Studies on the Structure of the Genome of *Haloferax volcanii*

by

Leonard Cornelis Schalkwyk

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
November, 1990

ISBN 0-315-64488-5

Table of contents

Figures

Tables

Abstract

Integrated bottom up and top down mapping has produced a nearly complete restriction map of the chromosome of *Haloferax volcanii* (3.5×10^6 bp) and complete maps of two plasmids of 4.2×10^5 bp and 8.9×10^4 bp, using six restriction enzymes. The map of the chromosome is composed of 131 cosmids grouped into nine contigs (groups of overlapping clones) which have been joined into two map fragments by pulsed field gel analysis. Regions with restriction site frequencies markedly different from that of the chromosome as a whole are located in one third of the genome,which is also very rich in copies of the insertion sequence ISH51 and two other, less well-characterized repetitive elements. Genes which can be located on the map by molecular hybridization are in general located outside of the site-rich region. The genes mapped include the two ribosomal RNA operons and 39 tRNA genes, of which ten had previously been sequenced, and a variety of protein-coding genes which have been cloned from *Haloferax volcanii* and other halobacteria. The two rRNA operons are opposite in orientation and nearly diametrically opposite each other on the chromosome, which is probably circular.

Abbreviations

bp	base pairs
BSA	bovine serum albumin
CHEF	contour-clamped homogeneous electric field
Cp	cytidine 3' monophosphate
cpm	counts per minute
DTT	dithiothreitol
FIGE	field inversion gel electrophoresis
kbp	thousand base pairs
KGB	potassium glutamate buffer
krpm	thousand revolutions per minute
mbp	million base pairs
mol%	moles per 100 moles
ND	not determined
NDS	0.5 M. Na_2EDTA. 1 M Tris base. 10 g/l sodium N-lauroyl sarcosinate
OFAGE	orthogonal-field alternation gel electrophoresis
pCp	[5'-^{32}P] cytidine 3'. 5' bisphosphate
PF	pulsed-field
PFGE	pulsed-field gel electrophoresis
PMSF	phenylmethylsulfonyl fluoride
rpm	revolutions per minute
SDS	sodium dodecyl sulfate
SSC	150 mM NaCl. 17.5 mM trisodium citrate
TE	10 mM Tris-HCl (pH 8.0). 1 mM EDTA (pH 8.0)
TEAS	50 mM Tris base. 20 mM sodium acetate. 30 mM acetic acid. 20 mM NaCl. 2 mM Na_2EDTA
YAC	yeast artificial chromosome

Acknowledgements

I have been financially supported at different times by studentships from the Natural Sciences and Engineering research Council, the Medical Research Council and The Killam Trust and from the operating grant of my supervisor, W. Ford Doolittle. In addition to his willingness to pay me, Ford has been a supportive and patient supervisor.

Most of this work was done in collaboration with Robert L. Charlebois, in particular the bottom-up mapping project, which was conceived and initiated by him. Cooperating with Rob has been enjoyable and productive even now that we are in different cities and communicating by electronic mail. Jason Hofman was also important in the first stage of bottom-up mapping, and has been a dependable source of technical advice.

INTRODUCTION

I. Map comparison

A. Maps and map comparisons

A map can be defined as a representation of a large number of observations arranged on a single sheet in such a way that broad features of the structure being observed can be discerned. The field of genetic mapping was founded in 1913 by Morgan, Bridges and Sturtevant (Sturtevant, 1913; Sturtevant *et al.*, 1919). They used an indirect measure (recombination frequency) to determine the relative locations of genes (of unknown physical nature) marked by mutation, along a linear structure thought to correspond to the X-chromosome of *Drosophila.* This resembles the work of (geographic) cartographers who have at least since Ptolemy drawn maps from an aerial perspective only seen by human eyes in modern times (Wilford, 1982).

The geographic analogy is not as superficial as it may at first seem – maps of both kinds are representations of large numbers of local observations in an integrated form that allows a higher level of structure to be understood. Each provides a framework upon which observations of many different kinds can be organized. Once the surface topology of a region has been mapped, for example, seismic, magnetic and other observations can be added, and used to study the underlying structure of the earth, as well as to choose places to drill for oil. Similarly, once there is a restriction map of a segment of DNA, observations made by Southern blotting, fragments of nucleotide sequence, and other observations can be added and studied for indications of underlying genome structure, and for interesting features for further study.

From the earliest days of genetic mapping, it was of interest to see if the order of genes was the same in different species. This was first done by Sturtevant (1921), who showed that some of the third chromosome markers in *Drosophila simulans* were in the opposite order to their counterparts in *D. melanogaster*, and one could postulate an inversion in the ancestor of one or the other. Comparative study of genome maps will be very informative about the functional and historical aspects of genome structure. In the study of gene structure by sequencing, the comparison of genes from the same species and of similar genes from different species has been a very powerful method of predicting which sequence features are functionally important.

Sequencing has also allowed inferences to be made about the history of the gene in question. In a similar fashion, comparison of maps should be informative about aspects of gene organization dictated by function and history. The depth to which we can see into the history of genomes will be determined by the extent to which gene order is maintained over geological time. Comparison of the detailed genetic maps of several eubacteria whose phylogeny can be independently inferred by comparison of their 16S ribosomal RNAs (Woese, 1987) gives a basis on which to make predictions. In this connection it is necessary to point out that genetic maps are only available for species representing a small part of the diversity of the eubacteria. All of the existing genetic maps are of gamma-subdivision purple bacteria or of Gram positive bacteria. There are likely many surprises left in more deeply branching groups, such as green non-sulfur bacteria, deinococci, and spirochetes.

B. Is map order conserved?

The most highly developed genetic map is that of *Escherichia coli* K12, which has been developed and refined over 40 years (Bachmann, 1990), with over a thousand markers, representing probably one half of the total number of genes (Kohara, 1990). The genetic map has recently been augmented with restriction maps produced by two different approaches, which I will review later. As the most nearly complete genetic map it can give us clues to structural features which may be functionally important, especially since some data exist on the effects of chromosomal rearrangements. This makes it a natural reference to which other maps of similar-sized genomes can be compared.

In comparing the genetic maps of *Escherichia coli* K12 and the closely related *Salmonella typhimurium* LT2, it is immediately evident that the two maps can be aligned. The two maps differ by an inversion of about 10% of their length symmetrically disposed about the terminus of replication (Sanderson and Hall. 1970) and by 14 (in *Escherichia coli* K12) and 15 (in *Salmonella typhimurium* LT2) loops of sequence not present in the other species (Riley and Anilionis. 1978).

The most detailed map of a Gram positive organism is that of *Bacillus subtilis* 168 (Piggot. 1990). The genome is about 20 percent larger than that of *Escherichia coli.* and also circular. but it is difficult to see any alignment with the *Escherichia coli* map. This impression. obtained by lengthy staring at the maps. agrees with the much more rigorous approach of Sankoff *et al* (1990). who have developed a measure of the number of rearrangements of one map order relative to another and compare this to what one would obtain by comparing with a selection of randomly shuffled maps. This is done by placing the two circular maps concentrically and connecting with an arc of less than 180° each pair of comparable loci. Each difference in relative map order is then signalled by intersections between arcs. Repeating this process as one map is rotated relative to the other allows the closest overall alignment to be chosen. The number of intersections can then be compared with that obtained by comparison of randomly shuffled genomes. To do the analysis. existing genetic maps must be reduced to a list of comparable marker names and coordinates. in itself a daunting task. Comparison of the *Escherichia coli* and *Salmonella typhimurium* maps gives 5% as many intersections as randomly shuffled genomes. and the comparison of *Escherichia coli* and *Bacillus subtilis*, 96%, or no detectable similarity. The analysis of Sankoff *et al.* (1990) also includes *Caulobacter crescentus* and *Pseudomonas aeruginosa*, two Gram negative (purple) bacteria whose behaviour in this analysis is much as one might predict from rRNA-based phylogenetic trees, even though the number of comparable markers available is small: *Caulobacter crescentus* and *Pseudomonas aeruginosa*, 40%; *Caulobacter crescentus* and *Escherichia coli*, 65%; *Pseudomonas aeruginosa* and *Escherichia coli*, 80%.

C. Are some general structural features conserved?

Transcription bias

In an actively growing cell, one can imagine a conflict between replication and transcription, especially if in opposing directions. Both processes involve complexes of proteins moving along the DNA. It is thus interesting to examine whether such conflicts are avoided by the orientation of genes such that their direction of transcription is the same as that of replication. This is the case for all seven ribosomal RNA operons in *Escherichia coli*. Direction of transcription is known for 707 *Escherichia coli* genes, and is predominantly in the direction of replication, in the ratio of 2.4:1. The subset of these genes encoding tRNAs, rRNAs, ribosomal proteins, and translation factors show this bias more strongly, with a ratio of 18.4:1 (Brewer, 1988; Brewer, 1990). The orientations of many of these genes could only be deduced after the completion of the restriction map of the entire *Escherichia coli* chromosome (Kohara *et ai.*. 1987) A more recent study has exhaustively searched for *Escherichia coli* tRNA genes and found 78, in 40 transcriptional units of which seven are the ribosomal RNA operons (Komine *et al.*, 1990). Orientations of all but two are known, and of the remaining 31 non-rRNA operons, fifteen are transcribed in the direction of replication, so the orientation bias does not hold for tRNA genes. The set of genes analyzed by Brewer contains some of these tRNA genes, so the bias is even stronger in the non-tRNA part of the sample. Orientation bias is likely to be widespread in the eubacteria, having also been detected in *Bacillus subtilis* (Zeigler and Dean, 1990), where 91 of 96 genes whose orientations are known are transcribed in the direction of replication. This number includes two of the *rrn* operons but no separate tRNA genes.

Position of genes on a bacterial chromosome may also have an effect on their expression since genes near the origin have a greater effective copy number than those near the terminus of replication. This is because a new round of replication can be begun before the previous one is finished. It may thus be advantageous for highly expressed genes to be located near the origin of replication. In *Salmonella typhimurium*, it has been shown that translocation of a gene (*his*D) to different regions of the chromosome causes differences in expression level. Expression is approximately proportional to the distance from the origin of replication and independent of orientation (Schmid and Roth, 1987). In *Escherichia coli*, the seven ribosomal RNA operons are all located in the origin-proximal half of the chromosome, as are the ten in *Bacillus subtilis* whose positions are known.

