The ISME Journal (2008) 2, 242–253

www.nature.com/ismej

ORIGINAL ARTICLE

Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter ruber*

Ramon Rosselló-Mora[1], Marianna Lucio[2], Arantxa Peña[3], Jocelyn Brito-Echeverría[1], Aránzazu López-López[1], Maria Valens-Vadell[1], Moritz Frommberger[2], Josefa Antón[3] and Philippe Schmitt-Kopplin[2]

[1]*Marine Microbiology Group, Institut Mediterrani d'Estudis Avançats, Esporles, Spain;* [2]*GSF—National Research Center for Environment and Health, Institute of Ecological Chemistry, Neuherberg, Germany and* [3]*División de Microbiología, Departamento de Fisiología, Genética y Microbiología and Instituto Multidisciplinar de Estudios del Medio, Ramon Margalef Universidad de Alicante, Alicante, Spain*

The biogeography of prokaryotes and the effect of geographical barriers as evolutionary constraints are currently subjected to great debate. Some clear-cut evidence for geographic isolation has been obtained by genetic methods but, in many cases, the markers used are too coarse to reveal subtle biogeographical trends. Contrary to eukaryotic microorganisms, phenotypic evidence for allopatric segregation in prokaryotes has never been found. Here we present, for the first time, a metabolomic approach based on ultrahigh resolution mass spectrometry to reveal phenotypic biogeographical discrimination. We demonstrate that strains of the cosmopolitan extremophilic bacterium *Salinibacter ruber*, isolated from different sites in the world, can be distinguished by means of characteristic metabolites, and that these differences can be correlated to their geographical isolation site distances. The approach allows distinct degrees of discrimination for isolates at different geographical scales. In all cases, the discriminative metabolite patterns were quantitative rather than qualitative, which may be an indication of geographically distinct transcriptional or posttranscriptional regulations.

The ISME Journal (2008) **2,** 242–253; doi:10.1038/ismej.2007.93; published online 31 January 2008
Subject Category: microbial population and community ecology
Keywords: biogeography; extremophilic bacteria; mass spectrometry; metabolomics; *Salinibacter ruber*

Introduction

Biogeography constitutes a cornerstone approach for studying biodiversity patterns at different taxonomic levels in the prokaryotic world (Ramette and Tiedje, 2006). In this regard, the importance of geographic barriers influencing microbial speciation is subjected to great debate (Whitaker *et al.*, 2003), and the old microbiological tenet of 'everything is everywhere, but, the environment selects' has served as a starting point for research on environmental forces that may lead to genetic and phenotypic allopatric segregation (de Wit and Bouvier, 2006). Most of the studies searching for the geographical isolation of prokaryotic populations and their divergence have been directed toward genetic characters (Whitaker *et al.*, 2003; Green and Bohannan, 2006; Huges-Martiny *et al.*, 2006; Ramette and Tiedje, 2006). The difficulties in retrieving phenotypic information, which is thought to be a stepping stone for taxonomic classifications (Staley, 2006), may hamper the discovery of divergences in prokaryotic populations at the phenotypic level. Clear biogeographical differences have been observed for some prokaryotes, such as thermophilic *Archaea* (Whitaker *et al.*, 2003) and *Cyanobacteria* (Papke *et al.*, 2003), and for fluorescent pseudomonads (Cho and Tiedje, 2000). However, the microbial counterpart of allopatric speciation in higher eukaryotes is still to be shown (Fenchel, 2003). Currently, only phenotypic differences matching biogeography have been reported for eukaryotic microorganisms (Fenchel and Finlay, 2006).

Prokaryote taxonomy, diversity and ecology have benefited from the developments of molecular

Correspondence: R Rosselló-Mora, Marine Microbiology Group, Recursos Naturals, Institut Mediterrani d'Estudis Avançats, IMEDEA (CSIC-UIB), C/Miquel Marqués 21, Esporles D-07190, Spain.
E-mail: rossello-mora@uib.es
Received 25 July 2007; revised 25 September 2007; accepted 26 September 2007; published online 31 January 2008

techniques. Ribosomal RNA–based approaches (Amann *et al.*, 1995), genomics, and currently, metagenomics (Green-Tringe and Rubin, 2005) are the major sources of information for understanding the diversity of the prokaryotic biome. These approaches give information of paramount importance, but only at the genetic level. However, analyzing the expression of the genotype may lead to a better understanding of the interactions microbes have with their environment. A microorganism is not only a composite of its genome, but also the multiple expressions of its genotype (Cavalier-Smith, 2007), and there is a significant part of the genome that might never be expressed (Ochman and Davalos, 2006). Approaches such as functional transcriptomics and proteomics may be considered as a dynamic link between the genome and the cellular phenotype (Singh and Nagaraj, 2006), the real interaction of the organism with its environment.

We recently showed that the extremely halophilic bacterium *Salinibacter ruber* (Antón *et al.*, 2000) can be isolated from different parts of the world in sites as diverse as Mediterranean coastal solar salterns (Peña *et al.*, 2005) and the remote Andean Peruvian salterns of Maras at 3380 m above sea level (Maturrano *et al.*, 2006). *S. ruber* growth is constrained to relatively small water bodies with high salt concentrations in restricted areas on the Earth. The extreme conditions and geographical isolation of its environments are optimal circumstances for observing allopatric speciation, as demonstrated for the hyperthermophilic archaeon *Sulfolobus* (Whitaker *et al.*, 2003), and thermophilic *Cyanobacteria* (Papke *et al.*, 2003). In this regard, initial genetic analyses based on fingerprinting genomic traits, such as pulsed field gel electrophoresis (PFGE) or randomly amplified polymorphic DNA (RAPD), although indicating a certain geographical differences, did not render a clear-cut geographical discrimination of *S. ruber* (Peña *et al.*, 2005).

In the present work we have undertaken a diversity study based on multilocus sequence analysis (MLSA) that had been successful for thermophilic *Archaea* (Whitaker *et al.*, 2003), and compared the results with a metabolomic approach based on high-resolution mass spectrometry. We used high-field ion cyclotron resonance Fourier transform mass spectrometry (12 T ICR-FT/MS) to evaluate the intraspecific diversity of *S. ruber* based on the differences of the chemical extracts from our strain collection.

ICR-FT mass spectrometry enables the assignment of thousands of elemental compositions of metabolites in a mass range from 120 to 800 kDa directly out of complex mixtures by virtue of ultrahigh mass accuracy (<100 parts per 10^9 (p.p.b.)) and ultrahigh resolution ($>1\,500\,000$ at mass 600) at high-field strength. This represents the initial, but crucial, step in metabolite annotation, for instance, by use of various targeted databases (that is, KEGG, Kyoto Encyclopedia of Genes and Genomes database). This technique is acquiring an increasingly important position in 'metabolomics' (Want *et al.*, 2007) together with spectroscopic methods, such as nuclear magnetic resonance spectroscopy (Nicholson *et al.*, 1999). However, high-field ICR-FT/MS (Marshall, 2004) showed to have the highest resolution among all spectrometric methods in revealing fine-scale diversity in complex mixtures. This method may help in revealing phenotypic patterns of geographically isolated organisms at the level of the direct interaction with the environment (phenotype) that may not be clearly indicated by indirect interaction (genotype).

