BMC Genomics

BioMed Central

Research article

Open Access

Low-pass sequencing for microbial comparative genomics

Young Ah Goo[1], Jared Roach[1], Gustavo Glusman[1], Nitin S Baliga[1], Kerry Deutsch[1], Min Pan[1], Sean Kennedy[2], Shiladitya DasSarma[2], Wailap Victor Ng[3] and Leroy Hood*[1]

Address: [1]Institute for Systems Biology, Seattle, Washington 98103, USA, [2]Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, Maryland 21202, USA and [3]Department of Biotechnology and Laboratory Science in Medicine, National Yang Ming University, Taipei, Taiwan

Email: Young Ah Goo - ygoo@systemsbiology.org; Jared Roach - jroach@systemsbiology.org; Gustavo Glusman - cgojp@systemsbiology.org; Nitin S Baliga - nbaliga@systemsbiology.org; Kerry Deutsch - kdeutsch@systemsbiology.org; Min Pan - mpan@systemsbiology.org; Sean Kennedy - kennedy@umbi.umd.edu; Shiladitya DasSarma - dassarma@umbi.umd.edu; Wailap Victor Ng - wvng@ym.edu.tw; Leroy Hood* - lhood@systemsbiology.org

* Corresponding author

Published: 12 January 2004

BMC Genomics 2004, **5**:3

Received: 02 July 2003
Accepted: 12 January 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/3

Abstract

Background: We studied four extremely halophilic archaea by low-pass shotgun sequencing: (1) the metabolically versatile *Haloarcula marismortui*; (2) the non-pigmented *Natrialba asiatica*; (3) the psychrophile *Halorubrum lacusprofundi* and (4) the Dead Sea isolate *Halobaculum gomorrense*. Approximately one thousand single pass genomic sequences per genome were obtained. The data were analyzed by comparative genomic analyses using the completed *Halobacterium sp.* NRC-1 genome as a reference. Low-pass shotgun sequencing is a simple, inexpensive, and rapid approach that can readily be performed on any cultured microbe.

Results: As expected, the four archaeal halophiles analyzed exhibit both bacterial and eukaryotic characteristics as well as uniquely archaeal traits. All five halophiles exhibit greater than sixty percent GC content and low isoelectric points (pI) for their predicted proteins. Multiple insertion sequence (IS) elements, often involved in genome rearrangements, were identified in *H. lacusprofundi* and *H. marismortui*. The core biological functions that govern cellular and genetic mechanisms of *H. sp.* NRC-1 appear to be conserved in these four other halophiles. Multiple TATA box binding protein (TBP) and transcription factor IIB (TFB) homologs were identified from most of the four shotgunned halophiles. The reconstructed molecular tree of all five halophiles shows a large divergence between these species, but with the closest relationship being between *H. sp.* NRC-1 and *H. lacusprofundi*.

Conclusion: Despite the diverse habitats of these species, all five halophiles share (1) high GC content and (2) low protein isoelectric points, which are characteristics associated with environmental exposure to UV radiation and hypersalinity, respectively. Identification of multiple IS elements in the genome of *H. lacusprofundi* and *H. marismortui* suggest that genome structure and dynamic genome reorganization might be similar to that previously observed in the IS-element rich genome of *H. sp.* NRC-1. Identification of multiple TBP and TFB homologs in these four halophiles are consistent with the hypothesis that different types of complex transcriptional regulation may occur through multiple TBP-TFB combinations in response to rapidly changing environmental conditions. Low-pass shotgun sequence analyses of genomes permit extensive and diverse analyses, and should be generally useful for comparative microbial genomics.

Background

Extremely halophilic archaea inhabit hypersaline environments containing three to five molar salts. These environments are widely distributed, and include solar saltern facilities, the Dead Sea coast, and brine inclusions. The environments are diverse with respect to salinity, pH, temperature, pressure, light, and oxygen. This wide range of conditions contributes to the diversity of halophile barophilic, alkaliphilic, and psychrophilic characteristics in addition to their obligate halophilicity. Extreme halophiles are classified as members of the *Halobacteriaceae* family and further organized into nineteen genera [1]. Halophiles are a relatively poorly studied branch of the archaea with only one complete genome sequence, that of *Halobacterium sp.* NRC-1 [2].

The genome of NRC-1 consists of a large chromosome of 2,014,239 bp and two additional minichromosomal replicons of 365,425 bp (pNRC200) and 191,346 bp (pNRC100). The two minichromosomes are relatively less GC rich than the largest chromosome (57.9% and 59.2% vs. 67.9%). Averaged across all three chromosomes, *H. sp.* NRC-1 has an overall 65.8% GC content. Identification of ninety-one insertion sequence (IS) elements representing twelve different IS families is consistent with the dynamic genome rearrangements observed, mediated by the multiple IS elements [3]. Genome sequence analysis identified 2,682 putative protein-coding genes. Among these, counting each identical multigene family as one gene, 2,413 genes are unique. Analysis of the genome sequence and the predicted proteome reveals that *H. sp.* NRC-1 is more similar to other archaeal species than to non-archaeals. *H. sp.* NRC-1 possesses similarities to bacteria in genes coding for aerobic respiration and in overall genomic structure, and to eukaryotes in genes coding for DNA replication, transcription, and translation [2,4], *H. sp.* NRC-1 predicted proteins are extremely acidic; protein acidity is associated with enhanced solubility and activity in hypersaline cytoplasm [4].

Nineteen genera of extreme halophiles have been reported: *Haloarcula, Halobacterium, Halobaculum, Halococcus, Haloferax, Halogeometricum, Halorhabdus, Halorubrum, Halosimplex, Haloterrigena, Natrialba, Natrinema, Natronobacterium, Natronococcu, Natronomonas, Natronorubrum, Haloalcalophilium, Halobiforma, and Halomicrobium.* Most halophilic species possess distinctive phenotypic features such as gas vesicles, purple membranes, or red-orange carotenoids. Many have the ability to grow phototropically in the absence of oxygen, using light energy transduced by bacteriorhodopsin (BR), halorhodopsin (HR), and ATP synthetase.

Some recently isolated species exhibit additional interesting features. *Halorubrum lacusprofundi*, isolated from Deep Lake, Antarctica, is psychrophilic (cold loving) and grows at temperatures as low as four degrees Celsius [5]. Some halobacterial species are alkaliphilic. Some are acid-tolerant. *Natronobacterium pharaonis* from Wadi Natrun, Egypt and *Natronococcus occultus* from Lake Magadi, Africa have pH optima ranging from 9.5 to 10 and do not grow below pH 8.5. Slight acidophiles, such as *Haloferax volcanii* and *Haloferax mediterranei*, grow at pH values as low as 4.5. Unlike other *Halobacterium* species, *Natrialba asiatica* and *Natrialba magadii* are not pigmented; they contain less than 0.1% of bacterioruberins, which give most halophiles their red-orange color [6].

Membership in a large and diverse family, adaptation to unique environments, ease of culture, and the availability of genetic tools make the halophiles one of the best model systems to study microbial diversity and adaptation. To advance this model system, we previously determined the first complete halobacterial genome sequence, that of *H. sp.* NRC-1 [2], In this present study, we analyzed the halophile family more broadly through a comparative genomic approach based on low-pass sequencing. We scanned the genome sequences of four additional species: (1) *Haloarcula marismortui*, a metabolically versatile halophile [7], (2) *Natrialba* asiatica, a non-pigmented alkaliphile [8], (3) *Halorubrum lacusprofundi*, a psychrophile [5], and (4) *Halobaculum gomorrense*, an isolate from a depth of 4 meters in the Dead Sea [9]. These organisms were studied by random shotgun sequence analysis. This analysis consisted of generating approximately one thousand genomic sequence reads, which provided partial genomic scans and permitted comparative genomic analyses with the completely sequenced *H. sp.* NRC-1 genome as a reference.

Low pass genomic sequencing is twenty-fold less expensive than complete genome sequencing, yet provides data for useful comparative genomic analyses. It was a natural choice for the study of the intriguing *Halobacteriaceae* family.

Results

Validation of low-pass sequencing

The genome of *H. sp.* NRC-1 is very compact (approximately 87% coding region) and the sequences of most random genomic fragments contain genes [2], To evaluate the utility of low-pass sequencing for the comparative analysis of *Halobacteriaceae* genomes, we hypothesized: 1) the partial genome sequences generated by shotgun sequencing were sufficient to identify the genes contained in most fragments and 2) a large number of different unique genes would be identified from ~1,000 sequence reads.

In order to test our hypotheses, 1,085 sequences of 450 bp average length were extracted from the actual *H. sp.* NRC-1 genome [10], with each sequence having a uniform probability of starting at any base of any of the three chromosomes, which is a reasonable approximation of the sonication process by which fragments are generated for shotgun sequencing. These 1,085 sequences represent approximately 0.2-fold coverage of the genome. The sequences were then used to search the protein non-redundant (nr) database on the NCBI server http://www.ncbi.nlm.nih.gov/BLAST with BLAST, using default parameters. Using expert curation of the graphically visualized alignments, a total of 1,094 genes were identified from the 1,057 sequences. Some reads spanned more than one gene. Twenty-eight of the 1,085 sequence reads did not have database gene matches; they were from the intergenic regions of the *H. sp.* NRC-1 genome. Several of the genes were overlapped by more than one random sequence; of the 1,094 genes identified, 820 were unique.

The validation study supports our hypothesis that many genes can be successfully identified using partial genome sequences, and also validates the gene identification methods which were further applied to analyze database search results with the actual partial genome sequence reads of the four halophiles, described in the next section.

The non-redundant database contained *H. sp.* NRC-1 proteins at the time this search was done. Therefore, we ran the risk of overestimating the number of genes that we might identify in the four halophiles to be shotgunned. However, excluding *H. sp.* NRC-1 proteins would create an underestimate, as there were few other sequences from *Halobacteriaceae* in Genbank. For example, only 329 TREMBL genes are identified with BLASTX using the BLOSUM80 matrix in our initial 1,085 random sequences when all *Halobacteriaceae* sequences are excluded (in this case, counting no more than one gene per sequence, and approximating by equating a gene with any hit with an e-value less than or equal to 0.01). We assayed variability of this statistic by repeating it on an independently generated set of another 1,085 random sequences; 304 TREMBL genes were identified. Including the *Halobacteriaceae* TREMBL proteins, 881 and 892 genes were identified respectively, with 95% of the additional identifications due to alignments with *H. sp.* NRC-1 proteins. For our experimental planning purposes, we felt that retaining the *H. sp.* NRC-1 proteins would result in less error than excluding them. The resulting approximation was sufficient for us to proceed with actual sequencing. However, our results would not necessarily be precisely predictive of gene prediction power in other organisms. Such predictive power would depend on the number of closely related sequences in Genbank.

Overview of sequencing

Sequencing was conducted with the M13 forward primer on shotgun libraries prepared using sonicated genomic DNA from each of the four halophiles. The traces were processed with PHREDPHRAP http://www.phrap.org. This procedure produced 1,097 high-quality sequences for *H. marismortui*, 1,085 for *H. lacusprofundi*, 1,104 for *N. asiatica*, and 1,170 for *H. gomorrense*. The initial 1,085 random *H. sp.* NRC-1 sequences described above in the validation study were used in conjunction with the newly sequenced halophile sequences in most of the subsequent analyses described here. The number of high-quality bases per read ranged from 40 to 750 for all genomes. The average lengths of the high-quality sequences were 524,441,469, and 415 bases, respectively (Figure 1). These represent approximately 15%, 18%, 17%, and 18% coverage of each genome, respectively.

