Title: Genomic islands predict functional adaptation in marine actinobacteria

Author(s): Kevin Penn[1], Caroline Jenkins[1], Markus Nett[1], Daniel W Udwary[1], Erin A Gontang[1], Ryan P McGlinchey[1], Brian Foster[2], Alla Lapidus[2], Sheila Podell[1], Eric E Allen[1], Bradley S Moore[1,3] and Paul R Jensen[1]

Author Affiliations: [1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA, [2]Department of Energy, Joint Genome Institute-Lawrence Berkeley National Laboratory, Walnut Creek, CA, USA, [3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA

Date: 06/10/09

Funding: This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

Genomic islands predict functional adaptation in marine Actinobacteria.

Kevin Penn[1], Caroline Jenkins[1], Markus Nett[1], Daniel W Udwary[1], Erin A Gontang[1], Ryan P McGlinchey[1], Brian Foster[2], Alla Lapidus[2], Sheila Podell[1], Eric E Allen[1], Bradley S Moore[1,3] and Paul R Jensen[1]

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

[2]Department of Energy, Joint Genome Institute-Lawrence Berkeley National Laboratory, Walnut Creek, CA, USA

[3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA

Linking functional traits to bacterial phylogeny remains a fundamental but elusive goal of microbial ecology [1]. Without this information, it becomes impossible to resolve meaningful units of diversity and the mechanisms by which bacteria interact with each other and adapt to environmental change. Ecological adaptations among bacterial populations have been linked to genomic islands, strain-specific regions of DNA that house functionally adaptive traits [2]. In the case of environmental bacteria, these traits are largely inferred from bioinformatic or gene expression analyses [2], thus leaving few examples in which the functions of island genes have been experimentally characterized. Here we report the complete genome sequences of *Salinispora tropica* and *S. arenicola*, the first cultured, obligate marine Actinobacteria [3]. These two species inhabit benthic marine environments and dedicate 8-10% of their genomes to the biosynthesis of secondary metabolites. Despite a close phylogenetic relationship, 25 of 37 secondary metabolic pathways are species-specific and located within 21 genomic islands, thus providing new evidence linking secondary metabolism to ecological adaptation. Species-specific differences are also observed in CRISPR sequences, suggesting that variations in phage immunity provide fitness advantages that contribute to the cosmopolitan distribution of *S. arenicola* [4]. The two *Salinispora* genomes have evolved by complex processes that include the duplication and acquisition of secondary metabolite genes, the products of which provide immediate opportunities for molecular diversification and ecological adaptation. Evidence that secondary metabolic pathways are exchanged by Horizontal Gene Transfer (HGT) yet are fixed among globally distributed populations [5] supports a functional role for their products and suggests that pathway acquisition represents a previously unrecognized force driving bacterial diversification.

Most bacterial diversity is delineated among clusters of sequences that share >99% 16S rRNA gene sequence identity [6]. These sequence clusters are believed to represent fundamental units of diversity (ie., species), while intra-cluster microdiversity is thought to persist due to weak selective pressures [6] suggesting little ecological or taxonomic relevance. The genus *Salinispora* is comprised of three species that collectively constitute a microdiverse sequence cluster [4]. Although taxonomic significance has been assigned to the microdiversity within this

cluster, it remains to be determined if the three species represent ecologically or functionally distinct lineages. A previous analysis of 40 *Salinispora* strains revealed that secondary metabolite production is the major phenotypic difference among the three species, an observation supported by the analysis of the *S. tropica* secondary metabolome [7]. Here we present a comparative analysis of the complete genome sequences of *S. tropica* (ST, strain CNB-440) and *S. arenicola* (SA, strain CNS-205) with the aim of defining the functional attributes that differentiate the two species.

The ST and SA genomes share 3606 orthologs, representing 79.4% and 73.2% of the respective genomes (Table 1). The average nucleotide identity among these orthologs is 87.2%, well below the 94% cut-off that has been suggested to delineate bacterial species [8]. Despite differing by only seven nucleotides (99.7% identity) in the 16S rRNA gene, SA is 603 kb (11.6%) larger and possesses 1505 species-specific genes relative to 987 in ST. Seventy-five percent of these species-specific genes are concentrated in 21 genomic islands (Tables 1, S1). These islands are enriched with large clusters of genes devoted to the biosynthesis of secondary metabolites (Figure 1). They house all 25 of the species-specific secondary metabolic pathways, while eight of 12 shared pathways occur in the genus-specific core (Table S2). We have isolated and identified the products of eight of these pathways, which include the highly selective proteasome inhibitor salinosporamide A, currently in clinical trials for the treatment of cancer, as well as sporolide A, which is derived from an enediyne polyketide precursor, one of the most potent classes of biologically active agents discovered to date [9].

Of the eight secondary metabolites that have been isolated from the two strains, all but salinosporamide A, sporolide A, and salinilactam have been reported from unrelated taxa (Figure 1), providing strong evidence of HGT. Further evidence for HGT comes from a phylogenetic analysis of the polyketide synthase (PKS) genes associated with the rifamycin biosynthetic gene cluster (*rif*) in SA and *Amycolatopsis mediterranei*, the original source of this compound [10]. This analysis reveals that all 10 of the ketosynthase domains are perfectly interleaved, as would be predicted if the entire PKS gene cluster had been exchanged between the two strains (Figure S1). Evidence of HGT coupled with the fixation of specific pathways such as *rif* among globally distributed SA populations [5] supports vertical inheritance following pathway acquisition [11] and is indicative of a selective sweep or ecotype diversification [12], either of which provide compelling evidence that secondary metabolites represent functional traits with important ecological roles. The hypothesis that gene acquisition provides a mechanism for ecological diversification that may ultimately drive the formation of independent bacterial lineages (*sensu* Ochman et al., 2005) sheds new light on the functional importance and evolutionary significance of secondary metabolism.

Ecological differences between the two species also appear linked to CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) sequences. CRISPRs are non-continuous direct repeats separated by variable (spacer) sequences that have been shown to confer immunity to phage [13]. The ST genome carries three intact prophage and three CRISPRs (35 spacers), while only one prophage has been identified in the genome of SA, which possesses eight different CRISPRs (140 spacers). The SA prophage is unprecedented among bacterial genomes in that it occurs in two adjacent copies that share 100% sequence identity. These copies are flanked by tRNA *att* sites and separated by an identical 45 bp *att* site, suggesting double integration as

opposed to duplication [14]. Remarkably, four of the SA CRISPRs possess a spacer that shares 100% identity with portions of three different genes found in ST prophage 1 (Figure 2). This observation provides evidence that SA was previously exposed to a phage related to one that currently infects ST and that SA now maintains acquired immunity to this phage genotype. This is a rare example of CRISPR-mediated acquired immunity to a prophage that resides in the genome of a closely related environmental bacterium. Given that SA strain CNS-205 was isolated from Palau while ST strain CNB-440 was recovered 15 years earlier from the Bahamas, it appears that actinophage have broad temporal-spatial distributions or that resistance is maintained on temporal scales sufficient for the global distribution of a bacterial species. Enhanced phage immunity coupled with its larger genome size and greater number of species-specific, secondary metabolic pathways may account for the cosmopolitan distribution of SA relative to ST, which to date has only been recovered from the Caribbean [4].

The 21 genomic islands are not contiguous regions of species-specific DNA but were instead created by a complex process of gene acquisition, loss, duplication, and inactivation (Figure 3). Interestingly, the overall composition, evolutionary history, and function of the island genes are similar in both strains, with duplication and HGT accounting for the majority of genes and secondary metabolism representing the largest functionally annotated category. Remarkably, 42% of the rearranged island orthologs fall within other islands indicating that inter-island movement or "island hopping" is common, thus providing support for the hypothesis that islands undergo continual rearrangement [2]. There is dramatic, operon-scale evidence of this process in the shared yersiniabactin (ST *sid2* and SA *sid1*) and unknown dipeptide (ST *nrps1* and SA *nrps3*) pathways, both of which occur in different islands in the two strains (Figure 1). There is also evidence of cluster fragmentation in the 10-membered enediyne gene set SA *pks3*, which contains the core set of genes associated with calicheamicin biosynthesis (Figure S3), [15], yet is split by the introduction of 145 kb of DNA from three different biosynthetic loci (island 10, Figure 1). The conserved fragments appear to encode the biosynthesis of a calicheamicin anolog, while flanking genes display a high level of gene duplication and rearrangement indicative of active pathway evolution. Cluster fragmentation is also observed in the 9-membered enediyne PKS cluster SA *pks1*, which is scattered across the genome in islands 4, 10, and 21.