Materials and methods

Strain isolation and culture conditions

Brine samples were directly plated onto 25% SW agar medium supplemented with 0.1% yeast extract (Antón *et al.*, 2002). Plates were incubated at 37 °C until growth was observed. Subsequent colonies were isolated in pure cultures, and those corresponding to *S. ruber* were studied further. Liquid 25% SW medium supplemented with 0.2% yeast extract was used to grow biomass in liquid conditions with vigorous shaking at 37 °C. For the metabolomic studies, all strains were inoculated and incubated for the same time under the same conditions. Biomass was harvested by centrifuga-

Table 1 List of *S. ruber* strains used in this study and their isolation origin

Strains	*Origin*	*Area considered*	*Year of isolation*
M8, M31	Mallorca, Llevant salterns	Mediterranean	2000
P13, P18	Alicante, Santa Pola salterns	Mediterranean	2000
E1, E3, E7, E12	Tarragona, Trinidad salterns	Mediterranean	2001
IL3	Ibiza, Ibiza salterns	Mediterranean	2001
ES4	Israel, Eilat salterns[a]	Mediterranean	2001
C3, C4Rj, C6, C9, C12, C14, C15, C17, C22, C25A, C26, C27, C29	Canary Islands, La Palma salterns	Atlantic	2001
PR1, PR3, PR2, PR6, PR8	Perú, Maras salterns	Peruvian	2003

[a]Eilat Salterns are located by the Red Sea, but we consider it as Mediterranean for proximity and climate similarities.

Table 2 List of genes used for the MLSA approach

Gene	*Str*	*DNA subst*	*Nsy*	*INDELS*
eno	685	11	9	—
pyrG	1645	53	25	—
tuf	961	21	3	—
glyA	1218	0	0	—
S5	486	13	1	—
rpsC	659	8	1	—
gap	903	10	8	—
Total homologous positions	6513	116	47	—
16S	1482	76 (44 without indels)	—	14 (44 nucl)
Total	7995	192 (148 without indels)	47	14 (44 nucl)

Str indicates the gene stretch sequenced. DNA subst indicates the total number of substitutions in each gene independently of whether they are silent or nonsynonymous. Nsy indicates the number of nonsynonymous substitutions after translation to amino acids. INDELS indicates the number of insertions/deletions observed among the sequences, and the number of nucleotides contained in the indels is in brackets. The newly sequenced housekeeping genes were deposited in the EMBL database with the accession numbers EF554259–EF554306, and EU003979–EU003986.

tion. Table 1 indicates the list of strains used in this study, their origin and year of isolation. Two growth batches were prepared to evaluate two simultaneous independent experiments: a complete set of strains from all different locations, and a second batch made up with 4–5 replicates of selected Mediterranean strains (that is, 13 and P18 from Alicante, M8 and M31 from Mallorca and IL3 from Ibiza).

DNA extraction, amplification and sequencing
DNA of the studied strains was isolated, amplified and sequenced following standard procedures (Antón *et al.*, 2002). The different genes selected for MLSA were amplified by the use of the primers and melting temperatures are shown in Supplementary Table S1. Primers were designed by searching for suitable DNA stretches in the complete sequence of *S. ruber* strain M31 (Mongodin *et al.*, 2005).

Phylogenetic reconstructions of concatenated genes
Phylogenetic reconstructions were based on multiple concatenated genes from different sequences (Table 2) and strain data sets. The analyses were performed by using the PHYML program package (Guindon and Gascuel, 2003) available from the http://atgc.lirmm.fr/phyml/ website, and they were compared with neighbor-joining and maximum-likelihood reconstructions, as implemented in the ARB program package (www.arb-home.de; Ludwig *et al.*, 2004). All analyses were performed by using the nucleotide sequence alignments, since their translation into amino acids rendered a very small number of informative positions. All alignments were calculated by using the HKY substitution model (Hasegawa *et al.*, 1985), and the proportion of invariable sites and the transition/transversion rates were estimated. The number of substitution rate categories was 4. Calculations were performed by using a BIONJ starting tree (Guindon and Gascuel, 2003). Topologies and branch lengths were optimized by the program. Multiple analyses were performed to find topology changes due to the effect of the gene composition of the alignments (Figure 1; Supplementary Figure S1), and to evaluate tree topology stabilities, as previously recommended (Ludwig and Klenk, 2001). Bootstrap values were obtained after the calculation of 100 replicates, as implemented in the PHYML program package.

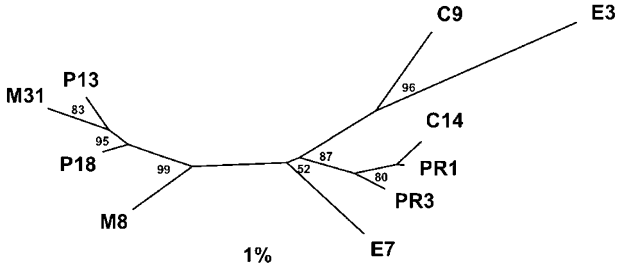

Figure 1 Phylogenetic reconstruction based on a PHYML algorithm of a 7995 nucleotide alignment corresponding to eight housekeeping genes, among which was SSU rRNA. The bar indicates 1% sequence divergence.

Metabolite extract preparation
A total of 3 ml of cell suspension grown on liquid media were collected by centrifugation. Cell-free supernatant (2 ml) was stored for further chromatographic extraction. Supernatant was acidified by the addition of 50 µl of 98–100% formic acid (Merck KGaA, Darmstadt, Germany). Pelleted biomass was then suspended in 1 ml of bidistilled water, and sonicated to obtain a clear lysate extract. The lysate was then acidified by the addition of 50 µl of 98–100% formic acid. After the acidification, the clear lysate formed insoluble aggregates that could be separated from the soluble fraction by centrifugation. The clear supernatant was stored for further fractionation, and the insoluble pellet was resuspended in 500 µl of methanol. Sample preparation resulted in three complementary fractions: the extracellular, cellular soluble and cellular insoluble fractions.

Solid-phase extraction
Both acidified extracellular and cellular soluble fractions were solid phase extracted using Bond Elut C18 columns (Varian Inc., Lake Forest, CA, USA). This chromatography enables the isolation of the organic molecules on the basis of their non-specific interaction and retention to the C18 material. This purification removes the high-salt charge of the media and extracts, which may interfere during the electrospray procedure by ion suppression (Li *et al.*, 2006). The retained fraction was recovered by the use of methanol.

ICR-FT/MS procedure
Broad band mass spectra were acquired on a Bruker (Bremen, Germany) APEX Qe ICR-FT/MS with 12 T superconducting magnet and an Apollo I electrospray (ESI) source, whereas high-resolution spectra were acquired with an Apollo II ESI source in positive and negative modes. The samples were infused in methanol with a microelectrospray source at a flow rate of 120 $\mu l\,h^{-1}$ with a nebulizer gas pressure of 20 p.s.i. and a drying gas pressure of 15 p.s.i. (200 °C). Spectra were externally calibrated on clusters of arginine (10 $mg\,l^{-1}$ in methanol), and calibration errors in the relevant mass ranges were always below 100 p.p.b., which is the prerequisite for an adequate elementary composition assignment. Relative standard deviation in the intensity values of the peaks was routinely lower than 5% in our analysis conditions. The spectra were acquired with a time domain of 1 megaword (where 1 data word corresponds to 32 bits) with a mass range of 150–2000 *m/z*. The spectra were zero filled to a processing size of 2 megawords. A sine apodization was performed before Fourier transformation of the time-domain transient. The ion accumulation time in the ion source was set to 0.2 s and 1024 scans were accumulated for one spectrum.