Constraints on genome reorganization

Both position and orientation effects may make rearrangements disadvantageous and thus favor conservation of overall genome structure. There may be other barriers to intrachromosomal recombination that favour conservation of gene order. These might include deleterious effects of disturbing the spacing of sites important in the organization of the folded chromosome or of disturbing the relative positions of the origin and terminus of replication. Details of the recombination and replication processes themselves may also make rearrangements which are not deleterious, nevertheless rare in occurrence. There are considerable data in *Escherichia coli* and *Salmonella typhimurium* LT2 on regions which do and do not permit inversions (François *et al.*, 1990; Mahan *et al.*, 1990)

D. Do some suites of genes keep together for functional reasons?

Having said that there is no detectable similarity between gene order in *Escherichia coli* and *Bacillus subtilis*, it is clear that it is a question of at what level of structure there is conservation. At the nucleotide sequence level, many genes are similar and can be aligned. Some operons have also genes in the same or similar order, for one example the rRNA operons in which the order is 16S-23S-5S throughout the eubacteria. Up to what level is order then conserved?

In comparing the sequence of the region of the origin of replication of *Bacillus subtilis* with the *rpn*A-*rpm*H-*dna*A-*dna*N-*rec*F-*gyr*B region of *Escherichia coli*, Ogasawara *et al.*(1985) found similar genes in similar relative positions. In this remarkable 10 kbp alignment, the order and orientation of 6 replication-related genes, though not the details of transcription or the size of the spacer sequences, are the same . The *Escherichia coli* genes are 45 kbp from *ori*C, but this may be an exceptional case in which the origin of replication and the associated genes have been separated relatively recently. In *Pseudomonas putida* and *Micrococcus luteus*, the same order of genes has been found, with the repeated consensus DnaA box found upstream of *dna*A in each case. In all of these organisms, transcription of *rpm*H and *dna*A is in opposite directions, and the

DnaA box repeats (the origin of replication) are in the middle, except in *Escherichia coli.*

Another case of possibly function-related gene order conservation is that of ribosomal protein and RNA polymerase operons in which the gene order in several archaebacteria is the same as it is in *Escherichia coli*, even though the sequences of the archaebacterial genes are more similar to those of the eukaryotes than eubacteria. Both *Halobacterium cutirubrum* and *Sulfolobus solfataricus* have operons with the same order as *rpl*KAJL of *Escherichia coli* (Shimmin *et al*., 1989). The spacing and details of transcription are different. In *Escherichia coli* the *rpo*BC operon is adjacent to *rpl*KAJL. This is not the case in the two archaebacteria, but they do both have RNA polymerase operons with the same gene order, even though the *Halobacterium* equivalent of *rpo*B is split into two genes (Zillig *et al*., 1989). *Methanococcus vannielii* has operons corresponding in a similar manner to the STR, S10, and SPC operons of *Escherichia coli* (Auer *et al*., 1989). This remarkable congruence must have to do with some functional reason for maintenance of the map order, but the expression and control of these genes is clearly not it. A hint is given by the finding of Rohl and Nierhaus (1982) that in *Escherichia coli*, groups of ribosomal protein genes whose products are interdependent for assembly are grouped together in operons. Some features of the ribosome assembly process might thus favor the maintenance of a particular operon structure.

E. Are there traces of genome building?

Even those who think of prokaryotes as lower forms of life and of their genomes as simple must admit that the bacterial or archaebacterial chromosome did not step from the ocean full blown. It is possible that there might be some traces of the process by which some most likely smaller, less refined and perhaps fragmentary chromosome precursor became a modern chromosome. Finding such traces is an ambitious goal, and it is quite possible that the traces have been completely lost.

One possible trace of chromosomal history is that the genes of several species seem to be organized into quadrants. This was first observed in *Streptomyces coelicolor* (Hopwood, 1967; Hopwood and Kieser, 1990). Almost all of the genes identified on the *Streptomyces coelicolor* map fall into two opposing quarters of the circular map. The other two quadrants are nearly silent. Furthermore, many genes of related function (belonging to the same biochemical pathway) are located opposite one another. It is not yet known whether the empty quadrants are really long stretches of silent DNA, or contain genes not easily marked by mutation, or are actually much shorter than their recombination map distances indicate. In any case, the fourfold arrangement of genes could be a trace of the growth of the genome from a smaller precursor by duplication, followed by loss or recruitment of some of the duplicated genes for other purposes.

A detailed search of the *Escherichia coli* map has revealed a tendency for metabolically related genes to be 90 or 180 degrees apart on the genetic map. This is the case to a statistically significant degree for genes grouped because they belong to the same pathway, and not for genes grouped on the basis of the type of reaction catalyzed by their products. The genes for glucose dissimilation in particular are organized in this way. Most of the 28 genes for the enzymes of the Entner-Doudoroff pathway, the tricarboxylic acid cycle, Embden-Meyerhof pathway and glyoxylate shunt are included in four clusters at four corners of the chromosome (Riley and Anilionis, 1978). Some other examples can be found but it is hard to judge their significance. Perhaps the analysis of the nucleotide sequence of whole genomes could eventually give some information on this point.

Traces of a more recent process can be seen with more confidence from the comparison of the genetic maps of *Pseudomonas putida* and *Pseudomonas aeruginosa*. In these species, auxotrophic markers (ie, genes for anabolic functions) are clustered in approximately half of the genome, and organized in a broadly similar way in both species. Many genes for catabolic functions can be mapped to the remaining part of the chromosome. Although the data are less extensive than for the anabolic genes, it looks like there is no similarity in gene order between the "catabolic" parts of the two genomes. The most likely reason for this is that the very versatile pseudomonads have received and integrated new catabolic capa bilities from plasmids as needed. Several transmissible plasmids encoding genes for the utilization of particular classes of substrates have been characterized, for example, the 115 kbp TOL plasmid, which contains the eleven genes involved in the degradation of toluene, xylenes and 1,2,4-trimethylbenzene. A strain of *Pseudomonas putida* has been characterized with a chromosomally integrated TOL plasmid (Holloway *et al.*, 1990)

F. What we can expect from an archaebacterial map.

It is not my objective in this work to produce a map sufficiently detailed to be able to determine whether the order of genes in *Haloferax volcanii* bears any resemblance to that of *Escherichia coli*. It is clear that there will be no resemblance, when enough markers are available for us to be able to see. On the other hand, there are some features of genome organization which we may find conserved, because they have functional importance in the operation of a compact circular genome, such as transcription bias and position preference, or because of the way the complement of genes was assembled. In isolation, a map of the *Haloferax volcanii* genome will be interesting because of features such as clustering of genes into certain regions, and a limited amount of information on conservation of gene clusters which may be functionally important. The map will become more and more interesting as further markers can be added to it. and especially as data from other archaebacteria can be compared.

II. Physical mapping

A. The development of physical mapping methods

While there is no fundamental difference between methods for genetic mapping of, say, mouse and maize, in the microbial world habits of growth and of genetic exchange are very diverse, and much has to be known about an organism before genetic mapping can be considered. For each case, a method of exchanging genes and measuring recombination (whether this is by conjugation, transformation, transduction or some other process) must be found, and an extensive set of marked strains isolated. If one hopes to be able to obtain enough maps to be able to compare and make conclusions about genome structure and its evolution, it is necessary to develop streamlined physical methods of mapping which are independent of the details of the biology of each organism. This technical development can also be seen as progress toward eventual comparative sequencing of whole genomes, which for prokaryotes is already in sight (Church and Kieffer-Higgins, 1988). For those studying the structure of large eukaryotic genomes, large-scale physical mapping techniques are desirable for a different reason, which is to bridge the gap between very coarse genetic and cytological methods and very fine molecular ones.

Until 1985, mapping methods using isolated DNA as the starting material were limited to fragments resolvable by agarose gel electrophoresis or clonable in *Escherichia coli* – at most 40 kbp.

Such units could be arduously strung together by chromosome walking (Bender *et al.*, 1983), in which an hybridization probe derived from the end of a characterized clone is used to isolate new clones. Restriction maps could also be made of larger DNAs that could be easily isolated (such as those of bacteriophages) or that could be specifically labelled such as the *Escherichia coli* *ter*C region (Bouché, 1982). Two approaches which emerged in 1985 have made restriction mapping of entire prokaryotic genomes a realistic goal.

The top-down approach is to divide the genome into a manageable number of pieces, determine the size and linkage of each, and so construct a map. This was made possible by the invention of pulsed-field electrophoresis (Schwartz and Cantor, 1984), which does not suffer from the same size limitation on resolution as conventional electrophoresis. The availability of a large variety of restriction enzymes is also necessary for this approach. For many DNAs, digests with a convenient number and distribution of fragments can be found.

The bottom-up approach, in contrast, does not depend on any specific technical advance, but on the realization that an efficient method of analyzing many randomly chosen clones in parallel could be reasonably used to piece together a map of any genome of modest size. Such an approach produces a set of characterized clones allowing any sequence of interest to be extracted with ease, in addition to a map.

B. Top-down mapping

Top-down mapping depends on pulsed-field gel electrophoresis, so I will review the development of the technique before describing the top-down mapping which has been published to date. I will confine myself to work on small genomes, on the grounds that the work on human and other highly complex genomes is too voluminous and too different in approach from the work I will describe for inclusion.

Top-down mapping depend on the ability to separate DNA fragments of a wide range of sizes. Conventional electrophoresis (as well as sedimentation) methods fail to separate DNAs of above 100 kbp and perform poorly above 20 kbp for linear molecules. In very general terms, this is thought to be because DNA molecules under the influence of a force such as that imposed by an electric field, assume an elongated conformation parallel to the direction of the force, and their passage is thus resisted by the medium to an extent little affected by their size (Lumpkin and Zimm, 1982). This end-on motion is known as "reptation" in electrophoresis and "tunneling" in zonal centrifugation. Smaller DNA molecules are relatively free to tumble with their own thermal motion and that of the surrounding water and thus present their broad side some fraction of the time, but as the length increases the rate of this motion decreases for the molecule as a whole, and in an agarose gel it is further constrained by the matrix, whose pores are far smaller than the length of the molecule. Within this conceptual framework it can be seen why some advantage in the separation of large DNAs can be had by electrophoresis at low field strengths, since at low fields random motion will be more important relative to electrophoresis. Low agarose concentrations have been used for these separations (Fangman, 1978; Serwer, 1981), in order that the length of the experiment can be kept within a reasonable limit.

