Sequencing Rare Marine Actinomycete Genomes Reveals High Density of Unique Natural Product Biosynthetic Gene Clusters

Supplementary Information

Michelle A. Schorn[1], Mohammad M. Alanjary[2], Kristen Aguinaldo[3], Anton Korobeynikov[4,5], Sheila Podell[1], Nastassia Patin[1], Tommie Lincecum[3], Paul R. Jensen[1,6] Nadine Ziemert[2] and Bradley S. Moore[1,6,7]*

[1] Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, USA.

[2] German Centre for Infection Research (DZIF), Interfaculty Institute for Microbiology and Infection Medicine Tuebingen (IMIT), University of Tuebingen, Germany

[3] Ion Torrent by Thermo Fisher Scientific, Carlsbad, California, USA.

[4] Center for Algorithmic Biotechnology, St. Petersburg State University, St. Petersburg, Russia.

[5] Department of Statistical Modeling, St. Petersburg State University, St. Petersburg, Russia.

[6] Center for Microbiome Innovation, University of California, San Diego, USA

[7] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, USA.

*e-mail: bsmoore@ucsd.edu

Table S1. RMA Strains in this Study

NCBI Accession Number	JGI Taxon OID	Strain	Genus	Sequencing	Collection	Reference(s)
MKJY01000000	2675903202	CNU-125	*Actinomadura*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKKH01000000	2675903201	CUA-896	*Cellulosimicrobium*	In-house	Mexico / sediment	(Patin *et al.*, 2016)
MKKI01000000	2675903203	CNJ-954	*Corynebacterium*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKKG01000000	2596583509	CNJ-863	*Gordonia*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
SRX873596*	2561511136	CNJ-787	*Kocuria*	JGI	Palau / sediment	(Gontang *et al.*, 2010)
MKJW01000000	2675903205	CNJ-770	*Kocuria*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKKB01000000	2675903206	CUA-901	*Kytococcus*	In-house	Mexico / sediment	(Patin *et al.*, 2016)
GCA_000374985.1	2517572149	CNB-394	*Micromonospora*	JGI	Bahamas / sediment	(Mincer *et al.*, 2002)
SRX873600*	2563366738	CNS-044	*Nocardia*	JGI	Palau / sediment	(Gontang *et al.*, 2010)
GCA_000482385.1	2528311129	CNY-236	*Nocardia*	JGI	Fiji / sediment	New to this study
MKKC01000000	2675903207	CNR-923	*Nocardiopsis*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
GCA_000381685.1	2519899670	CNS-639	*Nocardiopsis*	JGI	Fiji / sediment	New to this study
GCA_000515115.1	2515154089	CNT-312	*Nocardiopsis*	JGI	Fiji / sediment	New to this study
MKKA01000000	2675903208	CNJ-824	*Ornithinimicrobium*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKJV01000000	2675903200	CNS-004	*Pseudonocardia*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKJX01000000	2675903209	CNS-139	*Pseudonocardia*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKKD01000000	2675903210	CUA-806	*Rhodococcus*	In-house	Mexico / sediment	(Patin *et al.*, 2016)
MKKE01000000	2675903211	CUA-673	*Saccharomonospora*	In-house	San Diego / sponge	(Patin *et al.*, 2016)
GCA_000527075.1	2515154179	CNQ-490	*Saccharomonospora*	JGI	San Diego / sediment	(Yamanaka *et al.*, 2014)
MKKF01000000	2675903068	CNJ-927	*Serinicoccus*	In-house	Palau / sediment	(Trzoss *et al.*, 2014)
MKJZ01000000	2675903212	CUA-874	*Serinicoccus*	In-house	Mexico / sediment	(Patin *et al.*, 2016)

Table S1. NCBI accession numbers, JGI Organism ID (OID) numbers, genera, sequencing center, country and source of collection and previous references for all strains sequenced as part of this study are included in SI Table 1. Each strain is deposited and annotated in JGI and NCBI. *These accession numbers are for the NCBI Sequence Read Archive (SRA) database.

