J. Mol. Biol. (1991) 222, 509–524

Detailed Physical Map and Set of Overlapping Clones Covering the Genome of the Archaebacterium *Haloferax volcanii* DS2

Robert L. Charlebois†

Department of Biology
University of Ottawa, Ottawa, ON, Canada

Leonard C. Schalkwyk, Jason D. Hofman‡ and W. Ford Doolittle

Department of Biochemistry
Dalhousie University, Halifax, NS, Canada, B3H 4H7

(Received 3 May 1991; accepted 8 August 1991)

An integrated approach of "bottom up" and "top down" mapping has produced a minimal set of overlapping cosmid clones covering 96% of the 4140 kilobase-pairs (kbp) *Haloferax volcanii* DS2 genome and a completely closed physical map. This genome is partitioned into five replicons: a 2920 kbp chromosome and four plasmids, of 690 kbp (pHV4), 442 kbp (pHV3), 86 kbp(pHV1) and 6·4 kbp (pHV2). A restriction map for six infrequently-cutting restriction enzymes was constructed, representing a total of 903 sites in the cloned DNA. We have placed the two ribosomal RNA operons, the genes for 7 S RNA and for RNaseP RNA and 22 protein-coding genes on the map. Restriction site frequencies show significant variation in different portions of the genome. The regions of high site density correspond to halobacterial satellite or FII DNA which includes two small regions of the chromosome, the plasmids pHV1 and pHV2, and half of pHV4, but not pHV3.

Keywords: archaebacteria; *Haloferax*; mapping chromosome; megaplasmids

1. Introduction

Until recently, molecular biologists have looked at genomes one gene at a time. Except where nature has provided, or man has devised, effective methods for introduction, integration and phenotypic analysis of sizeable fragments of DNA (genetic approaches), we have only patchy information on overall genome organization and its possible functional or historical significance.

Purely physical approaches have changed this. With "top-down" methods made feasible by pulsed-field gel electrophoresis and "bottom-up" strategies made appealing by several model exercises (*Caenorhabditis*: Coulson *et al.*, 1986, 1988; yeast: Olson *et al.*, 1986; *Escherichia coli*: Kohara *et al.*, 1987; Birkenbihl & Vielmetter, 1989), we can look at whole genomes. Even in well-studied organisms, such as *E. coli*, with a well-defined genetic map, physical methods have provided additional detail, such as the ability to determine the orientation of many genes (Brewer, 1988, 1990), and a rapid way of mapping any sequence which can be used as a probe. For less well-characterized species, physical (mostly dop-down) methods have provided either confirmation or extension of classically-constructed genetic maps (*Pseudomonas aeruginosa*: Holloway *et al.*, 1990; *Caulobacter crescentus*: Ely & Gerardot, 1988; *Rhodobacter spheroides*: Suwanto & Kaplan, 1989; *Haemophilus influenzae*: Kauc *et al.*, 1989; Lee *et al.*, 1989; *Anabaena* PCC 7120: Bancroft *et al.*, 1989) or indeed the only available information on overall genome size and structure and gene location (*Bacillus cereus*: Kolstø *et al.*, 1990; *Clostridium perfringens*: Canard & Cole, 1989; *H. parainfluenzae*: Kauc & Goodgal, 1989; *Mycoplasma mycoides* and *M. mobile*: Pyle & Finch, 1988; *Mycoplasma pneumoniae*: Wenzel & Herrmann, 1988, 1989;

† Author to whom all correspondence should be addressed.

‡Present address: Department of Biology, Dalhousie University, Halifax, NS, Canada.

0022-2836/91/230509-16 $03.00/0

Campylobacter jejuni: Nuijten *et al.*, 1990; and the archaebacteria *Thermococcus celer*: Noll, 1989; and *Methanococcus voltae*: Sitzmann & Klein, 1991).

This growing prokaryotic database will help to define the new science of comparative genomics, with objectives of understanding: (1) global functional constraints on gene order and genome structure; (2) the significance of clusters of conserved order, such as those that are found for ribosomal protein genes between species such as *E. coli* and *Bacillus subtilis*, whose maps are otherwise thoroughly scrambled (Sankoff *et al.*, 1990); (3) the nature of mechanical restrictions on genomic rearrangements which should be frequent in the many bacterial genomes which boast repeated genes and transposable elements; and (4) the history of surviving rearrangements and plasmid or phage-derived accretions which are, at the molecular level, macroevolutionary events.

Here we present a physical map of the genome of the halophilic archaebacterium *Haloferax volcanii*. The archaebacteria (Archaea) are now thought to be more closely related to the eukaryotes than to the eubacteria (Woese *et al.*, 1990), so any information on archaebacterial genome structure has profound evolutionary as well as genetic implications. Halobacteria are the most experimentally tractable of the archaebacteria, and several interesting things are already known about their genomes. In most cases, halobacterial DNA can be resolved by density gradient centrifugation or malachite green bisacrylamide column chromatography into two fractions, a main band (Fraction I, or FI) of about 66 mol% (G+C), and a satellite (FII) fraction of about 56 mol% (G+C), representing 10 to 30% of the total mass (Joshi *et al.*, 1963; Moore & McCarthy, 1969; Pfeifer *et al.*, 1982; Tindall *et al.*, 1984; Ross & Grant, 1985). Plasmids found in halobacteria are usually part of the FII component. In *Halobacterium halobium*, the rest of this component has been shown to make up one or a few "islands" embedded within the chromosome, which is otherwise FI DNA (Pfeifer & Betlach, 1985).

Halobacterial insertion sequences may reside preferentially within FII DNA. Their transposition into genes in the FI fraction is nevertheless sufficiently frequent in *H. halobium* to result in very high rates of mutation (Pfeifer *et al.*, 1981*b*), and we showed earlier that this species carries insertion sequences of many different types, numbering in total several hundred (Sapienza & Doolittle, 1982). *H. volcanii*, on the other hand, may contain predominantly elements of one family, ISH51 (homologs of *H. halobium*'s ISH27; Pfeifer & Blaseio, 1990), numbering together fewer than 50 copies (Hofman *et al.*, 1986). This was one reason to choose *H. volcanii* for mapping. The others were that this species grows well on minimal media (Mevarech & Werczberger, 1985; Kauri *et al.*, 1990) and we had already selected it to develop other genetic tools, such as transformation methods (Charlebois *et al.*, 1987; Cline *et al.*, 1989) and shuttle vector technology (Lam & Doolittle, 1989).

Most physical maps made for bacteria have been top-down maps, because under favorable conditions a complete map can be produced with relatively little effort. Bottom-up maps require more persistence for completion but have a finer scale and have the advantage of providing as a by-product an ordered set of cosmid or phage clones, a valuable resource. As pointed out by Olson *et al.* (1986), the two approaches have complementary strengths: top-down mapping has superior continuity and bottom-up mapping has fine scale and indexed clones.

Bottom-up mapping requires detection of overlap between clones. Methods for detecting these overlaps can be divided into two classes whose characteristics differ in important ways. In methods of the fingerprint type, which includes the fingerprint method of Coulson *et al.* (1986), the fragment size matching method of Olson *et al.* (1986) and some hybridization-based methods (Herrero & Wolk, 1986; Birkenbihl & Vielmetter, 1989) the strength of the evidence for an overlap is proportional to its size. In landmark methods such as the present one, enough information is derived from small regions (single points, in practical terms) that any two clones containing the same landmark will be identified as overlapping, no matter how small the overlap. Overlaps can still be missed if they do not contain landmarks. This latter aspect of sensitivity can be increased by increasing the number (density) of landmarks assayed. Whereas fingerprint methods can only improve in sensitivity by improving the quality of the data, a landmark-based strategy can be adjusted at will. For example, in a first round of screening, a low density of landmarks may be assayed and then, in a second round, a less redundant subset of the clones can be assayed for more landmarks.

A new landmark method suitable for analysis of large genomes has recently been devised (Green & Olson, 1990). Although the method is designed to be used as part of a strategy also involving information from other mapping approaches, the method has the characteristics described above, and uses sequences of 200 to 500 base-pairs (bp†) which can be assayed by polymerase chain reaction as landmarks. These sequence tagged sites (STS) have been proposed as a "common language for physical mapping of the human genome" (Olson *et al.*, 1990). For applications such as the present one, our method (Charlebois *et al.*, 1989) has the advantage of great technical simplicity. We believe our map to be the most detailed prokaryotic genomic map yet assembled without the aid of genetic data.

2. Materials and Methods

(a) *Landmark analysis and chromosome walking*

The details and theory behind the landmark strategy (explained in Fig. 1) for determining when 2 clones

† Abbreviations used: bp, base-pair(s); STS, sequence tagged sites; kbp, 10^3 base-pairs; CHEF, contour-clamped homogeneous electric field electrophoresis.

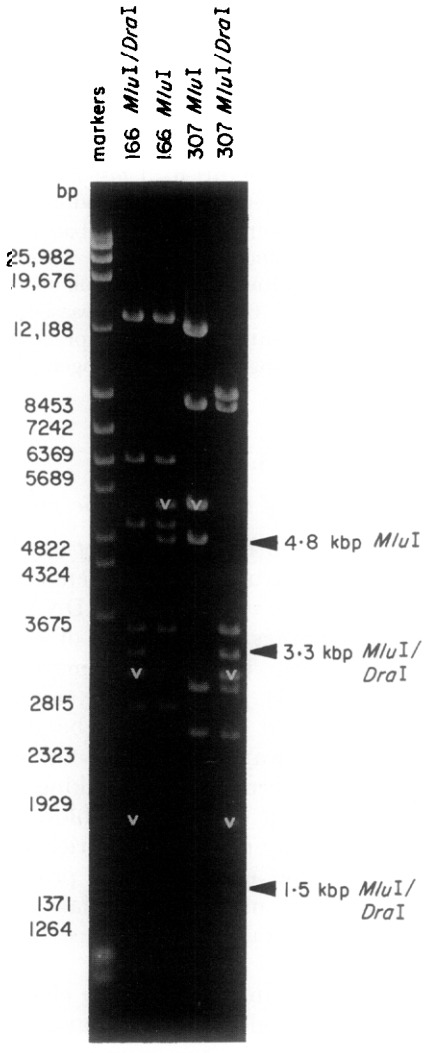

Figure 1. Example of the identification of overlapping clones by landmark analysis. For this figure, 2 clones (166 and 307) were chosen which share 1 landmark. V indicates vector fragments that are common to all clones. The 4·8 kbp *Mlu*I fragment in each clone is cut once by *Dra*I to give 2 *Mlu*I/*Dra*I fragments, of 3·3 kbp and 1·5 kbp. With fewer than 80 sites for *Dra*I in the genome, the presence of an *Mlu*I fragment of the same size cut into the same 2 subfragments by *Dra*I is quite distinctive and virtually proves overlap: it is a landmark. Cosmid 307, shown here, actually contains a 2nd *Dra*I landmark: the 12·2 kbp *Mlu*I fragment is cut by *Dra*I to give *Mlu*I/*Dra*I subfragments of 8·6 and 3·6 kbp. The examination of additional double digests, using *Mlu*I along with various other infrequently cutting enzymes, increases the density of landmarks and allows potentially shorter overlaps to be detected. We used 10 of these double digests: *Mlu*I with *Bam*HI, *Bgl*II, *Cla*I, *Dra*I, *Eco*RI, *Hin*dIII, *Nhe*I, *Pst*I, *Ssp*I and *Xba*I (see Charlebois *et al.*, 1989). Sites for *Eco*RI totalled several hundred in number but still generated useful landmarks.

