npg

ORIGINAL ARTICLE

Genomic islands link secondary metabolism to 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

Genomic islands have been shown to harbor functional traits that differentiate ecologically distinct populations of environmental bacteria. A comparative analysis of the complete genome sequences of the marine Actinobacteria *Salinispora tropica* and *Salinispora arenicola* reveals that 75% of the species-specific genes are located in 21 genomic islands. These islands are enriched in genes associated with secondary metabolite biosynthesis providing evidence that secondary metabolism is linked to functional adaptation. Secondary metabolism accounts for 8.8% and 10.9% of the genes in the *S. tropica* and *S. arenicola* genomes, respectively, and represents the major functional category of annotated genes that differentiates the two species. Genomic islands harbor all 25 of the species-specific biosynthetic pathways, the majority of which occur in *S. arenicola* and may contribute to the cosmopolitan distribution of this species. Genome evolution is dominated by gene duplication and acquisition, which in the case of secondary metabolism provide immediate opportunities for the production of new bioactive products. Evidence that secondary metabolic pathways are exchanged horizontally, coupled with earlier evidence for fixation among globally distributed populations, supports a functional role and suggests that the acquisition of natural product biosynthetic gene clusters represents a previously unrecognized force driving bacterial diversification. Species-specific differences observed in clustered regularly interspaced short palindromic repeat sequences suggest that *S. arenicola* may possess a higher level of phage immunity, whereas a highly duplicated family of polymorphic membrane proteins provides evidence for a new mechanism of marine adaptation in Gram-positive bacteria.

Introduction

Linking functional traits to bacterial phylogeny remains a fundamental but elusive goal of microbial ecology (Hunt *et al.*, 2008). Without this information, it becomes difficult to resolve meaningful units of diversity and the mechanisms by which bacteria interact with each other and adapt to environmental change. Most bacterial diversity is delineated among clusters of sequences that share >99% 16S rRNA gene sequence identity (Acinas *et al.*, 2004). These sequence clusters are believed to represent fundamental units of diversity, whereas intra-cluster microdiversity is thought to persist because of weak selective pressures (Acinas *et al.*, 2004), suggesting little ecological or taxonomic relevance. Recently, progress has been made in terms of delineating units of diversity that possess the fundamental properties of species by linking genetic diversity with ecology and evolutionary theory (Achtman and Wagner, 2008; Fraser *et al.*, 2009). Despite these advances, there remains no widely accepted species concept for prokaryotes (Gevers *et al.*, 2005), and sequence-based analyses reveal widely varied levels of diversity within assigned species boundaries.

The comparative analysis of bacterial genome sequences has revealed considerable differences among closely related strains (Joyce et al., 2002; Welch et al., 2002; Thompson et al., 2005) and provides a new perspective on genome evolution and prokaryotic species concepts. Genomic differences among closely related strains are concentrated in islands, strain-specific regions of the chromosome that are generally acquired by horizontal gene transfer (HGT) and that harbor functionally adaptive traits (Dobrindt et al., 2004) that can be linked to niche adaptation. The pelagic cyanobacterium Prochlorococcus is an important model for the study of island genes, which in this case are differentially expressed under low nutrient and high light stress in ecologically distinct populations (Coleman et al., 2006). Despite convincing evidence for the adaptive significance of island genes among environmental bacteria, the precise functions of their products have seldom been characterized and their potential role in the evolution of independent bacterial lineages remains poorly understood.

The marine sediment-inhabiting genus Salinispora belongs to the order Actinomycetales, a group of Actinobacteria commonly referred to as actinomycetes. Actinomycetes are a rich source of structurally diverse secondary metabolites and account for the majority of antibiotics discovered as of 2002 (Berdy, 2005). Salinispora spp. have likewise proven to be a rich source of secondary metabolites (Fenical and Jensen, 2006), including salinosporamide A, which is currently in clinical trials for the treatment of cancer (Fenical et al., 2009). At present, the genus is comprised of three species that collectively constitute a microdiverse sequence cluster (sensu (Acinas et al., 2004)), that is they share ⩾99% 16S rRNA gene sequence identity (Jensen and Mafnas, 2006). Although the microdiversity within this cluster has been formally delineated into species-level taxa (Maldonado et al., 2005), it remains to be determined if these taxa represent ecologically or functionally distinct lineages.

We here report the comparative analysis of the complete genome sequences of Salinispora tropica (strain CNB-440, the type strain for the species and thus a contribution to the Genomic Encyclopedia of Bacteria and Archaea project), hereafter referred to as ST, and Salinispora arenicola (strain CNS-205), hereafter referred to as SA, the first obligately marine Actinobacteria to be obtained in culture (Mincer et al., 2002). The aims of this study were to describe, compare and contrast the gene content and organization of the two genomes in the context of prevailing species concepts, identify the functional attributes that differentiate the two species, assess the processes that have driven genome evolution and search for evidence of marine adaptation in this unusual group of Gram-positive marine bacteria.

Materials and methods

Sequencing and ortholog identification
The sequencing and annotation of the SA genome was as previously reported for ST (Udwary et al., 2007). Both genomes were sequenced as part of the Department of Energy, Joint Genome Institute, Community Sequencing Program. Orthologs within the two genomes were predicted using the Reciprocal Smallest Distance method (Wall et al., 2003), 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 that are flanked by regions of conservation and within which <40% of the genes possess a positional ortholog in the reciprocal genome. Paralogs within each genome were identified using the blastclust algorithm (Dondoshansky and Wolf, 2000), with a cutoff of 30% identity over 40% of the sequence length. The automated phylogenetic inference system (APIS) was used to identify recent gene duplications (Badger et al., 2005).