ICR-FT/MS spectra were exported to peak lists at a signal-to-noise ratio (S/N) = 1. From these lists, possible elemental formulas were calculated for each peak in batch mode by a software tool written in-house. The generated formulas were validated by setting sensible chemical constraints (nitrogen rule, atomic oxygen to carbon ratio O/C $\leqslant$ 1, element counts: hydrogen H $\leqslant$ (2+C2), carbon C $\leqslant$ 100, oxygen O $\leqslant$ 80, nitrogen N $\leqslant$ 5 and sulfur S $\leqslant$ 1) and only the masses in conjunction with their automated generated theoretical isotope pattern (existence of the ^{13}C isotope) were taken into consideration (Hertkorn *et al.*, 2007). The obtained reduced peak lists were compared in *m/z* at 5 p.p.m. and the corresponding intensity matrices were generated for further statistical analysis.

Targeted approach
The targeted approach allowed a detailed analysis of specific metabolites following a specific chemical structures hypothesis after the previous metabolomic screening of the samples. The analysis, especially in high-resolution mode, enables a detailed description of the natural isotopic abundance that in addition allows confirmation of the elementary composition assignments. Figure 2 shows the assignment of the elementary compositions (including isotopic peaks) to the *m/z* as obtained in negative electrospray ICR-FT/MS in two different resolution modes. Mass intensity data related to sulfonolipids were analyzed statistically by one-way analysis of variance (ANOVA) with *post hoc* Bonferroni test for multiple comparisons (Holm, 1979). Probabilities less than 5% ($P<0.05$) were considered statistically significant.

Statistical analysis
The data were imported and analyzed in SIMCA-P 11.5 (Umetrics, Umea, Sweden). The statistical model used was partial least squares for discriminant analysis (PLS-DA) (Sjöström *et al.*, 1986; Stahle and Wold, 1987; Vong *et al.*, 1988; Kemsley, 1996). PLS-DA is a regression extension of the principal component analysis (Wold *et al.*, 1987). It takes advantage of class information (in this case the geographical origin of the samples) to maximize the separation between groups of masses. A list of masses (*m/z*) discriminative for the different geographical area is produced. The PLS-DA uses the *X* variables (matrix of masses) as predictors, and dummy variables (belonging or not belonging to a given class coded as 1/0; that is, origin of isolation) as response variables (*Y* variables). All three modalities (extracellular, cellular soluble and cellular insoluble) were calculated independently and cellular insoluble was chosen as the descriptive power of the model. The descriptive power can be defined by several terms, most directly the fraction of the sum of squares (SS) of all the *Y* explained by the current component (R^2Y(cum)) and Q^2(cum). R^2Y provides an estimate of how well the model fits the *Y* data and Q^2 provides an estimate of how well the model predicts the *Y* data (more detailed information in the Supplementary Materials and methods).

Pareto scaling of the intensity values with a logarithmic transformation of the data was chosen to consider all masses equally, including those with medium- and low-intensity values (Van den Berg *et al.*, 2006). The cellular insoluble metabolome data set contained 2099 variables (Table 3), from 28 observations measured in the three groups (Atlantic, Mediterranean and Peruvian). When analyzing this data set with PLS-DA using four significant components, R^2Y(cum) was equal to 0.98 and Q^2(cum) was equal to 0.45 both, with values indicating high predictive power.

The score scatter plot and loading plots are presented in Figures 3a and b, respectively. The score scatter plot (Figure 3a) presents a view of how well the classes (different geographical origin) are

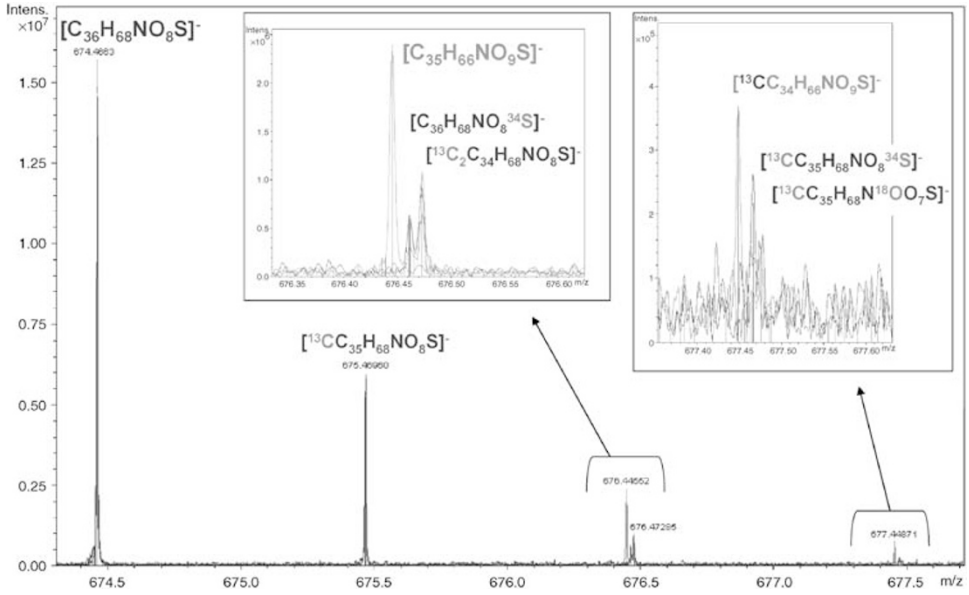

Figure 2 Detailed spectra on mass 674.4663 identified as a sulfonolipid in negative-mode electrospray ICR-FT/MS within the series described in Table 5 and its corresponding natural abundance isotopic pattern. Mass 676.4455 was found only in the Atlantic samples.

Table 3 Number of observed masses from the analysis, considered masses for statistics and masses for geographical discrimination (positive electrospray analysis)

m/z	*Insoluble cellular fraction*	*Soluble cellular fraction*	*Extracellular fraction*	*Sum of all masses*	*Cumulative unique masses*
Number of unique masses from raw data (S/N = 1)	168 444	157 378	161 322	487 144	247 655
Number of unique masses after CHONS calculation	3456	5293	5062	13 811	11 880
Number of masses used for statistical analysis ($m/z < 550$)	2099	3559	3450	9108	8873
Number of masses discriminative for Atlantic strains	181	74	80	335	333
Number of masses discriminative for Mediterranean strains	510	655	114	1279	1249
Number of masses discriminative for Peruvian strains	287	427	257	971	968
Number of masses from discriminative metabolome	1121	2403	2999	6523	6323

separated on the basis of their *X* variables. In the loading plot (Figure 3b), the different masses characteristic for each of the three classes are differently colored (red for Atlantic, green for Mediterranean and blue for Peruvian).

