Université d'Ottawa · University of Ottawa

Comparative genomics: Using the complete genome sequence of *Halobacterium salinarum* NRC-1 to assess gene order conservation within the *Halobacteriaceae*

Julie Chapados

Thesis submitted to the
School of Graduate Studies and Research
University of Ottawa
in partial fulfillment of the requirements for the
M.Sc. degree in the
Ottawa-Carleton Institute of Biology

Ottawa, Ontario, Canada
Jan 2002

ABSTRACT

Comparative mapping and comparative genomics can be used to study genome stability. The genome organization of different strains, species or genera are compared by using physical maps or complete genome sequences. A previous comparative study determined that there was no gene order conservation between *Haloferax volcanii* and *Halobacterium salinarum* at a 15-kbp resolution. My research was designed to assess gene order conservation between certain halophiles at a much finer resolution. This was accomplished by using the only available haloarchaeal genome sequence as a reference. Random gene fragments obtained from other related halophiles were compared to the complete genome sequence and the level of gene order conservation was estimated. The hypothesis of this study was that gene order will not be conserved among haloarchaeal genera. Significant linkage (conserved gene fragments) was detected between *Halobacterium salinarum* NRC-1 and the five analyzed species. The hypothesis was refuted as some gene fragments have maintained similar gene organization in different genera. Furthermore, many genes from the selected halophiles have no homologs in the NRC-1 genome. The use of gene order as a tool to measure phylogeny is discussed.

RÉSUMÉ

La stabilité des génomes peut être étudiée en comparant des cartes ou des séquences génomiques. L'organisation génomique entre souches, espèces et genres peut aussi être comparée en utilisant des cartes physiques ou des séquences génomiques complètes. Une étude faite auparavant a déterminé qu'il n'y avait aucune conservation dans l'ordre des gènes entre *Haloferax volcanii* et *Halobacterium salinarum* à une résolution de 15-kbp. Mon projet de recherche a été conçu afin d'évaluer la conservation dans l'ordre des gènes entre les halophiles à une plus haute résolution. Afin d'accomplir ceci, la seule séquence génomique complète qui est disponible pour un halophile a été utilisé comme point de repère. Des petits groupes de gènes aléatoires, obtenus de divers halophiles, ont été comparé à la séquence génomique complète et le niveau de conservation dans l'ordre des gènes a été estimé. L'hypothèse de cette étude était que l'ordre des gènes ne serait pas conservé entre les genres halophiliques. Un montant significatif de groupes de gènes sont conservés entre *Halobacterium salinarum* NRC-1 et les cinq espèces analysées. L'hypothèse est refutée car certains groupes de gènes ont maintenu leurs organisation dans divers genres halophiliques. En outre, plusieurs gènes isolés des cinq espèces halophiliques n'ont pas d'homologue dans la séquence génomique NRC-1. L'utilisation de l'ordre des gènes comme un outil pour mesurer la phylogénie est discutée

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Robert Charlebois. He never gave up on me even when I wanted to give up on myself. Through his funding from NSERC I was given a great opportunity to continue his work. I have learned many skills and have greatly benefited from his wisdom. I am especially grateful to him for all the time and effort he put into writing the numerous computer programs used in this study. Without him the data analysis would have been impossible.

I would like to thank the members of my research committee: Dr. Guy Drouin, Dr. Scott Findlay and Dr. Cam Wyndham for their helpful comments and suggestions.

I would like to acknowledge Claude Raymond and Vicky Sirois who helped my as temporary lab assistants.

I would like to thank my coworkers: Rob Beiko, Anick De Moors and Dave Sprott. Special thanks go to my family and last but not least to my life partner Pierre Bélanger. I could not have done this without any of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

I. *HALOBACTERIACEAE*

Based on ribosomal RNA phylogeny, living organisms on the tree of life are classified into three domains: Bacteria, Archaea and Eukarya (Woese *et al.*, 1990). Accumulating evidence of horizontal gene transfer occurring between organisms has led many to question the validity of the universal tree. If lateral transfer is in fact wide-spread, it may not be possible to determine the true evolutionary tree (Doolittle, 1999; Charlebois, 1999a).

The Bacteria and the Archaea make up the prokaryotes which characteristically lack a nuclear membrane. They also share numerous features such as genome topology, gene content and genome organization (Keeling *et al.*, 1994). Both domains package their DNA into a nucleoid structure while also having a high gene density along their DNA molecules. Prokaryotic genomes range in size from approximately 0.5 - 10 Mbp and are generally made up of a circular chromosome. However, it has been recently determined that numerous species have unconventional genome topologies (Cole and Saint-Girons, 1999; Casjens, 1998). Linear chromosomes, multiple chromosomes and numerous linear and/or circular plasmids have been observed in many organisms.

Based on 16S rRNA phylogeny, the Archaea cluster into two kingdoms (Aravalli *et al.*, 1998). The Euryarchaeotes are made up of the extreme halophiles, the methanogens and some thermophilic organisms. The

Crenarchaeotes contain only thermophiles. Figure 1 schematically represents the 3 domains of life with emphasis placed on the Archaea.

To date, the family *Halobacteriaceae* is divided into 10 genera: *Haloarcula, Halobacterium, Halobaculum, Halococcus, Haloferax, Halorubrum, Natrialba, Natronobacterium, Natronococcus* and *Natronomonas* (Kamekura, 1998). Figure 2 shows a phylogenetic tree of the *Halobacteriaceae* with emphasis on the particular species that were used in this study. The halophiles are mesophilic (optimum growth temperature ~40^{0}C), aerobic, neutro or alkaliphilic and require very salty (250-300 g/L of NaCl) environments (Oren, 1994). Table 1 lists other physical characteristics of the haloarchaea such as: G+C content, morphology and geographical locations.

Halophile genomes are made up of two DNA fractions (Charlebois, 1999b). The FI fraction (66-68% G+C) consists of 64-89% of the total DNA. The FII fraction (57-60% G+C) is mostly made up of satellite DNA consisting of plasmids and some A+T rich chromosomal islands. Insertion sequences are quite prevalent in the halophile genomes (Charlebois, 1999b). Numerous related families with high copy numbers have been identified in many species. Lateral transfer, occurring through cytoplasmic bridges, has been demonstrated between some halophilic species (Mevarech and Werczberger, 1985; Rosenshine *et al.*, 1989; Tchelet and Mevarech, 1994). There is therefore a great potential for genomic rearrangement and DNA exchange within and between the halophiles.

Figure 1: Schematic representation of the universal tree of life with its three domains: Bacteria, Archaea and Eukarya. The archaeal domain is divided in two kingdoms: Euryarchaeota and Crenarchaeota. This tree, based on small subunit rRNA sequence analysis, has been modified from Haas *et al.* (1996).

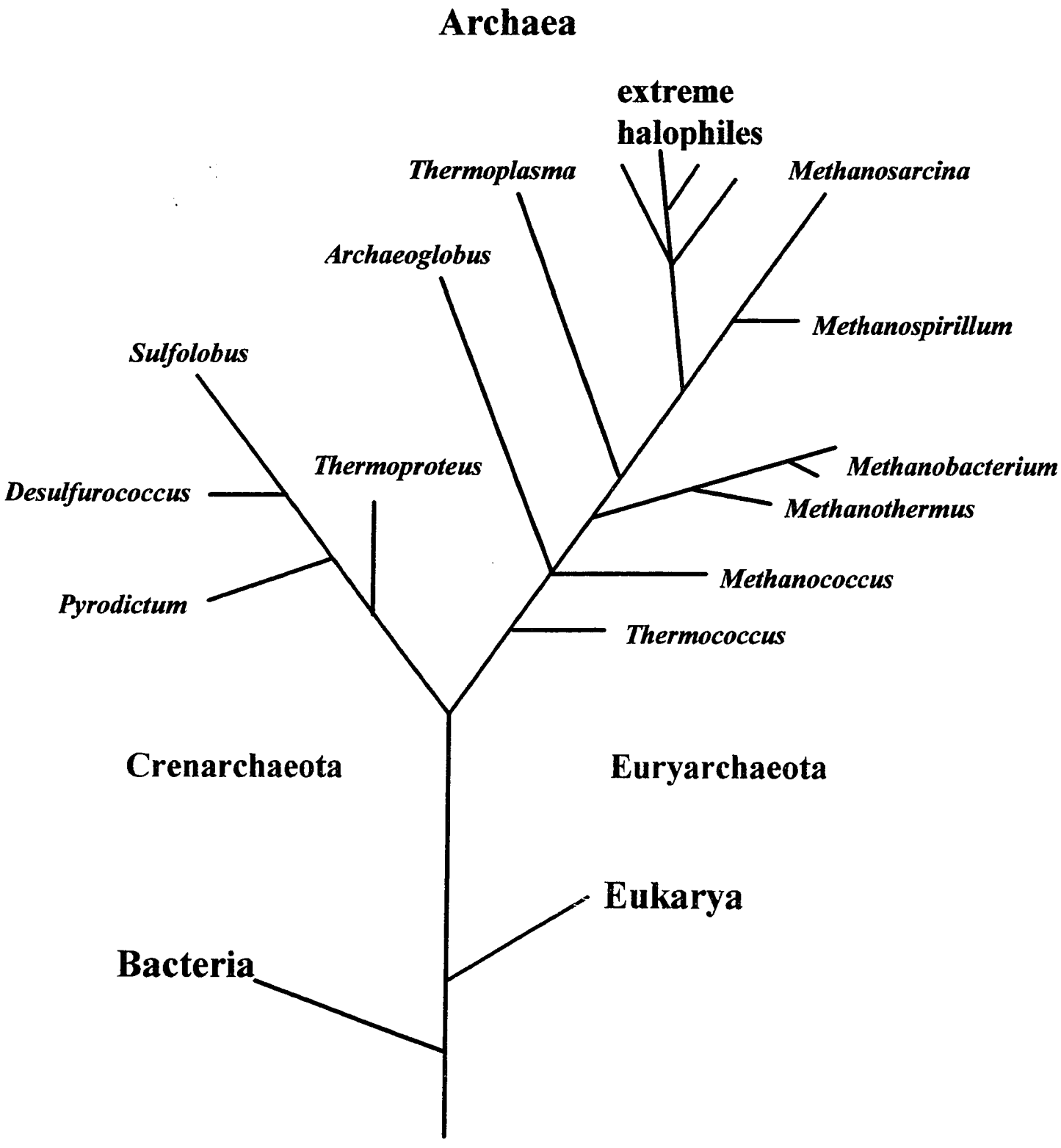

Archaea
extreme
halophiles
Thermoplasma
Methanosarcina
Archaeoglobus
Methanospirillum
Sulfolobus
Thermoproteus
Methanobacterium
Desulfurococcus
Methanothermus
Methanococcus
Pyrodictum
Thermococcus
Crenarchaeota
Euryarchaeota
Eukarya
Bacteria

Figure 2: Phylogenetic tree of the archaeal family *Halobacteriaceae*. The species used in this study are indicated with arrows. This 16S rRNA tree has been modified from Ventosa *et al.* (1999) and does not include species discovered or renamed after 1999. The scale bar represents a 5% nucleotide sequence difference as determined by measuring the length of horizontal lines connecting any two species.

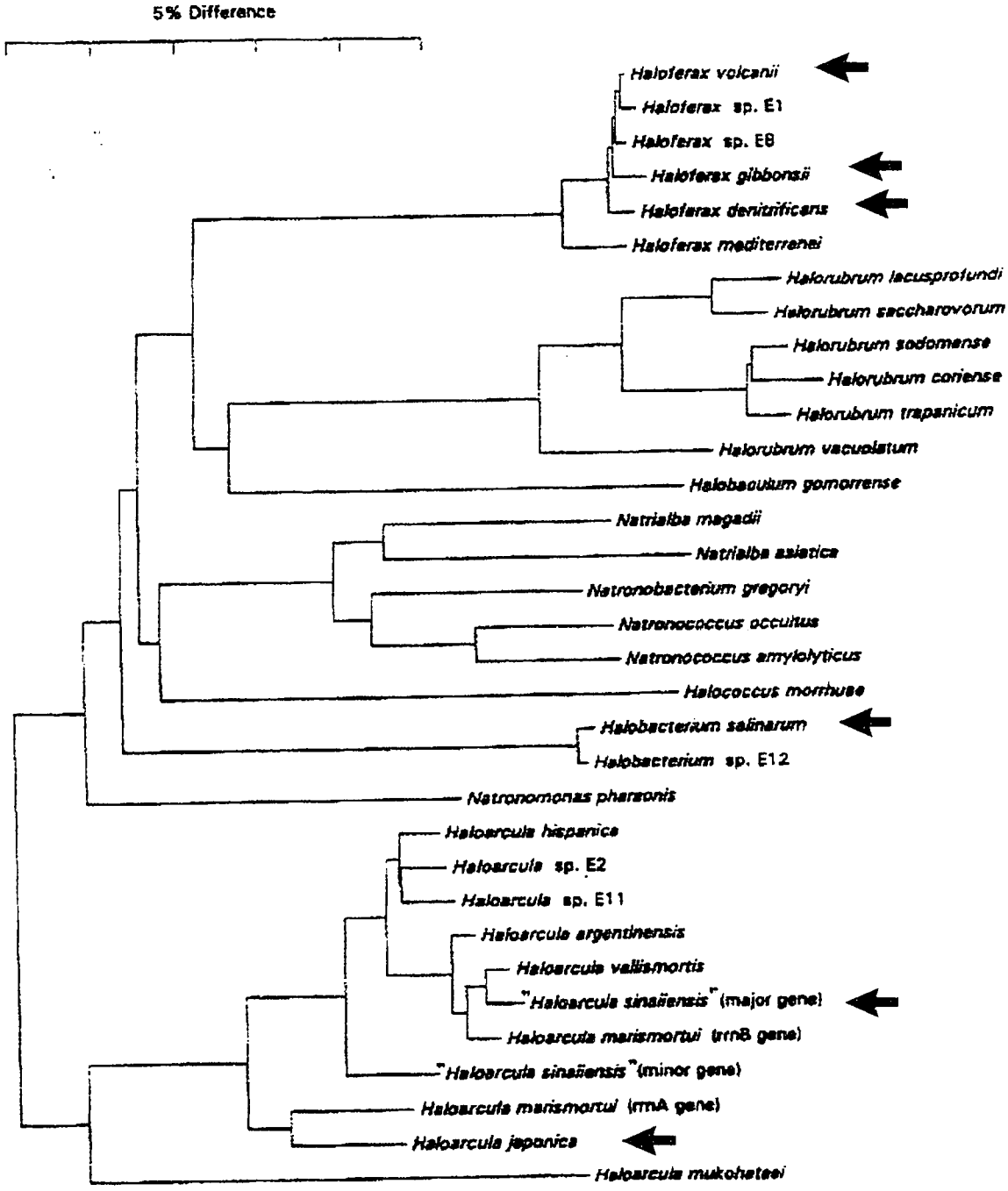

5% Difference
Haloferax volcanii
Haloferax sp. E1
Haloferax sp. E8
Haloferax gibbonsii
Haloferax denitrificans
Haloferax mediterranei
Halorubrum lacusprofundi
Halorubrum saccharovorum
Halorubrum sodomense
Halorubrum coriense
Halorubrum trapanicum
Halorubrum vacuolatum
Halobaculum gomorrense
Natrialba magadii
Natrialba asiatica
Natronobacterium gregoryi
Natronococcus occultus
Natronococcus amylolyticus
Halococcus morrhuae
Halobacterium salinarum
Halobacterium sp. E12
Natronomonas pharaonis
Haloarcula hispanica
Haloarcula sp. E2
Haloarcula sp. E11
Haloarcula argentinensis
Haloarcula vallismortis
"Haloarcula sinaiiensis" (major gene)
Haloarcula marismortui (rrnB gene)
"Haloarcula sinaiiensis" (minor gene)
Haloarcula marismortui (rrnA gene)
Haloarcula japonica
Haloarcula mukohataei

Table 1 : The geographical locations, the morphology and the G+C contents of the extreme halophiles used in this study. The G+C content of *Haloferax denitrificans* was not listed in the referenced paper.
* The G+C content for *Halobacterium salinarum* was calculated as shown in Table 3.

Organism	$G+C_{tot}$ (%)	Geographic location/ Morphology	Reference
Halobacterium salinarum	65.9*	salted hides and fish/ rods	Ng *et al.* (2000) Oren (1994)
Haloferax volcanii	63.4	Dead Sea/ pleomorphic disks	Mullakhanbhai and Larsen (1975)
Haloferax denitrificans	-	saltern, California/ pleomorphic disks	Tomlinson *et al.* (1986)
Haloferax gibbonsii	61.8	solar saltern, Spain/ pleomorphic rods	Juez *et al.* (1986)
Haloarcula japonica	63.3	Japanese saltern soil/ pleomorphic triangles	Takashina *et al.* (1990)
Haloarcula sinaiiensis	68.0	salt pool, Sinai/ box-shaped	Javor *et al.* (1982)

II. GENOMIC STABILITY

II. A. Rearrangements in the genome

Genetic material from the nucleoid can be gained or lost. The copy number, the spacing and the order of genes can also be altered. The prokaryotic genome can be rearranged by events such as duplications, deletions, translocations and inversions (Hughes, 1999; Roth *et al.*, 1996; Weinstock and Lupski, 1998). Rearrangements can be used as a way to adapt to stressful environments (Charlebois and St. Jean, 1995). Recombination is the mechanism used to mediate these rearrangements.

II. A. i. Recombination

Single-stranded DNA (generated during replication, transcription or in DNA breaks) can initiate recombination by pairing with a homologous double helix. Repeated elements like *rrn* genes, Rhs loci, repetitive sequences and transposable elements can also serve as substrates for homologous recombination (Hill *et al.*, 1990; Krawiec and Riley, 1990). Recombination between direct repeats leads to duplications and deletions. A DNA fragment deleted from one location can reintegrate at other loci by recombining with similar copies of the repeat (translocation). Duplication of certain genes can provide advantages. Klappenbach *et al.* (2000) found that the number of rRNA genes correlates with the rate at which bacteria respond to resource availability. Some duplicated gene copies can evolve into new genes by accumulating mutations and eventually providing novel functions. Recombination between

repeats in inverse orientation leads to inversion of the intervening sequence. It was demonstrated that some areas of the *Salmonella* chromosome can tolerate inversions (permissive) while other regions cannot (Mahan *et al.*, 1990; Segall *et al.*, 1988). Inversions with endpoints near the replication terminus (nondivisible zones) are highly disfavored (François *et al.*, 1990).

Recombination can also take place between regions with little or no homology. Illegitimate recombination can occur anywhere in the genome and its frequency increases at replication pauses or arrests (Ehrlich *et al.*, 1993; Michel, 1999). Site specific recombination, on the other hand, occurs at predetermined loci. Phenotype switching uses inversion to add or remove a gene's promoter. These reciprocal ON/OFF switches can initiate antigenic variation and provide phenotypic changes (Hallet and Sherratt, 1997; Dybvig, 1993).

II. A. ii. Horizontal transfer

Prokaryotes are able to transfer genetic information via three mechanisms: transformation, transduction and conjugation (Prozorov, 1999; Matic, 1995). These exchanges of DNA can rearrange the genome and promote rapid adaptation to novel environments. Recently introduced genes can be detected in the genome by their varying GC-content and codon bias. It is estimated that 18% of the *E. coli* genome has been horizontally transferred (Prozorov, 1999). Analyzing completely sequenced genomes can provide valuable information on lateral transfer. Nelson *et al.* (1999) found that 24% of *Thermotoga maritima*'s genes are most similar to hyperthermophilic archaeal

genes. Ruepp *et al.* (2000) discovered that over 250 *Sulfolobus solfataricus* genes are present in the genome of *Thermoplasma acidophilum*.

II. A. iii. Transposable elements (TEs)

Through the process of lateral transfer, transposable elements (insertion sequences and transposons) are introduced into the host's genetic material. TEs can alter the genome in numerous ways (Chalmers and Blot, 1999; Hallet and Sherratt, 1997). They can interrupt gene function or expression if integration occurs in the coding sequence or in the promoter region. TEs have the capacity to bring new genes into the genome; 68% of the insertion sequence elements in *Escherichia coli* are associated with foreign genes (Prozorov, 1999). If multiple copies of an element are present, homologous recombination can occur. The resulting deletions or inversions can significantly contribute to genome plasticity.

II. A. iv. Genome economization

Genome economization refers to the loss of DNA after exposure to an environmental stress. In unfavorable conditions, a faster reproduction rate confers a selective advantage. In response to this selection, chromosomal replication time is reduced by deleting non-essential genetic material. Vellai *et al.* (1999) found that *Escherichia coli* cells submitted to prolonged starvation underwent genome size reduction by accumulating deletions. It has been proposed that environmental stress induces genome economization followed by genome reloading of horizontally acquired genes which could then promote adaptation (Vellai *et al.*, 1999). This process has been observed in the obligate intracellular bacterial parasites *Rickettsiae prowazekii* and *Chlamydiae*

trachomatis (Wolf *et al.*, 1999). These species both underwent drastic genome reductions followed by an uptake of genes from their eukaryotic host. Wolf *et al.* (1999) also found 3 protein families that were clearly exchanged between *Rickettsiae* and *Chlamydiae* probably because of their shared niches.

II. A. v. SOS response

The SOS response accelerates the rate of evolution by increasing mutagenesis and recombination (Matic, 1995). The SOS response is induced when DNA replication is interrupted because of an excessive number of single-stranded DNA lesions (UV damage). If the DNA damage is minimal, excision and homologous recombination will suffice for repairs and the genetic information will be maintained. However, when environmental conditions deteriorate and the cell's life is at risk, SOS mutagenesis is induced. The DNA polymerase will try to repair the numerous lesions by inserting random nucleotides in the gaps without using a template (error-prone repair). The SOS response can thus promote genetic variability and allow adaptation to new environmental conditions (Matic *et al.*, 1995).

II. B. Forces that maintain genome organization

Numerous events can disrupt gene organization. Despite these rearrangements, the gene order of *Escherichia coli* K-12 and *Salmonella typhimurium* LT2 has remained strongly conserved (Sanderson *et al.*, 1999). Therefore, mechanisms must be in place to limit rearrangements and maintain genomic stability under certain conditions.

II. B. i. Chromosome symmetry

To obtain optimal replication, both chromosome arms must be of the same length. The origin of replication and the terminus are thus separated by 180°. Tolerated rearrangements must maintain the required symmetry around the origin/terminus axis (Sanderson *et al.*, 1999; Hughes, 1999). Large inversions which displace *oriC* with respect to *Ter* result in major growth rate reductions.

II. B. ii. Gene orientation

Chromosomal replication and gene transcription are two vital functions. If the DNA replication complex collides with the RNA polymerase while it's transcribing a gene in the opposite orientation, both polymerases could stall (Brewer, 1990). There is thus a tendency to arrange genes so their transcription proceeds in the same direction as chromosomal replication. In nine completely sequenced bacterial genomes, 50-78% of all genes and 83-98% of highly expressed genes (rRNA/tRNA) are transcribed in the same orientation as replication (McLean, 1998).

II. B. iii. Gene dosage

When DNA replication is initiated, the copy number (dosage) of the genes closer to the origin of replication increases when compared to the genes near the terminus (Sanderson *et al.*, 1999; Roth *et al.*, 1996; Segall *et al.*, 1988). Many rounds of replication can be initiated before cell division and these multiple forks can augment (up to four times) the copy number of a locus (Krawiec and Riley, 1990). By introducing *his* genes into various chromosomal locations, Schmid and Roth (1987) found that *his* expression increased with its

proximity to the origin of replication. In a stable genome, each gene and its promoter have been optimized to compensate for gene dosage. Rearrangements can be detrimental since they change the location of genes in relation to *oriC* thereby altering the expression level of certain genes.

II. B. iv. Gene clustering

Genes involved in a pathway were originally scattered around the genome. Eventually they were brought together by rearrangements to form operons (Roth *et al.*, 1996). It has been proposed by Roth *et al.* (1996) that operons are more easily horizontally transferred and are thus able to replace mutant versions of the gene sets in other cells. Once the operon is formed, the individual genes must remain together to retain cotranscriptional regulation. The organization of the genes encoding the L11, L1, L10 and L12 equivalent ribosomal proteins (Shimmin *et al.*, 1989) as well as the spectinomycin operon (Scholzen and Arndt, 1991) have been conserved in both bacteria and archaea.

II. B. v. Gene co-adaptation

Co-adaptation is a process which selects for the conservation of gene order within certain gene fragments. It is thought that the protein products of these genes physically interact. Any rearrangements disrupting this linkage would be detrimental. Dandekar *et al.* (1998) analyzed 9 bacterial and archaeal genomes and found that even in highly shuffled genomes, some gene pairs were always linked.

II. B. vi. **Local genetic context**

Every gene is adapted to its local superhelical context (Charlebois and St. Jean, 1995). Any rearrangement that modifies gene order could thus jeopardize the gene's normal expression pattern. Therefore, selective forces exist to reduce rearrangements and keep genes in their ideal genetic context where they work best (Charlebois and St. Jean, 1995).

Supercoiling, which imparts torsional stress to the DNA, is regulated by the opposing effects of topoisomerases (Drlica *et al.*, 1990). Gyrase (topoisomerase II) introduces negative supercoils whereas topoisomerase I relaxes the DNA. The level of global supercoiling can be affected by the expression of topoisomerases, by nicks in the chromosome, by DNA-binding proteins and by environmental stress (Dorman, 1996; Higgins *et al.*, 1990). Gene expression can also alter the local degree of supercoiling (St. Jean, 1999). During replication, the DNA is tightened ahead of the transcription bubble and relaxed behind it.

II. B. vii. **Mismatch repair**

Comparative analyses indicate that there has been limited gene exchange between *Escherichia coli* and *Salmonella typhimurium* (Matic *et al.*, 1995). Barriers like cell surface incompatibilities and restriction-modification systems are in place to restrict the entry of foreign DNA and prevent the incoming genes from being incorporated into the genome (Prozorov, 1999; Matic, 1995). However, Matic *et al.* (1995) found that the barrier limiting gene transfer between *E. coli* and *S. typhimurium* is genomic sequence divergence.