Gene prediction

The DNA sequences generated by the shotgun method are usually not long enough to contain complete genes including start and stop codons. For this study, "best match" open reading frames were determined as described in Methods. When multiple sequence reads with different reading frames matched to the same gene, the match results of the sequence reads were compared by the percent identity, length of the alignments, and *p* value. The best match was then selected and the reading frame of the sequence was considered the "best match" ORF representing that gene.

Of the 1,097 *H. marismortui* sequences, 810 sequences (74%) had matches to the nr protein database. Of these, 703 were unique. *H. gomorrense* had a total of 1,170 sequences, of which 1,006 sequences (86%) had matches to the nr protein database. Of these, 742 were unique. Of the 1,085 *H. lacusprofundi* sequences, 802 sequences (74%) matched to the nr protein database. Of these, 623 were unique. Of the 1,104 *N. asiatica* sequences, 862 (78%) were matched to the nr protein database. Of these, 678 were unique. The length range of partial ORFs was 30 to 256 residues with average lengths of 121,130,137, and 130 aa, respectively (Figure 2). The non-archaeal best matches were distributed between bacterial and eukaryotic sequences. Thus, like *H. sp.* NRC-1 and other archaea, these four halophiles have both bacterial and eukaryotic features [11,12].

Sequences with no database matches were not further analyzed. It is possible these sequences are new/unique genes to each halophile and therefore no match would be found in databases. Alternatively, these sequences may be non-coding, intergenic sequences. About 36% of predicted ORFs in *H. sp.* NRC-1 genome did not have any significant match to the nr protein database.

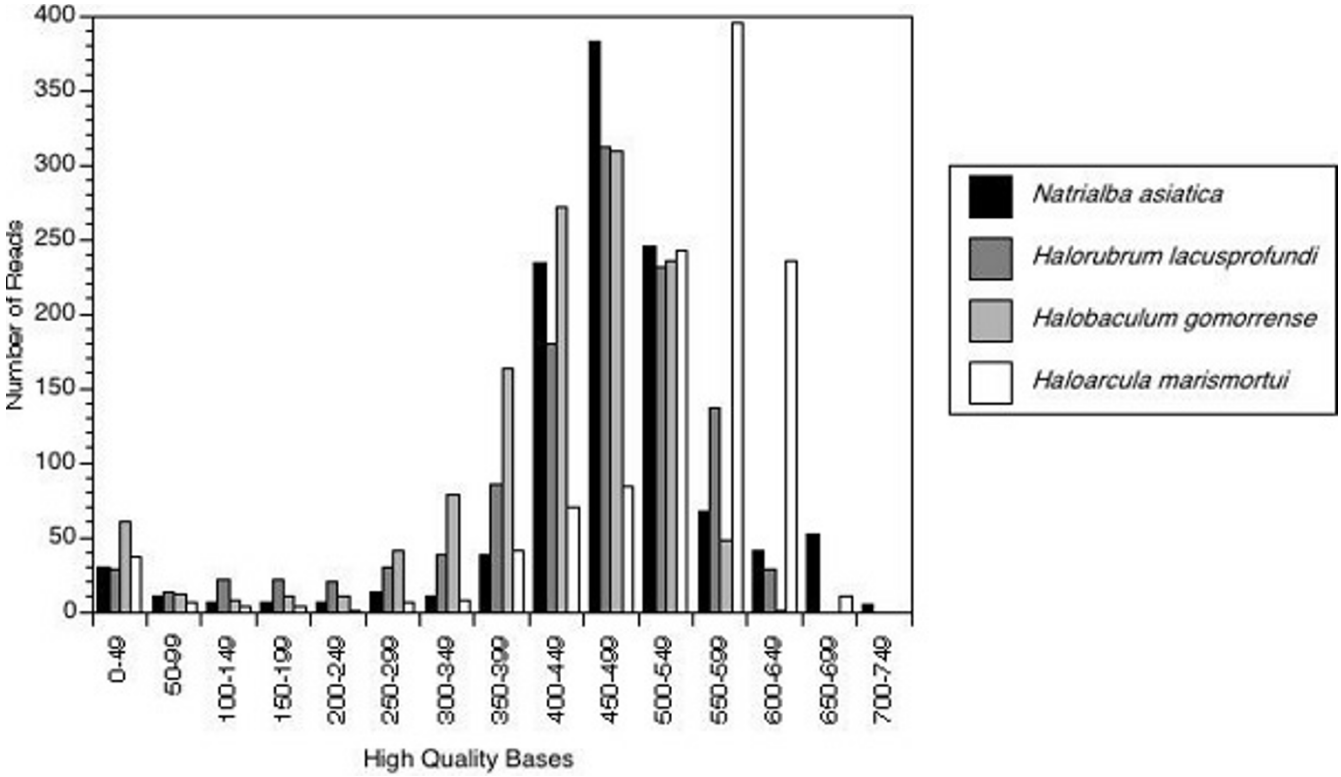

Figure 1
Distribution of lengths of low-pass sequence reads for the four archaeal genomes. High-quality sequences were obtained by processing the shotgun sequences with PHREDPHRAP.

Of the 1,085 randomly selected *H. sp.* NRC-1 sequences, 1,057 (97.4%) had matches to the nr protein database, while 28 (2.6%) had no matches to the database. Further analysis revealed that all of the 28 sequences were from intergenic regions, consistent with no significant matches with the nr protein database. The 1,057 sequences with database matches identified 820 unique proteins.

Insertion sequence elements

IS elements are repetitive sequences that can broker genome reorganization and mediate evolution by creating insertion mutations, recombinations, and gene conversions. In *Halobacterium* these elements have expanded to multiple families and subfamilies with multiple copy numbers [13]. Of the ninety-one IS elements representing twelve families in the *H. sp.* NRC-1 genome, sixty-nine were found in the minichromosomes including twenty-nine IS elements on pNRC100 and forty on pNRC200. In the *H. sp.* NRC-1, due to genetic instability caused by transposition of insertion sequences, typically about 1% of colonies in a plate are gas-vesicle deficient and 0.1% are carotenoid deficient mutants [13].

In order to determine whether the presence of multiple IS elements is a general property of halophiles, sequences of each halophile were searched for IS elements. *H. sp.* NRC-1 IS elements were searched against each halophile sequence to identify homologous IS elements. Halophile sequences were also searched against the nr protein database to identify any new IS elements. In the *H. lacusprofundi* genome, we identified 38 sequences homologous to six different IS families of NRC-1. Ten *H. marismortui* genome sequences that were homologous to six different IS families of *H. sp.* NRC-1 were also identified, while only one sequence each from *H. gomorrense* and *N. asiatica* was found. These matched to ISH4 and ISH9 of NRC-1, respectively (Table 1). All IS elements identified were homologous to one of the *H. sp.* NRC-1 IS elements, and no new IS elements were identified. In this analysis using partial genome sequences, multiple sequence reads often matched to a single IS family. This may represent multiple copies of the repeat or multiple sequence reads of the same repeat. Considering that the sequences were all generated from a random shotgun library and sequenced to relatively low redundancy, it is unlikely that several sequences would be derived from the same location. The results indicate that the *H. lacusprofundi* genome contains the highest number of IS elements.

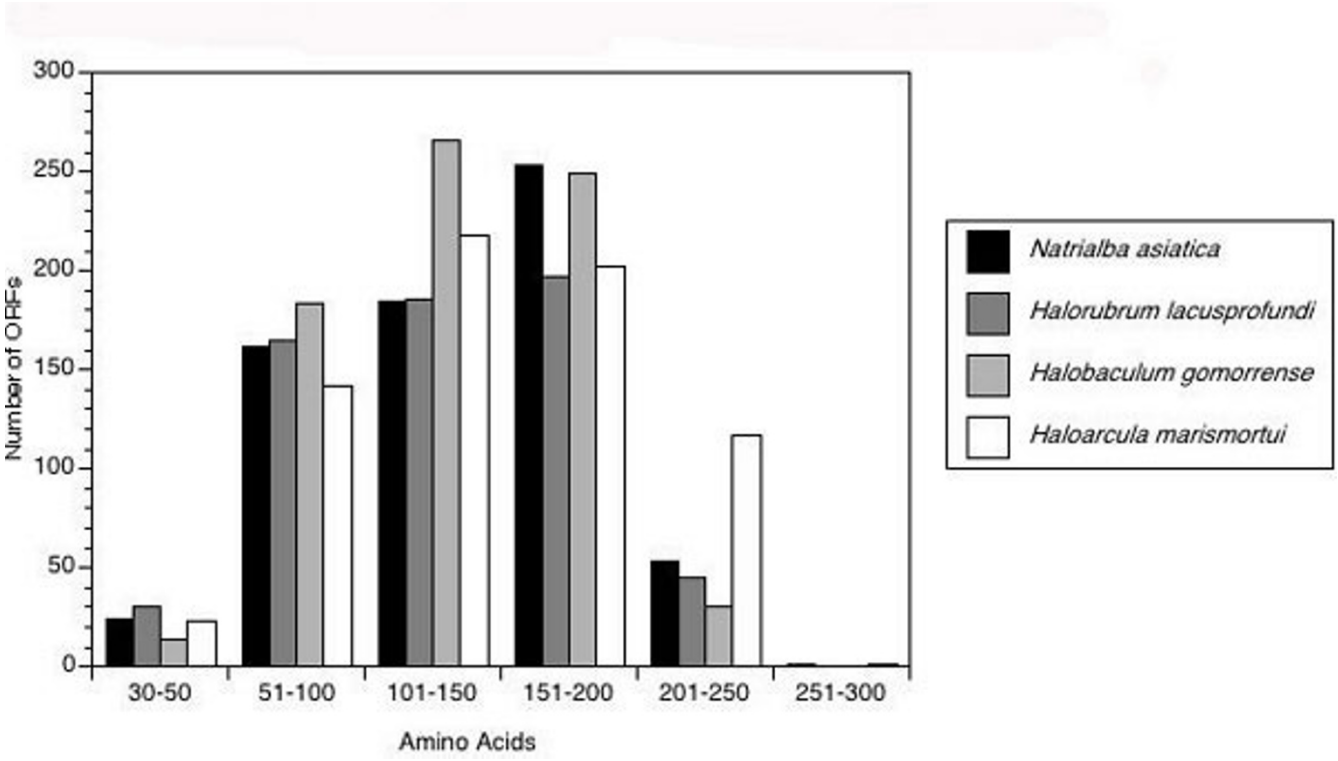

Figure 2
Distribution of lengths of partial ORFs derived from low-pass sequences.

