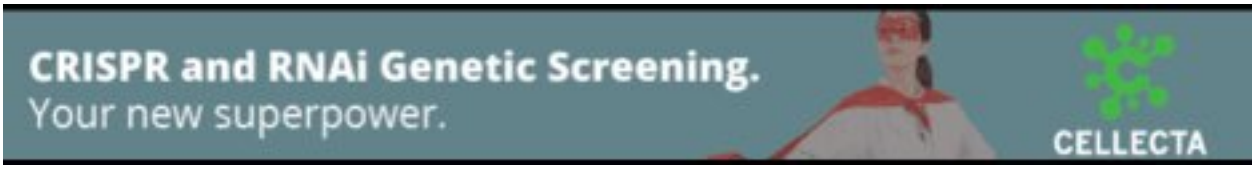
Snapshot of a Large Dynamic Replicon in a Halophilic Archaeon: Megaplasmid or Minichromosome?

WaiLap V. Ng, Stacy A. Ciufo, Todd M. Smith, et al.

Genome Res. 1998 8: 1131-1141
Access the most recent version at doi:10.1101/gr.8.11.1131



RESEARCH

Snapshot of a Large Dynamic Replicon in a Halophilic Archaeon: Megaplasmid or Minichromosome?

WaiLap V. Ng, Stacy A. Ciufo,[1] Todd M. Smith, Roger E. Bumgarner, Dale Baskin, Janet Faust, Barbara Hall, Carol Loretz, Jason Seto, Joseph Slagel, Leroy Hood,[2] and Shiladitya DasSarma[1,2]

Department of Molecular Biotechnology, University of Washington, Seattle Washington 98195 USA; [1]Department of Microbiology, University of Massachusetts, Amherst Massachusetts 01003 USA

Extremely halophilic archaea, which flourish in hypersaline environments, are known to contain a variety of large dynamic replicons. Previously, the analysis of one such replicon, pNRC100, in *Halobacterium* sp. strain NRC-1, showed that it undergoes high-frequency insertion sequence (IS) element-mediated insertions and deletions, as well as inversions via recombination between 39-kb-long inverted repeats (IRs). Now, the complete sequencing of pNRC100, a 191,346-bp circle, has shown the presence of 27 IS elements representing eight families. A total of 176 ORFs or likely genes of 850-bp average size were found, 39 of which were repeated within the large IRs. More than one-half of the ORFs are likely to represent novel genes that have no known homologs in the databases. Among ORFs with previously characterized homologs, three different copies of putative plasmid replication and four copies of partitioning genes were found, suggesting that pNRC100 evolved from IS element-mediated fusions of several smaller plasmids. Consistent with this idea, putative genes typically found on plasmids, including those encoding a restriction-modification system and arsenic resistance, as well as buoyant gas-filled vesicles and a two-component regulatory system, were found on pNRC100. However, additional putative genes not expected on an extrachromosomal element, such as those encoding an electron transport chain cytochrome d oxidase, DNA nucleotide synthesis enzymes thioredoxin and thioredoxin reductase, and eukaryotic-like TATA-binding protein transcription factors and a chromosomal replication initiator protein were also found. A multi-step IS element-mediated process is proposed to account for the acquisition of these chromosomal genes. The finding of essential genes on pNRC100 and its property of resistance to curing suggest that this replicon may be evolving into a new chromosome.

[The sequence data described in this paper have been submitted to GenBank under accession no. AF016485.]

Both archaeal and bacterial prokaryotes generally contain multiple circular replicons from only a few kilobases to several megabases in their genomes. Since early genetic work on *Escherichia coli* established the dogma of a single circular chromosome in prokaryotes, essentially all smaller replicons in cells have been relegated to the status of extrachromosomal elements or plasmids, irrespective of the species (Drlica and Riley 1990). Of the many plasmids that have been characterized, several are megaplasmids from a hundred kilobases to megabase sizes, including F (fertility) and R (resistance) factors, and toxin-bearing plasmids in *E. coli* and other pathogenic bacteria (Proter 1991), tumor-inducing plasmids and symbiotic plasmids in *Agrobacterium* and *Rhizobium* species (Van Larebeke et al. 1974; Banfalvi et al. 1985), and a variety of aromatic hydrocarbon degradation plasmids in *Pseudomonas* species and related bacteria (Franz and Chakrabarty 1986; Choudhary et al. 1997; Mouncey et al. 1997). Over the last few years, many of these and other large replicons have been studied and some found to contain genes that are normally thought to be essential for cell viability. The most striking example is for *Rhodobacter spheroides*, which contains a 900-kb replicon with two copies of the rRNA operon and a variety of other essential genes (Suwanto and Kaplan 1989; Choudhary et al. 1997). This finding led to the suggestion that prokaryotic genomes may

[2]Corresponding authors.
E-MAIL dassarma@microbio.umass.edu; FAX (413) 545-1578;
E-MAIL leehood@u.washington.edu; FAX (206) 616-5197.

be composed of multiple essential replicons and opened the possibility that some replicons originally classified as megaplasmids may in fact be chromosomes (Allardet-Servent et al. 1993; Michaux et al. 1993; Zuerner et al. 1993; Cheng and Lessie 1994). However, chromosomal status may depend not just on the occurrence of essential genes, but also on other criteria such as size, copy number and replication control, and evolutionary history.