The variables (single masses) discriminative for each class (origin of isolation) were chosen according to their correlation coefficient value. Those having the highest coefficients were considered to be relevant (that is, variables (m/z) with a correlation value higher than |0.002|). A total number of 180 out of 2099 masses were considered to be discriminative for the classes (values shown in Table 3). Interpretation of the regression coefficients provides information pertaining to the metabolic explanation of class differences (Holmes and Antti, 2002) based on the fact that each coefficient is related to a specific elemental composition. Those masses associated with the highest correlation coefficient were represented in the van Krevelen projection (H/C versus O/C on the basis of their elementary composition values; Figures 3c and d).

Table 4 lists the coordinate values along the first and second components that numerically represent the similarities and differences among the strains. These values represent the distances resulting from the projection of the points on the first and second components to the origin (0 value). They explain the magnitude (large or small correlation) and the nature (positive or negative correlation) of the samples. The scatter plot score of Supplementary Figure S2 summarizes the numerical coordinates

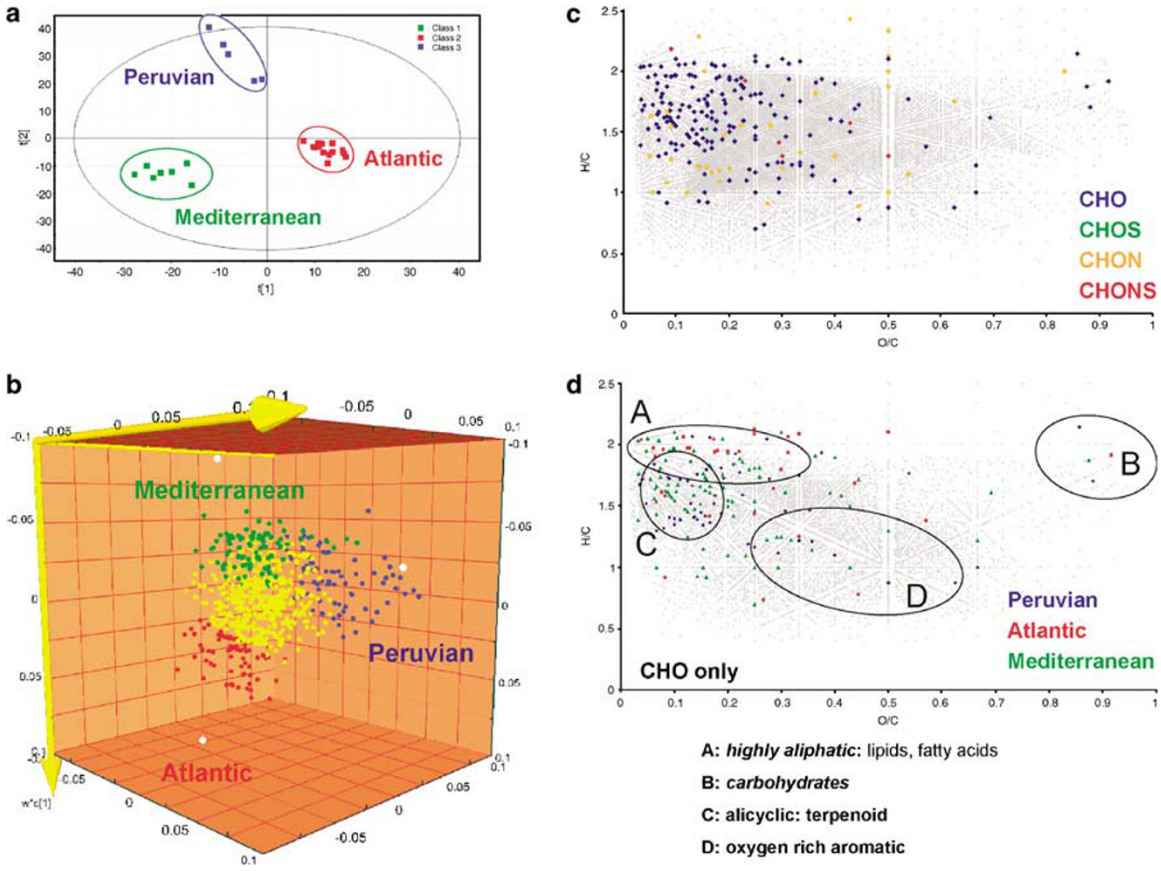

Figure 3 (**a**) Score plot of the PLS-DA of all cellular insoluble fractions analyzed with electrospray-positive mode ICR-FT/MS showing the differentiation based on geographical origin of the samples; (**b**) loading plot of the PLS-DA model correlating the 2099 *m*/*z* values of known elementary composition (C, H, O, N, S and $m/z < 550$) to the geographical origins. The *m*/*z* values having a high correlation with geographical origin are highlighted with a corresponding color; the nondiscriminating masses are represented in yellow; (**c**) all discriminating *m*/*z* values independent of their origin but colored only as a function of their attributed elementary composition (CHO, CHON, CHOS or CHONS) and visualized in a van Krevelen diagram (H/C versus O/C). Most of the discriminative metabolites contain only C, H and O (only a few metabolites contain sulfur or nitrogen) and these are compared within a van Krevelen type of diagram to the CHONS-containing metabolites of general metabolome databases (www.metabolome.jp, www.genome.jp/kegg/) shown in gray in the figure. Note that the triangular region corresponds to peptides (CHON and CHONS); (**d**) CHO metabolites in a van Krevelen diagram colored as a function of their origin.

present in Table 4 to provide a numerical perception of the group distance.

The similarity within strains M8, M31, P13, P18 and IL3 (intensities for each *m*/*z* value) was evaluated first by using the Levene's test (Malins *et al.*, 2002), which evaluates the differences (*P*-values) in the variances of each group of repetitions. Then, we used a one-way ANOVA, and a Tukey's test for repeated measures. All differences were considered to be significant when $P < 0.01$. The analyses were performed in SAS version 9.1 (SAS Institute Inc., Cary, NC, USA). At the $P < 0.01$ level, the population variations were not significantly different (Supplementary Table S2).

The discriminative analysis of the Mediterranean strains shown in Figure 4 was undertaken with orthogonal PLS-DA (OPLS-DA) based on the cellular soluble fraction. For this kind of sample, OPLS-DA rendered equivalent but clearer results than PLS-DA. In this case, OPLS-DA separates predictive from nonpredictive (orthogonal) variations (Bylesjö *et al.*, 2006; more information can be seen in the Supplementary Material).

Parsimony reconstruction of binary matrices

Raw matrices containing all variable characters among all strains studied, and coded as absence/presence of each peak, were reduced to an informative set by identifying all identical metabolites with different isotopic composition, and by reducing the background noise by the use of peak thresholds as described in the 'Materials and methods' section. Improved binary matrices were analyzed by the use of the parsimony tool in the Phylip program package (Felsenstein, 1981) using the default parameters (http://evolution.genetics.washington.edu/phylip.html).

Clustering analysis of binary matrices

Phenetic analyses were carried out by the use of the TREECON program version 1.3b (Van de Peer and De Wachter, 1994), and by using UPGMA.

