (2021) doi:10.1101/2021.01.20.427441.
21. Andreu, V. P. *et al.* A systematic analysis of metabolic pathways in the human gut microbiota. *bioRxiv* 2021.02.25.432841 (2021) doi:10.1101/2021.02.25.432841.
22. Sugimoto, Y. *et al.* A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science* **366**, (2019).
23. Chevrette, M. G. & Handelsman, J. Needles in haystacks: reevaluating old paradigms for the discovery of bacterial secondary metabolites. *Nat. Prod. Rep.* **38**, 2083–2099 (2021).
24. Zipperer, A. *et al.* Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* **535**, 511–516 (2016).
25. Claesen, J. *et al.* A Cutibacterium acnes antibiotic modulates human skin microbiota composition in hair follicles. *Sci. Transl. Med.* **12**, (2020).
26. Nakatsuji, T. *et al.* Antimicrobials from human skin commensal bacteria protect against Staphylococcus aureus and are deficient in atopic dermatitis. *Sci. Transl. Med.* **9**, (2017).
27. Wyatt, M. A. *et al.* Staphylococcus aureus nonribosomal peptide secondary metabolites regulate virulence. *Science* **329**, 294–296 (2010).
28. Ridaura, V. K. *et al.* Contextual control of skin immunity and inflammation by Corynebacterium. *J. Exp. Med.* **215**, 785–799 (2018).
29. Swaney, M. H. & Kalan, L. R. Living in Your Skin: Microbes, Molecules, and Mechanisms.

Infect. Immun. **89**, (2021).

30. Grice, E. A. *et al.* Topographical and temporal diversity of the human skin microbiome. *Science* **324**, 1190–1192 (2009).
31. Oh, J. *et al.* Temporal Stability of the Human Skin Microbiome. *Cell* **165**, 854–866 (2016).
32. Byrd, A. L., Belkaid, Y. & Segre, J. A. The human skin microbiome. *Nat. Rev. Microbiol.* **16**, 143–155 (2018).
33. Acosta, E. M. *et al.* Bacterial DNA on the skin surface overrepresents the viable skin microbiome. *bioRxiv* 2021.08.16.455933 (2021) doi:10.1101/2021.08.16.455933.
34. Saheb Kashaf, S. *et al.* Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions. *Nat Microbiol* **7**, 169–179 (2022).
35. Swaney, M. H., Sandstrom, S. & Kalan, L. R. Cobamide sharing drives skin microbiome dynamics. *bioRxiv* 2020.12.02.407395 (2021) doi:10.1101/2020.12.02.407395.
36. van Dijk, L. R. *et al.* StrainGE: a toolkit to track and characterize low-abundance strains in complex microbial communities. *Genome Biol.* **23**, 74 (2022).
37. Zimmermann, M. & Fischbach, M. A. A family of pyrazinone natural products from a conserved nonribosomal peptide synthetase in Staphylococcus aureus. *Chem. Biol.* **17**, 925–930 (2010).
38. Tauch, A. *et al.* Complete genome sequence and analysis of the multiresistant nosocomial pathogen Corynebacterium jeikeium K411, a lipid-requiring bacterium of the human skin flora. *J. Bacteriol.* **187**, 4671–4682 (2005).
39. Pelz, A. *et al.* Structure and biosynthesis of staphyloxanthin from Staphylococcus aureus. *J. Biol. Chem.* **280**, 32493–32498 (2005).
40. Balibar, C. J., Shen, X. & Tao, J. The mevalonate pathway of Staphylococcus aureus. *J. Bacteriol.* **191**, 851–861 (2009).
41. Clauditz, A., Resch, A., Wieland, K.-P., Peschel, A. & Götz, F. Staphyloxanthin plays a role in the fitness of Staphylococcus aureus and its ability to cope with oxidative stress. *Infect.*

Immun. **74**, 4950–4953 (2006).

42. Seel, W. *et al.* Carotenoids are used as regulators for membrane fluidity by Staphylococcus xylosus. *Sci. Rep.* **10**, 330 (2020).

43. Madhaiyan, M., Wirth, J. S. & Saravanan, V. S. Phylogenomic analyses of the Staphylococcaceae family suggest the reclassification of five species within the genus Staphylococcus as heterotypic synonyms, the promotion of five subspecies to novel species, the taxonomic reassignment of five Staphylococcus species to Mammaliicoccus gen. nov., and the formal assignment of Nosocomiicoccus to the family Staphylococcaceae. *Int. J. Syst. Evol. Microbiol.* **70**, 5926–5936 (2020).

