**BiosyntheticSPAdes:**

**Reconstructing Biosynthetic Gene Clusters From Assembly Graphs**

**(Supplementary Material)**

Dmitry Meleshko1,2, Hosein Mohimani3,4, Vittorio Tracanna5, Iman Hajirasouliha6,7, Marnix H. Medema5, Anton Korobeynikov1,8, Pavel A. Pevzner1,3,\*

1Center for Algorithmic Biotechnology, Institute for Translational Biomedicine,

St. Petersburg State University, St. Petersburg, Russia

2Tri-Institutional PhD Program in Computational Biology and Medicine,

Weill Cornell Medical College, New York, United States

3Department of Computer Science and Engineering, University of California, San Diego,

4Computational Biology Department, School of Computer Sciences,

Carnegie Mellon University

5Bioinformatics Group, Wageningen University, Wageningen, The Netherlands

6Institute for Computational Biomedicine, Department of Physiology and Biophysics,

Weill Cornell Medicine of Cornell University, New York, United States

7Englander Institute for Precision Medicine, Meyer Cancer Center,

Weill Cornell Medicine, New York, United States

8Department of Statistical Modelling, St. Petersburg State University, St. Petersburg, Russia

\*Corresponding author, ppevzner@ucsd.edu

**APPENDICES**

**A: Coupling biosyntheticSPAdes and NRPquest for PNP reconstruction**

**B: biosyntheticSPAdes output format**

**C: Putative NRP BGCs in the CYANO dataset**

**D: Biosynthetic capacity of the HMP datasets**

**E: Putative NRP synthetases in the subpravingal plaque samples from the HMP dataset**

**F: Reference-based putative BGC ranking algorithm**

**G: Ranking putative BGCs from *Streptomyces coelicolor* A3(2) and *Streptomyces avermitilis* MA-4680**

**References**

**Supplementary Figures and Tables**

**Appendix A: Coupling biosyntheticSPAdes and NRPquest for PNP reconstruction**

Each of the rural postman routes generated by biosyntheticSPAdes corresponds to a sequence of A-domains and thus allows one to generate putative NRPs encoded by this sequence using *nonribosomal code* (Stachelhaus and Marahiel 1999). Tandem mass spectra can be matched against these putative NRPs resulting in *Peptide-Spectrum Matches* (*PSMs*)with varying P-values (Mohimani and Pevzner, 2016). A PSM with the lowest P-value reveals the NRP (and thus the rural postman tour) that is more likely to be correct than others.

To demonstrate how this approach works, we matched both putative CALC BGCs (corresponding to two rural postman routes for the CALC BGC) against a high resolution mass spectral dataset from *S. coelicolor* deposited in the Global Natural Products Social (GNPS) molecular network (Wang et al. 2016) with MassiveID MSV000078839 (total of 11952 spectra). For each A-domain, we analyzed the top three candidate amino acids predicted by NRPSPredictor2, and considered linear, cyclic, and branch-cyclic structures. This resulted in 20720 candidate structures for each sequence, and we searched all those structures against all mass spectra of *S. coelicolor* using Dereplicator (Mohimani et al. 2017), allowing for a single blind modification. The correct sequence resulted in a score 16 (P-value 8.7 \* 10-15), while the incorrect sequence resulted in a score 15 (P-value 2.9 \* 10-14). This illustrates that coupling of biosyntheticSPAdes with peptidogenomics leads to elucidation of NRPs encoded by predicted NRP BGCs.

**Appendix B: biosyntheticSPAdes output format**

BiosyntheticSPAdes stores all output files in a user-specified folder.

* <output\_dir>/orderings.fasta contains putative sequences for all putative BGCs in the fasta format. Every header of a fasta record has the following format:

>NODE\_1\_length\_60699\_cluster\_3\_candidate\_2

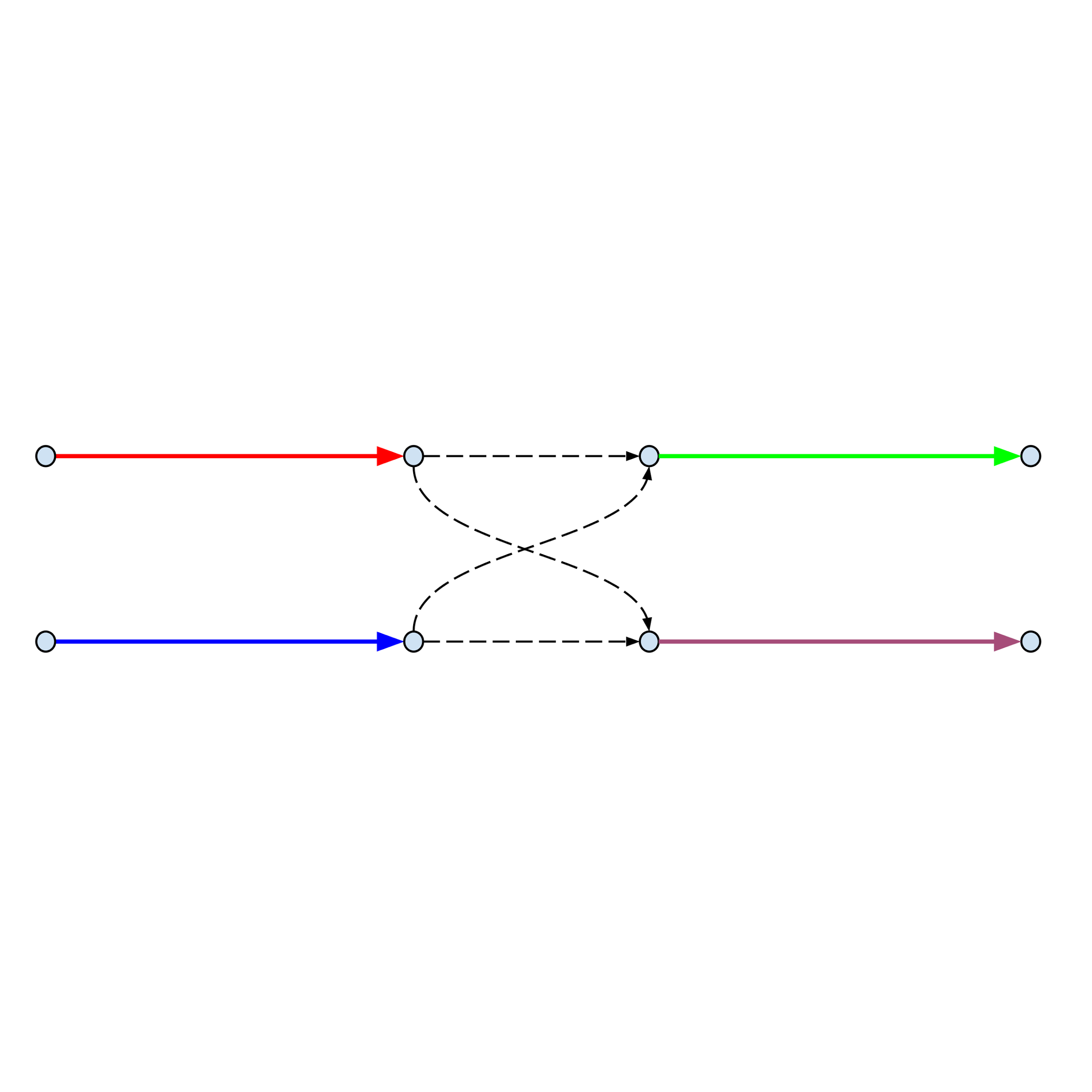
Here 1 is the identifier of the BGC sequence, 60699 is its length in nucleotides, 3 is the number of BGC subgraph that generated this sequence, and 2 is the number of rural postman routes generated from this subraph. We output at most 50 putative paths for each BGC subgraph according to their order in the Depth First Search traversal.