Hope of freedom from these limitations was offered by the discovery by Schwartz and Cantor (Schwartz and Cantor, 1984; Schwartz *et al.*, 1982) that DNA of over 100 kbp could be separated by electrophoresis when two fields at approximately right angles are alternately applied. Because it seemed at the time that an inhomogeneous field (field gradient) was essential, the term "pulsed field gradient gel electrophoresis" was coined and PFG of PFGE (now translated as pulsed field gel) has stuck as the general term for the varied methods which are based on this principle. The more euphonious and less inaccurate term "jiggle gel" coined by J. Hofman has unfortunately not caught on.

In order to analyze large DNAs, they must be obtained in intact form. DNA prepared by standard extraction methods derived from that of Marmur (1961) is extensively sheared and rarely has an average size exceeding 100 kbp, though when extreme care is taken, very large DNAs can be isolated and detected over a background of sheared molecules (for example Carle and Olson (1984)). Lysis of bacterial or some eukaryotic (Cook, 1984) cells in 2M NaCl had allowed the preparation of intact, supercoiled DNA in the form of "folded chromosomes" or "nucleoids". The DNA in these becomes sensitive to shearing when the protein and RNA components are extracted. Agarose can be used to replace the nuclear "cage" and protect the DNA from shearing, as demonstrated by Cook (1984). The cells were encapsulated in beads of agarose before lysis. Nucleoids produced in this way could be shown to have characteristics similar to those produced in solution, and the beads could be further extracted with detergents or proteases to produce nearly pure DNA in a form that could be pipetted without shearing. The beads are produced by homogenizing the cells and agarose with mineral oil to produce an oil-continuous emulsion, a method originally developed for making chromatography media. The method was simplified by Schwartz and Cantor (1984), who cast the agarose into blocks of a convenient size.

The first truly practical implementation of the pulsed-field method was that of Carle and Olson (Carle and Olson, 1984; Carle and Olson, 1987), which they termed Orthogonal Field Alternation Gel Electrophoresis (OFAGE). In contrast to the earliest designs of Schwartz and Cantor (Schwartz and Cantor, 1984), in the OFAGE device the direction of net migration is perpendicular to the wells of the gel, and useful comparisons can be made between samples run in parallel. The disposition of the electrodes is such that the angle between fields is 90° near the origin and decreases in the direction of migration. The electric field also decreases along this line providing a band-sharpening effect, since the leading edge of a band sees a smaller field than the trailing part does. The design embodies a sophisticated understanding of the behaviour of pulsed field electrophoresis, and was the first published in sufficient detail to be easily reproduced. Mobility of DNA over the range of sizes applicable for a given set of conditions, is a nearly linear, monotonic function of length, as judged by the behaviour of yeast chromosomes and of multimers of bacteriophage lambda DNA (Carle and Olson, 1984)

It was clear from soon after the development of PFGE that the the range of DNA sizes resolved depends on the frequency of switching of the field, the upper size limit increasing with increasing switching time. The physical theory of pulsed field electrophoresis is not yet fully developed, but it is clear that the switching time dependence has to do with the speed with which molecules can be reoriented and that this process resembles viscoelastic relaxation. Viscoelastic relaxation (Kavenoff and Zimm, 1973), was previously the only method of measuring the size of large DNA molecules without measuring individual molecules microscopically. Relaxation is measured as DNA in solution recoils after having been stretched by the application of a shearing force.

Emerging models of the mechanism of PFG electrophoresis (Lalande *et al.*, 1987; Mathew *et al.*, 1988; Southern *et al.*, 1987) were set on their ear when it was reported that an angle of 180^{o} between fields (*ie.* periodic reversal) would produce separation of large DNAs (different potentials or times are used in the two directions to cause net migration) (Carle *et al.*. 1986). This allowed for the first time perfectly homogeneous conditions. both from one lane to the next and over the length of the gel. Under such conditions. mobility is no' a monotonic function of length. but instead goes through a minimum at a value that varies with the switching frequency. This can be understood by imagining that at a given switching frequency. DNA molecules of a certain size will spend all of their time reorienting or relaxing. while those larger or smaller will partly retain or completely lose. respectively. their initial conformation during a period of reversed field and thus achieve some net migration. A more useful monotonic sizing curve can be achieved in Field Inversion Gel Electrophoresis (FIGE) by varying the switching frequency ("ramping") during the run. The FIGE technique has achieved some popularity but its utility was hindered. at least initially. by the complexly interacting variables of run length. ramp steepness and shape that had to be optimized in order to obtain a desired separation. With a sufficiently detailed understanding of the method. it should eventually be possible to tailor the separation to the detailed requirements of a given experiment.

One can imagine the inhomogeneity of field strength and angle over space in the original PFGE and OFAGE devices achieving the same end as the switching ramp does in FIGE. This has been shown to be an oversimplification by the success of Contour-Clamped Homogeneous Electric Field (CHEF) electrophoresis (Chu *et al.*. 1986). CHEF and its derivatives are currently the methods of choice in pulsed field electrophoresis. Itself a modification of OFAGE, CHEF uses homogeneous fields at an angle of 120^{o} , in such a way that neither the fields nor their angles vary appreciably over the area used for the separation. Homogeneity is achieved by the use of a contour clamp: in this case a hexagonal array of electrodes, connected to each other through resistors. The resistors, all of the same value, are chosen so that a large fraction of the current flows through this external circuit. The potential difference between two adjacent electrodes (the field strength) is then dictated by the resistor which connects them. The contour clamp is an elegant passive way of applying the principle that by controlling the potentials of points around the perimeter of a closed figure, any desired field can be produced within. The current is applied to two opposing sides of the hexagon to produce a homogeneous field for one switching period, and to another pair for the other. The third pair of sides is permanently idle. The performance of CHEF is similar to that of OFAGE except that the lanes are more nearly straight, and resolution of DNA molecules above 1 million bp (mbp) is somewhat better. The range of molecular weights separable on a single gel is also somewhat smaller, but this can be changed by switching frequency ramping.

Several more devices for PFGE have since been introduced, such as one using an actively controlled driver for each electrode to replace the contour clamp (Birren *et al.*, 1989), a vertical electrophoresis tank (without plates) in which the field is switched perpendicular to rather than in the plane of the gel (Gardiner *et al.*, 1986), a device which rotates the gel relative to fixed electrodes (Southern *et al.*, 1987), and one in which pulses from three directions are alternated (Bancroft and Wolk, 1988).

Electrophoretic karyotypes.

The karyotype of an eukaryotic organism is its complement of chromosomes, which can often be seen at metaphase in the light microscope. This is not possible in *Saccharomyces cerevisiae*, and Carle and Olson(1985) coined the phrase "electrophoretic karyotype" for their identification of the genetically characterized linkage groups with the DNA species which could be separated by OFAGE. Electrophoretic karyotyping was the first application of PFGE and it allowed direct examination of the karyotypes of several cytologically intractable eukaryotes with small genomes, such as the ascomycetes *Saccharomyces cerevisiae* (Carle and Olson, 1985), *Schizosaccharomyces pombe* (Smith *et al.*, 1987b), *Candida albicans* (Lasker *et al.*, 1989) and *Neurospora crassa* (Orbach *et al.*, 1988). The chromosomes of protozoa such as *Trypanosoma brucei* and *Plasmodium falciparum* can also be separated (Foote and Kemp, 1989; Van der Ploeg *et al.*, 1989). Each of these was in its turn a challenge to the upper limit of resolution of the technique.

I also include in this category those studies of prokaryotic genomes that, while they may involve restriction digests, determine few sites, or place few markers such as the several genome size and circularity determinations that have now been done (see table 1).

Top down maps

Several more detailed top down studies of prokaryotic genomes have been completed, including one archaebacterium, *Thermococcus celer* (Noll, 1989). The first map was that of *Escherichia coli* K-12 (Smith *et al.*, 1987a). This map is largely an alignment of the 22 NotI fragments with the existing genetic map, by probing of transfers of PFG gels with probes from representative genetically mapped genes. As such it does little to illustrate the potential and problems of the method for mapping unknown DNAs. Two important mapping approaches, junction cloning and partial digestion are used in the paper, so I will take this opportunity to introduce them.

In the junction cloning approach, clones are found which contain the rare restriction sites which are to be mapped. These can then be used as hybridization probes to demonstrate the contiguity of two fragments. Many clever ways to find junction clones have been proposed, but obtaining a complete set is an arduous proposition and no map has been completed by this method to date. In the case of the *Escherichia coli* map, several NotI junctions were obtained by screening the GenBank sequence database.

Partial digestion is used as in the Smith and Birnstiel (1976) method of indirect end-labelling, except that in top-down mapping

species	size	digests	circular?	ref
Myxococcus xanthus	9.4	2	N.D.	1
Borrelia burgdorferi	1	2	linear	2.3
Pseudomonas aeruginosa	5.4	1	N.D.	4
Caulobacter crescentus	4.0	2	yes, by correlation with (linear) genetic map	5
Coxiella burnetii	1.6	1	N.D.	6
Rickettsia melolonthae	1.7	2	yes, by partial digestion	7
6 Rickettsias/ Chlamydias	1.4-2.1	2	N.D.	7
7 Ureaplasma/ Mycoplasmas	0.9-1.28	1-2	N.D.	8
Sulfolobus acidocaldarius	2.1	1	yes, by junction clones	9

Table 1. Electrophoretic karyotypes of prokaryotes. N.D., not determined. Genome sizes are in millions of base pairs. References are:

1. (Chen *et al.*, 1990)
2. (Baril *et al.*, 1989)
3. (Ferdows and Barbour, 1989)
4. (Hector and Johnson, 1990)
5. (Ely *et al.*, 1990)
6. (Heinzen *et al.*, 1990)
7. (Frutos *et al.*, 1990)
8. (Pyle *et al.*, 1988)
9. (Yamagishi and Oshima, 1990)

one does not generally have the luxury of a unique starting site. Nevertheless, in this case many fragments were ordered by probing NotI partial digests with small gel-purified NotI fragments. This at least identifies the fragments immediately neighboring the probe.