overlap were described in the preliminary report of this project (Charlebois *et al.*, 1989). In short, a partial-*Mlu*I library of *H. volcanii* DS2 genomic DNA was constructed in the cosmid vector Lorist M. *Mlu*I was chosen as the cloning enzyme because it cuts *H. volcanii* DNA randomly with a frequency of about 1 site in 3·2 kilobase-pairs (kbp). Cosmid DNA was prepared by alkaline extraction, and subjected to restriction digest analysis on agarose gels. Single digests with *Mlu*I , as well as 10 double digests, with *Mlu*I plus *Bam*HI, *Bgl*II, *Cla*I, *Dra*I, *Eco*RI, *Hin*dIII, *Nhe*I, *Pst*I, *Ssp*I or *Xba*I, were run on 0·9% agarose gels and their fragments were sized. Analysis of the data included: (1) subtraction of vector fragments; (2) identification of which *Mlu*I fragments had been cut by which of the 10 infrequently cutting enzymes; and (3) assignment of double-digest *Mlu*I/2nd enzyme subfragments to *Mlu*I single-digest fragments.

In the 1st round of landmark analysis, 576 randomly chosen cosmid clones (reduced to 319 by elimination of duplicates) were analyzed, and could be sorted into 59 overlapping sets (contigs). These clones were named 1 to 576 (Charlebois *et al.*, 1989). A 2nd library was prepared (These cosmids have designations beginning with the letters A through H) by size fractionation of an *Mlu*I partial digest by CHEF (contour-clamped homogeneous electric field electrophoresis, Chu *et al.*, 1986), and packaging using an *Eco*K-free packaging extract (Stratagene), in an attempt to produce a more representative library. We transferred colonies from the *Mlu*I-partial cosmid library onto Colony/PlaqueScreen (DuPont). These were hybridized with probes prepared from the ends of cosmid inserts using SP6 or T7 RNA polymerase and the promoters contained within the Lorist vector. Many of the cosmids identified in this series of walking experiments were unrelated to the probes, but were nonetheless useful when analyzed. Part of the 2nd library was amplified on plates and stored at −70°C in 15% glycerol.

A 2nd round of landmark analysis was carried out on 384 more cosmids that were picked at random from the amplified library. These cosmid clones, designated 1A1 to 4H12, were digested with *Mlu*I, *Mlu*I/*Bam*HI and *Mlu*I/*Pst*I only. In most cases, the digests did not need sizing by the computer; i.e. the landmarks were quite distinctive even with a sizing error of 5% which is typical for estimations done by eye. Most of the new clones could be placed in their correct positions by comparison of their *Bam*HI and *Pst*I landmarks with the lists of *Bam*HI and *Pst*I landmarks tabulated for the previous clones. This very rapidly established linkage of the new clones to the old ones. Those clones which extended contigs, or which did not have informative *Bam*HI or *Pst*I landmarks, were analyzed further with additional landmarking digests using *Bgl*II, *Dra*I, *Eco*RI, *Hin*dIII, *Ssp*I and *Xba*I in combination with *Mlu*I.

The 2nd round of chromosome walking was performed on DNA dot blots, using pooled cosmid DNA samples. A total of 768 colonies picked from an amplified *Mlu*I-partial cosmid library were grown individually to saturation in Terrific Broth (Tartof & Hobbs, 1987) in Titertek tubes (Biorad), for a total of eight 8 × 12 boxes. The boxes were numbered 5 to 12, to correspond to clones 5A1 to 12H12. Samples (100 μl) of the cultures were then pooled in groups of 8 in such a way that each clone appeared in 2 pools. Preparation of these pools was simple using a multichannel pipettor.

Cosmid DNA was prepared from the pooled cultures in the Titertek boxes themselves. Since only crude DNA was required for dot blots, an abbreviated alkaline extraction procedure was performed. The Titertek boxes were spun in a Megafuge (Baxter) tabletop centrifuge using home-made swing-out aluminium holders, at 550 ***g***. Culture supernatants were aspirated away, the pellets were resuspended in TEG (20 mM-Tris·HCl (pH 7·6), 50 mM-EDTA, 1% glucose), treated with 0·2 M-NaOH, 1% (w/v) SDS, neutralized with 7·5 M-ammonium acetate, mixed well by

vortexing each tube individually, and centrifuged at 550 *g*. The supernatants were removed into a new box of tubes, and precipitated with isopropanol. A refinement and elaboration of this method will be published elsewhere. Pellets were dissolved, denatured with 0·4 M-NaOH, and spotted onto a number of pieces of GeneScreen nylon membrane. Hybridizations of the pooled DNA were performed using SP6 or T7-directed RNA probes synthesized from the ends of contigs. Positive signals in each of the 2 sets of pools indicated which clones were candidate extenders. These were pursued by full landmark analysis.

(b) *Restriction mapping of cosmids*

Cosmid clones included in the latter stages of the evolution of the minimal set of clones were mapped for *Bam*HI, *Bgl*II, *Dra*I, *Hin*dIII, *Pst*I and *Ssp*I sites. *Mlu*I fragment sizes were tabulated only and were not mapped. Single digests were run out on both 0·8% and 2% agarose gels; double-digests were run only on 0·8% gels. Markers used for the 0·8% gels were a combination of 4 digests: (1) λ–*Bst*EII; (2) λ–*Ppu*MI; (3) λ–*Sna*BI; and (4) pUC18, pUC19 or pIBI31 cut with both *Dra*I and *Hin*dIII. For the 2% agarose gels, markers were a mix of pUC or pIBI cut with both *Dra*I and *Hin*dIII and pUC19 cut with *Msp*I. Maps of sites for the 6 enzymes were deduced from the single and double digest fragment size lists using an algorithm similar to that described by Bellon (1988).

Some clones had too many sites for this approach to be practical, or had maps with unordered fragments. These were mapped by partial digestion, using the indirect end-labelling approach (Smith & Birnstiel, 1976). Enzymes such as *Dra*I which cut the vector, but whose sites are infrequent in the inserts, were used to linearize the clones. These were then partially digested with the enzyme of interest and the fragments were separated by conventional electrophoresis on 0·8% agarose gels. The same method was used in some cases to order *Mlu*I fragments for determining the degree of overlap between clones.

(c) *General methods*

Electrophoresis, transfer and hybridization methods were carried out as described by (Charlebois *et al.*, 1989). DNA–DNA hybridizations were done in a solution of 0·5 M-sodium phosphate (pH 7·2, 7% (w/v) SDS/l, at 65 to 68 °C for probes from *H. volcanii*, and 42 or 45 °C for probes from other species. Washing was in 1% (w/v) SDS, 50 mM-(in sodium) sodium phosphate (pH 7·2) at the same temperature as the hybridization, and for the low-stringency hybridizations increasing in 3 to 5 °C steps until the signal disappeared.

SP6 or T7 transcripts for use as probes were synthesized in a reaction (10 μl) containing about 100 ng of alkaline-extracted cosmid DNA (never treated with RNase), 20 μCi of [α-^{32}P]GTP at 3000 Ci/mmol (Amersham), 20 units of SP6 or T7 RNA polymerase (DuPont/NEN), 10 units RNAsin (Promega), 40 mM-Tris·HCl (pH 7·6, 6 mM-$MgCl_2$, 2 mM-spermidine, 0·1 mg acetylated BSA/ml, 0·4 mM each of ATP, CTP and UTP. The product was used without purification for hybridization in a solution of 0·5 M-(in sodium) sodium phosphate (pH 7·2), 7% (w/v) SDS, 0·1 mg each of salmon sperm DNA and yeast tRNA/ml, after an overnight prehybridization. Washing was in 1% (w/v) SDS, 50 mM-(in sodium) sodium phosphate (pH 7·2) at 70 °C.

Size estimation of fragments on photographs or Southern transfer autoradiograms of CHEF gels was by comparison with multimers and digests of bacteriophage lambda DNA. Migration distances were measured with the aid of a digitizing tablet and a polynomial curve was fit to the region of interest.

Malachite green bisacrylamide chromatography was as described by Bünemann & Müller (1978) using a 0·5 cm × 8 cm column. The solvent was 10 mM-(in phosphate) sodium phosphate (pH 6), 1 mM-EDTA (PE) and elution was with a gradient of 0 to 1 M-sodium perchlorate in PE. Fractions were surveyed by absorbance at 260 nm, by scintillation counting of the entire fraction, or by loading of samples of alternate fractions on an agarose gel, as necessary. Selected fractions were passed over 2 ml Sephadex columns in water before being precipitated with ethanol containing 20 mg of mussel glycogen (Boehringer) as carrier.

3. Results

(a) *Overview of the course of the project*

In a previous report (Charlebois *et al.*, 1989), we described the construction of a genomic cosmid clone library of *H. volcanii* DS2 DNA and the sorting of this library into 59 sets of overlapping clones, or contigs, by the landmark strategy (explained in Fig. 1). At that time, an estimated 500 kbp of sequence remained to be cloned and efforts were under way to complete the cloning of the genome by chromosome walking methods. Chromosome walking using colony blots gave us a set of candidates only moderately enriched in clones extending from the ends used as probes. When analyzed, however, many of the falsely positive clones proved to contain new sequence, and so were nonetheless useful.

The abundance of informative new clones showed us that we had ceased doing landmark analysis on randomly picked clones too soon, and that such analysis was still at this stage faster and easier than continued chromosome walking. The walking effort had nevertheless reduced the number of contigs to 25 and extended the amount of DNA cloned by 300 to 400 kbp. At this stage we began to determine the linkage of the contigs by pulsed-field gel analysis, while simultaneously analyzing more clones. Of the 25 contigs (50 ends), plus eight additional ends created by the discard of problem clones, 27 pairs were linked by the pulsed-field linking approach (see below). Seven of these pairs of ends actually overlapped, but had not been detected by landmark analysis.