* <output\_dir>/bgc\_in\_gfa/ folder contains the GFA file for each BGC subgraph. These files contain the assembly graph structures and can be visualized with tools such as Bandage (Wick et al. 2015)
* <output\_dir>/bgc\_statistics.txt contains information about each BGC subgraph and each rural postman route generated from this subgraph. For the BGC subgraph, it shows the number of domains, the number of strong and weak edges in the corresponding scaffolding graph, and the predicted BGC type (PK, NRP, PK/NRP, or not known). For each rural postman route, it shows an arrangement of domains and positions of domains on putative BGC sequence.

**Appendix C: Putative NRP BGCs in the CYANO dataset**

The CYANO dataset proved to be a rich source of natural products (Kleigrewe et al. 2015, Boudreau et al. 2015, Cummings et al. 2016). It is also a difficult test for the biosyntheticSPAdes algorithm for the following three reasons:

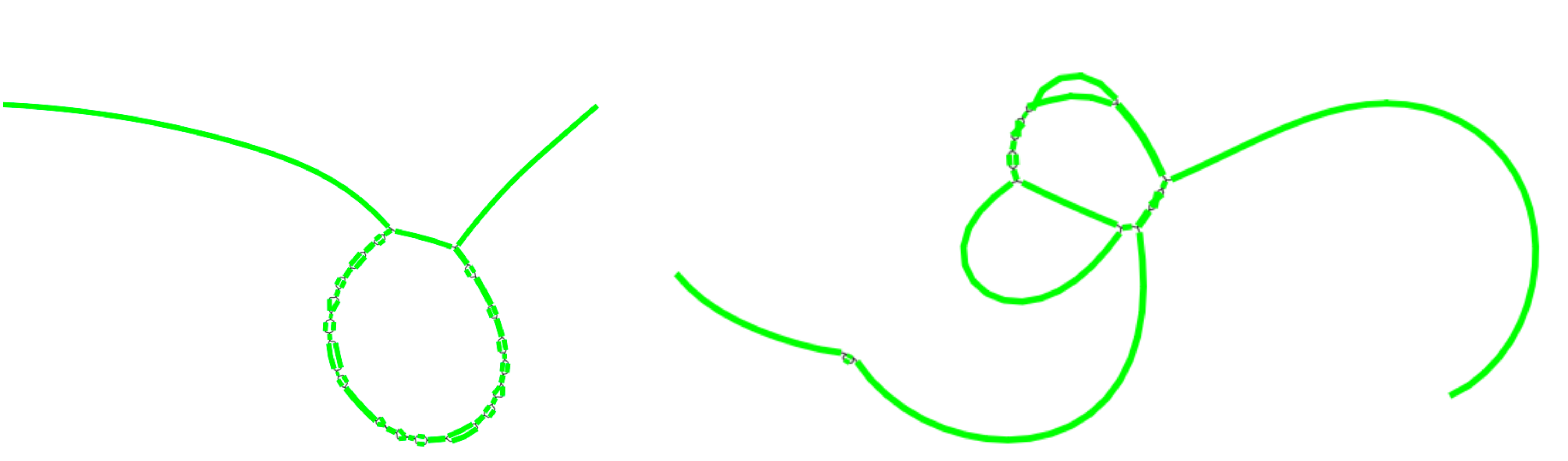
* Although the heterotrophic bacterial contaminants in this dataset encode some BGCs, these BGCs are difficult to reconstruct due to the low depth of coverage. E.g., biosyntheticSPAdes identified three NRP synthetases (108 kb, 44 kb, and 43 kb in length) with low mean coverage 8X arising from some low-abundance bacteria.
* Since some BGCs are located in close proximity to each other in the assembly graph, a single BGC subgraph and corresponding scaffolding graph may contain domains from several BGCs, thus preventing the rural postman approach from finding feasible routes (Supplementary Figure S3). Reconstruction of BGCs from such BGC subgraphs is challenging since they often span complex repeat structures. For example. biosyntheticSPAdes identified a BGC subgraph with more than 200 domains in the CYANO dataset. Such BGC subgraphs may encode dozens of BGCs.
* Many BGCs have highly similar domains resulting in domain collapsing. To perform domain restoration, one has to estimate how many domains were collapsed on a single edge in the assembly graph, which becomes challenging due to variations in the coverage depth.



**Supplementary Figure S3. An example of the scaffolding graph without a rural postman route.** This scaffolding graph is likely formed by two BGCs.

Despite the fact, that biosyntheticSPAdes faced all three challenges analyzing the CYANO dataset, it reconstructed five putative NRP synthetases with complexities 20, 9, 5, 5, and 2, respectively. Our analysis revealed that the CYANO sample contains novel BGCs that fell under the radar of previous studies (based on extensive manual curation) but were reconstructed by biosyntheticSPAdes.

The BGCs with complexities 20 and 9 likely originated from the low-coverage contaminant bacteria and their BGC subgraphs include multiple isolated edges. SPAdes/metaSPAdes combined them into a single scaffold but the nucleotide sequence of this scaffold contains stretches of Ns, making it difficult to infer the nucleotide sequences of the domains. Another possibility is that these two putative NRP synthetases represent parts of a single NRP that was not assembled into a single contig by SPAdes/metaSPAdes. Although two BGCs with multiplicity five have complex BGC subgraphs with loops and long repeats, (Supplementary Figure S4), there exist single rural postman routes in their scaffolding graphs. Since AntiSMASH analysis did not reveal any similarities with known BGCs, they likely represent novel NRP synthetases. The NRP synthetase with complexity 2 has a simple graph structure (all domains lie on a single edge of the assembly graph) and is similar to the known aeruginoside BGC (52% gene similarity and consistent gene order).

****

**Supplementary Figure S4. Two complex BGC subgraphs from the CYANO dataset visualized with the Bandage tool (Wick et al. 2015).** Grey edges represent edges of the assembly graph and each union of connected black edges represents a vertex of the assembly graph.