44. Dong, P.-T. *et al.* Photolysis of Staphyloxanthin in Methicillin-Resistant Staphylococcus aureus Potentiates Killing by Reactive Oxygen Species. *Adv. Sci.* **6**, 1900030 (2019).

45. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

46. Gande, R. *et al.* Acyl-CoA carboxylases (accD2 and accD3), together with a unique polyketide synthase (Cg-pks), are key to mycolic acid biosynthesis in Corynebacterianeae such as Corynebacterium glutamicum and Mycobacterium tuberculosis. *J. Biol. Chem.* **279**, 44847–44857 (2004).

47. Portevin, D. *et al.* A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 314–319 (2004).

48. Chen, Y. *et al.* Structural classification and properties of ketoacyl synthases. *Protein Sci.* **20**, 1659–1667 (2011).

49. Klaus, M., Buyachuihan, L. & Grininger, M. Ketosynthase Domain Constrains the Design of Polyketide Synthases. *ACS Chem. Biol.* **15**, 2422–2432 (2020).

50. Medema, M. H., Cimermancic, P., Sali, A., Takano, E. & Fischbach, M. A. A systematic

computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **10**, e1004016 (2014).

51. Palaniappan, K. *et al.* IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.* **48**, D422–D430 (2019).
52. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
53. Blin, K., Shaw, S., Kautsar, S. A., Medema, M. H. & Weber, T. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res.* **49**, D639–D643 (2020).
54. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
55. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *bioRxiv* 2022.02.14.480413 (2022) doi:10.1101/2022.02.14.480413.
56. Kloos, W. E. & Schleifer, K. H. Isolation and Characterization of Staphylococci from Human Skin II. Descriptions of Four New Species: Staphylococcus warneri, Staphylococcus capitis, Staphylococcus hominis, and Staphylococcus simulans. *Int. J. Syst. Bacteriol.* **25**, 62–79 (1975).
57. Becker, K., Heilmann, C. & Peters, G. Coagulase-negative staphylococci. *Clin. Microbiol. Rev.* **27**, 870–926 (2014).
58. Liu, C.-I. *et al.* A cholesterol biosynthesis inhibitor blocks Staphylococcus aureus virulence. *Science* **319**, 1391–1394 (2008).
59. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* **489**, 513–518 (2012).
60. Ghosh, S. & O'Connor, T. J. Beyond Paralogs: The Multiple Layers of Redundancy in Bacterial Pathogenesis. *Front. Cell. Infect. Microbiol.* **7**, 467 (2017).
61. Otto, M. Staphylococcus epidermidis--the 'accidental' pathogen. *Nat. Rev. Microbiol.* **7**,

555–567 (2009).

62. Brown, M. M. & Horswill, A. R. Staphylococcus epidermidis-Skin friend or foe? *PLoS Pathog.* **16**, e1009026 (2020).

63. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688–702.e13 (2020).

64. Melo, M. C. R., Maasch, J. R. M. A. & de la Fuente-Nunez, C. Accelerating antibiotic discovery through artificial intelligence. *Commun Biol* **4**, 1050 (2021).

65. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).

66. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

67. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab776.

68. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).

69. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

70. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 1–8 (2018).

71. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

72. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).

73. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo*. (2021). doi:10.5281/zenodo.5127899.
74. *seQuoia: A framework for diverse omics workflows, primarily focused on QC, processing and analysis of bacterial genomic sequencing sets*. (Github).
75. Krueger, F. *TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data*. (Github).
76. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
77. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
78. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
79. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
80. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
81. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* **1**, 1–6 (2016).
82. McDaniel, E. A., Anantharaman, K. & McMahon, K. D. metabolisHMM: Phylogenomic analysis for exploration of microbial phylogenies and metabolic pathways. *bioRxiv* 2019.12.20.884627 (2020) doi:10.1101/2019.12.20.884627.
83. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
84. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–12 (2006).
85. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large

phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

86. PareTree 1.0: Remove Sequences, Bootstraps, and Branch Lengths From Your Trees! http://emmahodcroft.com/PareTree.html.

87. Kania, S. A. Reclassification of Staphylococcus schleiferi by Madhaiyan et al. lacks key supporting data. *Int. J. Syst. Evol. Microbiol.* **72**, (2022).