Meanwhile, we returned to the library and analyzed 384 more cosmid clones for their landmarks. This joined seven pairs of contigs, and extended five additional ends. A second round of chromosome walking, on fewer start points, was now more productive. False-positives were also reduced in frequency by the use of DNA dot blots (in 2 sets of 96 pools each of 8 randomly picked clones), rather than colony blots. Where needed, unique probes from repeat-rich end clones were isolated from a gel. Walking joined six pairs of contigs and extended one end.

In the course of the whole project, over 4000 clones were screened, and over 2000 of these were

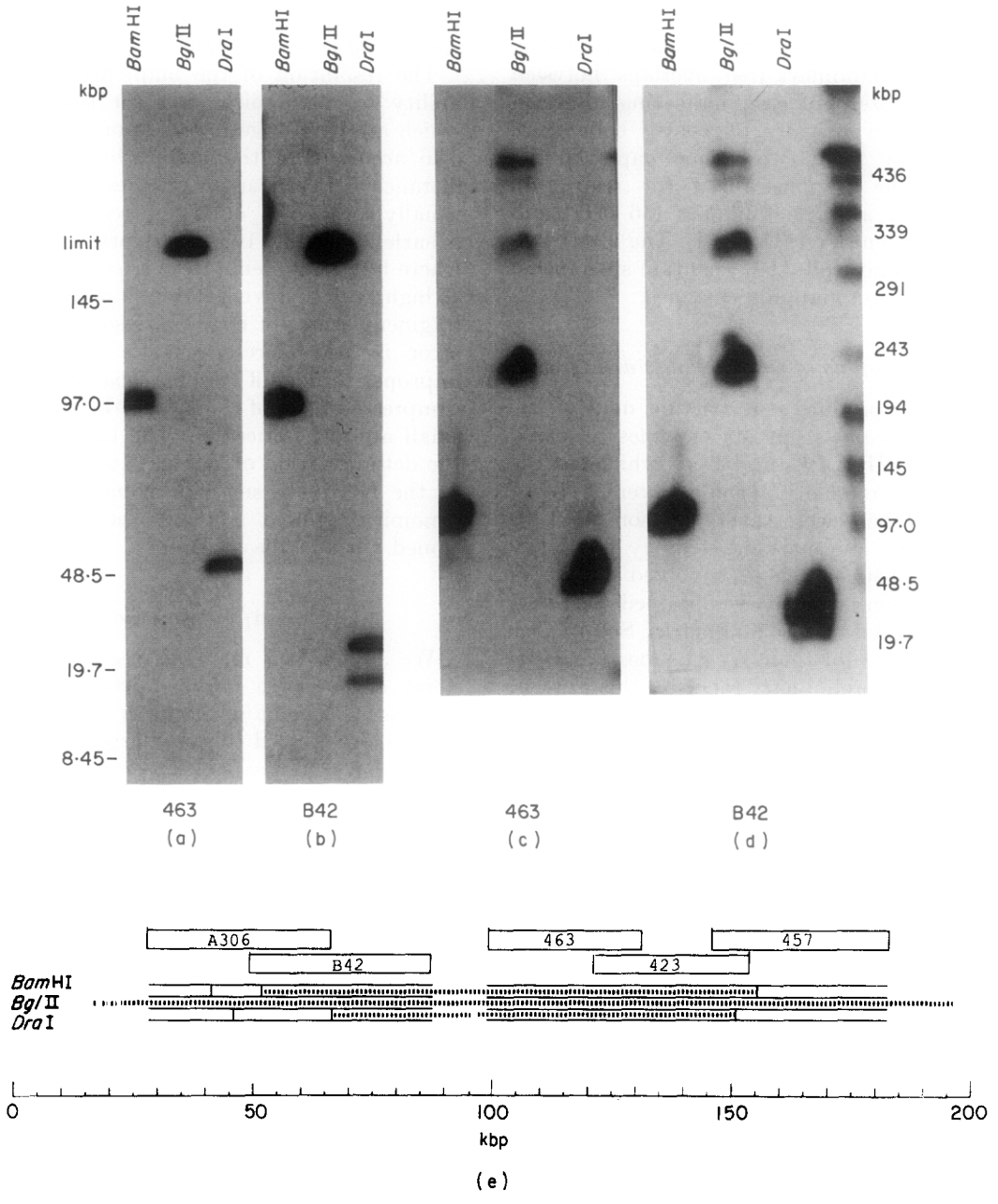

Figure 2. An example of link-up of contigs by pulsed-field gel electrophoresis. Two Southern transfers of CHEF gels of digests of *H. volcanii* DNA were each probed sequentially with cosmids B42 and 463. (a) and (b) The gel was run at a switching time of 5 s for 18 h and then 10 s for 12 h. (c) and (d) The gel was run at 20 s for 24 h in order to separate the largest fragments in these digests. (e) Map of the contig ends indicating the observed terminal fragments (dotted lines). The smaller of the 2 *Dra*I fragments that hybridized to cosmid B42 is the internal fragment. The 3 highest molecular weight bands in the *Bgl*II digest in (c) and (d) are partial digestion products. This gap was subsequently closed using cosmid 1F3.

subjected to landmark analysis. This number exceeded what should have been needed in an unbiased, statistical sense. Nevertheless, 4·3% of the genome (approx 178 kbp) remains uncloned, in seven gaps which might represent either unclonable DNA or the tail-end of a biased probability distribution.

(b) *Link-up of contigs by pulsed-field gel electrophoresis*

In order to complete the map, we turned to the top-down approach. Blots of digests of genomic DNA cut with infrequently cutting restriction enzymes were probed with the ends of each contig (for an example, see Fig. 2). When two contig ends produce identical patterns with several enzymes, it is strong evidence (because of the few, well-spaced fragments involved) that they are close together in the genome. Additional data from the cosmid restriction maps provides an estimate of the size of the gap between the ends since, when there are no sites within gaps, the probed fragment size should be equal to the sum of distances to the first sites within the end cosmids plus the distance between ends. Independent gap size estimates from different

enzymes should agree. In some cases, the gap was estimated either to be negative in size, indicating the presence of landmark-free overlaps between clones, or to be zero in size, indicating abutting clones. Twenty-seven pairs of contig ends were linked in this way, 24 by two or more gap-spanning fragments of the enzymes used for restriction mapping, one by a single fragment (56–5G7) and two by other enzymes (*Afl*II, *Asc*I). The latter two gaps (H3–D339 and G86–11A7) contain sites for all six of our restriction mapping enzymes.

(c) *Restriction mapping of cosmid clones and of gaps*

We elected to produce a restriction map of the genome of *H. volcanii* using six enzymes (*Bam*HI, *Bgl*II, *Dra*I, *Hin*dIII, *Pst*I and *Ssp*I) which cut the genome rather infrequently, each on average about every 30 kbp. There were several reasons for this choice. First, the landmarking strategy for determining overlap between clones involved the use of these enzymes; consequently we obtained data that could relate mapped sites to landmarks. Second, the infrequency of sites for these six enzymes was itself of interest. In the preliminary report of this project, we described the occurrence of "oases", areas of the genome strikingly richer in the landmark enzyme sites than the bulk of the genome. These infrequent enzymes were useful markers for studying the compositional sectoring of the genome. Third, a map of sites for these six enzymes would be highly useful for linking up contigs and for verification, providing enough fragments to be informative but not so many as to be confusing. Finally, the sites for these enzymes together could yield a map with reasonable detail (Fig. 3).

Restriction mapping of cosmid clones was done primarily by analyzing single and double digests. Sites which could not be ordered by this method, either because of multiple possible maps or because of too great a complexity in a particular clone, were ordered by partial digestion of cosmids linearized (usually only) at restriction sites within the vector. Of the 903 sites mapped, 16 sites remain with alternative possible placements; eight in the chromosome, six in pHV4, and two in pHV3.

Although the overlap between cosmid clones was shown by their sharing one or more landmarks or in some cases by hybridization, the size of this overlap had to be determined separately. This overlap was usually determined by alignment of the clone maps. In cases where the limited set of the six enzyme maps was insufficient, local maps of *Cla*I, *Eco*RI, *Nhe*I or *Xba*I were generated, or more usually, partial digest maps of *Mlu*I sites were compared.

The mapping of sites within uncloned DNA posed a special challenge. We obtained partial map information (whether or not 1 or more sites was present) in these regions by hybridization of probes from the ends of the contigs to genomic Southern blots of selected digests on pulsed-field gels. The presence of repeated sequences at the end of several of the contigs, and the richness of sites in two of the gaps, prevented complete mapping of sites in the gaps using this approach.

The resolution of the map, here defined as the ability to distinguish sites for the same enzyme which are close together, is approximately 0·2 kbp. The accuracy of the map, which is the error in distance between any two sites in the genome, is usually within 1%, our sizing accuracy of fragments (Charlebois *et al.*, 1989). In very site-poor regions, where large fragments were sized, the error may be as high as 5%; in very site-rich regions, where some fragment sizes are near the resolution of the map, error is also increased as is the chance of an improper order of closely spaced sites. At the compressed scale of the map in Figure 3, errors this small are not noticeable. The largest errors are in the determination of gap sizes which are arrived at as the relatively small difference between a large genomic fragment size and the large sizes of the cloned parts of the fragment.

(d) *Problem clones*

We found two main classes of problem clone. Some clones were rather unstable in *E. coli* and underwent specific deletions spontaneously, especially when we tried to grow large cultures for DNA extraction. Repeated attempts, and smaller cultures, provided us with enough intact cosmid DNA for mapping. Cosmid 280, a particularly unstable clone, had to be rescued by *in vitro* packaging (Yokobata *et al.*, 1991). In the second category were those rare clones that had landmarks from two different areas of the genome. These cloning artifacts were identified either through their being abnormal relative to the redundant clones, or by their anomalous hybridization patterns to genomic Southern blots. We also observed some site polymorphism in several cases, where a site was either lost or gained in a particular clone. Finally, small insertions were observed in a few clones.

(e) *Verification*

We believe that the sharing of landmark sites between two clones is very strong evidence of their overlap. The low cutting frequency of the enzymes chosen for landmark analysis makes the existence of twin landmarks unlikely in a genome this size (Charlebois *et al.*, 1989), and clones usually overlapped by more than one landmark. Of real concern was the possible existence of clones with multiple inserts, of clones which somehow became altered in the *E. coli* host and of clones which represent genomic rearrangements within the original *H. volcanii* culture. We needed a verification that our overlapping set of cosmid clones represents the true map of the genome of *H. volcanii*, or at least the consensus "snapshot" map if the genome is prone to rapid change. This was particularly important given that there exists no genetic map information for this organism, and given that halobacterial genomes contain numerous repeated sequences.