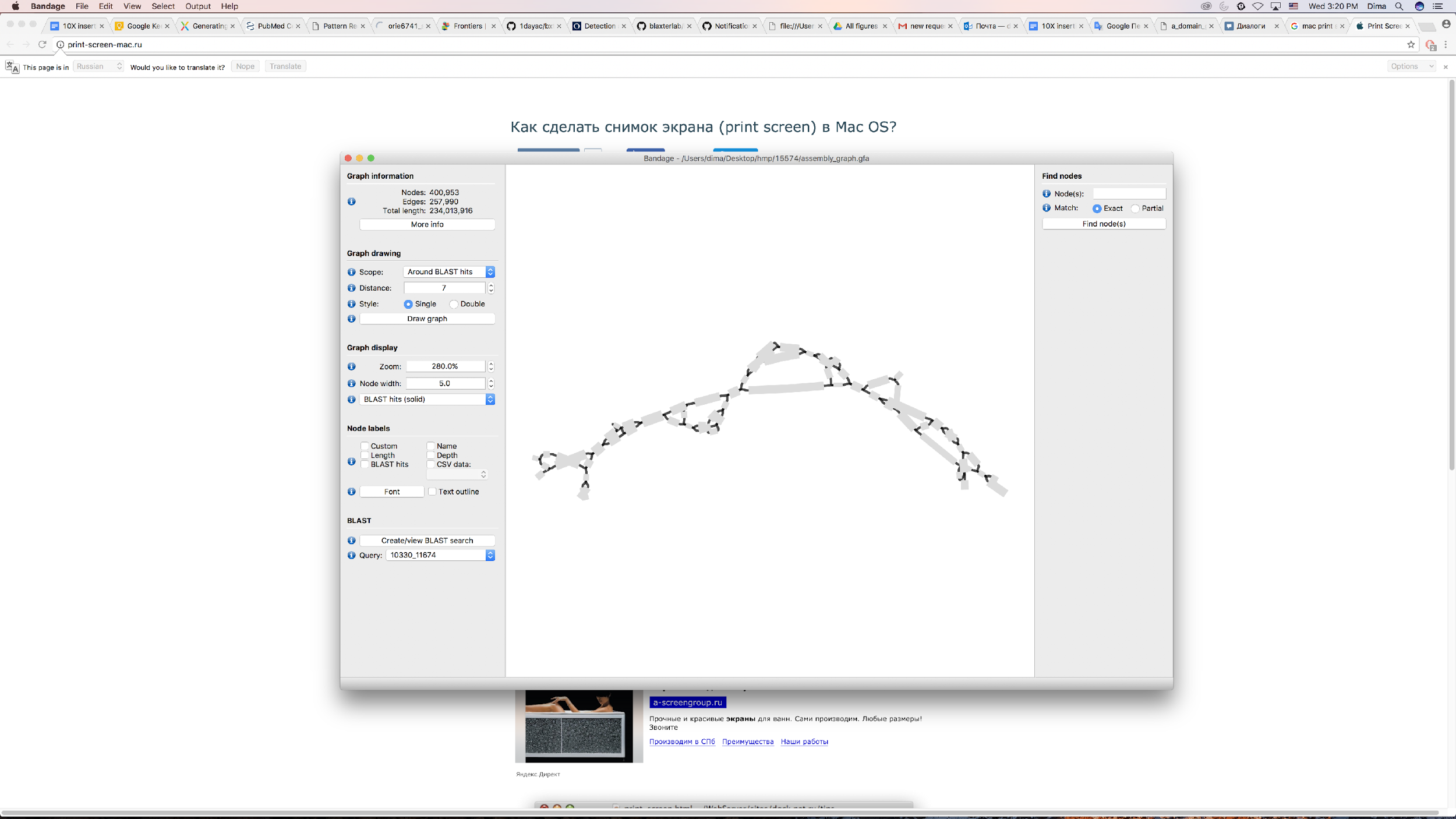
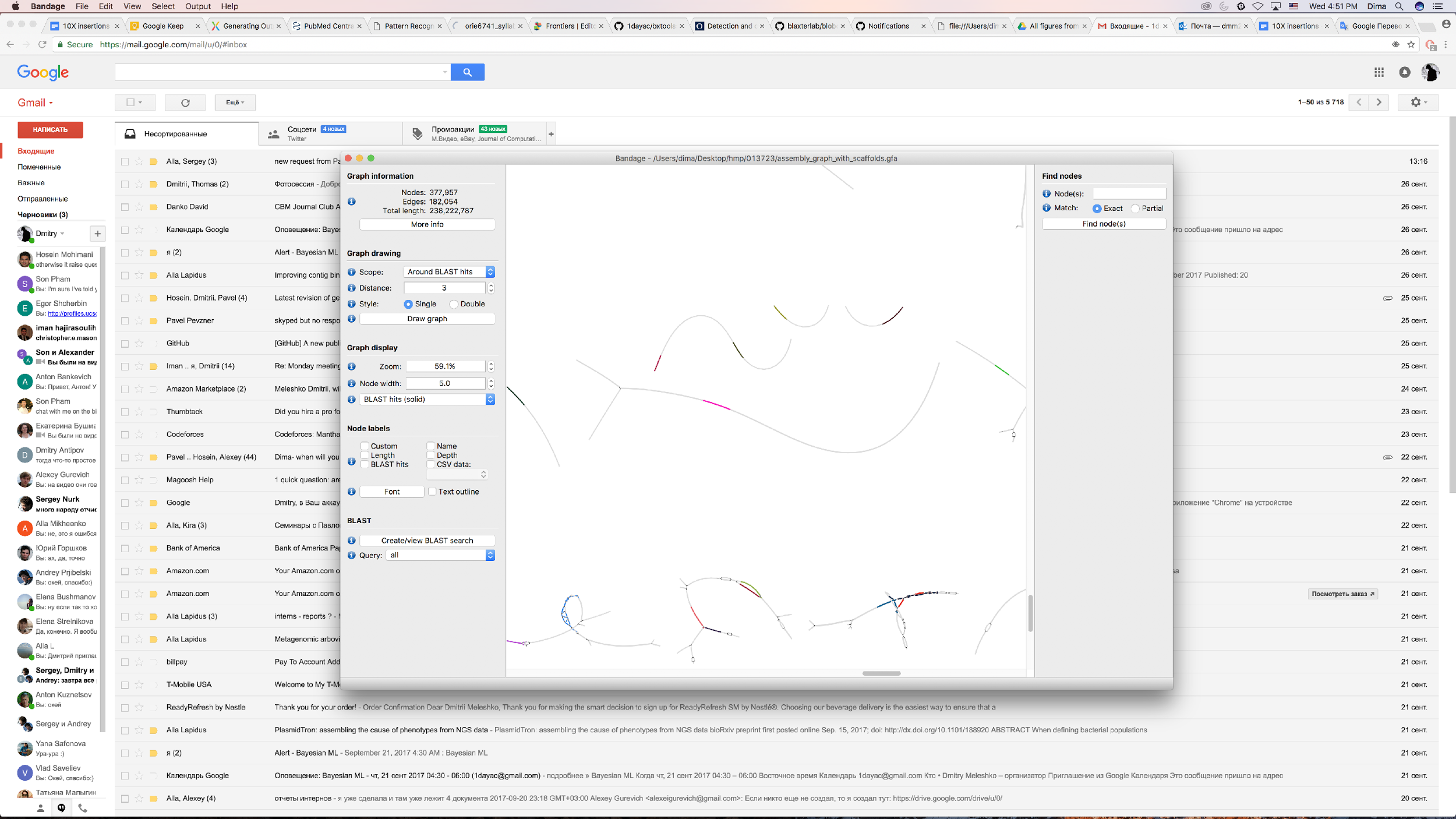
**Appendix D: Biosynthetic capacity of the HMP datasets**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | dataset  ID | total length of long contigs (Mb) | N50 (kb) | #A- domains | #AT-domains | # A/AT-domains per 1 Mb | # BGC subgraphs with complexity 1-3 | # BGC subgraphs with complexity 4-6 | # BGC subgraphs with complexity ≥7 |
| Keratinized gingiva | 019125 | 50.5 | 44,0 | 57 | 24 | 1.60 | 12 | 0 | 0 |
| 014473 | 41.2 | 6,2 | 60 | 36 | 2.33 | 15 | 0 | 0 |
| 015060 | 47.3 | 3,8 | 61 | 50 | 2.34 | 18 | 1 | 0 |
| Buccal mucosa | 018443 | 129.5 | 3,6 | 166 | 83 | 1.92 | 47 | 1 | 0 |
| 023930 | 29.3 | 12,1 | 41 | 19 | 2.05 | 10 | 0 | 0 |
| Stool | 052697 | 211.0 | 12,3 | 287 | 85 | 1.76 | 74 | 0 | 0 |
| 011239 | 136.2 | 8,4 | 215 | 48 | 1.93 | 45 | 0 | 0 |
| 016335 | 189,5 | 7,0 | 268 | 62 | 1.69 | 51 | 0 | 0 |
| Gingivival plaque | 013950 | 95.8 | 4,0 | 142 | 72 | 2.23 | 40 | 0 | 0 |
| 063215 | 76.7 | 3,3 | 169 | 52 | 2.88 | 26 | 4 | 0 |
| 019029 | 112.8 | 2,6 | 147 | 76 | 1.98 | 28 | 2 | 0 |
| Subpravingal plaque | 013723 | 149.0 | 3,4 | 242 | 117 | 2.41 | 59 | 7 | 0 |
| 015574 | 149.3 | 3,3 | 258 | 104 | 2.42 | 67 | 7 | 1 |
| 049318 | 221.2 | 3,9 | 300 | 124 | 1.92 | 80 | 3 | 0 |
| Tongue dorsum | 050244 | 174.3 | 5,8 | 204 | 83 | 1.65 | 47 | 4 | 0 |
| 024081 | 144.0 | 8,5 | 176 | 66 | 1.68 | 47 | 3 | 0 |
| 015762 | 168.3 | 6,3 | 208 | 85 | 1.74 | 57 | 3 | 0 |
| Throat | 019127 | 91.6 | 4,6 | 136 | 46 | 1.97 | 34 | 0 | 0 |
| 019027 | 76,4 | 4,2 | 90 | 40 | 1.70 | 23 | 1 | 0 |
| 014689 | 63,2 | 4,1 | 76 | 38 | 1.80 | 17 | 1 | 0 |

