AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

Genome-Guided Discovery of Natural Products through Multiplexed Low-Coverage Whole-Genome Sequencing of Soil Actinomycetes on Oxford Nanopore Flongle

Rahim Rajwani,[a] Shannon I. Ohlemacher,[a] Gengxiang Zhao,[a] Hong-Bing Liu,[a] Carole A. Bewley[a]

[a]Laboratory of Bioorganic Chemistry, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, USA

ABSTRACT Genome mining is an important tool for discovery of new natural products; however, the number of publicly available genomes for natural product-rich microbes such as actinomycetes, relative to human pathogens with smaller genomes, is small. To obtain contiguous DNA assemblies and identify large (ca. 10 to greater than 100 kb) biosynthetic gene clusters (BGCs) with high GC (>70%) and high-repeat content, it is necessary to use long-read sequencing methods when sequencing actinomycete genomes. One of the hurdles to long-read sequencing is the higher cost. In the current study, we assessed Flongle, a recently launched platform by Oxford Nanopore Technologies, as a low-cost DNA sequencing option to obtain contiguous DNA assemblies and analyze BGCs. To make the workflow more cost-effective, we multiplexed up to four samples in a single Flongle sequencing experiment while expecting low-sequencing coverage per sample. We hypothesized that contiguous DNA assemblies might enable analysis of BGCs even at low sequencing depth. To assess the value of these assemblies, we collected high-resolution mass spectrometry data and conducted a multi-omics analysis to connect BGCs to secondary metabolites. In total, we assembled genomes for 20 distinct strains across seven sequencing experiments. In each experiment, 50% of the bases were in reads longer than 10 kb, which facilitated the assembly of reads into contigs with an average N_{50} value of 3.5 Mb. The programs antiSMASH and PRISM predicted 629 and 295 BGCs, respectively. We connected BGCs to metabolites for *N,N*-dimethyl cyclic-di-tryptophan, two novel lasso peptides, and three known actinomycete-associated siderophores, namely, mirubactin, heterobactin, and salinichelin.

IMPORTANCE Short-read sequencing of GC-rich genomes such as those from actinomycetes results in a fragmented genome assembly and truncated biosynthetic gene clusters (often 10 to >100 kb long), which hinders our ability to understand the biosynthetic potential of a given strain and predict the molecules that can be produced. The current study demonstrates that contiguous DNA assemblies, suitable for analysis of BGCs, can be obtained through low-coverage, multiplexed sequencing on Flongle, which provides a new low-cost workflow ($30 to 40 per strain) for sequencing actinomycete strain libraries.

KEYWORDS actinomycetes, bioinformatics, biosynthetic gene cluster, cyclic dipeptide, genomics, glycopeptide, lasso peptide, mass spectrometry, metabolomics, multi-omics, natural products

Clinical pathogens are increasingly becoming resistant to currently used antimicrobials, causing over 700,000 deaths worldwide (1). New antimicrobials are urgently needed to alleviate antimicrobial resistance and prevent deaths per year from rising to over 10 million by 2050 (1). One of the prolific sources of new antimicrobials is a group of Gram-positive mycelium-forming bacteria, the actinomycetes. Several currently used antibiotics, including vancomycin, rifamycin, and streptomycin, are isolated from

Editor Gilles P. van Wezel, Leiden University

Address correspondence to Carole A. Bewley, caroleb@nih.gov.

Received 10 August 2021
Accepted 31 October 2021
Published 23 November 2021

actinomycetes, and they still hold enormous potential for the future discovery of new medicines (2).

Genome sequencing is now an important component of natural products research. Whole-genome sequencing (WGS) enables identification of the genes responsible for the biosynthesis of natural products (3). Often genes required for the biosynthesis of a natural product positionally cluster on the genome and are referred to as biosynthetic gene clusters (BGCs) (4). The BGC sequences can be used to predict possible structures of the resulting natural product (5), assess novelty of the compound (6), and dereplicate compounds from a strain collection (7). Despite the advantages offered by WGS, the number of actinomycete genomes remains limited. Several rare genera are not represented by a complete genome, and the majority of currently available genomes are sequenced using Illumina short-read technology that results in highly fragmented assemblies (see Fig. S1 in the supplemental material). BGCs span multiple contigs in fragmented genome assemblies and cannot be detected or analyzed by commonly used BGC prediction tools such as antiSMASH (8, 9).

Long-read sequencing technologies (e.g., PacBio or Oxford Nanopore Technologies [ONT]) produce contiguous genome sequences needed to analyze secondary metabolite gene clusters. Notably, PacBio assemblies achieve consensus accuracy of over 99.995% (10); however, the method is generally less accessible due to the upfront cost of sequencing instruments and higher per-sample sequencing costs. In contrast, ONT does not require an upfront cost of an expensive sequencing instrument and the devices are inexpensive. Nevertheless, ONT data result in a lower consensus accuracy (99.9%) (11) and often require polishing with Illumina reads to obtain reference-quality genomes. We hypothesized that while BGC identification requires a contiguous DNA sequence, it might be less affected by the lower consensus accuracy of a Nanopore assembly since most BGC analysis steps involve inferring homology between distantly related amino acid sequences using profile hidden Markov models. If this is true, contiguous DNA assemblies can be obtained at ca. 10× coverage using ONT, allowing complete genome sequencing at a significantly lower cost. While such ONT-sequenced genomes would still require error correction with Illumina reads, they could be used on their own to sequence a strain collection, build a catalog, and compare BGCs for dereplication or identification of potentially new compounds, which might be particularly useful to natural product research and drug discovery programs.

To assess the feasibility of obtaining contiguous assemblies from ca. 10× sequencing depth, predicting BGCs, and connecting BGCs to metabolites, we conducted the current multi-omics study. We sequenced 20 new soil-derived actinomycete strains and analyzed their metabolome using high-resolution mass spectrometry (HRMS). For sequencing, we specifically selected Flongle, a recently launched ONT sequencing device that costs $90 USD and can generate up to 1 to 2 Gb of sequence output. With a typical actinomycete genome being 8 to 10 Mb, a single Flongle experiment might be sufficient to sequence 3 to 4 strains at 20- to 30× coverage. Sequencing workflows based on Flongle could be broadly applicable to small and large studies due to the modular experimental design. In the current study, we obtained 300 to 850 Mb of data per experiment across 10 sequencing experiments with read-length N_{50} values over 10 kb. Assembling of reads resulted in contiguous assemblies (average contig N_{50} value = 3.5 Mb and average number of contigs = 47.3). AntiSMASH5 predicted a total of 629 BGCs from these assemblies. Through a combined analysis with metabolomics data, we were able to connect six BGCs to their secondary metabolites. The study demonstrates the utility of low-coverage Nanopore-only assemblies as a rapid and low-cost sequencing option to advance natural product research.

RESULTS

An *in silico* analysis to study the effect of sequencing coverage and read length on BGC detection. We first analyzed the level of sequencing coverage that would be sufficient for contiguous assemblies and BGC detection using Oxford Nanopore sequencing. For this purpose, three actinomycete genomes, previously sequenced at

high coverage, were downloaded from the European Nucleotide Archive, and assuming a genome size of 8 Mb, their reads were downsampled to 60×, 30×, 15×, and 7× coverage (assuming a genome size of 8 Mb) before assembling and detecting BGCs (see Table S1 at https://doi.org/10.6084/m9.figshare.16722961). These actinomycete strains were from the genera *Streptomyces* (genomes 1 and 3) and *Nocardia* (genome 2) with respective genome sizes of 6.8 Mb, 12.3 Mb, and 10.6 Mb (see Table S1 at https://doi.org/10.6084/m9.figshare.16722961). While their genome sizes were different (see Table S1 at https://doi.org/10.6084/m9.figshare.16722961), an assumption of a fixed expected genome size of 8 Mb allowed us to determine the utility of a prospective sequencing experiment where actinomycete genome sizes would not be known. In the downsampling analysis, we observed that actual mapped coverage for genomes 2 and 3 was approximately 50% of the expected coverage due to their larger genome sizes (see Fig. S2 in the supplemental material and also Table S2 at https://doi.org/10.6084/m9.figshare.16722934.v1). For the assemblies obtained from each genome at the tested coverages, we assessed quality from total assembly size, number of contigs, consensus accuracy, and number of BGCs detected by antiSMASH and PRISM (Fig. 1). We observed that assembly size and number of predicted genes nearly plateau at ca. 15× mapped coverage (Fig. 1). At ca. 15- to 20× mapped coverage, the assemblies reached 99% consensus accuracy, and further increases in coverage led only to minor improvements in consensus accuracy (Fig. 1). A sharp decline in the number of contigs was also noted at ca. 15- to 20× mapped coverage (Fig. 1). At 12× mapped coverage (15× estimated coverage), 15 out of 22 (72%) antiSMASH-predicted BGCs were detected in genome 1 (Fig. 1). Similarly, for genome 3 (10.6-Mb size), 36 BGCs were detected at 7× mapped coverage (or 15× estimated) (Fig. 1). The largest of the three genomes (genome 2, 12.3-Mb size) required 30× estimated coverage to reach 11× mapped coverage and detect 24 out of 29 BGCs (82%) (Fig. 1). Except for the 12.3-Mb genome 2, the majority of the BGCs were detected at ca. 15- to 20× mapped coverage and mapped coverage above 30× led to only 1 to 6 additional BGCs (Fig. 1). We also assessed whether these BGCs were located on the edge of a contig, which could result in fragmented or incomplete identification (Fig. S2). For the smallest genome (genome 1, 6.8-Mb size), none of the BGCs were on a contig edge at 12× mapped coverage (Fig. S2). For genome 3, 7× mapped coverage (or 15× estimated coverage) led to 21 out of 36 BGCs (58%) fragmented or located on the contig edge (Fig. S2). Genome 3 required 30× estimated coverage to reach approximately 20× mapped coverage and assemble 80% of the BGCs not on a contig edge (Fig. S2). For genome 2 as well, an estimated 30× coverage (11× mapped coverage) was needed for complete assembly of 99% of the BGCs (22 out of 23) (Fig. S2). Relative to antiSMASH, PRISM predicted fewer BGCs (Fig. 1). Nevertheless, the trend relative to coverage was similar between antiSMASH and PRISM (Fig. 1).