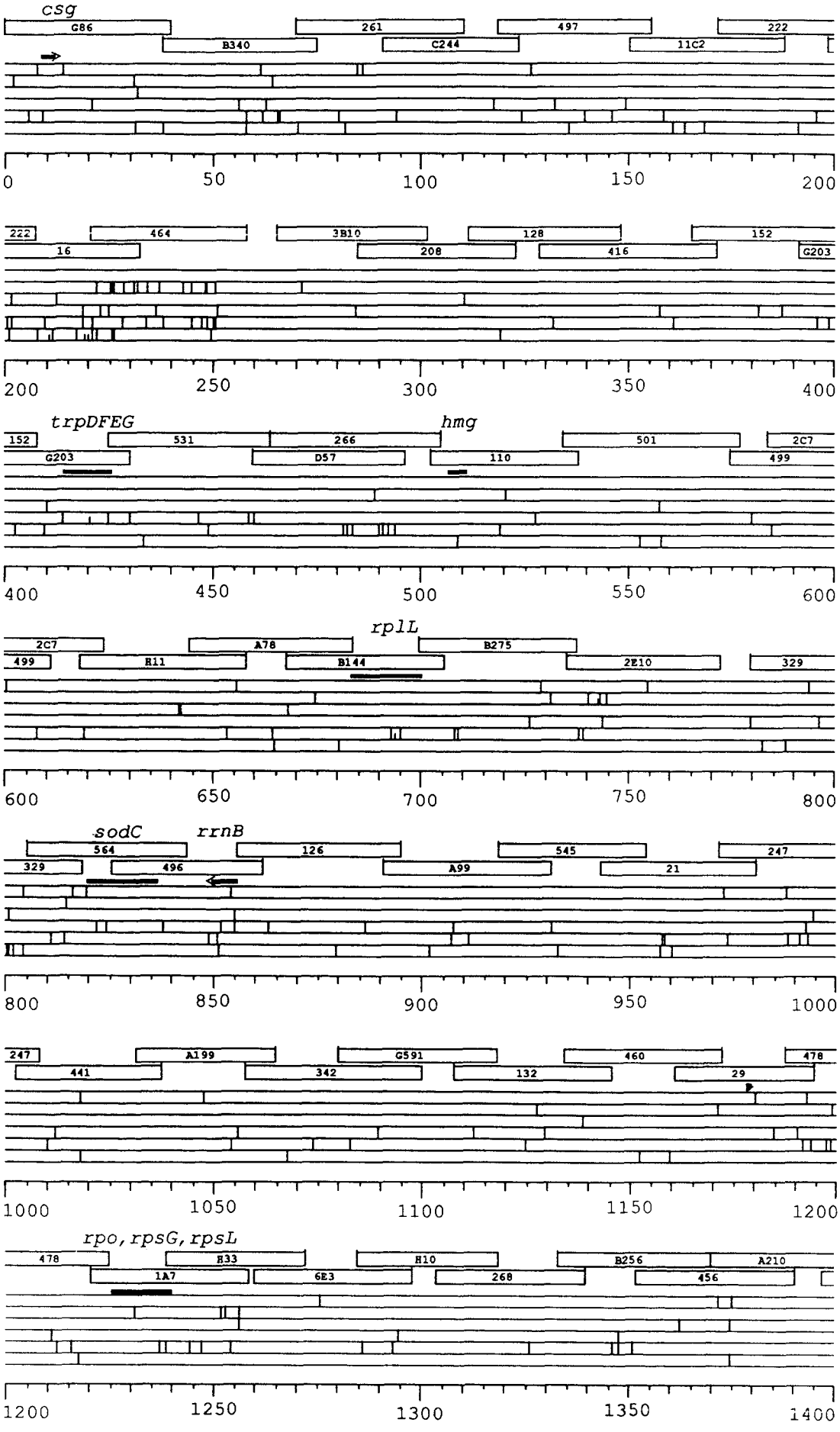
csg
G86
261
497
222
B340
C244
11C2
0
50
100
150
200
222
464
3B10
128
152
16
208
416
G203
200
250
300
350
400
trpDFEG
hmg
152
531
266
501
2C7
G203
D57
110
499
400
450
500
550
600
rplL
2C7
A78
B275
499
H11
B144
2E10
329
600
650
700
750
800
sodC
rrnB
564
126
545
247
329
496
A99
21
800
850
900
950
1000
247
A199
G591
460
478
441
342
132
29
1000
1050
1100
1150
1200
rpo, rpsG, rpsL
478
H33
H10
B256
A210
1A7
6E3
268
456
1200
1250
1300
1350
1400

Fig. 3.

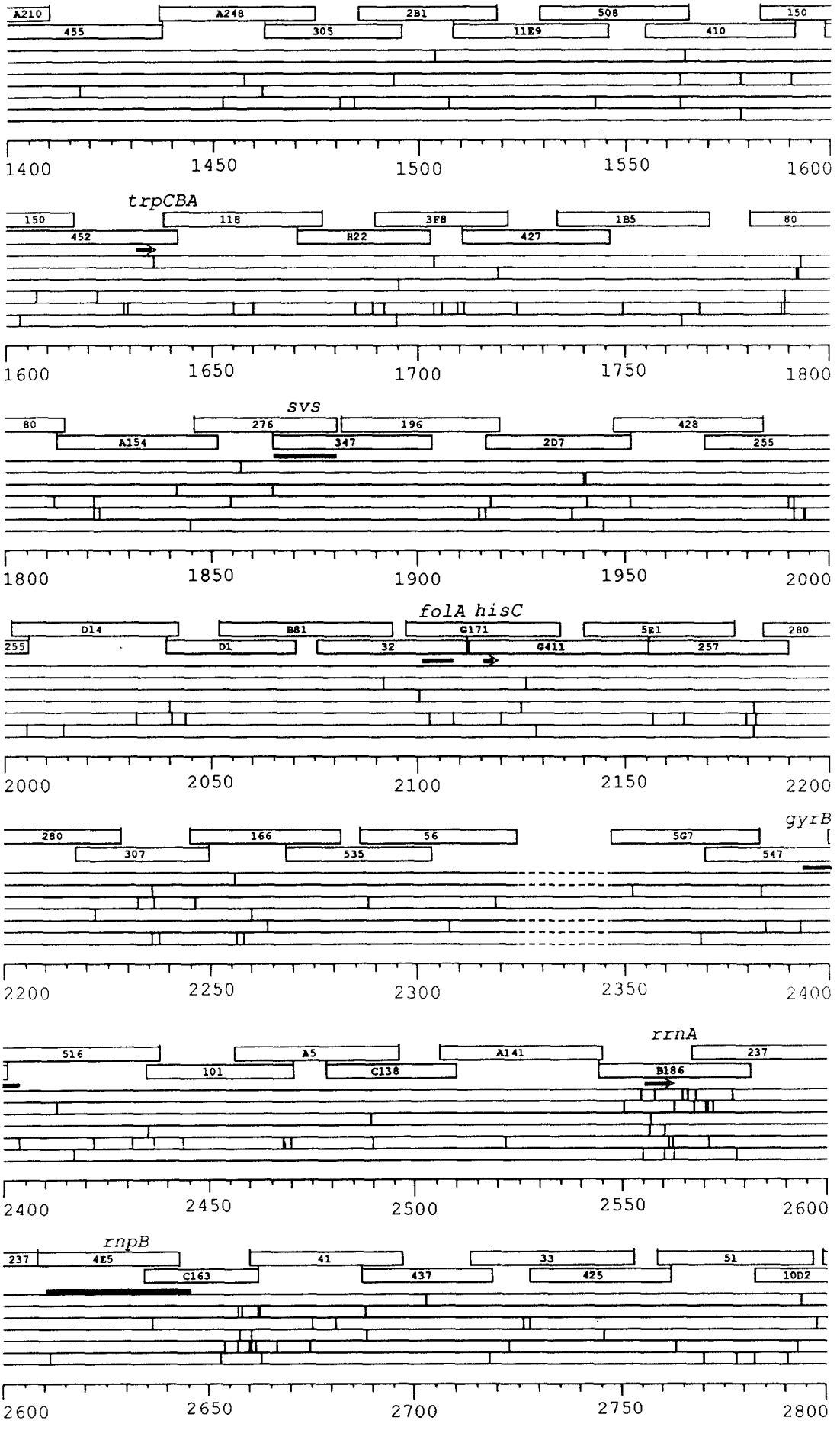
A210
A248
2B1
508
150
455
305
11E9
410
1400
1450
1500
1550
1600
trpCBA
150
118
3F8
1B5
80
452
H22
427
1600
1650
1700
1750
1800
svs
80
276
196
428
A154
347
2D7
255
1800
1850
1900
1950
2000
folA hisC
D14
B81
G171
5E1
280
255
D1
32
G411
257
2000
2050
2100
2150
2200
gyrB
280
166
56
5G7
307
535
547
2200
2250
2300
2350
2400
rrnA
516
A5
A141
237
101
C138
B186
2400
2450
2500
2550
2600
rnpB
237
4E5
41
33
51
C163
437
425
10D2
2600
2650
2700
2750
2800

Fig. 3.