Table S2. Genome Quality

Genome Name	% GC	Total length	Max scf length	Assembly n50	# Scaffolds	Genome Qual Score
Actinomadura sp. CNU-125	0.72	9,948,691	194,872	32,705	596	0.76
Cellulosimicrobium sp. CUA-896	0.75	3,725,808	196,551	92,221	78	0.88
Corynebacteria sp CNJ-954	0.65	3,773,651	433,327	251,665	45	0.92
Gordonia sp. CNJ-863	0.67	5,398,721	499,033	180,274	94	0.90
Kocuria sp. CNJ-787	0.71	3,637,109	473,511	141,127	83	0.80
Kocuria sp. CNJ-770	0.73	4,120,723	225,243	61,628	164	0.81
Kytococcus sp. CUA-901	0.71	3,637,109	301,417	110,371	51	0.79
Micromonospora sp. CNB-394	0.73	6,344,798	819,458	243,185	85	0.88
Nocardia sp. CNS-044	0.69	7,428,010	387,941	283,581	13	0.88
Nocardia sp. CNY-236	0.65	5,304,668	550,412	110,041	75	0.82
Nocardiopsis sp. CNR-923	0.71	5,546,007	230,239	88,494	149	0.81
Nocardiopsis sp. CNS-639	0.73	6,845,276	403,132	172,911	52	0.86
Nocardiopsis sp. CNT-312	0.73	4,681,465	403,132	172,911	62	0.89
Ornithinimicrobium sp. CNJ-824	0.73	3,445,055	450,180	132,299	62	0.90
Pseudonocardia sp. CNS-004	0.73	9,200,434	550,412	110,041	154	0.94
Pseudonocardia sp. CNS-139	0.74	7,140,900	385,218	99,408	134	0.87
Rhodococcus sp. CUA-806	0.64	5,797,761	424,519	136,942	66	0.87
Saccharomonospora sp. CUA-673	0.70	5,421,117	283,169	122,768	85	0.89
Saccharomonospora sp. CNQ-490	0.71	4,941,689	1,117,442	613,253	25	0.93
Serinicoccus sp. CNJ-927	0.72	3,438,061	409,679	214,309	52	0.82
Serinicoccus sp. CUA-874	0.72	3,521,025	464,901	242,965	39	0.84

Table S2. Assembly statistics and genome quality scores for each strain sequenced as part of this study. Quality scores were calculated according to the parameters set forth as standards for genome assembly quality in (Land *et al.*, 2014). All but two genomes have quality scores above 0.8, which are determined to be good quality assemblies. The two genomes with scores under 0.8 are still usable for analysis, as they are not under 0.6, the lower limit set for genome quality usable in analysis.

Table S3. antiSMASH 3.0 All Clusters

Strain	Genus	Total Clusters	Hybrid	Type 1 PKS	Type 2 PKS	Type 3 PKS	Other KS	NRPS	Terpene	Bacteriocin	Ectoine	Siderophore	Lantipeptide	Oligo-saccharide	Other	Butyrolactone	Phenazine	Nucleoside	Homoserine lactone	Arylpolyene	Indole	Hybrid Types
CNU-125	*Actinomadura*	44 (25)	2	4 (3)	3 (1)	2 (1)	2	18 (3)	6	1	1	1	3	0	0	0	0	0	0	1	0	OtherKS-T1PKS-NRPS, NRPS-Terpene
CUA-896	*Cellulosimicrobium*	3	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	
CNJ-954	*Corynebacterium*	5	0	1	0	0	0	1	2	0	0	0	0	0	1	0	0	0	0	0	0	
CNJ-863	*Gordonia*	20 (17)	2	1	1	0	0	9 (6)	2	1	1	0	0	0	2	0	0	0	0	1	0	NRPS-Siderophore, NRPS-T1PKS
CNJ-787	*Kocuria*	6	1	0	0	0	0	0	0	1	0	1	0	0	3	0	0	0	0	0	0	Terpene-T3PKS
CNJ-770	*Kocuria*	7	1	0	0	0	0	1	0	0	0	1	0	0	4	0	0	0	0	0	0	T3PKS-Terpene
CUA-901	*Kytococcus*	3	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	
CNB-394	*Micromonospora*	34 (27)	9	9 (4)	1	1	0	5 (3)	4	0	0	1	4	0	0	0	0	0	0	0	0	TransatPKS-NRPS-OtherKS, Oligosaccharide-NRPS-Terpene, NRPS-T1PKS, Siderophore-NRPS-Lantipeptide-T1PKS-OtherKS, NRPS-Lantipeptide-T1PKS, Lantipeptide-T2PKS, NRPS-T1PKS, Bacteriocin-Terpene
CNS-044	*Nocardia*	35	4	6	0	2	0	14	2	1	1	0	0	0	2	2	0	0	0	1	0	NRPS-Terpene, NRPS-T1PKS, T1PKS-TransatPKS-Oligosaccharide, Terpene-T1PKS-Butyrolactone
CNY-236	*Nocardia*	38 (31)	5	7 (5)	1	0	0	14 (9)	2	0	1	0	0	0	6	1	0	1	0	0	0	NRPS-Terpene, Amglyccycl-OtherKS, NRPS-Bacteriocin, NRPS-T1PKS, T1PKS-NRPS-OtherKS
CNR-923	*Nocardiopsis*	16	5	3	0	0	0	1	2	0	1	1	2	0	0	1	0	0	0	0	0	OtherKS-T2PKS, NRPS_Lantipeptide, T1PKS-NRPS, Thiopeptide-Lantipeptide, Lantipeptide-T1PKS
CNS-639	*Nocardiopsis*	24 (22)	6	1	1	1	0	5 (3)	2	0	1	1	1	1	1	2	1	0	0	0	0	Bacteriocin-Terpene, Lantipeptide-T1PKS, Lantipeptide-Oligosaccharide, Oligosaccharide-OtherKS-Lantipeptide, NRPS-T1PKS, Indole-NRPS
CNT-312	*Nocardiopsis*	22 (20)	0	3	0	0	0	7 (5)	2	3	3	0	3	0	0	0	1	0	0	0	0	
CNJ-824	*Ornithinimicrobium*	3	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	
CNS-004	*Pseudonocardia*	8	0	0	0	0	0	2	1	1	1	0	0	0	3	0	0	0	0	0	0	
CNS-139	*Pseudonocardia*	14 (13)	1	0	1	0	0	6 (5)	1	1	0	0	1	0	2	1	0	0	0	0	0	Arylpolyene-Butyrolactone
CUA-806	*Rhodococcus*	20 (15)	1	0	0	0	2	12 (7)	1	0	1	0	0	0	2	0	0	0	0	1	0	Oligosaccharide-T2PKS
CUA-673	*Saccharomonospora*	7	3	0	0	0	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	Oligosaccharide-T2PKS, Siderophore-Ectoine, NRPS-T1PKS
CNQ-490	*Saccharomonospora*	25 (21)	2	8 (4)	1	2	0	3	2	3	1	0	0	0	0	0	0	0	1	1	1	NRPS-T1PKS, T1PKS-Siderophore
CNJ-927	*Serinicoccus*	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
CUA-874	*Serinicoccus*	2	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	