Horizontal gene transfer
All genes were assessed for evidence of HGT based on abnormal DNA composition, phylogenetic, taxonomic and sequence-based relationships, and comparisons with known mobile genetic elements. Genes identified by ⩾2 different methodologies were counted as positive for HGT. To reflect confidence in the assignments, genes displaying positive evidence of HGT were color coded from yellow to red corresponding to total scores from 2 to 6. The results were mapped onto the genome to reveal HGT-clustering patterns and adjacent clusters were merged (Figure 1a). Four DNA compositional analyses included G + C content (obtained from the JGI annotation), codon adaptive index, calculated with the CAI calculator (Wu et al., 2005) using a suite of housekeeping genes as reference, dinucleotide frequency differences ($\delta*$), calculated using IslandPath (http://www.pathogenomics.sfu.ca/islandpath/update/IPindex.pl) (Hsiao et al., 2003), and DNA composition, calculated using Alien_Hunter (http://www.sanger.ac.uk/Software/analysis/alien_hunter/) (Vernikos and Parkhill, 2006). G + C content or codon usage values >1.5 s.d. from the genomic mean and dinucleotide frequency differences >1 s.d. from the mean were scored positive for HGT. Taxonomic relationships in the form of lineage probability index (LPI) values for all protein-coding genes were assigned using the Darkhorse algorithm (Podell and Gaasterland, 2007). Genes with an LPI of <0.5, indicating that the orthologs are not in closely related genomes, were scored positive for HGT. A reciprocal Darkhorse analysis (Podell et al., 2008) was then performed on the orthologs of all positives, and if these genes had an LPI score >0.5, indicating that the match
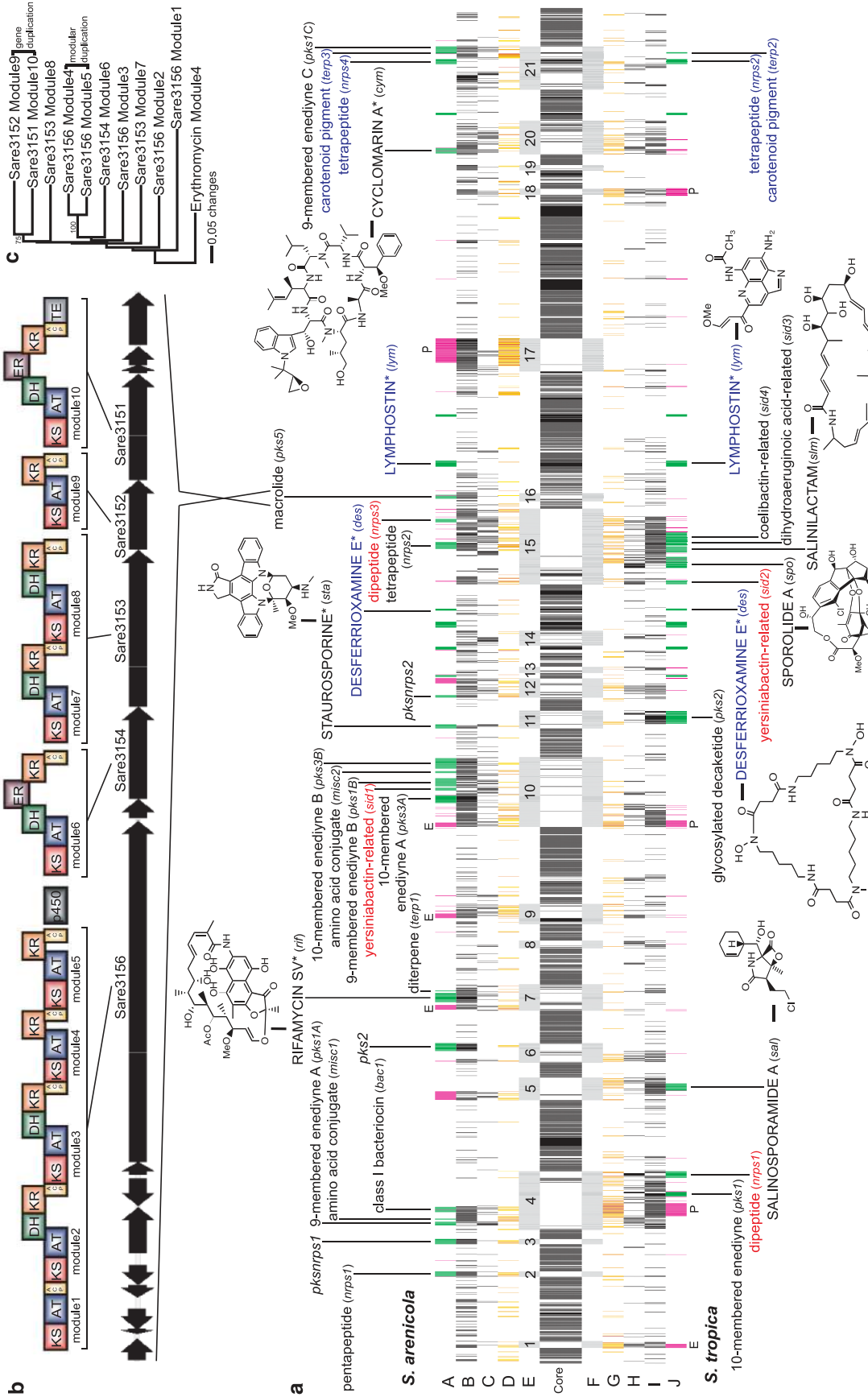
Figure 1 Linear alignment of the *Salinispora tropica* and *Salinispora arenicola* genomes starting with the origins of replication. (**a**) Positional orthologs (core) flanked by islands (E and F), heat-mapped HGT genes (D and G), rearranged orthologs (C and H), species-specific genes (B and I), secondary metabolite genes (green), MGEs (pink) with prophage (P) and AICEs (E) indicated (A and J). For genomic islands, predicted (lower case) and isolated (uppercase with structures) secondary metabolites 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 from SA *pks5* revealing gene and modular duplication events (erythromycin root, % bootstrap values from 1000 re-samplings). AICEs, actinobacterial integrative and conjugative elements; MGEs, mobile genetic elements.