**Supplementary Table S4: Statistics of A-domains and AT-domains in various samples from the HMP dataset.** Long contigs are defined as contigs longer than 1 kb. Dataset identifier is the numerical part of the SRX accession id.

**Appendix E: Putative NRP synthetases in the subpravingal plaque samples from the HMP dataset**

Subpravingal plaque samples from the HMP dataset contain more nontrivial BGCs as compared to the samples from other human body sites. biosyntheticSPAdes identified 18 non-trivial BGC subgraphs in three subpravingal plaque datasets, including (i) 5 BGCs with high-coverage edges, (ii) 10 BGCs without repetitive regions but with coverage gaps, and (iii) 3 BGCs with coverage gaps and complex BGC subgraphs. Supplementary Figure S5 provides examples of two BGC subgraphs from categories (ii) and (iii).



Supplementary Figure S5. Two low coverage BGC subgraphs from subpravingal plaque samples from the HMP dataset. Both subgraphs were visualized using Bandage tool (Wick et al. 2015). Grey edges represent edges of the assembly graph and each union of connected black edges represents a vertex of the assembly graph. (Left) A BGC subgraph for a low coverage region with coverage gaps. A and AT-domains are shown by different colors. (Right) A fragment of a BGC subgraph with low coverage and complex repeat structure. Each non-repetitive edge has coverage between 3X and 5X. This fragment of the assembly graph contains at least three AT-domains but none of them was assembled into a single contig. As the result, only parts of these domains were identified by HMMer. Corresponding scaffolding graph for this BGC subgraph doesn’t contain any rural postman routes.

The assembly graph in category (ii) are simple but their nucleotide sequences are incomplete with many gaps (represented as multiple stretches of Ns). These gaps lead to difficulties in the cases when the domain sequence falls into the gaps. Also, it is not clear how to determine whether a reconstructed putative BGC is complete in the case of low coverage. For example, if the first and the last domains are located near the end of the putative sequence of the BGC, it is not clear whether the BGC is complete as some of its domains can be located in another BGC subgraph.

In contrast, biosyntheticSPAdes recovered all BGC with high coverage, including the one with a complex repeat structure analyzed in the main text (Supplementary Table S5)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| BGC subgraph | predicted type | # domains | # rural postman routes | domain  arrangement |
| 1 | NRPS | 14 | 2 | TE-TE-A-C-A-C-A-C-A-C-A-C-A-C  TE-TE-A-C-A-C-A-C-A-C-A-C-A-C |
| 2 | NRPS/PKS | 12 | 1 | C-A-KS-AT-C-A-KS-C-A-KS-TE-KS |
| 3 | NRPS | 18 | 1 | A-C-A-C-A-C-A-C-A-C-A-C-A-C-A-C-TE-TE |
| 4 | NRPS/PKS | 13 | 1 | KS-AT-KR-TE-C-A-C-C-KS-AT-KR-C-A |
| 5 | NRPS/PKS | 10 | 1 | A-C-A-KS-AT-C-A-C-TE-TE |

**Supplementary Table S5. Statistics of five putative BGCs from the subpravingal plaque datasets with high coverage depth.**

**Appendix F: Reference-based putative BGC ranking algorithm**

If biosyntheticSPAdes outputs several putative BGCs (*pBGCs*) for a single BGC gene cluster, it is not clear which of them is correct. In such cases, biosyntheticSPAdes uses a BGC ranking algorithm to compare each putative BGC against all reference BGCs (rBGCs) from a database of all BGCs from the reference genome sequences, and report the pair of pBGC and rBGC that are most similar to each other.

First, the order and positions of all domains in a pBGCs and all reference rBGCs are predicted with antiSMASH. For each pBGC-rBGC pair, biosynthetiSPAdes constructs a bipartite graph, where nodes are domains and edges connect a domain in pBGC with a domain rBGC if both these domains have the same type, e.g., A-domains. The edge weight is defined as the amino acid sequence similarity for the corresponding domain pair. biosynthetiSPAdes further computes the *maximum-weight matching* in the constructed bipartite graph using the Hungarian algorithm (Kuhn et al., 1955) (Supplementary Figure S6). The matching nodes in the maximum-weight matching are referred to as the *domain twins*.



**Supplementary Figure S6. Reference-based ranking of two fictional putative BGCs (pBGC1 and pBGC2) according to their similarity to an rBGC in the antiSMASH-DB database.** To find which of two pBGC has a better match with the rBGC, the Hungarian algorithm determines domain twins between each pBGCs and the rBGC. Black and red arrows connect twin domains, red arrows further connect twin domains which will lower the score between rBGC and pBCG1 as the domain order in pBGC1 does not match the reference.

The closest rBCG from the database is taken based on the *Domain Sequence Similarity* (*DSS*) score described below.