The recent advent of high throughput sequencing technology has provided an outstanding opportunity for detailed analysis of prokaryotic genomes (Fleischmann et al. 1995; Fraser et al. 1995, 1997; Bult et al. 1996; Himmelreich et al. 1996; Kaneko et al. 1996; Sensen et al. 1996; Blattner et al. 1997; Klenk et al. 1997; Kunst et al. 1997; Smith et al. 1997; Deckert et al. 1998; Cole et al. 1998). In addition to permitting comparative phylogenetic analysis of individual genes across a wide spectrum of organisms, these studies also promise a better understanding of genome evolution and dynamics, including greater insights into the relationships between replicons resident within individual species. With the evolution of prokaryotic genomes in mind, we have focused on the extremely halophilic archaeon, *Halobacterium* sp. strain NRC-1, which grows optimally at a nearly saturated (4.5 M) NaCl concentration (Vreeland and Hochstein 1993; DasSarma and Fleischmann 1995). The genome of *Halobacterium* NRC-1 is notable for the presence of dynamic replicons and a variety of transposable elements that give rise to frequent DNA rearrangements (Charlebois and Doolittle 1989; DasSarma 1993). The genome of NRC-1 contains a circular 2-Mb chromosome and two other large replicons, pNRC100 and pNRC200 (Hackett et al. 1994). We have focused on pNRC100, a 191-kb replicon, which was restriction mapped and shown to contain at least 17 insertion sequence (IS) elements and large (>35 kb) inverted repeats mediating inversion isomerization (Ng et al. 1991). pNRC100 was also shown to contain a cluster of genes specifying buoyant intracellular gas-filled vesicles (Halladay et al. 1992, 1993; DasSarma et al. 1994). Because gas vesicles are necessary for flotation, which enhances both aerobic respiration and photophosphorylation, the finding of these important genes on an extrachromosomal element was surprising.

Here, we report the complete nucleotide sequence of pNRC100 and analysis of its IS elements and coding capacity. We also discuss a hypothetical model for its evolution and possible justification for its status as a minichromosome.

RESULTS

pNRC100 Sequencing and Assembly

Sequencing was conducted on shotgun libraries of *Halobacterium* sp. strain NRC-1 plasmid DNA and on cloned and ordered *Hin*dIII fragments of pNRC100. The initial period of automated sequence assembly produced nine contigs representing pNRC100 ranging from 1 to 69 kb. The assembly process was challenging because of the presence of a total of 27 copies of IS elements (17% of the replicon; Table 1) and the large (~39 kb) inverted repeat (IR). The contigs containing two or more physically unlinked segments of pNRC100 sequence were resolved by comparison of regions of IS element heterogeneities (e.g., for ISH3, ISH7, and ISH8; Charlebois and Doolittle 1989; DasSarma 1993) and alignment of known target-site duplications flanking IS elements (Ng et al. 1991) manually by use of the FINDPATTERNS, BESTFIT, SEQED, and WORDSEARCH pro-

Table 1. IS Elements in pNRC100

Element	Homolog(s)[a]	Size (kb)	Copy no.	Comments
ISH2	—	0.5	4	lacks ORF
ISH3	ISH27, ISH51	1.4	8	heterogeneous
ISH4	ISH50	1.0	1	
ISH5	—	1.5	2	interrupted by ISH11
ISH7	ISH24	2.9/3.3	2	heterogeneous
ISH8	ISH26	1.4	6	heterogeneous
ISH9	ISH28	0.9	2	
ISH11	—	1.1	2	

[a]Homologs are similar but nonidentical IS elements reported previously in other halophiles (Charlebois and Doolittle 1989; DasSarma 1993). ISH51 was reported from *Haloferax volcanii,* and the other ISH elements listed were from *Halobacterium* strains.

grams in the GCG software package (Devereux et al. 1984). End sequencing of the ordered pNRC100 *Hin*dIII fragment library (Ng et al. 1991) provided a scaffold for assembly. Appropriate segments of chimeric contigs were dissected and reassembled into a circular 191,346-bp sequence (GenBank Accession no. AF016485).

pNRC100 Putative Genes and Gene Products

A total of 1,965 individual ORFs 15 bp or larger were identified and analyzed by determination of their size, location, GC composition, codon third position GC bias, and isoelectric point of the predicted protein product by use of the GCG software package. These criteria were used to select 176 probable genes, 39 of which were repeated in the large IR (Fig. 1). Because one H0761 copy on the IR is interrupted by an ISH2 element, resulting in two smaller ORFs, the number of different ORFs is 136. Our analysis indicated that 72% of the replicon is coding and 28% is noncoding.

Of the 136 different probable gene products encoded by pNRC100, 62 or about 46% had statistically significant hits to the databases (Table 2). Of these, 14 hits were to gas vesicle proteins, 8 were to replication or partitioning proteins, 7 were to regulatory proteins, 4 were to transcription factors, 4 were to membrane components, 2 were to redox proteins, 2 were to heavy metal resistance proteins, 2 were to DNA endonucleases, and 1 was to a helicase. The remaining 18 hits were to putative transposases encoded by IS elements or proteins with unknown functions. Of the closest matches, 36 were to archaeal proteins, 23 were to bacterial proteins, and 3 were to eukaryotic proteins. Thus, like other archaea, *Halobacterium* clearly has both bacterial features, such as metabolic proteins, and eukaryotic features, such as the transcription system (Koonin et al. 1997). An additional interesting finding was that