Table 4 Coordinate value of the first and second components of the PLS-DA analysis (score)

Sites	*Strains*	*Coordinate value along the first component (score value)*	*Sites*	*Strains*	*Coordinate value along the second component (score value)*
Atlantic	C27	9.81995	Peruvian	PR3	28.1714
Atlantic	C3	9.65824	Peruvian	PR8	22.4758
Atlantic	C29	9.63297	Peruvian	PR2	19.0242
Atlantic	C22	9.25211	Peruvian	PR1	13.2149
Atlantic	C12	8.53592	Peruvian	PR6	12.1288
Atlantic	C25A	8.11802	Atlantic	C14	−0.254274
Atlantic	C9	7.54702	Atlantic	C17	−1.17895
Atlantic	C4_Rj	7.52264	Atlantic	C12	−1.28982
Atlantic	C15	7.46169	Atlantic	C26	−1.49031
Atlantic	C6	7.03174	Atlantic	C9	−1.53197
Atlantic	C17	5.94576	Atlantic	C6	−1.87301
Atlantic	C26	5.91143	Atlantic	C4Rj	−1.96649
Atlantic	C14	4.65194	Atlantic	C15	−2.42006
Peruvian	PR6	0.602852	Atlantic	C25A	−2.58963
Peruvian	PR1	−1.11698	Atlantic	C29	−3.75884
Peruvian	PR2	−4.17265	Atlantic	C22	−4.04727
Peruvian	PR8	−5.25783	Atlantic	C27	−4.47822
Peruvian	PR3	−7.62429	Atlantic	C3	−4.78183
Mediterranean	E1	−9.13346	Mediterranean	E1	−4.57788
Mediterranean	E7	−11.6219	Mediterranean	M31	−6.4909
Mediterranean	M31	−12.8236	Mediterranean	IL3	−6.56391
Mediterranean	P18	−14.623	Mediterranean	P18	−8.09294
Mediterranean	E12	−15.1815	Mediterranean	E12	−8.61029
Mediterranean	IL3	−16.8664	Mediterranean	ES4	−8.76525
Mediterranean	ES4	−20.3118	Mediterranean	E7	−9.96454

Strains close to each other have similar properties, common metabolites, whereas those far from each other are dissimilar with respect to the origin. From the inspection of the second component values, one factor that might contribute to the differentiation may be related to the geographical location of the origin of the strains. Peruvian is far away from Mediterranean but closer to Atlantic, and Peruvian is in fact negatively correlated with Mediterranean.

Results and discussion

Biogeographical comparisons at the genetic level

Representative members of *Salinibacter* spp. have been reported in several locations in the world either by molecular techniques (Antón *et al.*, 2000; Mesbah *et al.*, 2007) or by culturing approaches (Antón *et al.*, 2002; Maturrano *et al.*, 2006). These members of the Bacteria domain that do not show growth below 15% NaCl concentrations thrive in constrained environments that appear dotted on the earth's surface. An initial study based on our strain collection of about 17 strains isolated from several locations in Spanish coastal salterns indicated that a slight trend for geographic isolation could be discerned at the genetic level (Peña *et al.*, 2005). Contrary to observations for fluorescent pseudomonads (Cho and Tiedje, 2000), internal transcribed spacer (ITS) sequences were not suitable for studying biogeographical segregation due to their high sequence similarity. However, both PFGE and RAPD gave weak indications of geographical discrimination of genotypes. In no case were the analyses conclusive in proving allopatric segregation.

In this study, we enlarged the collection with about 28 strains isolated from five different locations in the world (Table 1). The isolates were obtained from five different locations in the Mediterranean area (Mallorca, Alicante, Tarragona, Ibiza and Israel), the Atlantic Canary archipelago (from a solar saltern on the island of La Palma) and from the 3500 m high salterns in the Peruvian Andes (Maras). Ten of the isolates were selected to undertake MLSA, which represented the three main geographical areas in the study (Mediterranean, Atlantic and Peruvian Andes). For all 10 strains, eight housekeeping genes that had been previously shown to be phylogenetically informative (Sória-Carrasco *et al.*, 2007) were sequenced (Table 2). The concatenated DNA stretch rendered an alignment of 7995 homologous sites, 6513 of them corresponding to seven protein gene sequences, with 129 of them being informative. Phylogenetic analyses were performed by including and excluding indels, as well as by using different data sets (including the 16S rRNA gene in the concatenate, Figure 1; or disregarding it, Supplementary Figure S1, respectively). In general, the trees agreed with regard to their topology, since only M8 acquired a stable position when including the 16S rRNA gene sequence in the analysis. Altogether, the reconstructions did not show clear geographical segregation of the selected strains, in contrast to observations made with other extremophiles (Whitaker *et al.*, 2003). Strains from Alicante (P13 and P18) affiliated together with that of Mallorca (M8 and M31). However, the Mediterranean strains E3 and E7 affiliated together with those from the Atlantic