The similarity score between two BGC clusters should take into account the sequence similarity, the domain composition, and the ordering of the domains. We also use a concept of highly similar domains – *domain twins* to find sequence similarity only between relevant domains of BGCs. We find a set of domain twins of a pBGC and rBGC as follows:

1. Construct a bipartite graph , where is the set of nodes that correspond to the domains of the first BGC, is the set of nodes that correspond to the domains of the second BGC, and is the set of edges that connecting pairs of domains from *U* and *V* of the same type (e.g. A-domains, C-domains, etc.)
2. Compute the similarity score between all pairs of domains of the same type as the amino acid sequence identity of their alignment. The weight of the edge between two domains in the bipartite graph is defined as the similarity score between these domains.
3. Find the maximum weight matching in the bipartite bipartite graph using the Hungarian algorithm (Kuhn, H. W., 1955). Pairs of domains connected by an edge from the maximum weight matching are called the *domain twins*.

To find a best matching pBGC-rBGC pair, we define the Domain Sequence Similarity (DSS) score. The DSS score is a measure of similarity between the amino acid sequences of twin domains between two BGCs. DSS also penalizes for domains that have no twin or different ordering of twin domains.

Let be the subset of edges in the maximum weight matching for anrBGC-pBGC Pair, and *DT* be a set of *domain types* (e.g. A-domains, C-domains, etc.). Given a BGC, we refer to the number of domains of the specific *type* in this BGC *Ntype(BGC).* Given the order of the twin domains in an rBGC and a pBGC , we analyze all domain twins and and classify a pair as an *inversion* if and . We define the *inversion index* as the total number of inversions between an *rBGC* and a *pBGC* divided by the, the maximum possible number of inversions between two permuatations of length |*M*|Given an rBGC-pBGC Pair (*rBGC, pBGC*) we define its *Domain Sequence Similarity* score *DSS*(*rBGC, pBGC*) as follows:

where is the subset of edges of the given *type* in the maximum weight matching and is the weight of an edge e in the bipartite graph. Note that the DSS score penalizes domains that do not participate in twin pairs.

Given a set of putative BGCs and a set of reference BGCs, biosyntheticSPAdes selects an rBGC-pBGC Pair with the maximum DSS score and outputs the pBGC from this pair as the most likely solution.

**Appendix G: Ranking putative BGCs from *Streptomyces coelicolor* A3(2) and *Streptomyces avermitilis* MA-4680**

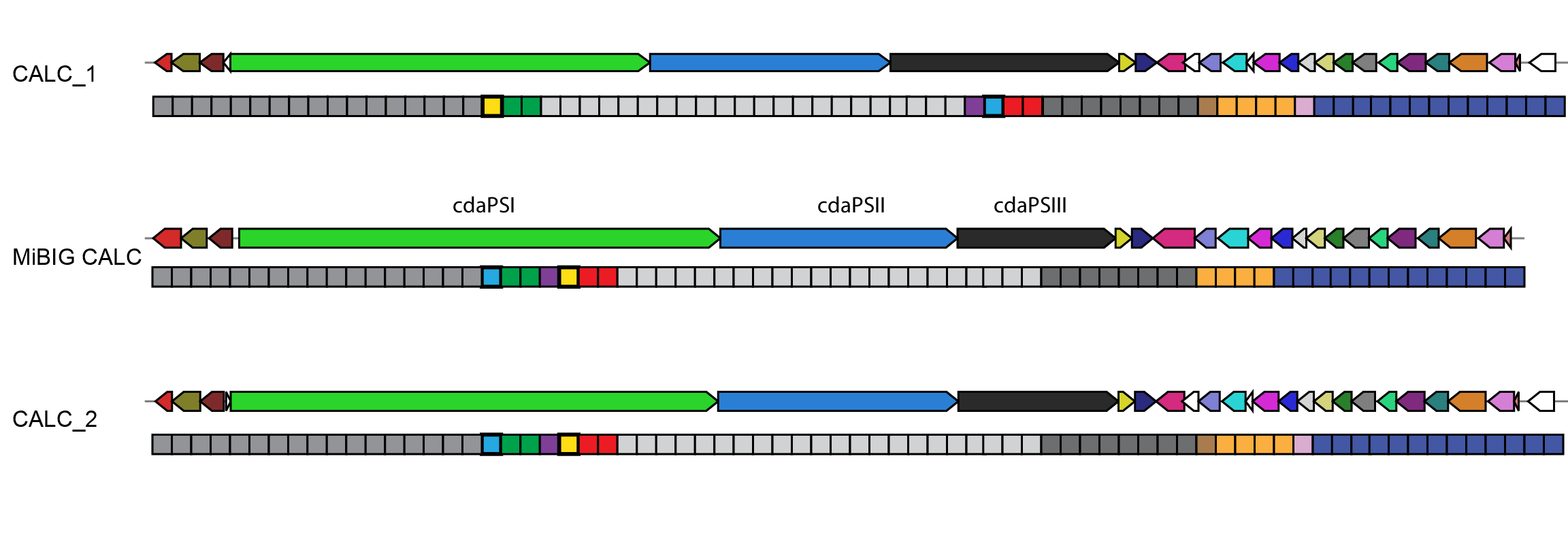
biosyntheticSPAdes assembly of the calcium dependent antibiotic (CALC) NRPS in *S. Coelicolor* produced two putative BGCs that we refer to as CALC\_1 and CALC\_2. These two putative BGCs were scored against all BGCs in antiSMASH-DB (excluding CALC itself) to identify which putative BGC is the most similar to known BGCs. To illustrate our approach, we analyzed an rBGC with the highest DSS scores against both CALC\_1 and CALC\_2: calcium dependent antibiotic BGC from *Streptomyces lividans* TK24. The rBGC chosen using the DSS score belong to the same genus suggesting that the concept of the DSS score helps to identify the correct domain order.

Since the MiBIG database contains the CALC BGC from *S. Coelicolor A3(2)* database, it was possible to also compare the two putative BGCs to their annotated version in MiBIG. Table S6 and Figure S6 illustrate that CALC\_2 has higher domain order consistency and achieves higher DSS score with both the rBGC from antiSMASH-DB and MiBIG making it the best candidate for the biosyntheticSPAdes assembly.

|  |  |  |
| --- | --- | --- |
|  | *S. Lividans* TK24 CALC | *S. coelicolor* CALC |
|  | DSS | DSS |
| CALC\_1 | 0.250 | 0.276 |
| CALC\_2 | 0.253 | 0.307 |

**Supplementary Table S6. Comparing the DSSs between the two putative BGCs and the two reference BGC from antiSMASH-DB and the reference BGC from MiBIG.**

The domain twins generated by the Hungarian algorithm reveal significant differences between the domain structures produced by the rural postman algorithm for the two putative CALC BGCs which affect the order of entire genes within the gene cluster (Supplementary Figure S7).



**Supplementary Figure S7. The domain orders of CALC\_1, CALC\_2 and reference CALC from MiBIG.**   
The cdaPSI and cdaPSIII genes from the reference were matched with the green and black labeled genes in CALC\_1 and CALC\_2. However, the cdaPSI gene in CALC\_1 is shorter than the corresponding gene in the reference and in CALC\_2, while the cdaPSIII gene (in black) is longer in CALC\_1 compared to the reference and CALC\_2. These differences are due to an incorrect assembly in CALC\_1. This indicates that CALC\_2 is the better candidate among the two.