A top-down map similarly dependent on an existing genetic map was that of the 35 DraI sites of *Caulobacter crescentus* (Ely and Gerardot, 1988), making use of TN5 insertions into many known genes. An engineered transposon was also used to introduce NotI sites (otherwise absent from the genome). Similarly Ventra and Weiss (1989) aligned 31 NotI fragments with the *Bacillus subtilis* chromosome by probing many genetically characterized TN917 insertion strains with the transposon. This produced a partial ordering of the fragments.

An early complete and independent map was that of *Anabaena* PCC7120 (Bancroft *et al.*, 1989). Many restriction enzymes cut this DNA infrequently, which may have to do with the many restriction enzymes found in *Anabaena* species, whose recognition sequences are among those under-represented (Herrero *et al.*, 1984). The mapping was primarily achieved by the use of restriction fragments isolated from pulsed field gels as probes on Southern transfers of digests of the chromosomal DNA made with other enzymes. This is a method of wide applicability in genomes without large numbers of repeated sequence elements. Use was also made of a SalI junction library prepared by *in vitro* circularization of (relatively small) HindIII fragments of the chromosome and then cutting with SalI and cloning the rare SalI- linearized fragments into a SalI-cut positive selection vector.

The distribution of the rarest restriction sites is interesting: the three SphI sites are each close to one of the five SstII sites. This may indicate their joint introduction on a prophage, transposable element or plasmid.

The genome of *Clostridium perfringens* has been mapped by comparison of single and double digests using six enzymes, and by indirect end labelling and partial digestion (Canard and Cole, 1989). Enough markers were placed on the map to allow the conclusion that the organization of this genome resembles that of *Bacillus subtilis* in that the ribosomal RNA and tRNA are concentrated in the third of the genome including the origin of replication. The origin was not located directly, but the order of genes *gyr*B-*gyr*A-*rrn* suggests a very likely region.

In the mapping of the *Rhodobacter sphaeroides* 2.4.1 genome (Suwanto and Kaplan, 1989), a more brute-force method of searching for junction clones was used. 350 cosmids were individually screened for AseI sites. Six site-containing fragments were found, of which 2 were apparently cloning artifacts, 2 were from the five plasmids present in the strain, and the other two were informative. Another junction fragment was serendipitously discovered when a cloned gene was used as a probe, and a further (uncloned) junction fragment was discerned from double digestions. Most of the *Rhodobacter* map was determined by probing of PFG separations of several digests with cloned genes, revealing which fragments overlap which others (from a purely restriction mapping point of view this could have equally well been done with anonymous clones). The genome unexpectedly turned out to consist in two circular chromosomes, of which the smaller one contains two of the three ribosomal RNA operons. Forty markers were placed by hybridization with cloned probes. This genome was a nearly ideal subject for top-down mapping, with a variety of restriction enzymes that cut it into a manageable number of fragments, allowing the construction of a quite dense map with a simple approach.

An extension of the two dimensional method of Yee and Inouye, (1982) to PFG was used by Bautsch (1988) to map the *Mycoplasma mobile* genome. The method involves cutting a lane from a gel (first dimension), redigesting the DNA *in situ* and electrophoresing again in a second dimension. A digest with one enzyme in the first dimension and another in the second, done in parallel with the reciprocal combination, identifies which fragments overlap and by how much. The remaining fragments were ordered by separating a partial digest in the first dimension and then completing the digestion for the second dimension. This very direct approach is attractive because it requires no cloning or hybridizations. The mapping of the *Pseudomonas aeruginosa* PAO genome by the same approach (Romling *et al.*, 1989) is an impressive achievement. The map produced is independent of the genetic map which already existed, and agrees with it.

Species	genome size	no. sites	no enzymes	markers	genetic map	reference
Escherichia coli K-12	4.7	22	1	entire genetic map aligned	yes	1
Anabaena PCC7120	6.4	59	5	30 incl 2 rrn	part	2
Thermococcus celer	1.9	22	3	3 stable RNA loci	no	3
Pseudomonas aeruginosa PAO	5.9	51	2	5 protein coding, 4 rrn	yes	4
Caulobacter crescentus	3.8	35	1	based on genetic map	yes	5
Mycoplasma mobile	0.78	19	4	no markers	no	6
Rhodobacter sphaeroides 2.4.1	3.0, 0.9	43	3	40 incl 3 rrn	part	7
Mycoplasma mycoides	1.2	32	8	2 rrn ori. and ter. of replication	no	8
Bacillus cereus	5.7	11	1	13 cloned genes	no	9
Haemophilus influenzae	2.0	37	3	5 antibiotic res. genes by transformation	part	10
Haemophilus influenzae	1.9	41	3	6 rrn,6 antibiotic res. genes by transformation,11 cloned genes	part	11
Clostridium perfringens	3.6	63	6	9 rrn,16 protein, tRNAs linked to rrn	no	12
Bacillus subtilis 168	4.7	31	1	based on genetic map	yes	13
Haemophilus parainfluenzae	2.3	35	3	6 antibiotic res. genes	no	14

Table 2. Top-down maps. Those with a small number of sites are found in table 1. Sizes are in millions of basepairs.
References: 1.(Smith *et al*., 1987a), 2.(Bancroft *et al*., 1989), 3.(Noll, 1989), 4.(Romling *et al*., 1989), 5.(Ely and Gerardot, 1988), 6.(Bautsch, 1988), 7.(Suwanto and Kaplan, 1989), 8.(Pyle and Finch, 1988), 9.(Kolstø *et al*., 1990), 10. (Kauc *et al*., 1989), 11.(Lee and Smith, 1988; Lee *et al*., 1989), 12. (Canard and Cole, 1989), 13. (Ventra and Weiss, 1989), 14.(Kauc and Goodgal, 1989)

In the mapping of the genome of *Haemophilus influenzae* RD. several infrequent cutting enzymes of this AT-rich DNA were found to have sites in each of the 6 ribosomal RNA operons. making it difficult to map across them (Lee and Smith, 1988: Lee *et al.*, 1989). Luckily. RsrII resolved this difficulty - there are four sites. none in the rRNA operons. Fragments isolated from PFG served as probes to link the RsrII fragments and further fragments could be aligned with this framework by hybridization and consideration of single and double digests. Markers were located on the map by hybridization and by transformation using fragments isolated from PFG gels. In one natural isolate, an inverted duplication of 43 kb was detected. A similar approach was used by Kauc and Goodgal (1989) to produce a map of the *Haemophilus parainfluenzae* genome.

The only map of an archaebacterial genome completed to date is that of *Thermococcus celer* (Noll. 1989). This relatively small genome was mapped by linking SpeI fragments with both cloned and uncloned junction fragments and alignment of other fragments to the map by hybridization and by digestion of isolated fragments. The markers on this map so far are the stable RNA genes, 16 and 23S rRNA. the unlinked 5S rRNA and 7S RNA.

The maps described above have in common a relatively crude scale, though in principle any number of additional sites can be added once a map is established. Markers (small in number in most cases so far) are located to the nearest interval between sites, though a promising method for locating hybridization markers relative to the ends of a fragment by X-ray breakage has been reported (Game *et al.*, 1990). In general, top-down mapping has not yet lived up to its initial promise of allowing the construction of maps of many organisms quickly and independently of previous detailed knowledge of the genome structure of the organism. The maps are still in the early stages of development, and they will be populated by many more markers in the future, allowing more detailed comparisons to be made.

Bottom up mapping has, in addition to its inherently greater detail, the advantage of producing a set of clones, which is a valuable resource. A set of clones minimally covering a whole genome can be readily screened for genes by hybridization or by biological assays such as transformation. The clones themselves can be used as probes for aligning maps of other species for comparison. It is at the same time true that cloning artifacts, extensive repeated sequences and other problems could fool the strictly bottom up mapper. The effort required to clone the last segments of a genome will also be out of proportion to their incremental value. These two points argue for an integrated approach, in which long-range and local mapping techniques are combined.

C. Bottom up maps

The first method of producing a map by assembling maps of clones was chromosome walking (Bender *et al.*, 1983). The idea of chromosome walking is to use a clone to screen a library for further fragments, take these and screen again. In practice, this involves analysis of numerous candidates at each step. Starting with a plasmid clone known to be in the region of interest (in this case near the *Ace* locus of *Drosophila melanogaster*) by *in situ* hybridization, Bender, Spierer and Hogness (1983) screened a library of *Drosophila melanogaster* fragments (in a lambda replacement vector) by hybridization, mapped the positive clones by restriction and heteroduplex analysis, then isolated a restriction fragment for use as a probe in the next round. Thus it was possible to walk along the chromosome until repeated sequences were encountered, and most of the clones isolated did not contain sequences from the region of interest. Chromosome walking has been widely used, but because of its linear nature it is very time consuming, in addition to being laborious. It has been used to construct one complete map of a bacterial genome, that of *Mycoplasma pneumoniae* (Wenzel and Herrmann, 1988a; Wenzel and Herrmann, 1989).

species	genome size, mbp	type of clones	number of clones analysed	number of contigs	number of single clones	method of analysis	reference
Saccharomyces cerevisiae	15	λ	4946	1422	742	fragment sizes	1
Anabaena variabilis	5.4	cosmid	960	40		hybridization	2
Escherichia coli K-12 (W3110)	4.7	λ 15.5 kbp	1056	63	7	partial digest 8 enzymes	3
Escherichia coli K-12 (W3110)	4.7	λ	2344	7		hybridization	3
Escherichia coli K-12 MG1655	4.7	λ	2000	90	300	rest.mapping 3 enzymes	4
Escherichia coli K-12 803	4.7	cosmid	2512	58		fingerprint	5
Escherichia coli K-12 BHB2600	4.7	cosmid	570	12		hybridization,Southerns	6
Escherichia coli K-12 (W3110)	4.7	cosmid	1300	31		hybridization, fingerprints	7
Mycoplasma pneumoniae	0.8	cosmid				walking	8
Caenorhabditis elegans	80	cosmid	8000	860		fingerprint	9
Caenorhabditis elegans	80	cosmid	17500	700		fingerprint	10
Caenorhabditis elegans	80	YAC	1000	346		hybridization	10
Myxococcus xanthus	10	YAC 110 kbp	409			partial digest 1 enzyme	11

Table 3. Summary of the bottom-up mapping projects published to date. Different stages of the same project are given as separate entries. References are: 1.(Olson *et al.*, 1986); 2.(Bancroft *et al.*, 1989); 3.(Kohara *et al.*, 1987); 4. (Daniels, 1990; Daniels and Blattner, 1987); 5.(Knott *et al.*, 1988); 6. (Birkenbihl and Vielmetter, 1989); 7.(Tabata *et al.*, 1989); 8. (Wenzel and Herrmann, 1988a; Wenzel and Herrmann, 1988b); 9.(Coulson *et al.*, 1986); 10.(Coulson *et al.*, 1988); 11.(Kuspa *et al.*, 1989)

The simultaneous innovation of several groups (Coulson *et al.*, 1986; Herrero and Wolk, 1986; Olson *et al.*, 1986) was to analyze many randomly chosen clones in parallel. Since in the early stages every clone isolated contains new sequence, and since cloning of the entire genome is the objective, efforts to direct acquisition of new clones are wasted. These three projects, as well as others published later, all involve the analysis of many clones in order to identify overlaps, a process to varying degrees independent of the restriction mapping of the individual clones. In the following discussion, I will use the word "contig" to mean a group of twc or more clones known to overlap. Contigs and lone clones are collectively known as "islands".