Table 1: Insertion sequence (IS) elements in five halophiles. For the reference genome of *Halobacterium sp.* NRC-1, the copy numbers of the IS are listed, while for the partial genome sequences of other halophiles, the sequence numbers that matched the IS elements are shown. More IS elements were identified from *H. lacusprofundi* and *H. marismortui* than from *H. gomorrense* and *N. asiatica.*

ISH elements	NRC-1 No. of copies	rNRC-1 No. of sequences	*H. lacusprofundi* No. of sequences	*H. gomorrense* No. of sequence	*H. marismortui* No. of sequence	*N. asiatica* No. of sequences
ISH1	1	1	-	-	-	-
ISH2	13	1	-	-	-	-
ISH3	23	11	13	-	-	-
ISH4	2	3	-	1	1	-
ISH5	6	4	-	-	-	-
ISH6	2	1	6	-	2	-
ISH7	4	3	-	-	-	-
ISH8	21	18	2	-	2	-
ISH9	4	-	6	-	2	1
ISH10	6	3	-	-	1	-
ISH11	7	6	8	-	2	-
ISH12	2	2	3	-	-	-

GC composition

Previous studies of several *Halobacterium* and *Halococcus* species found that their genomes contain both relatively GC-rich (66–68%) components and relatively GC-poor (57–60%) components [14]. The genome analysis of *H. sp.* NRC-1 revealed that the plasmids pNRC100 and pNRC200 contain lower GC content than the main chromosome (57.9% and 59.2% vs. 67.9%). To evaluate

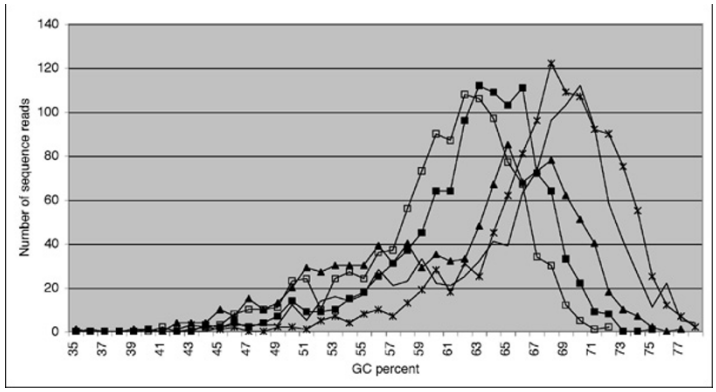

Figure 3
GC distribution for five partially sequenced genomes. GC distribution of *Halobacterium sp.* NRC-1 (unmarked line), *Halobaculum gomorrense* (asterisks), *Halorubrum lacusprofundi* (filled triangles), *Haloarcula marismortui* (squares), and *Natrialba asiatica* (filled squares) are indicated. Average GC contents for *H. sp.* NRC-1, *H. gomorrense, H. lacusprofundi, H. marismortui*, and *N. asiatica* were 66%, 65.1%, 61.4%, 60%, and 61.2%, respectively.

whether the presence of DNA components varying in their GC content is a general genome property of the halophiles, sequence reads from each of the halophiles were analyzed for their GC composition (Figure 3). GC analysis of random NRC-1 sequences shows two small peaks at about 56% and 59%, which likely reflect the DNA component of pNRC100 and pNRC200, and a large peak at about 70%, likely reflecting the DNA component of the chromosome. In *H. sp.* NRC-1, of the sequence reads in the peak of 56% GC content, 62% were from pNRC100; of the sequence reads in the peak of 59% GC content, 73% were from pNRC200. Similarly, lower GC peaks and higher GC peaks were observed from *H. lacusprofundi*, *H. gomorrense*, and *H. marismortui*. *H. lacusprofundi* showed lower GC peaks at about 56%–58% and a higher GC peak at about 65–68%. *H. gomorrense* had lower GC peaks at about 60%–62% and a higher GC peak at about 68%. *H. marismortui* showed lower GC peaks at about 51%–54% and a higher GC peak at about 62%. *N. asiatica* showed peaks at about 62–66%.

In the case of *H. sp.* NRC-1, the presence of three chromosomes (two of them mini) of differing GC contents is known. The presence of such plasmids in the four shotgunned halophiles has not been determined yet, as they are not, if present, resolvable with routine pulsed-field gel electrophoresis. It is not clear if the observed two or three peaks of differing GC contents in the shotgunned halophiles are due to the presence of plasmids, or to regions of varying GC contents within one replicon.

The genome analysis of *H. sp.* NRC-1 revealed an overall GC composition of 65.9%. The environments where halophiles are found are often areas exposed to extreme levels of UV solar radiation. Active DNA repair mechanisms for damage caused by UV irradiation, such as the formation of thymine dimers, have been previously observed in halophiles [4]. The four newly sequenced halophiles have 60% or higher overall GC content. The GC content of *Halobacteriaceae* thus may have evolved to minimize DNA damage caused by solar UV irradiation.

Isoelectric point of predicted proteins

Halophiles maintain osmotic balance between their cytoplasms and the hypersaline environment by maintaining ionic distributions of intracellular potassium of about four molar, equal to the external sodium concentration [2,15]. One of the most interesting properties of halophiles is the ability of their proteins to function in such a hypersaline cytoplasm, where mesophilic proteins would become denatured. Early findings showed that the surfaces of halophilic proteins are highly acidic, with negatively charged amino acid residues [16]. This was shown to enable proteins to stay soluble and active through the binding of hydrated salt ions in the solution [16-18]. Accordingly, recent investigations of predicted proteins from *H. sp.* NRC-1 demonstrate extremely acidic proteins with an overall median isoelectric point (pI) of 4.9 and a strong peak (mode) around 4.2 [4].

For our novel shotgunned sequences, the "best match" ORFs were used to estimate pI of their corresponding proteins. This approach was first validated using partial ORFs of 820 proteins from random *H. sp.* NRC-1 sequences. The predicted median pI of 4.4, calculated from partial ORFs, was somewhat lower to the median pI of 4.9 calculated

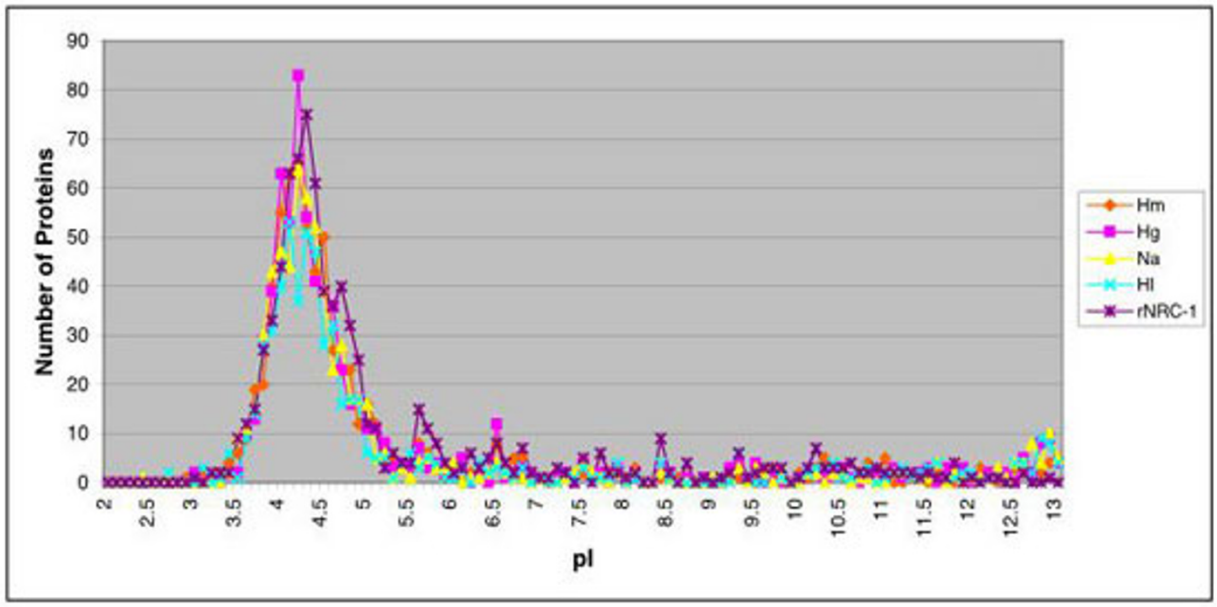

Figure 4
Calculated isoelectric point (pI) versus number of predicted proteins. Predicted isoelectric points from four halophiles and random *H. sp.* NRC-1 proteins were plotted [*H. gomorrense* (Hg), *H. lacusprofundi* (Hl), *H. marismortui* (Hm), *N. asiatica* (Na), and random *H. sp.* NRC-1 (rNRC-1)]. All halophiles show the highest peak at pI between ~4.1 and ~4.3 with median pI range from 4.3 to 4.4.

from the complete set of *H. sp.* NRC-1 proteins. However, both results had a distinct peak (mode) at pI 4.2. Therefore, we estimated pI from 703 *H. marismortui*, 623 *H. lacusprofundi*, 678 *N. asiatica*, and 742 *H. gomorrense* partial ORFs. All halophiles, including *H. sp.* NRC-1 recomputed from our random set of sequences, had their major pI peak between 4.1 and 4.3. The median pIs of *H. marismortui, H. lacusprofundi, N. asiatica, H. gomorrense*, and randomly recalculated *H. sp.* NRC-1 were 4.4, 4.4, 4.3, 4.3, and 4.4, respectively (Figure 4).

The prevalence of highly acidic proteins across the four newly sequenced halophiles is consistent with those halophiles employing the acidic protein stabilizing mechanisms for hypersaline environments. Highly acidic proteins are thus a general feature of extreme halophiles, providing the basis for enhanced solubility in the presence of high salts.

Conserved core biological functions

Of the unique genes identified through their "best match" ORFs in the newly sequenced halophiles, 56% or more were homologous to *H. sp.* NRC-1 genes. This finding suggests conservation of genes and their functions between *H. sp.* NRC-1 and each halophile. In order to identify conserved core biological functions between *H. sp.* NRC-1 and the newly sequenced halophiles, genes homologous to *H. sp.* NRC-1 were classified into twelve functional groups: amino acid metabolism, cell envelope components, cellular processes, cofactor metabolism, DNA replication repair and recombination, energy metabolism, nucleotide metabolism, regulation, transcription, translation, transport, and miscellaneous genes. These are the same classifications described for the *H. sp.* NRC-1 genome project [2]. Also, an identical comparison to a subset of only 820 random *H. sp.* NRC-1 genes was made to enable quantitative comparison of these results with those of the other interspecies results. Core biological functions that are essential for *H. sp.* NRC-1 to function also seem to be conserved in the four halophiles (Figure 5). Presence of genes for the citric acid cycle for uptake and utilization of amino acids, sodium-proton antiporters, potassium uptake systems, DNA replication, transcription, and translation systems resembling eukaryotes were observed in all species. Some of the aminoacyl-tRNA synthetases that were not found in some archaea (e.g. *Methanococcus jannaschii* [19]) but present in *H. sp.* NRC-1 were identified in *H. marismortui, H. gomorrense*, and *N. asiatica.* Identification of multiple proteins involved in DNA repair pathways suggests that halophiles utilize multiple pathways to repair DNA damage.