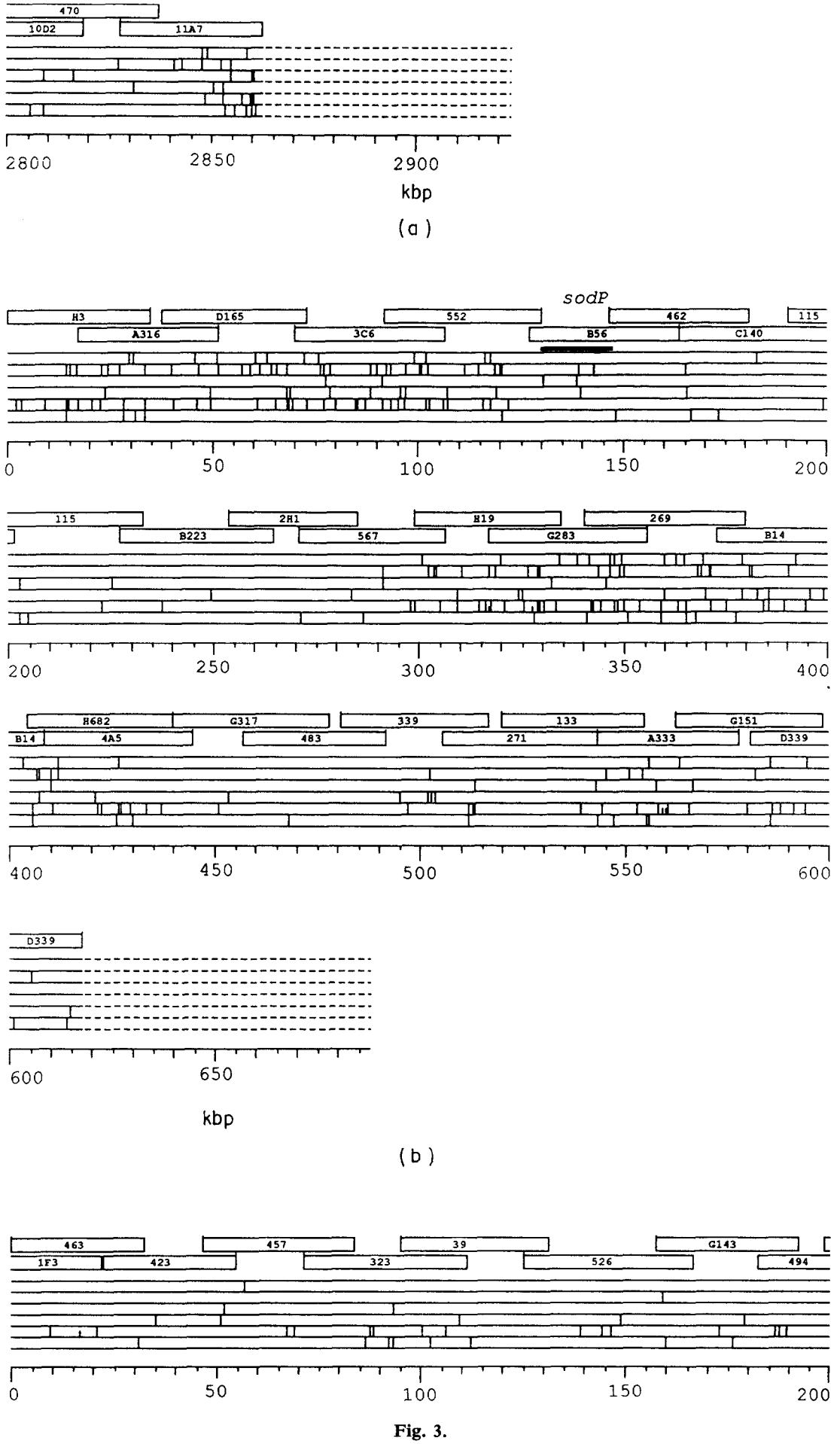
470
10D2
11A7
2800
2850
2900
kbp
(a)
sodP
H3
D165
552
462
115
A316
3C6
B56
C140
0
50
100
150
200
115
2H1
H19
269
B223
567
G283
B14
200
250
300
350
400
H682
G317
339
133
G151
B14
4A5
483
271
A333
D339
400
450
500
550
600
D339
600
650
kbp
(b)
463
457
39
G143
1F3
423
323
526
494
0
50
100
150
200

Fig. 3.

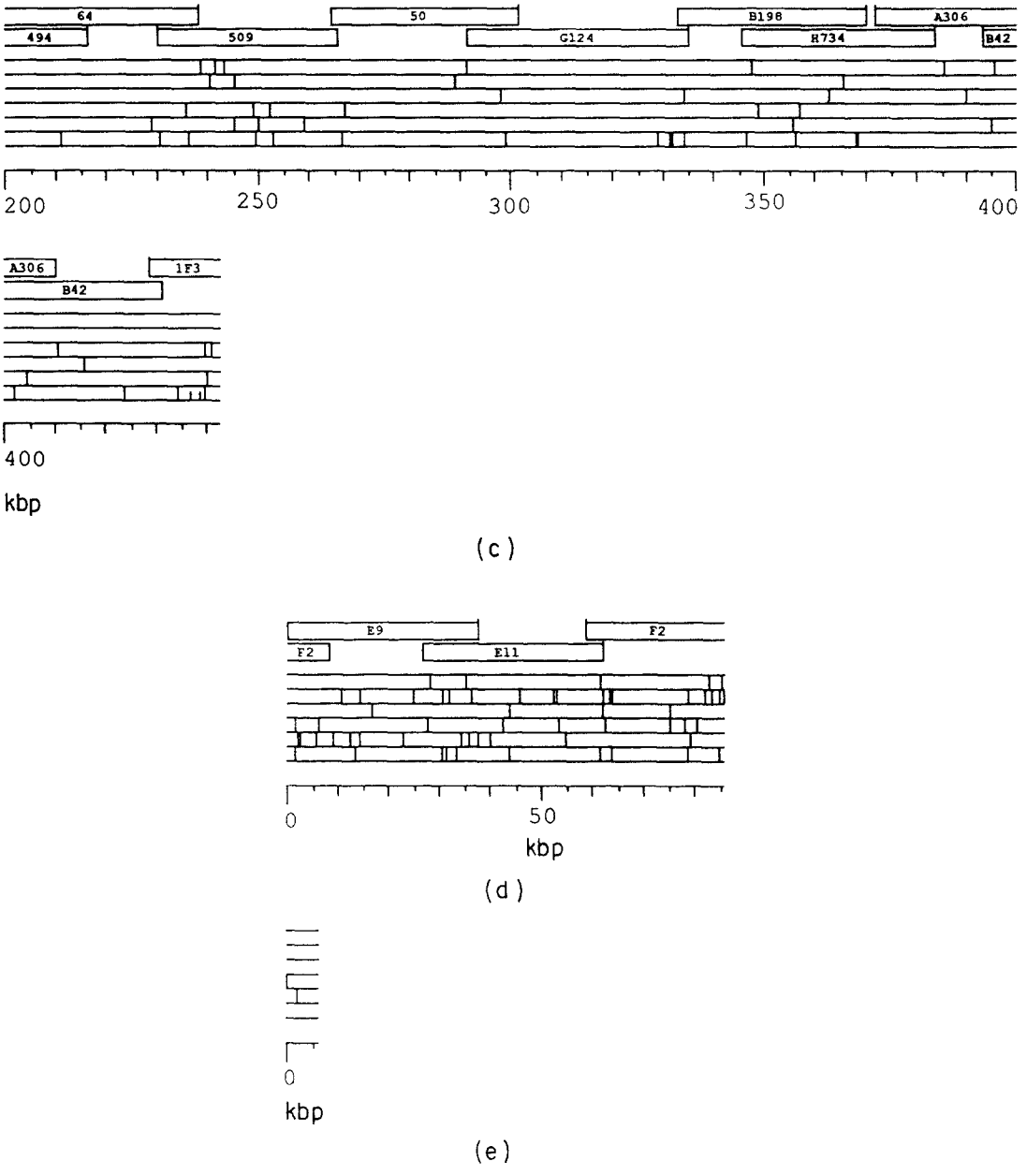

Figure 3. Physical map of the *Haloferax volcanii* DS2 genome. Within each staff are shown the positions of, from top to bottom, sites for *Bam*HI, *Bgl*II, *Dra*I, *Hin*dIII, *Pst*I and *Ssp*I. Some sites may represent multiple, closely spaced sites. Unordered fragments are indicated by half-height tick marks. Open boxes above each staff represent the overlapping cosmid clones, filled boxes indicate genetic loci; below each staff is a scale bar with numbers in kilobase-pairs. The raised edge on one side of the box representing each clone indicates the end of the insert nearest to the *Hin*dIII site of the Lorist M vector. (a) The circular chromosome is 2920 kbp in size, 96% of which is cloned. The map is incomplete within the 6 small uncloned regions; (b) pHV4, a 690 kbp plasmid, 90% cloned; (c) pHV3, a 442 kbp plasmid; (d) pHV1, an 86 kbp plasmid; and (e) pHV2, a 6354 bp plasmid, cloned in pUC. Refer to Table 3 for an explanation of the genetic loci shown on the map.

Verification of the map was based on finding, by hybridization to genomic Southern blots, restriction fragments which were predicted by the map. A set of 50 overlapping fragments averaging 120 kbp in length was checked. Small deletions, insertions, or other rearrangements could have been missed by this method, but the long-range continuity of the map was verified and any gross rearrangements were ruled out. Another source of reassurance was that most regions of the genome were cloned more than once.

Pfeifer *et al.* (1981*a*) reported the existence of two plasmids in *H. volcanii* DS2; pHV1 (for which we obtained a size of 86 kbp), and pHV2 (6·4 kbp), both of which we had cloned earlier. Gutiérrez *et al.* (1986) later reported two megaplasmids, one of 180 kbp and the other of 410 kbp. In our study, we obtained an overlapping set of clones forming a

Table 1

Restriction site frequency statistics for the cloned portions of the H. volcanii *genome*

Enzyme	No. of sites	Median (kbp)	Maximum (kbp)	Minimum (kbp)	Mean (kbp) ±S.D.	Obs./exp. means; FI portions	Obs./exp. means; FII portions	χ^2,5 partitions p =	χ^2,10 partitions p =
A. *Chromosome: 2815 kbp cloned (96% of the total); 97% FI DNA, 3% FII DNA*									
*Bam*HI	42	33·4	474	1·3	67·0±97·4	25·4	5·1	0·824	n.d.
*Bgl*II	63	6·0	463	0·2	44·7±81·2	5·7	0·79	0·9979	$0{\cdot}9^{6}41$
*Dra*I	44	43·4	236	0·2	64·0±57·2	1·5	2·7	0·817	n.d.
*Hin*dIII	76	23·4	167	1·4	37·0±37·0	3·4	2·4	0·958	0·823
*Pst*I	161	7·9	102	0·2	17·5±20·9	6·6	1·3	0·484	0·540
*Ssp*I	83	15·7	203	0·2	33·9±40·5	0·90	0·67	0·9966	$0{\cdot}9^{4}88$
Totals	469	2·9	50	<0·2	6·0±7·9	6·2	1·5	$0{\cdot}9^{4}40$	$0{\cdot}9^{9}76$
B. *pHV4: 618 kbp cloned (90% of the total); 60% FI DNA, 40% FII DNA*									
*Bam*HI	35	9·1	129	0·3	17·6±29·3	25·6	2·4	0·9966	n.d.
*Bgl*II	67	3·1	126	0·3	9·2±19·4	3·2	0·83	$0{\cdot}9^{7}64$	$0{\cdot}9^{8}26$
*Dra*I	14	34·4	103	8·0	44·1±35·8	0·91	6·2	n.d.	n.d.
*Hin*dIII	32	10·0	125	0·4	19·3±27·0	3·6	2·1	0·896	n.d.
*Pst*I	103	3·3	77	0·2	6·0±10·9	4·1	1·0	$0{\cdot}9^{5}50$	$0{\cdot}9^{9}44$
*Ssp*I	31	13·6	87	0·7	19·9±19·4	0·48	2·2	0·217	n.d.
Totals	282	0·9	27	<0·2	2·2±3·7	4·2	1·5	$0{\cdot}9^{13}62$	$0{\cdot}9^{25}0$