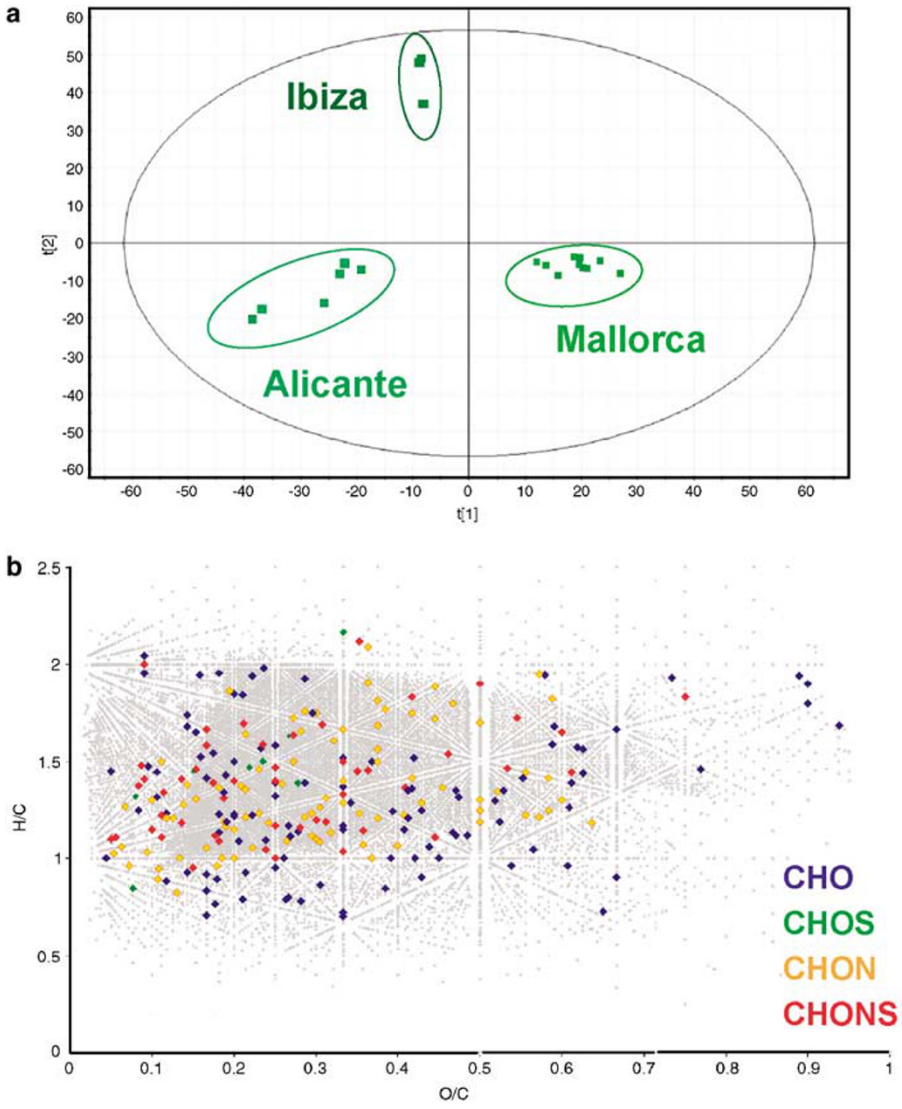

Figure 4 (**a**) Orthogonal partial OPLS-DA score plots of all cellular soluble fractions in the extracts of the Mediterranean strains from the locations of Alicante (P13 and P18), Ibiza (IL3) and Mallorca (M8 and M31). These metabolomes can be differentiated in a two component model with a high level of prediction ($R^2(Y) = 0.99$ $Q^2(\text{cum}) = 0.47$). (**b**) Representation of the discriminative masses in a van Krevelen diagram showing the high importance of nitrogen-containing metabolites representative of general metabolic pathways (confirmed by searching in public metabolite databases).

(C9 and C14) and Peru (PR1 and PR3). Neither our previous studies with fingerprinting techniques (Peña *et al.*, 2005), nor here with an MLSA of gene stretches of nearly 8000 homologous positions were informative enough to resolve biogeographical segregation. This could be an indication that the process of genetic divergence is still at an early stage and cannot render clearly resolvable trends. However, given that the size of the *S. ruber* genome is about 3000 open reading frames (Mongodin *et al.*, 2005), and despite the fact that we selected the genes to be sequenced from a set of putative phylogenetic markers (Sória-Carrasco *et al.*, 2007), the set of genes may not be adequate for understanding subtle geographical segregation. Intraspecific whole-genome comparisons with *S. ruber* might in future indicate which genes could be useful for understanding allopatric differentiation based on genetic drift.

Biogeographical comparisons at the phenotypic level

As stated above, genomic data are especially useful for solving the main problems in the classification of organisms, as well as understanding speciation processes (Staley, 2006; Ward *et al.*, 2007). In most of the fields related to prokaryote diversity (taxonomy, ecology, speciation), phenotype studies are being relegated in favor of those based on genome information, such as MLSA or other genome analyses, due to the ease of the latter. However, standard genotyping techniques may not always help in clearly resolving intraspecific diversity. As has already been requested (Ramette and Tiedje,

2006), there is a need to apply new approaches for understanding allopatric segregation of members of the same species. For this reason, we have evaluated the adequacy of a nontargeted metabolite profiling approach, using high-field ICR-FT/MS of the chemical extracts of our strain collection. Mass spectrometry has acquired a predominant position in metabolomics (Want *et al.*, 2007) and, especially, high-field ICR-FT/MS (Marshall, 2004). This technique provides ultrahigh resolved profiles with thousands of accurate mass values (m/z) that can be transformed into real elementary compositions.

For this study, a first experiment with 28 isolates of *S. ruber* from seven locations in the world (Table 1), divided into three geographical areas (Mediterranean (10 strains), Atlantic (13 strains) and Peruvian (5 strains)), were studied by ICR-FT/MS. All organisms were grown simultaneously under identical environmental conditions to avoid culture-dependent differences. Metabolome comparisons rendered a total of over 247 255 discriminative mass signals at $S/N = 1$ that could be attributed to distinct elementary compositions containing the elements C, H, O, N and S. Single-peak occurrence was reduced from 11 880 (verified by isotopic assignments of elementary composition) to a total of 8873 metabolites at an m/z lower than 550 a.m.u. (highest probable assignments). The core metabolome (that is, common peaks for all extracts) consisted of 2550 single masses, whereas the discriminative metabolome (that is, peaks not common to all extracts) consisted of 6323 single metabolites (Table 3). In all cases, the analyses were performed by using the whole metabolome.

With the raw information, the first comparative analyses were based on qualitative data coded as presence or absence of single metabolites. For this, the results were expressed in a binary matrix that was treated either cladistically using parsimony, or phenetically, using UPGMA (Supplementary Figure S3). However, in no case could the profile analysis, based on independent covariant characters, reveal clear geographical trends. Therefore, it seemed that the presence or absence of single metabolite comparisons did not reflect geographical isolation.

In contrast, weighing the relative intensity of each individual peak and treating the data by using multivariate analysis revealed statistically significant differences between the different samples. Thus, metabolome comparisons, focusing on geographically discriminative data, yielded clear-cut allopatric differences. The nontargeted analysis revealed unique features for each group of isolates (Figures 3a and b). In this regard, the most relevant markers were CHO molecules (Figure 3c). Furthermore, van Krevelen plots of atomic O/C versus H/C ratios (Wu *et al.*, 2004) and comparisons with the total metabolic spaces (Figure 3d; Supplementary Table S3) showed that the discriminative metabolites may be aliphatic in structure and depleted in oxygen. Thus, those components generally associated with cell membranes, such as fatty acids and terpenoids, could be responsible for the geographic differentiation. Between the two principal components that led to geographical discrimination, we could also find, for the second component, a relationship for geographical distance between the isolation sites (Table 4 ; Supplementary Figure S2). It seemed that for the second principal component the Atlantic strains might show intermediate differences with those of the Mediterranean and Peruvian strains.

When specifically directing the recognition of discriminative metabolites among geographically distinct metabolomes, we observed that a set of conspicuous compounds could be unambiguously assigned to a sulfonolipid family (Table 5; Figure 2). The members of this compound family have been observed to be major components of the cell envelope of *Cytophaga* (Godchaux and Leadbetter, 1984), a member of the same phylum as *S. ruber* (Antón *et al.*, 2002). One of these components ($C_{35}H_{67}NO_8S$, $m/z = 660.4505$) has been reported to be characteristic of *S. ruber* (Corcelli *et al.*, 2004).

Table 5 Proposed elemental compositions of various masses assigned to sulfonolipids with their structural variations from $C_{35}H_{67}NO_8S$, originally described by Corcelli *et al.* (2004) as $C_{35}H_{66}NO_8S$, where *n* indicates the number of strains

Exp. m/z	*Proposed composition as [M-H]⁻ Δm/z < 0.6 p.p.m.*	*Structural variation from $C_{35}H_{66}NO_8S$*	*Mediterranean (n = 6) Averaged intensity (× 10⁶)*	*Atlantic (n = 11) Averaged intensity (× 10⁶)*	*Peruvian (n = 5) Averaged intensity (× 10⁶)*
644.4195	$C_{34}H_{62}NO_8S$	$-CH_2$ & $-H_2$	1.27	1.14	1.37
646.4351	$C_{34}H_{64}NO_8S$	$-CH_2$	6.40	5.12	5.88
660.4505	$C_{35}H_{66}NO_8S$	(−)	67.20	76.60	**92.24**
672.4505	$C_{36}H_{66}NO_8S$	+C	6.42	5.88	7.10
674.4662	$C_{36}H_{68}NO_8S$	$+CH_2$	8.04	6.98	8.88
676.4454	$C_{35}H_{66}NO_9S$	+O	ND	**1.50**	ND
684.4508	$C_{37}H_{66}NO_8S$	+2C	0.30	0.24	0.26
686.4663	$C_{37}H_{68}NO_8S$	$+C_2H_2$	1.44	1.25	1.42
688.4455	$C_{36}H_{66}NO_9S$	+C & +O	ND	**0.40**	ND
688.4819	$C_{37}H_{70}NO_8S$	$+2CH_2$	0.95	0.87	0.97

Abbreviation: ND, not defined.
In bold are indicated those isomer intensities that provide discrimination for the given location.