88. Blin, K. *ncbi-genome-download: Scripts to download genomes from the NCBI FTP servers*. (Github).

89. Parks, D. *CompareM: A toolbox for comparative genomics*. (Github).

90. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).

91. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).

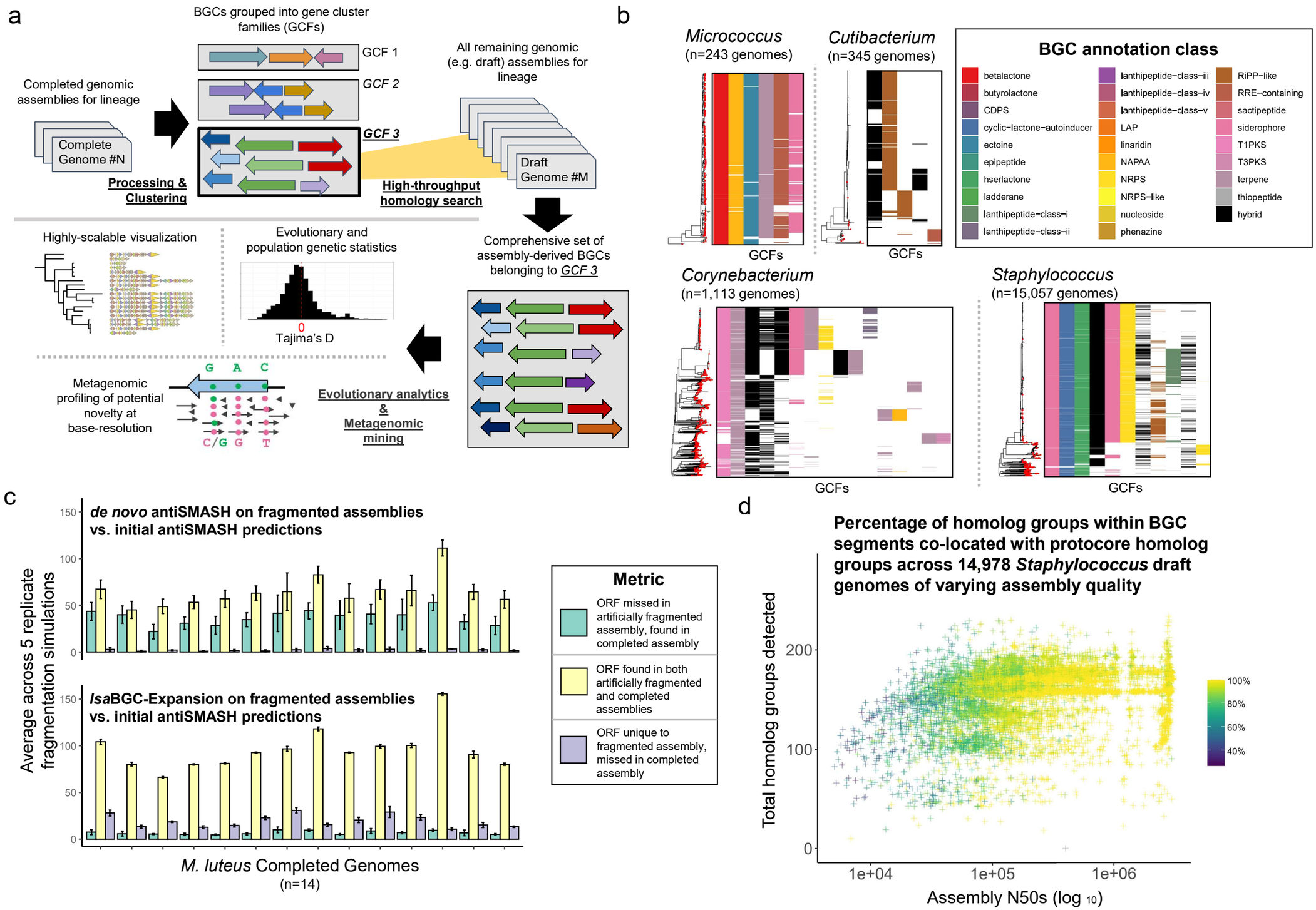

Figure 1: *lsa*BGC offers an efficient and sensitive means for GCF homology detection. a) A schematic overview of the full *lsa*BGC suite. **b**) Neighbor-joining trees were constructed for each of the four genera based on pairwise MASH ANI estimates between genomes and heatmaps showcase the presence of GCFs found in ≥ 5% of genomes. Red dots on neighbor-joining trees signify complete or chromosomal genomes used for initial identification of BGCs. Heatmap cell color indicates the BGC class as predicted by antiSMASH. **c**) Results from benchmarking *lsa*BGC-AutoExpansion sensitivity on 14 completed genomes for the species of *M. luteus*. AntiSMASH and *lsa*BGC-AutoExpansion were run on artificially fragmented genomes and resulting BGC predictions were compared to initial BGCs identified on the unfragmented, complete genomes. **d**) The relationship between the assembly quality measured by N50 and the total number of homolog groups detected across 14,978 *Staphylococcus* draft genomes is shown. The coloring corresponds to the percentage of BGC homolog groups co-located in a segment with a homolog group commonly found in protocore regions of BGC predictions by antiSMASH run on completed genomes.