The most straightforward approach for detecting overlaps, derived directly from chromosome walking, is to hybridize cosmid DNAs to dot blots of DNA of a large number of cosmids. This has been used to produce a partial set of ordered clones of *Anabaena variabilis* DNA (Herrero and Wolk, 1986). After initial random choices of probes, contigs could be identified and outer cosmids chosen as probes. The following criterion was chosen for defining a set of cosmids as linked: at least three cosmids must hybridize exclusively with the set. If something less than this rather stringent criterion was used, more links were made, some of which were anomalous, joining an end to the center of a contig. Without restriction mapping of the cosmids, it was not possible to distinguish causes such as chimaeric clones, repeated sequences, or rearrangements in the original culture. Hybridization methods can be further streamlined by pooling probes to reduce the number of hybridizations that need to be done, and the manipulations are simple and easily automated (Evans and Lewis, 1989).

Another alternative is to use a computationally intensive approach to get the most information possible from very simply obtained data. This was the approach used to obtain a partial bottom up map of the yeast genome (Olson *et al.*, 1986). The 5000 lambda clones analyzed were digested with EcoRI+HindIII (treated as a single enzyme), and the sizes of the fragments between 400 bp and 7.5 kbp were determined. Since the fragments originally cloned were produced by different enzymes or by mechanical shearing, the outermost part of each clone (which is fused to one or other arm of the vector) is ignored as are the (estimated 30) fragments above 7.5 kbp. The entry of band positions was automated and for each clone, previously analyzed clones were searched for matching fragments using an error window that depends on the size of the fragment. Overlapping cosmid candidates with more than five fragments in common were aligned in all of the possible ways, the best-fitting chosen and kept if it fulfilled a goodness of fit criterion. By this method, 85% of the clones were joined into 680 contigs, of which modelling had predicted 10 percent would include false linkages. Of the remaining isolated clones, many have less than 5 fragments and so cannot be linked up by this method. Some information is also available on the order of fragments from the comparison of many clones overlapping to different extents. This was used to give a partial map in the same operation, and also provided a sensitive method of detecting chimaeric clones, false overlaps, and some overlaps which escaped detection due to the five-fragment criterion.

A different approach was taken in the *Caenorhabditis* project (Coulson *et al.*, 1986) . Cosmid DNAs were digested into many fragments, of which a subset were labelled with ^{32}P. This allows the number of fragments analyzed to be independent of their size. The size was chosen so that the products could be separated on a sequencing gel, to nearly 1 bp resolution. This highly precise data simplifies the pattern-matching problem. The positions of the labelled bands were measured with the aid of a digitizing tablet and the previously analyzed cosmids were searched for matching fragments. The final decision on whether two cosmids overlap was made by direct comparison of the autoradiograms. This project suffered from an obvious cloning bias. For example ribosomal RNA genes, expected to be present on .05% of the clones, accounted for up to 5% of the primary cosmid banks.

Yet another approach was taken in the *Escherichia coli* mapping of Kohara (1987). This is the only case so far in which the detailed restriction map of each clone (for 8 enzymes) was used to identify overlaps. DNAs from lambda clones were partially digested with each enzyme, the digests separated on 0.4% agarose gels, transferred and hybridized with a probe derived from the right arm of the vector. The partial digest ladders that this produced were in the first instance simply used to determine the order of sites. Clones with sequences of five or more sites in common were taken to overlap, since such a sequence is expected less than once in the genome. This could be confirmed by comparing the maps complete with distances. This approach was further refined by considering the distances between sites and the possibility of reversal of the order of closely placed sites. This is by far the greatest density of information extracted by any of these approaches and thus is in principle capable of identifying the smallest overlap. 1025 clones were thus joined into 70 contigs, including 7 single clones. These were calculated to cover 94% of the genome, and approximately the same amount of effort again would be required to clone the rest by the same method, so a second stage was undertaken in which chromosome walking from the 70 ends (done in parallel) closed all but seven of the gaps. The restriction map could be aligned with the existing genetic map by comparison of published restriction maps and maps derived from published sequences. Three of the gaps were sequences already cloned by others. One contains the *lpp* gene, previously reported to be difficult to clone. Another contains *ori*C, which was probably not cloned because of the high density of Sau3AI (cloning enzyme) sites in this region.

These three projects demonstrated that the general approach is reasonable, even for a genome as large as that of *Caenorhabditis*. It is also clear that the design of a contig-building strategy is a complicated and subtle problem. A mathematical analysis of fingerprinting approaches by Lander and Waterman (1988) makes clear the most important considerations. The first is the sensitivity of detection of overlaps. The fraction of the length of a clone which must overlap in order to be detected (θ) depends on the amount of information the fingerprinting method has extracted from each clone and on the stringency of the criterion used to decide on overlaps. The stringency must be chosen so that false overlaps are rare. Assuming for the moment that θ does not vary from clone to clone (and that clones begin at random points in the genome), the expected length of an apparent contig at a given stage of the project is

$K=L[((e^{c\sigma}-1)/c)+\theta]$, where

K is the apparent length of a contig

L is the length of an insert

c is the number of genome lengths of cloned DNA analyzed

σ is $1-\theta$

Thus progress of the project per unit of effort (c) by this criterion goes with e^{σ}. Minimizing θ is thus of great importance. This must be balanced with the effort required to extract additional information from each clone. The other important consideration is L. The contig length goes linearly with L, for a given value of c, but this is expressed in terms of length of DNA analyzed. Effort expended is more nearly proportional to the number of clones isolated, regardless of their size, so the effect is much greater. Each cloning method has its own practical advantages and disadvantages, which must be considered against this.

The Lander and Waterman analysis fits well with the data of Kohara, Akiyama and Isono (1987) for which θ is about 0.2 (approximately 3 kbp overlap detectable in a clone of 15 kbp), and of Olson (1986) for which it is on the order of 0.6 (5 fragments required over 8.36 fragments in the average clone). In the case of *Caenorhabditis* (Coulson *et al.*, 1986) progress was less than expected both from the mathematical analysis and from the simulations done by the authors, most likely because of the strong cloning bias in their library. That the analysis fits the first two cases indicates that there is not a large deviation from random cloning, since the derivation depends on an assumption of randomness. Olson *et al.* (1986) comment that the depth of their contigs is greater than expected, leading them to suspect that the size of the genome is smaller than it has been thought to be. If this is the case, there might then be room for a certain amount of cloning bias in this analysis.

Four more assaults on the *Escherichia coli* K-12 map have appeared in the literature since the Kohara map. Although these are to some extent complementary to that work, none has been published in the same detail. Detailed information on these projects would be very useful for evaluating different approaches to bottom up mapping, since they differ in efficiency. This kind of comparison is useful, but much less interesting than that which would have been possible had the same effort been expended on producing maps of the genomes of a variety of organisms.

A map nearly as complete as that of Kohara was produced in a very efficient manner by Birkenbihl and Vielmetter (1989). Cosmid clones were isolated and purified DNAs were spotted on filters as well as being digested with EcoRI and run on agarose gels. The gels were used to screen for clones with suspiciously undersized or oversized inserts, poor DNA and other problems. Hybridization of individual clones to the dot blots identified overlapping clones, just as was done by Herrero and Wolk (1986), but then the overlaps could be further characterized by hybridization of selected probes to Southern transfers of the EcoRI gels. This approach required the preparation of only 570 cosmid DNAs and it leaves only 12 gaps of up to 40 kbp in size. The gaps include the *lpp* and *ori*C gaps also present in the Kohara map.

A more laborious *Escherichia coli* mapping project used the fingerprint method of Coulson on a total of 2512 cosmids from six different libraries in specially constructed cosmid vectors to produce 58 map fragments from 40-300 kbp in size (Knott *et al.*, 1988). The fragments cover an estimated 90% of the genome. It is difficult to estimate the amounts of overlap these different approaches can detect, but the difference in success between this project and that of Birkenbihl, with four times fewer clones, is probably too large to be accounted for by overlap alone. The quality of the initial bank of clones is very important, and must have been very good in the latter case. Birkenbihl and Vielmetter clearly demonstrate that cosmid clone libraries are practical for bottom-up mapping and need not be unstable or highly biased.

Conventional restriction mapping of lambda clones using single and double digests has also been used to produce a partial *Escherichia coli* encyclopaedia (ordered set of clones: Daniels, 1990; Daniels and Blattner, 1987). Analysis of some 2000 clones has left 90 contigs and 300 additional single clones. Such a high number of singletons indicates that many overlaps have not been detected.

Finally, Tabata *et al.* (1989) seek not so much to make an independent cosmid-based map of the *Escherichia coli* genome as to produce a set of cosmids complementary to Kohara's minimal set of lambda clones. The cosmid set would be useful for genetic analysis, for example complementation. 1300 cosmids were analyzed by a mixture of methods including fingerprint analysis by the method of Coulson (1986), hybridization with NotI fragments of genomic DNA, hybridization to clones of the Kohara set, and genetic complementation. These extensive efforts produced a set of clones covering about 70% of the genome, much sequence apparently having been absent from the starting bank.