The genomic islands are also enriched in mobile genetic elements including prophage, integrases, and Actinobacterial Integrative and Conjugative Elements (AICEs) [16] (Table S3), which are known to play a role in gene acquisition and rearrangement. The *Salinipora* AICEs possess *traB* homologs, which promote conjugal plasmid transfer in mycelial streptomycetes [17], suggesting that hyphal tip fusion is a prominent mechanism driving gene exchange in these bacteria. AICEs have been linked to the acquisition of secondary metabolite gene clusters [18] and their occurrence in island 7 (SA AICE1), which includes the entire 90 kb *rif* cluster, and island 10 (SA AICE3), which contains biosynthetic gene clusters for enediyne, siderophore, and amino acid-derived secondary metabolites, provides a mechanism for the acquisition of these pathways (Figure 1). Six additional secondary metabolite gene clusters (ST *nrps1*, ST *spo*, SA *nrps3*, SA *pks5*, SA *cym*, and SA *pks2*) are flanked by direct repeats, providing further support for HGT. In the case of *cym* [19], which is clearly inserted into a tRNA, the pseudogenes preceding and following it are all related to transposases or integrases providing a mechanism for chromosomal integration. Despite exhaustive analyses of HGT, only 22% of the 127 genes in the five

biosynthetic pathways whose products are shared with other bacteria scored positive in any of the tests applied. This observation suggests that much of the HGT in the two genomes has occurred among closely related bacteria and that this process likely accounts for many of the island genes for which no evidence of evolutionary origin could be detected (Figure 3). In support of an adaptive role for island genes, 7.6% (44/573) of the orthologs show evidence of positive selection (dN/dS >1) compared to 1.6% (49/3027) of the non-island pairs.

Functional differences between related organisms can be obscured when orthologs are taken out of the context of the gene clusters in which they reside. For example, the PKS genes Sare1250 and Stro2768 are orthologous, yet they reside in the *rif* and *slm* pathways, respectively, and thus contribute to the biosynthesis of dramatically different molecules. Likewise, intra-cluster PKS gene duplication (Sare3151 and Sare3152, Figure 1) has an immediate effect on the product of the pathway as opposed to the more traditional concept of parology facilitating mutation-driven functional divergence [20]. Sub-genic, modular duplications are also observed (Sare3156 modules 4 and 5, Figure 1), which likewise have an immediate effect on the small molecule product of the pathway. While HGT is considered a rapid method for ecological adaptation in bacteria [21], PKS gene duplication provides an effective evolutionary strategy that could lead to the rapid creation of new adaptive radiations in a manner akin to punctuated equilibrium, yet in the spatial and temporal context of bacteria.

Salinispora are the first marine Actinobacteria reported to require seawater for growth [22]. Unlike Gram-negative marine bacteria, in which seawater requirements are linked to a specific sodium ion requirement [23], *Salinispora* strains are capable of growth in osmotically adjusted, sodium-free media [24]. An analysis of the *Salinispora* core for evidence of genes associated with this unusual osmotic requirement reveals a highly duplicated family of 29 Polymorphic Membrane Proteins (PMPS) that include homologs associated with outer membrane proteins (POMPS). POMPS remain functionally uncharacterized however there is strong evidence that they are type V secretory systems [25], making this the first report of type V autotransporters outside of the Proteobacteria [26]. Phylogenetic analyses provide evidence that the *Salinispora* PMPs were acquired from aquatic, Gram-negative bacteria and that they have continued to undergo considerable duplication subsequent to divergence of the two species (Figure S2). The surprising occurrence of this large family of PMP autotransporters in marine Actinobacteria may represent a low nutrient adaptation that renders cells susceptible to lysis in low osmotic environments.

In conclusion, our comparative analysis of two closely related, marine Actinobacterial genomes provides new insight into the functional traits associated with genomic islands and evidence that secondary metabolism is a previously unrecognized force driving ecological diversification among closely related, sediment inhabiting bacteria. It has been possible to assign precise, physiological functions to island genes and link differences in secondary metabolism to fine-scale phylogenetic architecture in two distinct bacterial lineages, which by all available metrics maintain the fundamental characteristics of species-level units of diversity. It is clear that gene clusters devoted to secondary metabolite biosynthesis are dynamic entities that are readily acquired, rearranged, and fragmented in the context of genomic islands. The results of these processes create small molecule diversity that can have an immediate affect on fitness or niche utilization.

Methods Summary

Sequencing and annotation were completed by the Department of Energy, Joint Genome Institute [7]. Orthologs were predicted using the Reciprocal Smallest Distance method [27]. Genomic islands were defined as regions >20 kb where <40% of the genes lack a positional ortholog. Paralogs were identified using blastclust [28] with a cut-off of 30% identity over 40% of the sequence. APIS was used to identify recent gene duplications [29]. HGT was assessed using a variety of methods with genes scoring positive in $\geq$2 tests counted as positive. Methods included G+C content, Codon Adaptive Index [30], dinucleotide frequency differences [31], and DNA composition [32]. Lineage Probability Index (LPI) scores assigned using Darkhorse [33] were also used with values <0.5 scored as positive. An LPI score >0.5 in a reciprocal Darkhorse analysis was assigned an additional positive score. All genes clading with non-Actinobacterial homologs using APIS [29] were scored as positive. RSD analyses of 27 finished Actinobacterial genomes were also used with genes unique to SA or ST scored as positive. Genes identified as bacteriophage using Prophage [34] and Phage Finder [35] or other MGEs using blastX homology, PFAM, SPTR, KEGG, and COG databases were scored positive. The scores were amalgamated, mapped onto the genome, and genes scoring positive in only one test but associated with clusters of genes that scored in two or more tests were added to the total HGT pool.

CRISPRs were identified using CRISPR finder (http://crispr.u-psud.fr/Server/CRISPRfinder.php) while repeats larger than 35 bases were identified using Reputer [36]. Secondary metabolite gene clusters were manually annotated [7] and boundaries predicted using previously reported clusters or, for unknown clusters, loss of gene conservation across the Actinobacteria. The ratio of non-synonymous to synonymous mutations (dN/dS) was calculated using SNAP (http://www.hiv.lanl.gov).

References

1 Dana E. Hunt, Lawrence A. David, Dirk Gevers et al., *Science* **320** (5879), 1081 (2008).

2 Maureen L. Coleman, Matthew B. Sullivan, Adam C. Martiny et al., *Science* **311** (5768), 1768 (2006).

3 Tracy J. Mincer, Jensen, Paul R., Kauffman, Christopher A., Fenical, William, *Appl. Environ. Microbiol.* **68** (10), 5005 (2002).

4 Paul R. Jensen and Chrisy Mafnas, *Environmental Microbiology* **8** (11), 1881 (2006).

5 P. R.; Williams Jensen, P. G.; Oh, D.-C.; Zeigler, L.; Fenical, W., *Appl. Environ. Microbiol.* **73** (4), 1146 (2007).

6 Silvia G. Acinas, Vanja Klepac-Ceraj, Dana E. Hunt et al., *Nature* **430** (6999), 551 (2004).

7 Daniel W. Udwary, Lisa Zeigler, Ratnakar N. Asolkar et al., *Proceedings of the National Academy of Sciences* **104** (25), 10376 (2007).

8 Konstantinos T. Konstantinidis and James M. Tiedje, *PNAS* **102** (7), 2567 (2005).

9 W. Fenical, Jensen, P.R., *Nature Chemical Biology* **2**, 666 (2006).

10 T.-W. Yu, Shen, Y., Doi-Katayama, Y., Tang, L., Park, C., Moore, B.S., Hutchinson, C.R., Floss, H.G., *Proceedings of the National Academy of Sciences* **96** (16), 9051 (1999).