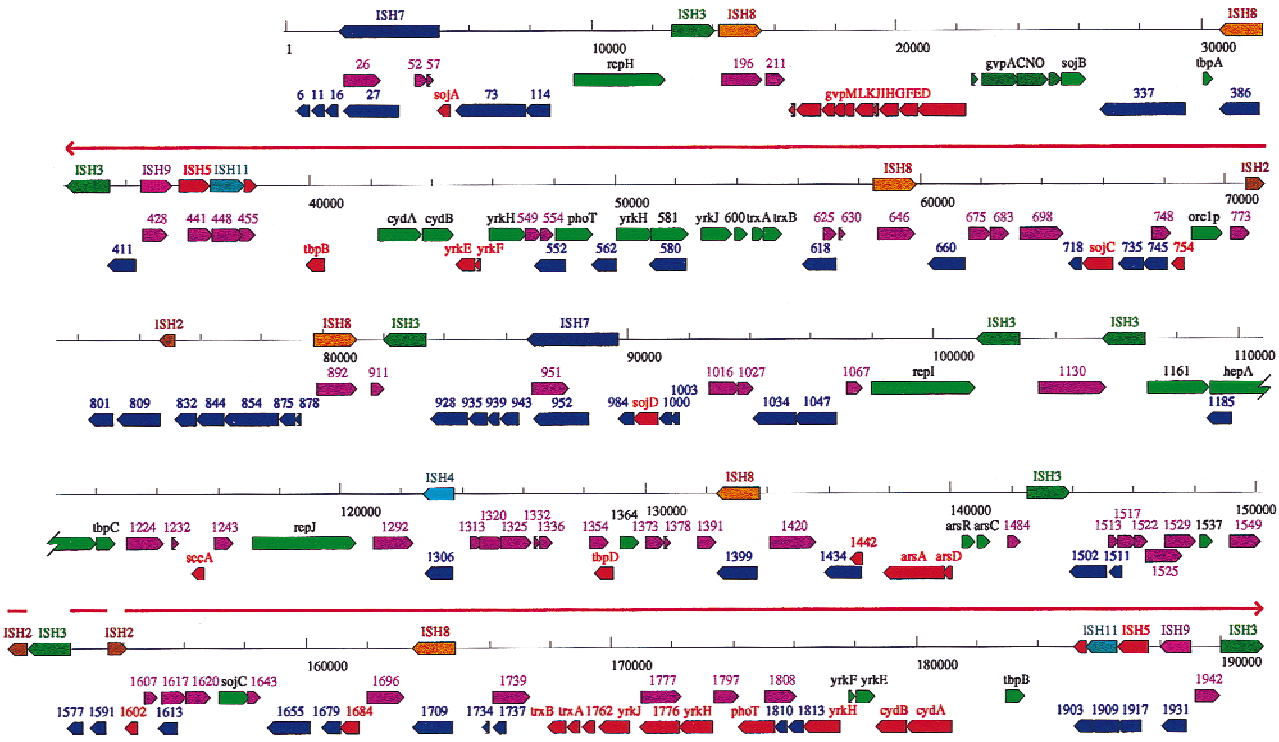


Figure 1 Linear representation of the genes, ORFs, and IS elements of the circular plasmid pNRC100. The plasmid is divided into five sections, the small single-copy region, the right IR (red arrow), the large single-copy region (divided into two sections), and the left IR (red arrow). IS elements (arrows indicating orientation) are shown along with a scale in nucleotides and genes and ORFs on each DNA strand are shown below. Identities and orientation of genes (red or green arrows) and ORFs (purple or blue arrows) are indicated. All 1965 pNRC100 ORFs five codons or larger have been assigned numerical designations on the basis of location; however, only those ORFs judged to represent probable genes are shown for clarity. Additional information is available at the following Web site: http://chroma.mbt.washington.edu/seq_www/.

Table 2. Summary of 62 Probable pNRC100 Gene Products with Homologs in GenPeptides Database

ORFs[a]	Proteins	Organisms	Homolog ID	Percent similarities	Percent identity
Membrane components					
H0511	CydA	*Synechocystis* sp.	gi-1652262	44.3	33.3
H0520	CydB	*Mycobacterium tuberculosis*	gi-2113906	35.0	27.8
H1238	SecA	*Bacillus subtilis*	gi-216334	37.5	25.8
H0556	PhoT	*Haemophilus influenzae*	gi-1574446	33.1	24.6
Endonucleases					
H1161	*Eco*571	*Escherichia coli*	gi-41321	49.6	30.4
H1537	*Cpa*I	*Chlamydomonas pallidostigmatica*	gi-845300	37.1	26.7
Gas vesicle proteins					
H0290	GvpA	*Halobacterium halobium*	gi-43530	100.0	100.0
H0293	GvpC	*H. halobium*	gi-43531	100.0	100.0
H0263	GvpD	*H. halobium*	gi-43504	100.0	100.0
H0255	GvpE	*H. halobium*	gi-43505	100.0	100.0
H0243	GvpF	*H. halobium*	gi-43519	99.5	99.5
H0240	GvpG	*H. halobium*	gi-455303	100.0	100.0
H0237	GvpH	*H. halobium*	gi-43521	100.0	100.0
H0232	GvpI	*H. halobium*	gi-43522	95.7	95.7
H0230	GvpJ	*H. halobium*	gi-43523	100.0	100.0
H0228	GvpK	*H. halobium*	gi-43524	100.0	100.0
H0221	GvpL	*H. halobium*	gi-43525	98.2	98.0
H0219	GvpM	*H. halobium*	gi-43526	100.0	100.0
H0310	GvpN	*H. halobium*	gi-43655	98.3	98.3
H0321	GvpO	*H. halobium*	gi-1154780	64.5	59.1
Heavy metal resistance proteins					
H1450	ArsA	*E. coli*	gi-151857	45.1	34.5
H1477	ArsC	*E. coli*	gi-1789916	38.6	33.3
Helicase					
H1186	HepA	*E. coli*	gi-1786245	33.7	23.5
Redox proteins					
H0606	TrxA	*Chlamydia psittaci*	gi-666879	54.5	42.6
H0610	TrxB	*Mycobacterium leprae*	gi-886315	57.3	49.5
Regulatory proteins					
H0337	protein kinase	*E. coli*	gi-606149	36.4	22.8
H1442	Arsefg	*B. subtilis*	gi-1881340	38.7	28.0
H1462	ArsD	*Salmonella typhimurium/E. coli*	gi-1061415	40.7	34.3
H1471	ArsR	*E. coli*	gi-1789916	38.9	33.3
H0675	ferric uptake	*Neisseria meningitidis*	gi-437629	34.1	23.0
H0562	bat	*H. halobium*	gi-148753	33.5	24.4
H0057	YorfE	*Streptococcus pneumoniae*	gi-1536960	56.3	37.5
Replication and partitioning proteins					
H0761	Orc1p	*Schizosaccharomyces pombe*	gi-1163108	35.2	24.0
H0136	RepH	*H. halobium*	gi-305350	100.0	100.0
H1080	RepI	*H. halobium*	gi-305350	43.9	34.8
H1260	RepJ	*H. halobium*	gi-305350	44.1	34.1
H0066	SojA	*B. subtilis*	gi-916914	39.8	35.7
H0324	SojB	*Methanococcus jannaschii*	gi-1522661	43.7	28.0
H0722	SojC	*M. jannaschii*	gi-1522661	34.4	23.3
H0991	SojD	*B. subtilis*	gi-580906	38.5	27.8

Table 2. (*Continued*)