Assembly contiguity and, therefore, BGC detection in a Nanopore sequencing experiment are also related to read length. In another computational experiment, we assessed which read length might be suitable for a contiguous DNA assembly and robust BGC detection at low sequencing coverage. For this purpose, simulated Nanopore reads of average lengths of 500, 1,000, 2,000, 4,000, 8,000, and 16,000 nucleotides (nt) were generated at 10× coverage of a *Streptomyces* sp. genome (GB4-14) using BadRead (11). The resulting reads were assembled and analyzed for assembly contiguity and BGC detection. We observed that an approximately 2-fold increase in average read length was associated with a 2-fold reduction in the number of contigs (Fig. S3). Short reads (up to 2,000 nt) led to a highly fragmented assembly with >100 contigs (Fig. S3) and >50% of BGCs on the edge of a contig (Fig. S4). With an average read length of 4,000 nt and 8,000 nt, the number of contigs declined to 55 and 22, respectively (Fig. S3), and the proportion of BGCs on a contig edge decreased to 38% and 16%, respectively (Fig. S4). Finally, the most contiguous assembly was obtained with <10 contigs and no BGCs on a contig edge using an average read length of 16,000 nt (Fig. S3 and S4).

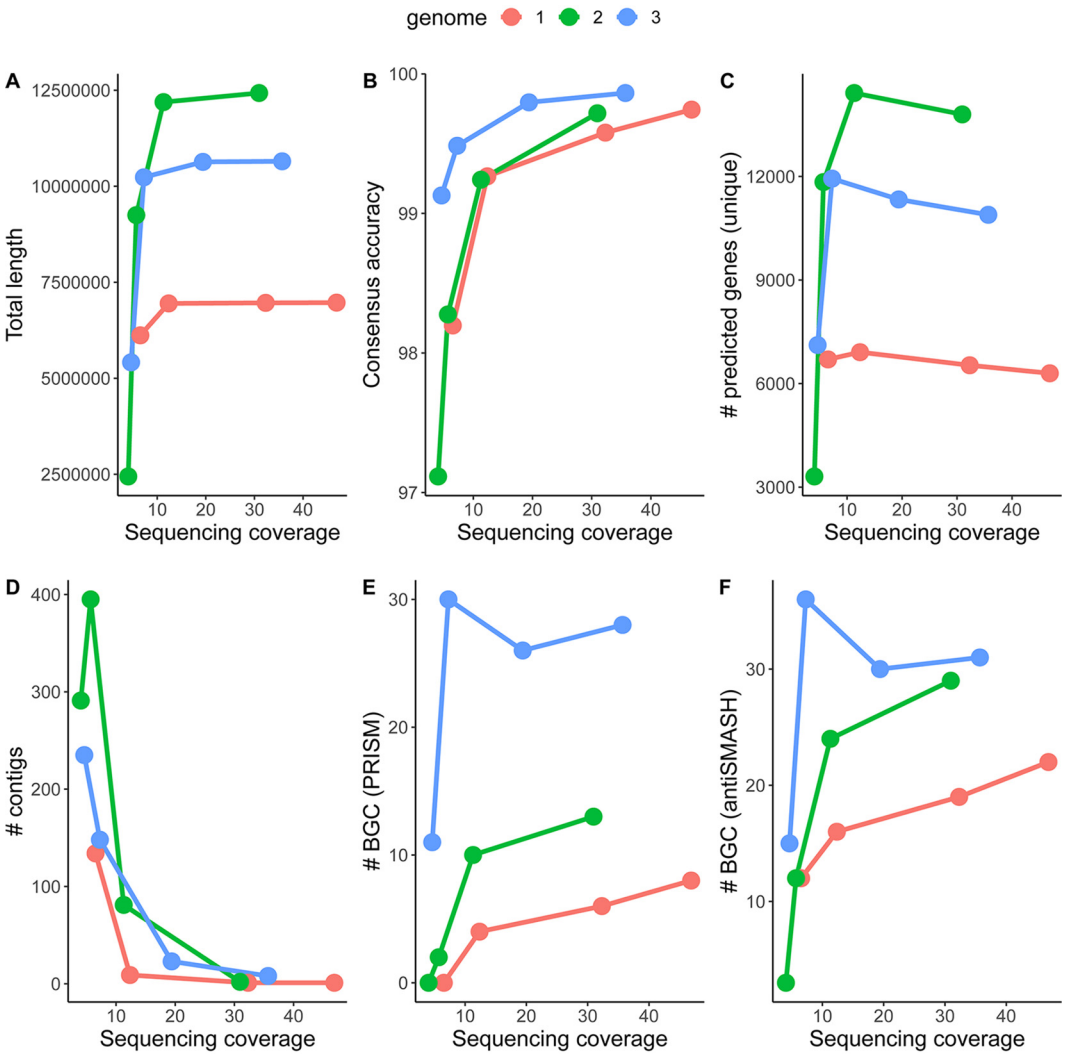

FIG 1 An *in silico* analysis of the effect of sequencing coverage on assembly quality and BGC prediction. Three genomes previously sequenced at high coverage on the ONT MinION or GridION platform were downsampled to the reduced coverage indicated and then assembled. The sequencing coverage on the *x* axis is the actual mapped coverage. (A to F) The quality of the assembly as indicated by total assembly length (A), number of mismatches per 100 kb shown as consensus accuracy (B), the number of unique predicted genes (C), the number of contigs (D), and the number of BGCs predicted by PRISM (E) and antiSMASH 5.0 (F). The sequencing data for this analysis were downloaded from the European Nucleotide Archive. Accession numbers are as follows: 1, SRR10597857; 2, SRR9710049; 3, DRR240480. The source data for the figure are provided in Table S2 at https://doi.org/10.6084/m9.figshare.16722934.v1.

Overall, these computational experiments suggested contiguous DNA assemblies can be obtained and complete BGCs can be detected at low sequencing coverage using long reads from Oxford Nanopore Technologies; this should allow for a dramatic reduction in cost per genome through multiplexing. The computational experiments were followed up with prospective sequencing of actinomycete genomes using Flongle, and more detailed analyses of their BGCs are described below.

Nanopore sequencing, genome assembly, and quality assessment. A total of 10 sequencing experiments were conducted—each attempting to sequence four actinomycete strains (see Table S3 at https://doi.org/10.6084/m9.figshare.16722889). Three experiments resulted in <100 Mb sequenced (Fig. 2). These experiments were considered unsuccessful and were not processed for genome assembly and BGC detection. The failure of these experiments could be attributed to impurities in the starting genomic DNA, as measured by a ratio of the UV absorbance at 260 and 280 nm (i.e., flow cell AET812 in Fig. 2), and low pore occupancy caused by insufficient loading of

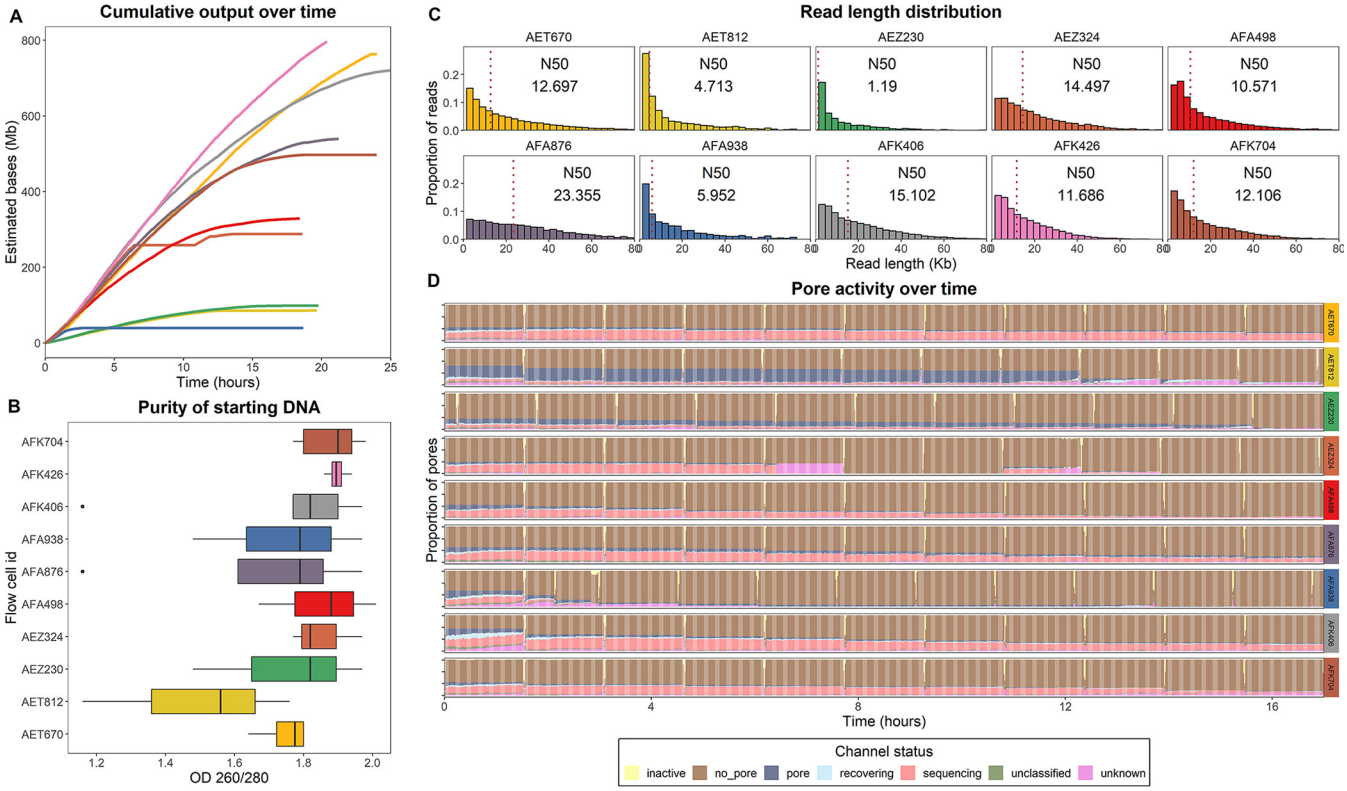

FIG 2 Library preparation and sequencing quality metrics. The data in panels A to D are colored and grouped by flow cell as indicated in panel B. (A) Cumulative output (estimated megabases) over time (hours) for 10 sequencing experiments. (B) DNA purity (as determined by A_{260}/A_{280} ratio) of the samples run in each experiment. (C) Read length distribution across experiments. The read N_{50} value (50% of bases are in reads longer than this value) for each experiment is labeled. (D) Performance of the flow cell at the pore level for all experiments except AFK4226 as indicated by the pore activity level. The AFK4226 experiment was interrupted at the end, and the instrument did not generate the pore activity metadata to include in this chart. "Sequencing" indicates that the pore is occupied with DNA and is sequencing. "Pore" indicates an empty pore with no DNA; "no pore" indicates an inactive pore that is unavailable for sequencing.

the library or inhibition of adapter ligation (i.e., AET812 and AEZ230 shown in Fig. 2). The remaining seven experiments yielded 288 to 797 Mb over 18 to 24 h. The longest read for each sample included in a multiplexed experiment was over 80 kb.