This is mediated by the mismatch repair system. The horizontally transferred genes are unable to find homologous sequences required for recombination and thus do not persist in the genome. The mismatch repair system plays an important role in maintaining genetic stability.

Promiscuous recombination can also be prevented by the RecQ/topoIII pathway (Wu *et al.*, 1999). RecQ is a helicase which can unwind DNA and topoisomerase III can catalyze the linking of DNA. Working together, RecQ and topoIII can remove an invading strand from the DNA duplex and restore the parental molecules before recombination occurs.

II. B. viii. Chi sites

Cross-over hot-spots instigator (chi) sites in *E. coli* are important for the repair of double-strand breaks during replication. The resulting recombination event acts to maintain genetic information after DNA damage (Roth *et al.*, 1996; Krawiec and Riley, 1990). RecBCD, which has helicase and exonuclease activity, binds the double-strand break. It proceeds towards the origin of replication by unwinding and degrading the DNA. When RecBCD reaches a chi sequence, it pauses then asymmetrically cleaves the site. This creates a single-stranded DNA length which can initiate recombination to refill the gap. The orientation of *chi* sites is essential for the activation of the double-strand repair system. Rearrangements which alter chi site orientation are thus counter-selected.

III. COMPARATIVE ANALYSES OF GENE ORGANIZATION

III. A. Genomic maps

III. A. i. Linkage maps

Genetic maps display the location and order of genes on the chromosome (Cole and Saint Girons, 1994; Fonstein and Haselkorn, 1995). Recombination frequencies are used to measure the distance between these genes. Complete linkage maps are available for *Bacillus subtilis* 168 (Piggot, 1990), *Escherichia coli* K-12 (Bachmann, 1983), *Pseudomonas aeruginosa* (Holloway *et al.*, 1990), *Pseudomonas putida* (Holloway *et al.*, 1990), *Salmonella typhimurium* LT2 (Sanderson and Roth, 1988), *Staphylococcus aureus* (Pattee, 1990) and *Streptomyces coelicolor* A3 (Hopwood and Kieser, 1990). It is unlikely that any new genetic maps will be established for prokaryotes. Their construction is labor-intensive and requires multiple genetic tools. As a substitute to linkage maps, physical maps demonstrating genetic markers are now commonly employed.

III. A. ii. Physical maps

A physical map is a restriction map of the entire genome for one or more enzymes. Two main strategies can be used for map construction (Cole and Saint-Girons, 1994; Fonstein and Haselkorn, 1995; Sanderson *et al.*, 1999). The top-down method is less time consuming but it provides a lower resolution. The intact genomic DNA is isolated and restricted in agarose plugs to prevent shearing. Endonucleases with long or rare restriction sites (i.e. *Not*I, *Avr*II) are

used since they cleave the genome into a limited number of fragments. The *I-CeuI* enzyme is also useful because it recognizes a specific 26 bp sequence located in the *rrl* gene. Since ribosomal RNA molecules are highly conserved among prokaryotes it is possible to determine the number and location of rRNA operons while mapping the genome. Pulsed-field gel electrophoresis (PFGE) is required to separate the mega-base DNA fragments by their size. These restriction fragments are then linked together to form a contiguous map. The bottom-up mapping method is labor-intensive but it provides a finer resolution and can be a useful start-point for a genome sequencing project. A genomic library representing numerous genome equivalents is prepared. These clones are assembled using methods such as chromosome walking and landmark analysis (Charlebois *et al.*, 1991) which detect regions of overlap. Cloned genes are often positioned on complete physical maps to increase their value for comparative mapping. This is done by hybridizing gene probes to Southern or colony blots.

III. A. iii. Restriction skeletons

This method is similar to the top-down approach used in physical mapping. The genome is restricted by a rare-cutting endonuclease (*I-CeuI*, *NotI*) and the DNA is separated by PFGE. The restriction patterns of a large number of strains can be easily compared. Any differences in fragment sizes indicate mutations or rearrangements in the genomes. These genetic polymorphisms are useful in comparative studies as well as in taxonomy and strain typing.

III. A. iv. Chromosomal surveying

Chromosomal surveying is a new comparative method developed to measure the distance between pairs of genes (Stibitz and Yang, 1997). A gene sequence is introduced into a vector containing the I-*Sce*I restriction site. The integration of a pair of these constructs is mediated by homologous recombination with the corresponding chromosomal gene copy. Since the genome lacks I-*Sce*I cleavage sites, an I-*Sce*I digest will produce a fragment whose length represents the distance between these two genes. Using this method, Stibitz and Yang (1997) mapped 11-13 genes in *Bordetella parapertussis* and 5 strains of *Bordetella pertussis*.

III. B. Comparative mapping

Genomic stability can be assessed by comparing the maps of two or more strains, species or genera. Physical maps are very useful for detecting genetic polymorphisms but more detailed maps are needed to compare gene order. The use of combined physical and genetic maps has greatly facilitated genome comparisons. Map resolution can be increased with the addition of genetic markers. In certain species, gene organization is strongly preserved whereas other species have highly rearranged genetic loci. It should be noted, however, that classifying a genome as being conserved or shuffled is quite subjective especially in genomes showing some conservation of gene order accompanied with numerous rearrangements. See Table 2 for a list of organisms whose genomic maps have been compared.

III. C. Genome sequencing

It is now possible to completely sequence an organism's genome. Different strategies have been developed to accomplish this task (Sensen, 1999; Fleischmann *et al.*, 1995). The random whole-genome shotgun sequencing strategy is often used when no detailed physical map is available. Small insert libraries are made and massive sequencing is initiated. These sequence fragments are then assembled into large contigs and directed sequencing is used to close any gaps. In other cases, a set of cosmid clones or restriction fragments covering most of the genome is established. These large ordered DNA segments can then be sheared, sequenced and assembled one at a time. To date, 8 archaeal (Table 3) and 27 bacterial (Table 4) genomes have been completely sequenced. Over 100 other prokaryotes have since been selected for complete genome sequencing.

III. D. Comparative genomics

This type of study usually involves the comparison of two completely sequenced genomes. All the open reading frames (ORFs) are found in both genomes and are then compared to each other and to sequence databases. Such queries can identify shared ORFs, strain specific ORFs, the sequence similarity of common ORFs, the ORF functions (if known) and any ORF fragments with conserved order. However, for most species, complete genomic sequences are not available and most of the genomes that are sequenced are only distantly

Table 2: Summary of published comparative mapping studies. Comparative mapping was used to assess the conservation of gene order between different strains, species or genera. The genomes are denoted as being either conserved or shuffled. This classification is based on the conclusions of each paper and refers to the overall genome organization. Findings which disagree with the major conclusions are indicated in these columns. Also shown is the type of map used in the comparison.

* physical/genetic maps
† restriction maps (restriction skeletons)
‡ chromosomal surveying maps
¥ genetic maps
§ subtractive hybridization

		Map comparison		
Organisms	**Map type**	**conserved**	**shuffled**	**Reference**
Bacillus cereus (4 strains)	*		similar chromosome sizes	Carlson *et al.* (1992)
Bacillus cereus and *Bacillus thuringiensis*	*	some simple rearrangements		Carlson and Kolsto (1993)
Bacillus subtilis (2 strains)	†	1700 kbp inversion		Toda and Itaya (1995)
Bacillus thuringiensis (12 strains)	†		some conserved restriction fragments	Carlson and Kolsto (1993)
Bordetella pertussis (6 strains)	‡		some conservation of gene order in 4 strains	Stibitz and Yang (1997)
Bordetella pertussis and *Bordetella parapertussis*	‡		additional Mbp of DNA in *B. parapertussis*	Stibitz and Yang (1997)
Borrelia afzelii, Borrelia burgdorferi, Borrelia garinii, Borrelia japonica and 4 other *Borrelia* strains	*	different restriction patterns		Casjens *et al.* (1995)
Borrelia afzelii, Borrelia burgdorferi and *Borrelia garinii*	*	no rearrangements detected at map resolution		Ojaimi *et al.* (1994)
Brucella suis (4 biovars)	†		a few conserved restriction sites	Jumas-Bilak *et al.* (1998)
Clostridium perfringens (10 strains)	*	3 variable regions found near *rrn* genes		Canard *et al.* (1992)
Escherichia coli (6 strains)	†	some simple rearrangements		Perkins *et al.* (1993)
Escherichia coli and *Salmonella typhimurium*	¥	inversion in the terminus region		Riley and Sanderson (1990)
Halobacterium salinarum (3 strains)	†	240 kbp variable region		Hackett *et al.* (1994)
Halobacterium salinarum and *Haloferax volcanii*	*		15 kbp map resolution	St. Jean and Charlebois (1996)
Haloferax mediterranei (7 strains)	†	some restriction polymorphisms		López-García *et al.* (1993)
Haloferax mediterranei and *Haloferax volcanii*	*	a few inversions and transpositions		López-García *et al.* (1995)
Helicobacter pylori (30 strains)	†		similar restriction fingerprints in 2 strains	Taylor *et al.* (1992)
Helicobacter pylori (5 strains)	*		a few conserved gene fragments	Jiang *et al.* (1996)
Lactococcus lactis subsp. *lactis* and *cremoris*	*	large inversion covering half the genome		Daveran-Mingot *et al.* (1998)
Legionella micdadei (9 strains)	†		some identical restriction fragments	Lück *et al.* (1995)
Leptospira interrogans (2 serovars)	*		some conserved gene fragments	Zuerner *et al.* (1993)
Methanobacterium wolfei and *Methanobacterium thermoautotrophicum*	*	different restriction patterns		Stettler *et al.* (1995)

Mycoplasma hominis (5 strains)	*	300 kbp inverted fragment; gene organization similar to *Clostridium perfringens*		Ladefoged and Christiansen (1992)
Neisseria gonorrhoeae and *Neisseria meningitidis*	*	complex rearrangements		Bautsch (1998)
Pasterella spp. (63 strains)	†		conserved genome structure in some strains	Liu *et al.* (1999)
Pseudomonas aeruginosa (97 strains)	†	some islands of rearrangement hotspots		Heuer *et al.* (1998)
Pseudomonas aeruginosa isolate C and *Pseudomonas aeruginosa* isolate SG17M	§	identified unique sequences in both genomes		Schmidt *et al.* (1998)
Pseudomonas aeruginosa and *Pseudomonas putida*	¥	a few small inversions		Holloway *et al.* (1990)
Pseudomonas aeruginosa, Pseudomonas fluorescens and *Pseudomonas putida*	*	a few small inversions		Ramos-Díaz and Ramos (1998)
Pseudomonas stutzeri (20 strains)	*		number of rRNA operons is conserved	Ginard *et al.* (1997)
Rhodobacter capsulatus (3 strains)	† (partial genome map)		restriction sites are highly conserved in some regions	Nikolskaya *et al.* (1995)
Salmonella spp. (32 strains)	†	some differences among subgenera		Liu *et al.* (1999)
Salmonella enteritidis and *Salmonella typhimurium*	*	inversion covering the terminus		Liu *et al.* (1993)
Salmonella typhi (127 strains)	†		all strains have 7 genomic segments separated by *rrn* operons	Liu and Sanderson (1996)
Salmonella typhi, Salmonella typhimurium and *Escherichia coli*	*		a total of 7 *rrn* operons are conserved	Liu and Sanderson (1995b)
Salmonella typhimurium (17 strains)	†	limited rearrangements in 2 strains		Liu and Sanderson (1995a)
Streptococcus thermophilus (3 strains)	*	some variable loci		Leblond and Decaris (1998)
Streptomyces ambofaciens (3 strains)	*	polymorphisms at ends of linear chromosome		Leblond and Decaris (1998)
Streptomyces ambofaciens, Streptomyces coelicolor and *Streptomyces lividans*	*	polymorphisms at ends of linear chromosome		Leblond and Decaris (1998)
Thermus thermophilus (2 strains)	*	a few gene rearrangements		Tabata and Hoshino (1996)

related. It is therefore quite difficult to perform comparative studies on closely related species.

New approaches are currently being developed where genomic subsets of multiple species are compared to one genome sequence. This reference genome helps to estimate the relatedness of these organisms and the level of gene order conservation between them. These methods allow closely related genomes, which will probably never be sequenced, to be compared.

III. D. i. Complete genome sequences

Klenk *et al.* (1997) studied the genomes of two archaea: *Archaeoglobus fulgidus* and *Methanococcus jannaschii.* They found that 51% of *A. fulgidus* ORFs have homologs in *M. jannaschii.* Most of these genes are involved in core functions such as replication, transcription, translation and biosynthetic pathways. Only 61% of *A. fulgidus* ORFs have been assigned a putative function.

Read *et al.* (2000) conducted comparative studies between *Chlamydia trachomatis* MoPn, *Chlamydia trachomatis* serovar D, *Chlamydia pneumoniae* AR39 and *Chlamydia pneumoniae* CWL029. The *C. pneumoniae* strains are 99% identical with only 161 variable ORFs. There is similar gene order between both *C. trachomatis* strains except for one 50 kbp plasticity zone. Multiple large inversions between *C. trachomatis* and *C. pneumoniae* are concentrated in variable zones near the terminus.

Tatusov *et al.* (1996) compared the genomes of *Haemophilus influenzae* and *Escherichia coli.* An average identity of 59% was found between the 1128

Table 3: Archaea whose genomes have been completely sequenced. The kingdom is denoted **C** for *Crenarchaeota* or **E** for *Euryarchaeota*. The genome size represents the sum of the chromosome and any sequenced extra-chromosomal elements. Also indicated is the number of plasmids (**P**).
* The G+C content of the entire genome was calculated with:

$$(G+C\%)_{tot} = \frac{(G+C\%)_1\,(size)_1 + (G+C\%)_2\,(size)_2 \ldots}{(size)_1 + (size)_2 \ldots}$$

where the size is in bp.

Organism	Kingdom & Group	Size (Mbp)	$(G+C)_{tot}$ (%)	Reference
Aeropyrum pernix K1	**C:** desulfurococcales	1.67	56.3	Kawarabayasi *et al.* (1999)
Archaeoglobus fulgidus VC-16	**E:** archaeoglobales	2.18	48.5	Klenk *et al.* (1997)
Halobacterium salinarum NRC-1	**E:** halobacteriales	2.57 2 **P**	65.9*	Ng *et al.* (2000)
Methanobacterium thermoautotrophicum ΔH	**E:** methanobacteriales	1.75	49.5	Smith *et al.* (1997)
Methanococcus jannaschii	**E:** methanococcales	1.74 2 **P**	29.5*	Bult *et al.* (1996)
Pyrococcus abyssi	**E:** thermococcales	1.77	44.7	www.ncbi.nlm.nih.gov/
Pyrococcus horikoshii OT3	**E:** thermococcales	1.74	41.9	Kawarabayasi *et al.* (1998)
Thermoplasma acidophilum	**E:** thermoplasmales	1.57	46	Ruepp *et al.* (2000)

Table 4: Bacteria whose genomes have been completely sequenced. The group classifications were obtained from http://www.tigr.org. The genome size represents the sum of the chromosome(s) and any sequenced extra-chromosomal elements. Also indicated is the number of chromosomes and plasmids (**C** and **P**, respectively).

* The G+C content of the entire genome was calculated as shown in Table 3.

† The size of one *Buchnera* plasmid could not be precisely determined because of numerous repeats. The genome size thus represents the chromosome and only one of the two plasmids.

Organism	Group	Size (Mbp)	G+C_{tot} (%)	Reference
Aquifex aeolicus VF5	aquificales	1.59 1 **P**	43.2*	Deckert *et al.* (1998)
Bacillus subtilis 168	firmicutes	4.22	43.5	Kunst *et al.* (1997)
Borrelia burgdorferi B31	spirochaetales	1.52 21 **P**	28.2*	Casjens *et al.* (2000) Fraser *et al.* (1997)
Buchnera sp. APS	gamma proteobacteria	0.65† 2 **P**	26.3*	Shigenobu *et al.* (2000)
Campylobacter jejuni NCTC11168	epsilon proteobacteria	1.64	30.6	Parkhill *et al.* (2000)
Chlamydia pneumoniae CWL029	chlamydiales	1.23	40.6	Read *et al.* (2000)
Chlamydia pneumoniae AR39	chlamydiales	1.23	40.6	Read *et al.* (2000)
Chlamydia trachomatis serovar D	chlamydiales	1.05 1 **P**	41.3*	Stephens *et al.* (1998)
Chlamydia trachomatis MoPn	chlamydiales	1.07 1**P**	40.3*	Read *et al.* (2000)
Deinococcus radiodurans R1	Thermus/ Deinococcus	3.28 2 **C** + 2 **P**	66.6*	White *et al.* (1999)
Escherichia coli K-12 MG1655	gamma proteobacteria	4.64	50.8	Blattner *et al.* (1997)
Haemophilus influenzae KW20	gamma proteobacteria	1.83	38	Fleischmann *et al.* (1995)
Helicobacter pylori 26695	epsilon proteobacteria	1.67	39	Tomb *et al.* (1997)
Helicobacter pylori J99	epsilon proteobacteria	1.64	39	Alm *et al.* (1999)
Mycobacterium tuberculosis H37Rv	firmicutes	4.41	65.6	Cole *et al.* (1998)
Mycoplasma genitalium G-37	firmicutes (mollicutes)	0.58	32	Fraser *et al.* (1995)
Mycoplasma pneumoniae M129	firmicutes (mollicutes)	0.82	40	Himmelreich *et al.* (1996)
Neisseria meningitidis MC58	beta proteobacteria	2.27	51.5	Tettelin *et al.* (2000)
Neisseria meningitidis Z2491	beta proteobacteria	2.18	51.8	Parkhill *et al.* (2000)
Pseudomonas aeruginosa PAO1	gamma proteobacteria	6.26	66.6	Stover *et al.* (2000)
Rickettsia prowazekii Madrid E	alpha proteobacteria	1.11	29.1	Andersson *et al.* (1998)
Synechocystis sp. PCC6803	cyanobacteria	3.57	47.7	Kaneko *et al.* (1996)
Thermotoga maritima MSB8	thermotogales	1.86	46	Nelson *et al.* (1999)
Treponema pallidum Nichols	spirochaetales	1.14	52.8	Fraser *et al.* (1998)
Ureaplasma urealyticum parvum biovar, serovar 3	firmicutes (mollicutes)	0.75	25.5	Glass *et al.* (2000)
Vibrio cholerae El Tor N16961	gamma proteobacteria	4.03 2 **C**	47.5*	Heidelberg *et al.* (2000)
Xylella fastidiosa 9a5c	gamma proteobacteria	2.73 2 **P**	52.6*	Simpson *et al.* (2000)

identified orthologs. Over 70% of orthologs were in short conserved strings, but overall there is no long range colinearity.

Tamames *et al.* (1997) studied neighboring pairs of genes in both these genomes and found a strong tendency for genes with similar function to cluster together. They also discovered that from a total of 800 neighbors in *H. influenzae* and 1971 neighbors in *E. coli*, 291 remained neighbors in both species.

Watanabe (1997) identified all the orthologous genes between *Haemophilus influenzae, Mycoplasma genitalium, Escherichia coli* and *Bacillus subtilis*. These gene sets were examined for any conservation in gene order and several conserved operons were discovered in all species.

Alm *et al.* (1999) compared the genomic sequences of *Helicobacter pylori* J99 and *Helicobacter pylori* 26695 and found similar proteomes (89% of ORFs are shared) despite one plasticity zone containing most of the strain-specific genes. Nine gene fragments of over 50 genes are conserved and 85% of the ORFs have the same neighbors in both genomes. It was determined that only 10 rearrangements are required to transform strain J99 into strain 26695.

Smith *et al.* (1997) sequenced the genome of *Methanobacterium thermoautotrophicum* and conducted a genome comparison with the *Methanococcus jannaschii* genome. Most orthologs in both methanogens are highly divergent since only 19% of *M. thermoautotrophicum* ORFs encode proteins that are over 50% identical to *M. jannaschii*. There is little conservation in gene order: 14% of orthologs have the same neighbor in both genomes.

Forty-six percent of *M. thermoautotrophicum* ORFs have been assigned putative functions and 59 gene families are conserved in both organisms. *M. thermoautotrophicum* has no known insertion sequences (IS) while *M. jannaschii* has 11 members of an IS family.

Makarova (1999) compared the four archaeal genomes of *Methanococcus jannaschii, Methanobacterium thermoautotrophicum, Archaeoglobus fulgidus* and *Pyrococcus horikoshii.* Five hundred and forty three common orthologs were identified most of which are involved in genome replication and expression.

Cole (1998) examined the genome sequences of *Mycobacterium tuberculosis* and *Mycobacterium leprae* and noticed that *M. leprae* contains numerous non-coding sequences (40-50% of the genome) while in *M. tuberculosis* 51% of genes have arisen from gene duplication. The *M. leprae* genome is 1.4 Mbp smaller than the *M. tuberculosis* genome and contains no insertion sequences.

Fraser *et al.* (1995) compared *Mycoplasma genitalium* and *Haemophilus influenzae* and determined that 34% of the *M. genitalium* ORFs have orthologs in *H. influenzae*. Short limited regions of conserved gene order were also found.

Hermann and Reiner (1998) studied the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* and discovered that all the *M. genitalium* genes (470 ORFs) are present in *M. pneumoniae*. Gene order is maintained within 6 large segments but these segments have been rearranged by homologous recombination.

Siefert *et al.* (1997) examined the bacterial and archaeal genomes of *Mycoplasma genitalium, Mycoplasma pneumoniae, Haemophilus influenzae, Synechocystis* spp. and *Methanococcus jannaschii.* Between the bacteria, 16 gene fragments are conserved and 8 gene fragments are conserved in the bacteria and the archaeon.

Tettelin *et al.* (2000) sequenced the genome of *Neisseria meningitidis* MC58 and compared it to the *Neisseria meningitidis* Z2491 genome. One large inversion (955 kbp) was discovered and 91% of strain MC58 ORFs are shared with strain Z2491. Both *N. meningitidis* strains have clustered pilin protein genes. Recombination occurs in these cassettes to promote antigenic variation.

Stover *et al.* (2000) studied the genomic sequences of *Pseudomonas aeruginosa* and *Escherichia coli* and found 33 conserved gene fragments. The *P. aeruginosa* genome (6.3 Mbp) is the largest sequenced to date and shows high genetic complexity.

Fraser *et al.* (1998) compared the genomes of *Treponema pallidum* and *Borrelia burgdorferi.* Forty-six percent of *T. pallidum* ORFs have orthologs in *B. burgdorferi* and 115 genes are spirochete specific. Some shared ORFs (304) are in clusters with conserved gene order, the largest of which contains 27 genes.

III. D. ii. Random genome subsets

Okstad *et al.* (1999) sequenced 43 random DNA segments from *Bacillus cereus* and compared them to the complete genome sequence of *Bacillus subtilis.* The length of the sequenced subsets totaled 80 kbp and 86 ORFs were identified, 54 of which could be assigned putative functions. Seven *B. cereus*

genes had no apparent homologues in other *Bacillus* species. The chromosomal locations of 30 putative orthologues were compared using the *B. cereus* physical map and the *B. subtilis* genome sequence. Some genes showed similar clustering patterns but different chromosomal locations. Overall no common gene arrangements were detected.

Bogush *et al.* (1999) used suppression subtractive hybridization (SSH) to create clone libraries containing unique sequences for *Escherichia coli* and *Salmonella typhimurium*. The identified sequences were then mapped to the complete genome sequence of *E. coli*. About 60% of the SSH clones were present in one genome but absent from the other. Many differential sequences (39% in *E. coli* and 54% in *S. typhimurium*) have unknown functional roles. Such sequence differences could be used to design diagnostic tests to differentiate pathogens and to study the genes involved in unique pathways.

IV. THE STUDY

Numerous genomes are now completely sequenced. Once annotated, these sequences reveal not only gene order but also genome organization. Within the halophilic archaea, only one complete genome sequence is currently available. I have used the complete genome sequence of *Halobacterium salinarum* NRC-1 (Ng *et al.*, 2000) as a reference point to assess gene order conservation between 5 halophilic archaea.

Previous comparative mapping studies have been conducted with the halophiles. The first compared the restriction maps of 3 *Halobacterium salinarum*

strains and found almost complete conservation of the restriction patterns (Hackett *et al.*, 1994). The second looked at the physical/genetic maps of *Haloferax volcanii* and *Haloferax mediterranei* (López-García *et al.*, 1995). The gene order was highly conserved between these two species. The third study compared the cosmid maps of *Haloferax volcanii* and *Halobacterium salinarum* (St. Jean and Charlebois, 1996). The gene order was not conserved between these two genera.

The goal of this study is to determine if any gene fragments are conserved among halophilic genera at a much finer resolution. Instead of looking at the order of a few isolated genes dispersed around the chromosome I will compare random sets of neighboring genes. To accomplish this a novel approach to compare one complete genome sequence with numerous genomic subsets (fragments) was developed.

MATERIALS AND METHODS[1]

I. GROWTH OF *HALOBACTERIACEAE*

I.A. Strains

Haloferax gibbonsii (DSM 4427^T), *Haloferax volcanii* DS2 (DSM 3757^T) and *Haloarcula japonica* (DSM 6131^T) were ordered from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSM, Braunschweig, Germany). *Haloferax denitrificans* (ATCC 35960) and *Haloarcula sinaiiensis* (ATCC 33800^T) were obtained from the American Type Culture Collection (ATCC, Maryland, USA).

I.B. Culture conditions

The ordered strains were rehydrated following DSM and ATCC protocols. One mL of "Happy medium" broth (modified from Mevarech and Werczberger, 1985) was added to each cell pellet for rehydration. This growth medium contains (per litre): 206 g of NaCl, 37 g of $MgSO_4 \cdot 7H_2O$, 3.7 g of KCl, 170 μL of 10 μM $MnCl_2$, 25 mL of 2M Tris•HCl (pH 7.2), 5 g of Bacto tryptone, 3 g of Bacto yeast extract and 5 mL of 10% $CaCl_2 \cdot 2H_2O$. The salts, tryptone/yeast extract and calcium chloride solutions were prepared separately and combined later. If plates were desired, 15 g of agar was added to the salt solution prior to autoclaving.