These compounds, which could account for 10% of total cellular lipids, have been proposed as signatures for *S. ruber* identification. The ICR-FT/MS approach, with a mass precision lower than 600 p.p.b., revealed that *S. ruber* may contain at least nine additional sulfonolipids analogous to $C_{35}H_{67}NO_8S$ in the mass range 644–688. These components differ from the originally described sulfonolipid in their elementary composition, with variations in their side chain length, insaturation or hydroxylation degree with variations in CH_2, H_2 and O, respectively, as described in Table 5. All these components were found in all of the analyzed samples with identical intensity ratios between isolates from the same location, except for *m*/*z* 676.4454 and *m*/*z* 688.4455. Both of these compounds ($C_{35}H_{68}NO_9S$ and $C_{36}H_{68}NO_9S$, respectively) seemed to be exclusive to the Atlantic strains.

The metabolomic approach allowed the targeted search for special metabolic traits considered to be relevant in the organisms' phenotype. Previous biochemical studies on *S. ruber*-type strain M31 revealed the presence of an active, hitherto unreported, rhodopsin type of membrane proton translocation system, the xanthorhodopsin, responsible for the putative phototrophy of *S. ruber* (Balashov *et al.*, 2005). In addition, the genome sequence of the same organism revealed the coding of one halorhodopsin (Peña *et al.*, 2005) and two sensory rhodopsin homologous genes (Mongodin *et al.*, 2005). Searching for an indication of the presence of retinal, the chromophore bound to rhodopsins, an experimental positive mass 285.22125 (theoretical 285.22129) was present in all samples. However, the *m*/*z* value was only discriminative for the Mediterranean strains.

An independent 'fine-tuning' experiment was undertaken by growing four replicates of five Mediterranean strains (P13 and P18 from Alicante, M8 and M31 from Mallorca and IL3 from Ibiza). Metabolome comparisons validated the replicates by first applying a Levene's test (Malins *et al.*, 2002) to evaluate differences in the variance, and after applying one-way ANOVA and the Tukey's test to evaluate the differences in the means of each replicate group. Nevertheless, the results between both latter tests were equivalent. At the $P<0.01$ level, population variations were not significantly different (see Supplementary Table S2 for the ANOVA results). Therefore, the differences observed between different strains could be attributed to strain-specific metabolisms rather than sample-to-sample variations. In contrast to previous results (Peña *et al.*, 2005), when searching for discriminative phenotypes at a more reduced geographical scale, we observed a phenotypic segregation in individual locations (Figure 4a), using the ICR-FT/MS approach. The main discriminative metabolomic profile features were different from those giving resolution at a larger geographical scale. In such cases, geographical differences were associated to strain-specific compositions of N-containing molecules (Figure 4b). The confrontation of their exact masses with the KEGG and the Japanese metabolome database (metabolome.jp) indicated that the discriminative molecules were involved primarily in the core metabolism (that is, carbohydrate, amino acid and fatty acid biosynthesis and metabolism).

Our findings reveal that intraspecific metabolic diversity of *S. ruber* can be readily detected by the ICR-FT/MS approach and that such diversity can be associated with different geographical patterns at different metabolic levels. In principle, the standard genetic methods used to assess biogeography (Whitaker *et al.*, 2003; Ramette and Tiedje, 2006) do not have the resolving power needed for a fine geographic discrimination of our model organism. The MLSA approach, based on different gene data sets, does not resolve putative genetic-geographic patterns, as the genetic divergence may be too subtle for the given selection of genes. However, one must take into account that, despite the fact that large sets of concatenated genes tend to reflect the organismal phylogeny (Sória-Carrasco *et al.*, 2007), perhaps only full genome sequences may reflect geographical isolation in the strain collection of *S. ruber*, in accordance with taxa segregation that correlates with the average nucleotide or amino-acid identity of shared genes (Konstantinidis and Tiedje, 2005). However, the backlogs in the current state of full genome sequencing make the metabolomic approach a fast and less expensive alternative for revealing prokaryotic biogeography, with the added value of being discriminative at different levels at the geographical scale.

It seems clear that different locations led to the isolation of strains sharing common metabolic traits, such as, the distinct production of sulfonolipid derivates. However, differences were generally related to quantitative composition yields, rather than qualitative production of distinct compounds. In addition, the metabolic differences correlated with the geographical locations, influenced perhaps by environmental conditions such as climate and distance, since in the second component Peruvian and Mediterranean strains were shown to be the most different. The discriminative metabolites were mainly aliphatic structures related to terpenoids or fatty acids, which might be membrane components, and these differences could be related to different environmental conditions (Sajbidor, 1997). Altogether, the results seem to indicate that the differences found could be attributed to transcriptional or posttranscriptional regulations rather than composition changes in genes at the genomic level. The major forces for these differences between strains should be related to their distinct response to the environmental conditions of the sites where they had been isolated, since, for example, the Peruvian salterns are not only over 10 000 km away from the rest of our sampling sites, but they are also at an altitude of 3500 m. At this site the temperature

changes and solar radiation are clearly different from those at sea level. ICR-FT/MS was shown to have a higher resolution in revealing fine-scale diversity. This method has a great potential for revealing biogeographical patterns in many other nonextremophilic microorganisms.

Acknowledgements

Funding for this study was provided by the Ministerio de Educación y Ciencia through the research projects CLG2006-12714-C02-01 and -02. The technical help provided by Cristina López and Mercedes Urdiain is warmly acknowledged. We thank Norbert Hertkorn and Karsten Suhre from GSF, and Silvia Marqués, Antonio Tovar, Rudolf Amann and Edward Moore for their helpful discussions and encouragements. New gene sequences were deposited in the EMBL database with the accession numbers EF554259 to EF554306, and EU003979 to EU003986.

References

Amann RI, Ludwig W, Schleifer KH. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59**: 143–169.

Antón J, Oren A, Benlloch S, Rodríguez-Valera F, Amann R, Rosselló-Mora R. (2002). *Salinibacter ruber* gen. nov., sp. nov., a novel, extremely halophilic member of the Bacteria from saltern crystallizer ponds. *Int J Syst Evol Microbiol* **52**: 485–491.

Antón J, Rosselló-Mora R, Rodríguez-Valera F, Amann R. (2000). Extremely halophilic bacteria in crystallizer ponds from solar salterns. *Appl Environ Microbiol* **66**: 3052–3057.