Enriching for long DNA fragments in a long-read library preparation is an important step in obtaining sequencing data that we require for assembly and BGC analysis. DNA fragment size selection is commonly applied using either an automated agarose gel electrophoresis system or solid-phase reversible immobilization (SPRI) beads. The gel-based size selection requires expensive equipment, whereas SPRI beads are economical and commonly used during cleanup procedures after enzymatic reactions in library preparations. Using commercial formulations of SPRI beads, a size cutoff of at most 400 bp is possible by reducing the bead-to-DNA ratio (12). Recent studies have shown that it might be possible to set the base pair cutoff higher, in the multikilobase range, by adjusting the composition of polyethylene glycol (PEG) and cations in the buffer (12). In the current study, we tried different previously reported buffers (12) for bead-based purification to apply size selection and increase the read length N_{50} values over those obtained using standard protocols (Fig. 2 and see Table S3 at https://doi.org/10.6084/m9.figshare.16722889). One of our initial experiments using a 0.5× bead-to-DNA ratio of the standard buffer concentration was not successful, resulting in read length N_{50} values of 1 to 1.6 kb for three out of five samples sequenced in the experiment (AET812). In two subsequent successful experiments (AET670 and AFK704), we used a 0.15× bead-to-DNA ratio of a modified buffer containing 0.5 M $MgCl_2$ plus 5% PEG in TE buffer (10 mM Tris-Cl, pH 8.0, 1 mM EDTA) for bead-based purification after barcode ligation as described previously (12). Read length N_{50} values in these experiments were 11.6 to 15.1 kb (Fig. 2). In three experiments (AEZ324, AFA498, and AFA876), the concentration of the

modified buffer for bead-based size selection was reduced to a 0.1× bead-to-DNA ratio, which led to a further increase in read length N_{50}s (10.5 to 23.3 kb) but was accompanied by increased sample loss. Application of size selection after barcode ligation ensures approximately equal fragment lengths for pooling of samples and adapter ligation. However, ligation of barcodes to longer fragments could be less efficient if shorter fragments are present in the mixture. Therefore, we tested bead-based size selection (0.1× bead-to-DNA ratio in the modified buffer) before barcode ligation in later experiments (flow cell identifiers [IDs] AFK426 and AFK406) (see Table S3 at https://doi.org/10.6084/m9.figshare.16722889). A more consistent output was observed in these experiments, possibly due to more efficient barcode/adapter ligation to longer DNA fragments.

Across the seven successful runs, 3,814,434,062 bases in 751,459 reads were generated. Upon demultiplexing, the median number of bases per sample was 77.5 Mb (theoretical coverage of 9.5× with an expected 8-Mb genome size). Three strains were sequenced at <2.5× theoretical coverage (<20 Mb per strain) and were excluded from further analysis. Subsequently, 25 samples (20 distinct isolates) were *de novo* assembled with Canu (13) and polished with Racon (14) and medaka (Fig. 3). The median length of the obtained assemblies was 8.5 Mb (average, 7.9 Mb; maximum, 9.4 Mb), typical of actinomycete genome size. The only exception was a 3-Mb assembly for GB8-002, which was also sequenced at the lowest coverage (4.0×) (Fig. 3).

We assessed the accuracy and quality of these low-coverage genomes by comparing them with genomes sequenced at high coverage on MinION or PacBio. Two strains, GA3-008 and GB4-14, were previously sequenced by our lab at 10-fold-higher coverage using MinION and PacBio, respectively (Table 1). Despite the lower coverage obtained with Flongle, the genome contiguity was only slightly affected, and both genomes were assembled into <10 contigs. The size of the assemblies differed by 6.1 kb (GA3-008) and 19.6 kb (GB4-14) due to insertion/deletion (indel) errors. Despite many mismatches and insertion and deletion errors, 87.5% and 100% of the respective BGCs detected in the MinION and PacBio assemblies were also detected in these Flongle assemblies by antiSMASH.

Taxonomy and BGCs. AntiSMASH 5.0 predicted 629 BGCs of 29 different types across all assembled genomes from this study (Fig. S5). Seventy-nine percent (497/629) of these BGCs were not located on a contig edge, suggesting their full-length or complete recovery from sequencing data. There was a median of 23 BGCs per strain. In addition to antiSMASH, 295 BGCs were predicted using PRISM software. A unique feature of PRISM is that it enables chemical structure prediction from BGCs (15). In the current data set, PRISM generated a predicted chemical structure for 180 out of 295 predicted BGCs.

The taxonomic identification based on 16S rRNA sequences extracted from the whole genomes revealed that the data set comprises 11 different actinomycete species belonging to four genera (see Table S4 at https://doi.org/10.6084/m9.figshare.16722952). It consisted of eight *Amycolatopsis*, nine *Streptomyces*, four *Lentzea*, and four *Nocardia* species. Some species were overrepresented in the data set. For example, *Amycolatopsis lurida* and *Streptomyces tendae* were each represented with four distinct strains and *Lentzea violacea* with two distinct strains (see Table S4 at https://doi.org/10.6084/m9.figshare.16722952). The biosynthetic diversity between strains was high with members of the same species (defined as sharing >99% identity in their 16S sequences) differing by up to 20 BGCs, depending on the species (Fig. S6). Strains of *Streptomyces tendae*, *Lentzea violacea*, or *Streptomyces kanamyceticus* were more diverse within the species than strains of *Amycolatopsis lurida*. Different species of the same genus also contained 10 to 15 different BGCs on average.

Predicting metabolites from BGCs—paired analysis of genome and secondary metabolites. (i) *In silico* PRISM-predicted chemical structures. The PRISM-predicted BGCs detected from low-coverage assemblies were analyzed further to determine whether they could be linked to a metabolite. We conducted a paired analysis by collecting tandem mass spectrometry (MS/MS) spectra of extracts from strain cultures grown in ISP1 and R2A media. MS/MS spectra were queried using molDiscovery

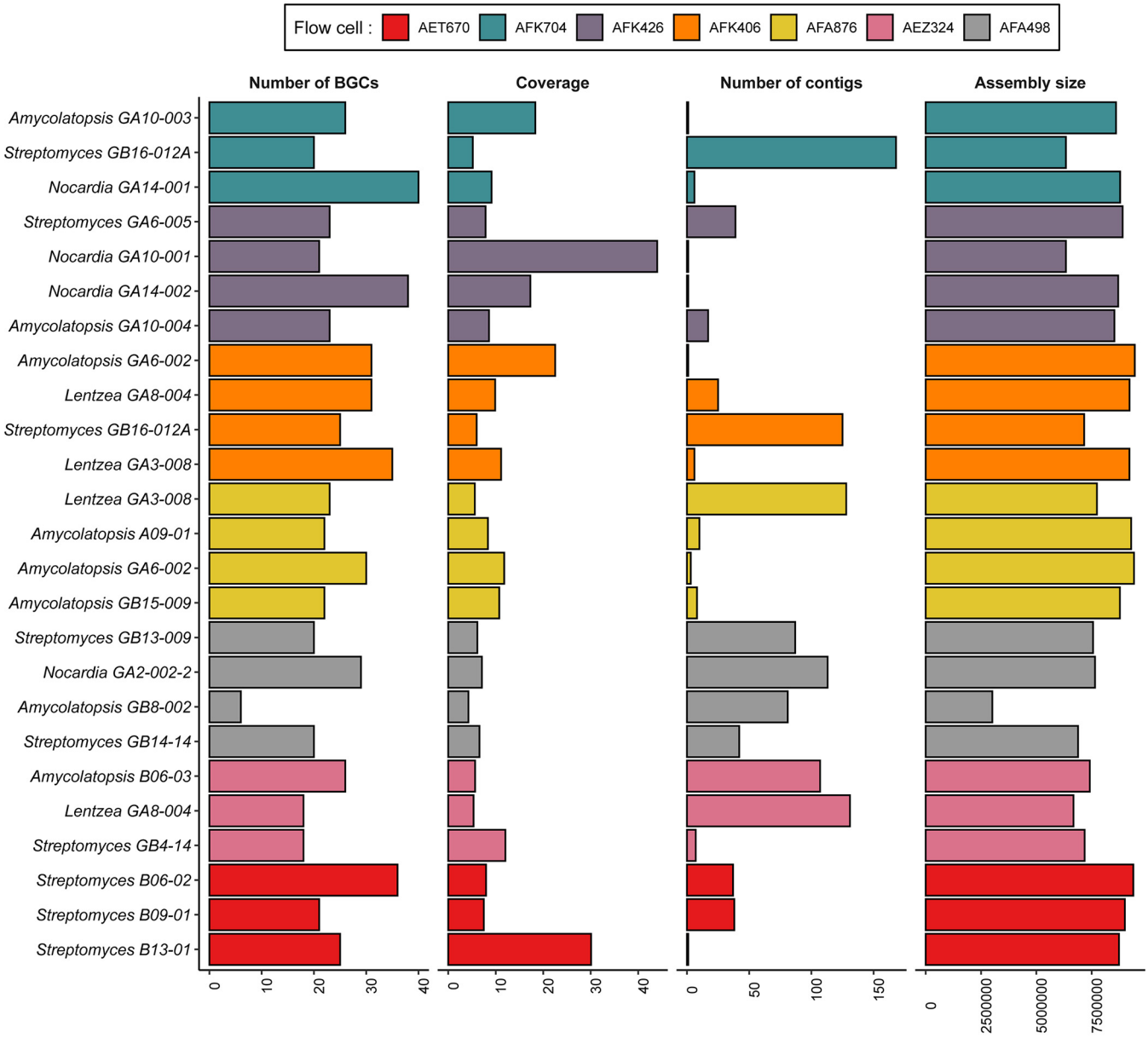

FIG 3 Characteristics of the genome assemblies obtained through low-pass multiplexed sequencing on Flongle. Each bar represents a sample. Bars are grouped and colored by flow cells (individual sequencing experiment).

against a database of PRISM-predicted chemical structures from the BGCs (16). We observed that three predicted structures matched MS/MS spectra at a false-discovery rate (FDR) of $<$1% and a *P* value of $<e^{-10}$.

Two of the matched structures were both predicted from a single cyclic dipeptide BGC in *Amycolatopsis* sp. strain GA6-002 (Fig. 4). The two-gene BGC encoded a cyclic dipeptide synthase (CDPS) and an *N*-methyltransferase. Both of the aminoacyl-tRNA binding pockets of the CDPS had a specificity signature for tryptophan tRNA. Based on this information, cyclic-di-tryptophan (cWW) and its *N*-methylated derivate (cWW-Me2) were predicted as possible metabolites (17). In all extracts from GA6-002, a metabolite matching the precursor *m/z* and predicted MS/MS fragmentation pattern for cWW was detected. In addition, the *N*-methylated metabolite cWW-Me2 was detected in the ethyl acetate extracts from ISP1 cultures.