[1] For recipes of solutions and media see Appendix I

Haloferax volcanii, *Haloferax denitrificans*, *Haloferax gibbonsii*, *Haloarcula japonica* and *Haloarcula sinaiiensis* were streaked onto plates and grown at 37°C for 4-6 days. A single colony was picked for each species and used to inoculate 10 mL of "Happy medium" broth. The liquid cultures were incubated aerobically with shaking for 3-6 days at 37°C.

I.C. Storage

For long term storage, 1 mL aliquots of liquid culture were dispensed into 2 mL screw cap tubes and mixed with 500 μL of 50% glycerol. The cells were then frozen at -80°C.

II. GENOMIC DNA

II.A. Isolation

Fifteen mL of "Happy medium" broth were inoculated with 100 μL of liquid culture and grown for 2-5 days until the medium was quite cloudy. The cultures were centrifuged at 4000 rpm in a Beckman Scientific Megafuge for 10 minutes. The cell pellets were then resuspended in 200 μL of TE (10 mM Tris•HCl pH 8 and 1 mM EDTA). The cells were lysed by adding 15 mL of STE (50 mM Tris•HCl pH 8, 50 mM EDTA and 1% sarkosil). To induce complete cell lysis, the mixtures were incubated at room temperature for 30 minutes with occasional gentle rocking. Three phenol/chloroform extractions were done by adding 20 mL of phenol/chloroform (1:1) followed by gentle mixing for 10 minutes. The tubes

were then centrifuged at 4000 rpm for 10 minutes and the organic phases were discarded. Ethanol precipitations were done by adding 1/10 volume of 3 M sodium acetate pH 5.5 and 2 volumes of 95% ethanol. The total genomic DNAs were washed in 70% ethanol, air dried, resuspended in 1 mL of H_2O and stored at 4°C.

II.B. Digestion

About 2 μg of genomic DNA from each species were mixed with 1 μL of 10X RE buffer, 10U of *Mlu* I and H_2O to a final volume of 10 μl. The genomic digests were incubated at 37°C for at least 5 hours (overnight when possible) followed by an electrophoresis on a 0.8% TAE agarose gel.

III. PREPARATION OF PLASMID LIBRARIES

III.A. Nebulization of genomic DNA

The total genomic DNA isolated from each species was randomly sheared into small fragments. Four μg of genomic DNA were mixed with 250 μL of 50% glycerol. This volume was then brought up to 500 μL with H_2O and transferred to a modified Hospitak "Up-Mist" medication nebulizer. The DNA was nebulized for 13 to 15 seconds using nitrogen at 7 p.s.i.. Once transferred to a new 1.5 mL tube, the sheared DNA was precipitated with 1/10 volume of 3 M sodium acetate and 2 volumes of 95% ethanol. The pellet was washed with 70% ethanol, air dried and resuspended in 20 μL of H_2O . Five μL of nebulized DNA mixed with

5 μL of H_2O were incubated with 1 μL of 0.1 μg/μL RNase for 15 minutes at room temperature. This sample was then run on a 0.8% TAE agarose gel to verify the nebulization. The DNA should produce a smear where the majority of the fragments are between 1 and 5 kbp.

III.B. End repair of DNA fragments

The nebulization process shears the DNA and can create staggered ends. These ends must be repaired before proceeding with the blunt end ligation. Two μL of nebulized DNA (0.4 μg) were combined with 11.7 μL of H_2O, 2 μL of 10X Klenow buffer, 1.5 μL of 1 mM dNTP, 3 U of unmodified T7 DNA polymerase (New England Biolabs) and 2.5 U of Klenow polymerase - sequencing grade (New England Biolabs). This mixture was then incubated at 12°C for 30 minutes. The enzymes were inactivated with a 5 minute incubation at 65°C. An ethanol precipitation was done by adding 10 μL of H_2O, 3 μL of 3 M sodium acetate pH 5.5 and 70 μL of 95% ethanol. The DNA fragments were then washed in 70% ethanol, air dried and resuspended in 13.6 μL of H_2O.

III.C. Phosphorylation of DNA fragments

The 5' ends of the DNA fragments must be phosphorylated since the vector is dephosphorylated to prevent self-ligation. The fragments were mixed with 2 μL of 10 mM ATP, 2 μL of 10X PNK buffer and 24 U of polynucleotide kinase (Boehringer-Mannheim) and incubated at 37°C for 30 minutes. The enzyme was denatured with an incubation at 65°C for 10 minutes.

III.D. Ligation

Ten μL of phosphorylated DNA fragments were combined with 1 μL of 10X ligase buffer without ATP, 0.3 μL of 10 mM ATP, 1.2 μL of H_2O, 1 μL (50 ng) of *Sma* I - cut and dephosphorylated pUC 18 vector DNA (Pharmacia) and 1.5 μL (1.5 U) of T4 DNA ligase (Boehringer-Mannheim). This mixture was incubated at 12°C for 18 hours.

III.E. Transformation

One tube of Epicurian coli XL10-Gold SoloPack supercompetent cells (Stratagene) was thawed on ice. One μL of XL10-Gold β-mercaptoethanol was added and the cells were incubated on ice for 10 minutes with swirling every 2 minutes. Three μL of ligation mix (40 ng of DNA) were added to the tube and the cells were incubated on ice for 30 minutes. The cells were then heat-pulsed at 54°C for 60 seconds and incubated on ice for 2 minutes. One hundred and fifty μL of pre-warmed NZY^+ broth were added and the cells were grown at 37°C for one hour with shaking. Between 5 and 20 μL of cell culture were spread onto YT-amp-Xgal-IPTG plates and incubated overnight at 37°C.

III.F. Storage of plasmid libraries

Using sterile toothpicks, 576 white colonies were picked for each library except for the *Haloferax volcanii* library which exceeded 1900 picked colonies. These clones were grown overnight at 37°C in 100 μL of YT-amp broth

dispensed in sterile 96-well lidded microtiter plates (Nunc). Fifty μL of 75% glycerol were then added to each well and the plates were frozen at -80°C.

IV. PREPARATION OF TEMPLATE FOR SEQUENCING

IV.A. Template isolation

Random subsets of clones were inoculated into 5 mL of YT-amp broth and grown overnight at 37°C with shaking. Plasmid DNA was extracted with the UltraClean mini plasmid prep kit (Mo Bio)[2]. Two mL of cell culture were dispensed in 2 mL tubes and centrifuged for 1 minute at 13 000 rpm in a Eppendorf benchtop centrifuge. The supernatants were discarded and the centrifugation was repeated in the same tubes with another 2 mL of cell culture. The cell pellets were resuspended in 50 μL of Solution 1 (Tris, EDTA and RNase A). Cell lysis was induced by adding 100 μL of Solution 2 (SDS/NaOH). The suspensions were then neutralized with 325 μL of Solution 3 (potassium acetate). The tubes were centrifuged for 1 minute at 13 000 rpm and the clear liquid supernatants were transferred to provided spin filters. The spin filters were centrifuged for 30 seconds at 13 000 rpm and the flow-throughs were discarded. The DNA in the filters was washed with 300 μL of Solution 4 (Tris/NaOH/ethanol) and centrifuged for 30 seconds at 13 000 rpm. The spin filters were transferred to new tubes and the DNA was eluted with 30 μL of H_2O.

[2] Mo Bio labs, when contacted, refused to specify the concentrations of the chemicals used in their kit solutions

IV.B. Size screening

The isolated plasmid samples (0.5 μL, about 2 μg) were digested with 2 U of *Hind* III for 2 hours at 37°C and run out on a 0.8% TAE agarose gel. The sizes of the inserts were determined and clones containing inserts between 1 and 5 kbp were selected for each species. The DNA concentrations and the purity of the plasmid preparations were quantified with a GeneQuant spectrophotometer.

V. TEMPLATE SEQUENCING

V.A. Sequencing parameters

Plasmid DNA was sequenced using the Applied Biosystems PRISM BigDye terminator cycle sequencing ready reaction kit by Perkin Elmer. Four μL of the terminator ready reaction mix, 500 ng of template DNA and 3.2 pmol of M13 universal forward or reverse sequencing primer were mixed in a 0.2 mL PCR tube. The final volume of the reaction was brought up to 20 μL with H_2O. The sequencing reaction went through the following cycles on the Gene Amp PCR system 9600: 96°C for 90 seconds; 25 cycles of: 96°C for 10 seconds, 50°C for 5 seconds, 60°C for 4 minutes; and hold at 4°C.

V.B. Purification of extension products

The unincorporated nucleotides were removed from the sequencing reactions with Centri-Sep spin columns (Princeton Separations). Nine hundred

μL of H_2O were added to the dry gel material (Sephadex G-50) in the column. This matrix was allowed to hydrate at room temperature for at least 2 hours. Excess water was drained out and the remaining interstitial fluid was removed by spinning the columns at 3000 rpm for 2 minutes in a Eppendorf benchtop centrifuge. The sequencing reactions were then loaded onto the hydrated gel matrices and centrifuged at 3000 rpm for 2 minutes. The cleaned flow-throughs were vacuum dried in a Concentrator 5301 (Eppendorf) and stored at -20°C until the sequencing gel could be run. André Bergeron (University of Ottawa) prepared the sequencing gels and ran them on a Applied Biosystems ABI373 sequencer. The trace files (raw data) were transferred to a Sun SPARCstation LX for editing and analysis.

VI. BIOINFORMATIC ANALYSES

VI.A. Sequence database comparisons

Each trace file was examined individually and edited to remove vector sequences and poor quality reads. These processed DNA sequences were then saved in Fasta format text files. All of the sequences were first compared to the publicly available non-redundant NCBI database (www.ncbi.nlm.nih.gov). These BLASTX results helped to ascertain the function of any protein coding DNA sequence by matching them to homologues in other species.

Each sequence was also compared to the complete genome sequence of *Halobacterium salinarum* NRC-1. To accomplish this, a BLAST searchable

database was created by Dr. Robert Charlebois using the "formatdb" program available from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov). These BLASTX searches identified matching protein homologs present in the NRC-1 genome. Each ORF in the sequenced NRC-1 genome has a sequential gene index (gi) number which can be used to mark its chromosomal position. The location of each matching homologue can then be obtained from the genome map. The distance between pairs of homologues is determined by the number of ORFs separating them.

VI.B. Published genome comparisons

VI.B.i. Genomic conservation

This computer simulation was designed to mimic the experimental data. Random gene fragments were selected from one sequenced genome and the gene order of these fragments was compared to the gene order found in another sequenced genome. The data points for the y axis were obtained from www.neurogadgets.com/julie.html. A pair of sequenced genomes was chosen for comparison and a BLASTP cutoff value of 1.0e-5 was selected. This query outputs random gene fragments from the first selected genome. To mimic the experimental data, 50 fragments of 3 genes were randomly picked. These fragments were then compared to the second genome, in order to obtain the proportion of gene fragments with conserved order in both genomes. The computer was instructed to perform this query 100 times and to calculate the mean proportion of conserved gene fragments.

The data points for the x axis were also obtained from the web site www.neurogadgets.com. The "Gene clusters" query allows the user to determine the degree of gene-order conservation within any two completely sequenced genomes. The BLASTP cutoff value of 1.0e-5 was again selected. A list of all the genes that are in clusters in both sequenced genomes was displayed. The number of fragments was then divided by the total number of genes in the first genome to obtain the global proportion of conserved gene fragments

VI.B.ii. Predicting phylogeny

The possible use of gene order as a tool to measure phylogeny was tested on the web site www.neurogadgets.com. The data used on the y axis are identical to the data used on the x axis in the previous computer simulation (section VI.B.i). The data points for the x axis were obtained from the www.neurogadgets.com's "Organism phylogenies" query and a BLASTP cutoff value of 1.0e-5 was selected. A distance matrix based on mean normalized BLASTP scores for all homologous ORFs was generated of all the published genomes. Each genome was thus compared to itself and to every other genome in the database. To determine the sequence similarity between two genomes, one minus the distance was calculated.

RESULTS

This research project required the isolation and identification of random gene fragments from 5 selected haloarchaea. The conservation of gene order within the *Halobacteriaceae* was assessed with the use of the *Halobacterium salinarum* NRC-1 genome sequence. The experimental goals of this study were to grow up these haloarchaeal species, isolate their genomic DNA, create a random plasmid library for each species and sequence both ends of selected clones to identify gene fragments. These fragments were then compared to all available sequences in the NCBI database and more specifically to the genome sequence of *H. salinarum* NRC-1. For a schematic overview of the main steps involved in this novel strategy see Figure 3.

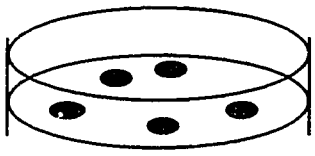
I. Isolation of random gene fragments

The genomic DNA of each haloarchaeon was examined and compared for any obvious signs of cross-contamination. Figure 4 shows an *Mlu* I digest of total DNA isolated from the 5 selected species. This genomic DNA was used to create the plasmid libraries which generated the needed clones. Random clone subsets from each library were grown up for plasmid isolation. The number of plasmid extractions totaled 276 for *Haloferax volcanii*, 120 for *Haloferax denitrificans*, 168 for *Haloferax gibbonsii*, 192 for *Haloarcula japonica* and 120 for *Haloarcula sinaiiensis*.

The clones were screened for insert size and the plasmid DNA was quantified. Only clones containing an insert size between 1 and 5 kbp were chosen to represent a

Figure 3: Schematic overview of the strategy employed to compare random gene fragments from five *Halobacteriaceae* with the complete genome sequence of *Halobacterium salinarum* NRC-1.

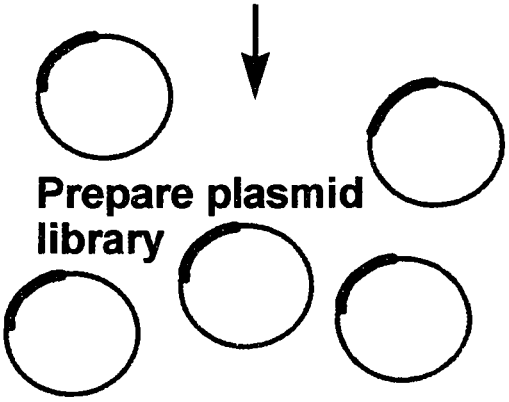
Grow *Hf. volcanii*
Hf. denitrificans
Hf. gibbonsii
Ha. japonica
Ha. sinaiiensis



Pick single colony

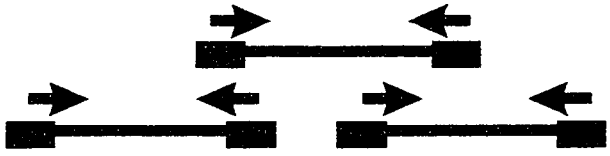
Extract genomic DNA

Shear genomic DNA by nebulization



Prepare plasmid library

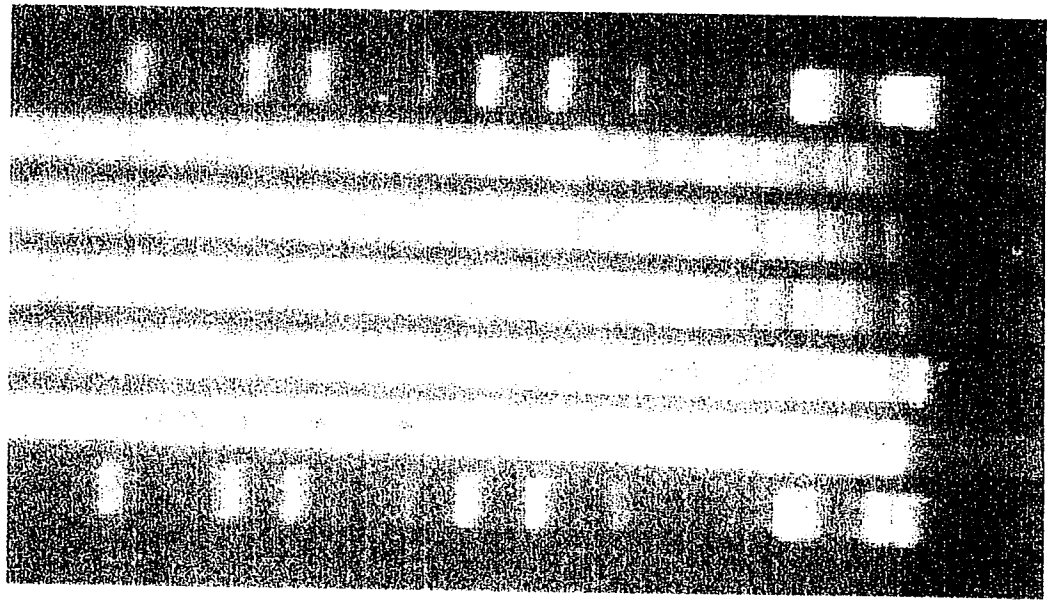
Screen library and isolate clones with inserts between 1- 5 kbp in length



Sequence the ends of 50 inserts using forward and reverse primers and identify genes using BLASTX

Compare sequenced gene pairs with the *H. salinarum* NRC-1 genome and determine if gene clusters are linked

Figure 4: *Mlu* I digest of total genomic DNA isolated from the 5 selected halophiles. The DNA size marker is lambda DNA digested with 3 restriction endonucleases: *Bst*E II, *Xho* I *and Xba* I.


marker
Haloferax volcanii
Haloferax denitrificans
Haloferax gibbonsii
Haloarcula japonica
Haloarcula sinaiiensis
marker

gene fragment. The lengths of all inserts were determined by comparing their migration distance to a DNA size marker (Appendix IIA-IIE).

II. Sequencing the gene fragments

Sequencing was initiated from both ends of the inserts using universal forward and reverse primers. More specifically, 100 clones were sequenced for *Haloferax volcanii*, 51 clones were sequenced for *Haloferax denitrificans*, 44 clones were sequenced for *Haloferax gibbonsii*, 48 clones were sequenced for *Haloarcula japonica* and 51 clones were sequenced for *Haloarcula sinaiiensis*. Each sequenced clone produced two distinct sequences. These two sequences represent the genes at the extremities of the gene fragment. Consequently, the number of genes in a particular fragment is not known. However, based on the insert sizes and on the average prokaryotic gene size of 1 kbp, it is assumed that each insert contains less than 6 genes. The lengths of the sequences vary between 300-750 nucleotides (see Appendix IIA-IIE) based on editing and raw sequence quality.

III. Similarity searches

In order to identify the genes at the extremities of the fragments, each sequence was subjected to a BLASTX search against the NCBI database. These results were used to assign a name and a function to each sequence that found a suitable match. BLASTX similarity searches were also done against the *Halobacterium salinarum*

NRC-1 database (see Appendix IIA-IIE). In each BLASTX search, the nucleotide sequence from each possible reading frame was converted into amino acids and compared to all the known proteins (ORFs) in the database. When performing these similarity searches, only matches with scores lower than $1.0e^{-5}$ were considered. The probability that any of these hits were due to chance increased with the BLASTX score.

IV. Comparative analyses

The BLASTX searches against the NRC-1 database provided the information required for the comparative analyses. The number of gene fragments, the number of sequences, the total number of hits and the number of hits on the different NRC-1 replicons are tabulated for each species in Table 5. By dividing the total number of hits by the number of sequences it is possible to obtain the proportion of sequences that are present in NRC-1. This indicates that between 30 and 40% of the sequences from the 5 selected haloarchaea are not present in NRC-1 and have unknown functions based on NCBI. Also shown in Table 5 and in Figure 5 are the 3 categories of gene fragments: those that matched 0 NRC-1 genes, those that matched 1 NRC-1 gene and those that matched 2 NRC-1 genes.

Some genes were found to be present in multiple halophiles. These were identified by their identical BLASTX results and gi numbers. This list of genes and the halophiles that possess them is shown in Table 6. It should be noted, however, that the information given in this table is purely for interest. The methodology used in this

Table 5: Gene fragment analysis based on the BLASTX searches done on the NRC-1 database. For raw data see Appendix IIA-IIE. All the numbers indicated here are based on unique gene fragments. All duplicate gene fragments were removed prior to the analysis. The hits on the NRC-1 database enumerate the number of sequences that match an NRC-1 gene with a BLASTX score better than 1.0e-5. In some cases, one sequence matched multiple NRC-1 genes. These were counted as one hit. In other cases, both sequences at the ends of one gene fragment matched the same gene. They too were counted as one hit. Also listed are the number of sequences that matched NRC-1 genes located on either plasmid.

Some gene fragments did not match anything in the NRC-1 database. These are referred to as fragments matching 0 NRC-1 genes. The following two rows (indicated in bold) refer to gene fragments matching only one NRC-1 gene. These include the fragments where one sequence matched an NRC-1 gene and the other sequence did not have any significant match. These also include the fragments where the two end sequences are part of the same gene. The last three rows (also indicated in bold) refer to gene fragments matching two different NRC-1 genes. These encompass clones that hit genes located on the same NRC-1 replicon, clones that hit genes located on different NRC-1 replicons, as well as clones that hit 3 distinct NRC-1 genes. Two sequences from a particular fragment are considered to be linked in *Halobacterium salinarum* NRC-1 when the distance (# of genes) between them is less than 6.

	Haloarchaeal species				
	Haloferax volcanii	*Haloferax denitrificans*	*Haloferax gibbonsii*	*Haloarcula japonica*	*Haloarcula sinaliensis*
Unique gene fragments	95	49	43	46	49
Duplicate gene fragments	5	2	1	2	2
Sequences	190	98	86	92	98
Hits (NRC-1 database)	114 (60%)	69 (70%)	53 (62%)	54 (59%)	67 (68%)
Hits on pNRC200	5 (4%)	2 (3%)	2 (4%)	7 (13%)	5 (8%)
Hits on pNRC100	1 (1%)	2 (3%)	1 (2%)	2 (4%)	3 (5%)
Sequences that hit multiple NRC-1 genes	1 (1%)	3 (4%)	2 (4%)	3 (6%)	3 (4%)
Fragments matching 0 NRC-1 genes	16 (17%)	6 (12%)	6 (14%)	11 (24%)	5 (10%)
Fragments matching 1 NRC-1 gene	44 (46%)	17 (35%)	21 (49%)	16 (35%)	21 (43%)
Clones where both sequences match same gene (D=0)	4 (9%)	2 (12%)	3 (14%)	2 (13%)	3 (14%)
Fragments matching 2 NRC-1 genes	35 (37%)	26 (53%)	16 (37%)	19 (41%)	23 (47%)
Linked sequences	18 (51%)	7 (27%)	5 (31%)	14 (74%)	7 (30%)
Hits on different replicons	2 (6%)	2 (8%)	0	2 (11%)	3 (13%)

Figure 5: Visual representation of the 3 gene fragment categories and the corresponding linkage data. This figure uses selected data from Table 5. The upper pie chart shows the percentage of gene fragments that hit either 0, 1 or 2 distinct genes in the BLASTX analysis on the NRC-1 database. Of these 3 categories, only the pie slice representing the fraction of gene fragments that hit 2 NRC-1 genes was used to create the lower pie chart. This linkage analysis can only be done on gene fragments that hit more than one NRC-1 gene. The lower pie chart thus represents the percentage of gene fragments that hit 2 NRC-1 genes which are either linked or rearranged. A gene fragment is considered to be linked if the distance (# of genes) separating the 2 NRC-1 homologs is less than 6. These pie charts were done for each halophile where **A** represents *Haloferax volcanii*, **B** represents *Haloferax denitrificans*, **C** represents *Haloferax gibbonsii*, **D** represents *Haloarcula japonica*, and **E** represents *Haloarcula sinaiiensis*.

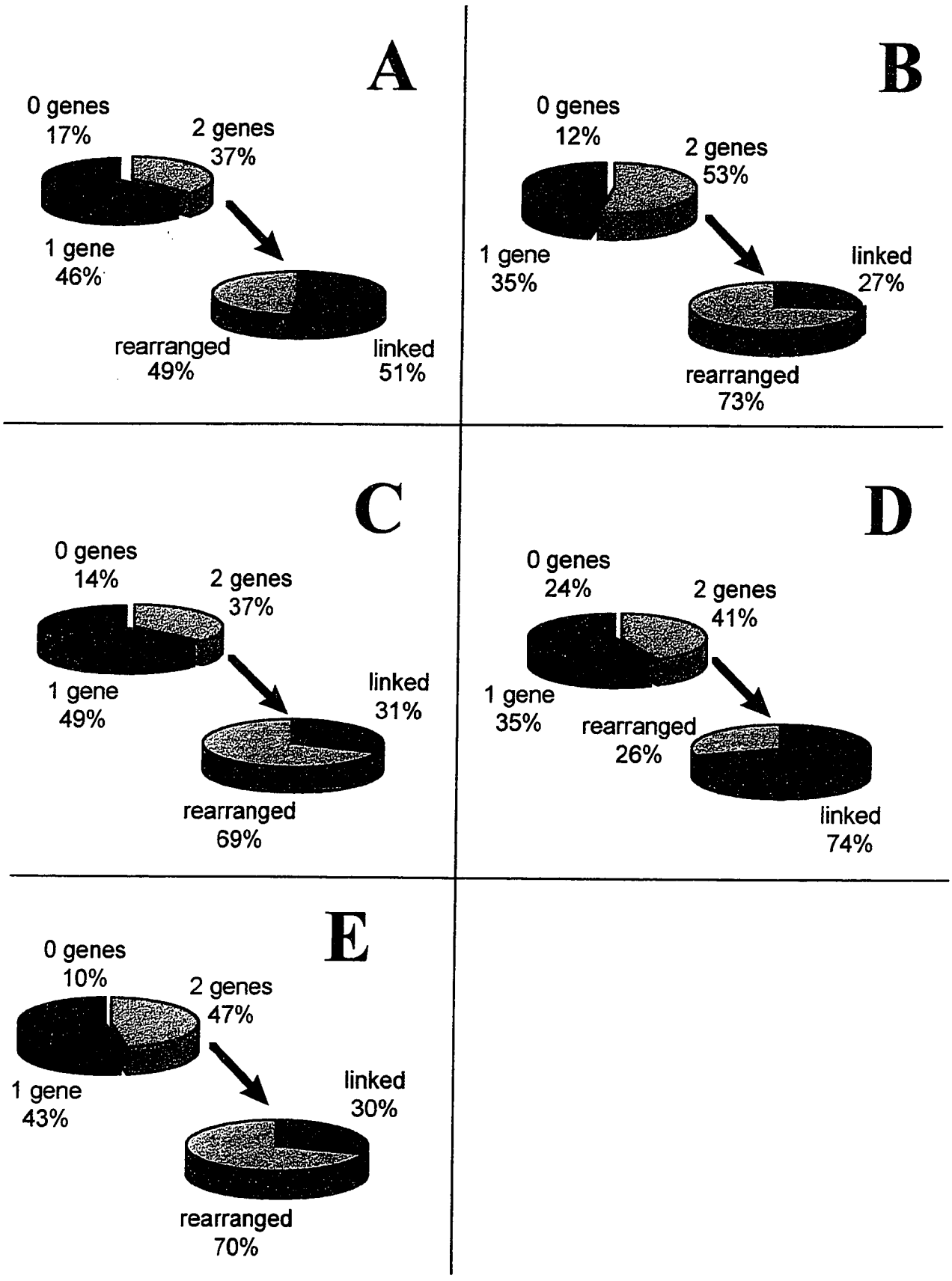

A
0 genes
17%
2 genes
37%
1 gene
46%
rearranged
49%
linked
51%
B
0 genes
12%
2 genes
53%
1 gene
35%
linked
27%
rearranged
73%
C
0 genes
14%
2 genes
37%
1 gene
49%
linked
31%
rearranged
69%
D
0 genes
24%
2 genes
41%
1 gene
35%
rearranged
26%
linked
74%
E
0 genes
10%
2 genes
47%
1 gene
43%
linked
30%
rearranged
70%

Table 6: List of all the NRC-1 genes that were identified in more than one halophile. This information was retrieved from Appendix IIA-IIE. For each gene (identified by its gi number), the corresponding halophile sequences are listed. Where ***Hf. vol.*** represents *Haloferax volcanii*, ***Hf. den.*** represents *Haloferax denitrificans*, ***Hf. gib.*** represents *Haloferax gibbonsii*, ***Ha. jap.*** represents *Haloarcula japonica* and ***Ha. sin.*** represents *Haloarcula sinaiiensis*. The lack of sequences does not mean that a particular halophile does not have this gene, simply that it was not found in this analysis.