[11] H. Ochman, Lerat, E., Daublin, V., *Proceedings of the National Academy of Sciences* **102**, 6595 (2005).
[12] F.M. Cohan, *Annual Review of Microbiology* **56**, 457 (2002).
[13] Rodolphe Barrangou, Christophe Fremaux, Helene Deveau et al., *Science* **315** (5819), 1709 (2007).
[14] Evelien M. te Poele, Markiyan Samborskyy, Markiyan Oliynyk et al., *Plasmid* **59** (3), 202 (2008).
[15] Joachim Ahlert, Erica Shepard, Natalia Lomovskaya et al., *Science* **297** (5584), 1173 (2002).
[16] Guillaume Pavlovic Bernard Decaris GÈrard GuÈdon Vincent Burrus, *Molecular Microbiology* **46** (3), 601 (2002).
[17] J. Reuther, Gekeler, C., Tiffert, Y., Wohlleben, W., Muth, G., *Molecular Microbiology* **61** (2), 436 (2006).
[18] Marrit N. Habets Geok Yuan Annie Tan Alan C. Ward Michael Goodfellow Henk Bolhuis Lubbert Dijkhuizen Evelien M. te Poele, *FEMS Microbiology Ecology* **61** (2), 285 (2007).
[19] A.W. Schultz, Oh, D.-C., Carney, J.R., Williamson, R.T., Udwary, D.W., Jensen, P.R., Gould, S.J., Fenical, W., Moore, B.S., *Journal of the American Chemical Society* **130** (13), 4507 (2008).
[20] V.E. Prince, Pickett, F.B., *Nat Rev Gen* **3**, 827 (2002).
[21] H. Ochman, Lawrence, J.G., Groisman, E.A., *Nature* **405**, 299 (2000).
[22] Luis A. Maldonado, William Fenical, Paul R. Jensen et al., *Int J Syst Evol Microbiol* **55** (5), 1759 (2005).
[23] K. Kogure, *Current Opinion in Biotechnology* **9**, 278 (1998).
[24] Ginger Tsueng and Kin Lam, *Applied Microbiology and Biotechnology* **78** (5), 821 (2008).
[25] I.R. Henderson, Lam, A.C., *Trends in Microbiology* **9**, 573 (2001).
[26] I. R. Henderson, Navarro-Garcia, F., Desvaux, M., Fernandez, R.C., Ala'Aldeen, D., *Microbiol. Mol. Biol. Rev.* **68** (4), 692 (2004).
[27] D. P. Wall, H. B. Fraser, and A. E. Hirsh, *Bioinformatics* **19** (13), 1710 (2003).
[28] I Dondoshansky and Y Wolf, BLASTCLUST (National Institutes of Health, Bethesda, MD, 2000).
[29] Jonathan H. Badger, Jonathan A. Eisen, and Naomi L. Ward, *Int J Syst Evol Microbiol* **55** (3), 1021 (2005).
[30] Gang Wu, David E. Culley, and Weiwen Zhang, *Microbiology* **151** (7), 2175 (2005).
[31] William Hsiao, Ivan Wan, Steven J. Jones et al., *Bioinformatics* **19** (3), 418 (2003).
[32] G. S. Vernikos and J. Parkhill, *Bioinformatics* **22** (18), 2196 (2006).
[33] S. Podell and T. Gaasterland, *Genome Biology* **8** (2) (2007).
[34] M. Bose, Barber, R.D., *In Silico Biology* **6**, 223 (2006).
[35] D. E. Fouts, *Nucleic Acids Research* **34** (20), 5839 (2006).
[36] Stefan Kurtz, Jomuna V. Choudhuri, Enno Ohlebusch et al., *Nucl. Acids Res.* **29** (22), 4633 (2001).

Supplementary Information is available online.

Acknowledgements This manuscript is dedicated to Professor William Fenical for his pioneering work on the secondary metabolites of marine actinomycetes. PRJ and BMS are funded by California Seagrant and NOAA. EEA thanks the Gordon and Betty Moore Foundation for funding through CAMERA.

Author Contributions KP conceived and performed the overall comparative and bioinformatics analyses, CJ assessed HGT and analyzed CRISPR and phage sequences, MN, DWU, RPM and BSM annotated and analyzed secondary metabolite clusters, EAG assessed secondary metabolite gene evolution, BF and AL sequenced and annotated the genomes, SP performed Darkhorse analyses and assisted with HGT assessment, EEA provided guidance and computational assistance with the overall bioinformatic analyses, PRJ assisted with data analysis and wrote the paper.

Author information Genome sequences have been deposited in GenBank under accession numbers CP000667 (ST) and CP000850 (SA).

Table 1. General genome features.

Feature	*S. tropica* (ST)	ST%	*S. arenicola* (SA)	SA %
No. base pairs	5183331	NA	5786361	NA
% G+C	69.4	NA	69.5	NA
Total genes	4536	NA	4919	NA
Pseudogenes	57	1.26%	192	3.90%
Hypotheticals (% genome)	1140	25.10%	1418	28.80%
No. rRNA operons (% identity)	3	100%	3	100%
Orthologs (% genome)	3606	79.40%	3606	73.20%
Positional orthologs (% genome)	3178	70.10%	3178	64.60%
Rearranged orthologs (% genome)	428	9.40%	428	8.70%
Species-specific genes (% genome)	987	21.80%	1505	30.60%
Island genes (% genome)	1350	29.80%	1690	34.30%
Total genes with evidence of HGT (% genome)	652	14.30%	750	14.70%
Species-specific genes with evidence of HGT (% species-specific)	405	41.00%	573	38.10%
Total island genes with evidence of HGT (% HGT)	473	72.50%	555	74.00%
Paralogs* (% genome)	1819	39.60%	2179	42.60%
Species-specific paralogs (% species-specific genes)	391	39.70%	647	43.00%
Secondary metabolism (% genome)	405	8.80%	556	10.90%

*Totals include parental gene.
NA: not applicable.

Figure legend

Figure 1. Linear alignment of the *S. tropica* and *S. arenicola* genomes. **a**, Positional orthologs (core) flanked by islands (E, F), heat-mapped HGT genes (D, G), rearranged orthologs (C, H), species-specific genes (B, I), secondary metabolite genes (green), MGEs (pink) with prophage (P) and AICES (E) indicated (A, J). For genomic islands, predicted (lower case) and isolated secondary metabolites (uppercase with structures) are given (not shown are six non-island secondary metabolic gene clusters of unknown function). Shared positional (blue) and rearranged (red) secondary metabolite clusters are indicated. *Previously isolated from other bacteria. **b**, Expanded view of SA *pks5* revealing gene and modular architecture. **c**, Neighbor-joining phylogenetic tree of KS domains (erythromycin root, % bootstrap values from 1000 re-samplings).

Figure 2. *S. tropica* prophage and *S. arenicola* CRISPRs. Four of 8 SA CRISPRs have spacers (color coded) that share 100% sequence identity with genes (Stro numbers given) in ST prophage 1 (inverted for visual purposes). SA CRISPRs 2-3 and 5-6 share the same direct repeats and may have at one time been a single allele. CRISPR Associated (CAS) genes (red), genes interrupting CRISPRs (black), and CRISPRs with no match to prophage in the NCBI and CAMERA databases (purple) are indicated.

Figure 3. Genomic islands. **a**, contribution of *S. tropica* (ST) and *S. arenicola* (SA) to island formation (gene totals presented in wedges). **b**, Evolutionary history and **c**, functional annotation of species-specific island genes. **d**, Distribution of species specific island genes that have no evidence for HGT or duplication among 27 actinobacterial genomes.

Figure S1. Polyketide synthase phylogeny. Neighbor-joining distance tree constructed in PAUP (Swafford) using the aligned amino acid sequences of the *rif* KS domains from *A. mediterranei* and *S. arenicola*. Bootstrap values (in percent) calculated from 1000 re-samplings are shown at their respective nodes for values greater than or equal to 60%. The KS domain from module 4 of the erythromycin biosynthetic pathway of *Saccharopolyspora erythraea* was used to position the root.

Figure S2. Polymorphic Membrane Protein (PMP) phylogeny. Neighbor-joining distance tree constructed in APIS (J. Badger, unpublished) using the aligned amino acid sequences of SA and ST PMPs as well as those observed in other genomes. Bold lines indicate boot-strap values >50% and blue indicates strains other than SA and ST that were derived from aquatic environments. Accession numbers in parentheses.