4

sequence is phylogenetically typical within its own lineage, they were assigned an additional positive score.

A phylogenetic approach using the APIS program (Badger et al., 2005) 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 cladding with non-actinobacterial homologs were binned into their respective taxonomic groups and given a positive HGT score. Evidence of HGT was also inferred from Reciprocal Smallest Distance 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 present in SA and/or ST and not observed among the 27 actinobacterial genomes were assigned a positive HGT score. Bacteriophages were identified using Prophage (http://bioinformatics.uwp.edu/~phage/Prophage Finder.php) (Bose and Barber, 2006) and Phage Finder (http://phage-finder.sourceforge. net/) (Fouts, 2006). 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 a positive HGT score. 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.

Clustered regularly interspaced short palindromic repeats (CRISPRs) were identified using CRISPR finder (http://crispr.u-psud.fr/Server/CRISPRfinder. php), whereas repeats larger than 35 bases were identified using Reputer (http://bibiserv.techfak. uni-bielefeld.de/reputer/) (Kurtz et al., 2001). Secondary metabolite gene clusters were manually annotated as in Udwary et al. (2007). Cluster boundaries were predicted using gene clusters reported earlier when available as in the case of rifamycin. For unknown clusters, loss of gene conservation across the Actinobacteria was used to aid boundary predictions. In the future, programs such as 'ClustScan' may prove useful for pathway annotation and product prediction (Starcevic et al., 2008). However, many biosynthetic genes are large (5–10 kb) and highly repetitive creating challenges associated with gene calling and assembly (Udwary et al., 2007) and the interpretation of operon structure. 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.

Results and discussion

The ST and SA genomes are circular and 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% cutoff that has been suggested to delineate bacterial species (Konstantinidis and Tiedje, 2005). Despite differing by only seven nucleotides (99.7% identity) in the 16S rRNA gene, the genome of SA is 603 kb (11.6%) larger and possesses 1505 species-specific genes compared

Table 1 General genome features

Feature	Salinispora tropica (ST)	ST%	Salinispora arenicola (SA)	SA%
No. of base pairs	5 183 331	NA	5 786 361	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. of 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[a] (% 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

Abbreviations: HGT, horizontal gene transfer; NA, not applicable.
[a]Totals include parental gene.

Table 2 Secondary metabolite gene clusters in *Salinispora tropica* (ST)

No.	Cluster name	Equivalent cluster	Biosynthetic class	Product	Biological activity/ target	Island	Gene start	Gene stop	No. of genes
1	ST *pks1*	None	Polyketide	*10-membered enediyne*	*Cytotoxin/DNA*	4	586	610	25
2	ST *nrps1*	SA *nrps3*[a]	Non-ribosomal peptide	*Dipeptide*	ND	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*	ND	11	2174	2227	54
5	ST *amc*	SA *amc*	Carbohydrate	*Aminocyclitol*	ND	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*	ND	NI/NI	2486	2510	25
8	ST *des*[b]	SA *des*	Hydroxamate	**Desferrioxamine**[c]	**Siderophore/iron chelation**	NI/NI	2541	2555	15
9	ST *sid2*	SA *sid1*[a]	Non-ribosomal peptide	*Yersiniabactin related*	*Siderophore/iron chelation*	15/10	2645	2659	15
10	ST *spo*	None	Polyketide	**Sporolide**	ND	15	2691	2737	47
11	ST *slm*	None	Polyketide	**Salinilactam**	ND	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**[c]	**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*	ND	21/21	4410	4429	20
19	ST *terp2*	SA *terp3*	Terpenoid	*Carotenoid pigment*	*Antioxidant*	21/21	4437	4441	5
								Total	407

Abbreviations: ND, not determined; NI, non-island.
Italics: predicted product or activity; bold: observed product or activity.
[a]Partial cluster.
[b]Previously designated ST Sid1 (32).
[c]Product observed in other bacteria.

with 987 in ST. Seventy-five percent of these species-specific genes are located in 21 genomic islands (Table 1, Supplementary Table S1), none of which are comprised of genes originating entirely from one genome (Figure 1). The presence of genomic islands in the same location on the chromosomes of closely related bacteria is well recognized (Coleman *et al.*, 2006) and facilitated by the presence of transfer RNAs (tRNAs) (Tuanyok *et al.*, 2008). Twelve islands in the *Salinispora* alignment share at least one tRNA between both genomes and, of those, four share two or more tRNAs within a single island, indicating multiple insertion sites. In addition to tRNAs, direct repeats detected in the same location in both genomes could also act as insertion sites to help create islands.