		Corresponding homologues				
Genes in NRC-1	**gi**	***Hf. vol.***	***Hf. den.***	***Hf. gib.***	***Ha. jap.***	***Ha. sin.***
*acs*2: acetyl-CoA synthetase	10580552			gib3B7	jap1B10	
*arc*R: transcription regulator	10584354	rDS23B8			rjap2D7	
*cys*S: cysteinyl-tRNA synthetase	10580644	rDS215B4			ujap1D2	
*fab*G: 3-oxoacyl reductase	10580854	rDS26A11			ujap2A5	
*gab*T: gamma-aminobutyrate aminotransferase	10584265	rDS25A4 rDS29A9		ugib3B5		
gsp: general stress protein 69	10581452		rden3A1	rgib3A11		
*hcp*B: halocyanin precursor-like	10581613	rDS25A2			ujap1C8	
*htl*D: Htr-like protein ORF H0337	10584098 10803577		uden2D4			usin1A12
*kin*A2: signal-transducing histidine kinase	10580314	rDS26A1 uDS213A8		rgib1A3		
*mcm*A1: methylmalonyl-CoA mutase, subunit alpha	10580085		uden3A9			usin3C2
*mmd*A: methylmalonyl-CoA decarboxylase, subunit alpha	10581017		den1C8		rjap2B9	
ntp: neutral proteinase	10579693	rDS24B4				usin3D1
ORF H0754 ORF H1607 Vng6090c	10803616 10803687 10584151				rjap2C3	rsin2A2 usin3D5
*prr*C: regulatory protein	10579799			rgib1A12		rsin2A1
*rad*24b: DNA repair protein	10581777	uDS212A3		rgib3A3		
*rsp*A: starvation sensing protein	10580052	rDS24A10				rsin1D10
sdh: succinate dehydrogenase	10581522	uDS215A5				sin3C12
*trm*1: N2,N2-dimethylguanosine tRNA methyltransferase	10581506	uDS22B11	uden3A5			
*trp*D2: phosphoribosyl transferase	10581352	rDS212B8			ujap1C5	
*uvr*C: excision nuclease chain C	10581790		rden2B3			usin1A2
Vng0668c	10580254	rDS213A1		ugib2E7		
Vng0725h	10580304		uden2A3			rsin2A5
Vng0727c	10580306			ugib3A2	rjap2D12	
Vng1092c	10580639	uDS212B3		rgib2E6	rjap1D2	
Vng1982c	10581415	uDS24A3	uden3A2			
Vng2121c	10581536	rDS22B12		rgib1B12		
Vng2199h	10581615	rDS210A9			rjap1C8	
*yfk*N: 2'3'-cyclic nucleotide 2'phosphodiesterase	10580946	rDS24A2	uden2B8			
*yjl*D: NADH dehydrogenase	10580455	rDS24A11	uden2D8			

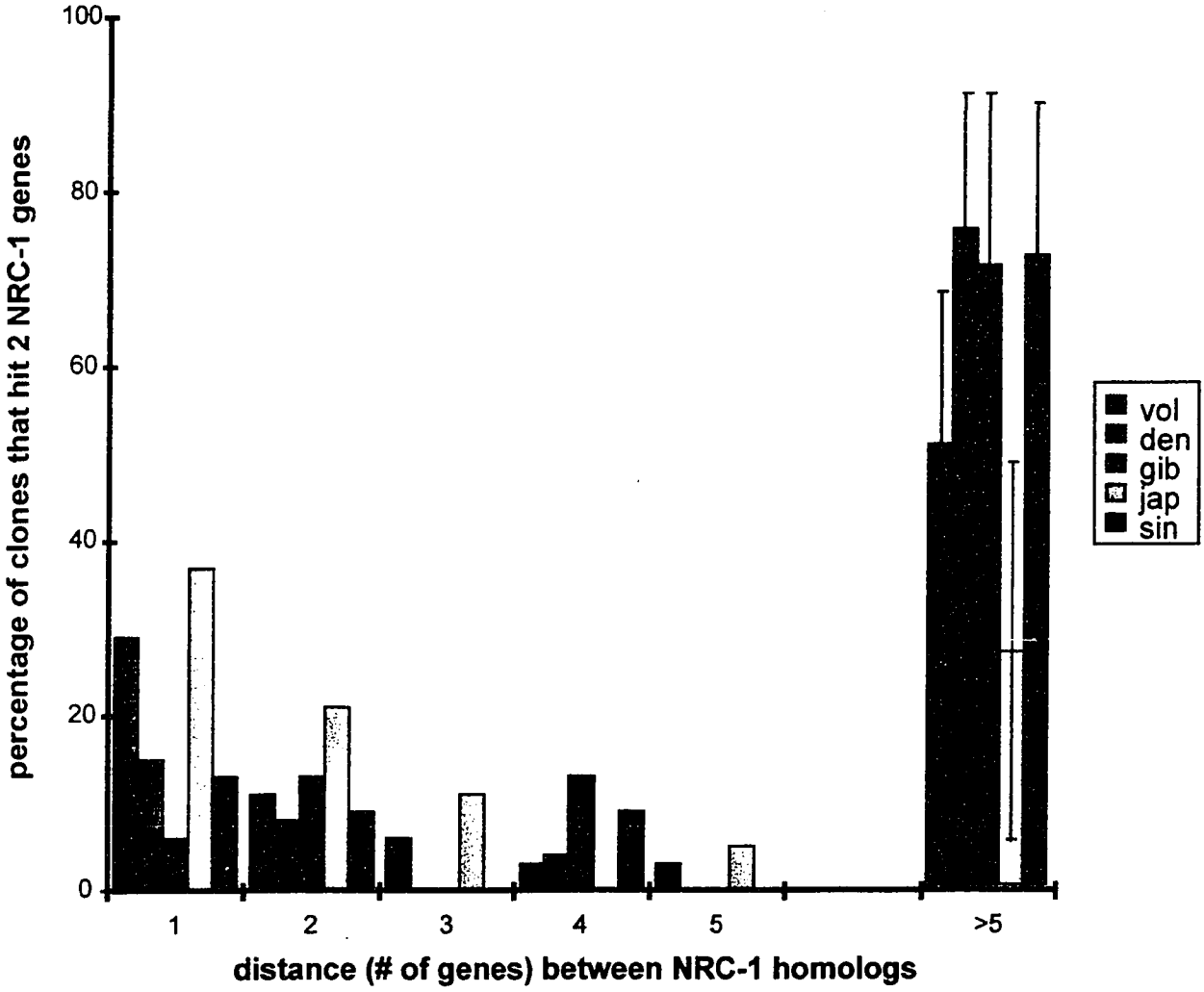
experiment provided completely random gene fragments. Therefore, it cannot be concluded that a gene is missing simply because it was not found in this analysis.

V. Linkage analyses

Only the gene fragments that matched 2 NRC-1 genes were considered for the linkage analysis. A gene fragment is considered to be linked when the distance (# of genes) between the two corresponding NRC-1 homologs is less than 6. Some data from Table 5 were used in Figure 5 to show the proportion of gene fragments matching two NRC-1 genes that are linked as opposed to rearranged. This is schematically represented by the lower pie charts. In the NRC-1 genome, the distance (# of genes) between any 2 genes was derived from the complete and sequential gene index (gi) list.

Some degree of linkage was observed in each halophile (between 27 and 74%). A statistical test was done to determine if the observed linkage was significant. Dr. Robert Charlebois wrote a computer program which shuffled all the BLASTX results and randomly paired them up again. The purpose of this was to determine if any linked gene fragments still existed. The distances (# of genes) between these shuffled gene pairs were then determined. After repeating this test 100 000 times, no linked gene fragments were observed ($p<10^{-5}$). This confirms that the linkage observed in the 5 halophiles is higher than what would be expected under random sampling conditions.

Figure 6: Linkage distribution of the gene fragments that hit 2 NRC-1 genes. The percentage of gene fragments that had a specific distance (1-5 genes) between the corresponding NRC-1 homologues are indicated. All the gene fragments that had a distance of over 5 genes are grouped together since they are considered to be rearranged. The 95% confidence intervals are shown on the rearranged gene fragments.


100
80
60
40
20
0
percentage of clones that hit 2 NRC-1 genes
1
2
3
4
5
>5
distance (# of genes) between NRC-1 homologs
vol
den
gib
jap
sin

V. A. Linkage distribution

The linkage distribution of the gene fragments that hit 2 NRC-1 genes is shown in Figure 6. The percentage of gene fragments that had a specific distance (1-5 genes) between the 2 NRC-1 homologues was calculated for each halophile. All the gene fragments that had a distance of over 5 genes where grouped together since these are considered to be rearranged.

At a preliminary glance it appears as if *Haloarcula japonica* has a more conserved gene order with *Halobacterium salinarum* than the other tested species. Of the gene fragments that matched 2 NRC-1 genes, over 70% were linked in *Halobacterium salinarum* NRC-1. A statistical test was done in order to determine if *H. japonica*'s genome shared more gene fragments with *H. salinarum* NRC-1 than with any of the other 4 halophiles. A computer program was written by Dr. Robert Charlebois to determine the 95% confidence intervals of the rearranged gene fragments (error bars shown in Figure 6). These error bars represent the distribution tails (2.5%) for each species. The distribution was obtained by sampling the data 10 000 times with a chance of getting "x" rearranged fragments. Where "x" is the proportion of rearranged fragments for each halophile (derived from Table 5). A significant difference was discovered between *H. japonica* and 2 other halophiles. The error bars of *Haloferax denitrificans* and *Haloarcula sinaiiensis* do not overlap those of *H. japonica.* This indicates that *H. japonica*'s gene order more closely resembles that of *H. salinarum* than does either of these other 2 halophiles.

V. B. Fragment analysis

BLASTX was used to find homologues in the NRC-1 genome that resembled the query sequences. When using this type of analysis, each query identifies numerous possible matches. It is thus left up to the user to select the most significant hits. Making the distinction between sequences that actually hit 2 different genes or just match multiple genes in a gene family or repetitive elements can get quite ambiguous. As to not bias the results, all matches with the best scores (e-values) were considered. These are referred to in Table 5 as sequences that hit multiple NRC-1 genes.

In some instances, different distances are given for one gene fragment (see Appendix IIA-IIE). These were determined by counting the number of genes between all the homologues identified in NRC-1. Even though multiple matches were found, these sequences were still considered to be part of one gene fragment. The same principle was applied to the linkage distribution analysis (Figure 6). If the distances separating the NRC-1 homologues were all under 6 (see fragment jap2C11 in Appendix IID), only the smallest distance was considered.

VI. Genomic conservation

The goal of this study was to estimate the gene order conservation between one completely sequenced genome and 5 other halophiles. However, unlike the sequenced NRC-1 genome, very little sequence information was obtained from each halophile. This sequence data was used to draw general conclusions about the entire genome. In order to determine if these data can accurately represent an organism, a computer

simulation was done. This simulation, shown in Figure 7, depicts the relationship between genomic conservation and the conservation in gene order of genomic subsets. Completely sequenced genomes were compared in such a way as to mimic the experimental data. A good correlation was obtained suggesting that a subset of gene fragments can be used to estimate gene order conservation between genomes. It should however be noted that the subset of 50 gene fragments in this simulation was randomly chosen 100 times. This simulation thus shows the mean of these 100 trials. This differs from the experimental protocol where only one genomic subset could be used.

VII. Predicting phylogeny

The possible use of gene order as tool to measure phylogeny was also assessed with a computer simulation on completely sequenced genomes. The same gene order data used in the previous simulation were compared to general sequence similarity (see Figure 8). A correlation was observed encouraging the further study of this new tool to study phylogeny.

Figure 7: Relationship between genomic conservation, and the conservation in gene order of genomic subsets, using published genomes. To mimic the experimental data, 50 fragments of 3 genes were randomly picked in the first genome. These fragments were then compared to the second genome, in order to obtain the proportion of gene fragments with conserved order in both genomes. The computer was instructed to perform this query 100 times and to calculate the mean. The standard deviations are indicated. These data were plotted against the gene order conservation between two completely sequenced genomes. A list of all the genes that are in linked in both genomes was displayed. This number was then divided by the total number of genes in the first genome to obtain the global proportion of conserved gene fragments (see Appendix II for raw data).

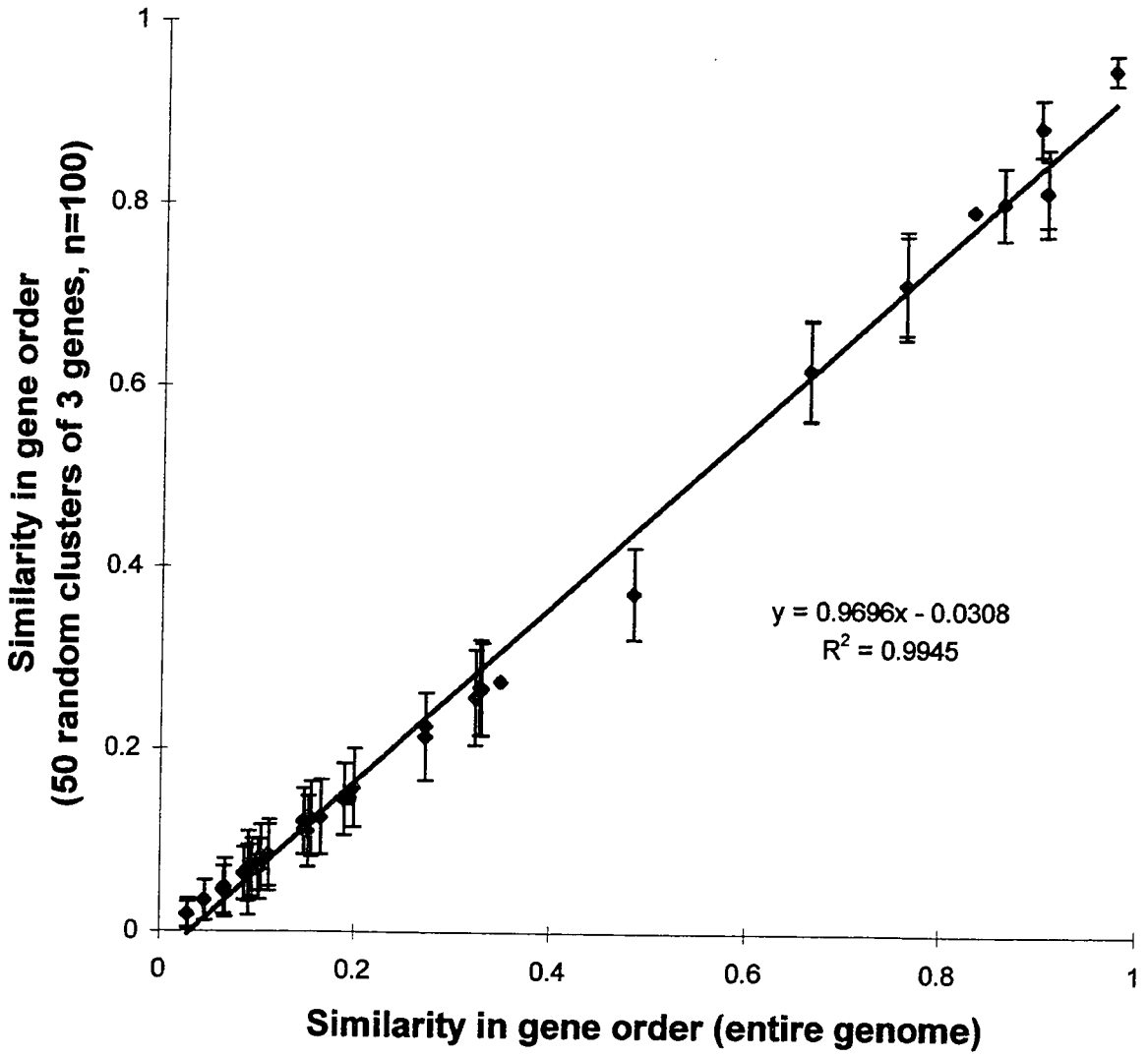
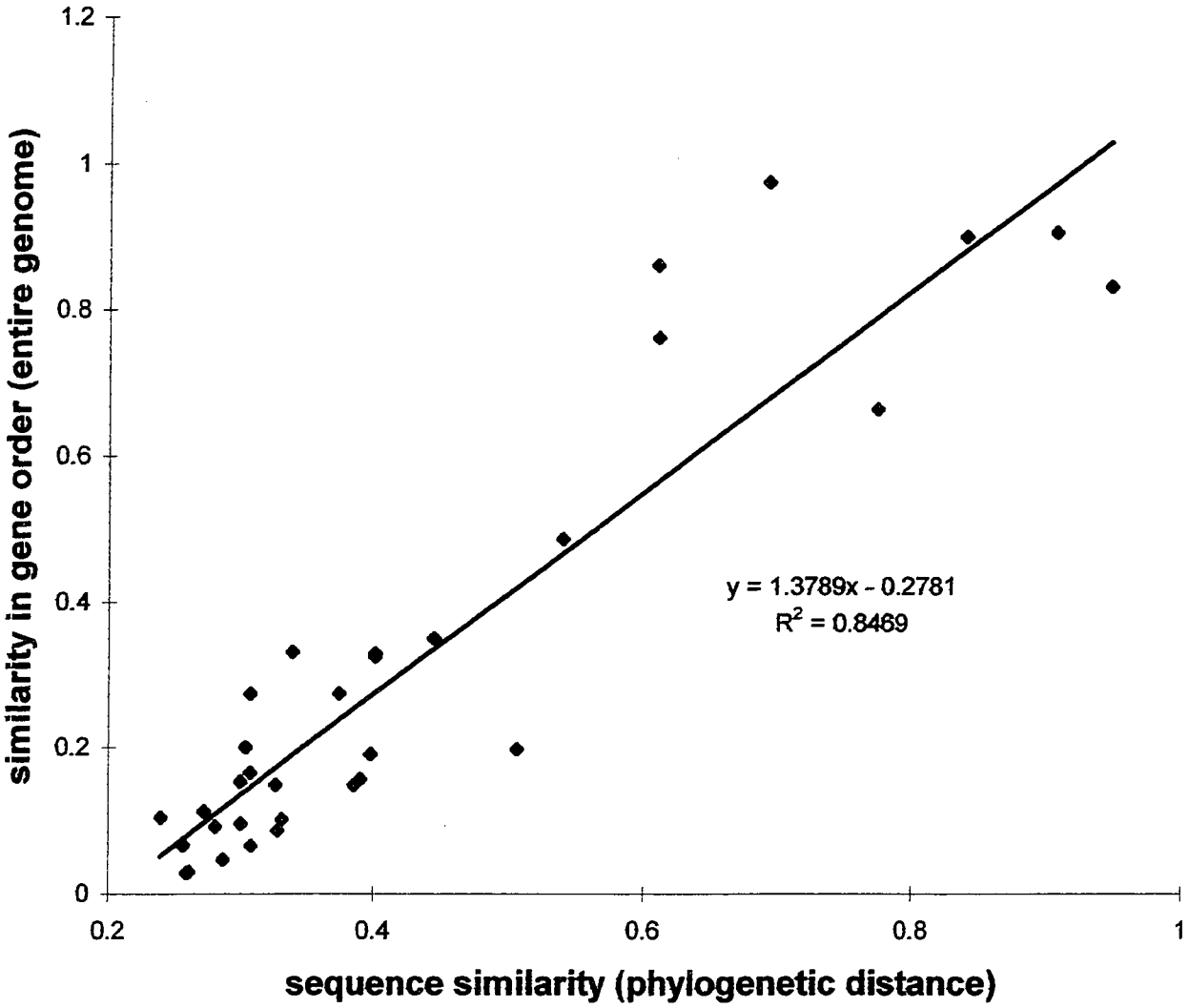

Similarity in gene order
(50 random clusters of 3 genes, n=100)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
y = 0.9696x - 0.0308
R2 = 0.9945
Similarity in gene order (entire genome)

Figure 8: Relationship between genomic conservation, and sequence similarity, using published genomes. The data used on the y axis is identical to the data used on the x axis in Figure 7. These same data were plotted against the data obtained from the distance matrix representing the BLASTP similarities of all the homologous ORFs from every published genome. Each genome was compared to itself and to every other genome in the database. To determine the sequence similarity between two species, one minus the distance was calculated.


similarity in gene order (entire genome)
sequence similarity (phylogenetic distance)
y = 1.3789x - 0.2781
R2 = 0.8469
1.2
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1

DISCUSSION

I. Major findings

The goal of this work was to use the complete genome sequence of *Halobacterium salinarum* NRC-1 to assess gene order conservation within the *Halobacteriaceae*. Gene fragments were randomly selected from the 5 halophiles and compared to the NRC-1 genome. If the gene order between the sequenced gene fragments and the NRC-1 homologues was maintained these fragments were considered to be linked as opposed to rearranged.

Significant linkage was found for all the tested halophiles. This means that more linkage was detected in the gene fragments than would be randomly expected ($p<10^{-5}$). This refutes the hypothesis that no linkage would be found between halophilic genera. Some differences were also found in the amount of linkage within the halophiles. The gene order of *Haloarcula japonica* more closely resembles *Halobacterium salinarum* than either *Haloferax denitrificans* or *Haloarcula sinaiiensis* resemble NRC-1.

Numerous sequences isolated from the various halophiles had no homologues in the NRC-1 database. These could be non-coding and thus would not match specific NRC-1 ORFs. Such sequences might simply be intergenic or could be RNA genes. Numerous genome sequencing projects have however determined that over 90% of most prokaryotic genomes are protein coding. Therefore the possibility that all these sequences are non-coding is unlikely.

What is more likely is that 30-40% of the isolated haloarchaeal sequences are not present in NRC-1 or in any other characterized organism. These sequences should be studied more closely as they could be species specific. Some genes might also possess novel functions and could help characterize unique biochemical pathways or regulatory systems.

II. Using gene order to measure phylogeny

II.A. Gene evolution versus genome evolution

Recently, the idea that the entire genome could be used to determine phylogeny has been gaining popularity. Gene order could be used in this fashion to represent the entire genome in phylogenies. In previous studies, most phylogenetic trees were constructed by comparing the sequence similarity of a single gene. However, with numerous genomic data becoming available, other methods of constructing phylogenies and studying evolution are being developed.

Fitz-Gibbon and House (1999) used genomic data from 11 completely sequenced free-living microorganisms. Using ORF sequence similarity they noted the presence or absence of certain gene families and established a distance matrix. Despite horizontal gene transfer, gene duplication and gene loss, they found a tree topology similar to the rRNA tree.

Huynen and Bork (1998) compared 9 completely sequenced genomes as "bags of genes". Their purpose was not to create a phylogenetic tree but rather

to analyze the conservation and differentiation of patterns between genomes. Orthologues were identified by their significant pairwise sequence similarity. This method only compared genes that were present in all 9 genomes. The conservation of gene order between these orthologues was then assessed. They found that gene order was related to genome divergence time.

Clarke *et al.* (2002; submitted for publication) have also developed a method for building phylogenies using all the genes present in the genome (unpublished data). A distance matrix was constructed using pairwise genome comparisons. Each genome was compared to itself and to every other sequenced genome in the study. All the common genes present in each pair of genomes were compared using BLASTP sequence similarity. This method has the advantage of using a higher number of genes for the comparison. It was found that the Bacteria had a phylogeny that closely resembles the rRNA tree. However, the archaeal phylogeny did not conform to the rRNA tree.

II.B. Does gene order mirror phylogeny?

To address this question, phylogenies built on the basis of gene order must be compared to standard phylogenies. The computer simulation done in Figure 8 seems to indicate a correlation between sequence similarity and the conservation of gene order. However, the data obtained in this study seem to contradict this finding.