Figure S3. Cluster SA *pks3A* and *pks3B* in comparison with the *cal* locus from *M. echinospora*. **a** Grey boxes indicate regions of gene conservation. Duplicated genes are circled in red with parologs identified by letter. Red arrows indicate pseudogenes (which are also checkered). Genes missing (green arrows) and unique (colored white) relative to the *cal* locus are indicated. **b** structure of calicheamicin.

Methods

Sequencing and ortholog identification

Sequencing and annotation were completed by the Department of Energy, Joint Genome Institute as part of the Community Sequencing Program using previously described methods [7]. Orthologs within the two genomes were predicted using the Reciprocal Smallest Distance method [27], which includes a maximum likelihood estimate of amino acid substitutions. A linear alignment of positional orthologs was created and the positions of rearranged orthologs and species-specific genes identified. Genomic islands were defined as regions >20 kb where <40 % of the genes lack a positional ortholog in the reciprocal genome and are flanked by regions of conservation. Paralogs within each genome were identified using the blastclust algorithm [28] with a cut-off of 30% identity over 40% of the sequence length. APIS was used to identify recent gene duplications [29].

Horizontal Gene Transfer

All genes were assessed for evidence of HGT using a variety of tests, with each positive result being assigned a specific score. Genes identified in $\geq$2 different tests were counted as positive for HGT and color coded from yellow to red corresponding to total scores from 2 to 6 (Figure 1a). Four DNA compositional analyses were combined to identify abnormalities relative to the genomic mean. These tests included G+C content (obtained from the JGI annotation), Codon Adaptive Index, calculated with the CAI calculator [30] using a suite of housekeeping genes as reference, dinucleotide frequency differences (δ^*), calculated using IslandPath [31], and DNA composition, calculated using Alien_Hunter [32]. G+C content or codon usage values >1.5 standard deviations from the genomic mean and dinucleotide frequency differences >1 standard deviation from the mean were scored positive for HGT.

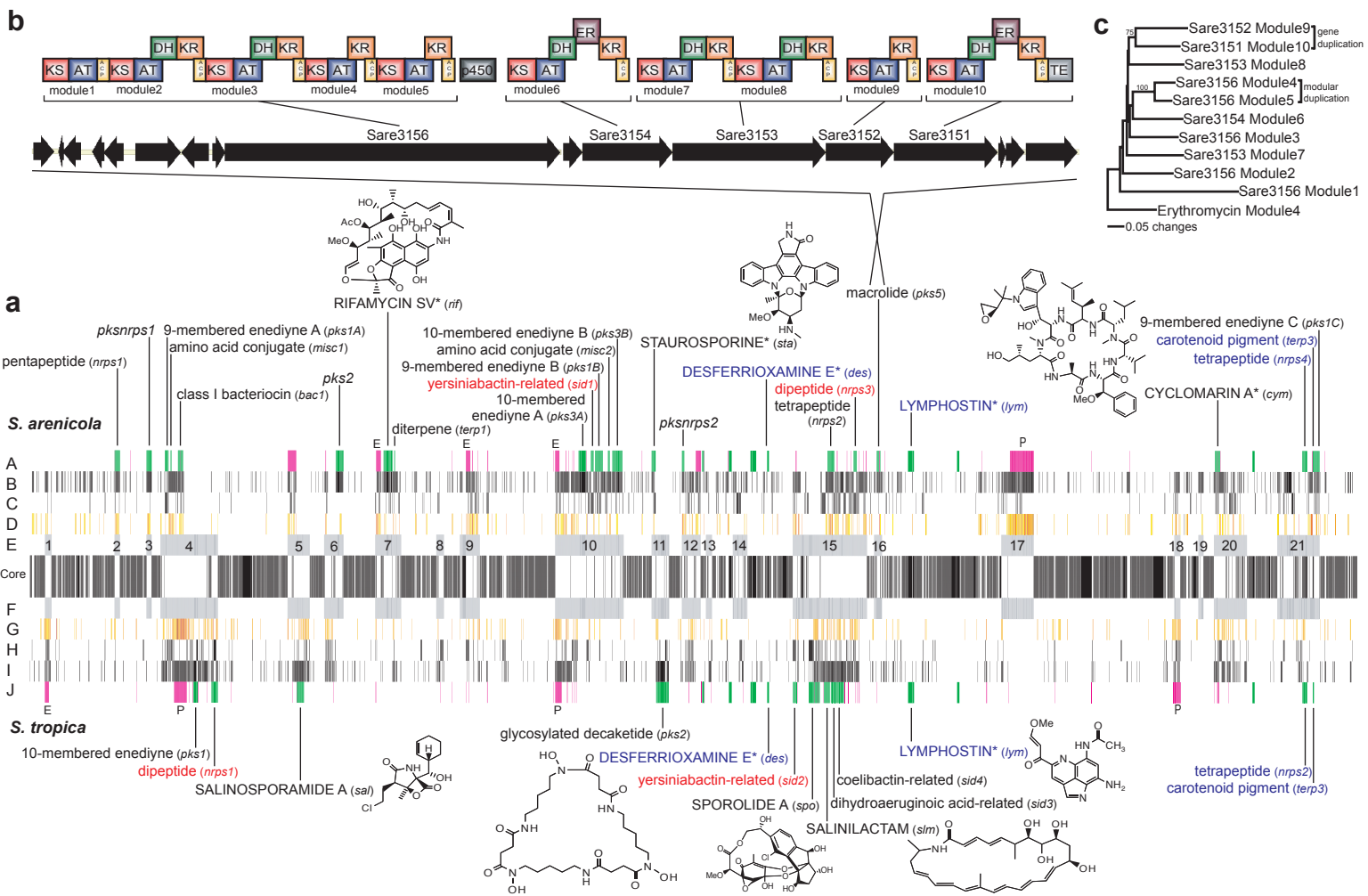
Taxonomic relationships in the form of Lineage Probability Index (LPI) scores for all protein coding genes were assigned using the Darkhorse algorithm [33]. Genes with an LPI score of <0.5 (indicating orthologs are not in closely related genomes) were scored positive for HGT. A reciprocal Darkhorse analysis was then performed on the orthologs of all positives and if these genes had an LPI score >0.5 (indicating both genes were in closely related organisms), they were assigned an additional positive score.