Table S3. This table includes the number of pathways for each category as output by antiSMASH 3.0. Genera are colored by total pathway number, with blue = low, green = medium, red = high amount of clusters. Each hybrid cluster is detailed in the last column. For NRPS and PKS clusters joined by NaPDoS (Fig. S1), the final predicted number of putative clusters is in parentheses.

Table S4: antiSMASH 3.0 Table used for Circos

Strain	Genus	Estimated Genome Size (MB)	NRPS-PKS Hybrid	All PKS	NRPS	Terpene	RiPPs	Siderophore	Arylpolyene	"Other"
CUA-874	*Serinicoccus*	3.7	0	0	0	1	1	0	0	0
CNJ-824	*Ornithinimicrobium*	3.7	0	0	0	2	0	0	0	1
CUA-896	*Cellulosimicrobium*	3.9	0	1	0	1	0	0	0	1
CNJ-954	*Corynebacterium*	3.9	0	1	1	2	0	0	0	1
CNJ-770	*Kocuria*	3.9	0	1	1	1	0	1	0	4
CUA-901	*Kytococcus*	4.4	0	1	0	1	0	0	0	1
CNJ-863	*Gordonia*	5.5	1	2	7	2	1	1	1	3
CUA-673	*Saccharomonospora*	5.9	1	1	2	1	1	1	0	2
CNR-923	*Nocardiopsis*	6.2	1	6	2	2	6	1	0	2
CNB-394	*Micromonospora*	6.3	3	10	6	6	8	2	0	1
CUA-806	*Rhodococcus*	7.2	0	3	7	1	0	0	1	4
CNS-044	*Nocardia*	7.4	1	11	15	4	1	0	1	6
CNS-139	*Pseudonocardia*	9.4	0	1	5	1	2	0	0	5
CNU-125	*Actinomadura*	14.6	1	7	5	7	4	1	1	1

Table S4. This table was used to create the Circos diagram (Fig. 1); it includes a representative genome from each genus sequenced and run through antiSMASH 3.0 without ClusterFinder. PKS and NRPS clusters that could be connected by NaPDoS are included in this table (Fig. S1). Hybrid clusters were separated into their composite categories (i.e. an NRPS-Siderophore hybrid is split into an NRPS cluster and a Siderophore cluster) to better assess the spread of cluster categories across genera. The hybrid category now only contains NRPS-PKS hybrids. All types

of PKS clusters were also collapsed into one category. Lantipeptide, Bacteriocin, Thiopeptide, and Lassopeptide are collapsed into the category "RiPPs" (Ribosomally synthesized and Post-translationally modified Peptides). All minor categories, present in less than 5 genomes, were collapsed into the "Other" category, along with clusters designated by antiSMASH as Other. Ectoine was also included in the "Other" category, although ectoine clusters were present in 13/21 genomes. Those categories included in the "Other" category are: Other, Ectoine, Oligosaccharide, Butyrolactone, Phenazine, Nucleoside, Homoserine lactone, Aminoglycoside and Indole.