We also assembled the genome of *Streptomyces avermitilis* MA-4680 (Ikegami et al., 2015), which contains a complex repeat-rich gene cluster that produced 6 candidate BGCs from the assembly graph. The ranking algorithm compared the pBGC structures with the filipin BGC, a polyketide synthase BGC, which is present in both antiSMASH-DB and MIBiG (accession: BGC0000059). Supplementary Table S7 illustrates that two out of six candidate BGCs (FILIPIN\_2 and FILIPIN\_6) produced an identical domain arrangement and the highest-ranking candidate was chosen based on small differences in amino acid sequence.

|  |  |
| --- | --- |
| **Putative BGC** | **Correctly ordered domain twins** |
| FILIPIN\_2 | 102/125 |
| FILIPIN\_6 | 102/125 |
| FILIPIN\_3 | 100/125 |
| FILIPIN\_1 | 100/125 |
| FILIPIN\_5 | 100/125 |
| FILIPIN\_4 | 99/125 |

**Supplementary Table S7.** Number of domain twins which had the same order between the putative BGC structure and the reference FILIPIN from antismash-db. The highest-ranking putative structures FILIPIN\_2 and FILIPIN\_6 have identical domain order. The tie is broken by the DSS score, which indicated that FILIPIN 2 putative BGC had higher sequence similarity to the reference.

Supplementary Figure S8 illustrtaes that the domain architecture for candidates 2 and 6 is more similar to the reference BGC domain architecture compared to lower-ranking pBGCs such as candidate FILIPIN\_5.



**Supplementary Figure S8. The domain orders of two of the FILIPIN putative BGCs and reference FILIPIN from AntismashDB.** The domains are color coded to represent blocks with conserved order in the three BGCs even when considering twin domains. The black arrows highlight an example of relocation of two domains for which the reference agrees on the placement for only one of the putative BGCs, notably the highest scoring putative FILIPIN.

As for other reference-based methods, the ranking is affected by database completeness and correctness. Also, the top-ranking pBGC is not necessarily 100% correct, as complex BGCs with high repeat content can result in misassemblies, even with biosyntheticSPAdes. Therefore, results from the ranking algorithm will give insight on which structure better matches the reference BGC but do not guarantee that the highest-ranking structure is also the actual sequence in the assembled genome. In the case of the filipin BGC, even the top-ranking pBGC has small differences with the reference, indicating that further analysis (e.g., by PCR) would be necessary to confirm the actual structure. We provide this example as a case in point to not blindly trust the results of biosyntheticSPAdes and instead verify them whenever possible.

**References**

Boudreau, P. D., Monroe, E. A., Mehrotra, S., Desfor, S., Korobeynikov, A., Sherman, D. H., ... & Gerwick, W. H. (2015). Expanding the described metabolome of the marine cyanobacterium Moorea producens JHB through orthogonal natural products workflows. *PLoS One*, 10(7), e0133297.

Cummings, S. L., Barbé, D., Leao, T. F., Korobeynikov, A., Engene, N., Glukhov, E., ... & Gerwick, L. (2016). A novel uncultured heterotrophic bacterial associate of the cyanobacterium Moorea producens JHB. *BMC Microbiology*, 16(1), 198.

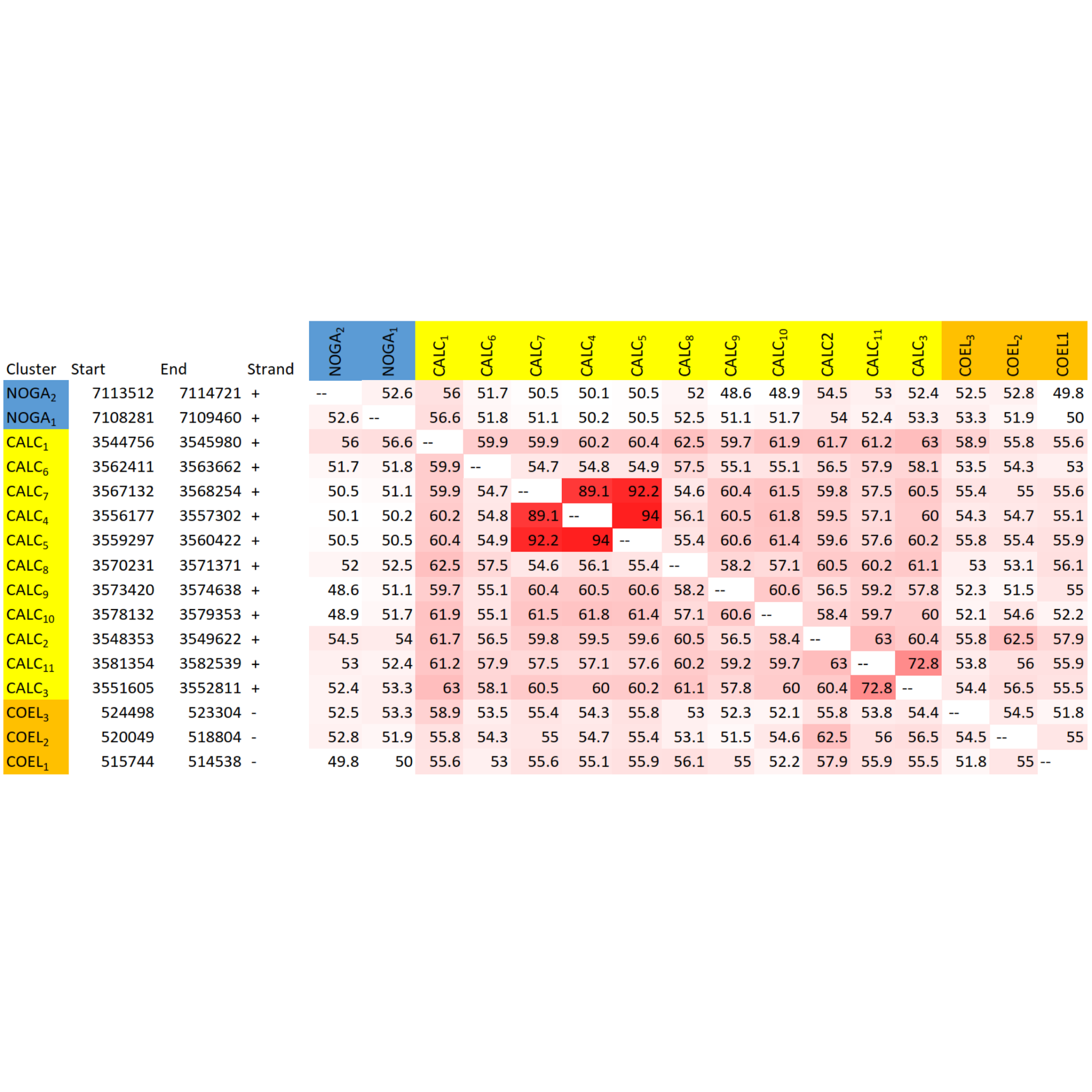
Kleigrewe, K., Almaliti, J., Tian, I. Y., Kinnel, R. B., Korobeynikov, A., Monroe, E. A., ... & Gerwick, L. (2015). Combining mass spectrometric metabolic profiling with genomic analysis: A powerful approach for discovering natural products from cyanobacteria. *Journal of Natural Products*, *78*(7), 1671-1682.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, *2*(1‐2), 83-97.