Enzyme	No. of sites	Median (kbp)	Maximum (kbp)	Minimum (kbp)	Mean (kbp) ±S.D.	Obs./exp. means; FI	χ^2, 2 partitions p =	χ^2, 3 partitions p =
C. *pHV3: size = 442 kbp; FI DNA*								
*Bam*HI	8	43·2	182	1·7	55·3±57·4	19·1	n.d.	n.d.
*Bgl*II	5	76·6	236	4·8	88·5±78·7	7·7	n.d.	n.d.
*Dra*I	9	29·3	205	1·2	49·2±56·6	1·1	n.d.	n.d.
*Hin*dIII	12	34·6	82	3·4	36·9±25·0	3·2	0·436	n.d.
*Pst*I	25	9·1	97	0·2	17·7±21·2	6·1	0·928	0·913
*Ssp*I	31	10·0	56	0·2	14·3±14·7	0·32	0·894	0·813
Totals	90	24	30	<0·2	4·9±6·0	4·5	0·908	0·352
D. *pHV1: size = 85·9 kbp; FII DNA*								
*Bam*HI	5	21·2	29·2	2·2	17·2±10·8	5·1	n.d.	n.d.
*Bgl*II	18	3·6	14·6	0·2	4·8±4·4	0·92	0·654	0·689
*Dra*I	4	22·6	27·4	13·4	21·5±6·0	2·7	n.d.	n.d.
*Hin*dIII	9	9·0	21·3	2·0	9·5±5·9	1·8	n.d.	n.d.
*Pst*I	14	3·1	24·2	0·3	6·1±6·6	1·8	0·9925	n.d.
*Ssp*I	10	8·3	17·3	0·7	8·6±6·2	1·1	0·0	n.d.
Totals	60	0·8	11·4	<0·2	1·4±1·9	1·8	0·204	0·295

Enzyme	No. of sites	Maximum (kbp)	Minimum (kbp)	Mean (kbp)	Obs./exp. means; FI portions	Obs./exp. means; FII portions	FI/FII ratio of means; absolute	FI/FII ratio of means; (G+C)-corrected
E. *Total genome: 3967 kbp cloned, (96% of the total); 89% FI; 11% FII including pHV2 (6·4 kbp, FII DNA) which has 1* Hin*dIII and 1* Pst*I site*								
*Bam*HI	90	474	0·3	44·1	24·4	3·1	6·7	7·9
*Bgl*II	153	463	0·2	25·9	5·4	0·85	14·0	6·4
*Dra*I	71	236	0·2	55·9	1·3	4·1	1·9	0·32
*Hin*dIII	130	167	0·4	30·5	3·4	2·0	3·7	1·7
*Pst*I	304	103	0·2	13·1	6·2	1·2	4·5	5·2
*Ssp*I	155	203	0·2	25·6	0·68	1·3	2·9	0·52
Totals	903	50	<0·2	4·4	5·7	1·6	4·9	3·6

Data for the 4 largest replicons, as well as for the total, are presented. Mean fragment sizes expected are based on FI = 66·5 mol% (G+C); FII = 55·3 mol% (G+C) (Tindall *et al.*, 1984), total = 64·9 mol% (G+C) (Gutiérrez *et al.*, 1989, 1990). The FI and FII regions of the genome are as defined in Fig. 4 and in the text. The χ^2 statistic tests the fit to an exponential distribution of fragment sizes. Two interval sizes were chosen to probe different levels of biased site distribution. In some cases the sample size was too small to permit the test (indicated by "n.d."). Values of p close to unity (e.g. $p = 0{\cdot}9^{4}40 = 0{\cdot}999940$) indicate the presence of clustering of sites; values close to zero indicate unexpected regularity in the site distribution (Churchill *et al.*, 1990).

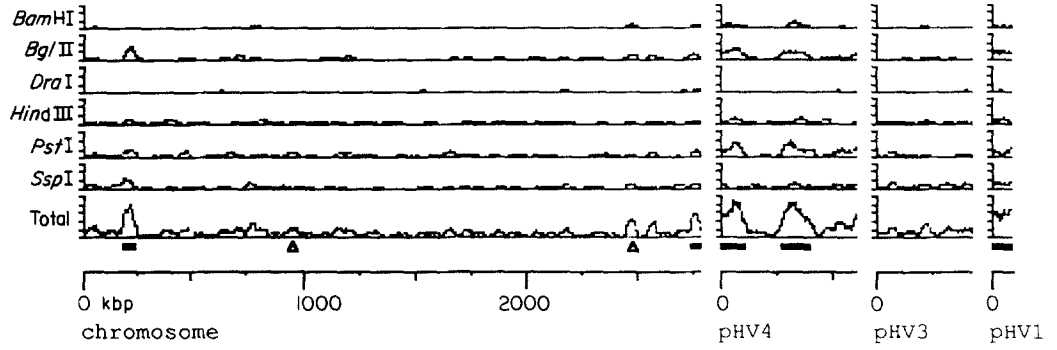

Figure 4. Restriction site frequency plot. The number of sites in a 40 kbp window was calculated at 1 kbp step intervals for each of the 6 enzymes separately and for the total of the 6 enzymes. Oases, regions of the genome particularly rich in sites (≥20 sites/40 kbp with a maximum density of at least 25 sites/40 kbp) are indicated by bars, and correspond to FII DNA. △ indicates the 2 rRNA operons.

442 kbp circle (pHV3), which also appears on pulsed-field gels (not shown), but we failed to find a 180 kbp plasmid. We conclude either that their 180 kbp plasmid was actually the 86 kbp pHV1, or that the strains are not identical in this respect.

(f) *Analysis of the physical map*

Table 1 is a presentation of summary statistics for the *H. volcanii* genomic restriction map. The most striking result is the non-uniform site density distribution which we had previously observed as "oases" of sites within "desert" (Charlebois *et al.*, 1989). The χ^2 test shows a marked clustering of sites for *Bgl*II and *Ssp*I in the chromosome, and for *Bam*HI, *Bgl*II and *Pst*I in the largest plasmid, pHV4. Moreover, the clusters of sites for individual enzymes are not independent of one another; the sites for different enzymes cluster together. These "oases" are evident in Figures 3 and 4, and are supported by the highly significant p-values presented in Table 1 for the total of the six enzymes in the χ^2 test of the chromosome and of pHV4. The plasmids pHV3 and pHV1 are much more homogeneous and show no overall clustering of sites. An additional test dividing pHV3 into ten and pHV1 into five intervals also did not show any clustering bias for the six enzymes in total ($p = 0{\cdot}732$ and $p = 0{\cdot}116$, respectively).

The regions of high site density (oases) correspond to the lower-mol% (G+C) fraction of the genome, the FII DNA. The proportion of the genome which is oasis is similar to the estimated proportion which is FII DNA. We arbitrarily defined an oasis as a region of DNA which has at least 20 sites per 40 kbp for the total of the six enzymes mapped, with a maximum site density of at least 25 sites per 40 kbp. The plasmid pHV1, known to be FII DNA by malachite green bisacrylamide chromatography, is entirely oasis DNA. Sample clones from desert and oasis regions of other replicons run through the (A+T) affinity column also confirm this correlation between oasis and FII DNAs. The chromosome is primarily (97%) FI DNA, with two small FII regions of 50 and 36 kbp, starting at positions 203 and 2841 kbp, respectively, on the map (Figs 3 and 4); pHV4 is 60% FI and 40% FII DNA and its two large FII regions of 119 and 128 kbp start at map positions 11 and 300 kbp, respectively; pHV3 is entirely FI DNA and pHV1 is FII DNA. The plasmid pHV2, its 6354 bp genome completely sequenced (Charlebois *et al.*, 1987), is FII DNA (56·0 mol% (G+C)).

Several small clusters of sites remain (see Figs 3 and 4) which were not elevated to the status of oasis. Two of these represent the two rRNA operons, which are rich in these sites. Others are difficult to judge in a statistical sense, e.g. the cluster in clone C163.

Mean fragment sizes, as shown in Table 1, vary greatly between FI and FII DNA but, except for *Ssp*I, there is no strong variation within FI DNA or within FII DNA among the different replicons. The difference between FI and FII DNA is more than just a difference in mol% (G+C), however. In Table 1(E), mean fragment sizes in FI DNA and FII DNA are compared. Sites are elevated in frequency for all six enzymes in FII relative to FI DNA. When expected mean fragment sizes based on mol% (G+C) are taken into account, sites for *Bam*HI, *Bgl*II, *Hin*dIII and *Pst*I are still relatively over-represented in FII DNA, but sites for *Dra*I and for *Ssp*I are under-represented. The recognition sequences for both *Dra*I and *Ssp*I consist of six (A·T) base-pairs, which is perhaps a factor. FII DNA must have a different overall structure from that of FI DNA, at the level of oligonucleotide composition, to account for all of these differences.

(g) *Gaps*

Six regions of the chromosome and one region of pHV4, totalling 178 kbp, remain uncloned (Table 2). The effects of these uncloned regions on conclusions derived from the restriction map should be minimal. For purposes of analysis of cloned DNA, gaps were set to zero length. The mean and median fragment sizes, the site rarity calculations and the test for clustering of sites should be good

Table 2
Size estimates for the seven uncloned regions of the genome

Gap	Position (kbp) (Fig. 3)	Flanking cosmids	Gap size estimate (kbp)	Size of contig (kbp)
1	258	464-3B10	7	258·2
2	772	2E10-329	7	507·2
3	1770	1B5-80	10	990·9
4	2324	56-5G7	23	543·5
5	2608	237-4E5	0	261·5
6	2862	11A7-G86	61	254·2
*	pHV4 618	H3-D339	70	617·5

Gap sizes were estimated by subtracting from the sizes of gap-spanning restriction fragments the distances to the 1st sites within the 2 flanking contigs.

estimates, given that the sample size represents 96% of the complete data set. Most affected by the gaps are the site counts and the estimates of maximum fragment size. In Table 1, the sizes of the largest fragments in the chromosome for *Bam*HI and *Hin*dIII were adjusted to include gap 1 and gap 3, respectively. The size of the largest *Hin*dIII fragment in pHV4 was estimated by hybridization of a pulsed-field gel blot, since one end of this fragment is in the uncloned portion of the plasmid. Site counts represent sites in cloned DNA only, hence they are underestimates concerning the complete genome.