The third mass spectral match for a small, glycosylated polyketide consisting of a propionate unit and an actinosamine sugar moiety was detected in the extract of strain

TABLE 1 Quality assessment of genomes of two strains (GA3-008 and GB4-14) obtained from low-coverage Flongle assemblies, compared to PacBio and MinION assemblies

	GA3-008		GB4-14	
Attribute	**Flongle multiplex**	**MinION singleplex**	**Flongle multiplex**	**PacBio**
Assembly size (nt)	9,205,503	9,199,325	7,183,038	7,163,416
Estimated coverage	15.3	121.65	14	200
No. of contigs	6	1	7	2
No. of mismatches[a]	3,000	–[b]	19,227	–
No. of insertions/deletions[c]	13,803	–	14,484	–
Consensus accuracy (%)[c]	99.79	–	99.48	–
No. of BGCs	35	40	18	18

[a]Number of mismatches and insertions and deletions in the Flongle multiplexed assemblies relative to the same strain sequenced by either MinION or PacBio.
[b]–, not applicable.
[c]Number of mismatches, insertions, and deletions and consensus accuracy in the Flongle multiplexed assemblies relative to the same strain sequenced by either MinION or PacBio.

GB4-14 (Fig. S7), and this structure was predicted by PRISM to be a shunt product of a larger polyketide BGC. Interestingly, PRISM was unable to predict the structure of a larger polyketide that would be the predicted product of all of the modules in this polyketide synthase (PKS). More complex structures that better resemble final products of PKS pathways were predicted from a more contiguous Flongle (Fig. 3) or PacBio assembly (Table 1) of the same strain, but they were not detected in the natural product extracts. The final product of this BGC, predicted from the PacBio-sequenced genome, was therefore regarded as not detected.

(ii) RiPPs. We conducted a second analysis to query MS/MS spectra for products of Ribosomally synthesized and Posttranslationally modified Peptide (RiPP) BGCs using MetaMiner (18). All open reading frames (ORFs) shorter than 600 nt were extracted from 43 antiSMASH-predicted RiPP BGCs [16 lanthipeptide, 4 Linear azol(in)e-containing peptide, 13 lasso peptides, 10 thiopeptides] and included in this analysis (Fig. S5). We observed a single high-confidence match for a class II lasso peptide BGC in the *Amycolatopsis* sp. strain GA6-002 (Fig. 5). The BGC encoded all essential elements for lasso peptide biosynthesis including precursor peptide, asparagine synthetase (SMCOG1177—essential for macrolactam formation), lasso peptide transglutamase protease (PF13471—leader peptide cleavage), RiPP recognition element (PF05402), and ABC transporter (SMCOG1288 and SMCOG1000) (Fig. 5) (19, 20). The precursor *m/z* ($[M + 2H]^{2+}$ = 1,041.504 and $[M + 3H]^{3+}$ = 694.672) of the matched spectra was consistent with the predicted core peptide after loss of one water molecule (−18.010). The MS^2 fragmentation pattern further indicated abundant ions matching *m/z* for y6 and y7 ions. The 17-amino-acid core peptide sequence (GYPWWDNRDIFGGRTFL) is a novel lasso peptide variant with 76% amino acid identity to propeptin, an endopeptidase inhibitor (21). The analysis was also repeated for RiPP BGCs predicted by PRISM, and no matches were detected with a *P* value lower than e^{-10}.

During the preparation of the manuscript, antiSMASH 6.0 was released and features improvements in the annotation of RiPPs (22). Importantly, antiSMASH 6.0 incorporates the program RRE-finder that allows one to search for RiPP recognition elements (RREs) in candidate tailoring enzymes, which enables more precise annotation of RiPP BGCs. It also includes new models to detect classes of RiPPs previously not detected in version 5.0 (e.g., epipeptide, ranthipeptide, and cyclic-lactone-autoinducer). To assess whether these improvements would uncover additional RiPP matches, we rescanned the genomes using antiSMASH 6.0 and detected 69 additional RiPP BGCs. We extracted the candidate precursor genes (<600-nt open reading frames) from these BGCs and screened the MS/MS data for the presence of their peptide products using MetaMiner. We observed two spectral matches for distinct lasso peptides with the antiSMASH6 RiPP BGCs. One of them corresponded to the lasso peptide in GA6-002 described above. Interestingly, the second lasso peptide was another novel class II lasso peptide candidate, and it was detected in the *n*-butanol extracts of *Nocardia* strain GA14-001 (Fig. 6). The GA14-001 BGC also encoded common components of a lasso peptide

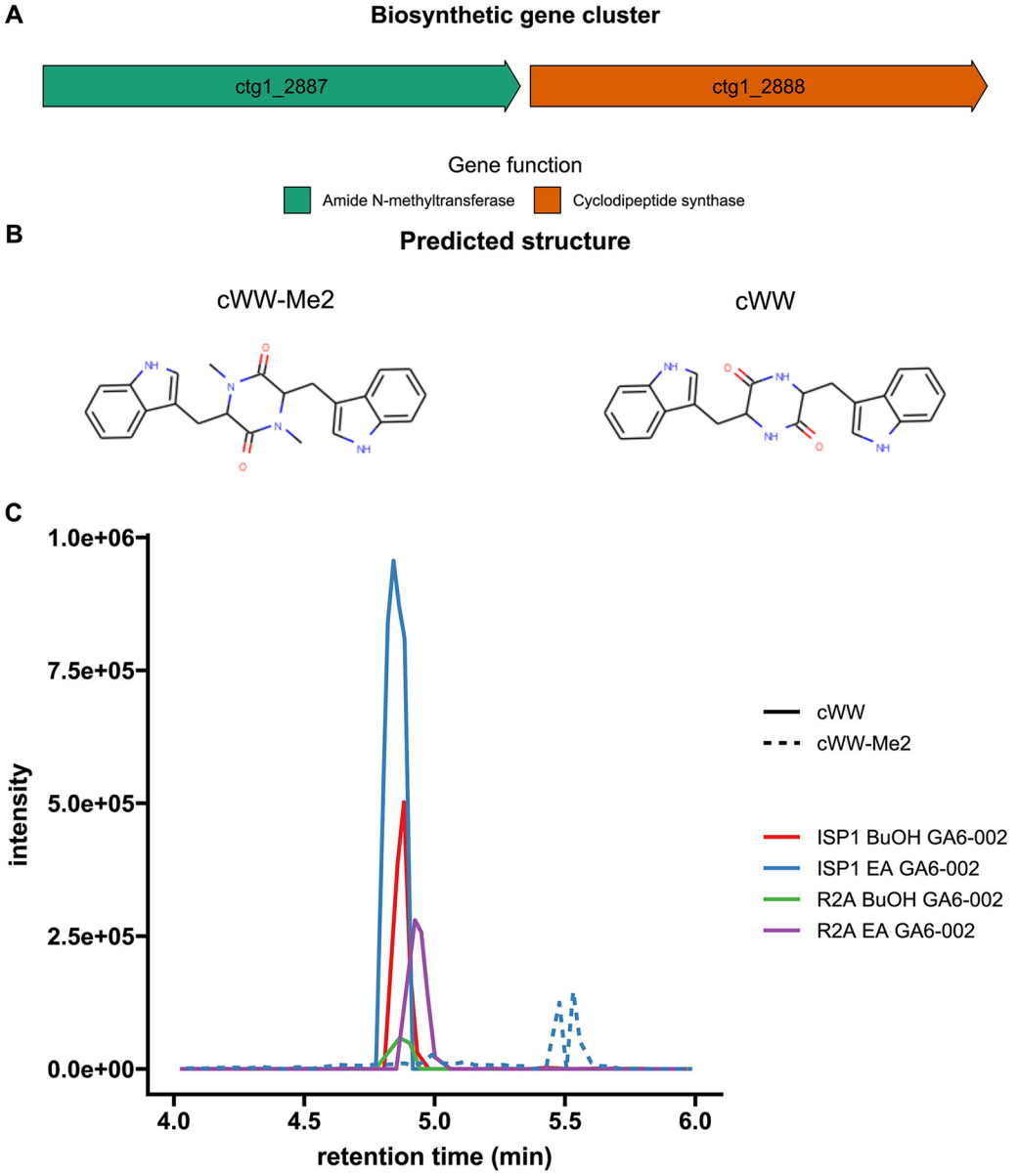

FIG 4 The cyclic dipeptide di-*N*-methylated cyclo(Trp-Trp) detected in *Amycolatopsis* sp. GA6-002 and its corresponding BGC. Cyclic dipeptide BGC (A) and the structures predicted by PRISM based on predicted specificity of the cyclic dipeptide synthase (B). (C) Extracted ion chromatogram for cyclic-di-tryptophan (cWW) and its *N*-methylated derivative (cWW-Me2) observed in four separate organic extracts (EA, ethyl acetate extract; BuOH, *n*-butanol extract).

biosynthetic pathway: an asparagine synthetase, transglutamase protease, a standalone RRE domain, and a precursor peptide (Fig. 6). The precursor peptide sequence did not share significant similarity with known lasso peptides. The peptide is predicted to harbor a unique isopeptide linkage between an N-terminal Ile and Asp residue, forming an eight-membered macrolactam ring (Fig. 6). In the MS data, we observed a peak eluting at 5.5 min in the extracted ion chromatogram (*m/z* 845.48434) corresponding to the $[M + 2H]^{2+}$ ion for the modified core peptide (Fig. 6). The MS^2 fragmentation spectra with ions corresponding to the *m/z* for b15, b12, b14, and y15 further supported the peptide detection in the strain (Fig. 6). Historically, it was thought that Gly is required at position 1 in a lasso peptide (19). This requirement has been updated by characterization of new lasso peptides showing diverse residues at position 1 exemplified by the presence of Ile in strain GA14-001 and Leu in citrulassin and lagmysin (19).