Two applications of yeast artificial chromosome (YAC) cloning techniques (Burke *et al.*, 1987) to bottom-up mapping have so far been published. YACs are attractive because of the large size of insert which can be obtained, but they have the disadvantage that the cloned DNA is not easily separated from that of the host, and the difficulty of analysis of the cloned DNA begins to approach that of ordinary chromosomal DNA. Coulson *et al.* (1988) have used YAC clones to join many of the the contigs of their *Caenorhabditis* map. After the initial *Caenorhabditis* cosmid mapping paper (Coulson *et al.*, 1986) was published, the total number of cosmids analyzed was brought to 17,500, and the contig count down to 700. This was judged to be the practical limit, and YAC libraries were constructed in order to span the gaps. This was promising not only because the inserts are large, but also because whatever cloning bias there might be, this is likely to differ from that seen in cosmid clone banks. Initial attempts to do fingerprint analysis on the YACs met with problems, some of which were necessary consequences of the complexity of these clones, so hybridization of entire YACs to a selected array of cosmids was used instead. This has allowed the reduction of the contig count to 346. One third of the joins proved to be previously undetected overlaps. The remaining YAC-hybridization linkages require some further confirmation because of the possibility of hybridization due to repeated sequences and of chimaeric YAC inserts. This confirmation is in some cases provided by several consistent YAC clones spanning the same gap but with different endpoints, and in other cases by comparison with the genetic map, by using cloned genes as probes.

A bottom up map using YACs from the beginning has been undertaken on *Myxococcus xanthus* (Kuspa *et al.*, 1989). This bacterium has a relatively large genome of 9.45 mbp (Chen *et al.*, 1990).

The YACs used by Kuspa *et al.* are of average size 111 kbp. The clones were analyzed by indirect end labelling (probing with vector sequences) of EcoRI partial digests, which does not require separation of the cloned DNA from that of the host. The overlap criterion used requires 4 fragments in common, except when a fragment is >60 kbp in length (on the basis of rarity). This has produced 60 contigs, averaging 150 kbp and 5.1 YACs each from the analysis of 409 YACs. These results are consistent with a Poisson distribution of clone positions, and from this the authors conclude that the cloning is roughly random and probably covers the complete genome. Two things are striking about this project, first that the YACs are relatively small, compared to the potential, and second that even though ordered fragments are being used, the sensitivity of overlap detection is quite low, with θ approximately 0.5. Six (of 18) gene probes tested demonstrated undetected overlaps, so it is quite likely that an organized attempt to link the rest up by hybridization will be quite successful.

Other high-capacity cloning methods are being developed, and methods of analyzing and sorting clones will certainly be improved, so that bottom-up mapping will become less arduous. A cloning system based on bacteriophage P2 (Sternberg, 1990) will allow the cloning of DNAs up to 100 kbp in size with techniques similar to those now very successfully being used for lambda- based vectors. Size limitations on cloning in *Escherischia coli* plasmid vectors are also being relaxed by the use of electroporation, which allows efficient introduction of constructs of up to 100 kbp into the cell (Leonardo and Sedivy, 1990).

Top-down and bottom-up approaches have been presented separately, and until now have been applied as such, but it is clear that the two have complementary strengths, as first pointed out by Olson (1986). Bottom up mapping gives a detailed map and the highly valuable side product of an indexed set of clones. It can be achieved by highly productive "factory work" methods. Top down mapping allows the establishment of long range linkage, which is the most difficult part of bottom up mapping. An integrated strategy is thus the most promising for producing large scale maps economically.

III. Archaebacteria

A. Archaebacteria

The concept of the archaebacteria as a group of prokaryotes forming a third primary lineage, with the bacteria and the eukaryotes, is now quite well established, and extensive efforts have gone into characterizing many aspects of their biochemistry (reviewed by Brown *et al.*, [1989]; Dennis, [1986]; Jones *et al.*, [1987]), much of it with the objective of making conclusions about the evolution of all life by comparison of the three lineages. Little is known to date about the genome structure of archaebacteria, and my goal in beginning this work was to produce a physical map of an archaebacterial genome, and investigate some general aspects of genome structure. Since part of the interest in the work stems from the phylogenetic position of the archaebacteria, which still arouses some controversy, I will begin with a review of phylogenetic analysis.

Linnaeus apparently grouped species in genera, genera in orders, orders into classes purely for convenience, without any stated rationale for such a hierarchy of relatedness (such as descent), saying "there are as many different species as there were different forms created in the beginning by the infinite being", though he did recognize hybrids as a source of new species. Darwin's principle of common descent made sense of hierarchical classification, in that a "Natural System of organisms acquires the significance of a real genealogical tree whose root is formed by those original archaic forms which have long since disappeared" (Haeckel, 1888), pp43-50.

Traditionally, the less morphologically complex and less well understood "lower" forms of life have been lumped together in taxa based on their common lack of certain features. Thus Haeckel, in discarding the traditional animal/ plant dichotomy, represented the family tree of all life as having three main branches-plants, animals, and unicellular forms (Stanier *et al.*, 1963). As more information was accumulated on subcellular structure, prokaryotes were similarly lumped together, for example in the five kingdom scheme (Whittaker, 1969), not because of specific evidence of their relatedness but because of their lack of the features used to differentiate other organisms.

While the problem of deriving a natural taxonomy of the bacteria was widely agreed to be hopeless, and the influential *Bergey's Manual* through the eighth edition (Buchanan and Gibbons, 1974) explicitly took up a policy of arbitrary classification, the molecular clock hypothesis of Zuckerkandl and Pauling (Zuckerkandl and Pauling, 1965) demonstrated the possibility of a quantitative approach. The fundamental idea is that if mutation rates are approximately constant, comparison of homologous genes from different species can be used to estimate how much time has passed since the ancestor of the gene in question existed in the common ancestor of the two species, from the number of accumulated substitutions.

The RNA analysis techniques developed by Sanger made a subset of the sequence of small subunit (SSU.16-18S) rRNA molecules accessible to this kind of analysis. Ribosomal RNA molecules are readily labelled and extracted and are present in all cells, where they clearly have a common function. Woese and Fox (1977) used ribonuclease T1 catalogues to produce a straightforward measure of sequence similarity between pairs of prokaryotes. This produced the unexpected result that a diverse assemblage of prokaryotes formed a group which were by this measure more closely related to each other than to any of the other organisms tested, both bacteria and eukaryotes. Because of its apparently ancient divergence from other forms of life, this group was informally named the archaebacteria.

The archaebacteria are a phenotypically diverse group including sulfur-metabolising thermophiles (some of which are chemolithotrophs), obligately anaerobic methanogens, and aerobic extreme halophiles. They have in common many features, including insensitivity to antibiotics inhibiting the activity of the ribosomes of both eukaryotes and and bacteria, unusual membrane lipids, and the lack of the peptidoglycan typical of the eubacterial cell wall (Brown *et al.*, 1989). Their molecular biology, reviewed later, combines eukaryote like, eubacterial, and uniquely archaebacterial features.

B. Are the archaebacteria a monophyletic group?

Sequence data allow more detailed analysis than did oligonucleotide catalogues, and this field of study is still developing and still controversial. Each method for making inferences about phylogeny from sequences contains assumptions about the characteristics of sequence change with time, which may be justified to different extents in different cases. Especially at the deepest level, the record of phylogeny contained in a set of homologous sequences can be obscured by multiple substitutions at a position. This and differences in the rate of sequence substitution (unclocklike behaviour), as well as the uncertainty inherent in aligning highly diverged sequences, are the fundamental problems which are still being addressed (Olsen, 1987).

While one might expect positions at which changes are selectively neutral or nearly so to behave in the most clock-like manner, in comparing distantly related organisms only the most conserved parts of the rRNA molecule can be aligned with confidence. These are likely to be conserved because of functional constraints, and this has been used to predict what aspects of secondary structure are functionally important (Gutell *et al.*, 1986; Gutell *et al.*, 1985; Woese and Gutell, 1989).

Complete sequencing of twelve archaebacterial 16S and eight 23S rRNA genes has allowed relationships within the archaebacteria to be examined. The group can be divided into two branches, with the halophiles and methanogens on one side, and the sulfur-thermophiles on the other (Woese, 1987). This corresponds with a number of molecular characteristics: the thermophiles differ from the methanogens and halophiles by their much higher level of modification of ribosomal RNAs (Woese *et al.*, 1984), by the subunit structure of their RNA polymerase (Zillig *et al.*, 1989) by the absence of a tRNA^{ala} gene in the 16-23S rRNA gene spacer (Achenbach-Richter and Woese, 1988), and in having 5S rRNA genes unlinked to the other rRNA genes (Kjems *et al.* , 1990). *Thermococcus celer* occupies an intermediate position in the tree, and rooting of the tree using eubacterial or eukaryotic SSU rRNA consensus sequences as outgroups places the root such that *Thermococcus celer* is on the methanogen-halophile side, even though it is an extreme thermophile. The presence of a tRNA^{ala} gene in the ribosomal RNA gene spacer agrees with this assignment (Achenbach-Richter and Woese, 1988), but several other informative features remain to be characterized.

In extreme cases of differing rates of nucleotide substitution. distance matrix methods such as that originally used by Woese (1977) can artificially cluster sequences which have in common only a relatively low rate of substitution. Parsimony analysis of aligned sequences (which assumes that changes are rare and uses this to reconstruct a phylogeny based on the minimum number of changes necessary to convert one sequence to another) can have the related problem of artificially grouping sequences with a relatively high rate of substitution (Felsenstein, 1988; Lake, 1987a; Olsen, 1987).

Evolutionary parsimony (Lake, 1987b) seeks to avoid this problem. Considering four taxa at a time, the method considers transversions among the nucleotide substitutions, on the principle that these are less frequent, and thus more informative in deep branches, and cleverly causes different cases in which subsequent transitions make the result misleading to cancel one another. This makes the analysis immune to the long branch problem of parsimony. Lake's analysis using this method has it that the archaebacteria are not a monophyletic group and that the halophiles and methanogens belong with the eubacteria and that the thermophilic archaebacteria (Eocytes) are the sister group of the eukaryotes. This conclusion does not agree with the many characteristics of archaebacteria which seem to make them a coherent group, some of which are mentioned above, and it has not found favor among archaebacteriologists. One characteristic that does seem to support Lake's hypothesis is the rRNA gene organization.