Figure S1. NaPDoS Cluster Connection

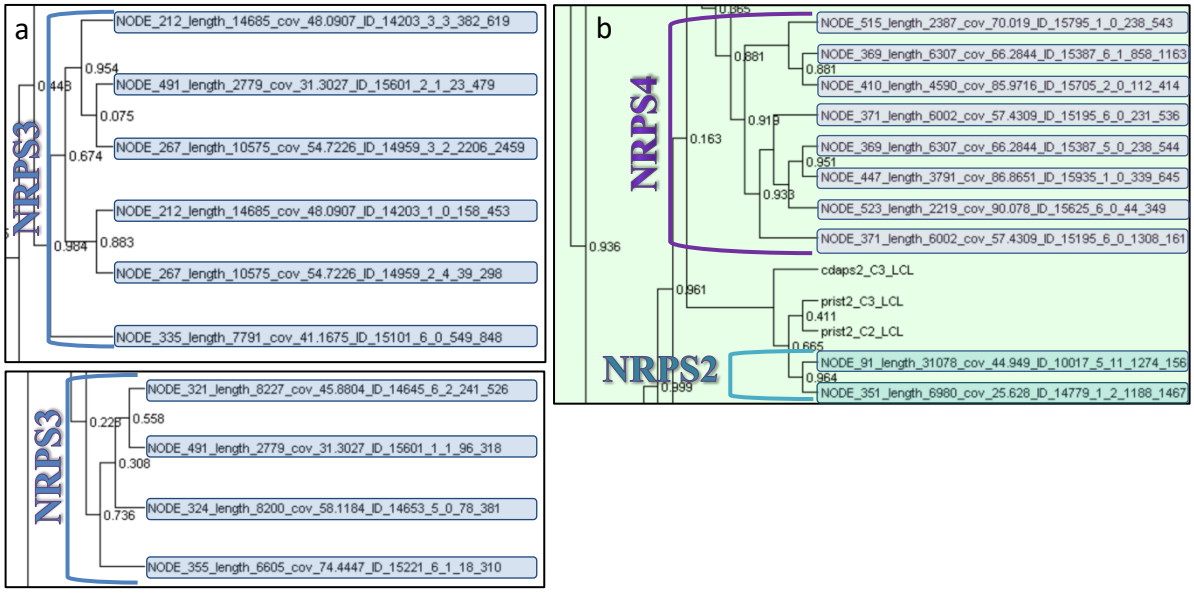
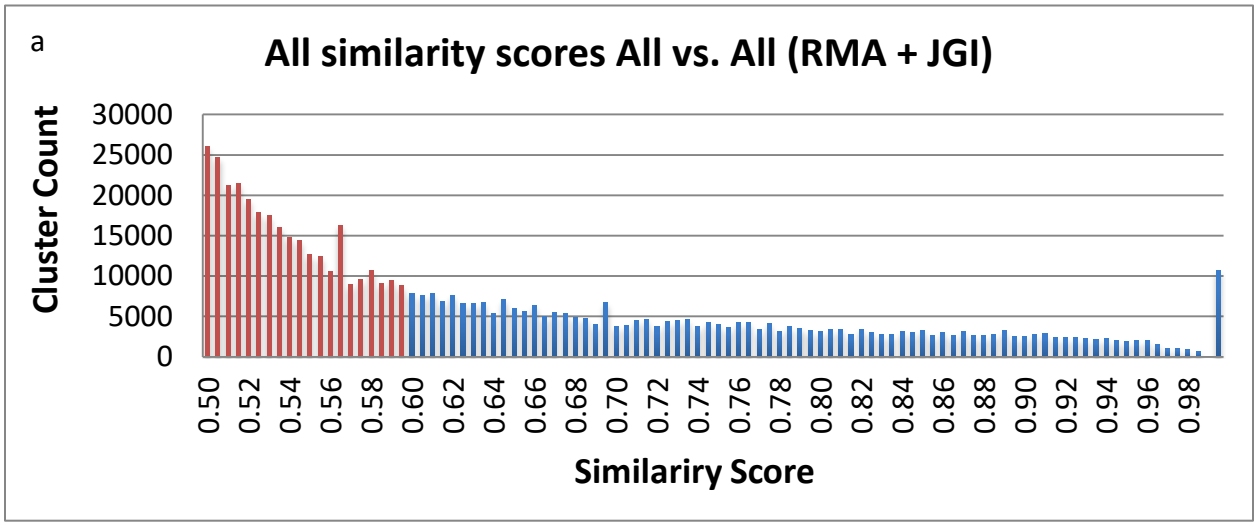


Figure S1. NaPDoS was used to connect NRPS and PKS clusters split onto two or more contigs. For example, *Actinomadura* CNU-125 NRPS3 (a), NRPS4 (b) and NRPS2 (b) clusters are made up of multiple sister taxa condensation domain sequences present on separate nodes (contigs). Secondary metabolite gene clusters can be inherited through horizontal gene transfer from other

phylogenetically distant bacteria. The transferred gene cluster harbors the genetic signature of its historical relative and thus contigs/scaffolds containing pieces of one gene cluster are likely to phylogenetically clade together. While this is not always the case, it is a good tool to narrow down a more accurate number of NRPS/PKS pathways present in fragmented next-generation sequencing assemblies. Genomes with more than three NRPS or PKS clusters, as identified by antiSMASH 3.0 without ClusterFinder, were submitted to NaPDoS and KS and/or C domains were identified and NaPDoS constructed a tree. If the cluster was in the middle of a contig (i.e. has sequence before and after the region antiSMASH identified), it is considered complete. If domains on different contigs were sister taxa in the NaPDoS outputted tree, the clusters on the two or more contigs were considered part of one cluster. The total length of the prospective gene cluster was also taken into consideration. For each genome, the sum of the lengths of all clusters was divided by the average length of all the complete clusters. The resulting measure is the expected number of clusters based on an average length, specific to each genome. These estimates support the joining of clusters using NaPDoS.