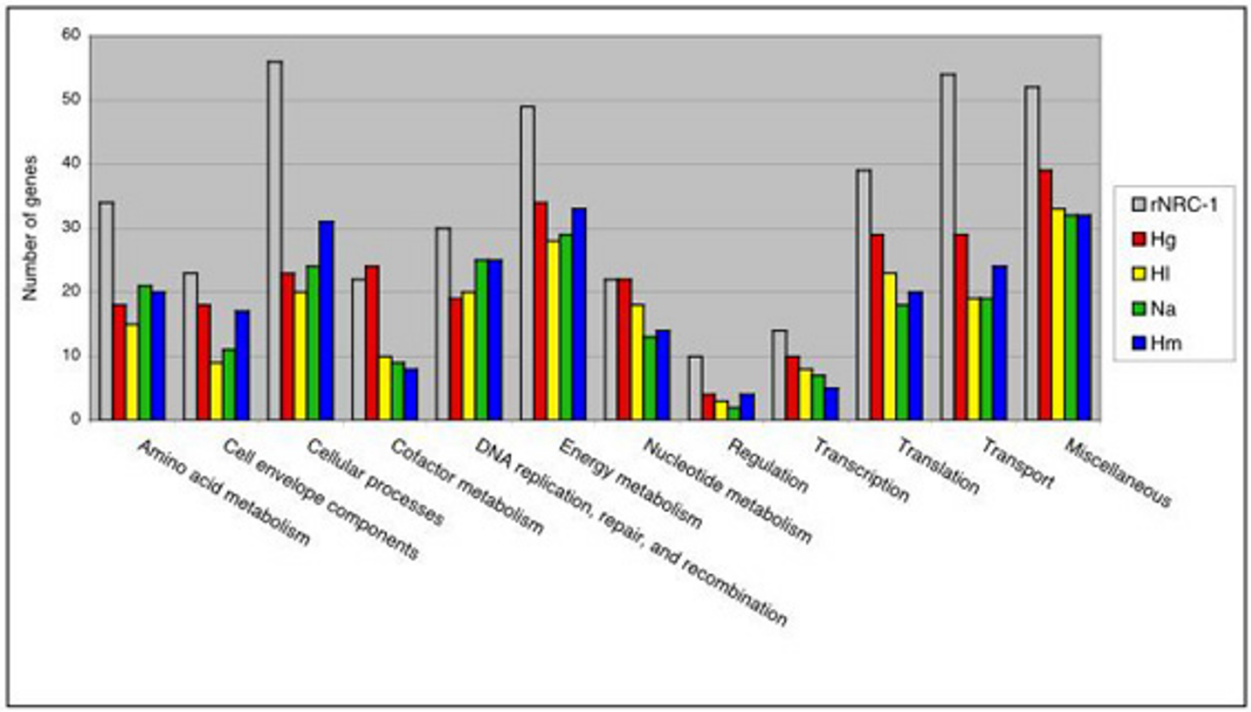

Figure 5
Distribution of core biological functions conserved in *Halobacterium sp.* NRC-1 and four other halophiles. Core biological functions shared between *H. sp.* NRC-1 and each halophile were classified into twelve classes and the number of genes in each category is plotted. Most of the functions are conserved with respect to *H. sp.* NRC-1. Mechanisms common to halophiles, such as active ion transporter system to maintain osmotic balances in and outside of the cells and DNA repair mechanism in response to possible UV – irradiation, seem well preserved.

Interestingly, bacteriorhodopsin (BR) and halorhodopsin (HR), the light driven ion transporters for protons and chloride, were identified in *H. marismortui*. This finding suggests the presence of a similar photoactive proton and chloride pumps for phototrophic growth in *H. marismortui* [20,21]. The kinase 2-keto-3-deoxygluconate (KdgK), an enzyme catabolizing 2-keto-3-deoxygluconate to 2-keto-3-deoxygluconate-6P in the glucose metabolism pathway was identified in *H. marismortui*. This finding suggests the presence of the Entner-Doudoroff pathway, observed in some halophiles including *H. sp.* NRC-1 [2,22,23]. The presence of similar pathways controlling flagellar rotation was also observed among all halophiles in this study. Identification of multiple cell-signalling transducer protein homologs in most of the newly sequenced halophiles suggests dynamic cellular functions in response to rapidly changing environmental conditions.

Conservation of gene order between halophiles

Gene clusters encoding proteins of related functions are often co-localized in prokaryotic genomes; this organization could assist assigning functions to unknown ORFs when gene clusters of two organisms are compared [24]. The prediction of functional linkage between the genes involved in several metabolic pathways across the multiple genomes has been demonstrated to correlate functional with physical linkages of gene clusters [25]. Co-localization of functionally related genes was also identified in the *H. sp.* NRC-1 genome. The *H. sp.* NRC-1 analysis revealed that nine of the ten genes involved in arginine metabolism cluster at three genomic loci. Co-induction of the *arg*S with the *arc*ABC gene cluster on repressing phototrophy has been recently discussed [26]. Co-localization of the gene clusters involved in isoprenoid biosynthesis, purple membrane biogenesis, and nucleotide metabolism are also observed at multiple loci across the *H. sp.* NRC-1 genome.

Table 2: Detection of the conserved gene order between *Halobacterium sp.* NRC-1 and halophiles. Sequence reads matched to observed *H. sp.* NRC-1 genes, number of physically linked genes, and gene names or numbers are listed for each halophile.

	Halobaculum gomorrense				
Hg11_0191	3	1384,rad3a, yvoF	Hg2_0160	2	rpl30p,rps5p
Hg12_0134	3	1564,1561,1562	Hg1_0175	2	2149,etlB
Hg8_0102	3	rpl1 4,prpl24,prps4e	Hg12_0110	2	cbiJ,cbiC
Hg12_0124	3	642,nolC,nuoL	Hg9_0124	2	1087,1086
Hg1_0132	2	rps13p,rps4p	Hg6_0157	2	ribG,1257
Hg13_0114	2	rpl24e,ndk	Hg9_0157	2	rpl13p,rps9p
Hg1_0114	2	5062,5063	Hg12_0125	2	ilVE1,388
Hg2_0136	2	2160,2162	Hg10_0188	2	cobH,cbiH
Hg8_0104	2	359,357	Hg8_0128	2	5191,5189
Hg3_0132	2	2089,2091	Hg1_0174	2	rpl32e,rpl6p
Hg13_0174	2	255,256	Hg1_0168	2	417,419
Hg12_0169	2	gpdB,gpdA2	Hg7_0133	2	cat,maoC1
Hg13_0179	2	gmd,graD3	Hg11_0101	2	sus,trn30
Hg5_0173	2	1641,1642	Hg6 0161	2	2316,cbiO1
Hg12_0172	2	uroM,hem3			
Hg10_0169	2	cbiL,cbiF		***Natrialba asiatica***	
Hg3_0178	2	riml,can	Na5_0165	3	rps3p,1699,rpl29p
Hg8_0132	2	1519,1518	Na7_0184	2	nolD,ndhG4
Hg1_0123	2	tfeA,755	Na11_0183	2	fbp,686
Hg7_0108	2	rpl13p, rpl18e	Na6_0131	2	2534,2535
Hg8_0147	2	trn24,1390	Na1_0196	2	1530,mmdA
Hg10_0108	2	801,trn12	Na3_0196	2	yusM,2121
Hg12_0123	2	rpl19e,rpl18p	Na8_0152	2	1900,1901
Hg5_0122	2	fbp,683	Na1_0112	2	po/B1,520
Hg13_0193	2	2637,bchP	Na11_0168	2	rad24c,2371
Hg5_0165	2	2098,trn36	Na2_0190	2	fbp,686
Hg9_0123	2	sus,trn30	Na1_0154	2	hUtG,hutl
Hg8_0119	2	1429,1431	Na4_0122	2	hpyA,aup
Hg12_0105	2	gatA,gatC	Na6_0176	2	6070,6071
Hg3_0115	2	gapB,hsp2	Na4_0158	2	rp/32e,rpl19e
Hg3_0187	2	*pyrE1,gcd*	Na5_0119	2	1934,ftsZ3
Hg10_0181	2	854,857	Na10_0123	2	1845,ansA
Hg13_0115	2	5069,5071	Na8_0119	2	1296,1297
Hg5_0113	2	hpyA,aup	Na3_0134	2	mamA,mamB
Hg8_0146	2	1783,panF	Na1_0138	2	2474,2475
Hg9_0194	2	1235,1236	Na10_0113	2	htlB,cmk
Hg6_0102	2	traB,875	Na12_0184	2	220,219
Hg9_0115	2	256,258	Na1_0125	2	525,527
Hg10_0131	2	1823,moaB	Na11_0176	2	404,405
Hg2_0106	2	alkK,gdhB	Na12_0110	2	dtd,247
Hg11_0194	2	2205,pmu1	Na8_0103	2	2369,rad24b
Hg5_0170	2	glyA,folD	Na9_0165	2	2306,2307
Hg2_0125	2	purK,purE	Na11_0175	2	rps6e,2515
Hg4_0125	2	*2098,trn36*	Na4_0146	2	gun,purM
Hg8_0107	2	rrlA,rrlB	Na5_0164	2	296,pnm
Hg2_0171	2	ndhG1,nolB	Na9 0164	2	rnh,1982
	Halorubrum lacusprofundi			***Haloarcula marismortui***	
Hl10_0165	3	rps14p,rpl6p,rps8p	Hm10_0182	3	767,768,769
Hl4_0114	2	5166,5167	Hm11_0113	3	1558,1559,cbiH
Hl6_0133	2	rad24c,2371	Hm8_0149	2	trkA4,phoT3
Hl12_0110	2	rps5p,rpl18p	Hm12_0196	2	gpdB,gpdA2
Hl11_0135	2	943,945	Hm7_0109	2	mal,mamB
Hl5_0115	2	alkK,gdhB	Hm8_0151	2	hpyA,rpa
Hl11_0152	2	alaS,2282	Hm4_0135	2	2555,trn44
Hl7_0101	2	pri,1471	Hm5_0163	2	1475,1476
Hl11_0159	2	981,982	Hm4_0159	2	2515,2516
Hl2_0111	2	6060,6061	Hm1_0167	2	576,578
Hl2_0107	2	pmm,2277	Hm6_0177	2	6434,6437

Table 2: Detection of the conserved gene order between *Halobacterium sp.* NRC-1 and halophiles. Sequence reads matched to observed *H. sp.* NRC-1 genes, number of physically linked genes, and gene names or numbers are listed for each halophile. *(Continued)*

HI7_0142	2	rpl3p,1688	Hm10_0123	2	trpE2,383
HI5_0125	2	trp1,124	Hm11_0186	2	2024,yyal
HI2_0193	2	rps4p,rps11p	Hm1_0106	2	rad3a,1382
HI8_0104	2	2333, cysG	Hm10_0192	2	sop1 ,htr1
HI7_0124	2	576,578	Hm9_0134	2	mal, mamB
HI9_0183	2	1613,mvaB	Hm8_0125	2	1845,ansA
HI12_0125	2	5131,5132	Hm1_0150	2	2369,rad24b
HI7_0158	2	2316,cbiO1	Hm4_0112	2	1823,moaB
HI10_0161	2	citE.,gdhA1	Hm6_0130	2	533,534
HI2_0112	2	2241,rfcC	Hm6_0112	2	1823,moaB
HI5_0112	2	2645, vacB	Hm6_0149	2	alkK,gdhB
HI9_0130	2	pgi,1993	Hm1_0165	2	266,267
HI5_0127	2	1578,1580	Hm11_0158	2	noxA,1806
HI6_0173	2	6198,6197	Hm7_0164	2	*dsa,lpdA*
HI11_0150	2	2089, trm1	Hm5_0174	2	2160,2162
HI11_0160	2	981,982	Hm11_0116	2	mal,mamB
HI5_0134	2	trn41,rrs	Hm12_0116	2	gpdB,gpdA2
HI1_0162	2	trn42,rrlB	Hm8_0137	2	dsa,lpdA
HI6_0103	2	1672,1670	Hm1_0104	2	1296,1297
HI6_0131	2	pri,1468	Hm12_0184	2	rps6e,2515
HI4_0176	2	trpB,trpC	Hm9_0104	2	758,759
HI10_0168	2	rpl14p,rps17p	Hm11_0176	2	ndhG1,ndhG5
HI2_0170	2	2149,ettB	Hm11_0135	2	trn46,rps10p
HI9_0170	2	ugpE,ugpC	Hm4_0154	2	yvrO,3
HI5_0149	2	rpl6p,rps8p	Hm4_0136	2	2370,rad24b
HI12_0187	2	2581,2582	Hm2_0113	2	hutG,hutI
HI6_0121	2	ndhG1,ndhG5	Hm12_0152	2	asnA,865
			Hm8_0109	2	1558,cdiH
			Hm7_0186	2	230,231
			Hm8_0134	2	361,nfi
			Hm8_0180	2	rpl14p,rpl24p
			Hm8_0177	2	711,713

We found many instances of two or more consecutive genes in *H. sp.* NRC-1 that may be similarly linked in the four newly sequenced halophiles (Table 2). Detection of the conserved gene orders among the genes associated with important biological functions such as amino acid metabolism, cofactor metabolism, energy metabolism, translation, and transport suggests selective pressure to maintain such gene arrangements.