ORFs[a]	Proteins	Organisms	Homolog ID	Percent similarities	Percent identity
Transcription factors					
H0378	TbpA	*Pyrococcus woesei*	gi-498649	51.5	36.6
H0486	TbpB	*P. woesei*	gi-498649	54.1	41.6
H1211	TbpC	*P. woesei*	gi-498649	41.8	32.2
H1356	TbpD	*P. woesei*	gi-498649	55.1	44.3
Putative transposases					
H0448	ISH11 ORF	*H. halobium*	gi-43508	100.0	100.0
H0196	ISH26 ORF[b]	*H. halobium*	gi-43511	93.9	92.5
H0386	ISH26 ORF[b]	*H. halobium*	gi-43511	93.9	93.2
H0646	ISH26 ORF[b]	*H. halobium*	gi-43511	87.7	86.4
H0892	ISH26 ORF[b]	*H. halobium*	gi-43511	85.0	81.6
H1399	ISH26 ORF[b]	*H. halobium*	gi-43511	88.4	85.7
H0428	ISH28 ORF	*Halobacterium salinarum*	gi-1353676	100.0	100.0
H1306	ISH50 ORF	*H. halobium*	gi-140827	81.2	80.5
Unknown functions					
H0581	C06004	*Sulfolobus solfataricus*	gi-1707683	43.8	31.2
H0660	C05008	*S. solfataricus*	gi-1707831	52.4	40.5
H0754	MJ0660	*M. jannaschii*	gi-1591373	45.5	37.7
H0600	MJ1503	*M. jannaschii*	gi-1500391	41.7	25.0
H1364	YGL101w	*Saccharomyces cerevisiae*	gi-1322641	36.2	25.5
H0526	YrkE	*B. subtilis*	gi-1303704	41.3	31.6
H0532	YrkF	*B. subtilis*	gi-1303705	55.1	43.6
H0539	YrkH	*B. subtilis*	gi-1303707	46.2	35.6
H0571	YrkH	*B. subtilis*	gi-1303707	47.4	34.2
H0594	YrkJ	*B. subtilis*	gi-1303709	36.8	25.6

Additional information can be found on the following web site: http://chroma.mbt.washington.edu/seq_www/.
[a]Only a single copy of ORFs in inverted repeats is listed.
[b]Similarity/identity between the ISH26 ORFs are indicated.

like most characterized proteins of halophiles, the great majority (91%) of the pNRC100 probable gene products are acidic (Fig. 2), a feature likely to be important for function at high salt concentrations known to be present in their cytoplasm (Lanyi 1974; Dym et al. 1995).

Among the most interesting putative gene products encoded by pNRC100 were the following: (1) H0378, H0486, H1211, and H1356 are probably transcription factors similar to the eukaryotic and archaeal general transcription factor TBP (TATA-binding protein; Thomm 1996). The multiplicity of TBPs suggested by this finding is highly unusual. However, not all of them may be functional. For example, H0378 appears to be only one-half the size of the others. (2) H0761, H0136, H1080, and H1260 are similar to replication proteins. H0761 is in the Orc1p family of eukaryotic origin-binding proteins (Gavin et al. 1995) whereas the other three are members of a family of replication proteins in halophiles (Ng and DasSarma 1993). The gene encoding

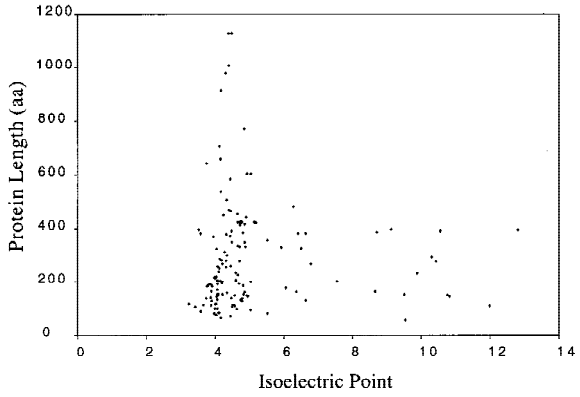


Figure 2 Plot of calculated isoelectric point versus length of 136 predicted proteins encoded by pNRC100. Isoelectric points were calculated by use of the GCG isoelectric program (Devereux et al. 1984). The range of calculated pI was 3.24–12.81 and the average pI was 5.05. The range of sizes was 56–1128 amino acids and the average size was 290 amino acids.

H0136, *repH*, was shown previously to be required for replication of pNRC100 minireplicons (Ng and DasSarma 1993). (3) H0066, H0324, H0722, and H0991 are similar to the ParA family in bacteria and to Soj in *Bacillus subtilis* and *Methanococcus jannaschii*, which are involved in plasmid and chromosomal partitioning, for orderly segregation of replicated molecules into daughter cells (Wheeler and Shapiro 1997). (4) H0511 and H0520 are similar to subunits I and II of *E. coli* cytochrome d oxidase, a terminal electron acceptor in the electron transport chain (Miller and Gennis 1983). (5) H0606 and H0610 are similar to thioredoxin and thioredoxin reductase, which are used for reduction of ribonucleotides to deoxyribonucleotides for DNA synthesis (Neuhard and Nygaard 1987). (6) H0337 is similar to several protein kinases of two-component regulatory systems, including both components of the FixL/J system probably involved in oxygen regulation (David et al. 1988; Lois et al. 1993). (7) H1450 and H1477 are similar to arsenic-resistance proteins, with the former similar to the catalytic subunit of the arsenite pump, and the latter similar to arsenate reductase (Rosen et al. 1988, 1991). (8) H1161 is similar to the type IV restriction-modification system protein *Eco*57I which contains both restriction and modification activity (Janulaitis et al. 1992). (9) H1537 is similar to an intron encoded endonuclease (Turmel et al. 1995). (10) Over a dozen Gvps are thought to be involved in gas-vesicle formation and have been studied genetically (Halladay et al. 1992, 1993; DasSarma et al. 1994).