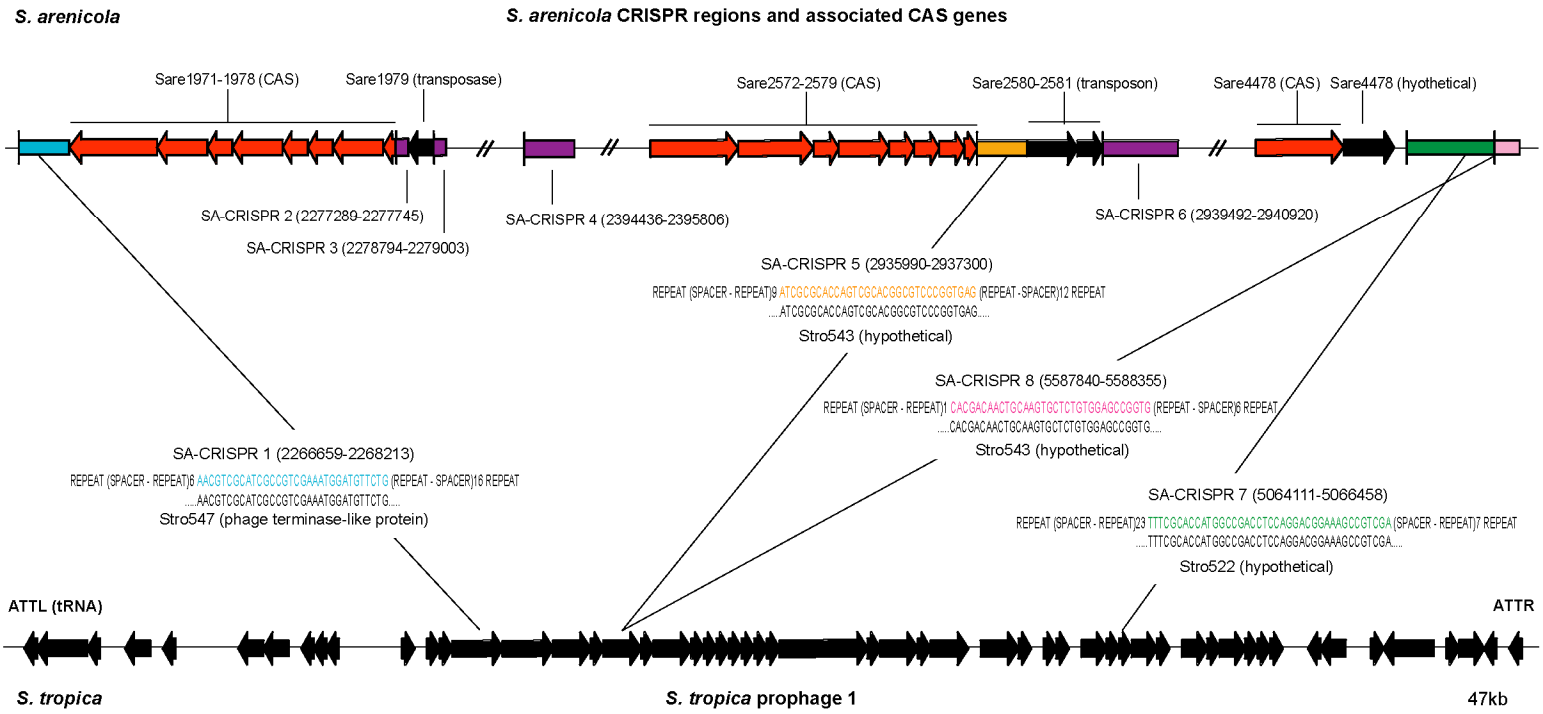
A phylogenetic approach using the APIS program [29] was also employed to assess HGT. Using this program, bootstrapped neighbor-joining trees of all predicted protein coding genes within each genome were created. All genes clading with non-Actinobacterial homologs were binned into their respective taxonomic groups and given a positive HGT score. Evidence of HGT was also inferred from RSD analyses of each genome against a compiled set of 27 finished Actinobacterial genomes that included at least two representatives of each genus for which sequences were available. Genes unique to SA or ST and not observed among the 27 Actinobacterial genomes were assigned a positive HGT score.

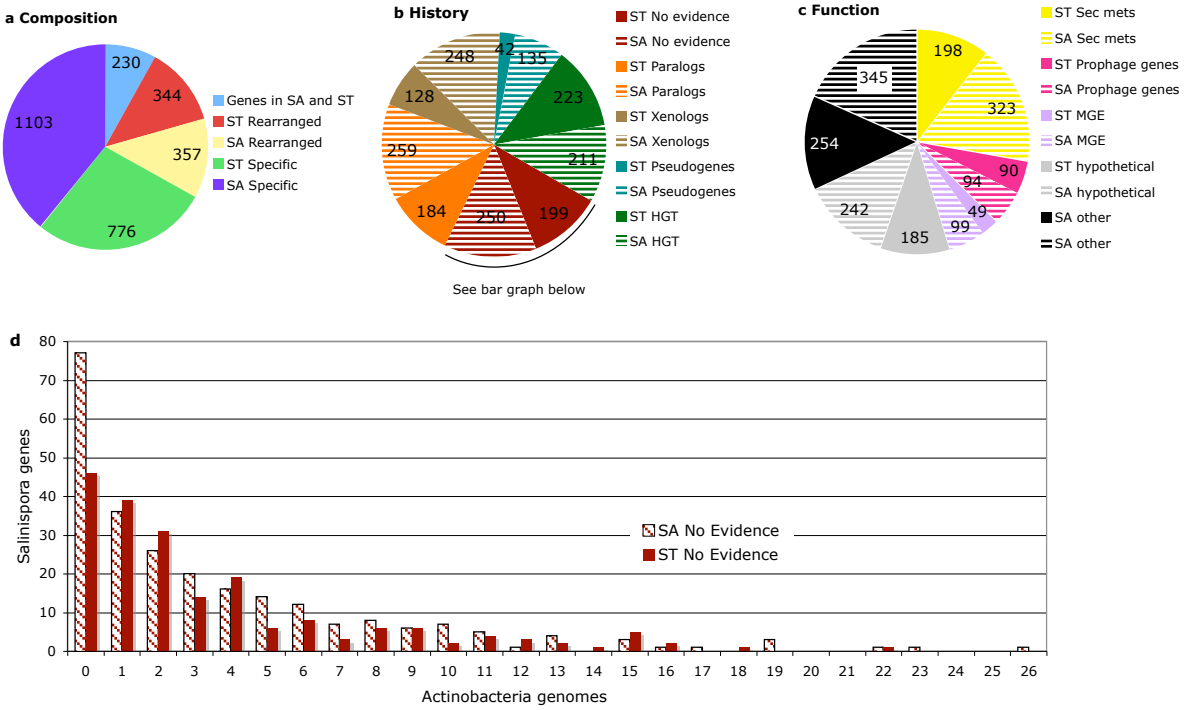
Bacteriophage were identified using Prophage [34] and Phage Finder [35]. Other insertion elements were identified as prophage or transposon in origin through blastX homology searches. Gene annotation based on searches for identity across PFAM, SPTR, KEGG and COG databases was also used to help identify mobile genetic elements. Each gene associated with a mobile

genetic element was assigned an HGT score of 1. Test scores were amalgamated and those genes showing evidence of HGT in two or more tests (maximum score 6) were classified as horizontally acquired. The results were mapped onto the genome and genes identified by only one test but associated with clusters of genes that scored in two or more tests were added to the total HGT pool. Adjacent clusters were merged.

CRISPRs were identified using CRISPR finder (http://crispr.u-psud.fr/Server/CRISPRfinder.php) while repeats larger than 35 bases were identified using Reputer [36]. Secondary metabolite gene clusters were manually annotated as in [7]. Cluster boundaries were predicted using previously reported gene clusters as in the case of rifamycin. For unknown clusters, loss of gene conservation across the Actinobacteria was used to aid boundary predictions. The ratio of non-synonymous to synonymous mutations (dN/dS) for all orthologs was calculated using the perl progam SNAP (http://www.hiv.lanl.gov) with the alignments for all values >1 checked manually.


b
KS AT KS AT DH KR KS AT DH KR KS AT KR KS AT KR p450
module1 module2 module3 module4 module5
KS AT DH ER KR KS AT DH KR KS AT DH KR KR KS AT DH ER KR KS AT TE
module6 module7 module8 module9 module10
Sare3156 Sare3154 Sare3153 Sare3152 Sare3151
c
Sare3152 Module9
Sare3151 Module10
Sare3153 Module8
Sare3156 Module4
Sare3156 Module5
Sare3154 Module6
Sare3156 Module3
Sare3153 Module7
Sare3156 Module2
Sare3156 Module1
Erythromycin Module4
75
100
gene duplication
modular duplication
0.05 changes
a
RIFAMYCIN SV* (rif)
STAUROSPORINE* (sta)
macrolide (pks5)
CYCLOMARIN A* (cym)
pksnrps1
9-membered enediyne A (pks1A)
amino acid conjugate (misc1)
pentapeptide (nrps1)
class I bacteriocin (bac1)
pks2
10-membered enediyne B (pks3B)
amino acid conjugate (misc2)
9-membered enediyne B (pks1B)
yersiniabactin-related (sid1)
10-membered enediyne A (pks3A)
diterpene (terp1)
pksnrps2
DESFERRIOXAMINE E* (des)
dipeptide (nrps3)
tetrapeptide (nrps2)
LYMPHOSTIN* (lym)
9-membered enediyne C (pks1C)
carotenoid pigment (terp3)
tetrapeptide (nrps4)
S. arenicola
A
B
C
D
E
Core
F
G
H
I
J
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
E P E E E P P
S. tropica
10-membered enediyne (pks1)
dipeptide (nrps1)
SALINOSPORAMIDE A (sal)
glycosylated decaketide (pks2)
DESFERRIOXAMINE E* (des)
yersiniabactin-related (sid2)
SPOROLIDE A (spo)
coelibactin-related (sid4)
dihydroaeruginoic acid-related (sid3)
SALINILACTAM (slm)
LYMPHOSTIN* (lym)
tetrapeptide (nrps2)
carotenoid pigment (terp3)