The *Salinispora* genomic 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, whereas 8 of the 12 shared pathways occur in the genus-specific core (Tables 2 and 3). We have isolated and identified the products of eight of these pathways, which include the highly selective proteasome inhibitor salinosporamide A (Feling *et al.*, 2003), as well as sporolide A (Buchanan *et al.*, 2005), which is derived from an enediyne polyketide precursor (Udwary *et al.*, 2007), one

of the most potent classes of biologically active agents discovered to date. An earlier analysis of 46 *Salinispora* strains revealed that secondary metabolite production is the major phenotypic difference among the three species (Jensen *et al.*, 2007), an observation supported at the genomic level by the analysis of the *S. tropica* genome (Udwary *et al.*, 2007) and now this study.

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 for 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 antibiotic (Yu *et al.*, 1999). This analysis confirms earlier observations of HGT in this pathway (Kim *et al.*, 2006) and 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 (Supplementary Figure S1). Evidence of HGT coupled with earlier evidence for the fixation of specific pathways, such as *rif,* among globally distributed SA populations (Jensen *et al.*, 2007) supports vertical inheritance following pathway

Table 3 Secondary metabolite gene clusters in *Salinispora arenicola* (SA)

No.	Cluster name	Equivalent cluster	Biosynthetic class	Product	Biological activity/ target	Island	Gene start	Gene stop	No. of genes
1	SA *nrps1*	None	Non-ribosomal peptide	*Pentapeptide*	ND	2	345	367	23
2	SA *pksnrps1*	None	Polyketide/non-ribosomal peptide	ND	ND	3	478	499	22
3	SA *pks1A*	None	Polyketide	*9-membered enediyne unit/kedarcidin-related, fragment A*	*Cytotoxin/DNA*	4	545	560	16
4	SA *misc1*	None	Aminoacyl tRNA synthetase derived	*Amino acid conjugate*	ND	4	570	573	4
5	SA *bac1*	None	Ribosomal peptide	*Class I bacteriocin (lantibiotic)*	*Antimicrobial*	4	602	623	22
6	SA *pks2*	None	Polyketide	ND	ND	6	1041	1073	33
7	SA *rif*	None	Polyketide	**Rifamycin**[b]	**Antibiotic/RNA polymerase**	7	1240	1278	39
8	SA *terp1*	None	Terpenoid	*Diterpene*	ND	7	1286	1288	3
9	SA *pks3A*	None	Polyketide	*10-membered enediyne unit/calicheamicin-related, fragment A*	*Cytotoxin/DNA*	10	2017	2049	33
10	SA *sid1*[b]	ST *sid2*	Non-ribosomal peptide	*Yersiniabactin related*	*Siderophore/iron chelation*	10/15	2070	2081	12
11	SA *pks1B*	None	Polyketide associated	*Modified tyrosine and deoxysugar units/ kedarcidin-related, fragment B*	*Cytotoxin/DNA*	10	2088	2121	34
12	SA *misc2*	None	Aminoacyl tRNA synthetase derived	*Amino-acid conjugate*	ND	10	2144	2151	8
13	SA *pks3B*	None	Polyketide related	*Aryltetrasaccharide unit/ calicheamicin-related, fragment B*	*Cytotoxin/DNA*	10	2163	2206	44
14	SA *sta*	None	Indolocarbazole	**Staurosporine**[b]	**Cytotoxin/ protein kinase**	11	2326	2342	17
15	SA *pksnrps2*	None	Polyketide/non-ribosomal peptide	ND	ND	12	2400	2409	10
16	SA *amc*	ST *amc*	Carbohydrate	*Aminocyclitol*	ND	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*	ND	NI/NI	2669	2694	26
19	SA *des*	ST *des*	Hydroxamate	**Desferrioxamine**[b]	**Siderophore/iron chelation**	NI/NI	2728	2744	17
20	SA *nrps2*	None	Non-ribosomal peptide	*Tetrapeptide*	ND	15	2939	2968	30
21	SA *nrps3*[a]	ST *nrps1*	Non-ribosomal peptide	*Dipeptide*	ND	15/4	3051	3063	13
22	SA *pks5*	None	Polyketide	*Macrolide*	ND	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 lym	Polyketide	**Lymphostin**[b]	**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**[b]	**Anti-inflammatory, antiviral**	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*	ND	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*	21	4932	4956	25
								Total	540

Abbreviations: ND, not determined; NI, non-island.
Italics: predicted product or activity; bold: observed product or activity.
[a]Partial cluster.
[b]Product observed in other bacteria.