DISCUSSION

The probable genes on pNRC100 with statistically significant homologs in the databases fall into three categories. Some genes appear to be typical for plasmids and extrachromosomal elements: for example, replication and partitioning genes, and arsenic resistance. Other genes, for example the gas vesicle genes and a number of regulatory genes, are not typically extrachromosomal but their extrachromosomal location is not surprising given that they may be dispensable for cell viability. Quite unexpectedly, however, another set of genes appears to encode essential proteins. For example, *cydA* and *cydB* encode a cytochrome d oxidase, which is used for the terminal step in respiration, and *trxA* and *trxB* encode thioredoxin and thioredoxin reductase enzymes, which participate in the pathway for synthesis of deoxyribonucleotides. These enzymes are involved in fundamental cellular processes and appear to be required for cell viability. Moreover, Southern hybridization analysis showed that these are unique genes present only on pNRC100 and are not repeated elsewhere in the genome (data not shown). Transcription factors, such as TBPs, four of which are encoded by pNRC100 (*tbpA*, *tbpB*, *tbpC*, and *tbpD*), may also be required under conditions in which they are used for transcription of critical genes. Therefore, pNRC100 appears to be a replicon with unique essential genes expected to reside on a chromosome rather than a plasmid.

The finding of chromosomal genes on pNRC100 raises the question of the mechanism of acquisition of these genes (Fig. 3). One possibility is that a smaller plasmid had integrated into the *Halobacterium* chromosome in the past, perhaps mediated by an IS element, and had subsequently excised imprecisely capturing a segment of chromosomal genes in the process. This type of process has been shown to be used to generate F′ plasmids in *E. coli* (Holloway and Low 1987). The IRs of pNRC100 would be prime candidates for recent acquisition from the chromosome as the cytochrome d oxidase (*cydA* and *cydB*) and thioredoxin/ thioredoxin reductase (*trxA* and *trxB*) genes, as well as one of the TBP genes (*tbpB*) are all located there. The presence of a eukaryotic-like chromosomal replication initiator protein gene (H0761) in the IRs is also intriguing in this respect (Gavin et al. 1995). This region also contains a large number of putative genes with homology to genes present on the chromosomes of other microorganisms, for example, the *B. subtilis yrkE*, *yrkF*, *yrkH*, and *yrkJ* genes, which are of unknown functions (Kunst et al. 1997), and HI1604, a *Haemophilus influenzae* putative phosphate transport gene (Fleischmann et al. 1995). In addition to bearing what appear to be chromosomal genes, the IR region is also relatively GC rich (Fig. 3) and IS-element poor, both characteristics of the *Halobacterium* chromosome (Charlebois and Doolittle 1989; DasSarma 1993; Hackett et al. 1994). An additional interesting possibility is that the inverted duplication of genes in the IR region may be important for stability. Duplicated genes present in inverted orientation are less likely to be lost by deletion and inactivated by deleterious mutation, as a result of repair by copy choice mechanisms. Analogous functions have been proposed for the large IRs in several chloroplast chromosomes, which are of similar size and arrangement to pNRC100 (Palmer 1985).

On the basis of the arrangement of IS elements and genes on pNRC100, we have hypothesized the pathway for evolution of the replicon. The original events leading to formation of the large IRs in pNRC100 probably involved multiple IS element-