It was found that the gene order of *Haloarcula japonica* more closely resembles that of *Halobacterium salinarum* than that of either *Haloferax denitrificans* or *Haloarcula sinaiiensis*. It is surprising that the gene order

between genera is not consistent with established phylogenies. Most halophile phylogenies were not only determined with ribosomal gene sequence similarities but also with morphological and biochemical characteristics such as shape, lipid composition, DNA homology and G+C content (Takashina *et al.*, 1990). It thus seems more likely that the original taxonomic classification is valid. It is more probable that the findings of this study are skewed by the linkage analysis or that in this case, gene order does not mirror phylogeny.

At this junction, there is still no concrete evidence that supports the use of gene order to estimate phylogeny. More studies need to be done before this question can be answered. The increasing number of sequenced genomes might help by increasing the sample size in such comparisons. However, an effective way to calculate phylogenies from gene order data will have to be established.

One area that will need more study is the conversion of gene order data into phylogenetic distances. This problem, unfortunately, cannot be resolved with this study. In order to build a distance matrix, a comparison must be done between all the implicated species. Since only one complete genome sequence was available, each species was compared to *Halobacterium salinarum* NRC-1 and not to each other. Since it was not possible to create a complete distance matrix , a phylogeny could not be built.

III. Ambiguities in comparative studies

Many comparative mapping studies have been done between two or more genomes. It is, however, very difficult to determine if any two maps are conserved or rearranged. No standards or guidelines are in place to assure an accurate conclusion.

Two comparative mapping experiments were done using 30 strains (Taylor *et al.*, 1992) and 5 strains (Jiang *et al.*, 1996) of *Helicobacter pylori*. Despite some conserved gene fragments and similar restriction patterns in a few strains, both studies concluded that the *H. pylori* strains had rearranged genomes. This contradicts the findings of the comparative study done on the two completely sequenced genomes of *H. pylori* strains 26695 and J99 (Alm *et al.*, 1999). It was found that the overall genomic organization, the gene order and the predicted proteomes of the two strains were in fact quite similar. It can thus be concluded that the comparative mapping studies done on this species overestimated the genomic diversity within the strains.

This example demonstrates some of the problems associated with comparative mapping. The data are fragmented and only represent small subsets of the genomes in question. The ideal solution would be to sequence every genome in order to compare their complete sets of ORFs. This, however, is not a feasible endeavor. Some other methods must be used to compare those genomes that will never be sequenced while still providing reliable and reproducible results.

IV. Evaluation of this novel comparative method

IV.A. Strengths

This novel approach would be ideal for assessing gene order conservation between families, genera, species and strains. It would be especially useful in comparing organisms that will probably never be sequenced and are relatively closely related. Previous comparative studies done with the halophiles found highly conserved genomes between strains (Hackett *et al.*, 1994) and species (López-García *et al.*, 1995). However, no gene order conservation was found between the physical maps of *Halobacterium salinarum* and *Haloferax volcanii* at a resolution of 15 kbp (St. Jean and Charlebois, 1996). In the present work, done at a much finer resolution, it was shown that gene order conservation does exist. This proves that in apparently scrambled genomes, some gene fragments remained linked and were previously undetected.

Sequencing the gene fragments was a fairly simple procedure. Since the fragments were all in a simple vector, universal primers could be used. Only two sequencing reactions were done for each fragment - one forward and one reverse. Only the end sequences were needed so this eliminated primer designing steps and contig alignment steps that would have been required to completely sequence each gene fragment.

As mentioned previously, other methods have been developed to compare random genome subsets to a completely sequenced genome.

However, some differences set these studies apart from the one undertaken here. Okstad *et al.* (1999) had the advantage of already having a physical map of *Bacillus cereus*. They simply sequenced the anonymous probes used in the construction of this map. They then proceeded to compare the *B. cereus* physical map to the complete genome sequence of *B. subtilis*. They, however, observed the overall genome conservation as opposed to specific gene fragment conservation.

In their search for unique sequences in *Escherichia coli* and *Salmonella typhimurium*, Bogush *et al.* (1999) didn't really compare gene order conservation. The goal of this study was to find sequences that could differentiate between these two species and since unique sequences were found, only one of the two genomes possessed them. It was thus impossible to compare gene order conservation.

IV.B. Weaknesses

This method will probably lose some of its effectiveness when used on very distantly related organisms. In order to assess gene order conservation, gene fragments matching at least 2 homologues in the reference genome are needed. Some species do not share many genes and some homologous genes have very divergent sequences. It would then become quite difficult and tedious to find such gene fragments for the linkage analysis.

In this approach, only the ends of the gene fragments were sequenced. This means that the actual number of genes per fragment is unknown. It was assumed that each gene fragment contained no more than 5 genes. This

number was chosen for two reasons. First, the gene fragments were size selected to be between 1-5 kbp and second, most prokaryotic genes have an average length of 1 kbp. If the distance between the two NRC-1 homologues was more than 5 the fragment was considered to be rearranged. This logic could have easily skewed the results. In the present linkage analysis some distances were of 5 genes (Figure 6) while other distances were of 6 or 7 genes (Appendix IIA-IIE). It was thus difficult to determine if these fragments were in fact conserved or if they had succumbed to small undetectable rearrangements

In such studies, the sample size is always an issue. It is usually assumed that having more gene fragments in the analysis will improve the reliability of the results. However, a final number of fragments must be chosen and it is often limited by time and money. A choice must be made between the number of species included in the study and the number of fragments sequenced per species. Since a previous comparative study had already been done between *Halobacterium salinarum* and *Haloferax volcanii* (St. Jean and Charlebois, 1996), 100 gene fragments were sequenced for *H. volcanii* in hopes of getting more useful data. Only 50 gene fragments were sequenced for the other 4 halophiles.

V. Future projects

It could be very useful to fully sequence some of the gene fragments isolated from the 5 halophiles. Primers would have to be designed from the 3' ends of both forward and reverse sequences before initiating another round of sequencing. These segments could then be aligned into a contiguous sequence

free of any gaps. If a single contig is not obtained more primers would have to be designed from the new sequences until all the gaps were closed. With a single contig, the number of genes in each fragment would then be known. This would permit a more accurate analysis to be done. The conclusions about each gene fragment would be more reliable since the actual linkage status would be known.

Another option could be to expand the size of the gene fragments used in this comparison. Colony blots would have to be made for all the halophile clones. These would then be probed with some fragment sequence fragments to identify clones extending the contig. By sequencing and aligning these into larger contigs, the gene order of bigger fragments could be assessed. More species could also be added to the comparative analysis in hopes of clarifying the usefulness of gene order in determining phylogeny.

Obtaining such information could allow for more detailed conclusions to be made about the forces rearranging and maintaining the genome.

VI. Conclusions

This study has successfully isolated gene fragments from various halophiles and compared them to the complete genome sequence of *Halobacterium salinarum* NRC-1. The hypothesis that no linkage would be found between genera was refuted since significant linkage was found in all the tested haloarchaea. Numerous sequences were found that had no homologues in

either databases. These might represent species specific genes and could be useful in further studies.

The field of comparative genomics is rapidly growing and a promising future lies ahead. Now that so much genomic data are becoming available, the key will be finding ways to accurately analyze and extract information from these genomes.

REFERENCES

Alm, R.A., Ling, L.-S.L., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L., Carmel, G., Tummino, P.J., Caruso, A., Uria-Nickelsen, M., Mills, D.M., Ives, C., Gibson, R., Merberg, D., Mills, S.D., Jiang, Q., Taylor, D.E., Vovis, G.F., and T.J. Trust. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*. **397**:176-180.

Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C.M., Podowski, R.M., Näslund, A.K., Eriksson, A.-S., Winkler, H.H., and C.G. Kurland. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*. **396**:133-140.

Aravalli, R.N., She, Q., and R.A. Garrett. 1998. Archaea and the new age of microorganisms. *TREE*. **13**:190-194.

Bachmann, B.J. 1983. Linkage map of *Escherichia coli* K-12, Edition 7. *Microbiological Reviews*. **47**:180-230.

Bautsch, W. 1998. Comparison of the genome organization of pathogenic *Neisseriae*. *Electrophoresis*. **19**:577-581.

Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science*. **277**:1453-1462.

Bobovnikova, Y., Ng, W.-L., DasSharma, S., and N.R. Hackett. 1994. Restriction mapping the genome of *Halobacterium halobium* strain NRC-1. *Systematic and Applied Microbiology*.**16**:597-604.

Bogush, M.L., Velikodvorskaya, T.V., Lebedev, Y.B., Nikolaev, L.G., Lukyanov, S.A., Fradkov, A.F., Pliyev, B.K., Boichenko, M.N., Usatova, G.N., Vorobiev, A.A., Andersen, G.L., and E.D. Sverdlov. 1999. Identification and localization of differences between *Escherichia coli* and *Salomonella typhimurium* genomes by suppressive subtractive hybridization. *Molecular and General Genetics*. **262**:721-729.

Brewer, B.J. 1990. Replication and the transcriptional organization of the *Escherichia coli* chromosome. p. 61-83. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.-F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S.M., Weidman, J.F., Fuhrmann, J.L., Nguyen, D., Utterback, T.R., Kelley, J.M., Peterson, J.D., Sadow, P.W., Hanna, M.C., Cotton, M.D., Roberts, K.M., Hurst, M.A., Kaine, B.P., Borodovsky, M., Klenk, H.-P., Fraser, C.M., Smith, H.O., Woese, C.R., and J.C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science.* **273**:1058-1072.

Canard, B., Saint-Joanis, B., and S.T. Cole. 1992. Genomic diversity and organization of virulence genes in the pathogenic anaerobe *Clostridium perfringens. Molecular Microbiology.* **6**:1421-1429.

Carlson, C.R., Gronstad, A., and A.-B. Kolsto. 1992. Physical maps of the genomes of three *Bacillus cereus* strains. *Journal of Bacteriology.* **174**:3750-3756.

Carlson, C.R., and A.-B. Kolsto. 1993. A complete physical map of a *Bacillus thuringiensis* chromosome. *Journal of Bacteriology.* **175**:1053-1060.

Casjens, S. 1998. The diverse and dynamic structure of bacterial genomes. *Annual Review of Genetics.* **32**:339-377.

Casjens, S., Delange, M., Ley III, H.L., Rosa, P., and W.M. Huang. 1995. Linear chromosomes of lyme disease agent spirochetes: genetic diversity and conservation of gene order. *Journal of Bacteriology.* **177**:2769-2780.

Casjens, S., Palmer, N., van Vugt, R., Huang, W.M., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R.J., Haft, D., Hickey, E., Gwinn, M., White, O., and C.M. Fraser. 2000. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the lyme disease spirochete *Borrelia burgdorferi. Molecular Microbiology.* **35**:490-516.

Chalmers, R., and M. Blot. 1999. Insertion sequences and transposons. p.151-169. In R.L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, DC.

Charlebois, R.L. 1999a. Archaea: whose sister lineage? p. 63-76. In R.L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, DC.

Charlebois, R.L. 1999b. Evolutionary origins of the haloarchaeal genome. p. 309-317. In A. Oren (ed.), Microbiology and biogeochemistry of hypersaline environments. CRC Press, Boca Raton, FL.

Charlebois, R.L., Schalkwyk, L.C., Hofman, J.D., and W.F. Doolittle. 1991. Detailed physical map and set of overlapping clones covering the genome of the archaebacterium *Haloferax volcanii* DS2. *Journal of Molecular Biology.* **222**:509-524.

Charlebois, R.L., and A. St. Jean. 1995. Supercoiling and map stability in the bacterial chromosome. *Journal of Molecular Evolution.* **41**:15-23.

Cole, S.T. 1998. Comparative mycobacterial genomics. *Current Opinion in Microbiology.* **1**:567-571.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., and B.G. Barrell. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature.* **393**:537-544.

Cole, S.T., and I. Saint-Girons. 1999. Bacterial genomes - all shapes and sizes. p.35-62. In R.L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, DC.

Cole, S.T., and I. Saint-Girons. 1994. Bacterial genomics. *FEMS Microbiology Reviews.* **14**:139-160.

Dandekar, T., Snel, B., Huynen, M., and P. Bork. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Science.* **23**:324-328.

Daveran-Mingot, M.-L., Campo, N., Ritzenthaler, P., and P. Le Bourgeois. 1998. A natural large chromosomal inversion in *Lactococcus lactis* is mediated by homologous recombination between two insertion sequences. *Journal of Bacteriology.* **180**:4834-4842.

Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M., Huber, R., Feldman, R.A., Short, J.M., Olsen, G.J., and R.V. Swanson. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus. Nature.* **392**:353-358.

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science.* **284**: 2124-2128.

Dorman, C.J. 1996. Flexible response: DNA supercoiling, transcription and bacterial adaptation to environmental stress. *Trends in Microbiology.* **4**:214-216.

Drlica, K., Pruss, G.J., Burger, R.M., Franco, R.J., Hsieh, L.-S., and B.A. Berger. 1990. Roles of DNA topoisomerases in bacterial chromosome structure and function. p. 195-204. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Dybvig, K. 1993. DNA rearrangements and phenotypic switching in prokaryotes. *Molecular Microbiology.* **10**:465-471.

Ehrlich, S.D., Bierne, H., d'Alençon, E., Vilette, D., Petranovic, M., Noirot, P., and B. Michel. 1993. Mechanisms of illegitimate recombination. *Gene.* **135**:161-166.

Fitz-Gibbon, S.T., and C.H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research.* **27**:4218-4222.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J.D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J.M., Weidman, J.F., Philips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O., and J.C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science.* **269**:496-512.

Fonstein, M., and R. Haselkorn. 1995. Physical mapping of bacterial genomes. *Journal of Bacteriology.* **177**:3361-3369.

François, V., Louarn, J., Rebollo, J.-E., and J.-M. Louarn. 1990. Replication termination, nondivisible zones, and structure of the *Escherichia coli* chromosome. p. 351-359. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.-F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M.D., Horst, K., Roberts, K., Hatch, B., Smith, H.O., and J.C. Venter. 1997. Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi. Nature.* **390**:580-586.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Philips, C.A., Merrick, J.M., Tomb, J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison III, C.A., and J.C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium. Science.* **270**:397-403.

Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G., Dodson, R., Gwinn, M., Hickey, E.K., Clayton, R., Ketchum, K.A., Sodergren, E., Hardham, J.M., McLeod, M.P., Salzberg, S., Peterson, J., Khalak, H., Richardson, D., Howell, J.K., Chidambaram, M., Utterback, T., McDonald, L., Artiach, P., Bowman, C., Cotton, M.D., Fujii, C., Garland, S., Hatch, B., Horst, K., Roberts, K., Sandusky, M., Weidman, J., Smith, H.O., and J.C. Venter. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science.* **281**:375-388.

Gaasterland, T. 1999. Archaeal genomics. *Current Opinion in Microbiology.* **2**:542-547.

Ginard, M., Lalucat, J., Tümmler, B., and U. Römling. 1997. Genome organization of *Pseudomonas stutzeri* and resulting taxonomic and evolutionary considerations. *International Journal of Systematic Bacteriology.* **47**:132-143.

Glass, J.I., Lefkowitz, E.J., Glass, J.S., Heiner, C.R., Chen, E.Y., and G.H. Cassell. 2000. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum. Nature.* **407**:757-762.

Haas, E.S., Armbruster, D.W., Vucson, B.M., Daniels, C.J., and J.W. Brown. 1996. Comparative analysis of ribonuclease P RNA structure in archaea. *Nucleic Acids Research.* **24**:1252-1259.

Hackett, N.R., Bobovnikova, Y., and N. Heyrovska. 1994. Conservation of chromosomal arrangement among three strains of the genetically unstable archaeon *Halobacterium salinarum*. *Journal of Bacteriology*. **176**:7711-7718.

Hallet, B., and D.J. Sherratt. 1997. Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. *FEMS Microbiology Reviews*. **21**:157-178.

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Umayam, L., Gill, S.R., Nelson, K.E., Read, T.D., Tettelin, H., Richardson, D., Ermolaeva, M.D., Vamathevan, J., Bass, S., Qin, H., Dragoi, I., Sellers, P., McDonald, L., Utterback, T., Fleishmann, R.D., Nierman, W.C., White, O., Salzberg, S.L., Smith, H.O., Colwell, R.R., Mekalanos, J.J., Venter, J.C., and C.M. Fraser. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*. **406**:477-483.

Herrmann, R., and B. Reiner. 1998. *Mycoplasma pneumoniae* and *Mycoplasma genitalium*: a comparison of two closely related bacterial species. *Current Opinion in Microbiology*. **1**:572-579.

Heuer, T., Bürger, C., and B. Tümmler. 1998. Smith/Birnstiel mapping of genome rearrangements in *Pseudomonas aeruginosa*. *Electrophoresis*. **19**:495-499.

Higgins, C.F., Dorman, C.J., and N. Ni Bhriain. 1990. Environmental influences on DNA supercoiling: a novel mechanism for the regulation of gene expression. p. 421-432. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Hill, C.W., Harvey, S., and J.A. Gray. 1990. Recombination between rRNA genes in *Escherichia coli* and *Salmonella typhimurium*. p. 335-340. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research*. **24**:4420-4449.

Holloway, B.W., Dharmsthiti, S., Krishnapillai, V., Morgan, A., Obeyesekere, V., Ratnaningsih, E., Sinclair, M., Strom, D., and C. Zhang. 1990. Patterns of gene linkages in *Pseudomonas* species. p. 97-105. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Hopwood, D.A., and T. Kieser. 1990. The *Streptomyces* genome. p. 147-162. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Hughes, D. 1999. Impact of homologous recombination on genome organization and stability. p. 109-128. In R.L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, DC.

Huynen, M.A., and P. Bork. 1998. Measuring genome evolution. *PNAS* USA. **95**:5849-5856.

Javor, B., Requadt, C., and W. Stoeckenius. 1982. Box-shaped halophilic bacteria. *Journal of Bacteriology.* **151**:1532-1542.

Jiang, Q., Hiratsuka, K., and D.E. Taylor. 1996. Variability of gene order in different *Helicobacter pylori* strains contributes to genome diversity. *Molecular Microbiology.* **20**:833-842.

Juez, G., Rodriguez-Valera, F., Ventosa, A., and D.J. Kushner. 1986. *Haloarcula hispanica* spec. nov. and *Haloferax gibbonsii* spec. nov., two new species of extremely halophilic archaebacteria. *Systematic and Applied Microbiology.* **8**:75-79.

Jumas-Bilak, E., Michaux-Charachon, S., Bourg, G., O'Callaghan, D., and M. Ramuz. 1998. Differences in chromosome number and genome rearrangements in the genus *Brucella. Molecular Microbiology.* **27**:99-106.

Kamekura, M. 1998. Diversity of extremely halophilic bacteria. *Extremophiles.* **2**:289-295.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., and S. Tabata. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Research.* **3**:109-136.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-No, K., Takahashi, M., Sekine, M., Baba, S.-i., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Nakazawa, H., Takamiya, M., Masuda, S., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K.-i., Kubota, K., Nakamura, Y., Nomura, N., Sako, Y., and H. Kikuchi. 1999. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Research.* **6**:83-101.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S.-i., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K.-i., Yoshizawa, T., Nakamura, Y., Robb, F.T., Horikoshi, K., Masuchi, Y., Shizuya, H., and H. Kikuchi. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Research.* **5**:55-76.

Keeling, P.J., Charlebois, R.L., and W.F. Doolittle. 1994. Archaebacterial genomes: eubacterial form and eukaryotic content. *Current Opinion in Genetics and Development.* **4**:816-822.

Kellenberger, E. 1990. Intracellular organization of the bacterial genome. p. 173-186. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Klappenbach, J.A., Dunbar, J.M., and T.M. Schmidt. 2000. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology.* **66**:1328-1333.

Klenk, H.-P., Clayton, R.A., Tomb, J.-F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., Richardson, D.L., Kerlavage, A.R., Graham, D.E., Kyrpides, N.C., Fleischmann, R.D., Quackenbush, J., Lee, N.H., Sutton, G.G., Gill, S., Kirkness, E.F., Dougherty, B.A., McKenney, K., Adams, M.D., Loftus, B., Peterson, S., Reich, C.I., McNeil, L.K., Badger, J.H., Glodek, A., Zhou, L., Overbeek, R., Gocayne, J.D., Weidman, J.F., McDonald, L., Utterback, T., Cotton, M.D., Spriggs, T., Artiach, P., Kaine, B.P., Sykes, S.M., Sadow, P.W., D'Andrea, K.P., Bowman, C., Fujii, C., Garland, S.A., Mason, T.M., Olsen, G.J., Fraser, C.M., Smith, H.O., Woese, C.R., and J.C. Venter. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature.* **390**:364-370.

Krawiec, S., and M. Riley. 1990. Organization of the bacterial chromosome. *Microbiological Reviews.* **54**:502-539.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.-K., Codani, J.-J., Connerton, I.F., Danchin, A. *et al.* 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*. **390**:249-256.

Ladefoged, S.A., and G. Christiansen. 1992. Physical and genetic mapping of the genomes of five *Mycoplasma hominis* strains by pulsed-field gel electrophoresis. *Journal of Bacteriology*. **174**:2199-2207.

Leblond, P., and B. Decaris. 1998. Chromosome geometry and intraspecific genetic polymorphism in Gram-positive bacteria revealed by pulsed-field gel electrophoresis. *Electrophoresis*. **19**:582-588.

Liu, S.-L., Hessel, A., and K.E. Sanderson. 1993. The *Xba*I-*Bln*I-*Ceu*I genomic cleavage map of *Salmonella enteritidis* shows an inversion relative to *Salmonella typhimurium* LT2. *Molecular Microbiology*. **10**:655-664.

Liu, S.-L., and K.E. Sanderson. 1996. Highly plastic chromosomal organization in *Salmonella typhi*. *PNAS* USA. **93**:10303-10308.

Liu, S.-L., and K.E. Sanderson. 1995a. I-*Ceu* I reveals conservation of the genome of independent strains of *Salmonella typhimurium*. *Journal of Bacteriology*.**177**:3355-3357.

Liu, S.-L., and K.E. Sanderson. 1995b. Rearrangements in the genome of the bacterium *Salmonella typhi*. *PNAS* USA. **92**:1018-1022.

Liu, S.-L., Schryvers, A.B., Sanderson, K.E., and R.N. Johnston. 1999. Bacterial phylogenetic clusters revealed by genome structure. *Journal of Bacteriology*. **181**:6747-6755.

López-García, P., Abad, J.P., and R. Amils. 1993. Genome analysis of different *Haloferax mediterranei* strains using pulsed-field gel electrophoresis. *Systematic and Applied Microbiology*. **16**:310-321.

López-García, P., St. Jean, A., Amils, R., and R.L. Charlebois. 1995. Genomic stability in the archaeae *Haloferax volcanii* and *Haloferax mediterranei*. *Journal of Bacteriology*. **177**:1405-1408.

Lück, P.C., Helbig, J.H., Drasar, V., Bornstein, N., Fallon, R.J., and M. Castellani-Pastoris. 1995. Genomic heterogenicity amongst phenotypically similar *Legionella micdadei* strains. *FEMS Microbiology Letters.* **126**:49-54.

Mahan, M.J., Segall, A.M., and J.R. Roth. 1990. Recombination events that rearrange the chromosome: barriers to inversion. p. 341-349. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I., and E.V. Koonin. 1999. Comparative genomics of the archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Research.* **9**:608-628.

Matic, I. 1995. Les mécanismes du contrôle des échanges génétiques interspécifiques et de la variabilité génétique chez les bactéries. *Bulletins de l'Institut Pasteur.* **93**:187-219.

Matic, I., Rayssiguier, C., and M. Radman. 1995. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. *Cell.* **80**:507-515.

McLean, M.J., Wolfe, K.H., and K.M. Devine. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *Journal of Molecular Evolution.* **47**:691-696.

Mevarech, M., and R. Werczberger. 1985. Genetic transfer in *Halobacterium volcanii. Journal of Bacteriology.* **162**:461-462.

Michel, B. 1999. Illegitimate recombination in bacteria. p. 129-150. In R.L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, DC.

Mullakhanbhai, M.F., and H. Larsen. 1975. *Halobacterium volcanii* spec. nov., a Dead Sea *Halobacterium* with a moderate salt requirement. *Archives of Microbiology.* **104**:207-214.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C., and C.M. Fraser. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*. **399**:323-329.

Ng, W.V., Kennedy, S.P., Mahairas, G.G., Berquist, B., Pan, M., Shukla, H.D., Lasky, S.R., Baliga, N.S., Thorsson, V., Sbrogna, J., Swartzell, S., Weir, D., Hall, J., Dahl, T.A., Welti, R., Goo, Y.A., Leithauser, B., Keller, K., Cruz, R., Danson, M.J., Hough, D.W., Maddocks, D.G., Jablonski, P.E., Krebs, M.P., Angevine, C.M., Dale, H., Isenbarger, T.A., Peck, R.F., Pohlschroder, M., Spudich, J.L., Jung, K.-H., Alam, M., Freitas, T., Hou, S., Daniels, C.J., Dennis, P.P., Omer, A.D., Ebhardt, H., Lowe, T.M., Liang, P., Riley, M., Hood, L., and S. DasSarma. 2000. Genome sequence of *Halobacterium* species NRC-1. *PNAS* USA. **97**:12176-12181.

Nikolskaya, T., Fonstein, M., and R. Haselkorn. 1995. Alignment of a 1.2-Mb chromosomal region from three strains of *Rhodobacter capsulatus* reveals a significantly mosaic structure. *PNAS* USA. **92**:10609-10613.

Ojaimi, C., Davidson, B.E., Saint Girons, I., and I.G. Old. 1994. Conservation of gene arrangement and an unusual organization of rRNA genes in the linear chromosomes of the Lyme disease spirochaetes *Borrelia burgdorferi, B. garinii* and *B. afzelii*. *Microbiology*. **140**:2931-2940.

Okstad, O.A., Hegna, I., Lindbäck, T., Rishovd, A.-L., and A.-B. Kolsto. 1999. Genome organization is not conserved between *Bacillus cereus* and *Bacillus subtilis*. *Microbiology*. **145**:621-631.

Oren, A. 1994. The ecology of the extremely halophilic archaea. *FEMS Microbiology Reviews*. **13**:415-440.

Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., Davies, R.M., Davis, P., Devlin, K., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Leather, S., Moule, S., Mungall, K., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., Simmonds, M., Skelton, J., Whitehead, S., Spratt, B.G., and B.G. Barrell. 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*. **404**:502-505.

Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A.V., Moule, S., Pallen, M.J., Penn, C.W., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., van Vliet, A.H.M., Whitehead, S., and B.G. Barrell. 2000. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature.* **403**:665-668.

Pattee, P.A. 1990. Genetic and physical mapping of the chromosome of *Staphylococcus aureus* NCTC 8325. p. 163-169. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Perkins, J.D., Heath, J.D., Sharma, B.R., and G.M. Weinstock. 1993. *Xba*I and *Bln*I genomic cleavage maps of *Escherichia coli* K-12 strain MG1655 and comparative analysis of other strains. *Journal of Molecular Biology.* **232**:419-445.

Piggot, P.J. 1990. Genetic map of *Bacillus subtilis* 168. p. 107-145. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Prozorov, A.A. 1999. Horizontal gene transfer in bacteria: laboratory modeling, natural populations, and data from genome analysis. *Microbiology.* **68**:551-564.

Ramos-Díaz, M.A., and J.L. Ramos. 1998. Combined physical and genetic map of the *Pseudomonas putida* KT2440 chromosome. *Journal of Bacteriology.* **180**:6352-6363.

Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S.L., Eisen, J., and C.M. Fraser. 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Research.* **28**:1397-1406.

Riley, M., and K.E. Sanderson. 1990. Comparative genetics of *Escherichia coli* and *Salmonella typhimurium.* p. 85-95. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Rosenshine, I., Tchelet, R., and M. Mevarech. 1989. The mechanism of DNA transfer in the mating system of an archaebacterium. *Science.* **245**:1387-1389.

Roth, J.R., Benson, N., Galitski, T., Haack, K., Lawrence, J.G., and L. Miesel. 1996. Rearrangements of the bacterial chromosome: formation and applications. p. 2256-2276. In F.C. Neidhardt, R. Curtiss III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Resnikoff, M. Riley, M. Schaechter, and H.E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. American Society for Microbiology, Washington, DC.

Ruepp, A., Graml, W., Santos-Martinez, M.-L., Koretke, K.K., Volker, C., Mewes, H.W., Frishman, D., Stocker, S., Lupas, A.N., and W. Baumeister. 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*. **407**:508-513.

Sanderson, K.E., McClelland, M., and S.-L. Liu. 1999. "Stable" genomes. p. 217-233. In R.L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, DC.

Sanderson, K.E., and J.R. Roth. 1988. Linkage map of *Salmonella typhimurium*, Edition VII. *Microbiological Reviews*. **52**:485-532.

Schmid, M.B., and J.R. Roth. 1987. Gene location affects expression level in *Salmonella typhimurium*. *Journal of Bacteriology*. **169**:2872-2875.

Schmidt, K.D., Schmidt-Rose, T., Römling, U., and B. Tümmler. 1998. Differential genome analysis of bacteria by genomic subtractive hybridization and pulsed field gel electrophoresis. *Electrophoresis*. **19**:509-514.

Scholzen, T., and E. Arndt. 1991. Organization and nucleotide sequence of ten ribosomal protein genes from the region equivalent to the spectinomycin operon in the archaebacterium *Halobacterium marismortui*. *Molecular and General Genetics*. **228**:70-80.

Segall, A., Mahan, M.J., and J.R. Roth. 1988. Rearrangement of the bacterial chromosome: forbidden inversions. *Science*. **241**:314-318.

Sensen, C.W. 1999. Sequencing microbial genomes. p. 1-9. In R.L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, DC.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*. **407**:81-86.

Shimmin, L.C., Newton, C.H., Ramirez, C., Yee, J., Downing, W.L., Louie, A., Matheson, A.T., and P.P. Dennis. 1989. Organization of genes encoding the L11, L1, L10, and L12 equivalent ribosomal proteins in eubacteria, archaebacteria, and eucaryotes. *Canadian Journal of Microbiology.* **35**:164-170.

Siefert, J.L., Martin, K.A., Abdi, F., Widger, W.R., and G.E. Fox. 1997. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *Journal of Molecular Evolution.* **45**:467-472.

Simpson, A.J.G., Reinach, F.C., Arruda, P., Abreu, F.A., Acencio, M., Alvarenga, R., Alves, L.M.C., Araya, J.E., Baia, G.S., Baptista, C.S., Barros, M.H., Bonaccorsi, E.D., Bordin, S., Bové, J.M., Briones, M.R.S., Bueno, M.R.P., Camargo, A.A., Camargo, L.E.A., Carraro, D.M., Carrer, H., Colauto, N.B., Colombo, C., Costa, F.F., Costa, M.C.R., Costa-Neto, C.M., Coutinho, L.L., Cristofani, M., Dias-Neto, E., Docena, C., El-Dorry, H., Facincani, A.P., Ferreira, A.J.S. *et al.* 2000. The genome sequence of the plant pathogen *Xylella fastidiosa. Nature.* **406**:151-156.

Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Safer, H., Patwell, D., Prabhakar, S., McDougall, S., Shimer, G., Goyal, A., Pietrokovski, S., Church, G.M., Daniels, C.J., Mao, J.-I., Rice, P., Nölling, J., and J.N. Reeve. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *Journal of Bacteriology.* **179**:7135-7155.

Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q., Koonin, E.V., and R.W. Davis. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis. Science.* **282**:754-759.

Stettler, R., Erauso, G., and T. Leisinger. 1995. Physical and genetic map of the *Methanobacterium wolfei* genome and its comparison with the updated genomic map of *Methanobacterium thermoautotrophicum* Marburg. *Archives of Microbiology.* **163**:205-210.

Stibitz, S., and M.-S. Yang. 1997. Genomic fluidity of *Bordetella pertussis* assessed by a new method for chromosomal mapping. *Journal of Bacteriology.* **179**:5820-5826.

St. Jean, A. 1999. Local genetic context, supercoiling, and gene expression. p. 203-215. In R.L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, DC.

St. Jean, A., and R.L. Charlebois. 1996. Comparative genomic analysis of the *Haloferax volcanii* DS2 and *Halobacterium salinarium* GRB contig maps reveals extensive rearrangement. *Journal of Bacteriology.* **178**:3860-3868.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S.L., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K.-S., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E.W., Lory, S., and M.V. Olson. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature.* **406**:959-964.

Tabata, K., and T. Hoshino. 1996. Mapping of 61 genes on the refined physical map of the chromosome of *Thermus thermophilus* HB27 and comparison of genome organization with that of *T. thermophilus* HB8. *Microbiology.* **142**:401-410.

Takashina, T., Hamamoto, T., Otozai, K., Grant, W.D., and K. Horikoshi. 1990. *Haloarcula japonica* sp. nov., a new triangular halophilic archaebacterium. *Systematic and Applied Microbiology.* **13**:177-181.

Tamames, J., Casari, G., Ouzounis, C., and A. Valencia. 1997. Conserved clusters of functionally related genes in two bacterial genomes. *Journal of Molecular Evolution.* **44**:66-73.

Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M., Rudd, K.E., and E.V. Koonin. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli. Current Biology.* **6**:279-291.

Taylor, D.E., Eaton, M., Chang, N., and S.M. Salama. 1992. Construction of a *Helicobacter pylori* genome map and demonstration of diversity at the genome level. *Journal of Bacteriology.* **174**:6800-6806.

Tchelet, R., and M. Mevarech. 1994. Interspecies genetic transfer in halophilic archaebacteria. *Systematic and Applied Microbiology.* **16**:578-581.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., Nelson, W.C., Gwinn, M.L., DeBoy, R., Peterson, J.D., Hickey, E.K., Haft, D.H., Salzberg, S.L., White, O., Fleischmann, R.D., Dougherty, B.A., Mason, T., Ciecko, A., Parksey, D.S., Blair, E., Cittone, H., Clark, E.B., Cotton, M.D., Utterback, T.R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, V., Masignani, V., Pizza, M., Grandi, G., Sun, L., Smith, H.O., Fraser, C.M., Moxon, E.R., Rappuoli, R., and J.C. Venter. 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science.* **287**:1809-1815.

Toda, T., and M. Itaya. 1995. I-*Ceu*I recognition sites in the *rrn* operons of the *Bacillus subtilis* 168 chromosome: inherent landmarks for genome analysis. *Microbiology.* **141**:1937-1945.

Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E.F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H.G., Glodek, A., McKenney, K., Fitzegerald, L.M., Lee, N., Adams, M.D., Hickey, E.K., Berg, D.E., Gocayne, J.D., Utterback, T.R., Peterson, J.D., Kelley, J.M., Cotton, M.D., Weidman, J.M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W.S., Borodovsky, M., Karp, P.D., Smith, H.O., Fraser, C.M., and J.C. Venter. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori.* *Nature.* **388**:539-547.

Tomlinson, G.A., Jahnke, L.L., and L.I. Hochstein. 1986. *Halobacterium denitrificans* sp. nov., an extremely halophilic denitrifying bacterium. *International Journal of Systematic Bacteriology.* **36**:66-70.

Vellai, T., Kovács, A.L., Kovács, G., Ortutay, C., and G. Vida. 1999. Genome economization and a new approach to the species concept in bacteria. *Proceedings of the Royal Society of London-series B: biological sciences.* **266**:1953-1958.

Ventosa, A., Arahal, D.R., and B.E. Volcani. 1999. Studies on the microbiota of the Dead Sea - 50 years later. p.139-147. In A. Oren (ed.), Microbiology and biogeochemistry of hypersaline environments. CRC Press, Boca Raton, FL.

Watanabe, H., Mori, H., Itoh, T., and T. Gojobori. 1997. Genome plasticity as a paradigm of eubacteria evolution. *Journal of Molecular Evolution.* **44** (Suppl 1):S57-S64.

Weinstock, G.M., and J.R. Lupski. 1998. Chromosomal rearrangements. p.112-118. In F.J. de Bruijn, J.R. Lupski and G.M. Weinstock (ed.), Bacterial genomes physical structure and analysis. Chapman and Hall, New York, NY.

White, O., Eisen, J.A., Heidelberg, J.F., Hickey, E.K., Peterson, J.D., Dodson, R.J., Haft, D.H., Gwinn, M.L., Nelson, W.C., Richardson, D.L., Moffat, K.S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J.J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K.S., Aravind, L., Daly, M.J., Minton, K.W., Fleischmann, R.D., Ketchum, K.A., Nelson, K.E., Salzberg, S., Smith, H.O., Venter, J.C., and C.M. Fraser. 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*. **286**:1571-1577.

Woese, C.R., Kandler, O., and M.L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *PNAS* USA. **87**:4576-4579.

Wolf, Y.I., Aravind, L., and E.V. Koonin. 1999. *Rickettsiae* and *Chlamydiae* evidence of horizontal gene transfer and gene exchange. *Trends in Genetics*. **15**:173-175.

Wu, L., Karow, J.K., and I.D. Hickson. 1999. Genetic recombination: helicases and topoisomerases link up. *Current Biology*. **9**:R518-R520.

Zuerner, R.L., Herrmann, J.L., and I. Saint Girons. 1993. Comparison of genetic maps for two *Leptospira interrogans* serovars provides evidence for two chromosomes and intraspecies heterogeneity. *Journal of Bacteriology*. **175**:5445-5451.

APPENDIX

I. RECIPES

I.A. Solutions

10X Klenow buffer: 100 mM Tris•HCl (pH 7.6) and 50 mM $MgCl_2$

0.5 mL of 1 M Tris•HCl, pH 7.6
0.25 mL of 1 M $MgCl_2$
Bring volume to 5 mL with water
Filter sterilize

10X Ligase buffer without ATP: 200 mM Tris•HCl (pH 7.6), 100 mM $MgCl_2$ and 100 mM dithiothreitol (DTT)

2 mL of 1 M Tris•HCl, pH 7.6
1 mL of 1 M $MgCl_2$
0.16 g of DTT
Bring volume to 10 mL with water
Filter sterilize

10X RE: 500 mM Tris•HCl (pH 8), 500 mM KCl, 100 mM $MgCl_2$ and 100 mM DTT

5 mL of 1 M Tris•HCl, pH 8
0.4 g of KCl
1 mL of 1 M $MgCl_2$
0.16 g of DTT
Bring volume to 10 mL with water
Filter sterilize

Size marker (λ × *Bst*E II, *Xho* I and *Xba* I):

λ DNA digested with *Bst*E II:
120 μg of λ DNA
80 μL of 10X RE buffer
20 μL of 4 M NaCl
10 μL of *Bst*E II (100 U)
Bring volume to 800 μL with water
Incubate at 62°C for 3 hours

λ DNA digested with *Xho* I:
40 μg of λ DNA
80 μL of 10X RE buffer
5 μL of *Xho* I (50 U)
Bring volume to 800 μL with water
Incubate at 37°C for 3 hours

λ DNA digested with *Xba* I:
40 μg of λ DNA
80 μL of 10X RE buffer
8 μL of *Xba* I (80 U)
Bring volume to 800 μL with water
Incubate at 37°C for 3 hours

Combine together in a tube:
100 μL of λ × *Bst*E II
100 μL of λ × *Xho* I
100 μL of λ × *Xba* I
100 μL of TE
100 μL of 0.25 M EDTA
150 μL of loading buffer
Verify bands on an agarose gel

STE: 50 mM Tris•HCl (pH 8), 50 mM EDTA and 1% sarkosil (N-lauroylsarcosine)

25 mL of 1 M Tris•HCl, pH 8
50 mL of 0.5 M EDTA
5 g of N-lauroylsarcosine
Bring volume to 500 mL with water
Filter sterilize

40X TAE: 2 M Tris, 80 mM EDTA and 0.8 M sodium acetate

484.5 g of Tris
59.9 g of EDTA
217.8 g of sodium acetate
Bring volume to 1858 mL with water
Adjust pH to 7.2 by adding 142 mL of acetic acid

TE: 10 mM Tris•HCl (pH 8) and 1 mM EDTA

10 mL of 1 M Tris•HCl, pH 8
2 mL of 500 mM EDTA
Bring volume to 1 litre with water
Autoclave

I.B. Media

"Happy medium" broth: (per litre)

Combine
950 mL of "Happy medium" salt solution
50 mL of "Happy medium" food solution

"Happy medium" food solution:

5 g of Bacto tryptone
3 g of Bacto yeast extract
Bring volume to 50 mL with water
Autoclave

"Happy medium" plates: (per litre)

950 mL of "Happy medium" salt solution
15 g of agar
Autoclave
Add 50 mL of autoclaved "Happy medium" food solution
Pour plates

"Happy medium" salt solution:

206 g of NaCl
37 g of $MgSO_4 \cdot 7H_2O$
3.7 g of KCl
170 μL of 10 μM $MnCl_2$
25 mL of 2 M Tris•HCl, pH 7.2
Bring volume to 945 mL with water
Autoclave
Add 5 mL of sterile 10% $CaCl_2 \cdot 2H_2O$

NZY broth: (per 500 mL)

5 g of NZ amine (casein hydrolysate)
2.5 g of yeast extract
2.5 g of NaCl
Adjust pH to 7.5 with NaOH
Autoclave

NZY$^+$ broth:

10 mL of NZY broth
125 μL of 1 M $MgCl_2$
125 μL of 1 M $MgSO_4$
200 μL of sterile 20% glucose (w/v)
Filter sterilize

YT-Ampicillin broth: (per 500 mL)

YT broth
Autoclave
Cool to 55°C
Add 35 mg of filter-sterilized ampicillin

YT-Ampicillin-Xgal-IPTG plates: (per 500 mL)

YT plates
Autoclave
Cool to 55°C
Add 35 mg of filter-sterilized ampicillin
Add 2 mL of 2% Xgal
Add 0.4 mL of filter-sterilized 1 M IPTG
Pour plates

YT broth: (per 500 mL)

4 g of tryptone
2.5 g of yeast extract
2.5 g of NaCl
Autoclave

YT plates: (per 500 mL)

YT broth
7.5 g of agar
Autoclave
Pour plates

II. CLONE ANALYSIS

Each clone (fragment) is represented by 2 sequences: the universal forward sequence (u) and the universal reverse sequence (r).Indicated for each sequence are: the total insert length (I), the edited sequence length (S), the BLASTX results for searches against the NCBI database, the BLASTX results for searches against the *Halobacterium salinarum* NRC-1 database, the gene index number (gi) of the best NRC-1 match and the distance (number of genes) between NRC-1 homologs (D).

* homologs located on pNRC200
† homologs located on pNRC100
‡ homologs on different replicons

II.A. *Haloferax volcanii*

Clone	I (kbp)	S (bp)	NCBI database	NRC-1 database	gi	D
uDS22B1	1.6	608	thiamine biosynthesis protein (e-25)	Vng0730c (e-22)	10580308	510
rDS22B1	1.6	640	oxidoreductase (e-6)	Vng0061c (e-11)	10579715	
uDS22B2	1.8	596				
rDS22B2	1.8	676	hypothetical protein (e-9)	Vng0383h (e-7)	10579997	
uDS22B3	1.0	583	hypothetical protein (e-8)			
rDS22B3	1.0	651	altronate hydrolase (e-34)			
uDS22B6	1.0	629	hypothetical protein (e-9)	Vng0354c (e-37)	10579975	
rDS22B6	1.0	712	sulfur deprivation response regulator (e-46)	*ars*B: arsenite transporter protein (e-79)	10581983	338
uDS22B7	2.3	571	hypothetical protein (e-13)	Vng6293c (e-31)* ORF H1626c (e-28)†	10584331 10581101	
rDS22B7	2.3	685	ferredoxin-nitrite reductase (e-41)			
uDS22B8	2.1	604		*glc*D: glycolate oxidase subunit (e-21)	10581829	
rDS22B8	2.1	346	glycerol-3-phosphate dehydrogenase (e-6)	*glc*D: glycolate oxidase subunit (e-27)	10581829	0

uDS22B11	1.8	333		*trm*1: N2,N2-dimethylguanosine tRNA methyltransferase (e-8)	10581506	1
rDS22B11	1.8	537	hypothetical protein (e-6)	Vng2089h (e-33)	10581507	
uDS22B12	3.0	438				
rDS22B12	3.0	559	hypothetical protein (e-14)	Vng2121c (e-9)	10581536	
uDS23A2	1.3	569				
rDS23A2	1.3	624	transposase (e-15)			
uDS23A8	3.7	555				
rDS23A8	3.7	601	sodium-dependent transporter (e-25)	*nac*: sodium- and chloride-dependent transporter (e-53)	10581016	
uDS23A11	1.9	575	activator 1, replication factor (e-21)	*rfc*A: replication factor C small subunit (e-37)	10581691	2
rDS23A11	1.9	627	hypothetical protein (e-8)	Vng2277h (e-21)	10581689	
uDS23B2	1.9	484				
rDS23B2	1.9	613	secretory protein kinase (e-19)	*gsp*E1: type II secretion system protein (e-81)	10579863	
uDS23B5	2.6	602	S-adenosyl-L methionine: uroporphyrinogen (e-45)	*uro*M: S-adenosyl-L methionine: uroporphyrinogen III methyltransferase (e-82)	10581741	
rDS23B5	2.6	605	hypothetical protein (e-8)			
uDS23B6	2.1	544	glucodextranase precursor (e-25)			
rDS23B6	2.1	578	glucodextranase (e-36)			
uDS23B7	1.9	547	hypothetical protein (e-57)	Vng2281c (e-70)	10581692	
rDS23B7	1.9	488				
uDS23B8	2.1	586				
rDS23B8	2.1	514	transcriptional regulator (e-9)	*arc*R: transcription regulator (e-7)*	10584354	
uDS23B9	1.7	533				
rDS23B9	1.7	475				
uDS23B10	1.2	549	ribosomal protein S19 (e-62)	*rps*19p: 30S ribosomal protein S19P (e-63)	10581163	1
rDS23B10	1.2	565	ribosomal protein L2 (e-81)	*rp*12p: 50S ribosomal protein L2P (e-76)	10581162	
uDS23B11	1.6	479	ATP-dependent protease (e-43)	*lon*: lipoate protein ligase (e-66)	10579932	
rDS23B11	1.6	421				

uDS24A1	2.9	514				
rDS24A1	2.9	500	deoxyhypusine synthase (e-21)	*dhs*: deoxyhypusine synthase (e-12)	10580933	
uDS24A2	4.0	478		Vng1451c (e-9)	10580947	1
rDS24A2	4.0	450	hypothetical protein (e-11)	*yfk*N: 2'3'-cyclic nucleotide 2'phosphodiesterase (e-25)	10580946	
uDS24A3	2.9	416	hypothetical protein (e-22)	Vng1982c (e-49)	10581415	
rDS24A3	2.9	612				
uDS24A5	1.0	516	hypothetical protein (e-12)			
rDS24A5	1.0	532				
uDS24A6	3.3	617	molybdopterin converting factor (e-9)	*moa*E: molybdenum cofactor biosynthesis protein (e-21)	10579734	592
rDS24A6	3.3	481	lysyl-tRNA synthetase (e-25)	*lys*A: lysyl-tRNA synthetase (e-70)	10581446	
uDS24A8	2.4	514				
rDS24A8	2.4	468				
uDS24A9	3.3	505				
rDS24A9	3.3	424	DNA ligase (e-23)	*lig*: DNA ligase (e-50)	10580445	
uDS24A10	3.3	656				
rDS24A10	3.3	732	2-oxo-3-deoxygalactonate 6-phosphate (e-41)	*rsp*A: starvation sensing protein (e-11)	10580052	
uDS24A11	1.6	387	UDP-glucose 4-epimerase (e-13)	*gmd*: GDP-D-mannose dehydratase (e-13)	10579719	629
rDS24A11	1.6	642	NADH dehydrogenase (e-9)	*yjl*D: NADH dehydrogenase (e-34)	10580455	
uDS24B3	1.6	682	hypothetical protein (e-8)	Vng0359c (e-44)	10579978	0
rDS24B3	1.6	669	hypothetical protein (e-10)	Vng0359c (e-14)	10579978	
uDS24B4	1.5	588	4-aminobutyrate aminotransferase (e-30)	Vng1524c (e-57)	10581011	927
rDS24B4	1.5	705		*ntp*: neutral proteinase (e-23)	10579693	
uDS24B6	1.3	675	hypothetical protein (e-6)	Vng6441h (e-12)*	10584457	
rDS24B6	1.3	430	aminoacylase (e-12)			
uDS24B8	3.3	650	hypothetical protein (e-10)			
rDS24B8	3.3	563	hypothetical protein (e-10)	Vng1198c (e-69)	10580732	
uDS24B9	2.9	608	hypothetical protein (e-12)			
rDS24B9	2.9	540	hypothetical protein (e-9)			

uDS24B10	2.4	499		Vng1425h (e-13)	10580927	
rDS24B10	2.4	478				
uDS24B11	1.9	618				
rDS24B11	1.9	575	hypothetical protein (e-20)	Vng1905c (e-61)	10581347	
uDS25A2	2.9	511				
rDS25A2	2.9	490		*hcp*B: halocyanin precursor-like (e-9)	10581613	
uDS25A4	2.1	525	N-acetyl-gamma-glutamyl-phosphatase (e-40)			
rDS25A4	2.1	574	acetylornithine aminotransferase (e-32)	*gabT:* gamma-aminobutyrate aminotransferase (e-12)*	10584265	
uDS25A7	1.2	576	hypothetical protein (e-6)	*ywa*D: aminopeptidase (e-30)	10580170	1
rDS25A7	1.2	557	hypothetical protein (e-9)	Vng0576c (e-16)	10580171	
uDS25A11	2.3	559	phosphatidylserine synthase (e-12)	*pss*A: CDP-diacylglycerol-serine O-phosphatidyltransferase (e-45)	10580359	4
rDS25A11	2.3	551	hypothetical protein (e-8)	Vng0789c (e-22)	10580363	
uDS26A1	1.9	531	hypothetical protein (e-14)			
rDS26A1	1.9	507	hypothetical protein (e-7)	*kin*A2: signal-transducing histidine kinase homolog (e-22)	10580314	
uDS26A6	4.6	573	asparaginase (e-45)	*ans*A: L-asparaginase (e-76)	10581290	34
rDS26A6	4.6	542	2,3 biphosphoglyceromutase (e-41)	*gmp*: phosphoglycerate mutase (e-72)	10581330	
uDS26A9	4.6	586	N-ethylammeline chlorohydrolase (e-20)	Vng1205c (e-62)	10580737	
rDS26A9	4.6	428				
uDS26A11	1.9	597	superoxide dismutase (e-18)	Vng1330h (e-18)	10580847	7
rDS26A11	1.9	498	oxidoreductase (e-20)	*fab*G: 3-oxoacyl (acyl-carrier protein) reductase (e-15)	10580854	
uDS26A12	4.6	590	sialidase (e-11)			
rDS26A12	4.6	422	ORF H0337 from *Halobacterium* (e-6)	Vng1120h (e-11)	10580662	
uDS26B1	1.4	550	hypothetical protein (e-8)	Vng0405c (e-35)	10580018	0
rDS26B1	1.4	555	hypothetical protein (e-8)	Vng0405c (e-37)	10580018	
uDS26B2	2.1	405	hypothetical protein (e-23)			
rDS26B2	2.1	394	hypothetical protein (e-18)	Vng0096c (e-16)	10579744	