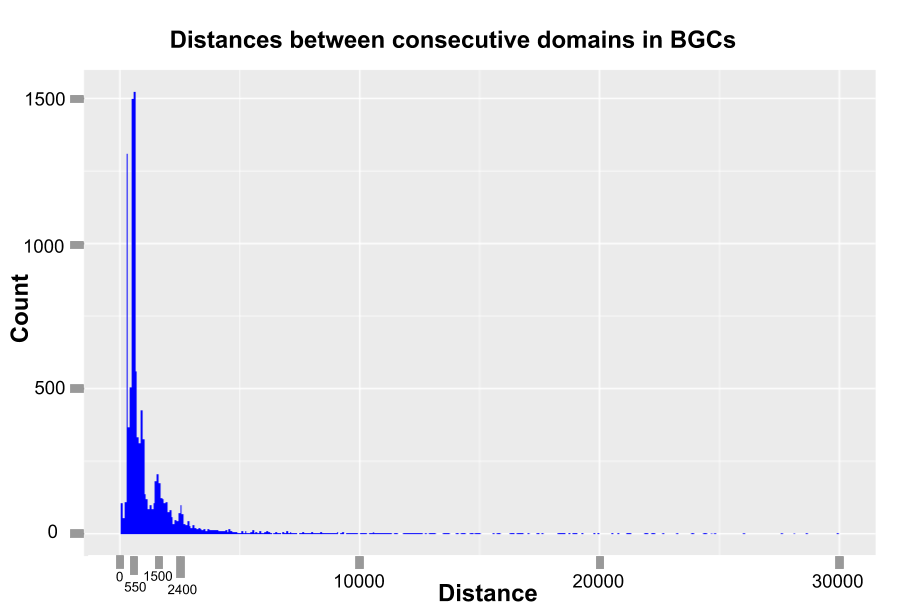
Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L. F., Ninomiya, A., ... & Pevzner, P. A. (2017). Dereplication of peptidic natural products through database search of mass spectra. *Nature Chemical Biology*, *13*(1), 30-37.

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., ... & Porto, C. (2016). Sharing and community curation of mass spectrometry data with GNPS. *Nature Biotechnology, 34(8)*, 828.

Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics, 31(20),* 3350-3352.



**Supplementary Figure S1:** **The percent identity matrix for A-domains in three BGCs in *S. coelicolor A3(2)*.** A-domains CALC4, CALC5, CALC7 (numbers denotes index in the CALC BGC sequence) in the CALC gene cluster (shown as red entries) share identical segments of 96 nucleotides or longer.



**Supplementary Figure S2. Histogram of distances between consecutive domains in NRP and PK BGCs from the MIBIG database.** The distances are computed for A, C, TE, AT, KS and KR domains.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#edges in the assembly graph** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ≥10 |
| **#genes** | 7625 | 112 | 90 | 23 | 35 | 5 | 10 | 1 | 1 | 8 |
| **median gene length** | 849 | 855 | 843 | 1443 | 1158 | 1020 | 1422 | 1011 | 2889 | 1656 |

**Supplementary Table S1.** Number of genes in S. coelicolor A3(2) categorized by the number of edges they traverse in the SPAdes assembly graph (total of 7910 genes). Even after repeat resolution in the assembly graph using exSPAnder, 54 genes in the S. coelicolor genome remain split over multiple scaffolds.

|  |  |  |
| --- | --- | --- |
| locus ID | Gene | gene length |
| **SCO6275** | **type I polyketide synthase** | **13674** |
| **SCO3231** | **CDA peptide synthetase II** | **11013** |
| **SCO0492** | **peptide synthetase** | **10932** |
| **SCO6274** | **type I polyketide synthase** | **10731** |
| **SCO3232** | **CDA peptide synthetase III** | **7254** |
| **SCO6827** | **polyketide synthase** | **7077** |
| SCO6428 | hypothetical protein | 6945 |
| **SCO5892** | **polyketide synthase** | **6894** |
| **SCO0127** | **beta keto-acyl synthase** | **6723** |
| **SCO7682** | **non-ribosomal peptide synthase** | **6690** |
| SCO6220 | hypothetical protein | 6552 |
| **SCO6273** | **type I polyketide synthase** | **6459** |
| **SCO0126** | **beta keto-acyl synthase** | **6249** |
| **SCO7683** | **non-ribosomal peptide synthase** | **5529** |
| SCO5748 | sensory histidine kinase | 5490 |
| SCO2226 | bifunctional alpha-amylase/dextrinase | 5397 |
| SCO3285 | large glycine/alanine rich protein | 5319 |
| SCO1182 | hypothetical protein | 5181 |
| SCO5761 | ATP-dependent DNA helicase | 5073 |
| SCO6687 | DNA-binding protein | 5037 |
| SCO3869 | WD-40 repeat-containing protein | 5031 |
| SCO2999 | hypothetical protein | 4962 |
| SCO6626 | protein kinase | 4674 |
| SCO1407 | hypothetical protein | 4662 |
| SCO2383 | hypothetical protein | 4638 |
| SCO4508 | cell division-like protein | 4578 |
| **SCO2026** | **glutamate synthase** | **4545** |
| SCO2499 | transport ATPase | 4419 |
| SCO4009 | bifunctional histidine kinase and regulator | 4392 |
| SCO7015 | glycosyl hydrolase | 4302 |
| **SCO6432** | **peptide synthase** | **4224** |
| SCO5710 | large Pro/Ala/Gly-rich protein | 4101 |
| SCO6348 | hypothetical protein | 4086 |
| SCO2450 | Ser/Thr protein kinase (regulator) | 4050 |
| SCO2975 | hypothetical protein | 4038 |
| SCO2599 | hypothetical protein | 4023 |
| SCO2259 | multidomain-containing protein family | 4005 |
| SCO7327 | two-component system sensory histidine kinase | 3996 |
| SCO5544 | hypothetical protein | 3990 |
| SCO4092 | ATP-dependent helicase | 3984 |
| SCO5397 | large Ala/Glu-rich protein | 3981 |
| SCO5734 | ATP/GTP binding protein membrane protein | 3966 |
| SCO1184 | hypothetical protein | 3963 |
| SCO6457 | beta-galactosidase | 3924 |
| SCO4655 | DNA-directed RNA polymerase subunit beta' | 3900 |
| SCO6635 | bacteriophage resistance gene pglY | 3885 |
| kgd | alpha-ketoglutarate decarboxylase | 3819 |
| SCO6004 | ATP/GTP binding protein | 3807 |
| SCO3033 | integral membrane regulatory protein | 3807 |
| SCO6219 | Ser/Thr protein kinase | 3786 |
| SCO7176 | peptidase | 3762 |
| SCO4263 | transcriptional regulator | 3756 |
| SCO0432 | peptidase | 3738 |
| SCO2763 | ABC transporter ATP-binding protein | 3732 |
| SCO7188 | peptidase | 3720 |
| SCO6572 | glycosyl hydrolase | 3717 |
| SCO0216 | nitrate reductase subunit alpha NarG2 | 3702 |
| SCO6535 | nitrate reductase subunit alpha NarG | 3696 |
| SCO4947 | nitrate reductase subunit alpha NarG3 | 3684 |
| SCO5184 | ATP-dependent DNA helicase | 3669 |
| SCO2446 | peptidase | 3663 |
| cobN | cobaltochelatase subunit CobN | 3654 |
| SCO1554 | nicotinate-nucleotide-dimethylbenzimidazole phosphoribosyltransferase | 3639 |
| SCO6627 | hypothetical protein | 3633 |
| SCO5331 | DNA methylase | 3603 |
| SCO2590 | glycosyltransferase | 3594 |
| SCO5577 | chromosome associated protein | 3561 |
| SCO1739 | DNA polymerase III subunit alpha | 3558 |
| SCO3109 | transcriptional-repair coupling factor | 3555 |
| dnaE | DNA polymerase III subunit alpha | 3540 |
| SCO3947 | ABC transporter | 3519 |
| SCO6688 | hypothetical protein | 3516 |
| **SCO6431** | **peptide synthase** | **3516** |
| SCO3168 | protease | 3516 |
| **SCO1657** | **methionine synthase** | **3513** |
| SCO4969 | regulatory protein | 3504 |
| rpoB | DNA-directed RNA polymerase subunit beta | 3486 |
| SCO5183 | ATP-dependent DNA helicase | 3480 |
| SCO6198 | hypothetical protein | 3471 |
| SCO5280 | ATP-binding protein | 3447 |
| SCO6593 | hypothetical protein | 3444 |
| SCO0488 | hydrolase | 3417 |
| SCO0370 | DNA-binding protein | 3405 |
| SCO7037 | hypothetical protein | 3396 |
| SCO0546 | pyruvate carboxylase | 3375 |
| SCO0072 | hypothetical protein | 3354 |
| SCO4116 | AfsR-like regulatory protein | 3345 |
| SCO2672 | hypothetical protein | 3342 |
| SCO5540 | hypothetical protein | 3336 |
| SCO4250 | hypothetical protein | 3336 |
| SCO5511 | membrane associated phophodiesterase | 3327 |
| **carB** | **carbamoyl phosphate synthase large subunit** | **3309** |
| SCO2637 | serine protease | 3297 |
| SCO5271 | hypothetical protein | 3291 |
| SCO5506 | regulatory protein | 3276 |
| SCO3542 | integral membrane protein with kinase activity | 3270 |
| SCO6994 | hypothetical protein | 3261 |
| SCO0369 | hypothetical protein | 3258 |
| SCO5717 | hypothetical protein | 3252 |
| SCO2549 | Protease | 3204 |