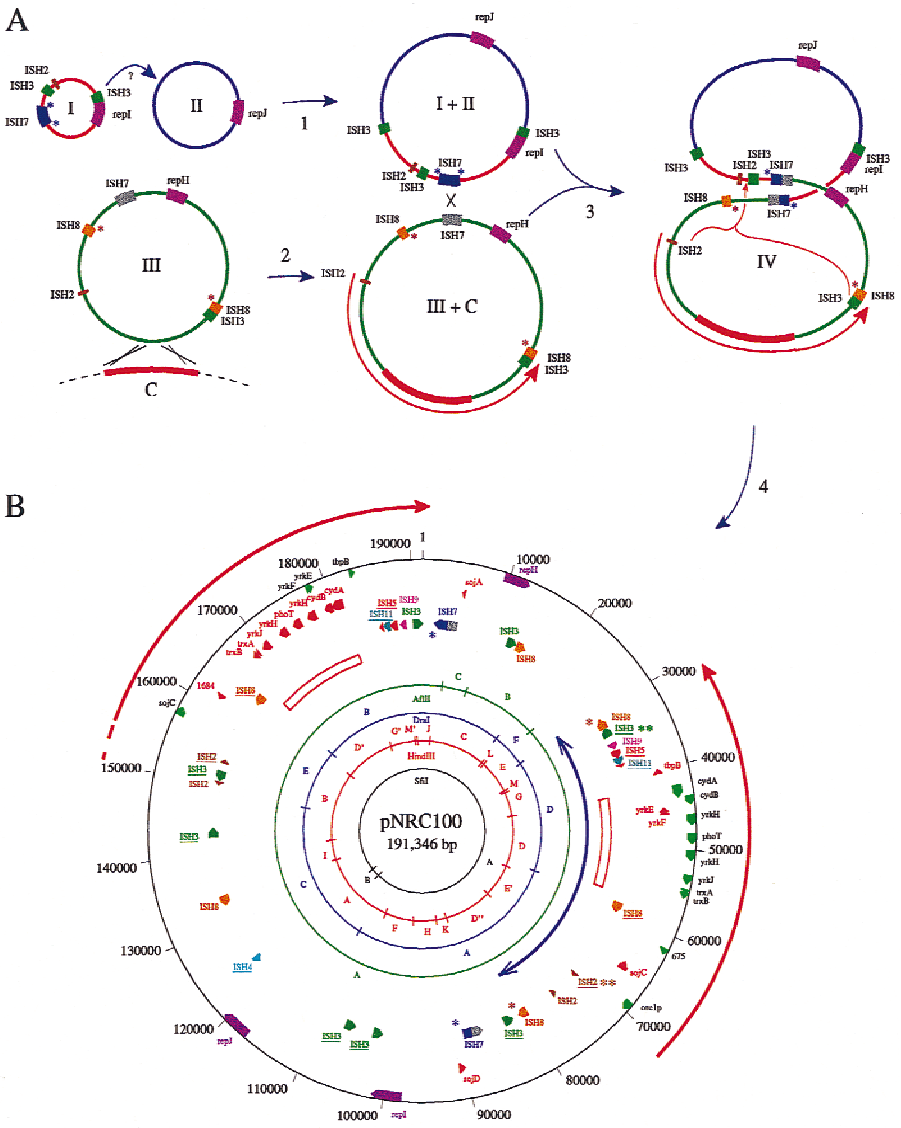


Figure 3 (*A*) Hypothetical pathway for evolution of pNRC100 (not drawn to scale). Step 1: Two small plasmids (I + II) fused, probably mediated by an ISH3 element, to form a larger plasmid (I + II) representing most of the large single-copy region of pNRC100. Step 2: A region of the chromosome (C, thick red line), 15 kb or larger was acquired on a plasmid (III) by integration into and aberrant excision out of the chromosome. The acquired chromosomal region became associated with a 50-kb composite transposon bounded by IRs of ISH8 on a large plasmid (III + C) containing one copy of the 39-kb IR region (red arrow) and most of the small single-copy region of pNRC100. Step 3: The intermediate plasmids fused, probably through ISH7-mediated homologous or site-specific recombination. Step 4: The IR region of pNRC100 was duplicated by a gene conversion-like mechanism (indicated by large orange bracket with arrow) involving ISH2 and ISH3 elements at the ends of the IRs and between the small and large single-copy regions. Subsequent insertion of an ISH2 and ISH3 element into one copy of the IRs is not shown.The relative locations of *rep* genes, IS elements, and the acquired chromosomal region are indicated by colored boxes. Heterogeneity of the two ISH7 copies is indicated by shading. Target-site duplications are indicated by asterisks. (*B*) Circular representation of general chromosomal features of pNRC100. The large IRs are indicated on the outside by red arrows and the scale is indicated on the outer circle in nucleotides. The position and orientation of genes are indicated by wide colored arrows. The locations of 15-kb GC-rich (64% G+C) regions within the IRs are indicated by open rectangles. The concentric circles near the center indicate the *Sfi*I, *Hin*dIII, *Dra*I, and *Afl*II restriction maps. Restriction fragments are labeled according to size. IS elements flanked by target-site duplication sequences are underlined. Only the genes and ORFs involved in duplicated plasmid features are indicated. (Two-headed blue arrow) Location of a putative 50-kb transposon.

mediated rearrangements. The present location of one copy of the IRs may be explained by the transposition of a ~50 kb composite transposon. This putative transposon is bounded by IRs of ISH8, which are flanked by short direct repeats (5′-CGTATCGGAG-3′), suggestive of a recent transposition event (asterisks and double arrows in Fig. 3B). Both copies of the IRs are bounded by two other IS elements, ISH2 and ISH3, which were probably involved in the duplication events resulting in formation of the IRs. A possible mechanism for duplication is via gene conversion (or gap repair) of sequences located between the ISH2 and ISH3 elements flanking the original copy of the IRs to another pair of elements located between the small and large single-copy regions (step 4, Fig. 3). Either site-specific or homologous recombination within ISH3 and ISH2 element pairs may have initiated the conversion process, involving either a single molecule or sister plasmids. Such a mechanism is consistent with the finding of short direct repeats flanking one copy of the terminal ISH2 and ISH3 elements but not the other copy of the elements (double asterisks in Fig. 3B). After duplication, two additional transposition events also occurred, insertion of ISH2 and ISH3 elements near one end of one copy of the IR, and resulted in the observed heterogeneity in the IRs. The lack of other sequence heterogeneity between the two copies of IRs and the report of a pNRC100-like plasmid lacking the IRs in another *Halobacterium* strain suggests that the inverted duplication occurred recently in the evolutionary history of NRC-1 (Pfeifer et al. 1981).

The events preceding the acquisition of chromosomal genes and duplication of the IRs in evolution of pNRC100 were likely to be fusions of three smaller plasmids. One replicon fusion was probably mediated by recombination between two nonidentical ISH7 elements present on distinct replicons (I + II and III + C in Fig. 3A). This possibility is suggested by the occurrence of short direct repeats (5′-CGAAGCG-3′) flanking the two copies of ISH7 in pNRC100, which are 85 kb apart, one located to the left of the ISH7 element in the small single-copy region and the second located to the right of the ISH7 element in the large single-copy region (asterisks in Fig. 3A). The smaller of the two replicons hypothesized may have been formed by another replicon fusion, a step suggested by the presence of two very similar replication genes, *repI* and *repJ*, in the large single-copy region of pNRC100. Although a specific mechanism is not obvious in this step, the presence of several ISH3 elements in this region suggests their involvement in the fusion process.

The sequence of pNRC100 has provided valuable insights into the structure and evolution of a dynamic replicon in an unstable archaeal genome. A total of 136 probable genes (with 39 duplicated in the inverted repeat) and 27 IS elements have been found. On the basis of the arrangement of IS elements and flanking sequences, we have been able to hypothesize a series of recombinational events that explains the evolution of pNRC100. Several replicon fusions occurred to form a large and complex plasmid. During this process, one plasmid probably inserted into the resident chromosome and excised aberrantly, taking with it a number of genes that were necessary for viability. To stabilize the required chromosomal genes, a 39-kb segment of pNRC100 was duplicated to form IRs, a structure reminiscent of the chloroplast chromosome.

The finding of essential genes on pNRC100 raises an interesting question on the precise distinction between plasmids and chromosomes. Classically, prokaryotic genomes have been thought as being composed of a single large (megabase-plus size) chromosome containing all essential genes, and a diversity of small (<~100 kb in size) multicopy plasmids containing accessory genes. Plasmids have been shown to recombine with each other or with the chromosome, but are not thought to be involved in the formation of new chromosomes. Not fully consistent with these ideas are pNRC100 and a variety of other essential replicons that have been studied (Van Larebeke et al. 1974; Banfalvi et al. 1985; Franz and Chakrabarty 1986; Suwanto and Kaplan 1989; Allardet-Servent et al. 1993; Michaux et al. 1993; Zuerner et al. 1993; Cheng and Lessie 1994; Choudhary et al. 1997; Mouncey et al. 1997). Many of these seem to occupy an intermediate status between plasmids and chromosomes and may represent evolutionary intermediates in the formation of new chromosomes (or the breakdown of old ones). More detailed analysis of such replicons, including the elucidation of their copy number and replication control as well as their evolutionary history may provide a more meaningful distinction between megaplasmids and minichromosomes. Such scrutiny will likely lead to a deeper understanding of the mechanisms of prokaryotic genome evolution.