uDS26B9	2.5	514	transcriptional regulatory protein (e-9)	Vng1237c (e-36)	10580765	
rDS26B9	2.5	415				
uDS26B10	1.0	384	hypothetical protein (e-11)			
rDS26B10	1.0	468	cell division inhibitor (e-23)	*min*D2: cell division inhibitor (e-27)	10581544	
uDS26B11	1.0	320				
rDS26B11	1.0	531				
uDS28A10	2.6	695	hypothetical protein (e-17)	Vng0743h (e-8)	10580322	
rDS28A10	2.6	637	hypothetical protein (e-53)			
uDS29A7	3.6	528	hypothetical protein (e-10)	Vng1101c (e-31)	10580647	50
rDS29A7	3.6	571	LSU ribosomal protein (e-21)	Vng1169c (e-32)	10580704	
uDS29A9	1.6	629	acetylglutamate kinase (e-34)	Vng2433h (e-10)	10581839	
rDS29A9	1.6	574	N-acetylornithine aminotransferase (e-32)	*gabT*: gamma-aminobutyrate aminotransferase (e-27)*	10584265	‡
uDS29A11	2.1	548	type II secretion system protein (e-40)	*gsp*E2: type II secretion protein (e-39)	10580393	1
rDS29A11	2.1	665	hypothetical protein (e-27)	Vng0822c (e-19)	10580392	
uDS29A12	3.1	564	hypothetical protein (e-8)	Vng1782c (e-13)	10581236	1
rDS29A12	3.1	720	hypothetical protein (e-34)	Vng1784c (e-39)	10581237	
uDS210A7	3.6	606	inosine-5'-monophosphate dehydrogenase (e-30)	*gua*B: inosine-monophosphate dehydrogenase (e-51)	10580558	2
rDS210A7	3.6	622	hypothetical protein (e-13)	Vng1003h (e-28)	10580560	
uDS210A8	1.5	667				
rDS210A8	1.5	602		Vng1375c (e-15)	10580884	
uDS210A9	1.8	567	cytochrome c oxidase subunit (e-23)	*cox*B2: cytochrome c oxidase subunit II (e-32)	10581612	3
rDS210A9	1.8	575	hypothetical protein (e-8)	Vng2199h (e-43)	10581615	
uDS210A11	2.6	510	hypothetical protein (e-7)	*uvr*D: repair helicase (e-26)	10582002	3
rDS210A11	2.6	668	carboxypeptidase (e-45)	*cxp*: carboxypeptidase (e-65)	10581997	
uDS210A12	4.3	487	halobacterial transducer protein (e-28)	*htr*15: Htr15 transducer (e-41)	10580513	978
rDS210A12	4.3	617	hypothetical protein (e-7)	Vng2366c (e-25)	10581772	
uDS211A9	1.0	634	gluconolactonase (e-11)			
rDS211A9	1.0	588				

uDS211A10	3.1	632		duplicate DS28A10		
rDS211A10	3.1	520				
uDS211A11	2.6	639	hypothetical protein (e-12)	Vng1898c (e-32)	10581338	
rDS211A11	2.6	562	hypothetical protein (e-9)			
uDS211A12	2.2	594	hypothetical protein (e-9)			
rDS211A12	2.2	513	indoleglycerol phosphate synthetase (e-60)	*trp*C: indole-3-glycerol-phosphate synthase (e-34)	10579934	
uDS212A2	3.0	320				
rDS212A2	3.0	402				
uDS212A3	4.4	399	DNA helicase (e-30)	*rad*24b: DNA repair protein (e-45)	10581777	25
rDS212A3	4.4	444	hypothetical protein (e-8)	*scm*: 24-sterol C-methyltransferase (e-47)	10581805	
uDS212A5	1.5	466	phosphoenolpyruvate carboxylase (e-35)			
rDS212A5	1.5	417	phosphoenolpyruvate carboxylase (e-20)			
uDS212A6	2.4	565	glucose-1-phosphate thymidylyltransferase (e-10)	*gra*D5: glucose-1-phosphate thymidylyltransferase (e-12)	10579655	1
rDS212A6	2.4	501	glucosamine-fructose-6-phosphatase (e-40)	*glm*S: glutamine-fructose-6-phosphate transaminase (e-36)	10579654	
uDS212A7	3.0	486	alanyl-tRNA synthetase (e-48)	*ala*S: alanyl-tRNA synthetase (e-64)	10581694	169
rDS212A7	3.0	454	hypothetical protein (e-8)	Vng2501c (e-22)	10581894	
uDS212A9	2.4	516				
rDS212A9	2.4	351				
uDS212A10	4.4	431		Vng1743c (e-18)	10581201	
rDS212A10	4.4	492	hypothetical protein (e-24)			
uDS212B1	3.3	503	nicotinate phosphoribosyltransferase (e-49)	Vng2541c (e-70)	10581933	
rDS212B1	3.3	520				
uDS212B2	3.0	470	hypothetical protein (e-9)	*hpb*: phosphate binding protein (e-51)	10581502	2
rDS212B2	3.0	522	inner membrane component (e-8)	*phn*E: transport protein (e-24)	10581500	
uDS212B3	2.9	413		Vng1092c (e-24)	10580639	313
rDS212B3	2.9	348	transducer protein (e-11)	*htr*10: Htr10 transducer (e-13)	10580997	

uDS212B4	3.3	389		duplicate DS212B3		
rDS212B4	3.3	349				
uDS212B8	1.7	458	hypothetical protein (e-29)	*nir*D: heme biosynthesis protein (e-56)	10581351	1
rDS212B8	1.7	340	hypothetical protein (e-7)	*trp*D2: phosphoribosyl transferase (e-20)	10581352	
uDS212B10	2.4	428				
rDS212B10	2.4	323				
uDS213A1	2.4	539	hypothetical protein (e-25)			
rDS213A1	2.4	529	hypothetical protein (e-9)	Vng0668c (e-24)	10580254	
uDS213A2	2.4	570		duplicate DS213A1		
rDS213A2	2.4	525				
uDS213A3	2.9	523		Vng0669h (e-31)	10580255	‡
rDS213A3	2.9	453	hypothetical protein (e-14)	Vng6319h (e-16)*	10584355	
uDS213A6	2.9	572		duplicate DS23B8		
rDS213A6	2.9	518				
uDS213A8	1.2	452	hypothetical protein (e-12)	*kin*A2: signal-transducing histidine kinase homolog (e-10)	10580314	898
rDS213A8	1.2	558	phosphoglucomutase (e-21)	*pmm*: phosphoglucomutase/phospho-mannomutase (e-47)	10581688	
uDS213A10	2.9	405	hypothetical protein (e-11)	*hjr*: Holliday junction resolvase (e-20)	10581665	
rDS213A10	2.9	559	N-ethylammeline chlorohydrolase (e-11)	*trz*A: N-ethylammeline chlorohydrolase (e-25)	10581663	2
uDS213A11	3.9	428	5-methyltetrahydropteroyl-triglutamate (e-15)			
rDS213A11	3.9	472		*hsp*3: small heat-shock protein (e-36)	10580692	
uDS213A12	2.4	557				
rDS213A12	2.4	523	hypothetical protein (e-14)			
uDS213B2	1.9	544	hypothetical protein (e-6)	Vng0280h (e-6)	10579913	
rDS213B2	1.9	538	hypothetical protein (e-21)			
uDS213B3	1.9	482	hypothetical protein (e-7)	*cta*B: heme synthase (e-30)	10580252	1
rDS213B3	1.9	606	hypothetical protein (e-8)	*trp*4: ABC transporter, ATP-binding protein homolog (e-31)	10580253	

uDS213B5	1.9	566	hypothetical protein (e-8)	Vng0555c (e-25)	10580152	916
rDS213B5	1.9	402	dolichol-P-glucose synthetase (e-8)	*dgs*: dolichol-P-glucose transferase (e-53)	10581484	
uDS213B6	1.4	516	sugar fermentation stimulation protein (e-9)			
rDS213B6	1.4	601				
uDS213B10	2.4	597				
rDS213B10	2.4	577				
uDS213B12	3.6	591	LSU ribosomal protein (e-23)	*rp*121e: 50S ribosomal protein L21E (e-36)	10580705	51
rDS213B12	3.6	542	hypothetical protein (e-18)	Vng1101c (e-42)	10580647	
uDS215A2	2.7	366	hypothetical protein (e-6)	*ykh*A: acyl-CoA hydrolase (e-8)	10580436	
rDS215A2	2.7	505	glycerate kinase (e-16)			
uDS215A3	1.0	489		Vng1617h (e-24)	10581092	
rDS215A3	1.0	467				
uDS215A5	2.7	503	hypothetical protein (e-8)	*sdh*: succinate dehydrogenase subunit (e-28)	10581522	
rDS215A5	2.7	546				
uDS215A8	2.1	553	hypothetical protein (e-9)	Vng0188h (e-70)	10579835	
rDS215A8	2.1	562				
uDS215A9	2.1	365		duplicate DS215A8		
rDS215A9	2.1	493				
uDS215A11	2.1	381	carbamoyl-phosphate synthetase (e-36)	*car*B: carbamoyl-phosphate synthase large subunit (e-54)	10581263	0
rDS215A11	2.1	610	carbamoyl-phosphate synthetase (e-16)	*car*B: carbamoyl-phosphate synthase large subunit (e-22)	10581263	
uDS215A12	1.3	563	cell division protein (e-44)	*fts*Z3: cell division protein (e-85)	10581370	1
rDS215A12	1.3	578		Vng1934h (e-20)	10581371	
uDS215B1	1.6	558				
rDS215B1	1.6	558	archaeal ABC-transporter system (e-23)	*app*C: oligopeptide transporter permease protein (e-15)	10581769	
uDS215B2	2.1	555	cytochrome b6 (e-19)	*cyb*: cytochrome b6 (e-48)	10580178	5
rDS215B2	2.1	490	hypothetical protein (e-8)	Vng0576c (e-15)	10580171	

uDS215B4	2.7	428				
rDS215B4	2.7	504	cysteinyl-tRNA synthetase (e-16)	*cys*S: cysteinyl-tRNA synthetase (e-49)	10580644	

II.B. *Haloferax denitrificans*

Clone	I (Kb)	S (bp)	NCBI database	NRC-1 database	gi	D
uden1A3	4.0	526	cobalamin biosynthesis precorrin (e-23)	*cbi*C: precorrin isomerase (e-79)	10581052	1
rden1A3	4.0	489	cobalamin biosynthesis protein (e-16)	*cob*N: cobalamin biosynthesis protein (e-68)	10581051	
uden1A6	4.5	606	hypothetical protein (e-15)	*vac*B: ribonuclease II family protein (e-90)	10582029	944
rden1A6	4.5	517	hypothetical protein (e-15)	Vng1422h (e-22)	10580925	
uden1A7	4.5	609				
rden1A7	4.5	492	phototaxis transducer II (1e-10)	*htr2*: Htr2 transducer (e-12)	10581218	
uden1A8	2.8	615		Vng0853c (e-30)	10580421	
rden1A8	2.8	558				
uden1A9	4.5	493	ORF H1336 *Halobacterium* (e-9)	ORF H1336 (e-10)†	10803659	
rden1A9	4.5	568	fatty acid acyl-CoA dehydrogenase (e-31)	*acd4*: acyl-CoA dehydrogenase (e-31)	10580266	‡
uden1A10	2.5	598	ORF H0337 *Halobacterium* (e-29)	*ark*: adaptive-response sensory-kinase (e-33)	10580474	893
rden1A10	2.5	489	hypothetical protein (e-34)	Vng2154c (e-61)	10581570	
uden1A11	3.6	613		Vng6370h (e-24)*	10584404	
rden1A11	3.6	620	hypothetical protein (e-30)			
uden1B4	3.6	520	chloride channel (e-24)	*clc*: chloride channel (e-67)	10581031	
rden1B4	3.6	546				
uden1B5	5.0	598	aldehyde dehydrogenase (e-106)	*ald*Y1: aldehyde dehydrogenase (retinol) (e-57)	10581906	711
rden1B5	5.0	545	hypothetical protein (e-18)	Vng1595c (e-32)	10581075	
uden1B10	5.0	529				
rden1B10	5.0	611	hypothetical protein (e-13)	Vng0879c (e-25)	10580443	
uden1C2	4.5	579		duplicate den1A9		
rden1C2	4.5	670				
uden1C4	3.8	512	uracil phosphoribosyltransferase (e-6)	Vng2304h (e-14)	10581716	705
rden1C4	3.8	641	hypothetical protein (e-12)	Vng0537c (e-44)	10580137	

uden1C8	2.7	547	propionyl-CoA carboxylase (e-69)	*mmd*A: methylmalonyl-CoA decarboxylase, subunit alpha (e-89)	10581017	0
rden1C8	2.7	615	carboxyl transferase (e-10)	*mmd*A: methylmalonyl-CoA decarboxylase, subunit alpha (e-19)	10581017	
uden1C12	2.1	647	ATP-dependent oligopeptide ABC transporter (e-44)	*opp*D2: oligopeptide ABC transporter (e-43)	10581757	139
rden1C12	2.1	657	ATP-dependent oligopeptide ABC transporter (e-40)	*dpp*F: dipeptide ABC transporter, ATP binding (e-40)	10581920	
uden1D5	4.6	663	enoyl-CoA hydratase (e-34)	*fad2*: enoyl-CoA hydratase (e-40)	10580038	45
rden1D5	4.6	616	acetyl-CoA synthetase (e-18)	*acs*1: acetyl-CoA synthetase (e-11)	10580090	
uden1D7	3.5	633		Vng0631c (e-29)	10580222	1
rden1D7	3.5	639	phosphoribosyl aminoimidazole (e-42)	*pur*K: phosphoribosyl aminoimidazole carboxylase ATP binding subunit (e-63)	10580223	
uden1D9	3.7	673	ORF H0581/H1776 from *Halobacterium* (e-48)	Vng6457c (e-49)* Vng6068c (e-49)* ORF H0581 (e-49)† ORF H1776 (e-49)†	10584475 10584127 10803598 10803705	‡
rden1D9	3.7	627	replication factor C, activator 1 (e-31)	*rfc*B: replication factor C large subunit (e-60)	10581095	
uden1D10	3.9	664	hypothetical protein (e-10)	Vng0322h (e-38)	10579950	3
rden1D10	3.9	643		Vng0319h (e-20) Vng0320h (e-18)	10579947 10579948	2
uden1D11	3.7	624		duplicate den1D9		
rden1D11	3.7	632				
uden1D12	3.5	675	hypothetical protein (e-42)			
rden1D12	3.5	606	probable RNA 3'-terminal phosphatase (e-38)	*tpc*: RNA 3'-terminal phosphate cyclase (e-45)	10580310	
uden2A3	2.7	648	hypothetical protein (e-10)	Vng0725h (e-48)	10580304	764
rden2A3	2.7	511	hypothetical protein (e-8)	Vng1798h (e-32)	10581252	
uden2A4	2.9	595	hypothetical protein (e-8)			
rden2A4	2.9	615	cell surface glycoprotein (e-13)			

uden2A5	2.1	679	glutamine ABC transporter (e-59)	yvrO: amino acid ABC transporter, ATP-binding protein (e-32)	10579651	
rden2A5	2.1	557	zinc-binding alcohol dehydrogenase (e-18)	*adh*2: alcohol dehydrogenase (e-16)	10581998	50
uden2A8	2.5	645	threonine synthase (e-21)	*thr*C3: threonine synthase (e-59)	10581523	
rden2A8	2.5	581	hypothetical protein (e-9)			
uden2A10	2.3	639				
rden2A10	2.3	659		Vng1518h (e-7)	10581007	
uden2A12	3.9	374				
rden2A12	3.9	466	pyridoxal phosphate biosynthetase (e-13)			
uden2B2	3.0	427	uridylyl transferase (e-8)	Vng1536c (e-12)	10581024	
rden2B2	3.0	352	electron transfer flavoprotein (e-22)	*etf*A: electron transfer flavoprotein alpha subunit (e-23)	10581567	467
uden2B3	3.4	307	malate dehydrogenase (e-53)	*mdh*A: L-malate dehydrogenase (e-51)	10581775	
rden2B3	3.4	410	ABC excision nuclease subunit C (e-11)	*uvr*C: excision nuclease chain C (e-48)	10581790	13
uden2B5	2.3	427	hypothetical protein (e-29)	Vng1303c (e-35)	10580823	0
rden2B5	2.3	402		Vng1303c (e-27)	10580823	
uden2B8	2.5	433	hypothetical protein (e-10)	*yfk*N: 2'3'-cyclic-nucleotide 2'-phosphodiesterase (e-29)	10580946	
rden2B8	2.5	454				
uden2B11	3.7	486	molybdopterin oxidoreductase (e-11)	*dms*A: dimethylsulfoxide reductase (e-40)	10580398	2
rden2B11	3.7	387		Vng0826c (e-18)	10580396	
uden2C2	3.7	659	putative oxidoreductase (e-49)	Vng2488c (e-63)	10581883	6
rden2C2	3.7	615		Vng2477h (e-7)	10581877	
uden2C4	5.0	653				
rden2C4	5.0	722	hypothetical protein (e-12)			
uden2C9	5.0	653		Vng1362h (e-13)	10580872	911
rden2C9	5.0	655	valyl-tRNA synthetase (e-57)	*val*S: valyl-tRNA synthetase (e-58)	10581937	

uden2C10	3.9	736	tryptophanyl-tRNA synthetase (e-60)	*trp*S1: tryptophanyl-tRNA synthetase (e-109)	10581625	1
rden2C10	3.9	674	tRNA intron endonuclease (e-43)	*end*A: tRNA intron endonuclease (e-22)	10581626	
uden2C11	3.1	672	hypothetical protein (e-6)	*sir*R: transcription repressor (e-18)	10580136	
rden2C11	3.1	536				
uden2D4	3.7	635	ORF H0337 *Halobacterium* (e-51)	ORF H0337 (e-53)† *htl*D: Htr-like protein (e-53)*	10803577 10584098	
rden2D4	3.7	652				
uden2D8	5.0	698	NADH dehydrogenase (e-26)	*yjl*D: NADH dehydrogenase (e-42)	10580455	
rden2D8	5.0	735	hypothetical protein (e-8)			
uden3A1	2.1	459		Vng1034h (e-14)	10580586	693
rden3A1	2.1	466	oxidoreductase (e-22)	*gsp*: general stress protein 69 (e-10)	10581452	
uden3A2	2.9	500	hypothetical protein (e-40)	Vng1982c (e-58)	10581415	
rden3A2	2.9	568	hypothetical protein (e-10)			
uden3A3	2.8	526	probable RNA 3'-terminal phosphatase (e-20)	*tpc*: RNA 3'-terminal phosphate cyclase (e-33)	10580310	619
rden3A3	2.8	545	RNase L inhibitor (e-14)	*rli*: RNase L inhibitor homolog (e-16)	10581993	
uden3A5	3.4	460	N2,N2-dimethylguanosine tRNA (e-30)	*trm*1: N2,N2-dimethylguanosine tRNA methyltransferase (e-55)	10581506	99
rden3A5	3.4	559	ribose-phosphate pyrophosphokinase (e-20)	*prs*A: ribose-phosphate pyrophosphokinase (e-30)	10581620	
uden3A7	2.9	451		Vng1476c (e-49)	10580974	1
rden3A7	2.9	560	hypothetical protein (e-6)	Vng1475c (e-46)	10580973	
uden3A8	4.5	465				
rden3A8	4.5	505		Vng1842h (e-22)	10581288	
uden3A9	4.6	422	methylmalonyl-CoA mutase (e-52)	*mcm*A1: methylmalonyl-CoA mutase, subunit alpha (e-65)	10580085	62
rden3A9	4.6	501	NADH dehydrogenase (e-33)	Vng0560c (e-19)	10580156	
uden3A10	3.1	593	fructose-biphosphate aldolase (e-28)			
rden3A10	3.1	571	clumping factor (e-22)			
uden3B4	2.5	446	recombinase (e-15)			
rden3B4	2.5	527				

uden3B7	2.1	462				
rden3B7	2.1	526				
uden3B9	2.1	710	triosephosphate isomerase (e-21)	*tpi*A: triosephosphate isomerase (e-17)	10580581	
rden3B9	2.1	611	hypothetical protein (e-16)			
uden3B10	3.3	754		*rpa*: replication A related protein (e-29)	10579780	4
rden3B10	3.3	709		Vng0128c (e-46)	10579776	
uden3B11	3.9	773	ribosomal protein (e-90)	*rps*4e: 30S ribosomal protein S4E (e-82)	10581171	7
rden3B11	3.9	692	ribosomal protein (e-43)	*rpl18p: 50S* ribosomal protein L18P (e-73)	10581178	

II.C. *Haloferax gibbonsii*

Clone	I (Kb)	S (bp)	NCBI database	NRC-1 database	gi	D
ugib1A3	4.5	596	chemotaxis protein (e-27)	*che*D: chemotaxis protein (e-12)	10580524	
rgib1A3	4.5	516	signal-transducing histidine, ORF H0337 (e-13)	*kin*A2: signal-transducing histidine kinase (e-17)	10580314	177
ugib1A9	4.5	614	flavoprotein; electron transporter (e-23)			
rgib1A9	4.5	574				
ugib1A10	2.9	627	amidophosphoribosyl transferase (e-50)	*pur*F: amidophosphoribosyl-pyrophosphate amidotransferase (e-95)	10580989	1
rgib1A10	2.9	562		Vng1492c (e-22)	10580988	
ugib1A12	4.9	573	hypothetical protein (e-11)	Vng0150h (e-44)	10579797	2
rgib1A12	4.9	581	hypothetical protein (e-6)	*prr*C: regulatory protein (e-49)	10579799	
ugib1B3	1.6	489	hypothetical protein (e-7)	Vng1788c (e-59)	10581242	0
rgib1B3	1.6	439		Vng1788c (e-14)	10581242	
ugib1B4	1.0	588	hypothetical protein (e-9)			
rgib1B4	1.0	476	hypothetical protein (e-10)	Vng2297h (e-8)	10581708	
ugib1B9	1.7	530				
rgib1B9	1.7	511	N-carbamyl-L-amino acid amidohydantoin (e-20)			
ugib1B12	4.1	500	GTP cyclohydrolase II (e-30)	*rib*A: GTP cyclohydrolase II (e-33)	10581540	4
rgib1B12	4.1	490	similar to proline dehydrogenase (e-6)	Vng2121c (e-15)	10581536	
ugib1C5	3.0	527	hypothetical protein (e-9)	*rfb*Q: rhamnosyl transferase (e-10)	10580611	
rgib1C5	3.0	420	purine/xanthine permease (e-23)			
ugib1D2	3.8	363				
rgib1D2	3.8	390	hemolysin (e-19)	Vng6302c (e-34)	10584340	
ugib1D5	2.1	461	LSU ribosomal protein (e-54)	*rpl*19e: 50S ribosomal protein L19E (e-52)	10581177	2
rgib1D5	2.1	455	SSU ribosomal protein (e-40)	*rps*5p: 30S ribosomal protein S5P (e-42)	10581179	

ugib1D6	2.6	538				
rgib1D6	2.6	404	alkaline phosphatase (e-8)			
ugib2B4	1.0	629	archaeal transcription initiation factor (e-65)	*tbf*G: transcription initiation factor IIB (e-94)	10579892	0
rgib2B4	1.0	588	archaeal transcription initiation factor (e-45)	*tbf*G: transcription initiation factor IIB (e-67)	10579892	
ugib2C2	1.5	608				
rgib2C2	1.5	547	ribose 5-phosphate isomerase (e-15)	*rpi*: ribose 5-phosphate isomerase (e-24)	10581685	
ugib2C3	3.0	604	molybdopterin converting factor (e-9)	Vng1848h (e-23)	10581293	
rgib2C3	3.0	581	phosphoenolpyruvate carboxylase (e-11)			
ugib2C5	1.2	573	2-aminomuconic acid semialdehyde (e-9)	Vng0248c (e-28)	10579887	
rgib2C5	1.2	537	hemin-binding lipoprotein transporter (e-19)			
ugib2C7	3.2	618	ORF H0027/H0952 *Halobacterium* (e-34)	ORF H0027 (e-35)† ORF H0952 (e-35)† Vng6007h (e-35)* Vng6119h (e-35)*	10803552 10803632 10584069 10584173	
rgib2C7	3.2	618	encapsulation protein (e-10)			
ugib2C8	1.5	600	DNA polymerase II (e-24)	*pol*A2: DNA polymerase II, large chain (e-46)	10581749	
rgib2C8	1.5	466				
ugib2C11	3.6	472				
rgib2C11	3.6	596	cationic amino acid transporter (e-16)	*yhd*G: amino acid transporter (e-22)	10580770	
ugib2D2	2.5	612	deoxyribose aldolase (e-6)	Vng1852h (e-22) *deo*C: deoxyribose-phosphate aldolase (e-22)	10581298 10581305	943 936
rgib2D2	2.5	523	hypothetical protein (e-16)	*caa*: cation antiporter (e-14)	10579995	
ugib2D3	2.1	533	hydantoin utilization protein (e-44)			
rgib2D3	2.1	465	hypothetical protein (e-14)			

ugib2D6	3.0	637				
rgib2D6	3.0	591	thymidine kinase (e-16)	*tdk*: thymidine kinase (e-29)	10581006	
ugib2D7	3.0	607	putative ABC transporter, ATP binding (e-20)	*trp*3: daunorubicin resistance ABC transporter ATP-binding protein (e-22)	10580075	4
rgib2D7	3.0	556	pyruvate synthase (e-9)	*por*A: pyruvate ferrodoxin oxidoreductase, subunit alpha (e-26)	10580081	
ugib2D10	5.0	560	UDP-N-acetylglucosamine (e-28)			
rgib2D10	5.0	657	aspartate aminotransferase (e-35)	*yvo*F: acetyltransferase homolog (e-7)	10580893	
ugib2E2	5.0	634	hypothetical protein (e-16)	Vng0740c (e-33)	10580319	320
rgib2E2	5.0	582	hypothetical protein (e-43)	Vng1149c (e-92)	10580687	
ugib2E6	3.8	651	sensory rhodopsin II transducer protein (e-25)	*htr*12: Htr12 transducer (e-27)	10580941	212
rgib2E6	3.8	543	hypothetical protein (e-9)	Vng1092c (e-33)	10580639	
ugib2E7	4.6	679	ATP-binding transport protein (e-9)	Vng0668c (e-17)	10580254	676
rgib2E7	4.6	500	oxidoreductase (e-26)	*aad*: aryl-alcohol dehydrogenase (e-28)	10581102	
ugib2E10	2.2	613	hypothetical protein (e-12)	*put*P: proline permease (e-8)	10580468	
rgib2E10	2.2	611				
ugib2F6	4.2	642	methionyl-tRNA synthetase (e-33)	*met*S: methionine-tRNA synthetase (e-65)	10579953	697
rgib2F6	4.2	479	GTP cyclohydrolase II (e-30)	*rib*A: GTP cyclohydrolase II (e-22)	10581540	
ugib3A2	1.9	578	DNA damage-inducible protein (e-18)	Vng0727c (e-39)	10580306	551
rgib3A2	1.9	577		Vng0016h (e-13)	10579664	
ugib3A3	4.8	631	aspartate aminotransferase (e-11)	*asp*B2: aspartate aminotransferase (e-23)	10580220	726
rgib3A3	4.8	553	DNA helicase related protein (e-8)	*rad*24b: DNA repair protein (e-14)	10581777	
ugib3A4	3.0	501	hypothetical protein (e-6)	Vng2133h (e-12)	10581549	504
rgib3A4	3.0	593	bacterio-opsin activator (e-24)	*bat*: bacterio-opsin activator (e-26)	10580961	
ugib3A6	5.0	645	protease I (e-8)	duplicate gib3B7		
rgib3A6	5.0	648	acetyl-CoA synthetase (e-61)			
ugib3A8	4.8	570				
rgib3A8	4.8	567				

ugib3A11	3.7	625				
rgib3A11	3.7	473	oxidoreductase (e-14)	*gsp*: general stress protein 69 (e-19)	10581452	
ugib3A12	1.6	381				
rgib3A12	1.6	603	similar to adhesion protein (e-12)	*ycd*H: adhesion protein (e-65)*	10584307	
ugib3B2	3.7	604	hypothetical protein (e-14)	Vng2021c (e-26)	10581449	
rgib3B2	3.7	586				
ugib3B5	4.0	669	N-acetylornithine aminotransferase (e-29)	*gab*T: gamma-aminobutyrate aminotransferase (e-24)*	10584265	
rgib3B5	4.0	565	N-acetyl-gamma-glutamyl-phosphatase (e-38)			
ugib3B6	3.3	660	glutamate synthase (e-9)	Vng2306h (e-50)	10581718	98
rgib3B6	3.3	645	threonine synthase (e-57)	*thr*C1: threonine synthase (e-87)	10581834	
ugib3B7	3.0	667	acetyl-CoA synthetase (e-26)	*acs*2: acetyl-CoA synthetase (e-51)	10580552	0
rgib3B7	3.0	534	acetyl-CoA synthetase (e-34)	*acs*2: acetyl-CoA synthetase (e-56)	10580552	
ugib3B8	2.7	476				
rgib3B8	2.7	402				
ugib3B10	4.8	567	endo-glucanase (e-8)			
rgib3B10	4.8	389	thioredoxin (e-15)	*trx*A2: thioredoxin (1e-36)	10581982	
ugib3B11	2.7	482				
rgib3B11	2.7	375	sensory rhodopsin II transducer protein (e-19)	*htr*6: Htr6 transducer (e-21)	10580365	
ugib4A4	5.0	527	mannose-6-phosphate isomerase (e-30)	*man*C: mannose-1-phosphate guanylyltransferase (e-29)	10581577	8
rgib4A4	5.0	527	hypothetical protein (e-14)	Vng2168c (e-28)	10581585	