S. arenicola
S. arenicola CRISPR regions and associated CAS genes
Sare1971-1978 (CAS)
Sare1979 (transposase)
Sare2572-2579 (CAS)
Sare2580-2581 (transposon)
Sare4478 (CAS)
Sare4478 (hyothetical)
SA-CRISPR 2 (2277289-2277745)
SA-CRISPR 3 (2278794-2279003)
SA-CRISPR 4 (2394436-2395806)
SA-CRISPR 6 (2939492-2940920)
SA-CRISPR 5 (2935990-2937300)
REPEAT (SPACER - REPEAT)9 ATCGCGCACCAGTCGCACGGCGTCCCGGTGAG (REPEAT - SPACER)12 REPEAT
....ATCGCGCACCAGTCGCACGGCGTCCCGGTGAG....
Stro543 (hypothetical)
SA-CRISPR 8 (5587840-5588355)
REPEAT (SPACER - REPEAT)1 CACGACAACTGCAAGTGCTCTGTGGAGCCGGTG (REPEAT - SPACER)8 REPEAT
....CACGACAACTGCAAGTGCTCTGTGGAGCCGGTG....
Stro543 (hypothetical)
SA-CRISPR 1 (2266659-2268213)
REPEAT (SPACER - REPEAT)8 AACGTCGCATCGCCGTCGAAATGGATGTTCTG (REPEAT - SPACER)16 REPEAT
....AACGTCGCATCGCCGTCGAAATGGATGTTCTG....
Stro547 (phage terminase-like protein)
SA-CRISPR 7 (5064111-5066458)
REPEAT (SPACER - REPEAT)23 TTTCGCACCATGGCCGACCTCCAGGACGGAAAGCCGTCGA (SPACER - REPEAT)7 REPEAT
....TTTCGCACCATGGCCGACCTCCAGGACGGAAAGCCGTCGA....
Stro522 (hypothetical)
ATTL (tRNA)
ATTR
S. tropica
S. tropica prophage 1
47kb


a Composition
230
344
357
776
1103
Genes in SA and ST
ST Rearranged
SA Rearranged
ST Specific
SA Specific
b History
42
135
223
211
199
250
184
259
128
248
ST No evidence
SA No evidence
ST Paralogs
SA Paralogs
ST Xenologs
SA Xenologs
ST Pseudogenes
SA Pseudogenes
ST HGT
SA HGT
See bar graph below
c Function
198
323
90
94
49
99
185
242
254
345
ST Sec mets
SA Sec mets
ST Prophage genes
SA Prophage genes
ST MGE
SA MGE
ST hypothetical
SA hypothetical
SA other
SA other

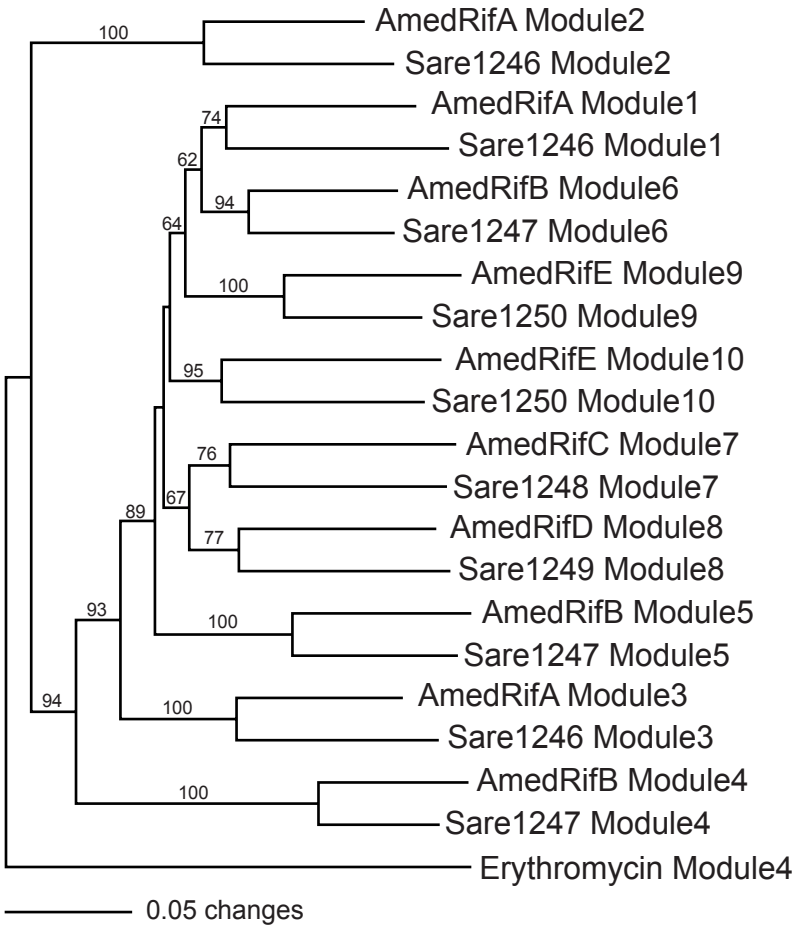
d
Salinispora genes
Actinobacteria genomes
SA No Evidence
ST No Evidence
0
10
20
30
40
50
60
70
80
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

Table 1. General genome features.

Feature	*S. tropica* (ST)	ST%	*S. arenicola* (SA)	SA %
No. base pairs	5183331	NA	5786361	NA
% G+C	69.4	NA	69.5	NA
Total genes	4536	NA	4919	NA
Pseudogenes	57	1.26%	192	3.90%
Hypotheticals (% genome)	1140	25.10%	1418	28.80%
No. rRNA operons (% identity)	3	100%	3	100%
Orthologs (% genome)	3606	79.40%	3606	73.20%
Positional orthologs (% genome)	3178	70.10%	3178	64.60%
Rearranged orthologs (% genome)	428	9.40%	428	8.70%
Species-specific genes (% genome)	987	21.80%	1505	30.60%
Island genes (% genome)	1350	29.80%	1690	34.30%
Total genes with evidence of HGT (% genome)	652	14.30%	750	14.70%
Species-specific genes with evidence of HGT (% of species-specific)	405	41.00%	573	38.10%
Total island genes with evidence of HGT (% HGT)	473	72.50%	555	74.00%
Paralogs* (% genome)	1819	39.60%	2179	42.60%
Species-specific paralogs (% species-specific genes)	391	39.70%	647	43.00%
Secondary metabolism (% genome)	405	8.80%	556	10.90%

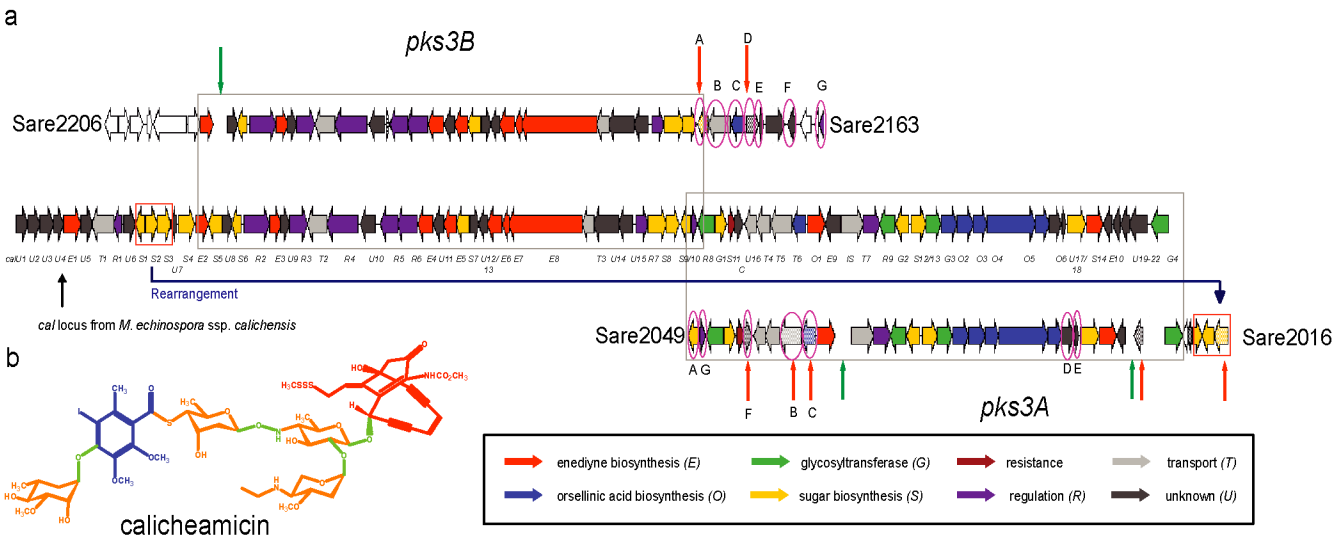
*Paralog totals include parental gene.
NA not applicable.