Figure S2. Similarity Scores and Mismatch BGC Type suggest 0.6 cutoff



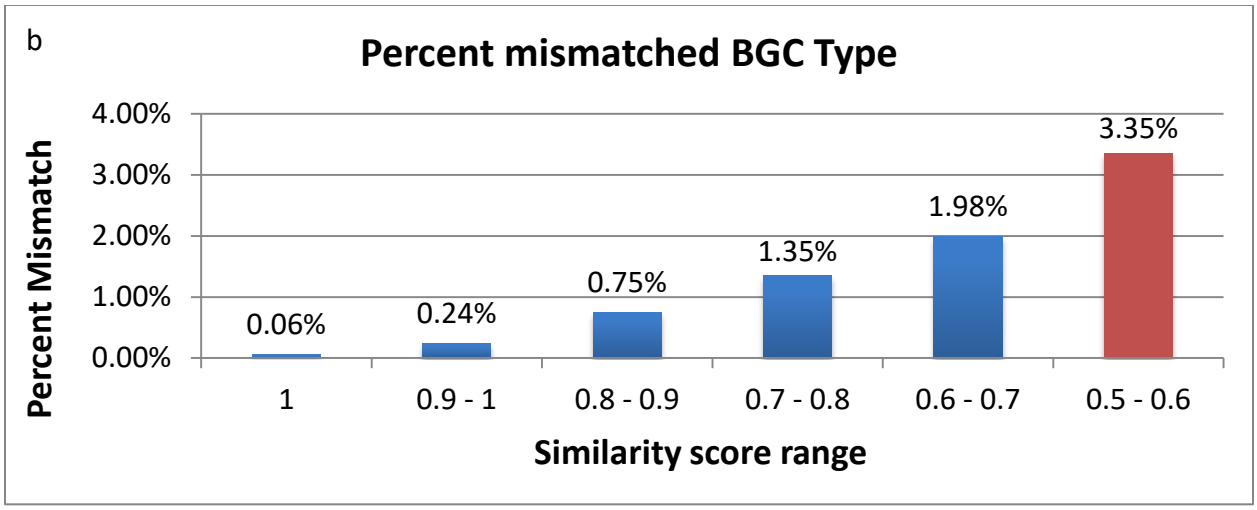
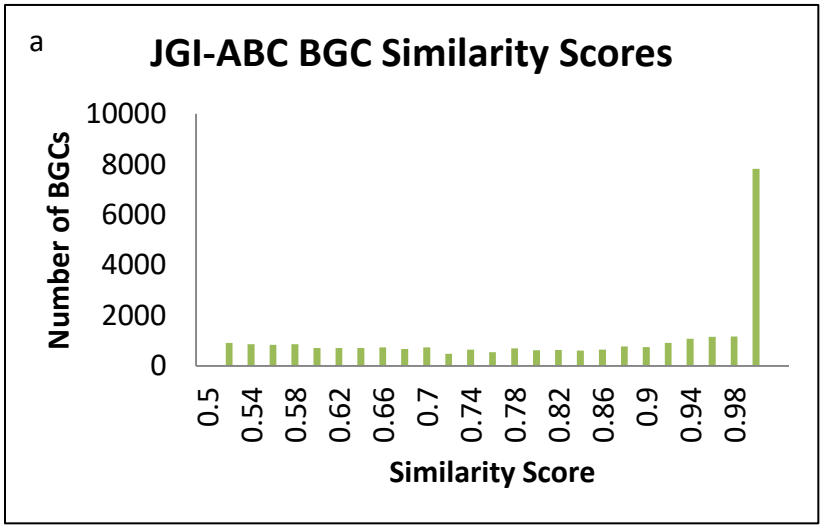


Figure S2. (a) All pairwise similarity scores are shown from 0.5 to 1. A threshold at 0.6 was chosen to significantly reduce the number of edges in the network. (b) Percent of pairwise connections where BGC Type did not match for nodes with BGC Type annotated by JGI. The raise in mismatches between connected nodes for the 0.5 to 0.6 similarity score range corroborates the 0.6 cutoff for clustering.