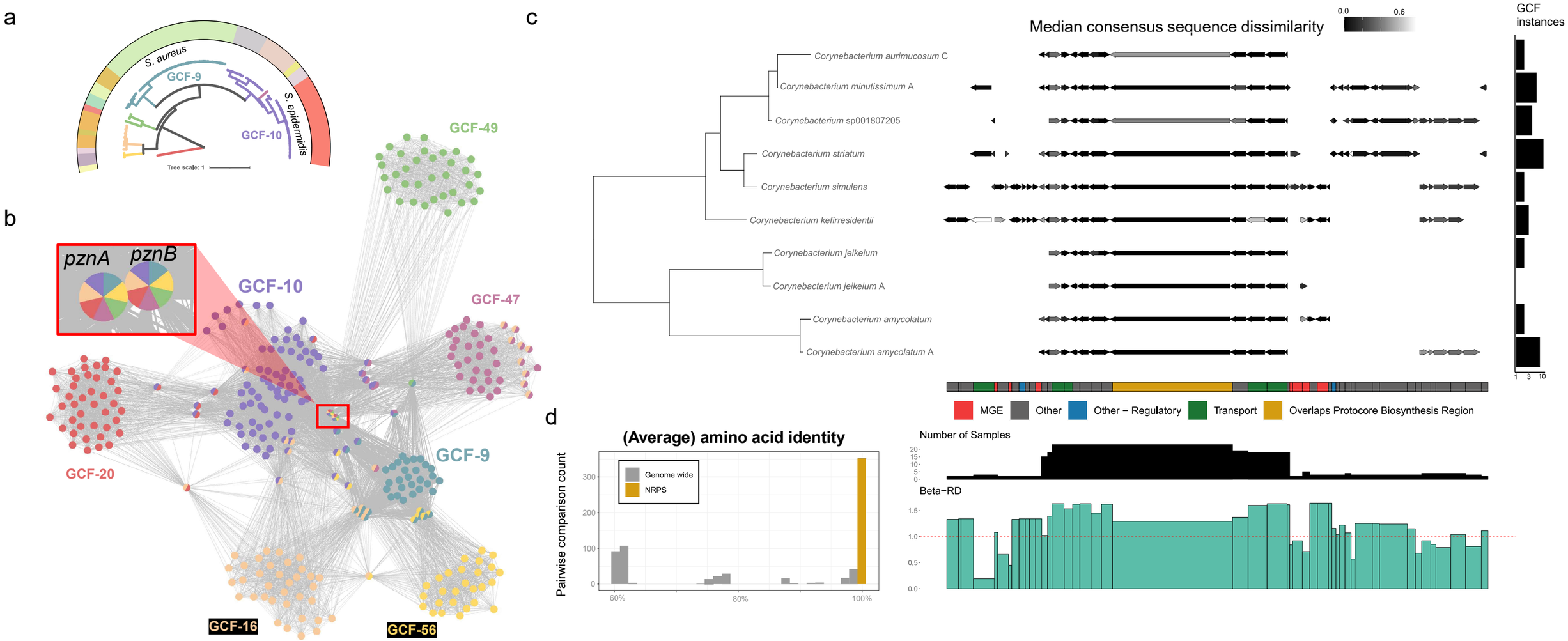

Figure 2: *Isa*BGC reveals new insight about characterized and unstudied BGCs from skin-associated species. a) A maximum-likelihood phylogeny was constructed from PznA and PznB showing the relationship between GCFs encoding for pyrazinones. Branch color corresponds to different GCFs and the outer circular color denotes the species classification for the genome from which the sequences were gathered. **b**) Homolog groups across the seven pyrazinone encoding GCFs are depicted as a network where edges indicate that two homolog groups co-occur in the same GCF. Homolog group color of nodes depict distinct GCFs. **c**) Phylogeny constructed from ribosomal proteins showing the conservation of *Corynebacterium* GCF-50 across the diverse species encoding the NRPS using results from *Isa*BGC-PopGene analysis. The gene schematics show the consensus gene order for the GCF. Gene lengths correspond to the median length across all BGC instances and only genes found in two or more genomes are shown. For each species, the gene is colored based on the median similarity of gene sequences from the species to the consensus sequence across all species. The barplot displays the total number of GCF instances identified per species across the full set of *Corynebacterium* genomes. Broad annotations for homolog groups are shown as a color strip. Below it are bar plots corresponding to the conservation of homolog groups across samples and their median Beta-RD. **d**) A histogram of the amino acid identities between NRP synthase instances is compared to the average amino acid identities (AAI) between genomes they were gathered from.