Transcription factor family

Archaea possess a simplified version of the eukaryotic RNA polymerase II-like transcription system including highly conserved TBP and TFB proteins, but no eukaryal-like TBP-associated factors (TAFs) [27,28]. Recently, multiple TBP and TFB proteins have been reported in several archaea including *H. sp.* NRC-1 [12] and another halophilic archaea isolated from the Dead Sea, *Haloferax volcanii* [29]. A novel mechanism of transcriptional regulation by forty-two different combinations of multiple TBP and TFB interactions has been recently suggested for NRC-1 [30].

In our database searches, we identified several TFB homologs including TfbF and TfbC homologs from *H. gomorrense*, TfbF and TfbG homologs from *H. lacusprofundi*, and a TfbF homolog from *N. asiatica.* Pairwise amino acid identity of these newly identified TFBs to *H. sp.* NRC-1 TFBs demonstrated a high conservation of more than 83% except for the TfbC (64%) homolog from *H. gomorrense.* Sequence alignments between the newly identified TFBs and the seven *H. sp.* NRC-1 TFB proteins revealed conserved functional domains with eukaryal TFIIB proteins (Figure 6).

Among the newly identified TFB proteins, one *H. gomorrense* protein (TfbC) had sequence spanning the amino-terminal sequences. However, the possession of a likely zinc finger domain with conserved patterns of $(CXXC)X_{15-17}(CXXC)$ found in *H. sp.* NRC-1 [30] was not observed. A helix-turn-helix motif, the TFIIB recognition element (BRE) binding domain, was identified in every new TFB protein with a sequenced carboxy-terminus [31]. The amino acid residues glycine and lysine/arginine,

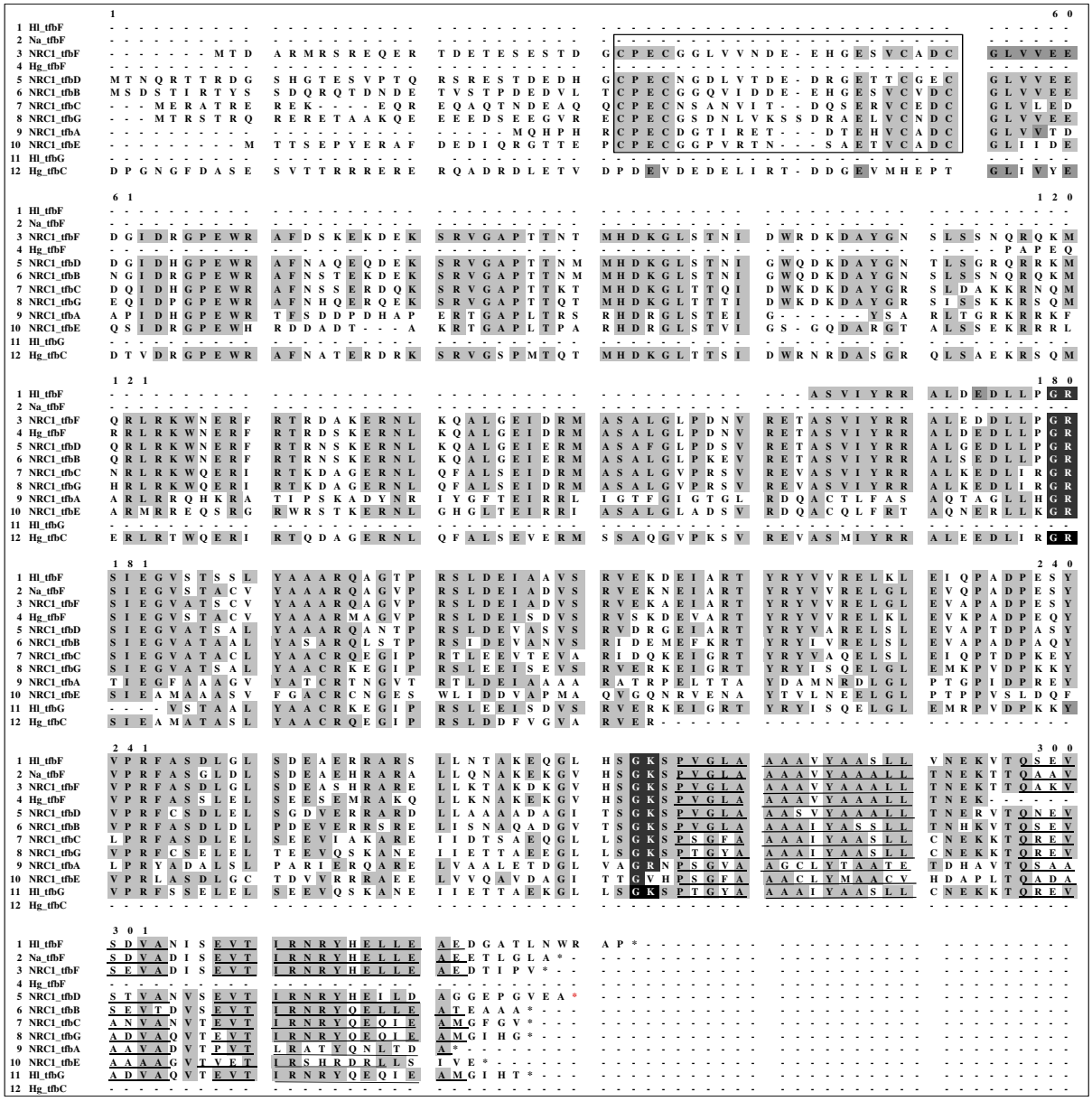

Figure 6
Sequence alignments for TFB proteins. Sequences for seven *Halobacterium sp.* NRC-1 TFBs and five predicted TFBs from *H. lacusprofundi* (Hl), *H. gomorrense* (Hg), and *N. asiatica* (Na) were aligned. Amino acids conserved in at least six of the twelve sequences are shaded (the amino acid involved in interaction with TBP are shaded in black). The boxed region indicates likely Zn-finger domains [(CXXC)X{15–17}(CXXC)]. Helix-turn-helix DNA binding motifs are underlined.

involved in TBP interaction [30], were conserved in the newly identified TFB proteins.

In order to study TFB family evolution, orthology relationships between the newly identified TFBs and *H. sp.* NRC-1 TFBs were analyzed by constructing a molecular tree from protein sequence distances. Of the five newly

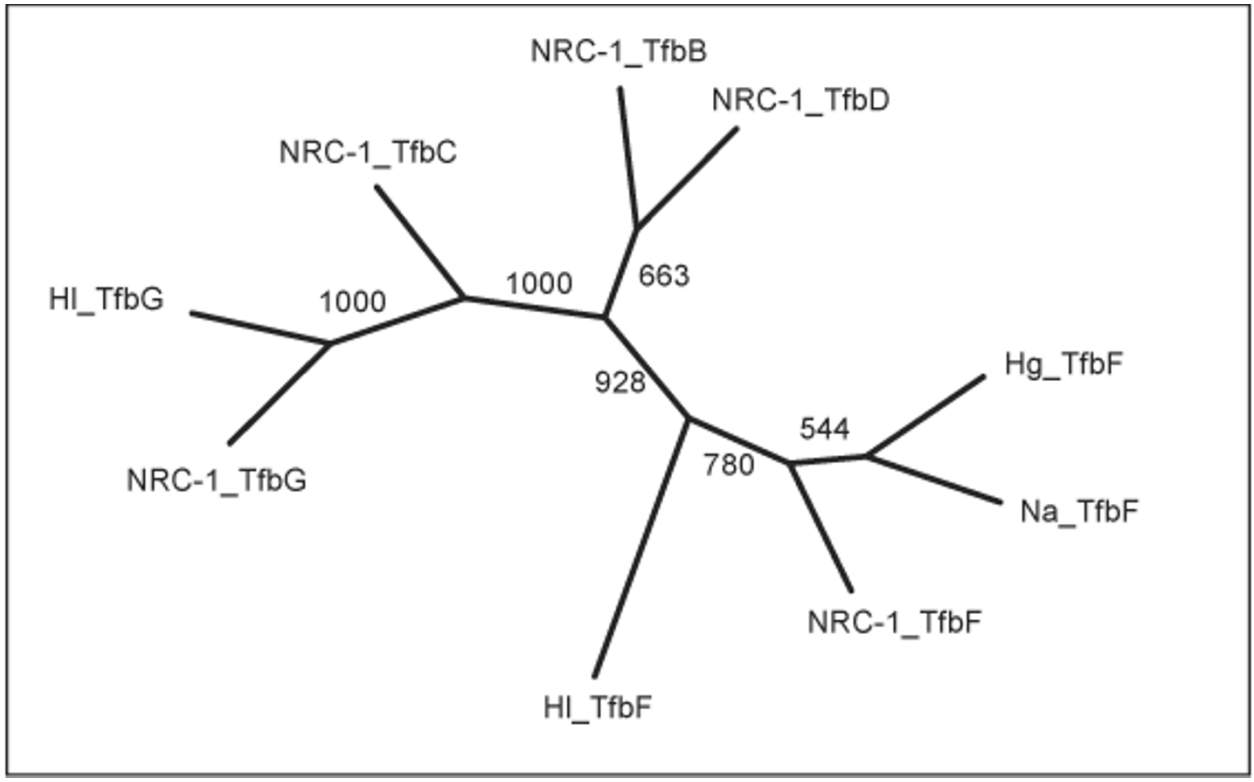

Figure 7
Unrooted molecular tree of TFB protein family. Newly identified TfbF proteins from *H. lacusprofundi* (Hl), *N. asiatica* (Na), and *H. gomorrense* (Hg) and TfbG from *H. lacusprofundi* show orthology with TfbF and TfbG of *H. sp.* NRC-1. Orthology was supported by the high bootstrap scores (numbers on the branch) generated by 1000 replicates. The results suggest similar structural and functional characteristics between TFB proteins among the halophiles.

identified TFBs, TfbC of *H. gomorrense* and TfbA and TfbE of *H. sp.* NRC-1 were excluded from the molecular tree generation because these sequences had high sequence divergence from all other members. The resulting molecular tree lends support to orthologous relationships among the newly identified TFBs and those of *H. sp.* NRC-1, suggesting similar structural and functional characteristics between these TFB proteins (Figure 7).