Figure S3. BGC Similarity Score Distributions



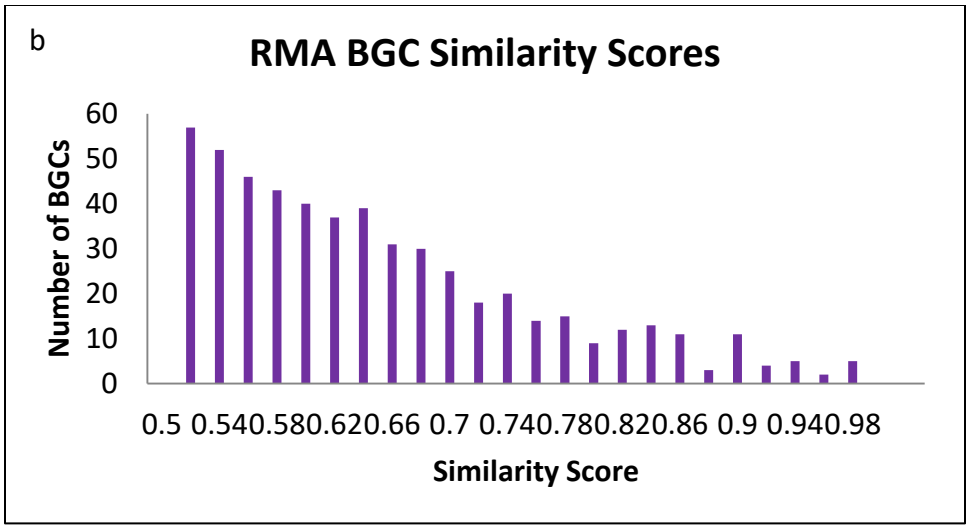


Figure S3. Calculated similarity scores >0.5 are shown for: (a) JGI-ABC clusters versus all clusters. Note the relatively flat distribution of scores. The spike in scores at 1 is due to replicate sequencings of the same strain, and were de-replicated in the similarity network. Similarity scores shown in (b) are from RMA clusters versus all clusters, including self-similarity between RMA clusters. Note that the RMA scores skew more heavily toward lower similarity scores, suggesting that they are more unique than what is currently available in the JGI-ABC database.

The 24 marine *Streptomyces* strains included in the comparison against RMA strains (Table 1) are: *Streptomyces* spp. CNB-091, CNB-632, CNH-099, CNH-189, CNH-287, CNQ-525, CNQ-329, CNQ-766, CNQ-865, CNR-698, CNS-335, CNS-606, CNS-615, CNT-302, CNT-318, CNT-360, CNT-371, CNT-372, CNX-435, CNY-228, CNY-243, TAA-040, TAA-204, and TAA-486.

Table S5. Networking Breakdown by Genus

	RMA							In-Network Cluster Diversity		JGI - Not Marine						In-Network Cluster Diversity	
Genus	**Total Clusters**	**# of Strains**	**Clusters in network**	**# of GCFs**	**% Clusters in Network**	**# Unique genus GCFs**	**% New genus GCFs**	**True Diversity (q=1, D1)**	**True Diversity / Cluster**	**Total Clusters****	**# of Strains**	**Clusters in Network**	**# of GCFs**	**% Clusters in Network**	**RMA GCFs shared**	**True Diversity (q=1, D1)**	**True Diversity / Cluster**
Serinicoccus	99	4	43	21	43%	21	100 %	20.49	**0.4764**	0	0	0	0	0%	0	N/A	**N/A**
Nocardiopsis	292	4	57	38	20%	26	68%	34.93	**0.6128**	1147	18	163	51	14%	12	25.99	**0.1595**
Actinomadura	153	1	27	8	18%	5	63%	4.46	**0.1653**	696	7	81	27	12%	3	20.80	**0.2568**
Saccharomonospora	221	4	40	28	18%	20	71%	24.23	**0.6057**	365	8	77	35	21%	8	26.46	**0.3436**
Pseudonocardia	186	2	11	7	6%	4	57%	6.64	**0.6040**	613	7	35	20	6%	3	16.19	**0.4625**
Micromonospora	209	4	100	52	48%	13	25%	41.97	**0.4197**	1886	40	701	146	37%	39	76.49	**0.1091**
Ornithinimicrobium	35	1	2	2	6%	2	100%	2.00	**1.0000**	27	1	1	1	4%	0	1.00	**1.0000**
Kocuria	123	3	27	17	22%	11	65%	16.16	**0.5984**	131	6	17	13	13%	6	12.27	**0.7217**
Kytococcus	49	2	5	3	10%	3	100%	2.59	**0.5173**	0	0	0	0	0%	0	N/A	**N/A**
Cellulosimicrobium	36	1	2	2	6%	0	0%	2.00	**1.0000**	170	7	39	19	23%	2	16.04	**0.4114**
Nocardia	289	3	57	28	20%	11	39%	18.28	**0.3207**	3718	36	716	176	19%	17	49.23	**0.0688**
Corynebacterium	68	3	19	17	28%	4	24%	17.66	**0.9296**	2165	143	291	84	13%	13	46.52	**0.1598**
Gordonia	66	1	20	17	30%	1	6%	15.16	**0.7579**	1665	31	410	87	25%	16	37.83	**0.0923**
Rhodococcus	244	3	69	46	29%	5	11%	26.72	**0.3872**	3565	46	1094	242	31%	41	95.52	**0.0873**

**After de-replication

Table S5. This table breaks down the number of BGCs and GCFs by genus, comparing the RMA strains (from this study, as well as those labelled as marine in JGI: Figure S5) against the same genera from the JGI-ABC database. Novel contributions, in the form of GCFs not previously present in the JGI-ABC database for that genus, can be seen for each genus sampled in this study. True diversity was calculated according to equation (3) in (Jost & Baños, 2016). This equation is the exponent of the Shannon Index when q = 1.