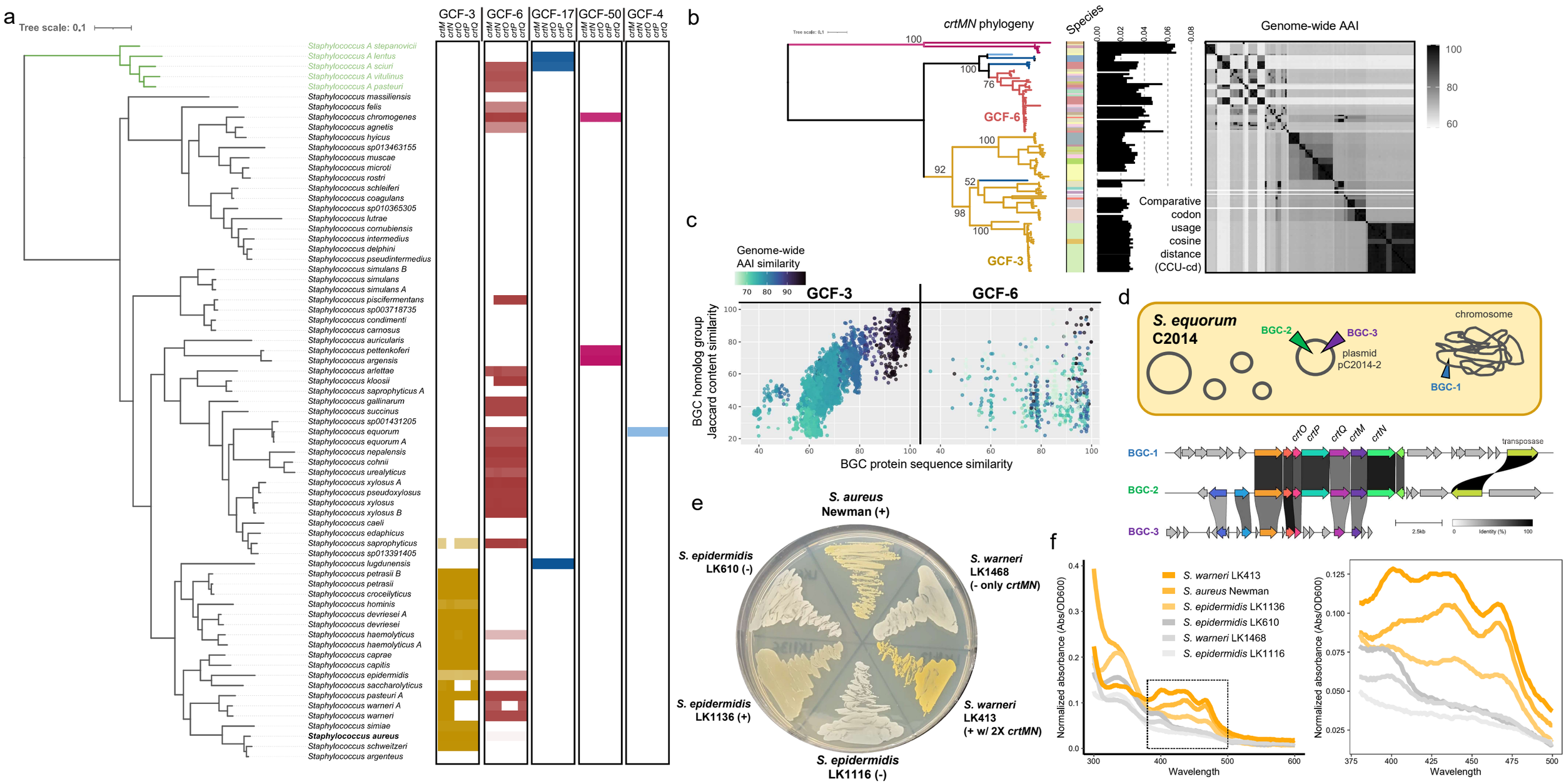

Figure 3: Staphyloxanthin is ubiquitous across *Staphylococcus* and found on plasmids. a) A maximum-likelihood phylogeny of the full *Staphylococcus* genus was constructed from ribosomal protein encoding genes. The clade in green represents the newly recognized genus of *Mammaliicoccus*. Heatmaps showcase the presence of the five *crt* genes for each species for five separate GCFs containing the *crt* operon. The shading indicates the proportion of a species found to carry homolog groups within a particular GCF context in $\log_{10}$ scale. **b**) A maximum-likelihood phylogeny was built from CrtM and CrtN protein alignments. Branch colors represent the GCF classification of the CrtMN sequences and bootstrap values are shown for key nodes where GCFs partition. The species the CrtMN sequence was extracted from is shown as a color strip followed by a bar plot depicting the comparative codon usage cosine distance (CCU-cd). CCU-cd was only calculated for the full five gene *crt* operon and represents the codon frequency dissimilarity with the codon frequency of the background genome. The heatmap showcases the genome-wide average amino acid identity (AAI) between pairs of genomes from which CrtMN sequences were gathered. **c**) AAI and shared homolog group content were calculated between pairs of GCF instances from different genomes for GCF-3 and GCF-6 individually. The coloring represents genome-wide AAI similarity. **d**) A schematic of the completed genome for *S. equorum* C2014, which was predicted to feature GCF-6 in three distinct copies. Bottom panel shows the BGC alignment between the three instances of GCF-6 found within this genome. **e**) Representative isolates with and without staphyloxanthin encoding GCFs were grown on 20% BHI agar to visualize pigmentation production. **f**) Methanol extractions and wavelength absorption analysis were performed to identify signature peaks in the 400 to 500 nm range associated with the presence of staphyloxanthin in *S. aureus*[39,44]. Spectra are normalized to cell density.