METHODS

Plasmid Preparation, Library Construction, and Sequencing

A modified Currier and Nester procedure followed by CsCl–

ethidium bromide equilibrium gradient centrifugation was used to purify covalently closed circular DNA from *Halobacterium* NRC-1 (Ng et al. 1995). A shotgun library was prepared from the purified DNA, which consisted largely of pNRC100, and minor quantities of deletion derivatives of pNRC100, as well as the larger resident plasmid, pNRC200, by sonic shearing and cloning of 1- to 2.5-kb fragments into the *Sma*I site of M13mp18 (Messing 1983). Sequencing was also conducted at the ends of the cloned *Hin*dIII-A to -K fragments of pNRC100 and on shotgun libraries of the pNRC100 *Hin*dIII fragments cloned in M13mp18 (Ng et al. 1991). A total of 4,606 sequencing reactions were conducted on pNRC100 subclones by use of fluorescent dye primers or dideoxy-terminators, or both and analyzed on ABI 373 and 377 sequencers (Smith et al. 1986).

Sequence Assembly

The sequencing results were assembled by use of the PHRED/PHRAP/CONSED base-calling and sequence-assembly software (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). The contig sequences were analyzed and merged together by use of FINDPATTERNS, BESTFIT, SEQED, and WORDSEARCH programs in the GCG software package (Devereux et al. 1984).

Sequence Analysis

The complete pNRC100 nucleotide sequence was analyzed for genes by use of a modified GCG FRAMES program (S. Ciufo and S. DasSarma, unpubl.). ORFs 15 bp or larger were identified and analyzed by determining their size, location, GC composition, codon third position GC bias, and isoelectric point of the predicted protein product by use of the GCG software package. The pNRC100 nucleotide sequence and all of the predicted ORF products were used as queries to search for homologous entries in the GenBank and GenPeptide databases with the NCBI BLASTN, BLASTP, and BLASTX programs (Lipman and Pearson 1985). Significance of possible similarities was evaluated by use of the GCG Gap program with multiple randomizations of the query sequence; those homologies with 99% or higher confidence level were included in Table 2.

ACKNOWLEDGMENTS

We thank Dr. Samuel Kaplan for critical reading of the manuscript. This work was supported by NSF grants BIR9214821 to L.H. and MCB-9604443 to S.D.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Allardet-Servent, A., S. Michaux-Charachon, E. Jumas-Bilak, L. Karayan, and M. Ramuz. 1993. Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. *J. Bacteriol.* 175: 7869–7874.

Banfalvi, Z., E. Kondorosi, and A. Kondorosi. 1985. *Rhizobium meliloti* carries two megaplasmids. *Plasmid* 13: 129–138.

Blattner, F.R., G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453–1474.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* 273: 1058–1073.

Charlebois, R.L. and W.F. Doolittle. 1989. In *Mobile DNA* (ed. D.E. Berg and M.M. Howe), pp. 297–307. American Society for Microbiology, Washington, DC.

Cheng, H.P. and T.G. Lessie. 1994. Multiple replicons constituting the genome of *Pseudomonas cepacia* 17616. *J. Bacteriol.* 176: 4034–4042.

Choudhary, M., C. Mackenzie, K. Nereng, E. Sodergren, G.M. Weinstock, and S. Kaplan. 1997. Low-resolution sequencing of *Rhodobacter sphaeroides* 2.4.1T: Chromosome II is a true chromosome. *Microbiology* 143: 3085–3099.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry III et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537–544.

DasSarma, S. 1993. Identification and analysis of the gas vesicle gene cluster on an unstable plasmid of *Halobacterium halobium*. *Experientia* 49: 482–486.

DasSarma, S. and E.M. Fleischmann. 1995. *Archaea: A laboratory manual—Halophiles.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

DasSarma, S., P. Arora, F. Lin, E. Molinari, and L.R. Yin. 1994. Wild-type gas vesicle formation requires at least ten genes in the gvp gene cluster of *Halobacterium halobium* plasmid pNRC100. *J. Bacteriol.* 176: 7646–7652.

David, M., M.L. Daveran, J. Batut, A. Dedieu, O. Domergue, J. Ghai, C. Hertig, P. Boistard, and D. Kahn. 1988. Cascade regulation of *nif* gene expression in *Rhizobium meliloti. Cell* 54: 671–683.

Deckert, G., P.V. Warren, T. Gaasterland, W.G. Young, A.L. Lenox, D.E. Graham, R. Overbeek, M.A. Snead, M. Keller, M. Aujay et al. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus. Nature* 392: 353–358.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12: 387–395.

Drlica, K. and M. Riley. 1990. *The bacterial chromosome.* American Society for Microbiology, Washington, DC.

Dym, O., M. Mevarech, and J.L. Sussman. 1995. Structural

features that stabilize halophilic malate dehydrogenase from an archaebacterium. *Science* 267: 1344–1346.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 8: 186–194.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* 8: 175–185.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496–512.

Franz, B. and A.M. Chakrabarty. 1986. In *The bacteria* (ed. J.R. Sokatch and L.N. Ornston), pp. 295–323. Academic Press, New York.

Fraser, C.M., S. Casjens, W.M. Huang, G.G. Sutton, R. Clayton, R. Lathigra, O. White, K.A. Ketchum, R. Dodson, E.K. Hickey et al. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390: 580–586.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397–403.

Gavin, K.A., M. Hidaka, and B. Stillman. 1995. Conserved initiator proteins in eukaryotes. *Science* 270: 1667–1671.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* 8: 195–202.

Hackett, N.R., Y. Bobovnikova, and N. Heyrovska. 1994. Conservation of chromosomal arrangement among three strains of the genetically unstable archaeon *Halobacterium salinarium*. *J. Bacteriol.* 176: 7711–7718.

Halladay, J.T., W.L. Ng, and S. DasSarma. 1992. Genetic transformation of a halophilic archaebacterium with a gas vesicle gene cluster restores its ability to float. *Gene* 119: 131–136.

Halladay, J.T., J.G. Jones, F. Lin, A.B. MacDonald, and S. DasSarma. 1993. The rightward gas vesicle operon in *Halobacterium* plasmid pNRC100: Identification of the *gvpA* and *gvpC* gene products by use of antibody probes and genetic analysis of the region downstream of *gvpC*. *J. Bacteriol.* 175: 684–692.