II.D. *Haloarcula japonica*

Clone	I (Kb)	S (bp)	NCBI database	NRC-1 database	gi	D
ujap1A2	2.1	458		Vng6372h (e-19)*	10584406	
rjap1A2	2.1	568	ORF H1502 from *Halobacterium* (e-44)	ORF H1502 (e-45)†	10803675	‡
ujap1A12	2.5	484				
rjap1A12	2.5	494				
ujap1B1	2.7	592				
rjap1B1	2.7	520		Vng0851c (e-12)	10580419	
ujap1B8	2.3	589				
rjap1B8	2.3	509	hypothetical protein (e-8)			
ujap1B9	3.7	621				
rjap1B9	3.7	510	flagellin precursor (e-10)	*fla*B2: flagellin B2 precursor (e-11)	10580518	
ujap1B10	4.0	595	acetyl-CoA synthetase (e-70)	*acs*2: acetyl-CoA synthetase (e-83)	10580552	0
rjap1B10	4.0	607	acetyl-CoA synthetase (e-38)	*acs*2: acetyl-CoA synthetase (e-90)	10580552	
ujap1C3	2.4	622				
rjap1C3	2.4	587	DNA-binding protein (e-8)			
ujap1C5	1.7	533		*trp*D2: phosphoribosyl transferase (e-43)	10581352	0
rjap1C5	1.7	577	phosphoribosyl anthranilate synthase (e-6)	*trp*D2: phosphoribosyl transferase (e-43)	10581352	
ujap1C6	2.6	581	S-adenosyl-methionine precorrin (e-15)	*cbi*L: cobalamin biosynthesis (e-51)	10581038	2
rjap1C6	2.6	587		*cbi*G: cobalamin biosynthesis (e-34)	10581040	
ujap1C8	1.6	537	halocyanin precursor (e-12)	*hcp*B: halocyanin precursor-like (e-16)	10581613	2
rjap1C8	1.6	619	hypothetical protein (e-8)	Vng2199h (e-38)	10581615	
ujap1C11	4.5	492				
rjap1C11	4.5	666				
ujap1D1	3.7	508				
rjap1D1	3.7	650				

ujap1D2	4.0	558	cysteinyl - tRNA synthetase/ligase (e-51)	*cys*S: cysteinyl-tRNA synthetase (e-76)	10580644	5
rjap1D2	4.0	543	hypothetical protein (e-13)	Vng1092c (e-67)	10580639	
ujap1D6	2.4	557	glycerol-3-phosphate dehydrogenase (e-26)	*gpd*A2: glycerol-3-phosphate dehydrogenase chain A (e-88)	10581403	1
rjap1D6	2.4	579	glycerol-3-phosphate dehydrogenase (e-17)	*gpd*B: glycerol-3-phosphate dehydrogenase chain B (e-71)	10581404	
ujap1G6	3.7	588		duplicate jap2D12		
rjap1G6	3.7	591				
ujap1G8	2.1	536				
rjap1G8	2.1	485	beta-methylaspartase ammonia-lyase (e-8)	*mal*: methylaspartate ammonia-lyase (e-17)	10581701	
ujap1G12	3.7	397				
rjap1G12	3.7	607				
ujap1H3	2.6	505				
rjap1H3	2.6	674	hypothetical protein (e-11)	Vng0170c (e-12)	10579818	
ujap1H5	4.5	489		Vng6143h (e-43)*	10584199	
rjap1H5	4.5	601				
ujap1H6		451	enoyl-CoA hydratase (e-30)	*hbd*2: 3-hydroxyacyl-CoA dehydrogenase (e-64)	10580830	1052
rjap1H6		492	glucose-1-phosphate thymidylyltransferase (e-37)	*gra*D2: glucose-1-phosphate thymidylyl transferase (e-52)	10579656	
ujap2A5	2.6	604	probable oxoacyl (e-13)	*fab*G: 3-oxoacyl (acyl-carrier-protein) reductase (e-36)	10580854	110
rjap2A5	2.6	519	acyl-CoA dehydrogenase (e-12)	*acd*5: acyl-CoA dehydrogenase (e-14)	10580979	
ujap2A6	2.0	585				
rjap2A6	2.0	565	exodeoxyribonuclease (e-23)			
ujap2A10	1.7	534				
rjap2A10	1.7	548				
ujap2B9	2.3	553	biotin carboxylase (e-28)	*acc*: biotin carboxylase (e-40)	10581019	
rjap2B9	2.3	578	propionyl-CoA carboxylase (e-73)	*mmd*A: methylmalonyl-CoA decarboxylase, subunit alpha (e-85)	10581017	2

ujap2C1	2.3	559		duplicate jap2E6		
rjap2C1	2.3	568				
ujap2C2	2.7	548	argininosuccinate lyase (e-27)	*arg*H: argininosuccinate lyase (e-27)	10581840	1
rjap2C2	2.7	578	argininosuccinate synthetase (e-66)	*arg*G: argininosuccinate synthetase (e-77)	10581841	
ujap2C3	1.3	585	bile acid-inducible operon protein (e-14)			
rjap2C3	1.3	577	ORF H0754/H1607 *Halobacterium* (e-46)	ORF H0754 (e-48)† ORF H1607 (e-48)† Vng6090c (e-48)*	10803616 10803687 10584151	
ujap2C5	1.1	460	ferredoxin (e-7)	Vng1561c (e-71)	10581045	1
rjap2C5	1.1	557	hypothetical protein (e-7)	Vng1562h (e-41)	10581046	
ujap2C6	4.8	434	carbamoyl-phosphate synthase (e-11)	*car*A: carbamoyl-phosphate synthase, subunit alpha (e-22)	10581264	413
rjap2C6	4.8	532	hypothetical serine rich protein (e-6)	Vng2334c (e-20) Vng2037c (e-20)	10581746 10581463	171
ujap2C7	4.6	583				
rjap2C7	4.6	667	TRK potassium uptake system (e-10)	*trk*H1: TRK potassium uptake system protein (e-30)	10580807	
ujap2C8	3.0	539	phytoene synthase (e-8)	*fdf*T: farnesyl-diphosphate farnesyl transferase (e-59)	10580267	
rjap2C8	3.0	513	levansucrase (e-28)			
ujap2C9	2.4	492				
rjap2C9	2.4	570	acetylpolyamine aminohydolase (e-6)	*aup*: acetoin utilization protein (e-32)	10579782	
ujap2C11	3.7	568	dihydroxynapthoic acid synthetase (e-12)	*cin*R: transcription repressor (e-42)	10580622	1
rjap2C11	3.7	658	hypothetical protein (e-10)	*lfl*1: long-chain fatty-acid-CoA ligase (e-18) *ykf*B2: chloromuconate cycloisomerase (e-20)	10580623 10580624	2
ujap2D1	2.0	516	hypothetical protein (e-9)			
rjap2D1	2.0	620	hypothetical protein (e-26)	Vng0683c (e-74)	10580270	

ujap2D7	2.1	573				
rjap2D7	2.1	431	transcriptional regulatory protein (e-9)	*arc*R: transcription regulator (e-10)*	10584354	
ujap2D12	1.8	509	hypothetical protein (e-11)			
rjap2D12	1.8	568	probable DNA-damage-inducible protein (e-15)	Vng0727c (e-49)	10580306	
ujap2E1	1.7	529				
rjap2E1	1.7	625	hypothetical protein (e-43)	Vng6183c (e-50)*	10584239	
ujap2E2	1.4	552				
rjap2E2	1.4	649				
ujap2E3	1.2	464				
rjap2E3	1.2	610	spore coat protein precursor (e-13)			
ujap2E4	1.4	525				
rjap2E4	1.4	651				
ujap2E5	1.5	505	iron ABC transporter (e-29)	*hem*U: iron (III) ABC transporter permease (e-51)	10580878	
rjap2E5	1.5	667	iron ABC transporter (e-18)	*hem*VI: iron (III) ABC transporter ATP binding (e-52)	10580877	1
ujap2E6	2.4	519		*cat*4: cationic amino acid transporter (e-10)*	10584240	‡
rjap2E6	2.4	642	cationic amino acid transporter (e-10)	*cat*1: cationic amino acid transporter (e-74)	10579821	
ujap2E8	1.0	542	chitinase (e-9)	Vng6145h (e-9)*	10584201	1
rjap2E8	1.0	640	serine-rich protein (e-6)	*trs*E: transfer complex protein (e-12)*	10584200	
ujap2E9	2.1	472	hypothetical protein (e-11)	Vng1297c (e-23)	10580816	2
rjap2E9	2.1	497	serine-rich protein (e-6)	Vng1.3h (e-8)	10580820	
ujap2E12	2.8	588	hypothetical protein (e-23)	Vng0674c (e-38)	10580261	3
rjap2E12	2.8	582	ketoacyl-CoA thiolase (e-43)	*aca*B1: 3-ketoacyl-CoA thiolase (e-81)	10580264	
ujap2F1	3.7	587		*prp*2: phosphate regulatory protein homolog (e-36)	10580345	3
rjap2F1	3.7	556		Vng0769h (e-35)	10580348	

ujap2F2	1.6	552				
rjap2F2	1.6	573	ethylene-inducible protein (e-32)	Vng1793c (e-41)	10581247	
ujap2F4	2.1	570	hypothetical protein (e-8)	*pot*A1: spermidine/putrescine ABC transporter, ATP-binding (e-43)	10580481	1
rjap2F4	2.1	544	spore coat protein precursor (e-8)	Vng0920h (e-36)	10580480	

II.E. *Haloarcula sinaiiensis*

Clone	l (Kb)	S (bp)	NCBI database	NRC-1 database	gi	D
usin1A2	3.8	505	excinuclease ABC subunit C (e-12)	*uvr*C: excision nuclease chain C (e-60)	10581790	334
rsin1A2	3.8	567	heat shock protein (e-16)	*hsp*4: heat shock protein (e-21)	10579777	
usin1A3	4.8	442				
rsin1A3	4.8	608				
usin1A4	4.8	506				
rsin1A4	4.8	496	phosphoribosyl formyl-glycinamidine synthase (e-30)	*pur*L: phosphoribosyl formyl-glycinamidine synthase (e-51)	10580429	
usin1A11	3.1	528				
rsin1A11	3.1	526		Vng6292c (e-30)*	10584330	
usin1A12	3.1	454	signal-transducing histidine kinase (e-46)	*htl*D: Htr-like protein (e-48)* ORF H0337 (e-48)†	10584098 10803577	293 ‡
rsin1A12	3.1	610		Vng6419h (e-19)*	10584442	
usin1B6	5.0	467	H+ transporting ATP synthase subunit (e-36)	*atp*D: H+ transporting ATP synthase subunit D (e-68)	10581550	4
rsin1B6	5.0	607	H+ transporting ATP synthase (e-107)	*atp*A: H+ transporting ATP synthase subunit A (e-109)	10581554	
usin1B9	3.3	491	alcohol dehydrogenase (e-9)			
rsin1B9	3.3	550	phosphoribosyl aminoimidazole carboxamide (e-41)	*pur*H: phosphoribosylaminoimidazole-succinocarboxamide formyl-transferase (e-70)	10580025	
usin1C1	4.8	474	hypothetical protein (e-12)	Vng1503c (e-49)	10580996	761
rsin1C1	4.8	536	dihydrodipicolinate synthase (e-19)	*dap*A: dihydrodipicolinate synthase (e-16)	10580053	
usin1C4	2.2	468		duplicate sin2D3		
rsin1C4	2.2	594				
usin1C7	2.9	522	isoaspartyl protein carboxyl methyltransferase (e-15)	Vng0094c (e-34)	10579743	
rsin1C7	2.9	599				

usin1C8	2.8	337		Vng0077h (e-24)	10579731	
rsin1C8	2.8	360				
usin1C9	2.8	436		duplicate sin1C8		
rsin1C9	2.8	387				
usin1D4	3.0	380		Vng2490h (e-27)	10581884	202
rsin1D4	3.0	580	UDP-glucose epimerase (e-45)	*galE2*: UDP-glucose 4-epimerase (e-93)	10579717	
usin1D7	1.6	388				
rsin1D7	1.6	467				
usin1D10	3.3	358				
rsin1D10	3.3	576	putative dehydratase (e-27)	*rspA*: starvation sensing protein (e-97)	10580052	
usin1E1		357	hypothetical protein (e-20)			
rsin1E1		503				
usin1F3	4.6	346	sugar ABC transporter (e-19)		10581074	307
rsin1F3	4.6	450		Vng1999h (e-17)	10581431	
usin1F4	4.6	370				
rsin1F4	4.6	491				
usin1F6	1.3	407	hypothetical protein (e-7)	Vng2454c (e-24)	10581856	
rsin1F6	1.3	551	farnesylgeranyl diphosphate synthatase (e-16)	*fps*: putative isopentenyl pyrophosphate isomerase (e-55)	10581569	243
usin1F8	3.7	375		*lfl2*: long-chain fatty-acid-CoA ligase (e-8)	10581487	
rsin1F8	3.7	549				
usin1F10	1.3	404	bacterioferritin comigratory protein (e-15)	*hyrA*: monooxygenase (e-23)	10581280	634
rsin1F10	1.3	393			10582022	
usin2A1	4.8	328				
rsin2A1	4.8	565		*prrC*: regulatory protein (e-37)	10579799	

usin2A2	4.8	291	ORF H0754 *Halobacterium* (e-31)	*mam*B: methylaspartate mutase (e-22)	10581700	
rsin2A2	4.8	465	ORF H0754 *Halobacterium* (e-30)	ORF H1607 (e-32)† ORF H0754 (e-32)† Vng6090c (e-32)*	10803687 10803616 10584151	‡
usin2A5	4.6	482	ATP-binding protein (e-19)	*phn*P: ATP-binding protein (e-54)	10581480	1050
rsin2A5	4.6	503	DNA binding protein (e-9)	Vng0725h (e-50)	10580304	
usin2A9	3.7	459	ORF H1373 *Halobacterium* (e-53)	ORF H1373 (e-54)†	10803663	
rsin2A9	3.7	481				
usin2A10	4.8	536	transcriptional regulator (e-8)		10581265	
rsin2A10	4.8	517	ABC transporter (e-23)			
usin2A12	2.6	506				
rsin2A12	2.6	457	mercuric reductase (e-20)	*lpd*A: dihydrolipoamide dehydrogenase (e-17)	10581636	
usin2B5	4.6	546	hypothetical protein (e-56)			
rsin2B5	4.6	562	SDH subunit B-homolog (e-21)	*sdh*B: succinate dehydrogenase subunit B (e-21)	10580826	
usin2B6	3.8	619	transposase (e-24)			
rsin2B6	3.8	615	thymidylate kinase (e-19)	Vng1031c (e-44)	10580584	
usin2C4	5.0	397	pyruvate dehydrogenase (e-18)	*pdh*A2: pyruvate dehydrogenase alpha subunit (e-38)	10581633	
rsin2C4	5.0	532	dihydrolipoamide acetyltransferase (e-47)	*dsa*: dihydrolipoamide S-acetyltransferase (e-61)	10581635	2
usin2C12	4.3	471	DNA helicase (e-16)	*hel*: DNA helicase (e-52)	10580995	
rsin2C12	4.3	627	methanol dehydrogenase (e-38)	*mox*R: methanol dehydrogenase regulatory protein (e-57)	10579869	1089
usin2D3	1.7	584	dihydrolipoamide dehydrogenase (e-10)	Vng6254c (e-23)*	10584299	0
rsin2D3	1.7	588		Vng6254c (e-16)*	10584299	
usin2D5	3.7	557		Vng2233c (e-16)	10581647	4
rsin2D5	3.7	485	hypothetical protein (e-27)	Vng2227c (e-46)	10581643	

usin2D6	4.6	516	glycine betaine transporter (e-28)	*opu*D: glycine betaine transporter (e-22)	10580626	715
rsin2D6	4.6	598	protein translation inhibitor (e-11)	Vng2099c (e-18)	10581514	
usin2D7	3.0	567	ferrodoxin oxidoreductase (e-6)	Vng0240c (e-51)	10579882	2
rsin2D7	3.0	563	ribosomal protein (e-20)	Vng0239c (e-35)	10579880	
usin2D11	3.3	538	hypothetical protein (e-6)	Vng6251h (e-13)*	10584298	
rsin2D11	3.3	457				
usin3A2	2.2	545	NADH dehydrogenase (e-27)	*ndh*G2: NADH dehydrogenase/ oxidoreductase (e-60)	10580160	1
rsin3A2	2.2	643	NADH dehydrogenase (e-23)	Vng0562c (e-86)	10580157	
usin3A3	3.0	562	keto deoxygluconate kinase (e-57)	*kdg*K: 2-keto-3-deoxygluconate kinase (e-59)	10579806	
rsin3A3	3.0	576				
usin3A7	3.7	600	response regulator protein (e-41)		10580475	48
rsin3A7	3.7	609	hypothetical protein (e-9)	Vng0854c (e-12)	10580422	
usin3A8	4.8	573	GTP-binding protein (e-21)	*gbp*1: GTP-binding protein (e-37)	10581205	
rsin3A8	4.8	619	*mcc*F from *Escherichia coli* (e-14)			
usin3A10	1.2	594	hypothetical protein (e-23)	*nop*56/58: archaeal nucleolar protein (e-50)	10580709	0
rsin3A10	1.2	560		*nop*56/58 archaeal nucleolar protein (e-24)	10580709	
usin3B2	4.3	579				
rsin3B2	4.3	414		*mrr*: restriction system protein (e-56)*	10584405	
usin3B3	3.0	623				
rsin3B3	3.0	608				
usin3B4	4.3	585	eukaryotic initiation factor of translation (e-82)	*eif*2g: translation initiation factor eIF-2 subunit gamma (e-84)	10581478	1
rsin3B4	4.3	455		Vng2059h (e-38)	10581479	
usin3C2	4.4	614	methylmalonyl-CoA mutase (e-23)	*mcm*A1: methylmalonyl-CoA mutase subunit alpha (e-37)	10580085	
rsin3C2	4.4	535	ABC transporter ATP binding (e-18)	*rbs*A: ribose ABC transporter ATP-binding (e-12)	10580461	321

usin3C9	3.8	634	hypothetical protein (e-6)	Vng1287c (e-26)	10580809	804
rsin3C9	3.8	592	proteasome, subunit alpha (e-31)	*psm*B: proteasome, subunit beta (e-56)	10579815	
usin3C10	4.2	625	hypothetical protein (e-10)	*boa*4: bacterio-opsin activator-like protein (e-8)	10580551	888
rsin3C10	4.2	549	hypothetical protein (e-6)	Vng2523h (e-10)	10581917	
usin3C11	3.9	579	hypothetical protein (e-10)	Vng0222c (e-19)	10579868	
rsin3C11	3.9	509				
usin3C12	1.2	575		*sdh*: succinate dehydrogenase (e-53)	10581522	0
rsin3C12	1.2	542	hypothetical protein (e-11)	*sdh*: succinate dehydrogenase (e-73)	10581522	
usin3D1	3.8	522		*ntp*: neutral proteinase (e-44)	10579693	1
rsin3D1	3.8	632		Vng0043h (e-7)	10579694	
usin3D5	1.8	628	ORFs H0754 and H1607 *Halobacterium* (e-41)	ORF H0754 (e-42)† ORF H1607 (e-42)† Vng6090c (e-42)*	10803616 10803687 10584151	‡
rsin3D5	1.8	561	TATA box binding protein - archael transcription factor (e-34)	*tbp*E: transcription initiation factor IID (e-55)	10581658	

III. QUERIES DONE ON PUBLISHED GENOME SEQUENCES

Compared genome sequences	Similarity in gene order (50 random fragments of 3 genes,n=100	Similarity in gene order (entire genome)	Sequence similarity
Archaeoglobus fulgidus **vs** *Methanococcus jannaschii*	0.0677	0.1018	0.3318
Bacillus subtilis **vs** *Mycoplasma pneumoniae*	0.0296	0.0468	0.2868
Bacillus subtilis **vs** *Thermotoga maritima*	0.0706	0.0952	0.2999
Borrelia burgdorferi **vs** *Treponema pallidum*	0.1114	0.1527	0.2995
Campylobacter jejuni **vs** *Helicobacter pylori* 26695	0.2565	0.3238	0.4010
Campylobacter jejuni **vs** *Helicobacter pylori* J99	0.2584	0.3286	0.4009
Chlamydia muridarumus **vs** *Chlamydia pneumoniae*	0.8058	0.8594	0.6102
Chlamydia muridarumus **vs** *Chlamydia trachomatis*	0.8920	0.8973	0.8394
Chlamydia pneumoniae **vs** *Chlamydia trachomatis*	0.7122	0.7605	0.6107
Chlamydia pneumoniae **vs** *Haemophilus influenzae*	0.0816	0.1122	0.2721
Deinococcus radiodurans **vs** *Thermotoga maritima*	0.0480	0.0667	0.2562
Haemophilus influenzae **vs** *Escherichia coli*	0.3776	0.4851	0.5393
Helicobacter pylori 26695 **vs** *Haemophilus influenzae*	0.0697	0.0914	0.2806
Helicobacter pylori 26695 **vs** *Helicobacter pylori* J99	0.8053	0.9035	0.9059
Methanobacterium thermoautotrophicum **vs** *Methanococcus jannaschii*	0.1202	0.1482	0.3854
Mycobacterium tuberculosis **vs** *Aquifex aeolicus*	0.0178	0.0296	0.2611

Mycobacterium tuberculosis **vs** *Mycoplasma pneumoniae*	0.0184	0.0286	0.2586
Mycoplasma genitalium **vs** *Mycoplasma pneumoniae*	0.9492	0.9722	0.6925
Neisseria meningitidis **vs** Escherichia coli	0.1438	0.1901	0.3980
Neisseria meningitidis **vs** *Haemophilus influenzae*	0.1137	0.1559	0.3905
Neisseria meningitidis **vs** Rickettsia prowazekii	0.0463	0.0659	0.3080
Pyrococcus horikoshii **vs** *Pyrococcus abyssi*	0.6193	0.6623	0.2262
Rickettsia prowazekii **vs** *Escherichia coli*	0.1534	0.2002	0.3034
Rickettsia prowazekii **vs** *Haemophilus influenzae*	0.1222	0.1655	0.3070
Saccharomyces cerevisiae **vs** *Pyrococcus horikoshii*	0.0002	0.0013	0.2068
Thermotoga maritima **vs** *Archaeoglobus fulgidus*	0.0801	0.1040	0.2391
Ureaplasma urealyticum **vs** *Bacillus subtilis*	0.2120	0.2733	0.3072
Ureaplasma urealyticum **vs** *Mycoplasma genitalium*	0.0696	0.0867	0.3284
Ureaplasma urealyticum **vs** *Mycoplasma pneumoniae*	0.2725	0.3306	0.3392
Vibrio cholerae **vs** *Escherichia coli*	0.2746	0.3493	0.4444
Vibrio cholerae **vs** *Haemophilus influenzae*	0.1460	0.1967	0.5057
Vibrio cholerae **vs** *Pseudomonas aeruginosa*	0.2257	0.2733	0.3743
Vibrio cholerae **vs** *Xylella fastidiosa*	0.1115	0.1489	0.3268