Balashov S, Imasheva ES, Boichenko VA, Antón J, Wang JM, Lanyi JK. (2005). Xanthorhodopsin: a proton pump with a light-harvesting antenna. *Science* **309**: 2061.

Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. (2006). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemometrics* **20**: 341–351.

Cavalier-Smith T. (2007). Concept of a bacterium still valid in prokaryote debate. *Nature* **446**: 257.

Cho JC, Tiedje JM. (2000). Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. *Appl Environ Microbiol* **66**: 5448–5456.

Corcelli A, Lattanzio VM, Mascolo G, Babudri F, Oren A, Kates M. (2004). Novel sulfonolipid in the extremely halophilic bacterium *Salinibacter ruber*. *Appl Environ Microbiol* **70**: 6678–6685.

de Wit R, Bouvier T. (2006). 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environ Microbiol* **8**: 755–758.

Felsenstein J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**: 368–376.

Fenchel T. (2003). Biogeography for bacteria. *Science* **301**: 925–926.

Fenchel T, Finlay BJ. (2006). The diversity of microbes: resurgence of the phenotype. *Philos Trans R Soc Lond B Biol Sci* **361**: 1965–1973.

Godchaux III W, Leadbetter ER. (1984). Sulfonolipids of gliding bacteria. Structure of the *N*-acylaminosulfonates. *J Biol Chem* **259**: 2982–2990.

Green J, Bohannan BJ. (2006). Spatial scaling of microbial biodiversity. *Trends Ecol Evol* **21**: 501–517.

Green-Tringe S, Rubin WM. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* **6**: 805–814.

Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Hasegawa M, Kishino H, Yano T. (1985). Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**: 160–174.

Hertkorn N, Meringer M, Gugisch R, Ruecker C, Frommberger M, Perdue EM *et al.* (2007). High-precision frequency measurements: indispensable tools at the core of molecular-level analysis of complex systems. *Anal Bioanal Chem* (doi:10.1007/s00216-007-1577-4).

Holm S. (1979). A simple sequentially rejective multiple test procedure. *Scand J Stat* **6**: 65–70.

Holmes E, Antti H. (2002). Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. *Analyst* **127**: 1549–1557.

Huges-Martiny J, Bohannan B, Brown J, Colwell R, Fuhrman J, Green J *et al.* (2006). Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102–112.

Kemsley EK. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemom Intell Lab Syst* **33**: 47–61.

Konstantinidis K, Tiedje JM. (2005). Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**: 6258–6264.

Li X, Fekete A, Englmann M, Götz Ch, Rothballer M, Frommberger M *et al.* (2006). Development and application of a method for the analysis of *N*-acylhomoserine lactones by solid-phase extraction and ultra high pressure liquid chromatography. *J Chromatogr A* **1134**: 184–193.

Ludwig W, Klenk HP. (2001). Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone D, Castenholz R, Garrity G (eds), *Bergey's Manual of Systematic Bacteriology*, 2nd edn. Springer: New York, pp 49–65.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Malins DC, Hellstrom KE, Anderson KM, Johnson PM, Vinson MA. (2002). Antioxidant-induced changes in oxidized DNA. *Proc Natl Acad Sci USA* **99**: 5937–5941.

Marshall AG. (2004). Accurate mass measurement: taking full advantage of nature's isotopic complexity. *Physica B* **346–347**: 503–508.

Maturrano L, Santos F, Rosselló-Mora R, Antón J. (2006). Microbial diversity in Maras salterns, a hypersaline environment in the Peruvian Andes. *Appl Environ Microbiol* **72**: 3887–3895.

Mesbah NM, Abou-El-Ela SH, Wiegel J. (2007). Novel and unexpected prokaryotic diversity in water and sediments of the alkaline, hypersaline lakes of the Wadi An Natrun, Egypt. *Microb Ecol* (doi:10.1007/s00248-006-9193-y).

Mongodin EF, Nelson KE, Daugherty S, Deboy RT, Wister J, Khouri H *et al.* (2005). The genome of *Salinibacter ruber*: convergence and gene exchange among hyper-

halophilic bacteria and archaea. *Proc Natl Acad Sci* **102**: 18147–18152.
Nicholson JK, Lindon JC, Holmes E. (1999). Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29**: 1181–1189.
Ochman H, Davalos LM. (2006). The nature and dynamics of bacterial genomes. *Science* **311**: 1730–1733.
Papke RT, Ramsing NB, Bateson MM, Ward DM. (2003). Geographical isolation in hot spring cyanobacteria. *Environ Microbiol* **5**: 650–659.
Peña A, Valens M, Santos F, Buczolits S, Antón J, Kämpfer P *et al.* (2005). Intraspecific comparative analysis of the species *Salinibacter ruber*. *Extremophiles* **9**: 151–161.
Ramette A, Tiedje JM. (2006). Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microb Ecol* **53**: 197–211.
Sajbidor J. (1997). Effect of some environmental factors on the content and composition of microbial membrane lipids. *Crit Rev Biotechnol* **17**: 87–103.
Singh OV, Nagaraj NS. (2006). Transcriptomics, proteomics and interactomics: unique approaches to track the insights of bioremediation. *Brief Funct Genomic Proteomic* **4**: 355–362.
Sjöström M, Wold S, Söderström B. (1986). PLS discriminant plots. In: Gelsema ES, Kanal LN (eds). *Pattern Recognition in Practice II*. Elsevier: Amsterdam, pp 486.
Sória-Carrasco V, Valens-Vadell M, Peña A, Antón J, Amann R, Castresana J *et al.* (2007). Phylogenetic position of *Salinibacter ruber* based on concatenated protein alignments. *Syst Appl Microbiol* **30**: 171–179.
Stahle L, Wold S. (1987). Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J Chemometrics* **1**: 185–196.
Staley J. (2006). The bacterial species dilemma and the genomic-phylogenetic species concept. *Phil Trans R Soc B* **361**: 1899.
Van de Peer Y, De Wachter R. (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci* **10**: 569–570.
Van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, Van der Werf MJ. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**: 142–157.
Vong R, Geladi P, Wold S, Esbensen K. (1988). Source contributions to ambient aerosol calculated by discriminant partial least squares regression (PLS). *J Chemometrics* **2**: 281–286.
Want EJ, Nordstro A, Morita H, Siuzdak G. (2007). From exogenous to endogenous: the inevitable imprint of mass spectrometry in metabolomics. *J Proteome Res* **6**: 459–468.
Ward D, Cohan F, Bhaya D, Heildelberg J, Kühl M, Grossman A. (2007). Genomics, environmental genomics and the issue of microbial species. *Heredity*, 1–13 (doi:10.1038/sj.hdy.6801011).
Whitaker RJ, Grogan DW, Taylor JW. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* **301**: 976–978.
Wold S, Esbensen K, Geladi P. (1987). Principal component analysis. *Chemom Intell Lab Syst* **2**: 37–52.
Wu Z, Rodgers RP, Marshall AG. (2004). Two- and three-dimensional van Krevelen diagrams: a graphical analysis complementary to the Kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband Fourier-transform ion cyclotron resonance mass measurements. *Anal Chem* **76**: 2511–2516.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)