**Supplementary Table S2.** List of 100 longest genes in the Streptomyces coelicolor A3(2) genome. Genes forming BGC genes are shown in bold.

|  |  |  |  |
| --- | --- | --- | --- |
| locus ID | gene | gene length | #contigs |
| **SCO6274** | **type I polyketide synthase** | **13674** | **9** |
| **SCO6273** | **type I polyketide synthase** | **10731** | **7** |
| **SCO3232** | **CDA peptide synthetase III** | **7254** | **2** |
| SCO6270 | oxidoreductase alpha-subunit | 6457 | 2 |
| SCO2599 | hypothetical protein | 4021 | 2 |
| SCO6836 | transcription regulator ArsR | 3994 | 2 |
| SCO5540 | hypothetical protein | 3334 | 2 |
| SCO2000 | ATP-binding RNA helicase | 2997 | 2 |
| SCO6789 | fatty oxidation protein | 2202 | 2 |
| **SCO6275** | **type I polyketide synthase** | **2159** | **3** |
| SCO6082 | glycogen debranching protein | 2107 | 2 |
| SCO5443 | alpha-amylase | 2026 | 3 |
| SCO7327 | two-component system sensory histidine kinase | 2008 | 2 |
| SCO4595 | Oxidoreductase | 1936 | 2 |
| SCO4777 | protein Ser/Thr kinase | 1800 | 2 |
| SCO6661 | glucose-6-phosphate 1-dehydrogenase | 1777 | 4 |
| SCO6659 | glucose-6-phosphate isomerase | 1651 | 3 |
| SCO4296 | chaperonin GroEL | 1626 | 2 |
| SCO4762 | chaperonin GroEL | 1626 | 2 |
| SCO6832 | methylmalonyll-CoA mutase | 1596 | 3 |
| SCO4258 | hydrolytic protein | 1458 | 2 |
| SCO4257 | hydrolytic protein | 1443 | 2 |
| SCO5087 | actinorhodin polyketide beta-ketoacyl synthase subunit alpha | 1404 | 2 |
| SCO2931 | ABC transporter ATP-binding protein | 1275 | 2 |
| SCO5393 | ABC transporter ATP-binding protein | 1270 | 2 |
| SCO2366 | hypothetical protein | 1141 | 3 |
| SCO6837 | arsenic resistance membrane transport protein | 1107 | 2 |
| SCO4594 | 2-oxoglutarate ferredoxin oxidoreductase subunit beta | 1057 | 2 |
| SCO6269 | 2-oxoglutarate ferredoxin oxidoreductase subunit beta | 1051 | 2 |
| SCO4885 | lipoprotein | 1047 | 2 |
| SCO1471 | transposase | 1020 | 3 |
| SCO2632 | transposase | 1020 | 4 |
| SCO4370 | transposase | 1020 | 3 |
| SCO4698 | IS1652 transposase | 1020 | 3 |
| SCO4183 | transposase | 1018 | 3 |
| SCO5514 | ketol-acid reductoisomerase | 999 | 3 |
| SCO0091 | IS1652 transposase | 957 | 2 |
| SCO0368 | transposase | 957 | 2 |
| SCO7335 | alpha-amylase | 957 | 3 |
| SCO7803 | insertion element transposase | 957 | 2 |
| SCO5641 | transposase | 955 | 2 |
| SCO7819 | hypothetical protein | 847 | 2 |
| SCO5634 | pseudo | 596 | 2 |
| SCO5292 | ATP/GTP-binding protein | 576 | 2 |
| SCO4061 | hypothetical protein | 556 | 2 |
| SCO6395 | pseudo | 379 | 3 |
| SCO7805 | hypothetical protein | 336 | 2 |
| SCO6403 | hypothetical protein | 309 | 2 |
| SCOr15 | 5S ribosomal RNA | 121 | 2 |
| SCOr04 | 5S ribosomal RNA | 119 | 2 |
| SCOr01 | 5S ribosomal RNA | 118 | 2 |
| SCOr10 | 5S ribosomal RNA | 117 | 2 |
| SCOt05 | tRNA | 72 | 2 |
| SCOt07 | tRNA | 72 | 2 |

**Supplementary Table S3.** The list of 54 genes from Streptomyces coelicolor A3(2) that span multiple contigs even after repeat resolution in the SPAdes assembly graph. The length of genes in this table varies from 72 to 13762 (average length is 2997 nucleotides). Multiple biosynthetic genes (e.g., the genes encoding the calcium-dependent antibiotic) are split over several contigs (shown in bold). Note that in addition to NRPSs and PKSs, e.g. other long genes including 16S RNA genes are also highly fragmented in metagenomic assemblies.