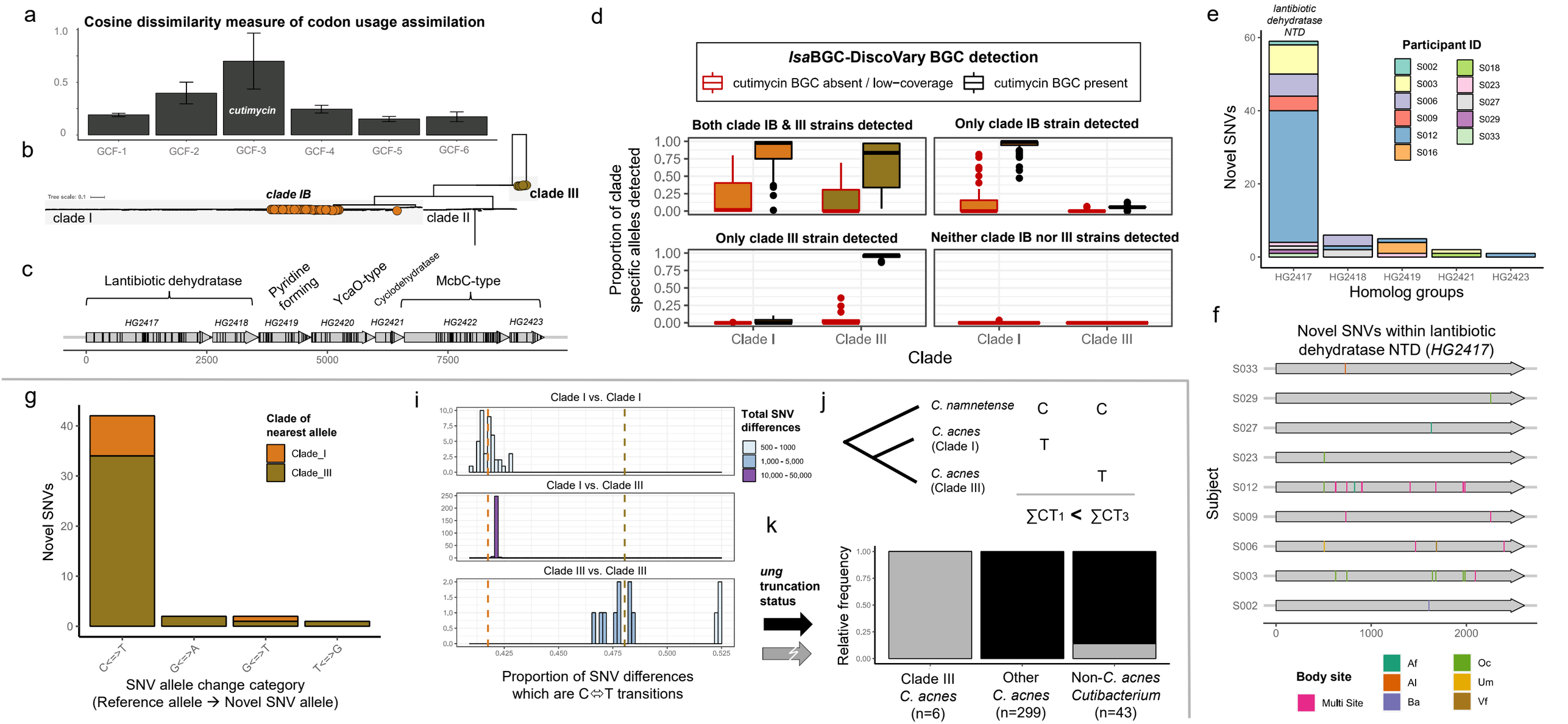

Figure 4: Mining for microdiversity within the cutimycin BGC leads to genetic insights for clade III *C. acnes*. **a**) Codon usage dissimilarity between BGC regions and background genomes shown for six common GCFs in *C. acnes*. **b**) The cutimycin encoding GCF is found primarily in clade I (subclade IB; orange) and clade III (tan) *C. acnes*. **c**) A schematic of the cutimycin encoding GCF. Lines in black correspond to sites of nearly-fixed differences between clades of *C. acnes*. **d**) *lsa*BGC-DiscoVary was used to identify the presence of the cutimycin GCF in 270 skin metagenomes. The proportion of alleles representative of clade IB or clade III *C. acnes* at sites of nearly fixed-differences was quantified. High-resolution strain quantification within metagenomes was similarly performed with StrainGE and used to overlay clade-specific typing results from *lsa*BGC-DiscoVary analysis. **e**) The distribution of novel SNVs identified by *lsa*BGC-DiscoVary within metagenomes, post filtering for removal of false positives, is shown across homolog groups and subjects. **f**) The location of novel SNVs is shown along the lantibiotic dehydratase N-terminal encoding gene (*HG2417*) for each subject. Coloring of the novel SNVs corresponds to which body-site(s) they were identified within. **g**) The clade distribution of reference genes on which novel SNVs were identified across the cutimycin GCF for subject S012 are shown individually for the predicted transitions or transversions. **i**) The proportion of differences which corresponded to cytosine to thymine or thymine to cytosine transitions along a core-genome alignment was calculated between pairs of *C. acnes* and categorized by clade comparisons. **j**) Comparative analysis of presumed directional deaminations from *C. namnetense*, treated as ancestral, to *C. acnes* clade IB and clade III were estimated using a multi-iteration sub-sampling based approach for sites along the core-genome alignment, whereby a greater number of suspected C to T transitions were estimated for clade III *C. acnes*. **k**) The presence of a complete uracil DNA glycosylase (*ung*), from clade I *C. acnes*, is shown across available *Cutibacterium* genomes.