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

Holloway, B.W. and K.B. Low. 1987. In Escherichia coli *and* Salmonella typhimurium *cellular and molecular biology* (ed. F.C. Neidhardt, J.L. Ingraham, K.B. Low, B. Magasanik, M. Schaechter, and H.E. Umbarger), pp. 1145–1153. American Society for Microbiology, Washington, DC.

Janulaitis, A., M. Petrusyte, Z. Maneliene, S. Klimasauskas, and V. Butkus. 1992. Purification and properties of the *Eco57*I restriction endonuclease and methylase-prototypes of a new class (type IV). *Nucleic Acids Res.* 20: 6043–6049.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto et al. 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 Res.* 3: 109–136.

Klenk, H.P., R.A. Clayton, J.F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390: 364–370.

Koonin, E.V., A.R. Mushegian, M.Y. Galperin, and D.R. Walker. 1997. Comparison of archaeal and bacterial computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* 25: 619–637.

Kunst, F., N. Ogasawara, I. Moszer, A.M. Albertini, G. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bolotin, S. Borchert et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390: 249–256.

Lanyi, J.K. 1974. Salt-dependent properties of proteins from extremely halophilic bacteria. *Bacteriol. Rev.* 38: 272–290.

Lipman, D.J. and W.R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science* 227: 1435–1441.

Lois, A.F., G.S. Ditta, and D.R. Helinski. 1993. The oxygen sensor FixL of *Rhizobium meliloti* is a membrane protein containing four possible transmembrane segments. *J. Bacteriol.* 175: 1103–1109.

Messing, J. 1983. New M13 vectors for cloning. *Methods Enzymol.* 101: 20–78.

Michaux, S., J. Paillisson, M.J. Carles-Nurit, G. Bourg, A. Allardet-Servent, and M. Ramuz. 1993. Presence of two independent chromosomes in the *Brucella melitensis* 16M genome. *J. Bacteriol.* 175: 701–705.

Miller, M.J. and R.B. Gennis. 1983. The purification and characterization of the cytochrome d terminal oxidase complex of the *Escherichia coli* aerobic respiratory chain. *J. Biol. Chem.* 258: 9159–9165.

Mouncey, N.J., M. Choudhary, and S. Kaplan. 1997. Characterization of genes encoding dimethyl sulfoxide reductase of *Rhodobacter sphaeroides* 2.4.1T: An essential metabolic gene function encoded on chromosome II. *J. Bacteriol.* 179: 7617–7624.

Neuhard, J. and P. Nygaard. 1987. In Escherichia coli *and* Salmonella typhimurium *cellular and molecular biology* (ed. F.C. Neidhardt, J.L. Ingraham, K.B. Low, B. Magasanik, M. Schaechter, and H.E. Umbarger), pp. 445–473. American Society for Microbiology, Washington, DC.

Ng, W.-L. and S. DasSarma. 1993. Minimal replication origin of the 200-kilobase *Halobacterium* plasmid pNRC100. *J. Bacteriol.* 175: 4584–4596.

Ng, W.-L., S. Kothakota, and S. DasSarma. 1991. Structure of the gas vesicle plasmid in *Halobacterium halobium*: *Inversion isomers, inverted repeats, and insertion sequences. J. Bacteriol.* 173: 1958–1964.

Ng, W.-L., C.-F. Yang, J.T. Halladay, P. Arora, and S. DasSarma. 1995. In *Archaea: A laboratory manual—Halophiles* (ed. S. DasSarma and E.M. Fleischmann), pp. 179–184. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Palmer, J.D. 1985. Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* 19: 325–354.

Pfeifer, F., G. Weidinger, and W. Goebel. 1981. Genetic variability in *Halobacterium halobium*. *J. Bacteriol.* 145: 375–381.

Proter, R.D. 1991. In *Modern microbial genetics* (ed. U.N. Streips and R.E. Yasbin), pp. 157–189. Wiley-Liss, New York.

Rosen, B.P., U. Weigel, C. Karkaria, and P. Gangola. 1988. Molecular characterization of a unique anion pump: The ArsA protein is an arsenite (antimonate)-stimulated ATPase. *Prog. Clin. Biol. Res.* 273: 105–112.

Rosen, B.P., U. Weigel, R.A. Monticello, and B.P. Edwards. 1991. Molecular analysis of an anion pump: Purification of the ArsC protein. *Arch. Biochem. Biophys.* 284: 381–385.

Sensen, C.W., H.P. Klenk, R.K. Singh, G. Allard, C.C. Chan, Q.Y. Liu, S.L. Penny, F. Young, M.E. Schenk, T. Gaasterland et al. 1996. Organizational characteristics and information content of an archaeal genome: 156 kb of sequence from *Sulfolobus solfataricus* P2. *Mol. Microbiol.* 22: 175–191.

Smith, D.R., L.A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional analysis and comparative genomics. *J. Bacteriol.* 179: 7135–7155.

Smith, L.M., J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, C. Heiner, S.B. Kent, and L.E. Hood. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* 321: 674–679.

Suwanto, A. and S. Kaplan. 1989. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. *J. Bacteriol.* 171: 5850–5859.

Thomm, M. 1996. Archaeal transcription factors and their role in transcription initiation. *FEMS Microbiol. Rev.* 18: 159–171.

Turmel, M., V. Cote, C. Otis, J.P. Mercier, M.W. Gray, K.M. Lonergan, and C. Lemieux. 1995. Evolutionary transfer of ORF-containing group I introns between different subcellular compartments (chloroplast and mitochondrion). *Mol. Biol. Evol.* 12: 533–545.

Van Larebeke, N., G. Engler, M. Holsters, S. Van den Elsacker, I. Zaenen, R.A. Schilperoort, and J. Schell. 1974. Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-inducing ability. *Nature* 252: 169–170.

Vreeland, R.H. and L.H. Hochstein. 1993. *Biology of halophilic bacteria.* CRC Press, Boca Raton, FL.

Wheeler, R.T. and L. Shapiro. 1997. Bacterial chromosome segregation: Is there a mitotic apparatus? *Cell* 88: 577–579.

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

Received June 24, 1998; accepted in revised form September 21, 1998.