Supplementary Figures and Legends



Supplementary Figure 1. Polyketide synthase phylogeny. Neighbor-joining distance tree constructed using the aligned amino acid sequences of the *rif* KS domains from *A. mediterranei* and *S. arenicola*. Bootstrap values (in percent) calculated from 1000 re-samplings are shown at their respective nodes for values greater than or equal to 60%. The KS domain from module 4 of the erythromycin biosynthetic pathway (*Saccharopolyspora erythraea*) was used to position the root.

Supplementary Figure 2. Polymorphic Membrane Protein (PMP) phylogeny. Neighbor-joining distance tree constructed in APIS (J. Badger, unpublished) using the aligned amino acid sequences of SA and ST PMPs as well as those observed in other genomes. Bold lines indicate boot-strap values >50% and blue indicates strains other than SA and ST that were derived from aquatic environments. Accession numbers in parentheses.



Supplementary Figure S3. Cluster SA *pks3A* and *pks3B* in comparison with the *cal* locus from *M. echinospora*. **a** Grey boxes indicate regions of gene conservation. Duplicated genes are circled in red with parologs identified by letter. Red arrows indicate pseudogenes (which are also checkered). Genes missing (green arrows) and unique (colored white) relative to the *cal* locus are indicated. **b** structure of calicheamicin.

Supplementary Tables

Supplementary Table 1. Genomic islands.

Island # strain	Start position	Stop position	Size bp	Total bp's	Start gene	Stop gene	# of genes	Total genes
1 ST	67688	92154	24,466		58	83	26	
1 SA	73610	95007	21,397	45,863	63	80	18	44
2 ST	340915	355342	14,427		300	304	5	
2 SA	381999	427379	45,380	59,807	345	367	23	28
3 ST	471193	472396	1,203		410	411	2	
3 SA	547253	570209	22,956	24,159	478	499	22	24
4 ST	512154	781349	269,195		449	694	246	
4 SA	608623	723155	114,532	383,727	537	641	105	351
5 ST	1107318	1215323	108,005		988	1068	81	
5 SA	1040020	1082195	42,175	150,180	924	958	35	116
6 ST	1271151	1324135	52,984		1127	1180	54	
6 SA	1139966	1202883	62,917	115,901	1018	1073	56	110
7 ST	1477636	1495949	18,313		1315	1357	43	
7 SA	1354284	1521916	167,632	185,945	1204	1314	111	154
8 ST	1702965	1734332	31,367		1492	1524	33	
8 SA	1685105	1694512	9,407	40,774	1457	1466	10	43
9 ST	1803340	1855755	52,415		1585	1631	47	
9 SA	1771879	1850282	78,403	130,818	1536	1617	82	129
10 ST	2206426	2298319	91,893		1931	2067	137	
10 SA	2218415	2546802	328,387	420,280	1922	2210	289	426
11 ST	2460444	2522674	62,230		2172	2230	59	
11 SA	2672473	2701243	28,770	91,000	2326	2347	22	81
12 ST	2575986	2626461	50,475		2282	2333	52	
12 SA	2756098	2832944	76,846	127,321	2400	2476	77	129
13 ST	2650556	2667541	16,985		2355	2373	19	
13 SA	2856961	2871093	14,132	31,117	2500	2512	13	32
14 ST	2750480	2781381	30,901		2445	2473	29	
14 SA	2967508	3029499	61,991	92,892	2601	2656	56	85
15 ST	2968640	3325240	356,600		2645	2909	265	
15 SA	3227645	3533832	306,187	662,787	2842	3109	268	533
16 ST	3357796	3368101	10,305		2937	2946	10	
16 SA	3568463	3657421	88,958	99,263	3143	3170	33	43
17 ST	3910860	3921251	10,391		3407	3417	11	
17 SA	4217838	4322435	104,597	114,988	3655	3794	140	151
18 ST	4543960	4565093	21,133		3991	4016	26	
18 SA	4942782	4969601	26,819	47,952	4375	4397	23	49
19 ST	4634866	4636953	2,087		4077	4077	1	
19 SA	5057490	5084903	27,413	29,500	4476	4497	22	23
20 ST	4688038	4738420	50,382		4121	4239	119	
20 SA	5136948	5290253	153,305	203,687	4543	4669	127	246
21 ST	4936430	4954143	17,713		4357	4441	85	
21 SA	5432928	5629894	196,966	214,679	4799	4956	158	243

Supplementary Table 2. Secondary metabolite gene clusters in *S. tropica* (ST) and *S. arenicola* (SA).

#	Strain Cluster name	Equivalent cluster	Biosynthetic class	Product(s)	Biological activity/target	Island	Gene start	Gene stop	# of Genes
1	ST *pks1*	none	polyketide	10-membered enediyne	cytotoxin/DNA damage	4	586	**610**	25
2	ST *nrps1*	SA *nrps3**	non-ribosomal peptide	dipeptide	N/D	4/15	**667**	**694**	28
3	ST *sal*	none	polyketide/non-ribosomal peptide	**salinosporamide**	**cytotoxin/proteasome**	5	1012	1043	32
4	ST *pks2*	none	polyketide	glycosylated decaketide	N/D	11	**2174**	**2227**	54
5	ST *amc*	SA *amc*	carbohydrate	aminocyclitol	N/D	NI/NI	2340	2346	7
6	ST *bac1*	SA *bac2*	ribosomal peptide	class I bacteriocin (non-lantibiotic)	antimicrobial	NI/NI	**2428**	**2440**	13
7	ST *pks3*	SA *pks4*	polyketide	aromatic polyketide	N/D	NI/NI	2486	**2510**	25
8	ST *des***	SA *des*	hydroxamate	**desferrioxamine**+	**siderophore/iron chelation**	NI/NI	**2541**	**2555**	15
9	ST *sid2*	SA *sid1**	non-ribosomal peptide	yersiniabactin-related	iron chelation	**15/10**	**2645**	**2659**	15
10	ST *spo*	none	polyketide	**sporolide**	N/D	**15**	**2691**	**2737**	47
11	ST *slm*	none	polyketide	**salinilactam**	N/D	**15**	2757	2781	25
12	ST *sid3*	none	non-ribosomal peptide	dihydroaeruginoic acid-related siderophore	siderophore/iron chelation	**15**	2786	2813	28
13	ST *sid4*	none	non-ribosomal peptide	coelibactin-related siderophore	siderophore/iron chelation	**15**	2814	2842	29
14	ST *bac2*	SA *bac3*	ribosomal peptide	class I bacteriocin (non-lantibiotic)	antimicrobial	NI/NI	3042	**3054**	13
15	ST *lym*	SA *lym*	polyketide/non-ribosomal peptide	**lymphostin**+	**immunosuppressant**	NI/NI	3055	3066	12
16	ST *terp1*	SA *terp2*	terpenoid	carotenoid pigment	antioxidant	NI/NI	3244	3253	10
17	ST *pks4*	SA *pks6*	polyketide	phenolic lipids	cell wall lipid	NI/NI	**4264**	**4267**	4
18	ST *nrps2*	SA *nrps4*	non-ribosomal peptide	tetrapeptide	N/D	21/21	**4410**	4429	20
19	ST *terp2*	SA *terp3*	terpenoid	carotenoid pigment	antioxidant	21/21	4437	**4441**	5
								Total	407