If the SSU RNA sequence data contains phylogenetic information sufficient to unequivocally determine the topology of the universal tree, one might expect that any reasonable method of analysis, properly applied, should give a similar result. Similarly, different subsets of the data should give similar results. Since no glaring logical fault has been found in evolutionary parsimony, its result must be addressed. A problem with the approach is that it considers a small number of positions. Distance methods consider all positions, parsimony all "cladistically informative" positions, but evolutionary parsimony considers only informative transversions. This makes the conclusions sensitive to a small number of noisy positions. Indeed, the Eocyte tree (Lake, 1989) seems to be a result not of the improved treeing algorithm (Lake, 1987b), but of the choice of aligned positions included in the analysis (Olsen and Woese, 1989). When applied to large subunit ribosomal RNAs, evolutionary parsimony strongly supports the archaebacterial tree, as do other methods (Gouy and Li, 1989). As more and more data become available, the picture will become clearer, but it does seem that the archaebacterial tree is robust. It will be particularly interesting to find out whether trees constructed from different macromolecules are congruent. A tree based on eight archaebacterial 23S rRNA gene sequences, for example, closely resembles that constructed from 16S rRNA gene sequences (Kjems *et al.* , 1990), as does one constructed using those parts of the DNA dependent RNA polymerases which can be aligned (715 amino acids - Pühler *et al.*, 1989a). 5S RNA (Hori and Osawa, 1987) gives broadly the same result, although with such a short sequence resolution is poor at the deepest level.

C. The root of the universal tree

The position of the root of the universal tree is of great interest, but it is difficult to derive because of the lack of an outgroup. Iwabe *et al.*, (1989) propose that this can be overcome by considering pairs of genes descended from an ancestral gene duplicated before the divergence of the three primary kingdoms. Such a pair of genes, if present in representatives of all three primary kingdoms could be used to produce a composite tree whose deepest branching would be between the two genes. That is to say that one might expect the genes A from many species to cluster on one side, and the genes B on the other, and in the middle, representing a divergence which happened within the common ancestor, would be the root. This result is obtained with the alignable parts of EF-Tu and EF-G, ATPase F1 subunits A and B, and initiator and elongator methionine tRNAs (by neighbor-joining, a distance method), and this gives a position for the root between the eubacteria on one side and the archaebacteria and eukaryotes on the other. In the first two cases the bootstrap method (computation of the frequency of obtaining the same result in many analyses of subsets of the data chosen at random [Felsenstein, 1988]) indicates that the trees obtained are very robust. Average branch lengths from the root to extant organisms are nearly equal, indicating similar rates of substitution in these genes in the three lineages. The sequences of several genes of archaebacteria are more similar to their eukaryotic than eubacterial counterparts, including 5S rRNA (Hori and Osawa, 1987), RNA polymerase large subunit (Puhler *et al.*, 1989b), and ribosomal A protein (Itoh, 1988) genes. On the other hand, small and large subunit ribosomal RNA genes of archaebacteria are most like those of eubacteria.

Considering the evidence that the archaebacteria are not specifically related to the eubacteria, and proposing that the similarity in names is a barrier to correct understanding of the relationship. Woese *et al.*, (1990) have proposed a new top level taxon, the domain. The names for the three domains are to be *Eukarya*, *Bacteria* and *Archaea*. This proposal is likely to be adopted, but I will continue to use the name archaebacteria throughout this work, for convenience. It is my opinion that the name *Archaea* fosters an equally serious and more insidious misunderstanding, that these are ancient or primitive organisms. On the contrary, they are alive and highly adapted to a variety of environments in the modern world, and there is no *a priori* reason to believe that their characteristics are more primitive (in the strict sense of being more similar to our ultimate ancestor) than those of the other groups. One of the chief benefits of the emergence of the archaebacterial tree was to begin to dislodge the commonly-held notion that we are evolved from an ancestor much like *Escherichia coli*, through intermediates similar to yeast. This cause is not much advanced if the role of living fossil is thrust upon, say, *Thermoplasma*.

While it is clear from the foregoing that the archaebacteria are not specifically related to the bacteria more than to the eukaryotes, much of this work will be a comparison of archaebacteria with the eubacteria. This is because in order to make instructive comparisons, there must be detectable similarity. Above the level of gene sequence, it is hard to detect similarities between the structure of eukaryotic genomes and those of prokaryotes. Eukaryotic karyotypes, with multiple linear chromosomes and DNA contents often a thousand fold more than those of bacteria, in themselves argue for a vastly different scheme of gene organization. The typically prokaryotic *Bauplan* of the archaebacteria, however, leads one to expect that their genomes can be instructively compared to those of eubacteria.

D.The halobacteria

The halophilic archaebacteria differ from eubacterial halophiles by their high internal ionic strength. Eubacteria may for osmotic reasons accumulate non-ionic solutes such as glycerol or betaines, but not salts. The non-alkaliphilic halobacteria are now divided into three genera based on membrane lipid and nutritional characteristics: *Halobacterium*, *Haloferax*, and *Haloarcula* (Torreblanca *et al.*, 1986). The most studied genus of halobacteria to date is *Halobacterium*. The species *H. cutirubrum*, *H. halobium* and *H. salinarium*, although still commonly known by these names, are more reasonably all classified as strains of one species, *H. salinarium* (Fox *et al.*, 1980). *Halobacterium salinarium* has an unusual retinal-containing light-driven proton pump (Stoeckenius, 1979), bacteriorhodopsin, whose characteristics have been extensively studied. Bacteriorhodopsin is important as a model membrane protein, and characterization has included protein and DNA sequencing, crystal structure at 2.8 Å resolution, complete synthesis of the gene and investigation of the properties of many amino acid substituted derivatives expressed in *Escherichia coli*, reviewed by Khorana, (1988). The bacteriorhodopsin gene was the first archaebacterial protein-coding gene cloned (Dunn *et al.*, 1981), and it allowed an entry into the study of insertion sequences described later.

Haloferax species do not produce the bacteriorhodopsin-containing purple membrane. *Haloferax volcanii* (originally named *Halobacterium volcanii*) is a moderately halophilic species isolated from mud of the Dead Sea (Mullakhanbhai and Larsen, 1975). Unlike *Halobacterium halobium*, it can use simple carbohydrates and will grow on a minimal medium free of amino acids.

E. What is known about archaebacterial genes and genomes

Genome size

Genome sizes of a variety of archaebacteria have been determined by renaturation ($Cot_{0.5}$) analysis and more recently by determination of large restriction fragment sizes using pulsed field gel electrophoresis. (See table 4). The sizes fall within the eubacterial range indicated by the last three entries in the table. The $Cot_{0.5}$ -derived sizes are based on the *Escherichia coli* chromosome as a standard, and have been adjusted using a current estimate of the *Escherichia coli* genome size of 4.7 mbp (Kohara *et al.*, 1987; Smith *et al.*, 1987a).

	mol% G+C	size, mbp	method	reference
Methanobacterium thermoautotrophicum ΔH	51-55	1.5	C_0t	Mitchell *et al.* 1979
Methanobacterium thermoautotrophicum ΔH	45	1.8	C_0t	Klein and Schnorr, 1984
Methanobrevibacter arboriphilicus AZ	27*	3.3	C_0t	Klein and Schnorr, 1984
Methanococcus voltae	30*	3.3	C_0t	Klein and Schnorr, 1984
Methanococcus thermolithotrophicus	35	2 0	C_0t	Klein and Schnorr, 1984
Methanosarcina barkeri	41	2.0	C_0t	Klein and Schnorr, 1984
Halobacterium halobium	66.5*	4.3	C_0t	Klein and Schnorr, 1984
Halococcus morrhuae	64.5	4.3	C_0t	Klein and Schnorr, 1984
Thermplasma acidophilum	46	0.81	C_0t	Searcy and Doyle, 1975
Thermococcus celer	56	1.89	PFG	Noll, 1989
Sulfolobus acidocaldarius	37	2.1	PFG	Yamagishi and Oshima, 1990
Ureaplasma urealyticum	27-28	0.9	PFG	Pyle *et al.* 1988
Escherichia coli	48-52	4.7	cloning	Kohara *et al.* 1987
Myxococcus xanthus	68-71	10	PFG	Chen *et al.* 1990

Table 4. Genome sizes and base compositions

Values marked with an asterisk are overall values for genomes containing two fractions of DNA separable on the basis of composition.

Replication of circular DNAs is simpler than of linear molecules in that no separate provision for maintaining the ends is necessary (Watson, 1972), although there are other requirements, such as for a topoisomerase. On balance, it seems most likely that circular DNA is the ancestral form and that archaebacteria have this characteristic in common with the eubacteria. Although there are linear plasmids in *Streptomyces* and *Borrelia*, in the mitochondria of many filamentous fungi and plants, in the chloroplast of *Chlamydomonas moewusii* and in the cytoplasm of several yeasts (Meinhardt *et al.*, 1990), circular chromosomal DNA is apparently universal among the eubacteria. Circularity has been determined by genetic mapping in *Escherichia coli* (Bachmann, 1990), *Pseudomonas aeruginosa* (Holloway *et al.*, 1990), *Bacillus subtilis* (Piggot, 1990), *Streptomyces coelicolor* (Hopwood, 1965), by physical mapping in *Staphylococcus aureus* (Stahl and Pattee, 1983), *Mycoplasma pneumoniae*(Piggot, 1990; Wenzel and Herrmann, 1988a; Wenzel and Herrmann, 1989) and directly by autoradiography in *Escherichia coli* (Cairns, 1963) and *Bacillus subtilis* (Wake, 1973). *Borrelia burgdorferi* may be an exception - by electrophoretic methods its chromosome appears to be linear (Ferdows and Barbour, 1989); Baril,1989]. Numbers of chromosomes greater than one occur in the eubacteria, though the distinction between a large plasmid and a small chromosome is an arbitrary one. In the case of *Rhodobacter sphaeroides* (Suwanto and Kaplan, 1989) two circular DNAs of different size each bear a ribosomal RNA operon, giving the authors the confidence to call the smaller molecule, which might otherwise have been known as a plasmid, a second chromosome.