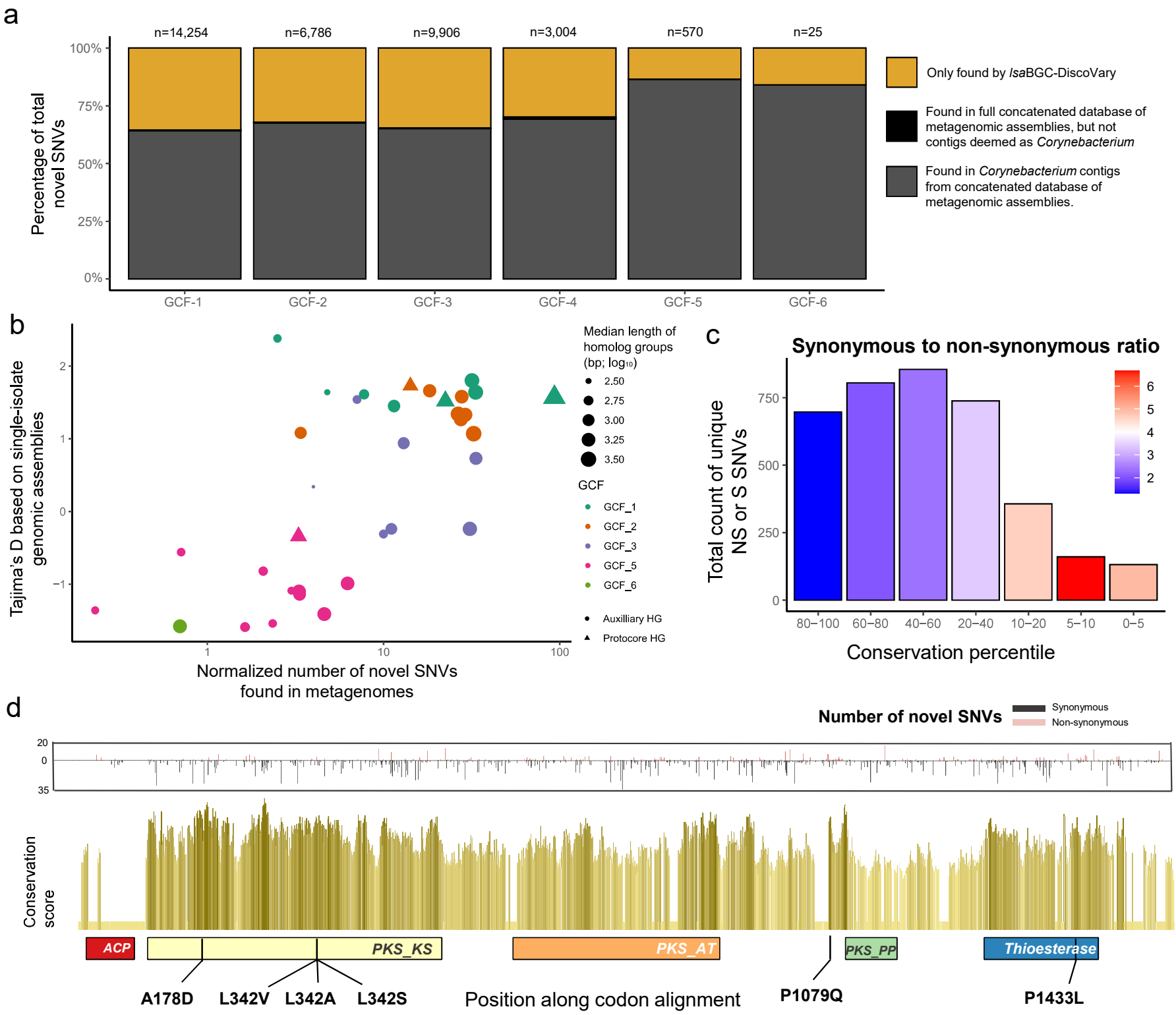

Figure 5: Thousands of novel SNVs identified within BGCs from the *C. kefirresidentii* species complex. a) Assessment of novel SNVs identified by *Isa*BGC-DiscoVary, post-filtering, for each BGC from the *C. kefirresidentii* species complex for presence within metagenomic assemblies. **b**) Tajima's D statistic, calculated for homolog groups using sequences in available genomic assemblies, is compared to the normalized number of novel SNVs found per homolog groups within metagenomes. **c**) The number of novel SNVs is shown for ranges of conservation percentile across all homolog groups and metagenomes. The coloring corresponds to the ratio of synonymous to non-synonymous novel SNVs within each conservation range. **d**) Novel SNVs are shown for the polyketide synthase involved in mycolic acid biosynthesis. The top panel showcases the non-synonymous (up-facing) and synonymous (down-facing) novel SNVs found along the T1PKS gene, with heights corresponding to the number of metagenomes with the novel SNV. The middle panel shows the conservation scores colored by percentile ranges (lightest = least conserved sites; darkest = most conserved sites). Key domains common to T1PKS are shown underneath along with the location of predicted non-synonymous mutations within the top 5% of conserved sites for the gene. Two adjacent SNVs affecting the coding of the residue at position 342 were identified and these could result in a combined effect of encoding for alanine.