acquisition (Ochman *et al.*, 2005). This evolutionary history is what might be expected if pathway acquisition fostered ecotype diversification or a selective sweep (Cohan, 2002) resulting from strong selection for the acquired pathway, either of which provides compelling evidence that secondary metabolites represent functional traits with important ecological roles. The concept that gene acquisition
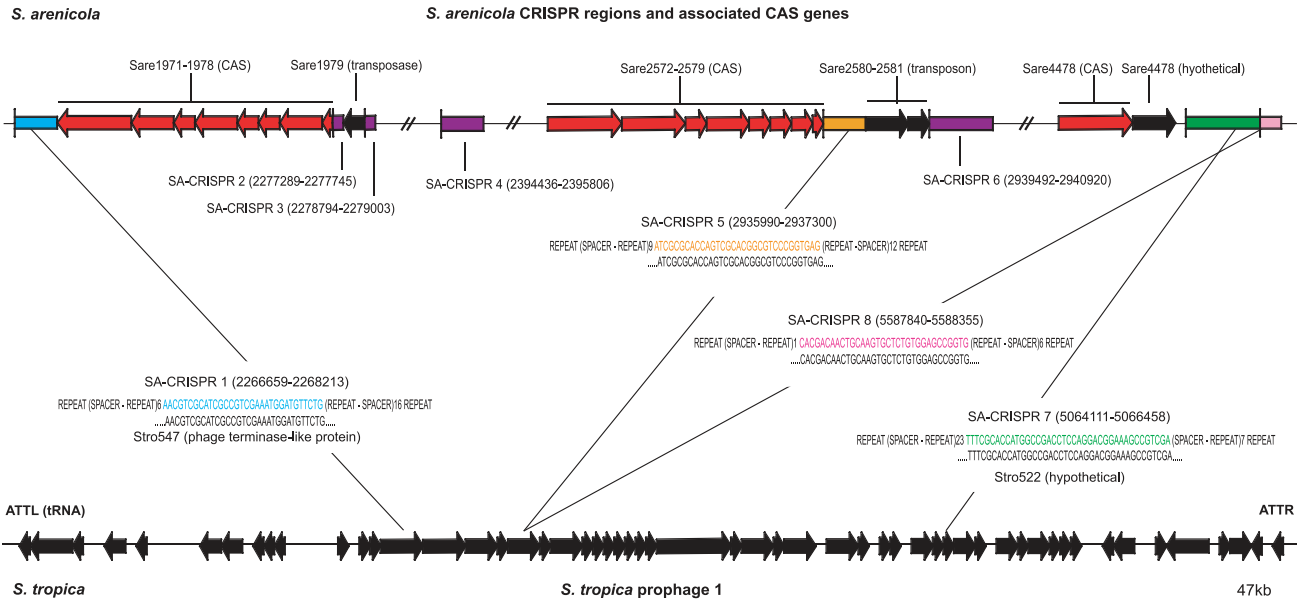
Figure 2 *Salinispora tropica* prophage and *Salinispora arenicola* CRISPRs. Four of eight SA CRISPRs (1, 5, 7 and 8) have spacers (color coded) that share 100% sequence identity with genes (Stro numbers and annotation given) in ST prophage 1 (Supplementary Table S2, inverted for visual purposes). Other CRISPRs are colored purple. 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) and genes interrupting CRISPRs (black) are indicated. None of the spacer sequences possessed 100% identity to prophage in the NCBI non-redundant sequence database, the SDSU Center for Universal Microbial Sequencing database or the CAMERA metagenomic database. CRISPRs, clustered regularly interspaced short palindromic repeats.

provides a mechanism for ecological diversification that may ultimately drive the formation of independent bacterial lineages has been proposed earlier (Ochman *et al.*, 2000). The inclusion of secondary metabolism among the functional categories of acquired genes that may have this effect sheds new light on the functional importance and evolutionary significance of this class of genes. Although the ecological functions of secondary metabolites remain largely unknown, and thus it is not clear how these molecules might facilitate ecological diversification, there is mounting evidence that they play important roles in chemical defense (Haeder *et al.*, 2009) or as signaling molecules involved in population or community communication (Yim *et al.*, 2007).