1	SA *nrps1*	none	non-ribosomal peptide	pentapeptide	N/D	2	345	367	23
2	SA *pksnrps1*	none	polyketide/non-ribosomal peptide	N/D	N/D	3	478	499	22
3	SA *pks1A*	none	polyketide	9-membered enediyne unit/kedarcidin-related, fragment A	cytotoxin/DNA damage	**4**	545	560	16
4	SA *misc1*	none	aminoacyl tRNA synthetase-derivec	amino acid conjugate	N/D	**4**	570	573	4
5	SA *bac1*	none	ribosomal peptide	class I bacteriocin (lantibiotic)	antimicrobial	**4**	602	623	22
6	SA *pks2*	none	polyketide	N/D	N/D	6	1041	1073	33
7	SA *rif*	none	polyketide	**rifamycin**+	**antibiotic/RNA polymerase**	**7**	1240	1278	39
8	SA *terp1*	none	terpenoid	diterpene	N/D	**7**	1286	1288	3
9	SA *pks3A*	none	polyketide	10-membered enediyne unit/calicheamicin-related, fragment A	cytotoxin/DNA damage	**10**	2017	2049	33
10	SA *sid1**	ST *sid2*	non-ribosomal peptide	yersiniabactin-related	siderophore/iron chelation	**10/15**	2070	2081	12
11	SA *pks1B*	none	polyketide-associated	nodified tyrosine and deoxysugar units/kedarcidin-related, fragment I	cytotoxin/DNA damage	**10**	2088	2121	34
12	SA *misc2*	none	aminoacyl tRNA synthetase-derivec	amino acid conjugate	N/D	**10**	2144	2151	8
13	SA *pks3B*	none	polyketide-related	aryltetrasaccharide unit/calicheamicin-related, fragment B	cytotoxin/DNA damage	**10**	2163	2206	44
14	SA *sta*	none	indolocarbazole	**staurosporine**+	**cytotoxin/protein kinase**	11	2326	2342	17
15	SA *pksnrps2*	none	polyketide/non-ribosomal peptide	N/D	N/D	12	2400	2409	10
16	SA *amc*	ST *amc*	carbohydrate	aminocyclitol	N/D	NI/NI	2483	2491	9
17	SA *bac2*	ST *bac1*	ribosomal peptide	class I bacteriocin (non-lantibiotic)	antimicrobial	NI/NI	2583	2595	13
18	SA *pks4*	ST *pks3*	polyketide	aromatic polyketide	N/D	NI/NI	2669	2694	26
19	SA *des*	ST *des*	hydroxamate	**desferrioxamine**+	**siderophore/iron chelation**	NI/NI	2728	2744	17
20	SA *nrps2*	none	non-ribosomal peptide	tetrapeptide	N/D	15	2939	2968	30
21	SA *nrps3**	ST *nrps1*	non-ribosomal peptide	dipeptide	N/D	15/4	3051	3063	13
22	SA *pks5*	none	polyketide	macrolide	N/D	16	3148	3163	16
23	SA *bac3*	ST *bac2*	ribosomal peptide	class I bacteriocin (non-lantibiotic)	antimicrobial	NI/NI	3268	3280	13
24	SA *lym*	ST l*ym*	polyketide	**lymphostin**+	**immunosuppressant**	NI/NI	3281	3293	13
25	SA *terp2*	ST *terp1*	terpenoid	carotenoid pigment	antioxidant	NI/NI	3471	3480	10
26	SA *cym*	none	non-ribosomal peptide	**cyclomarin**+	**nti-inflammatory, antivira**	20	4547	4569	23
27	SA *pks6*	ST *pks4*	polyketide	phenolic lipids	cell wall lipid	NI/NI	4694	4697	4
28	SA *nrps4*	ST *nrps2*	non-ribosomal peptide	tetrapeptide	N/D	21/21	4885	4904	20
29	SA *terp3*	ST *terp2*	terpenoid	carotenoid pigment	antioxidant	21/21	4927	4931	5
30	SA *pks1C*	none	polyketide	naphthoic acid unit/kedarcidin-related, fragment C	cytotoxin/DNA damage	21	4932	4956	25
								Total	540

NI: non-island. Italics: predicted product or activity. Bold: observed product or activity. * Partial cluster. ** Previously designated ST Sid1 (32). + Product observed in other bacteria. N/D: not determined.

Supplementary Table 3. Mobile Genetic Elements (MGEs).

S. tropica	Start gene	Stop gene	# Genes	Island
AICE1	58	74	17	1
Phage integrase	505	505	1	4
Prophage 1	507	559	53	4
IS1380	570	570	1	4
IS256	586	587	2	4
ISNCY	608	608	1	4
ISNCY	609	609	1	4
IS3	648	648	1	4
Unknown MGE	988	994	7	5
IS1380	1014	1014	1	5
IS3	1164	1165	2	6
IS5	1315	1315	1	7
phage gene	1317	1317	1	7
IS701	1506	1518	13	8
Unknown MGE	1602	1609	8	9
IS5	1614	1614	1	9
Prophage 2	1931	1957	27	10
Phage gene	1980	1980	1	10
Phage gene	1983	1983	1	10
Phage gene	2002	2002	1	10
Phage gene	2013	2013	1	10
IS630	2021	2022	2	10
Tn3	2304	2304	1	12
IS110	2305	2305	1	12
Tn3	2369	2369	1	13
IS110	2466	2466	1	14
IS5	2661	2661	1	15
IS1380	2716	2717	2	15
IS630	2729	2730	2	15
IS701	2752	2753	2	15
IS630	2845	2846	2	15
IS110	2861	2861	1	15
IS630	2891	2891	1	15
unk IS	2899	2899	1	15
Unknown MGE	2908	2909	2	15
IS5	2941	2941	1	16
Phage gene	3417	3417	1	17
Prophage 3	3986	4017	32	18
IS630	4122	4123	2	20
Tn3	4134	4134	1	20
Tn3	4137	4137	1	20
ISL3	4138	4138	1	20
Rev transcriptase	4139	4139	1	20
IS5	4140	4140	1	20
IS3	4141	4141	1	20
transposase	4142	4142	1	20
IS5	4179	4179	1	20
IS30	368	368	1	NI
Unknown MGE	749	756	8	NI
IS66	1556	1556	1	NI
IS110	1662	1662	1	NI
Phage gene	2334	2334	1	NI
Phage gene	2347	2347	1	NI
IS3	3350	3351	2	Ni
Phage gene	3352	3352	1	NI
IS5	3501	3506	6	NI
IS630	3656	3662	7	NI
Total			235	

S. arenicola	Start gene	Stop gene	# Genes	Island
Tn3	346		1	2
Recombinase	612		1	4
Plasmid	925	958	34	5
IS21	1024	1025	2	6
ICE1	1208	1227	20	7
ICE2	1562	1580	19	9
IS21	1590	1591	2	9
Phage gene	1612	1613	2	9
IS701	1650	1650	1	10
IS256	1651	1651	1	10
ICE3	1922	1939	18	10
IS21	1968	1969	2	10
IS5	1979	1979	1	10
IS3	1991	1991	1	10
IS5	1998	1998	1	10
Recombinase	2051	2051	1	10
Unknown MGE	2456	2477	22	12
IS21	2854	2855	2	15
Phage gene	2857	2857	1	15
Plasmid gene	2979	2979	1	15
IS110	2982	2982	1	15
IS5	3023	3023	1	15
IS4	3041	3041	1	15
Phage gene	3074	3074	1	15
Recombinase	3094	3094	1	15
IS4	3105	3105	1	15
IS4	3106	3106	1	15
IS630	3107	3107	1	15
IS21	3160	3161	2	16
Prophage 1A	3692	3743	52	17
Prophage 1B	3744	3794	51	17
IS5	4558	4558	1	20
IS630	4571	4571	1	20
IS21	4925	4926	2	21
Recombinase	413	413	1	NI
Old Plasmid	1501	1502	2	NI
IS630	1649	1649	1	NI
Recombinase	1915	1915	1	NI
IS630	2285	2285	1	NI
Recombinase	2492	2492	1	NI
IS21	2580	2581	2	NI
IS630	3178	3178	1	NI
IS630	3576	3576	1	Ni
IS	4038	4038	1	NI
ISL3	4192	4192	1	NI
Phage gene	4977	4977	1	NI
Total			264	