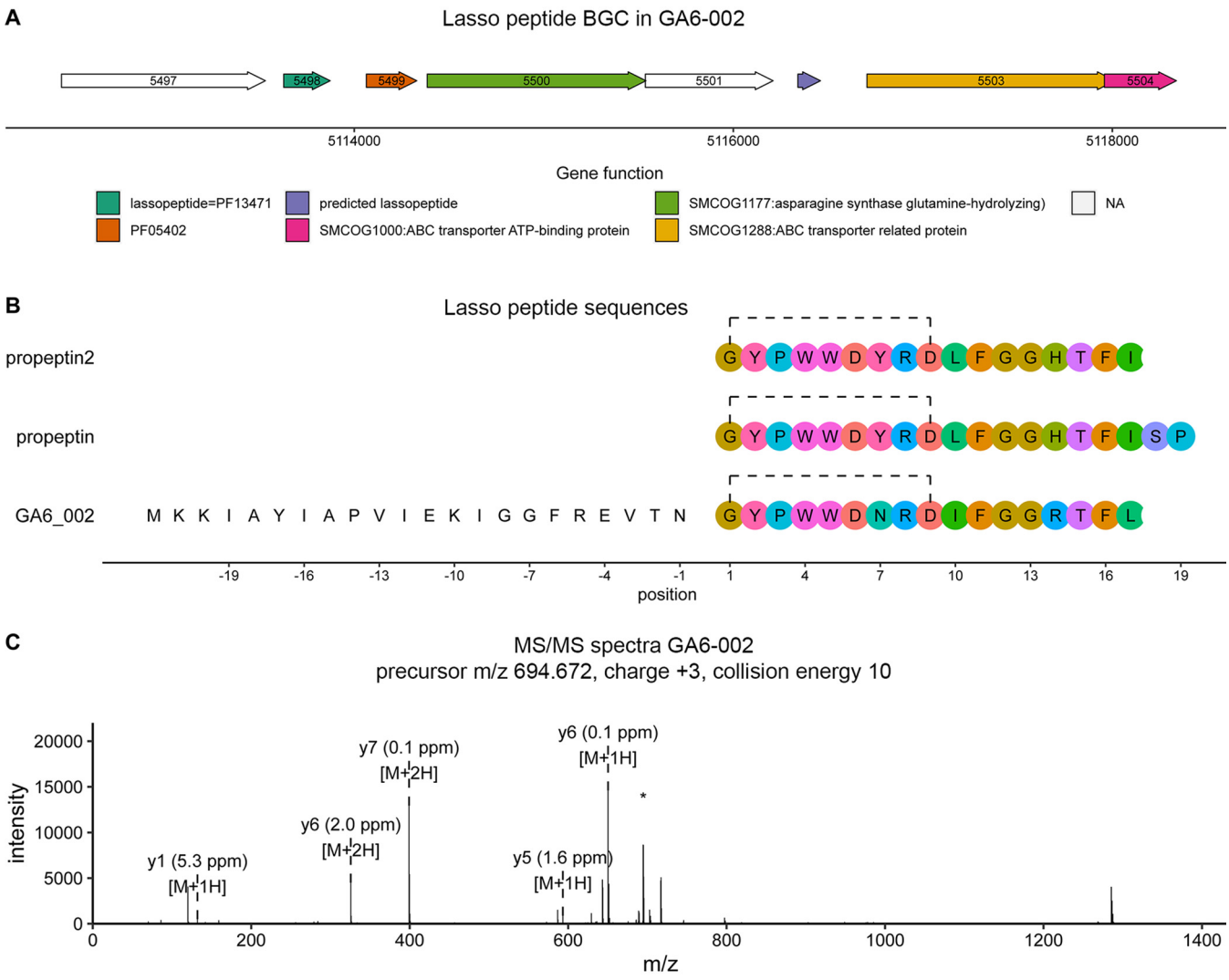

FIG 5 A new lasso peptide variant identified in GA6-002. (A) Schematic of the lasso peptide BGC in GA6-002. Genes are colored according to their functions. (B) Lasso peptide sequences in propeptin 2, propeptin, and GA6-002 (this study). Amino acid positions for the leader sequence are shown as negative numbers. A BGC for propeptin 2 or propeptin has yet to be characterized, and therefore, the leader sequences are not shown. Dashed line between Gly and Asp shows the position of the predicted cross-link. (C) MS/MS spectra that match the posttranslationally modified predicted core sequence in GA6-002. The precursor ion is indicated by an asterisk.

(iii) Known metabolites and their BGCs. An important application of genome sequencing is to understand the biosynthesis of known natural products. Genome sequences can also be used for dereplication of strains and/or compounds in a natural product discovery program. To assess this application with the current sequencing data, we screened the MS/MS spectra for known natural products present in the Natural Product Atlas database (3) (29,006 compounds) using molDiscovery (16). A total of 324 significant matches to known compounds were detected (P value $< e^{10}$). Of these, 30 had a reference MS/MS spectrum available in the Global Natural Product Social Molecular Networking (GNPS) database. We compared the spectra observed in our data set to the reference spectra available in GNPS and found MS/MS spectral matches, defined as sharing 5 or more fragment ions (Fig. S8).

Twenty-one of the 324 identified compounds had a previously characterized BGC in the MiBiG database (4). Twelve out of the 21 compounds were known natural products previously isolated from an actinomycete strain. The other nine were compounds isolated from diverse bacterial genera, including several Gram-negative bacteria and others from the phylum *Cyanobacteria*. Using the antiSMASH known cluster blast module that compares BGCs to those characterized and entered into the MiBiG database, we were able to confirm the

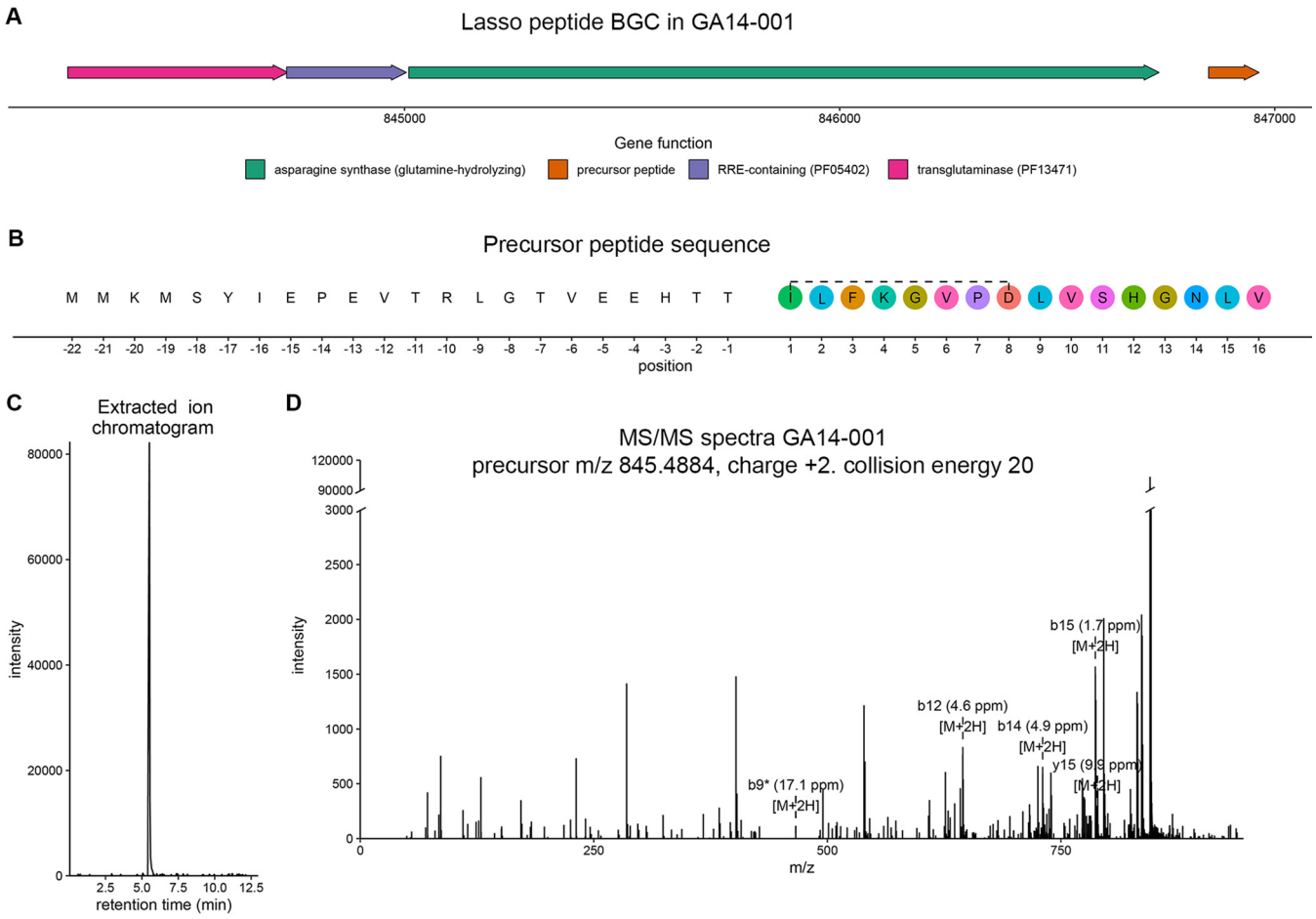

FIG 6 A new lasso peptide identified in GA14-001. (A) Schematic of the lasso peptide BGC in GA14-001. (B) Amino acid sequence of the precursor peptide. Leader sequence is shown with negative numbering and core peptide with positive numbering. The dashed line shows the predicted macrolactam ring. (C) Extracted ion chromatogram for *m/z* 845.48 (±20 ppm) from an *n*-butanol extract of GA14-001 cultured in ISP1 medium. (D) MS/MS spectra for the detected peptide. y and b ions for the modified core peptide are annotated with their charge and calculated ppm difference. * indicates loss of ammonia from the b9 ion.

presence of homologous BGCs for four known actinomycetes in our genome set, and the production of their corresponding compounds could be confirmed by MS/MS (Fig. 7). These compounds included *N*-acetyltryptophan (strain GA10-001) and the siderophores heterobactin A (strain GA14-001 and strain GA14-002), mirubactin (strain GA10-003), and salinichelin (strain GA10-001). From the BGC comparisons illustrated in Fig. 7, it is evident that the Flongle-sequenced genomes from this study may have sequencing artifacts resulting in fragmentation of large genes into multiple small open reading frames (ORFs) (see, for example, the comparison of mirubactin).

BGCs encoding known antibiotic classes with no metabolites detected—glycopeptide, aminoglycoside, and aminocoumarin. The strains sequenced in the current study were resistant to streptomycin, gentamicin, vancomycin, or novobiocin (see Materials and Methods). Enriching for antibiotic resistance in actinomycete collections often enriches for strains with the biosynthetic capacity to produce an antibiotic of the same scaffold (23). Nonetheless, in the above-described analyses, we did not detect a glycopeptide, aminoglycoside, or aminocoumarin antibiotic, although we were able to confirm the presence of BGCs potentially encoding antibiotics of these families through a manual inspection of the antiSMASH known cluster blast results. Fourteen BGCs from the sequenced strains that shared similarity to previously characterized BGCs encoding aminoglycoside, aminocoumarin, and glycopeptide antibiotics are described below.