Our findings of multiple TFB homologs in *H. lacusprofundi* and *H. gomorrense* suggest the presence of multiple transcription factors in other halophiles as well. In order to evaluate such a possibility, enzyme-digested genomic DNA of five halophiles, including *H. sp.* NRC-1, were subjected to Southern hybridization to estimate the number of related or identical sequences to *tfb* genes. By counting only highly distinguishable bands in the Southern blot (data not shown), at least four *tfb*-related sequences were identified in all newly sequenced halophiles but *N. asiatica* (three). *H. sp.* NRC-1 genomic DNA, used as a positive control, hybridized to seven bands, presumably corresponding to the seven known *tfbs*.

In our low-pass sequence analysis, the presence of a TBP homolog was not observed in the newly sequenced halophiles. Using Southern hybridization, we estimated the numbers of sequences related to *tbp* genes. The results suggest the possible presence of two *tbp* related sequences from *H. gomorrense* and *H. lacusprofundi* and one *tbp*-related sequence from *H. marismortui* and *N. asiatica*. These four halophiles contain fewer copies of *tbp* genes than *H. sp.* NRC-1, which showed six *tbp*-related sequences in Southern analysis, corresponding to the six known *tbp*s.

Identification of the presence of multiple transcription factor genes suggests transcriptional regulation through a variety of TBP-TFB combinations in the four newly sequenced halophiles.

Global similarity estimation to NRC-1

The four newly sequenced halophiles, as well as the previously sequenced *H. sp.* NRC-1, belong to the *Halobacteriaceae* family. More recent isolates *H. marismortui* and *H. gomorrense* are from the Dead Sea, *H. lacusprofundi* from Antarctica, and *N. asiatica* from Japan [6]. Sequence similarity is expected between closely related species. However, adaptation to their unique environments may have contributed to sequence divergence in these organisms. Similarity between organisms may be estimated among orthologous genes of all analyzed species, but this limits analysis to only a small number of genes, which may not be representative of global similarity.

In order to estimate sequence divergence of our four shotgunned species from *H. sp.* NRC-1, overall sequence similarity was determined from the BLAST DNA and protein sequence alignment identities to *H. sp.* NRC-1. The species with the highest average sequence similarity to *H. sp.* NRC-1 was *H. lacusprofundi*, with 62% identity in DNA and 46% in protein sequences. *H. gomorrense* had 62% and 45%, *H. marismortui* had 60% and 43%, and *N. asiatica* had 60% and 43% identity to DNA and protein sequences of *H. sp.* NRC-1, respectively. The sequence divergence may be due to different selective pressures in the unique environments of each species.

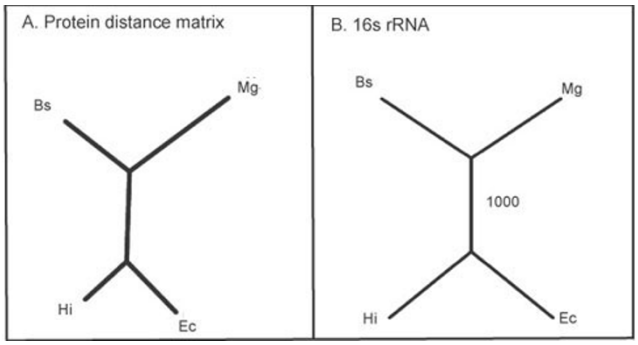

Figure 8

Molecular tree of four completed genomes. Molecular tree of *E. coli* (Ec), *H. influenzae* (Hi), *M. genitalium* (Mg), and *B. subtilis* (Bs) were reconstructed by the protein distance matrix (A) and the 16S rRNA (B) analysis. Molecular trees reconstructed by two different methods are compared and show agreement each other. The methodology using protein distances of whole genome sequences could successfully estimate the evolutionary relationship among the four complete genomes. The bootstrap score (1000) is shown for the 16S rRNA analysis.

Estimation of evolutionary distance between five halophiles

16S rRNA is useful for studying evolution because it is ubiquitous in all prokaryotes and contains highly conserved as well as variable regions, which can be used for determination of divergences at different taxonomic levels. For instance, the highly conserved region can be used to compare distantly related organisms whereas the variable region can be used to look at the divergence between closely related species. Although analysis based on 16S rRNA is useful, it does not yield unambiguous conclusions. Phylogenetic analysis of individual genes (such as 16S rRNA) is based upon only a small portion of an entire genome. Since a massive amount of sequence data (including the whole genome sequences) is now available, alternative methods of phylogenetic analysis that make use of these data are possible.

One method developed in this study for phylogeny reconstruction utilizes the protein distances calculated between the species. The method was first tested on completed genomes including *Escherichia coli* K-12 Strain MG1655, *Haemophilus influenzae* Rd KW20, *Mycoplasma genitalium* G-37, and *Bacillus subtilis* 168. *E. coli* and *H. influenzae* belong to the same bacterial phylum, gamma proteobacteria, while *M. genitalium* and *B. subtilis* belong together in the low-GC gram-positive phylum. These genomes were selected to determine if the newly developed method is able to identify distance between closely related species (in the same phylum), and also to test its ability to discriminate between different phyla. These two phyla are closely related based on 16S rRNA phylogeny. Unrooted molecular trees were computed; the tree computed by our novel protein distance matrix was compared to the tree computed from 16S rRNA (Figure 8). The reconstructed molecular trees are in good agreement. Our simple methodology thus demonstrates an alternative way of calculating protein distances using more sequence data, and validates the current proposed relationship of four genomes.

Knowing that evolutionary relationships can be closely estimated using protein distance matrix methods, we then employed the method to calculate distances between the halophiles. The molecular tree obtained was compared to the 16S rRNA tree (Figure 9). Our analysis indicates *H. sp.* NRC-1 is more closely related to *H. lacusprofundi*, but the 16S rRNA analysis shows *N. asiatica* is more closely related to *H. sp.* NRC-1. It is not clear which molecular tree more likely represents the true evolutionary relationships among these 5 halophiles. However, closer relationship between *H. sp.* NRC-1 and *H. lacusprofundi* were supported by the identification of the highest number of IS elements and genes that are homologs to *H. sp.* NRC-1 in *H. lacusprofundi.*

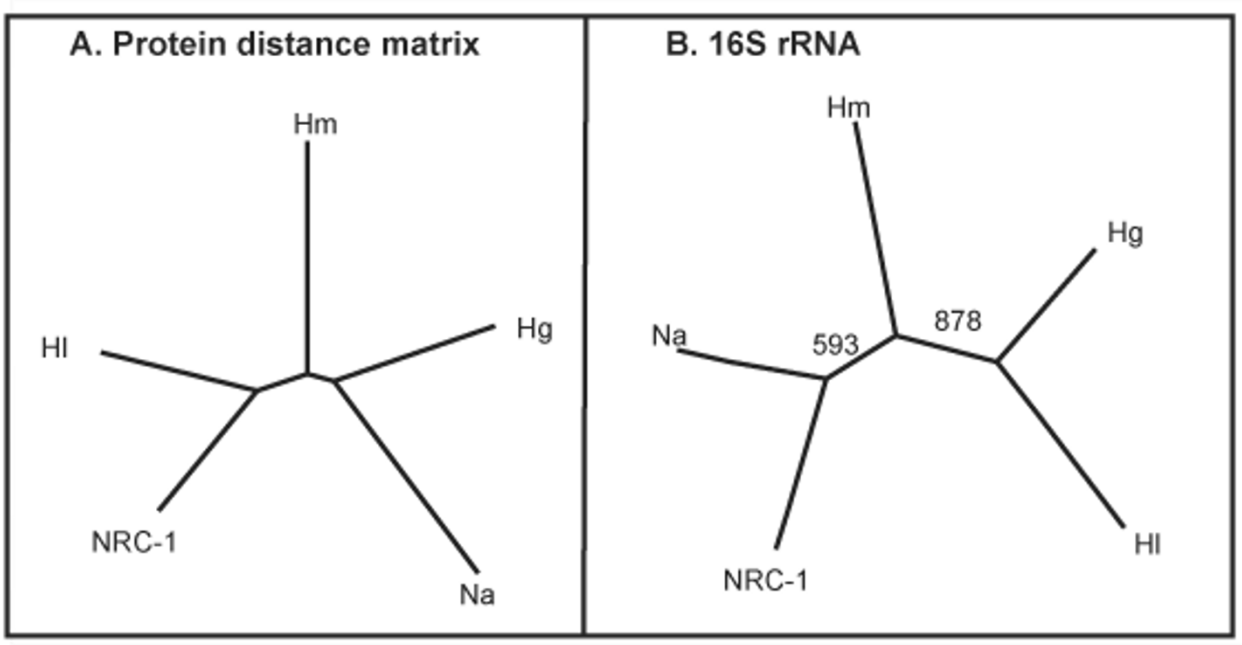

Figure 9
Molecular trees of five halophiles. Molecular tree was reconstructed by the protein distance matrix (A) and the 16S rRNA (B) analysis. The comparison between molecular trees obtained by two different methods showed lack of congruence among halophiles. Both trees resemble a "star" phylogeny among the five species, suggesting significant molecular divergence between species. The bootstrap scores are shown on 16S rRNA analysis.

Discussion

Orthology and divergence between the genomes of NRC-1 and other halophiles

We have studied four extremely halophilic archaea by low-pass genome sequencing and compared the results to a reference genome (*H. sp.* NRC-1) to better understand their orthologous characteristics. Despite the geographic remoteness of the habitats from which each of these species was isolated, and the evolutionary pressure to adapt to their unique environments, some unique core characteristics of extremely halophilic archaea seem conserved among the five species. Strikingly, the four newly sequenced halophiles also exhibit low pI of their proteins. The result suggests the presence of acidic proteins is an essential mechanism to function in hypersaline environments and that these four halophiles also maintain high salt concentration in their cytoplasm to maintain ionic balance in hypersaline environments.

Approximately 60% or higher overall GC content was observed in the halophile species analyzed in our study. This observation suggests that halophiles evolved to maintain such GC content as a survival mechanism in order to counteract DNA damage, mostly in the form of thymine dimers, caused by solar UV irradiation. Together with the observed low pI of the proteins, this finding suggests that additional characteristics are shared among the five halophiles, corresponding to life in UV irradiation and high salinity. Furthermore, analysis of putative genes homologous to *H. sp.* NRC-1 genes indicates that core biological functions governing cellular and genetic mechanisms of halophiles are well conserved between *H. sp.* NRC-1 and the four halophiles.

The newly identified TFBs included in the molecular tree analysis had pairwise amino acid identity of more than 83% with those in *H. sp.* NRC-1. The molecular tree of the newly identified TFBs and *H. sp.* NRC-1 TFBs suggests an orthologous relationship of the TfbG protein of *H. lacusprofundi* with *H. sp.* NRC-1. TfbF proteins identified from *H. lacusprofundi*, *H. gomorrense*, and *N. asiatica* also show orthologous relationships with TfbF of *H. sp.* NRC-1. These relationships were supported by high bootstrap values. Identification of multiple TFB proteins in all newly sequences halophiles suggests transcriptional regulation through a variety of TBP-TFB combinations, as previously suggested [30]. Different TBP-TFB combinations are likely to serve in response to different cellular and environmental stimuli.