Figure S4. RMA and Marine-derived *Streptomyces* Network

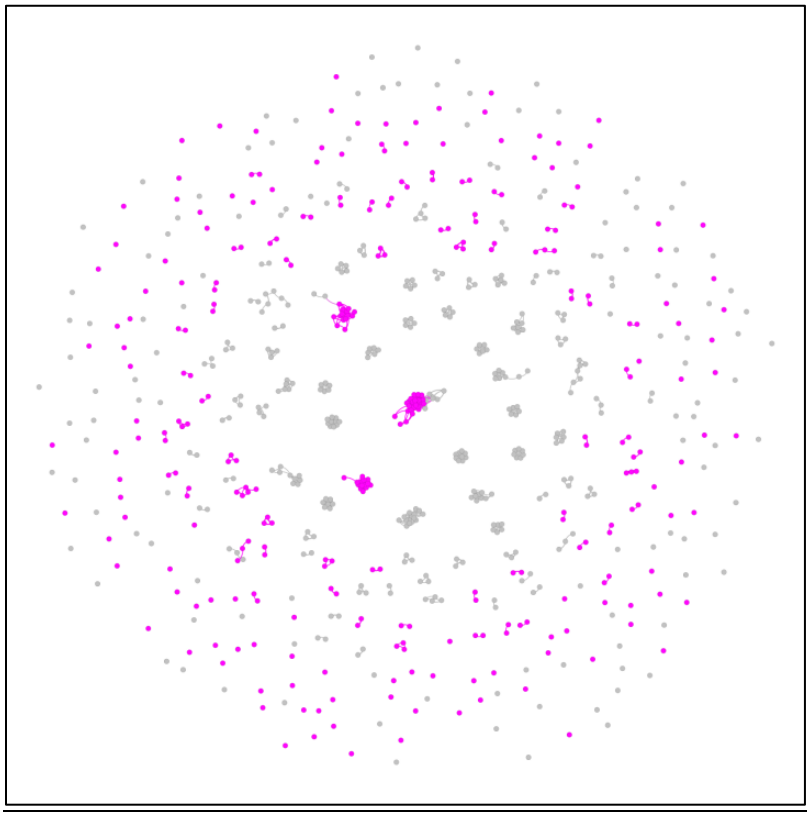


Figure S4. This BGC network includes the 21 RMA strains sequenced as part of this study and 24 marine-derived *Streptomyces* strains from the JGI database. Those nodes colored in pink are BGCs from RMAs and marine-derived *Streptomyces* BGCs are in grey. Notice that there is little overlap between RMA and marine-derived *Streptomyces* BGCs.

Figure S5. Environments/Sources of Genomes in each Genus

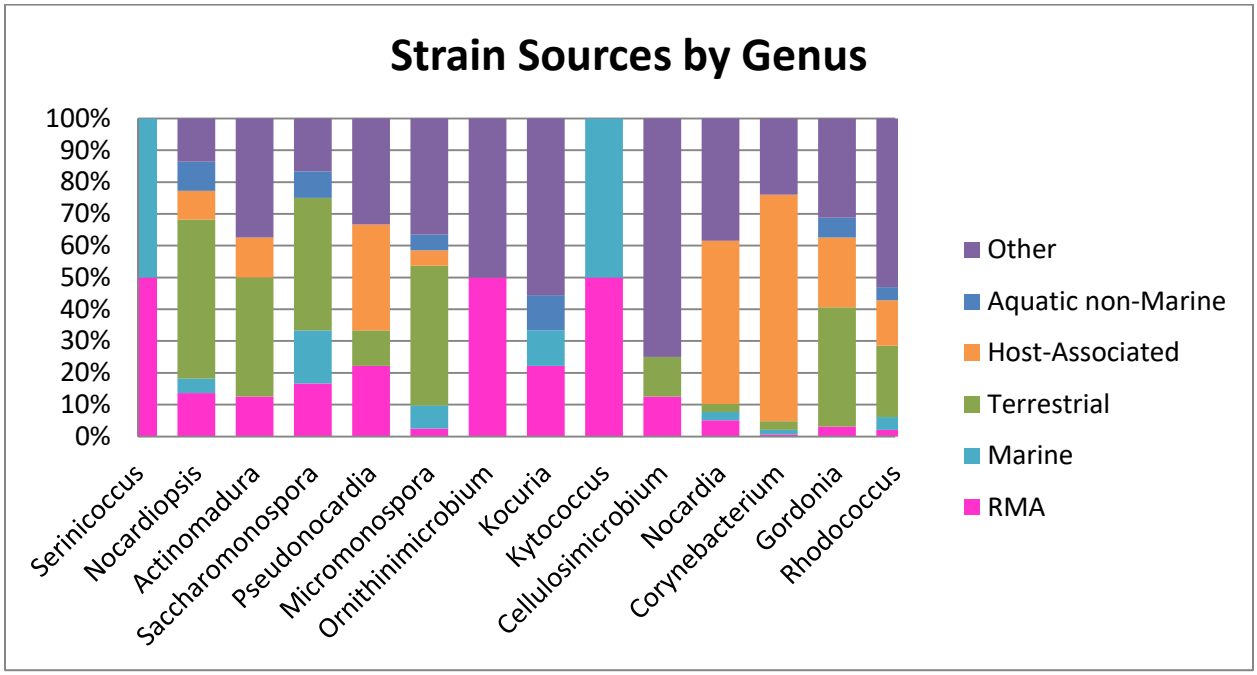


Figure S5. This stacked bar graph shows the sources of genome sequenced strains in JGI for each genus studied. RMA genomes from this study are colored in pink. JGI genomes were categorized by scanning all metadata fields. If no metadata was present, the genome was categorized as Other, so it is possible that marine genomes were included in the "non-marine" JGI-ABC calculations of True Diversity. Strains with species name "marina" (i.e. *Micromonospora marina*) with no metadata were looked up and determined as marine. These designated marine genomes were excluded when calculating Total Diversity in SI Table 6.