Differences between the two species also occur in CRISPR sequences, which are non-continuous direct repeats separated by variable (spacer) sequences that have been shown to confer immunity to phage (Barrangou *et al.*, 2007). The ST genome carries three intact prophages and three CRISPRs (35 spacers), whereas 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 (te Poele *et al.*, 2008). 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). These spacer sequences have no similar matches to genes in the SA prophage or in any prophage sequences deposited in the NCBI, CAMERA or the SDSU Center for Universal Microbial Sequencing databases. The detection of these spacer sequences provides evidence that SA has been exposed to a phage related to one that currently infects ST and that SA now maintains acquired immunity to this phage genotype as has been reported earlier in other bacteria (Barrangou *et al.*, 2007). This is a rare example in which evidence has been obtained for 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, whereas ST strain CNB-440 was recovered 15 years earlier from the Bahamas, it appears that actinophages 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, as evidenced by 140 relative to 35 CRISPR spacer sequences, coupled with a larger genome size and a 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 (Jensen and Mafnas, 2006). Also included among the SA-specific gene pool is a complete phosphotransferase system (Sare4844–4850). Phosphotransferase systems are centrally involved in carbon source uptake and regulation
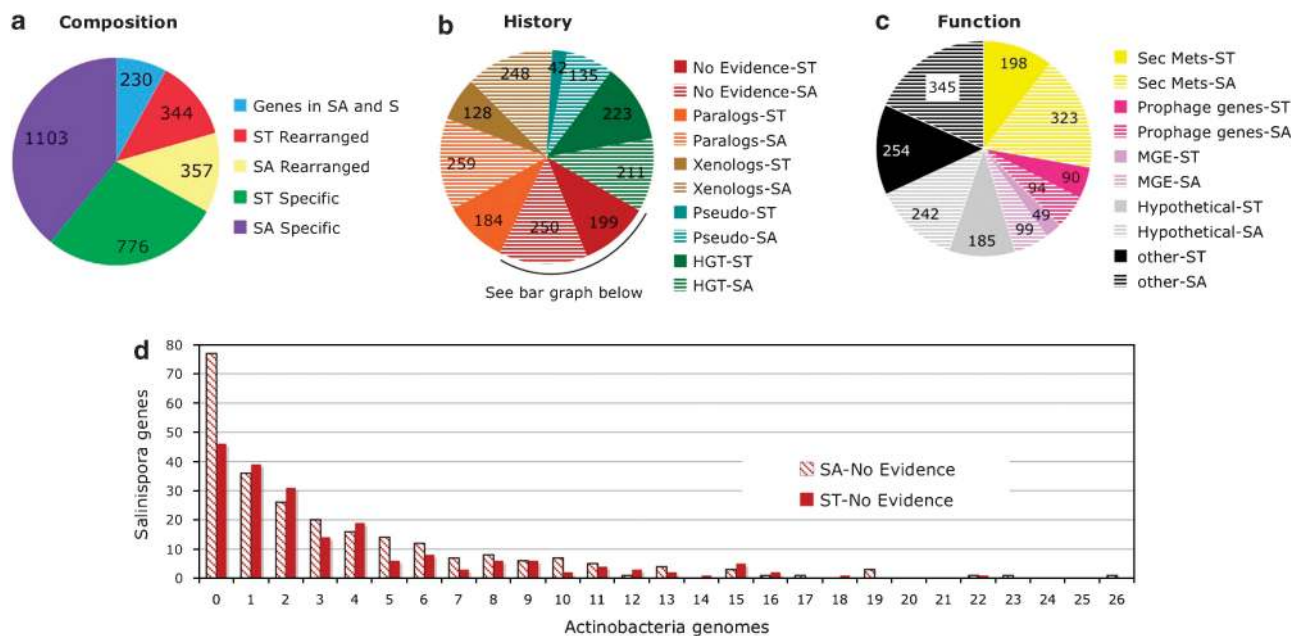
Figure 3 Composition, evolutionary history and function of island genes in *Salinispora tropica* (ST) and *Salinispora arenicola* (SA). (**a**) In all, 3040 genes comprising 21 genomic islands were analyzed for positional orthology (that is, the gene is part of the shared 'core' genome), re-arranged orthology (that is, the gene is present in the other genome but not in the same position or island) and species specificity (gene totals presented in wedges). (**b**) The ST and SA species-specific island genes were analyzed for evidence of paralogy, xenology and HGT. Pseudogenes and the number of genes with no evidence for any of these processes were also identified. (**c**) Functional annotation of the species-specific island genes. (**d**) Distribution of species-specific island genes that have no evidence for HGT or parology among 27 actinobacterial genomes. HGT, horizontal gene transfer.

(Parche *et al.*, 2000) and may provide growth advantages that also factor into the relatively broad distribution of SA. However, additional strains will need to be studied before any of these differences can be firmly linked to species distributions.

The 21 genomic islands are not contiguous regions of species-specific DNA but are instead created by a complex process of gene acquisition, loss, duplication and inactivation (Figure 3). 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 (Coleman *et al.*, 2006). There is dramatic, operon-scale evidence of this process in the shared yersiniabactin-related pathways (ST *sid2* and SA *sid1*), which occur in islands 15 and 10, respectively, and in the unknown dipeptide pathways (ST *nrps1* and SA *nrps3*), which occur in islands 4 and 15, respectively. In both cases, these pathways remain largely intact yet are located in different islands in the two strains (Figure 1, Tables 2 and 3). 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 (Supplementary Figure S2) (Ahlert *et al.*, 2002), 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 analog, whereas flanking genes display a high level of gene duplication and rearrangement indicative of active pathway evolution. Cluster fragmentation is also likely observed in the nine-membered enediyne PKS cluster SA *pks1A–C*, which is scattered across the genome in islands 4, 10 and 21 (Figure 1 and Table 3).

The genomic islands are also enriched in mobile genetic elements including prophage, integrases, and actinobacterial integrative and conjugative elements (AICEs) (Burrus *et al.*, 2002) (Supplementary Tables S2, S3), the later of which are known to play a role in gene acquisition and rearrangement. The *Salinispora* AICEs possess *traB* homologs, which promote conjugal plasmid transfer in mycelial streptomycetes (Reuther *et al.*, 2006), 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 (te Poele *et al.*, 2007), 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* (Schultz *et al.*, 2008), 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 (*rif, sta, des, lym* and *cym*) whose products have also been observed in other bacteria (Figure 1 and Table 3) scored positive for HGT. This observation suggests either that the pathways originated in *Salinispora* or that the exchange of these biosynthetic genes has occurred largely among closely related bacteria and therefore gone undetected with the HGT methods applied in this study. The latter scenario is supported by the observation that all five of the shared biosynthetic pathways were previously reported in other actinomycetes. The acquisition of genes from closely related bacteria likely accounts for many of the species-specific island genes for which no evidence of evolutionary history could be determined (Figure 3b). These genes were poorly conserved among 27 actinobacterial genomes (Figure 3d), providing additional support that they were acquired most likely from environmental Actinobacteria that are not well represented among sequenced genomes. Although gene loss was not quantified, this process is also a likely contributor to island formation. 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 with 1.6% (49/3027) of the non-island pairs. Given that the majority of island genes display evidence of HGT, the increased dN/dS ratio is in agreement with the observation that acquired genes experience relaxed functional constraints (Hao and Golding, 2006).