Three *Amycolatopsis lurida* strains (GB15-009, GA10-003, and GA10-004) harbored a nearly identical aminocyclitol gene cluster which encoded a homolog of 2-epi-5-epi-

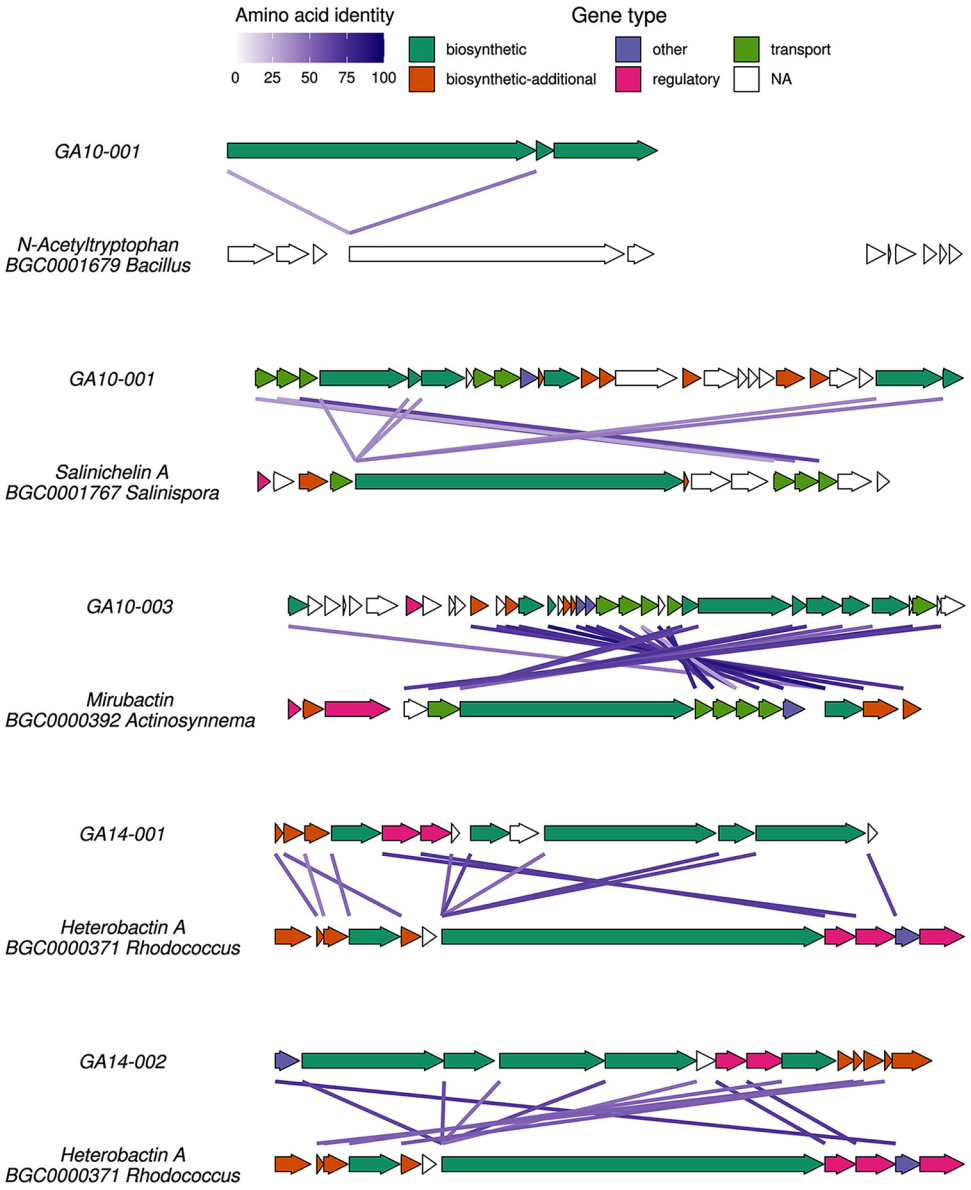

FIG 7 Alignments and pairwise comparisons of homologous BGCs mapped for known metabolites detected by LC-MS/MS. BGCs used for comparisons to those identified in this study are from the MiBiG database. Genes are colored according to function (top). The purple lines connecting homologous genes indicate percent similarity with color intensity being proportional to the BLAST amino acid identity between the BGCs in any pair of genomes.

valiolone synthase (salQ) which is responsible for the first step in the biosynthesis of C7N-aminocyclitols (24) (Fig. S9). Aminocyclitols are biosynthesized from sugars through cyclization by a sugar phosphate cyclase (SPC) such as dehydroquinate (DHQ) synthase. The BGCs were highly homologous to cetoniacytone A, sharing 70 to 82% amino acid identity for core biosynthetic genes (25, 26).

In *Amycolatopsis coloradensis* B06-03, an aminocoumarin BGC was detected (Fig. S10). Aminocoumarin antibiotics are biosynthesized from L-tyrosine (27). Tyrosine is activated by an adenylation domain and covalently attached to a peptidyl carrier protein (PCP). A NovH-like cytochrome P450 hydroxylates PCP-bound tyrosine to β-hydroxy tyrosyl-S-PCP. A 3-oxoacyl-acyl carrier protein (ACP) reductase converts it to a β-keto-tyrosyl intermediate that undergoes cyclization to form 3-amino-4,7-dihydroxycoumarin. In B06-03, downstream of the core aminocoumarin biosynthesis genes, a type I polyketide BGC encoding

coenzyme A ligase (CAL) domain specific for 3-amino-5-hydroxybenzoic acid (AHBA) was present, as in rubradirin (24) and chaxamycin (28).

A total of seven BGCs similar to previously characterized glycopeptide BGCs were detected (see Fig. S11 at https://doi.org/10.6084/m9.figshare.16722835.v1). Glycopeptides are biosynthesized through a multimodular nonribosomal peptide synthetase (NRPS) assembly line. Glycopeptide BGCs encode additional tailoring enzymes such as P450 monooxygenases and glycosyltransferases that result in amino acid cross-linking and glycosylation, respectively, to yield the complex multicyclic antibiotics exemplified by vancomycin. The predicted glycopeptides from the four glycopeptide-like BGCs (from strains GA10-004, GA10-003, GB8-002, and GB15-009) were similar in amino acid composition to ristocetin (29). These four strains primarily contained butylated hydroxytoluene (Bht), dihydroxyphenylglycine (Dhpg), and 4-hydroxyphenylglycine (Hpg) as seen in ristocetin. The predicted glycopeptide from B06-03 is predicted to contain Trp, Hpg, and Tyr as seen in complestatin (30).

DISCUSSION

Soil actinomycetes hold enormous potential for the discovery of new antibiotics. However, the number of genome-sequenced actinomycetes in the public domain is still limited, partly due to the cost of long-read next-generation sequencing. In this study, we assessed the capability of the ONT Flongle platform as a low-cost sequencing option to obtain multiple nearly complete genomes of actinomycetes, identify BGCs, and connect them to metabolites through a paired genome-metabolome analysis.

Our sequencing and assembly results showed that up to four nearly complete genomes of actinomycete strains could be sequenced on a single Flongle device. Skipping an optional DNA fragmentation step enabled read lengths up to 80 kb in each sample. Bead-based size selection further depleted shorter DNA fragments and enriched sequencing reads in longer sequences (10 kb+). The long reads enabled contiguous DNA assemblies at lower sequencing depth. The sizes of the assemblies were typical of soil actinomycetes, suggesting that they represent nearly complete genomes of the strains. There were several mismatches in the accuracy comparison analysis in Flongle genomes relative to PacBio or high-coverage MinION genomes. However, the contiguity of the genomes was only slightly affected (1 to 2 contigs versus 6 to 7 contigs), indicating that important structural information about the genome (e.g., position and organization of genes) can be inferred from these sequences.

A common strategy to obtain contiguous and accurate genome assemblies is through polishing contiguous Nanopore assemblies with Illumina reads. One of the significant findings of this study is that BGC predictions and their connection to metabolites were performed without the need for error correction using Illumina reads. Based on downsampling analysis of public data sets, it was initially hypothesized that low-coverage Nanopore-only assemblies could be used to predict and analyze BGCs. Through prospectively sequencing actinomycetes using Flongle, it was empirically evaluated in the current study.

An interesting observation upon BGC analysis was that active site specificity for various BGC classes (NRPS, PKS, and CDPS) in the Flongle assemblies was correctly predicted. The active site specificities were used by PRISM to generate possible structures, which were then used to query MS/MS data for potential spectral matches. A spectrum match to a predicted structure indirectly proves that active site specificities were correct, for instance, in the case of cWW. However, we also observed that frameshifts and sequencing errors affected *in silico* prediction of accurate structures for some BGCs.

The analysis of RiPP BGCs in Flongle assemblies was relatively less affected by sequencing. RiPP metabolite prediction is based on short precursor peptides, and BGC prediction relies on detecting posttranslational modifying enzymes through error-tolerant profile hidden Markov models. The chance of a mismatch underlying a 100-nucleotide (30-mer) core peptide sequence is low. For example, 19,227 mismatches were detected in total in a Flongle assembly relative to PacBio, which corresponds to a chance of less than one mismatch per 100 nt (19,227 mismatches/7,183,038-nt

genome size × 100 nt = 0.26 per 100 nt). This is evident through accurate prediction of two new lasso peptide BGCs and matching of experimental MS/MS spectra to their predicted products.

Similarly, BGCs homologous to previously characterized BGCs for known metabolites can be identified through sequence similarity searches. The consensus accuracy of the assemblies was observed to be 99.5%, which makes a genome sequence suitable for comparison with known BGC sequences or for computing average nucleotide identity to published genomes. We demonstrated this through the rediscovery of BGCs and selected actinomycete siderophores.

While our results suggest Flongle is a useful platform for sequencing actinomycetes, increased consistency in total sequencing output might enable further optimized workflows. For instance, Flongle flow cells were inconsistent in the number of starting pores (<60 out of an expected 126 in most cases), which affected total sequencing output and yielded lower-than-desired coverage for a few samples. A consistent number of pores closer to the anticipated total number (>100) across experiments would allow for higher sequencing coverage using the same experimental workflow or allow more genomes to be sequenced in an experiment.

While a few BGCs could be connected to the metabolites in the current study, most remained unconnected. Connecting BGCs to metabolites is a multifactor problem not limited by sequencing accuracy alone. Improvements in experiments and computational algorithms would be needed to circumvent this issue in the future. First, it is highly unlikely that all BGCs will be expressed when strains are grown in only two culture media as tested here; thus, additional media and growth conditions or genetics-free elicitor screens should be used (31–33). Second, only PRISM *in silico* predicted structures were used. In the future, a more extensive *in silico* structure generation that addresses ambiguous active site specificities (e.g., two or more possible amino acids at a site in NRPS) could be used. Third, MS/MS data were queried for exact compound spectral matches. Minor differences between predicted and expressed metabolites (e.g., single-site methylation or hydroxylation) would result in a mass shift, and a match would not be possible.