Analysis of IS elements showed that more sequence reads for *H. lacusprofundi* (38 reads) had similarity to NRC-1 IS elements, followed by *H. marismortui* (10 reads). Considering that the sequences are generated from a random shotgun library and that similar coverage of each genome sequence was used for analysis, these results suggest that the *H. lacusprofundi* genome contains the highest number of IS sequences homologous to *H. sp.* NRC-1. This conservation may be due to these transposon elements entering the two genomes by multiple transposition processes, or simply due to a more recent divergence from a common ancestor.

The overall DNA sequence identity of the four halophiles to *H. sp.* NRC-1 was 60–62%. Similarly, an average of 39% of putative genes identified from the four halophiles were not homologous to *H. sp.* NRC-1 genes. The sequence divergence may conceivably be due to different selective pressures in the environment of each species. The four newly sequenced halophiles also exhibit homology to both bacteria and eukaryotes, consistent with similar observations in *H. sp.* NRC-1 and other archaea.

Molecular tree reconstruction

The evolutionary relationships of the five halophiles were assessed by two methods: a protein distance matrix and a 16S rRNA analysis. The protein distance matrix method we have developed for phylogeny reconstruction utilizes all available sequence information and calculates protein distances from the results of alignments of protein sequences between species. The method is very simple, as it relies on mean or median similarity of the entire complement of predicted proteins, but appears to offer sufficient resolution to provide useful insight across relevant evolutionary time frames. The molecular trees obtained by the protein distance matrix and the 16S rRNA lacked congruence. The protein distance matrix analysis indicates that *H. sp.* NRC-1 is more closely related to *H. lacusprofundi*, while the 16S rRNA analysis suggests that *N. asiatica* is more closely related to *H. sp.* NRC-1. Although 16S rRNA is the commonly used sequence to study evolutionary divergence, some discrepancies between the rRNA trees and gene trees in the investigation of individual genes and proteins have been previously discussed [32,33]. In contrast, the result of the protein distance matrix is supported by our findings of DNA and protein sequence similarity analysis, which indicated the highest sequence similarity between *H. lacusprofundi* and NRC-1. This finding was further supported by the presence of the highest number of IS elements that are homologous to IS elements found in *H. sp.* NRC-1. In addition, 63 % of putative genes identified from *H. lacusprofundi* sequences were homologous to NRC-1 genes, the highest percent among the four halophiles in this study. Although the observed "star" phylogeny suggests significant molecular divergence between species, considered together, our findings suggest the closest of the relationships is between NRC-1 and *H. lacusprofundi*.

Utility of low-pass sequencing for comparative genomics

Our analysis of low-pass sequencing demonstrated possible ways of identifying coding regions from sequence reads and allowed comparison between related species. Low-pass sequencing becomes more useful when a closely related reference genome is available. Some of our findings include: 1) identification of 60% or higher GC content and low protein pI among five halophiles, 2) identification of IS elements in all halophiles, 3) detection of gene conservation among halophiles, including approximately 60% genes shared with *H. sp.* NRC-1, spanning many core biological functions; 4) identification of genes belonging TFB families, and 5) detection of conserved gene orders between *H. sp.* NRC-1 and other halophiles.

Low-pass sequencing, however, exhibited limited utility for some genome analyses. Partial ORFs of predicted proteins limit detailed analysis on such proteins. Large-scale genomic topology cannot be inferred from low-pass sequencing. The evidence of common genetic exchange within a species and among closely and distantly related microbes by lateral gene transfer is hard to elucidate. A complete list of genes in a gene family or elements involved in a pathway could not be obtained. The basis of additional extremophilic characters, such as psychrophily, alkaliphily, barophily, and acidophily are difficult to determine. Low-pass sequencing has limited ability to resolve multigene families, pseudogenes, and allelic variations. These last are seldom issues in microbial genetics, but could be problematic in diploid organisms.

Nevertheless, to obtain a general comparative view on genomes of related species, low-pass sequencing can be an effective way of providing information in a time- and effort-efficient manner [34].

Conclusions

Low-pass sequencing is useful for comparative genomics, and will be increasingly so as sequencing technologies improve. It could be both cost- and time-effective for comparative genomics. Low-pass sequence data enables diverse analyses, especially identifying orthologous characteristics between a reference genome, such as that of *H. sp.* NRC-1 and related species. Comparative analyses show shared mechanisms governing cellular and genetic functions among these five halophiles. High GC contents, low pI of predicted proteins, multiple IS elements, biological pathways, and multiple transcription factors are conserved among the species despite geographic diversity of habitats.

Methods

Strains

The halophiles used in the study for sequencing were obtained from American Type Culture Collection (ATCC, http://www.atcc.org). *Haloarcula marismortui* (ATCC#: 43049), *Halobaculum gomorrense* (ATCC#: 700876), *Halorubrum lacusprofundi* (ATCC#: 49239), and *Natrialba asiatica* (ATCC#: 700177) were cultured in ATCC media 1218, 2169, 1682, and 805, respectively, prepared according to the directions from ATCC. The estimated genome sizes of the halophiles by pulse field gel electrophoresis were: *H. marismortui*, ~3,900 kb; *H. lacusprofundi*, ~2,600 kb; *N. asiatica*, ~3,100 kb; and *H. gomorrense*, ~2,700 kb (data not shown).

Preparation of templates for DNA sequencing

White colonies from a blue/white screened shotgun library were picked and transferred to 96-well PCR plates containing 5 μl deionized water. Cells were resuspended thoroughly by multiple pipetting and a 45 μl of PCR mixture containing 1X buffer, 0.25 mM dNTP, 1.5 mM $MgCl_2$, 8% DMSO, 0.2 μM each forward and reverse primer and 1.25 unit of Taq polymerase was added into each well. Taq polymerase, 10X buffer (without $MgCl_2$) and $MgCl_2$ were purchased from Promega (Promega Corporation, Madison, WI). Primers used in the study were modified M13 forward and reverse primers (Forward primer: 5'-TTT CCC AGT CAC GAC GTT GTA-3'. Reverse primer: 5'-GTG AGC GGA TAA CAA TTT CAC3'). PCR was carried out by denaturing at 92°C for 10 seconds, annealing at 58°C for 45 seconds, and extension at 72°C for 3 minutes for a total of 35 cycles. PCR products with clear bands were further selected and then purified using MutiScreen-FB, Opaque plates (Millipore Corporation, Bedford, MA) according to the manufacturer's direction. Purified PCR products were analyzed by agarose gel electrophoresis to assess DNA quality. The average size of DNA was ~2.5 kb.

Automated DNA sequencing and sequence processing

Sequencing was carried out using ABI PRISM® BigDye™ Terminator cycle-sequencing ready-reaction kits, Version 2.0 and 3.0 (PE Corporation, Foster City, CA). The cycle sequencing reaction was performed in 96-well plates according to the procedures provided by the manufacturer using 3.3 pmol of M13 forward primer, up to 1 μg template DNA, and 4 μl BigDye™ Terminator in a total of 10 μl reaction volume. Thermal cycling was set for denaturing at 96°C for 10 seconds, annealing at 50°C for 5 seconds, and extension at 60°C for 2 minutes for a total of 50 cycles. Sequencing electrophoresis was performed on the ABI PRISM 3700 automated DNA analyzer (PE Corporation, Foster City, CA). Using PHREDPHRAP http://www.phrap.org, DNA chromatograms were analyzed, bases called, and quality values assigned.

Gene prediction

Sequence reads were translated into six reading frames and compared to the nr protein database at NCBI http://www.ncbi.nlm.nih.gov using the BLASTX program, with default parameters. The best match of each search result with a minimum 30 amino acid alignment, protein identity match ≥ 30%, and BLAST *P* value ≤ 10^{-5} was considered to possibly represent a protein encoding gene. In prokaryotes, there is a high prior probability of the presence of a gene on a sequence read, so a 30 amino acid alignment yields few false positive gene/ORF assignments; this might not be the case in vertebrates, for example. Our criteria were selected based on validation test results from the random *H. sp.* NRC-1 sequences searched against the nr protein database at NCBI. Using these criteria, sequences encoding the genes and sequences derived from intergenic regions resulting in no match to the nr protein database were identified, using manual curation to eliminate obviously spurious results (e.g., those resulting from low-complexity alignments).

If more than one "best match" ORF mapped to the same homolog, then only the longest match was retained as a representative of that homolog. Most of the "best match" ORFs lacked a start codon, stop codon, or both. One sequence read could contain two "best match" ORFs, with the ends of the sequence matching to two different proteins.

Identification of Insertion sequence elements

A database containing IS element sequences found in the *H. sp.* NRC-1 genome was created. All *H. sp.* NRC-1 IS sequences were searched against each halophile sequences using BLASTN http://www.ncbi.nlm.nih.gov/blast. A minimum 90 bp nucleotide match with ≥ 70% identity was considered to represent a true match. A search of the nr-protein database was performed to identify IS elements present in other species besides *H. sp.* NRC-1; no additional IS elements were identified. Sequence reads identified from the search results as IS elements were then compared to a list of unknown function *H. sp.* NRC-1 genes (named "*vng*" genes [2]) of containing IS elements.

GC composition analysis

To avoid repetitive counting of overlapping nucleotides common to more than one sequence, PHREDPHRAP http://www.phrap.org assembled sequences into contigs based on overlap regions between sequences. After the assembling process, PHREDPHRAP creates output files for assembled sequences as contigs, and for the unassembled sequences, as singlets. The global GC composition of each halophile was analyzed from contigs and singlets files by calculating the frequency of each nucleotide.

Prediction of isoelectric point of predicted protein

"Best match" ORFs were extracted from each sequence read by a Perl script. The sequences were converted into GCG format and submitted to the GCG.® Wisconsin Package™ program PEPTIDESORT http://www.accelrys.com to obtain isoelectric point values.

Analysis of gene order conservation

In order to detect instances of two or more consecutive genes in *H. sp.* NRC-1 that may be similarly linked in the four newly sequenced halophiles, all sequence reads were compared to the *H. sp.* NRC-1 genome as a reference sequence. Due to the sequence divergence between these genomes, this comparison requires the most sensitive alignment algorithm in order to be successful. The FASTA3 [35] program was used with the "A" parameter to force the use of unlimited Smith-Waterman [36,37] alignment for the final DNA sequence alignment. Using a Perl script for automation, one or more pairs (due to multiple matches to NRC-1) of start and end positions were defined for each sequence read, rejecting segments with expectation values above 0.1. The most significant matching sequence read (by calculated expectation value) was compared to the gene annotation of NRC-1 to determine which genes, if any, overlap.

Analysis of sequence similarity to NRC-1

The DNA sequence similarity to *H. sp.* NRC-1 was analyzed by comparing all of the sequence reads against the *H. sp.* NRC-1 genomic DNA sequence database using BLASTN http://www.ncbi.nlm.nih.gov/BLAST. Protein sequence similarity to *H. sp.* NRC-1 was analyzed by comparing all the sequence reads against the *H. sp.* NRC-1 genomic DNA sequence database using TBLASTX http://www.ncbi.nlm.nih.gov/BLAST. A Perl script was written to extract the best match result including E and *p* value, % identity, and the query sequence reading frame matched to the subject. A minimum 90 bp nucleotide match for DNA, or 30 amino acid match with ≥ 30 % identity for protein between the query and the subject sequences was required for inclusion in the analysis. The "percent identity" value from all the search results was used to calculate average similarity.