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 and perform similar functions, yet they reside in the *rif* and *slm* pathways, respectively, and thus contribute to the biosynthesis of dramatically different secondary metabolites. Likewise, intra-cluster PKS gene duplication (Sare3151 and Sare3152; Figure 1) has an immediate effect on the product of the pathway by the introduction of an additional acyl group into the carbon skeleton of the macrolide as opposed to the more traditional concept of duplication facilitating mutation-driven functional divergence (Prince and Pickett, 2002). Subgenic, modular duplications are also observed (Sare3156 modules 4 and 5; Figure 1), which likewise have an immediate effect on the structure of the secondary metabolite produced by the pathway. Although HGT is considered a rapid method for ecological adaptation in bacteria (Ochman *et al.*, 2000), PKS gene duplication provides a complementary evolutionary strategy (Fischbach *et al.*, 2008) that could lead to the rapid production of new secondary metabolites that subsequently drive the creation of new adaptive radiations.

Salinispora species are the first marine Actinobacteria reported to require seawater for growth (Maldonado *et al.*, 2005). Unlike Gram-negative marine bacteria, in which seawater requirements are linked to a specific sodium ion requirement (Kogure, 1998), *Salinispora* strains are capable of growth in osmotically adjusted, sodium-free media (Tsueng and Lam, 2008). 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 that include homologs associated with polymorphic outer membrane proteins. Polymorphic outer membrane proteins remain functionally uncharacterized; however, there is strong evidence that they are type V secretory systems (Henderson and Lam, 2001), making this the first report of type V autotransporters outside of Proteobacteria (Henderson *et al.*, 2004). Phylogenetic analyses provide evidence that the *Salinispora* polymorphic membrane proteins were acquired from aquatic, Gram-negative bacteria and that they have continued to undergo considerable duplication subsequent to divergence of the two species (Supplementary Figure S3). The occurrence of this large family of polymorphic membrane protein autotransporters in marine Actinobacteria may represent a low-nutrient adaptation that renders cells susceptible to lysis in low osmotic environments.

Conclusion

In conclusion, the comparative analysis of two closely related marine actinobacterial genomes provides new insights into the functional traits associated with genomic islands. 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, and that the results of these processes create natural product diversity that can have an immediate effect on fitness or niche utilization. The high level of species specificity associated with secondary metabolism suggests that this functional trait may represent a previously unrecognized force driving ecological diversification among closely related, sediment-inhabiting bacteria.

References

Achtman M, Wagner M. (2008). Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* **6**: 431–440.

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551–554.

Ahlert J, Shepard E, Lomovskaya N, Zazopoulos E, Staffa A, Bachmann BO *et al.* (2002). The calicheamicin gene cluster and its iterative type I enediyne PKS. *Science* **297**: 1173–1176.

Badger JH, Eisen JA, Ward NL. (2005). Genomic analysis of *Hyphomonas neptunium* contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders '*Rhodobacterales*' and *Caulobacterales*. *Int J Syst Evol Microbiol* **55**: 1021–1026.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S *et al.* (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709–1712.

Berdy J. (2005). Bioactive microbial metabolites. A personal view. *J Antibiot* **58**: 1–26.

Bose M, Barber RD. (2006). Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* **6**: 223–227.

Buchanan GO, Williams PG, Feling RH, Kauffman CA, Jensen PR, Fenical W. (2005). Sporolides A and B: structurally unprecedented halogenated macrolides from the marine actinomycete *Salinispora tropica*. *Org Lett* **7**: 2731–2734.

Burrus V, Pavlovic G, Decaris B, GuÈdon G. (2002). Conjugative transposons: the tip of the iceberg. *Mol Microbiol* **46**: 601–610.

Cohan FM. (2002). What are bacterial species? *Annu Rev Microbiol* **56**: 457–487.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF *et al.* (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768–1770.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**: 414–424.

Dondoshansky I, Wolf Y. (2000). *BLASTCLUST*, 2.2 edn. National Institutes of Health: Bethesda, MD.

Feling RH, Buchanan GO, Mincer TJ, Kauffman CA, Jensen PR, Fenical W. (2003). Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *Salinospora. Angewandte Chemie* **115**: 369–371.

Fenical W, Jensen PR. (2006). Developing a new resource for drug discovery: marine actinomycete bacteria. *Nat Chem Biol* **2**: 666–673.

Fenical W, Jensen PR, Palladino MA, Lam KS, Lloyd GK, Potts BC. (2009). Discovery and development of the anticancer agent salinosporamide A (NPI-0052). *Bioorg Med Chem* **17**: 2175–2180.

Fischbach MA, Walsh CT, Clardy J. (2008). The evolution of gene collectives: how natural selection drives chemical innovation. *Proc Natl Acad Sci USA* **105**: 4601–4608.

Fouts DE. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* **34**: 5839–5851.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**: 741–746.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ *et al.* (2005). Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3**: 733–739.