In summary, multiplexed low-coverage sequencing of actinomycete genomes on Flongle is a promising option for the genome-guided discovery of natural products. Numerous research laboratories house valuable bacterial strain collections (34–39). Limited by the costs of large-scale long read sequencing, genome sequencing of natural product producing bacteria usually occurs on a strain-by-strain basis (40–43). The future of natural product research is expected to involve analysis of genomics and metabolomics data using genome mining (e.g., antiSMASH and PRISM) and mass spectrum matching tools (such as molDiscovery integrated with NPAtlas-like databases used in this study). Indeed, such efforts are already taking place on metagenomic data sets (44, 45); while those studies provide vast amounts of data on the natural productsome, a key limitation is that the data are not connected to archived bacterial strains. It is our hope that low-cost sequencing workflows such as the one described here may allow for access to genome sequencing on a larger scale and/or by a broader community of researchers, especially in resource-limited settings.

MATERIALS AND METHODS

Strain isolation. The 20 sequenced strains were a subset of streptomycin-, novobiocin-, or vancomycin-resistant strains from an in-house actinomycete strain library housed in the Laboratory of Bioorganic Chemistry, National Institutes of Health. The strains were isolated from soil specimens collected from deserts in Arizona, California, and Nevada through standard procedures described in a previous study (40).

Nanopore sequencing. Each strain was cultivated for 3 to 7 days in 10 ml of tryptic soy broth (BD Diagnostic, catalog no. 211768) with 0.5% (wt/vol) glycine from frozen glycerol stocks. The cultures were centrifuged at 10,000 × *g* for 10 min, and cell pellets were resuspended into 250 μl Tris-EDTA (TE) buffer followed by addition of 50 μl of lysosome (100 mg/ml). The mixture was incubated overnight (16 h) at 37°C. The next morning 10 μl of RNase A (10 mg/μl) was added to the cell lysate and incubated for an additional 20 min after which 250 μl of proteinase K (400 μg/μl) was added and incubated for 2 h. DNA was purified from cell lysates using 1:1 (vol/vol) phenol-chloroform extraction, and the DNA was collected from the upper phase. Genomic DNA was precipitated with 0.7 volume of isopropanol, washed with 80% ethanol, and resuspended into 50 μl TE.

DNA libraries were prepared using the Oxford Nanopore ligation sequencing kit (SQK-LSK109) and the native barcoding kit (NBD104) protocol for Flongle with some modifications. A DNA fragmentation step was not performed. Five hundred nanograms of genomic DNA was directly processed for DNA end repair with the NEBNext Ultra II end repair/dA-tailing module (New England Biolabs, catalog no. E7546). Barcodes were ligated to the end-repaired DNA and purified with 0.1× or 0.15× beads (Omega Bio-Tek Inc., catalog no. M1378-01), resuspended in a custom buffer (10 mM Tris-Cl, pH 8.0, 1 mM EDTA, 0.5 M $MgCl_2$, and 5% PEG). A pooled library was prepared by combining 62.5 ng of each barcoded DNA. Nanopore adapters were ligated to the pooled library followed by library loading and sequencing according to the manufacturer's instructions.

Data-dependent untargeted LC-MS/MS. Each strain was cultivated in deep-well plates containing 400 μl of ISP1 (BD Diagnostic, catalog no 276910) or R2A (Teknova, catalog no. R0005) medium. The cultures were incubated at 30°C with shaking at 200 rpm for 1 week before extraction with an equal volume of ethyl acetate followed by extraction with *n*-butanol. Uninoculated media were used as blanks/negative control, and any metabolite observed in a blank run was excluded from interpretation. The liquid chromatography (LC)-MS/MS data were collected using an Agilent 1290 Infinity II ultraperformance liquid chromatography (UPLC) system equipped with an Agilent 6545 quadrupole time of flight (qTOF) mass spectrometer. Samples were chromatographed on an Agilent Eclipse Plus C_{18} 2.1- by 50-mm column (3-μl injections) using a gradient of 99% A (0.1% formic acid in water) to 95% B (acetonitrile) at a flow rate of 0.5 ml/min over 10 min. MS/MS fragmentation was carried out in auto mode with collision energies of 10, 20, and 40 keV excluding precursor ions in the range of 40 to 180 *m/z* and abundance below 7,000 counts.

Data analysis. (i) Genomics. Primary genomic data analysis was conducted by basecalling with guppy (version 4.2.2, model dna_r9.4.1_450bps_hac.cfg), demultiplexing with Qcat (version 1.0.6), and assembling with Canu (version 2.0) (13). Canu assemblies were constructed with an expected genome size of 8 Mb, minimum read length threshold of 1 kb, minimum coverage of 2, and high Mhap error correction sensitivity (13). The genome assemblies were polished by aligning reads to the assembly and calling consensus with Racon and Medeka (14). Genes and secondary metabolite gene clusters were predicted using the programs antiSMASH (version 5) (46) and PRISM (version 4.4.5) (15).

(ii) Assessing effect of sequencing coverage on BGC detection. FastQ reads for previously sequenced actinomycete genomes were downloaded from the European Nucleotide Archive, downsampled to an estimated coverage of 60×, 30×, 15×, and 7× using Seqtk (assuming an 8-Mb genome as in prospective sequencing) (see Table S1 at https://doi.org/10.6084/m9.figshare.16722961). Seqtk allows random subsampling of reads. Reads were subsampled to desired coverage according to the following estimation: number of reads for Q coverage = (Q/original coverage) × original number of reads. The downsampled FastQ files were assembled with Canu and polished with Medeka, and BGCs were predicted using antiSMASH. The number of mismatches in each assembly relative to original coverage was calculated using Quast (47). For consistency with data from this study (presented in Fig. 3), the coverage presented is aligned coverage, taking into account the size of the final assembly and not the expected size (i.e., 8 Mb). The mapped coverage was extracted from Canu assembler tig information files.

(iii) BGC comparison between strains. The antiSMASH-predicted BGC sequences were extracted from each strain's genome assemblies and aligned in all possible strain pair combinations using minimap2, allowing for 5% sequence divergence (48). If a BGC from strain 1 did not align to any of the BGCs in strain 2, it was considered absent in strain 2.

(iv) Homologous BGCs of previously characterized metabolites. Homologous BGCs of previously characterized metabolites were obtained through the 'known cluster blast' module of antiSMASH. The output of the program contains gene-wise blast hits for each BGC in the genome to BGCs in MiBiG (4). To identify the best hit in the MiBiG database, output was first filtered to obtain MiBiG BGCs that share the largest number of genes, highest mean percent identity, and highest mean coverage with genes in the query BGC. The results were subsequently filtered to retain only BGCs where the ratio of lengths between query and MiBiG BGCs was between 0.7 and 1.1.

(v) LC-MS/MS analysis. Analysis of LC-MS/MS data was conducted by conversion of the vendor .d format to mzXML files using the GNPS conversion utility. These mzXML files were subsequently used for all analyses.

(vi) Spectrum matching—known or unknown structures. Spectrum matches for known metabolites using the Natural Product Atlas or unknown metabolites (PRISM-predicted structures) were identified by using molDiscovery (3, 16). molDiscovery computes theoretical MS/MS spectra of compounds in the database, identifies spectrum matches at user-defined mass tolerance, and subsequently calculates statistical significance by matching the spectrum against a decoy database. In this analysis, mass tolerance of 20 ppm, *P* value less than e^{-10}, and FDR less than 1% were considered.

(vii) RiPPs. Spectrum matches for RiPPs were detected using MetaMiner (18). Given a list of short peptides, MetaMiner constructs possible RiPP products based on knowledge of posttranslational modifications within RiPPs. It then predicts an MS/MS spectrum for each predicted RiPP product and conducts a search of experimentally collected MS/MS spectra for potential matches. All open reading frames (ORFs) shorter than 600 nt (200 amino acids) located within RiPP BGCs predicted by antiSMASH or PRISM were used for this analysis.

Integration of the mass spectrometry and genomic sequences was achieved through R scripts and several packages including MSnbase (49) and Open Babel (50).

Data availability. Raw sequencing data are available at NCBI, project accession no. PRJNA752621. Genome assemblies and additional data are available at Figshare (https://doi.org/10.6084/m9.figshare.15094044.v1). The MS/MS spectra have been uploaded to GNPS with accession no. MSV000087950. Scripts used in data analysis and preparation of figures are available at https://github.com/rajwanir/flongle_actinomycetes_paper.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.3 MB.

FIG S2, PDF file, 0.1 MB.

FIG S3, PDF file, 0.1 MB.

FIG S4, PDF file, 0.1 MB.

FIG S5, PDF file, 0.2 MB.

FIG S6, PDF file, 0.1 MB.

FIG S7, PDF file, 0.1 MB.

FIG S8, PDF file, 0.1 MB.

FIG S9, PDF file, 0.1 MB.

FIG S10, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

This work was supported by the NIH Intramural Research Program (NIDDK) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