Southern hybridization analysis

Enzyme-digested genomic DNA of five halophiles including *H. sp.* NRC-1 were subject to Southern hybridization to identify the copy numbers of *tbp* and *tfb* genes. Aliquots of 5 μg of genomic DNA were digested with *Bam*HI, *cla*I, *Kpn*I, *Pst*I, and *Sma*I restriction enzymes and fractionated by agarose gel electrophoresis in 1X TAE buffer (40 mM Tris-acetate, 2 mM EDTA, pH 8.0). The genes *tfbF* and *tbpE* found in *H. sp.* NRC-1 were used for the Southern blot analysis probe. The *tfbF* gene was selected as a probe because our findings show homologous sequences in most of the halophiles used in this study. The *tbpE* gene was selected as a probe because the expression of this protein, but not of the other TBP, had been confirmed in our proteomics study [38]. The genes were PCR-amplified from *H. sp.* NRC-1 genomic DNA and the PCR amplified genes were excised and purified from the gel by the Sephaglas™BandPrep Kit (Amersham Pharmacia Biotech Inc. Piscataway, NJ). The gene sizes for *tfbF* and *tbpE* are 950 bp, 550 bp, respectively. DNA labelling, hybridization, and detection were carried out using DIG High Prime DNA Labelling and Detection Starter Kit II (Roche Diagnostics GmbH, Mannheim, Germany) according to the manufacturer's directions.

Molecular tree construction

ClustalX ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/ was used to construct multiple alignments. The range of the multiple alignments to be included in the analysis was decided based upon alignment quality. Multiple sequence alignments were visually inspected, and regions with the least number of gaps were selected. Different alignment lengths were tested for each tree construction and found not to significantly affect the results. Sequences were then bootstrapped with thousand replicates using the SEQBOOT algorithm implemented in the PHYLIP 3.573c package http://evolution.genetics.washington.edu/phylip.html. For DNA sequences, distances and the molecular tree were computed with the DNAML algorithm with the global rearrangements heuristic option of PHYLIP. Protein distances were computed with the PROTDIST algorithm using the Dayhoff matrix, and the unrooted molecular tree was computed with the FITCH algorithm.

Multiple Sequence Alignments and Identification of protein motifs

Conserved sequences of the multiple alignment results were shaded using the multiple-alignment printing and shading program, BOXSHADE 3.21 http://www.ch.embnet.org/software/BOX_form.html. Conserved motifs were identified by searching sequences against the Pfam database http://pfam.wustl.edu.

Estimation of evolutionary distance by means of protein distance

Annotated gene sequences of four complete microbial genomes, *E. coli*, *H. influenzae*, *M. genitalium*, and *B. subtilis* were retrieved from the TIGR microbial database http://www.tigr.org/tdb/mdb/mdbcomplete.html. DNA sequences were then translated into six reading frames. CROSS_MATCH http://www.phrap.org was used to compare the sets of sequences with other sets of sequences. Several combinations of command line options, such as MINMATCH, BANDWIDTH, and MINSCORE, were tested. The cut-off criteria selected in the study was MIN-

SCORE 250, MINMATCH 15, and BANDWIDTH 1. The output file of CROSS_MATCH contains the substitution score which represents the distance between two species calculated from the differences in a given protein sequence alignment. The median values of the substitution scores were then calculated and stored as a pairwise distance matrix.

For rRNA tree construction, 16S rRNA sequences of halophiles and four complete genomes used in the study were retrieved from NCBI (*H. sp.* NRC-1, AE005128; *H. lacusprofundi*, X82170; *N. asiatica*, AB046875; *H. gomorrense*, L37444; *H. marismortui*, AF034620; *B. subtilis*, NC000964; *H. influenzae*, NC000907; *E. coli*, NC000913; and *M. genitalium*, X77334).

Author's contributions

YG: carried out sequencing, annotation, analyses, study design and drafted the manuscript. JR: performed statistical analysis of low-pass sequencing, developed the protein matrix phylogeny estimation method, and drafted the manuscript. GG: carried out identification of conserved gene orders. NB: participated in the design of the study. KD: wrote Perl scripts. MP: participated in sequencing. SK: provided feedback for the manuscript. SD: conceived of the study, and participated in the study design. WN: participated in the design of the study and coordination. LH: conceived of the study, and participated in its design and coordination.

All authors have read and approved the final manuscript.

Acknowledgements

This work was supported by grants from the National Science Foundation (MCB-9900497 to LH and MCB-0135595 to SD) and funds from Merck & Co. Inc. to the Institute for Systems Biology.

References

1. Tindall BJ: **The family *Halobacteriaceae*.** In: *The Prokaryotes: a Handbook of Bacteria. Ecophysiology, Isolation, Identification, Applications Volume 1. New York, Springer*; 1992:768-808.
2. Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J *et al.*: **Genome sequence of *Halobacterium* species NRC-1.** *Proc Natl Acad Sci USA* 2000, **97:**12176-12181.
3. DasSarma S: **Mechanisms of genetic variability in *Halobacterium halobium:* the purple membrane and gas vesicle mutations.** *Can J Microbiol* 1989, **35:**65-72.
4. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S: **Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence.** *Genome Res* 2001, **11:**1641-1650.
5. Franzmann PD, Stackebrandt E, Sanderson K, Volkman JK, Cameron DE, Stevenson PL, McMeekin TA, Burton HR: ***Halobacterium lacusprofundi* sp. nov., a Halophilic Bacterium Isolated from Deep Lake, Antarctica.** *Syst Appl Microbiol* 1988:20-27.
6. Kamekura M: **Diversity of extremely halophilic bacteria.** *Extremophiles* 1998, **2:**289-295.
7. Oren A, Ginzburg M, Ginzburg BZ, Hochstein LI, Volcani BE: ***Haloarcula marismortui* (Volcani) sp. nov., nom. rev., an extremely halophilic bacterium from the Dead Sea.** *Int J Syst Bacteriol* 1990, **40:**209-210.
8. Kamekura M, Dyall-Smith ML: **Taxonomy of the family *Halobacteriaceae* and the description of two new genera Halorubrobacterium and Natrialba.** *J Gen Appl Microbiol* 1995, **41:**333-350.
9. Oren A, Gurevich P, Gemmell RT, Teske A: ***Halobaculum gomorrense* gen. nov., sp. nov., a novel extremely halophilic archaeon from the Dead Sea.** *Int J Syst Bacteriol* 1995, **45:**747-754.
10. Roach JC: **Random subcloning.** *Genome Res* 1995, **5:**464-473.
11. Koonin EV, Galperin MY: **Prokaryotic genomes: the emerging paradigm of genome-based microbiology.** *Curr Opin Genet Dev* 1997, **7:**757-763.
12. Ng WV, Ciufo SA, Smith TM, Bumgarner RE, Baskin D, Faust J, Hall B, Loretz C, Seto J, Slagel J *et al.*: **Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome?** *Genome Res* 1998, **8:**1131-1141.
13. DasSarma S: **Identification and analysis of the gas vesicle gene cluster on an unstable plasmid of *Halobacterium halobium*.** *Experientia* 1993, **49:**482-486.
14. Moore RL, McCarthy BJ: **Characterization of the deoxyribonucleic acid of various strains of halophilic bacteria.** *J Bacteriol* 1969, **99:**248-254.
15. Lanyi JK: **Light energy conversion in *Halobacterium halobium*.** *Microbiol Rev* 1978, **42:**682-706.
16. Bonnete F, Madern D, Zaccai G: **Stability against denaturation mechanisms in halophilic malate dehydrogenase "adapt" to solvent conditions.** *J Mol Biol* 1994, **244:**436-447.
17. Eisenberg H: **Life in unusual environments: progress in understanding the structure and function of enzymes from extreme halophilic bacteria.** *Arch Biochem Biophys* 1995, **318:**1-5.
18. Madern D, Zaccai G: **Stabilisation of halophilic malate dehydrogenase from *Haloarcula marismortui* by divalent cations – effects of temperature, water isotope, cofactor and pH.** *Eur J Biochem* 1997, **249:**607-611.
19. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD *et al.*: **Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*.** *Science* 1996, **273:**1058-1073.
20. Oesterhelt D: **The structure and mechanism of the family of retinal proteins from halophilic archaea.** *Curr Opin Struct Biol* 1998, **8:**489-500.
21. Lanyi JK: **Crystallographic studies of the conformational changes that drive directional transmembrane ion movement in bacteriorhodopsin.** *Biochim Biophys Acta* 2000, **1459:**339-345.
22. Rawal N, Kelkar SM, Altekar W: **Alternative routes of carbohydrate metabolism in halophilic archaebacteria.** *Indian J Biochem Biophys* 1988, **25:**674-686.
23. Tomlinson GA, Strohm MP, Hochstein LI: **The metabolism of carbohydrates by extremely halophilic bacteria: the identification of lactobionic acid as a product of lactose metabolism by *Halobacterium saccharovorum*.** *Can J Microbiol* 1978, **24:**898-903.
24. Suter-Crazzolara C, Kurapkat G: **An infrastructure for comparative genomics to functionally characterize genes and proteins.** *Genome Inform Ser Workshop Genome Inform* 2000, **11:**24-32.
25. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
26. Baliga NS, Pan M, Goo YA, Yi EC, Goodlett DR, Dimitrov K, Shannon P, Aebersold R, Ng WV, Hood L: **Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach.** *Proc Natl Acad Sci U S A* 2002, **99:**14913-14918.
27. Bell SD, Magill CP, Jackson SP: **Basal and regulated transcription in Archaea.** *Biochem Soc Trans* 2001, **29:**392-395.
28. Littlefield O, Korkhin Y, Sigler PB: **The structural basis for the oriented assembly of a TBP/TFB/promoter complex.** *Proc Natl Acad Sci USA* 1999, **96:**13668-13673.
29. Thompson DK, Palmer JR, Daniels C: **Expression and heat-responsive regulation of a TFIIB homologue from the archaeon *Haloferax volcanii*.** *Mol Microbiol* 1999, **33:**1081-1092.
30. Baliga NS, Goo YA, Ng WV, Hood L, Daniels CJ, DasSarma S: **Is gene expression in *Halobacterium* NRC-1 regulated by multiple TBP and TFB transcription factors?** *Mol Microbiol* 2000, **36:**1184-1185.
31. Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH: **New core promoter element in RNA polymerase II-dependent**

transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* 1998, **12:**34-44.
32. Brown JR, Doolittle WF: **Archaea and the prokaryote-to-eukaryote transition.** *Microbiol Mol Biol Rev* 1997, **61:**456-502.
33. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA *et al.*: **Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima.** *Nature* 1999, **399:**323-329.
34. Bouck J, Miller W, Gorrell JH, Muzny D, Gibbs RA: **Analysis of the quality and utility of random shotgun sequencing at low redundancies.** *Genome Res* 1998, **8:**1074-1084.
35. Pearson WR: **Flexible sequence similarity searching with the FASTA3 program package.** *Methods Mol Biol* 2000, **132:**185-219.
36. Smith TF, Waterman MS: **Overlapping genes and information theory.** *J Theor Biol* 1981, **91:**379-380.
37. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
38. Goo YA, Yi E, Tao WA, Baliga NS, Pan M, Goodlett DR, Aebersold R, Hood L, Ng WV: **Proteomic analysis of an extreme halophilic archaeon, *Halobacterium sp.* NRC-1.** *Mol Cell Proteomics* 2003, **2:**506-524.