Haeder S, Wirth R, Herz H, Spiteller D. (2009). Candicidin-producing *Streptomyces* support leaf-cutting ants to protect their fungus garden against the pathogenic fungus *Escovopsis. Proc Natl Acad Sci USA* **106**: 4742–4746.

Hao W, Golding GB. (2006). The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res* **16**: 636–643.

Henderson IR, Lam AC. (2001). Polymorphic proteins of *Chlamydia* spp.—autotransporters beyond the Proteobacteria. *Trends Microbiol* **9**: 573–578.

Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala'Aldeen D. (2004). Type V protein secretion pathways: the autotransporter story. *Microbiol Mol Biol Rev* **68**: 692–744.

Hsiao W, Wan I, Jones SJ, Brinkman FSL. (2003). IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* **19**: 418–420.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**: 1081–1085.

Jensen PR, Mafnas C. (2006). Biogeography of the marine actinomycete *Salinispora. Environ Microbiol* **8**: 1881–1888.

Jensen PR, Williams PG, Oh D-C, Zeigler L, Fenical W. (2007). Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora. Appl Environ Microbiol* **73**: 1146–1152.

Joyce EA, Chan K, Salama NR, Falkow S. (2002). Redefining bacterial populations: a post-genomic reformation. *Nat Rev Genet* **3**: 462–473.

Kim TK, Hewavitharana AK, Shaw PN, Fuerst JA. (2006). Discovery of a new source of rifamycin antibiotics in marine sponge actinobacteria by phylogenetic prediction. *Appl Environ Microbiol* **72**: 2118–2125.

Kogure K. (1998). Bioenergenics of marine bacteria. *Curr Opin Biotechnol* **9**: 278–282.

Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**: 2567–2572.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. (2001). REPuter: the manifold

applications of repeat analysis on a genomic scale. *Nucleic Acids Res* **29**: 4633–4642.

Maldonado LA, Fenical W, Jensen PR, Kauffman CA, Mincer TJ, Ward AC *et al.* (2005). *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. *Int J Syst Evol Microbiol* **55**: 1759–1766.

Mincer TJ, Jensen PR, Kauffman CA, Fenical W. (2002). Widespread and persistent populations of a major new marine actinomycete taxon in ocean sediments. *Appl Environ Microbiol* **68**: 5005–5011.

Ochman H, Lawrence JG, Groisman EA. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.

Ochman H, Lerat E, Daublin V. (2005). Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci USA* **102**: 6595–6599.

Parche S, Nothaft H, Kamionka A, Titgemeyer F. (2000). Sugar uptake and utilisation in *Streptomyces coelicolor*: a PTS view to the genome. *Antonie van Leeuwenhoek* **78**: 243–251.

Podell S, Gaasterland T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* **8**: R16.

Podell S, Gaasterland T, Allen EE. (2008). A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinformatics* **9**: 419.

Prince VE, Pickett FB. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* **3**: 827–837.

Reuther J, Gekeler C, Tiffert Y, Wohlleben W, Muth G. (2006). Unique conjugative mechanism in mycelial streptomycetes: a DNA-binding ATPase translocates unprocessed plasmid DNA at the hyphal tip. *Mol Microbiol* **61**: 436–446.

Schultz AW, Oh D-C, Carney JR, Williamson RT, Udwary DW, Jensen PR *et al.* (2008). Biosynthesis and structures of cyclomarins and cyclomarazines, prenylated cyclic peptides of marine actinobacterial origin. *J Am Chem Soc* **130**: 4507–4516.

Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. (2008). ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Res* **36**: 6882–6892.

te Poele E, Bolhuis H, Dijkhuizen L. (2008). Actinomycete integrative and conjugative elements. *Antonie van Leeuwenhoek* **94**: 127–143.

te Poele EM, Habets MN, Tan GYA, Ward AC, Goodfellow M, Bolhuis H *et al.* (2007). Prevalence and distribution of nucleotide sequences typical for pMEA-like accessory genetic elements in the genus *Amycolatopsis*. *FEMS Microbiol Ecol* **61**: 285–294.

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J *et al.* (2005). Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**: 1311–1313.

Tsueng G, Lam K. (2008). A low-sodium-salt formulation for the fermentation of salinosporamides by *Salinispora tropica* strain NPS21184. *Appl Microbiol Biotechnol* **78**: 821–826.

Tuanyok A, Leadem BR, Auerbach RK, Beckstrom-Sternberg SM, Beckstrom-Sternberg JS, Mayo M *et al.* (2008). Genomic islands from five strains of *Burkholderia pseudomallei*. *BMC Genomics* **9**: 14.

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W *et al.* (2007). Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc Natl Acad Sci USA* **104**: 10376–10381.

Vernikos GS, Parkhill J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics* **22**: 2196–2203.

Wall DP, Fraser HB, Hirsh AE. (2003). Detecting putative orthologs. *Bioinformatics* **19**: 1710–1711.

Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D *et al.* (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* **99**: 17020–17024.

Wu G, Culley DE, Zhang W. (2005). Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology* **151**: 2175–2187.

Yim G, Wang HMH, Davies J. (2007). Antibiotics as signalling molecules. *Philos Trans R Soc Lond B Biol Sci* **362**: 1195–1200.

Yu T-W, Shen Y, Doi-Katayama Y, Tang L, Park C, Moore BS *et al.* (1999). Direct evidence that rifamycin polyketide synthase assembles polyketide chains processively. *Proc Natl Acad Sci USA* **96**: 9051–9056.