Figure S6. Phylogenetic Similarity vs. Shared GCFs

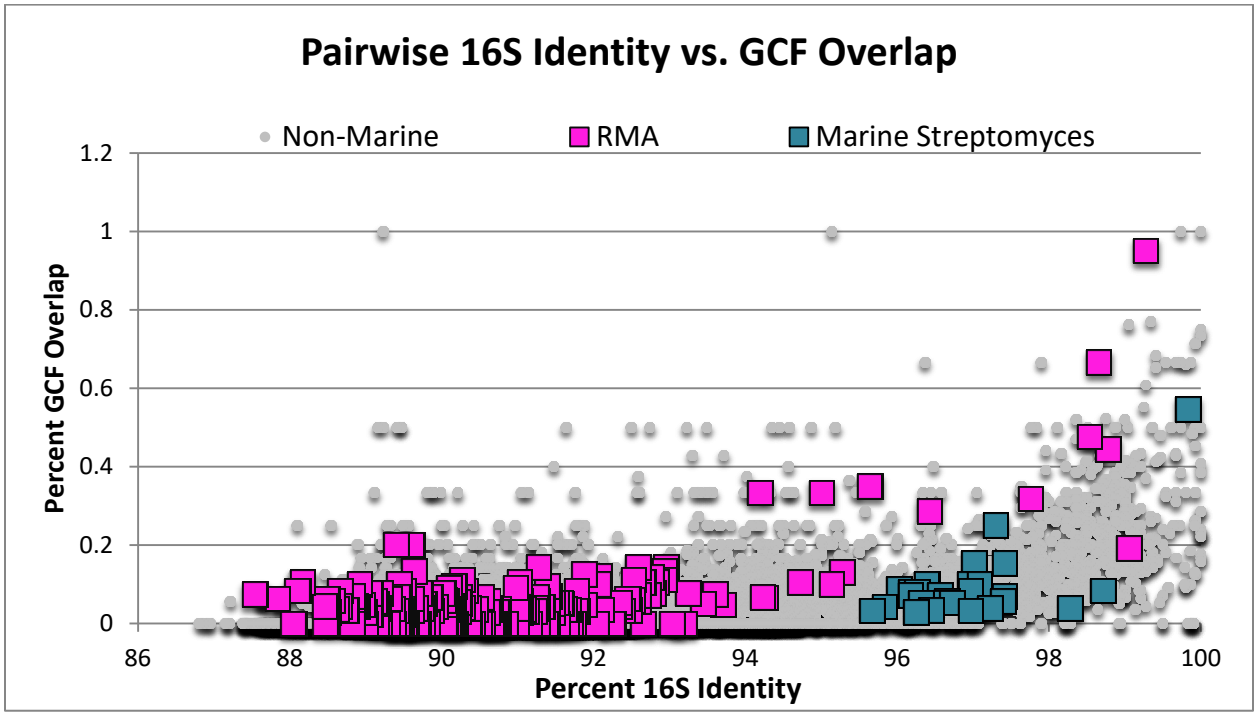


Figure S6. Each point represents a pairwise distance of 16S rRNA percent identity vs the GCF overlap between two genomes. Each pair is part of a larger group: non-marine JGI genomes from the genera examined in this study (grey), RMA genomes from this study (pink), and marine streptomycetes (teal).

Supplementary References

Gontang, E., Gaudencio, S., Fenical, W. & Jensen, P. (2010). Sequence-based analysis of secondary metabolite biosynthesis in marine Actinobacteria. *Appl Environ Microbiol* **76**, 2487-2499.

Jost, L. & Baños, T., Ecuador (loujost@yahoo.com). (2016). Entropy and diversity. *Oikos* **113**, 363-375.

Land, M. L., Hyatt, D., Jun, S. R., Kora, G. H., Hauser, L. J., Lukjancenko, O. & Ussery, D. W. (2014). Quality scores for 32,000 genomes. *Stand Genomic Sci* **9**, 20.

Mincer, T. J., Jensen, P. R., Kauffman, C. A. & Fenical, W. (2002). Widespread and persistent populations of a major new marine actinomycete taxon in ocean sediments. *Appl Environ Microbiol* **68**, 5005-5011.

Patin, N. V., Duncan, K. R., Dorrestein, P. C. & Jensen, P. R. (2016). Competitive strategies differentiate closely related species of marine actinobacteria. *Isme j* **10**, 478-490.

Trzoss, L., Fukuda, T., Costa-Lotufo, L. V., Jimenez, P., La Clair, J. J. & Fenical, W. (2014). Seriniquinone, a selective anticancer agent, induces cell death by autophagocytosis, targeting the cancer-protective protein dermcidin. *Proc Natl Acad Sci U S A* **111**, 14687-14692.

Yamanaka, K., Reynolds, K. A., Kersten, R. D., Ryan, K. S., Gonzalez, D. J., Nizet, V., Dorrestein, P. C. & Moore, B. S. (2014). Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A* **111**, 1957-1962.