(h) *Markers*

Few *H. volcanii* protein-coding genes have been characterized to date, but we have begun to populate the map with markers by hybridizing dot blots of cosmid DNAs and genomic Southern blots with cloned genes from *H. volcanii* and other (not closely related) halobacteria. Most genes from *Halobacterium* spp. can be hybridized to *H. volcanii* clones to give an unambiguous signal. We expect more distant cross-hybridization to work only in exceptional cases (such as *nif* (Sibold *et al.*, 1985) and gas vacuole (Horne *et al.*, 1988) genes) and, even in these cases, sequencing may be required in order to be certain of the identity of the gene. Markers located by hybridization are shown in Table 3 and on Figure 3, along with *his* and *trp* genes, which were cloned by complementation of auxotrophs with cosmid DNAs. The positions of these genes (Conover & Doolittle, 1990; Lam *et al.*, 1990), along with that of the RNaseP RNA (Nieuwlandt *et al.*, 1991) were reported using the names of cosmids which are not part of our minimal set.

Table 3
Identity, location and source of genetic markers

Locus	Cosmid	Position	Description	Probe source	Reference
csg	G86	8–11	Cell surface glycoprotein	*Hf.v.*†	Sumper *et al.*, 1990
folA	32/G171	2102–2109	Dihydrofolate reductase	*Hf.v.*	Zusman *et al.*, 1989
gyrB	547	2394–2404	DNA gyrase, B subunit	*Hf.*sp.	Holmes & Dyall-Smith, 1991
hisC	G171/G411	2116–2117	Histidinol-phosphate aminotransferase	*Hf.v.*	Conover & Doolittle, 1990
hmg	110	507–511	3-Hydroxy-3-methylglutaryl coenzyme A reductase	*Hf.v.*	Lam & Doolittle, 1989
rnpB	4E5 (and or) C163	2611–2645	RNase P RNA	*Hf.v.*	Nieuwlandt *et al.*, 1991
rplL	B144	684–700	Ribosomal protein L12	*Hb.c.*	Shimmin *et al.*, 1989
rpo	1A7	1225–1239	RNA polymerase subunits B″, B′, A, C; plus *rpsG*, *rpsL* and the evolutionarily conserved ORF75	*Hb.h.*	Leffers *et al.*, 1989
rrfA	B186	2562	5 S rRNA	*Hf.v.*	Charlebois *et al.*, 1989
rrfB	496	850	5 S rRNA	*Hf.v.*	Charlebois *et al.*, 1989
rrlA	B186	2559–2562	23 S rRNA	*Hf.v.*	Charlebois *et al.*, 1989
rrlB	496	850–853	23 S rRNA	*Hf.v.*	Charlebois *et al.*, 1989
rrsA	B186	2557–2559	16 S rRNA	*Hf.v.*	Charlebois *et al.*, 1989
rrsB	496	854–855	16 S rRNA	*Hf.v.*	Charlebois *et al.*, 1989
sodC	564	819–826	Superoxide dismutase	*Hb.c.*	May & Dennis, 1989; P. Joshi, pers. comm.
sodP	B56	pHV4 130–147	Superoxide dismutase	*Hb.c*	May & Dennis, 1989; P. Joshi, pers. comm.
svs	276/347	1865–1880	7 S RNA (presumed homologue of *E. coli ffs*)	*Hf.v.*	This work
trpA	452	1635–1636	Tryptophan synthase, A subunit	*Hf.v.*	Lam *et al.*, 1990
trpB	452	1633–1635	Tryptophan synthase, B subunit	*Hf.v.*	Lam *et al.*, 1990
trpC	452	1633	Indole 3-glycerol-phosphate synthetase	*Hf.v.*	Lam *et al.*, 1990
trpD	G203	414–425	Phosphoribosyl anthranilate transferase	*Hf.v.*	Lam *et al.*, 1990; pers. comm.
trpE	G203	414–425	Anthranilate synthase, α subunit	*Hf.v.*	Lam *et al.*, 1990; pers. comm.
trpF	G203	414–425	*N*-(5-phosphoribosyl) anthranylate isomerase	*Hf.v.*	Lam *et al.*, 1990; pers. comm.
trpG	G203	414–425	Anthranilate synthase, β subunit	*Hf.v.*	Lam *et al.*, 1990; pers. comm.

† Abbreviations: *Hf.v.*, *Haloferax volcanii*; *Hf.*sp., *Haloferax* sp.; *Hb.c.*, *Halobacterium cutirubrum*; *Hb.h.*, *Halobacterium halobium*; pers. comm., personal communication.

4. Discussion

We report here a genome mapping effort for the archaebacterium *Haloferax volcanii* DS2 in which we generated a set of minimally overlapping clones covering nearly all (96%) of the genome, and a physical map with 903 restriction sites, linking all of these clones. We have located, in addition to the two ribosomal RNA operons and the genes for RNaseP RNA and 7 S RNA, 22 protein-coding genes, thus beginning the development of a rich physical–genetic map of this genome. (It is premature to attempt comparison with other prokaryotic maps, since there is extensive scrambling even between *E. coli* and *B. subtilis*.) Work in preparation includes the complementation of hundreds of auxotrophs with cosmid DNA (A. Cohen, W. L. Lam, S. Cline & W. F. Doolittle, unpublished results) and the location by hybridization of all the tRNA genes and of all the copies of several insertion sequence families (L. C. Schalkwyk, R. L. Charlebois & W. F. Doolittle, unpublished results).

The *H. volcanii* genome is not physically uniform in structure. The 11% of the genome which we estimate to be of the lower (G+C) content, FII type, is present in the form of an 86 kbp plasmid, a six-copy 6·4 kbp plasmid, two regions within the large plasmid pHV4 of 119 kbp and 128 kbp and two regions, near each other within the chromosome, of 50 and 36 kbp. There is a striking discontinuity in site frequency at the boundaries of the FI and FII regions in the chromosome, and it is one of our hypotheses that these regions represent plasmids (or bacteriophage genomes) which have been inserted. The plasmid pHV4 also contains two distinct FII regions flanked by FI DNA, and one could propose a number of possible scenarios responsible for its creation, involving replicon fusion and resolution, perhaps mediated by insertion element activity. The origin of pHV3, another megaplasmid present in this strain, is as perplexing: its DNA composition, as judged by restriction enzyme site frequencies, is comparable to that of FI chromosomal DNA. This plasmid could represent a class of halobacterial plasmids altogether different from FII-DNA plasmids, or it may simply be derived from a replicable portion of chromosomal FI DNA by an excision event.

An extension of the hypothesis of large integrating plasmids is that the plasmids might also be carriers and preferred targets of halobacterial insertion elements, and that they might carry these elements from one strain to another horizontally, by some form of mating or by other means. A mating system has been described for *H. volcanii* (Rosenshine *et al.*, 1989), although (small) plasmids did not seem to be involved. Newly isolated *H. volcanii* strains used in that study contained plasmids unlike each other or anything else known. The distribution of insertion elements among taxa also seems not to make much phylogenetic sense, and all of this evidence supports horizontal transfer. This idea is in agreement with thoughts on eubacterial systems (Craig & Kleckner, 1987).

A competing hypothesis for the existence of FII DNA, based on the markedly increased frequency of insertion sequences in FII DNA (Pfeifer, 1986; Schalkwyk *et al.*, unpublished results), is that these regions might represent a gradual buildup of insertion sequence upon insertion sequence, essentially producing a junkyard of moribund transposable elements. The dynamics of such a system would be rather interesting.

One of the uses of a physical–genetic map is for the comparison of genome structure between different species. Efforts are currently under way (R.L.C., unpublished results) to map the genome of *Halobacterium* sp. GRB, a genetically stable relative of *H. halobium* (Soppa & Oesterhelt, 1989). Comparison of maps may go a long way towards measuring the genomic stability, or instability, of the bulk of the DNA, the FI DNA, in halobacteria. FII DNA has already been shown to be rather unstable because of the activity of insertion elements (Pfeifer, 1986). We wish to understand the evolutionary events, and the forces behind them, that have shaped the dynamic genome of the extremely halophilic archaebacteria.

We thank Antoine Morin for providing the program to calculate the *p*-values in the chi-square test, and R. Keith Conover for the program to size fragments of pulsed-field gels. R.L.C. is funded by NSERC; W.F.D. is funded by MRC and is a fellow of the Canadian Institute for Advanced Research. L.C.S. was a recipient of an MRC studentship and an I.W. Killam Memorial Scholarship.

References

Bancroft, I., Wolk, C. P. & Oren, E. V. (1989). Physical and genetic maps of the genome of the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* **171**, 5940–5948.

Bellon, B. (1988). Construction of restriction maps. *CABIOS*, **4**(1), 111–115.

Birkenbihl, R. P. & Vielmetter, W. (1989). Cosmid derived map of *E. coli* strain BHB2600 in comparison to the map of strain W3110. *Nucl. Acids Res.* **17**, 5057–5069.

Brewer, B. (1988). When polymerases collide: replication and the transcriptional organisation of the *Escherichia coli* chromosome. *Cell*, **53**, 679–686.

Brewer, B. J. (1990). Replication and the transcriptional organisation of the *Escherichia coli* chromosome. In *The Bacterial Chromosome* (Drlica, K. & Riley, M., eds), pp. 61–83, American Society for Microbiology, Washington, DC.

Bünemann, H. & Müller, W. (1978). Base specific fractionation of double-stranded DNA: affinity chromatography on a novel type of adsorbant. *Nucl. Acids Res.* **5**, 1059–1074.

Canard, B. & Cole, S. T. (1989). Genome organisation of the anaerobic pathogen *Clostridium perfringens*. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 6676–6680.

Charlebois, R. L., Lam, W. L., Cline, S. W. & Doolittle, W. F. (1987). Characterization of pHV2 from *Halobacterium volcanii* and its use in demonstrating transformation of an archaebacterium. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 8530–8534.

Charlebois, R. L., Hofman, J. D., Schalkwyk, L. C., Lam,

W. L. & Doolittle, W. F. (1989). Genome mapping in halobacteria. *Can. J. Microbiol.* **35**, 21–29.

Chu, G., Vollrath, D. & Davis, R. W. (1986). Separation of large DNA molecules by countour clamped homogeneous electric fields. *Science,* **234**, 1582–1585.

Churchill, G. A., Daniels, D. L. & Waterman, M. S. (1990). The distribution of restriction enzyme sites in *Escherichia coli*. *Nucl. Acids Res.* **18**, 589–597.

Cline, S. W., Lam, W. L., Charlebois, R. L., Schalkwyk, L. C. & Doolittle, W. F. (1989). Transformation methods for halophilic archaebacteria. *Can. J. Microbiol.* **35**, 148–152.

Conover, R. K. & Doolittle, W. F. (1990). Characterization of a gene involved in histidine biosynthesis in *Halobacterium (Haloferax) volcanii*: isolation and rapid mapping by transformation of an auxotroph with cosmid DNA. *J. Bacteriol.* **172**, 3244–3249.

Coulson, A., Sulston, J., Brenner, S. & Karn, J. (1986). Toward a physical map of the nematode *Caenorhabditis elegans*. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 7821–7825.