REFERENCES

1. World Health Organization. 2019. No time to wait: securing the future from drug-resistant infections. World Health Organization, Geneva, Switzerland.
2. Genilloud O. 2017. Actinomycetes: still a source of novel antibiotics. Nat Prod Rep 34:1203–1232. https://doi.org/10.1039/c7np00026j.
3. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, Neto FC, Castaño-Espriu L, Chang C, Clark TN, Cleary Little JL, Delgadillo DA, Dorrestein PC, Duncan KR, Egan JM, Galey MM, Haeckl FPJ, Hua A, Hughes AH, Iskakova D, Khadilkar A, Lee J-H, Lee S, LeGrow N, Liu DY, Macho JM, McCaughey CS, Medema MH, Neupane RP, O'Donnell TJ, Paula JS, Sanchez LM, Shaikh AF, Soldatou S, Terlouw BR, Tran TA, Valentine M, van der Hooft JJJ, Vo DA, Wang M, Wilson D, Zink KE, Linington RG. 2019. The Natural Products Atlas: an open access knowledge base for microbial natural products discovery. ACS Cent Sci 5:1824–1833. https://doi.org/10.1021/acscentsci.9b00806.
4. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal AV, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 48:D454–D458. https://doi.org/10.1093/nar/gkz882.
5. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16:60–68. https://doi.org/10.1038/s41589-019-0400-9.
6. Kloosterman AM, Cimermancic P, Elsayed SS, Du C, Hadjithomas M, Donia MS, Fischbach MA, van Wezel GP, Medema MH. 2020. Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lantibiotics. PLoS Biol 18:e3001026. https://doi.org/10.1371/journal.pbio.3001026.
7. Ganley JG, Pandey A, Sylvester K, Lu K-Y, Toro-Moreno M, Rütschlin S, Bradford JM, Champion CJ, Böttcher T, Xu J, Derbyshire ER. 2020. A systematic analysis of mosquito-microbiome biosynthetic gene clusters reveals antimalarial siderophores that reduce mosquito reproduction capacity. Cell Chem Biol 27:817–826. https://doi.org/10.1016/j.chembiol.2020.06.004.
8. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39:W339–W346. https://doi.org/10.1093/nar/gkr466.
9. Meleshko D, Mohimani H, Tracanna V, Hajirasouliha I, Medema MH, Korobeynikov A, Pevzner PA. 2019. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. Genome Res 29:1352–1362. https://doi.org/10.1101/gr.243477.118.
10. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155–1162. https://doi.org/10.1038/s41587-019-0217-9.
11. Wick RR, Holt KE. 2019. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Res 8:2138–2138. https://doi.org/10.12688/f1000research.21782.4.
12. Stortchevoi A, Kamelamela N, Levine SS. 2020. SPRI beads-based size selection in the range of 2-10kb. J Biomol Tech 31:7–10. https://doi.org/10.7171/jbt.20-3101-002.
13. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116.
14. Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27:737–746. https://doi.org/10.1101/gr.214270.116.
15. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, Li H, Ranieri MRM, Webster ALH, Cao MPT, Pfeifle A, Spencer N, To QH, Wallace DP, Dejong CA, Magarvey NA. 2020. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun 11:6058. https://doi.org/10.1038/s41467-020-19986-1.
16. Mohimani H, Cao L, Guler M, Tagirdzhanov A, Gurevich A. 2020. MolDiscovery: learning mass spectrometry fragmentation of small molecules. Preprint, version 1. Research Square https://doi.org/10.21203/rs.3.rs-71854/v1.
17. Giessen TW, von Tesmar AM, Marahiel MA. 2013. A tRNA-dependent two-enzyme pathway for the generation of singly and doubly methylated ditryptophan 2,5-diketopiperazines. Biochemistry 52:4274–4283. https://doi.org/10.1021/bi4004827.
18. Cao L, Gurevich A, Alexander KL, Naman CB, Leão T, Glukhov E, Luzzatto-Knaan T, Vargas F, Quinn R, Bouslimani A, Nothias LF, Singh NK, Sanders JG, Benitez RAS, Thompson LR, Hamid M-N, Morton JT, Mikheenko A, Shlemov A, Korobeynikov A, Friedberg I, Knight R, Venkateswaran K, Gerwick WH, Gerwick L, Dorrestein PC, Pevzner PA, Mohimani H. 2019. MetaMiner: a scalable peptidogenomics approach for discovery of ribosomal peptide natural products with blind modifications from microbial communities. Cell Syst 9:600–608.e604. https://doi.org/10.1016/j.cels.2019.09.004.
19. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai H-C, Zakai UI, Mitchell DA. 2017. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat Chem Biol 13:470–478. https://doi.org/10.1038/nchembio.2319.
20. Metelev M, Tietz JI, Melby JO, Blair PM, Zhu L, Livnat I, Severinov K, Mitchell DA. 2015. Structure, bioactivity, and resistance mechanism of

streptomonomicin, an unusual lasso peptide from an understudied halophilic actinomycete. Chem Biol 22:241–250. https://doi.org/10.1016/j.chembiol.2014.11.017.
21. Kimura K, Kanou F, Takahashi H, Esumi Y, Uramoto M, Yoshihama M. 1997. Propeptin, a new inhibitor of prolyl endopeptidase produced by microbispora. I. Fermentation, isolation and biological properties. J Antibiot 50:373–378. https://doi.org/10.7164/antibiotics.50.373.
22. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T. 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res 49:W29–W35. https://doi.org/10.1093/nar/gkab335.
23. Thaker MN, Wang W, Spanogiannopoulos P, Waglechner N, King AM, Medina R, Wright GD. 2013. Identifying producers of antibacterial compounds by screening for antibiotic resistance. Nat Biotechnol 31:922–927. https://doi.org/10.1038/nbt.2685.
24. Choi WS, Wu X, Choeng YH, Mahmud T, Jeong BC, Lee SH, Chang YK, Kim CJ, Hong SK. 2008. Genetic organization of the putative salbostatin biosynthetic gene cluster including the 2-epi-5-epi-valiolone synthase gene in *Streptomyces albus* ATCC 21838. Appl Microbiol Biotechnol 80:637–645. https://doi.org/10.1007/s00253-008-1591-2.
25. Schlorke O, Krastel P, Muller I, Uson I, Dettner K, Zeeck A. 2002. Structure and biosynthesis of cetoniacytone A, a cytotoxic aminocarba sugar produced by an endosymbiontic Actinomyces. J Antibiot (Tokyo) 55:635–642. https://doi.org/10.7164/antibiotics.55.635.
26. Wu X, Flatt PM, Xu H, Mahmud T. 2009. Biosynthetic gene cluster of cetoniacytone A, an unusual aminocyclitol from the endosymbiotic bacterium *Actinomyces* sp. Lu 9419. Chembiochem 10:304–314. https://doi.org/10.1002/cbic.200800527.
27. Heide L. 2009. The aminocoumarins: biosynthesis and biology. Nat Prod Rep 26:1241–1250. https://doi.org/10.1039/b808333a.
28. Castro JF, Razmilic V, Gomez-Escribano JP, Andrews B, Asenjo JA, Bibb MJ. 2015. Identification and heterologous expression of the chaxamycin biosynthesis gene cluster from *Streptomyces leeuwenhoekii*. Appl Environ Microbiol 81:5820–5831. https://doi.org/10.1128/AEM.01039-15.
29. Truman AW, Kwun MJ, Cheng J, Yang SH, Suh J-W, Hong H-J. 2014. Antibiotic resistance mechanisms inform discovery: identification and characterization of a novel *Amycolatopsis* strain producing ristocetin. Antimicrob Agents Chemother 58:5687–5695. https://doi.org/10.1128/AAC.03349-14.
30. Culp EJ, Waglechner N, Wang W, Fiebig-Comyn AA, Hsu Y-P, Koteva K, Sychantha D, Coombes BK, Van Nieuwenhze MS, Brun YV, Wright GD. 2020. Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. Nature 578:582–587. https://doi.org/10.1038/s41586-020-1990-9.
31. Bode HB, Bethe B, Hofs R, Zeeck A. 2002. Big effects from small changes: possible ways to explore nature's chemical diversity. Chembiochem 3:619–627. https://doi.org/10.1002/1439-7633(20020703)3:7<619::AID-CBIC619>3.0.CO;2-9.
32. Liu M, Grkovic T, Liu X, Han J, Zhang L, Quinn RJ. 2017. A systems approach using OSMAC, log p and NMR fingerprinting: an approach to novelty. Synth Syst Biotechnol 2:276–286. https://doi.org/10.1016/j.synbio.2017.10.001.
33. Xu F, Wu Y, Zhang C, Davis KM, Moon K, Bushin LB, Seyedsayamdost MR. 2019. A genetics-free method for high-throughput discovery of cryptic microbial metabolites. Nat Chem Biol 15:161–168. https://doi.org/10.1038/s41589-018-0193-2.
34. Amiri Moghaddam J, Crusemann M, Alanjary M, Harms H, Davila-Cespedes A, Blom J, Poehlein A, Ziemert N, Konig GM, Schaberle TF. 2018. Analysis of the genome and metabolome of marine myxobacteria reveals high potential for biosynthesis of novel specialized metabolites. Sci Rep 8:16600. https://doi.org/10.1038/s41598-018-34954-y.
35. Bader CD, Panter F, Muller R. 2020. In depth natural product discovery - myxobacterial strains that provided multiple secondary metabolites. Biotechnol Adv 39:107480. https://doi.org/10.1016/j.biotechadv.2019.107480.
36. Hernandez A, Nguyen LT, Dhakal R, Murphy BT. 2021. The need to innovate sample collection and library generation in microbial drug discovery: a focus on academia. Nat Prod Rep 38:292–300. https://doi.org/10.1039/d0np00029a.
37. Jensen PR, Moore BS, Fenical W. 2015. The marine actinomycete genus Salinispora: a model organism for secondary metabolite discovery. Nat Prod Rep 32:738–751. https://doi.org/10.1039/c4np00167b.
38. Steele AD, Teijaro CN, Yang D, Shen B. 2019. Leveraging a large microbial strain collection for natural product discovery. J Biol Chem 294:16567–16576. https://doi.org/10.1074/jbc.REV119.006514.
39. Waglechner N, McArthur AG, Wright GD. 2019. Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. Nat Microbiol 4:1862–1871. https://doi.org/10.1038/s41564-019-0531-5.
40. Sun J, Zhao G, O'Connor RD, Davison JR, Bewley CA. 2021. Vertirhodins A-F, *C*-linked pyrrolidine-iminosugar-containing pyranonaphthoquinones from *Streptomyces* sp. B15-008. Org Lett 23:682–686. https://doi.org/10.1021/acs.orglett.0c03825.
41. Li H, Zhang M, Li H, Yu H, Chen S, Wu W, Sun P. 2021. Discovery of venturicidin congeners and identification of the biosynthetic gene cluster from *Streptomyces* sp. Nrrl s-4. J Nat Prod 84:110–119. https://doi.org/10.1021/acs.jnatprod.0c01177.
42. Yang J, Song Y, Tang M-C, Li M, Deng J, Wong N-K, Ju J. 2021. Genome-directed discovery of tetrahydroisoquinolines from deep-sea derived Streptomyces niveus SCSIO 3406. J Org Chem 86:11107–11116. https://doi.org/10.1021/acs.joc.1c00123.
43. Braesel J, Crnkovic CM, Kunstman KJ, Green SJ, Maienschein-Cline M, Orjala J, Murphy BT, Eustaquio AS. 2018. Complete genome of *Micromonospora* sp. strain b006 reveals biosynthetic potential of a Lake Michigan actinomycete. J Nat Prod 81:2057–2068. https://doi.org/10.1021/acs.jnatprod.8b00394.
44. Sharrar AM, Crits-Christoph A, Méheust R, Diamond S, Starr EP, Banfield JF. 2020. Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type. mBio 11:e00416-20. https://doi.org/10.1128/mBio.00416-20.
45. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen IM, Huntemann M, Palaniappan K, Ladau J, Mukherjee S, Reddy TBK, Nielsen T, Kirton E, Faria JP, Edirisinghe JN, Henry CS, Jungbluth SP, Chivian D, Dehal P, Wood-Charlson EM, Arkin AP, Tringe SG, Visel A, IMG/M Data Consortium, Woyke T, Mouncey NJ, Ivanova NN, Kyrpides NC, Eloe-Fadrosh EA. 2021. A genomic catalog of earth's microbiomes. Nat Biotechnol 39:499–509. https://doi.org/10.1038/s41587-020-0718-6.
46. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. Antismash 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81–W87. https://doi.org/10.1093/nar/gkz310.
47. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. Quast: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
48. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
49. Gatto L, Gibb S, Rainer J. 2021. MSnbase, efficient and elegant R-based processing and visualization of raw mass spectrometry data. J Proteome Res 20:1063–1069. https://doi.org/10.1021/acs.jproteome.0c00313.
50. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. 2011. Open Babel: an open chemical toolbox. J Cheminform 3:33. https://doi.org/10.1186/1758-2946-3-33.