Coulson, A., Waterston, R., Kiff, J., Sulston, J. & Kohara, Y. (1988). Genome linking with yeast artificial chromosomes. *Nature (London),* **335**, 184–186.

Craig, N. L. & Kleckner, N. (1987). Transposition and site-specific recombination. In Escherichia coli *and* Salmonella typhimurium (Neidhardt, F. C., ed.), pp. 1054–1070. American Society for Microbiology, Washington, DC.

Ely, B. & Gerardot, C. J. (1988). Use of pulsed-field-gradient gel electrophoresis to construct a physical map of the *Caulobacter crescentus* genome, *Gene,* **68**, 323–333.

Green, E. D. & Olson, M. V. (1990). Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping. *Science,* **250**, 94–98.

Gutiérrez, M. C., García, M. T., Ventosa, A., Nieto, J. J. & Ruiz-Berraquero, F. (1986). Occurrence of megaplasmids in halobacteria. *J. Appl. Bacteriol.* **61**, 67–71.

Gutiérrez, M. C., Ventosa, A. & Ruiz-Berraquero, F. (1989). DNA–DNA homology studies among strains of *Haloferax* and other halobacteria. *Curr. Microbiol.* **18**, 253–256.

Gutiérrez, M. C., Ventosa, A. & Ruiz-Berraquero, F. (1990). Deoxyribonucleic acid relatedness among species of *Haloarcula* and other halobacteria. *Biochem. Cell Biol.* **68**, 106–110.

Herrero, A. & Wolk, C. P. (1986). Genetic mapping of the chromosome of the cyanobacterium, *Anabaena variabilis*. Proximity of the structural genes for nitrogenase and ribulose-bisphosphate carboxylase. *J. Biol. Chem.* **261**, 7748–7754.

Hofman, J. D., Schalkwyk, L. C. & Doolittle, W. F. (1986). ISH51: a large, degenerate family of insertion sequence-like elements in the genome of the archaebacterium, *Halobacterium volcanii*. *Nucl. Acids Res.* **14**, 6983–7000.

Holloway, B. W., Dharmsthiti, S., Krishnapillai, V., Morgan, A., Obeyesekere, V., Ratnaningsib, E., Sinclair, M., Strom, D. & Zhang, C. (1990). Patterns of gene linkages in *Pseudomonas* species. In *The Bacterial Chromosome* (Drlica, K. & Riley, M., eds), pp. 147–162. American Society for Microbiology, Washington, DC.

Holmes, M. L. & Dyall-Smith, M. L. (1991). Mutations in DNA gyrase result in novobiocin resistance in halophilic archaebacteria. *J. Bacteriol.* **173**, 642–648.

Horne, M., Englert, C. & Pfeifer, F. (1988). Two genes encoding gas vacuole proteins in *Halobacterium halobium*. *Mol. Gen. Genet.* **213**, 459–464.

Joshi, J. G., Guil, W. R. & Handler, P. (1963). The presence of two species of DNA in some halobacteria. *J. Mol. Biol.* **6**, 34–38.

Kauc, L. & Goodgal, S. H. (1989). The size and a physical map of the chromosome of *Haemophilus parainfluenzae*. *Gene,* **83**, 377–380.

Kauc, L., Mitchell, M. & Goodgal, S. H. (1989). Size and physical map of the chromosome of *Haemophilus influenzae*. *J. Bacteriol.* **171**, 2474–2479.

Kauri, T., Wallace, R. & Kushner, D. J. (1990). Nutrition of the halophilic archaebacterium, *Haloferax volcanii*. *System. Appl. Microbiol.* **13**, 14–18.

Kohara, Y., Akiyama, K. & Isono, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell,* **50**, 495–508.

Kolstø, A.-B., Grønstad, A. & Oppegaard, H. (1990). Physical map of the *Bacillus cereus* genome. *J. Bacteriol.* **172**, 3821–3825.

Lam, W. L. & Doolittle, W. F. (1989). Shuttle vectors for the archaebacterium *Halobacterium volcanii*. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 5478–5482.

Lam, W. L., Cohen, A., Tsouluhas, D. & Doolittle, W. F. (1990). Genes for tryptophan biosynthesis in the archaebacterium *Haloferax volcanii*. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 6614–6618.

Lee, J. J., Smith, H. O. & Redfield, R. J. (1989). Organisation of the *Haemophilus influenzae* Rd genome. *J. Bacteriol.* **171**, 3016–3024.

Leffers, H., Gropp, F., Lottspeich, F., Zillig, W. & Garrett, R. A. (1989). Sequence, organization, transcription and evolution of RNA polymerase subunit genes from the archaebacterial extreme halophiles *Halobacterium halobium* and *Halococcus morrhuae*. *J. Mol. Biol.* **206**, 1–17.

May, B. P. & Dennis, P. P. (1989). Evolution and regulation of the gene encoding superoxide dismutase from the archaebacterium *Halobacterium cutirubrum*. *J. Biol. Chem.* **264**, 12253–12258.

Mevarech, M. & Werczberger, R. (1985). Genetic transfer in *Halobacterium volcanii*. *J. Bacteriol.* **162**, 461–462.

Moore, R. L. & McCarthy, B. J. (1969). Characterisation of the deoxyribonucleic acid of various strains of halophilic bacteria. *J. Bacteriol.* **99**, 248–254.

Nieuwlandt, D. T., Haas, E. S. & Daniels, C. J. (1991). The RNA component of RNase P from the archaebacterium *Haloferax volcanii*. *J. Biol. Chem.* **266**, 5689–5695.

Noll, K. M. (1989). Chromosome map of the thermophilic archaebacterium *Thermococcus celer*. *J. Bacteriol.* **171**, 6720–6725.

Nuijten, P. J., Bartels, C., Bleumink-Pluym, N. M., Gaastra, W. & van der Zeijst, B. A. (1990). Size and physical map of the *Campylobacter jejuni* chromosome. *Nucl. Acids Res.* **18**, 6211–6214.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. & Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 7826–7830.

Olson, M., Hood, L., Cantor, C. R. & Botstein, D. (1990). A common language for physical mapping of the human genome. *Science,* **245**, 1434–1435.

Pfeifer, F. (1986). Insertion elements and genome organization of *Halobacterium halobium*. *System. Appl. Microbiol.* **7**, 36–40.

Pfeifer, F. & Betlach, M. (1985). Genome organization in *Halobacterium halobium*: a 70 kb island of more (AT)-rich DNA in the chromosome. *Mol. Gen. Genet.* **198**, 449–455.

Pfeifer, F. & Blaseio, U. (1990). Transposition burst of the ISH27 insertion element family in *Halobacterium halobium*. *Nucl. Acids Res.* **18**, 6921–6925.

Pfeifer, F., Weidinger, G. & Goebel, W. (1981*a*). Characterization of plasmids in halobacteria. *J. Bacteriol.* **145**, 369–374.

Pfeifer, F., Weidinger, G. & Goebel, W. (1981*b*). Genetic variability in *Halobacterium halobium*. *J. Bacteriol.* **145**, 375–381.

Pfeifer, F., Ebert, K., Weidinger, G. & Goebel, W. (1982). Structure and functions of chromosomal and extrachromosomal DNA in halobacteria. *Zbl. Bakt. Hyg., I. Abt. Orig. C* **3**, 110–119.

Pyle, L. E. & Finch, L. R. (1988). A physical map of the genome of *Mycoplasma mycoides* subspecies *mycoides* Y with some functional loci. *Nucl. Acids Res.* **16**, 6027–6039.

Rosenshine, I., Tchelet, R. & Mevarech, M. (1989). The mechanism of DNA transfer in the mating system of an archaebacterium. *Science*, **245**, 1387–1389.

Ross, H. N. M. & Grant, W. D. (1985). Nucleic acid studies on halophilic archaebacteria. *J. Gen. Microbiol.* **131**, 165–173.

Sankoff, D., Cedergren, R. & Abel, Y. (1990). Genomic divergence through gene rearrangement. *Methods Enzymol.* **183**, 428–438.

Sapienza, C. & Doolittle, W. F. (1982). Unusual physical organization of the *Halobacterium* genome. *Nature (London)*, **295**, 384–389.

Shimmin, L. C., Newton, C. H., Ramirez, C., Yee, J., Downing, W. L., Louie, A., Matheson, A. T. & Dennis, P. P. (1989). Organization of genes encoding the L11, L1, L10 and L12 equivalent ribosomal proteins in eubacteria, archaebacteria, and eucaryotes. *Can. J. Microbiol.* **35**, 164–170.

Sibold, L., Pariot, D., Bhatnagar, L., Henriquet, M. & Aubert. J.-P. (1985). Hybridization of DNA from methanogenic bacteria with nitrogenase structural genes (*nif*HDK). *Mol. Gen. Genet.* **200**, 40–46.

Sitzman, J. & Klein, A. (1991). Physical and genetic map of the *Methanococcus voltae* chromosome. *Mol. Microbiol.* **5**, 505–513.

Smith, H. O. & Birnstiel, M. L. (1976). A simple method for DNA restriction mapping. *Nucl. Acids Res.* **3**. 2387–2394.

Soppa, J. & Oesterhelt, D. (1989). *Halobacterium* sp. GRB: a species to work with!? *Can. J. Microbiol.* **35**, 205–209.

Sumper, M., Berg, E., Mengele, R. & Strobel, I. (1990). Primary structure and glycosylation of the S-layer protein of *Haloferax volcanii*. *J. Bacteriol.* **172**, 7111–7118.

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

Tartof, K. D. & Hobbs, C. A. (1987). Improved media for growing plasmid and cosmid clones. *BRL Focus*, **9**(2). 12.

Tindall, B. J., Ross, H. N. M. & Grant, W. D. (1984). *Natronobacterium* gen. nov. and *Natronococcus* gen. nov., two new genera of haloalkaliphilic archaebacteria. *System. Appl. Microbiol.* **5**, 41–57.

Wenzel, R. & Herrmann, R. (1988). Physical mapping of the *Mycoplasma pneumoniae* genome. *Nucl. Acids Res.* **16**, 8323–8336.

Wenzel, R. & Herrmann, R. (1989). Cloning of the complete *Mycoplasma pneumoniae* genome. *Nucl. Acids Res.* **17**, 7029–7043.

Woese, C. R., Kandler, O. & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 4576–4579.

Yokobata, K., Trenchak, B. & de Jong, P. J. (1991). Rescue of unstable cosmids by *in vitro* packaging. *Nucl. Acids Res.* **19**, 403–404.

Zusman, T., Rosenshine, I., Boehm, G., Jaenicke, R., Leskiw. B. & Mevarech, M. (1989). Dihydrofolate reductase of the extremely halophilic archaebacterium *Halobacterium volcanii*. *J. Biol. Chem.* **264**, 18878–18883.

Edited by J. Karn