The expectation that archaebacterial chromosomes would be circular has now been borne out in two cases mentioned above. In the case of *Thermococcus celer* (Noll, 1989), restriction mapping of the entire chromosome with three rare-cutting enzymes clearly demonstrated circularity. (The formal possibility exists that such a circular map could be obtained from a linear, randomly circularly permuted chromosome. This also applies to circular genetic maps (Hopwood and Kieser, 1990)). *Sulfolobus solfataricus* (Yamagishi and Oshima, 1990) DNA gives two fragments when digested with Not1. Cloned DNAs containing Not1 sites indicates that each end of one fragment links an end of the other, so barring accidental ligation together of unrelated Not1 arms in the cloning (and the existence of further, undetected Not1 fragments), this organism also has a circular chromosome.

Ribosomal RNA

Archaebacterial ribosomal RNAs resemble their eubacterial counterparts in size, number, and nucleotide sequence. Those of the halophiles and methanogens are also transcribed from genes organized in operons similar to those of eubacteria, in the order 5'-16S-tRNA (ala where known)-23S-5S. In halobacteria, there is a tRNA cys gene downstream of the 5S gene. Several methanogenic species have additional unlinked 5S rRNA genes. In the sulfur-dependent branch, there is no tRNA in the 16S-23S spacer and 5S RNA genes are unlinked (Eggen *et al.*, 1990; Kjems *et al.*, 1987b), which might be considered an eukaryote- like arrangement.

Although the unprocessed precursor has not been detected, archaebacterial rRNA operons have flanking and spacer regions which can base pair to form structures similar to those important in eubacterial rRNA transcript processing (Achenbach-Richter and Woese, 1988; Chant and Dennis, 1986; Chant *et al.*, 1986; Hui and Dennis, 1985; Kjems *et al.*, 1987a; Kjems *et al.*, 1987b; Larsen *et al.*, 1986; Lechner *et al.*, 1985; Mankin and Kagramanova, 1986). In *Halobacterium cutirubrum*, there is a specific endonucleolytic cleavage site in each stem (Dennis, 1985), the characteristic bulge-helix-bulge structure of which is also found in *Thermoproteus tenax* (Kjems *et al.*, 1987a), *Desulfurococcus mobilis* (Kjems *et al.*, 1987a), *Thermofilum pendens* (Kjems *et al.* , 1990), and *Methanobacterium thermoautotrophicum* (Østergaard *et al.*, 1987). This processing occurs at such a rate that a full length precursor is probably not formed.

In the nuclei of eukaryotes, ribosomal RNA genes are linked and transcribed in the order 5'-18S-5.8S-28S-3'. Many copies of this unit are tandemly disposed. The 5S genes are typically unlinked, and in the sporadic cases where they are included in the rDNA, some are on the same strand, some on the other. In all cases the 5S RNA is produced by RNA polymerase III. while the 18S-5.8S-23S precursor is an RNA polymerase I product.

The number of rRNA operons in archaebacteria is typically one. though *H.volcanii*. *Mb. thermoautotrophicum* and *Mt. fervidus* have two and *Mb. vannielii* has four (reviewed by (Brown *et al*.. 1989)). On the basis of Southern transfers of pulsed field gels. Sanz *et al*.. (1988) report that the halophilic archaebacteria have from one to four rRNA operons: *Haloarcula californiae*. four: *Haloferax gibbonsii*. four: *Halobacterium halobium* NCMB 777.three: *Halobacterium marismortui*. three: *Halococcus morrhuae*. two: and *Halobacterium salinarium*. 1. This result has to be treated as preliminary without any restriction mapping of the rRNA operons themselves. Eubacteria have from one to eleven copies of the ribosomal RNA genes for example *Bacillus subtilis*. eleven (Widom *et al*.. 1988): *Escherichia coli*. seven (Bachmann. 1990): *Streptomyces coelicolor*. six (Hara *et al*.. 1983): *Haemophilus influenzae*. six (Lee *et al*.. 1989b): *Pseudomonas aeruginosa*. four (Romling *et al*..): *Mycoplasma capricolum*. two (Glaser *et al*.. 1984) and *Mycoplasma pneumoniae*. one (Wenzel and Herrmann. 1988a).

Introns

Introns have been found in the tRNA^{trp} genes of *Haloferax volcanii* (Daniels *et al*.. 1985a). *Haloferax mediterranei*. and *Halobacterium cutirubrum* (Daniels *et al*.. 1986). in the elongator methionine tRNAs of *Haloferax volcanii* (Datta *et al.*. 1989) and *Desulfurococcus mobilis* . in five *Sulfolobus solfataricus* tRNAs of six sequenced (Kaine. 1987: Kaine *et al.*. 1983). including elongator methionine. and three *Thermoproteus tenax* tRNAs (Wich *et al.*. 1987). These short introns are placed (except for two of those in *T. tenax*) in exactly the same position (one nucleotide 3' of the anticodon) as the introns of eukaryotic nuclear tRNA genes. but the mechanism of splicing is different (Perlman *et al.*. 1990: Thompson and Daniels. 1988). A similar intron which also has some sequence features in common with class I introns has been found in the 23S rRNA gene of *Desulfurococcus mobilis* (Kjems and Garrett. 1985). These introns have in common the possibility of forming a characteristic base paired stem with the splice sites in staggered bulges. These sites resemble the exonuclease III sites involved in 16 and 23S rRNA processing. and cleavage could be achieved by the same enzyme (Thompson and Daniels. 1988). Cleavage at splice sites can be achieved *in vitro* and the use of various synthetic substrates has demonstrated the importance of the bulge-helix bulge structure (Kjems and Garrett. 1988: Kjems *et al.*. 1989: Thompson and Daniels. 1988).

Transcription

The single known DNA-dependent RNA polymerase of archaebacteria is probably used for all purposes. like its eubacterial homologue, but the complex subunit structure and immunological cross-reactivity are more eukaryote-like (Huet *et al.*, 1983). The RNA polymerases of *Sulfolobus acidocaldarius* and *Halobacterium halobium* resemble most the RNA polymerase II of eukaryotes, with subunits A and C corresponding to subunit A of the eukaryotic polymerase (eubacterial subunit β') and subunit B (*Sulfolobus*) or B'+B" (*Halobacterium*) corresponding to subunit B (eubacterial β) (Zillig *et al.*, 1988). Correspondingly, archaebacterial promoters are similar to eukaryotic pol II promoters, with a TATA-like "box A" sequence at -26bp and another "box B" motif surrounding the transcription start (Zillig *et al.*, 1988). These motifs, detected by sequence comparison, have now been functionally dissected using *in vitro* transcription of *Sulfolobus* rRNAs (Hudepohl *et al.*, 1990).

Translation

In contrast to transcription, most aspects of translation in archaebacteria are eubacterial in character. The ribosomal RNAs are, unlike several protein-coding genes, most similar to their eubacterial counterparts. Shine-Dalgarno complementarity can be found in translated archaebacterial genes (Zillig *et al.*, 1988) Messages can be polycistronic (*e.g. rpl*1e,10e,12e in *H. cutirubrum*), and translation can begin at UUG and GUG as well as AUG (Shimmin *et al.*, 1989). The standard genetic code is used. In vitro transcription of synthetic polynucleotides by a *Halobacterium cutirubrum* extract indicated that this was the case (Bayley *et al.*, 1978) and it was further determined by comparison of the amino acid sequence of bacteriorhodopsin with the DNA sequence of its gene (Dunn *et al.*, 1981), although the gene contains no cys, his, or AGR arginine codons. Still more confirmation was had from the characterization of 41 tRNAs from *Haloferax volcanii* (Gupta, 1984; Gupta, 1986), which can be charged with amino acids corresponding to their anticodons, except for a Gln tRNA which can be charged with glutamic acid and is apparently amidated *in situ*, and an asparagine tRNA which could not be charged *in vitro*.

The sequencing of 41 *Haloferax volcanii* tRNAs, as well as several other archaebacterial tRNAs (reviewed by (Gupta, 1985)) has indicated that these fit the generalized structure known chiefly from *Escherichia coli* and yeast tRNAs, but in the sequence and pattern of modification there is a mixture of eukaryotic, eubacterial and unique features, including that, like in eukaryotes, leucine and serine but not tyrosine tRNAs are of class II (large extra arm). Cysteine tRNA has an unusual six base extra arm. The archaebacterial tRNAs examined to date lack ribothymidine and 7-methyl guanosine, and dihydrouridine has so far been found only in *Methanosarcina barkeri*. The D and T of "D-loop" and "TΨC loop" are thus not universal features. The initiator methionyl tRNA of *Halobacterium cutirubrum* is not n-formylated as its eubacterial counterpart is (Bayley *et al.*, 1978). The initiator methionine tRNA of *Haloferax volcanii* is the only tRNA known to have a 5'-triphosphate. It is likely that this is the 5'-end of the primary transcript.

Ten *Haloferax volcanii* tRNA genes have been cloned and sequenced (Daniels *et al.*, 1986; Datta *et al.*, 1989, R.Gupta pers. comm.) as have a number from other archaebacteria. None has the 3' CCA encoded; these nucleotides are added after transcription, as they generally are in eukaryotes.

7S RNA

Archaebacteria contain a stable 7S RNA species of unknown function, which is not associated with the ribosome. It resembles the 7SL RNA of eukaryotes (Daniels *et al.*, 1985b) and the 4.5S (*Escherichia coli*) or sc (*Bacillus subtilis*) RNAs of eubacteria (Struck *et al.*, 1988) in structure. The 7SL RNA forms part of the signal recognition particle, involved in protein translocation. The function of the 4.5S RNA is not exactly known, but it is known to be essential (Brown, 1987).

Plasmids

Circular plasmids are widely distributed among the archaebacteria, having been found in the thermoacidophiles *Sulfolobus* B12 (Yeats *et al.*, 1982) and *Desulfurococcus ambivalens* (Zillig *et al.*, 1985), in methanogens of the genera *Methanococcus* (Wood *et al.*, 1985). *Methanolobus* (Thomm *et al*., 1983) and *Methanothermus* (Meile *et al*., 1983). Plasmids have been detected in most of the halobacterial species studied, and are of a wide range of sizes. Most studied are the plasmids of *Halobacterium halobium*, in which are found the only plasmid-borne genes characterized to date. pHH1 (Pfeifer *et al*., 1981a) bears one (DasSarma *et al*., 1987) of the two gas vacuole genes in this organism (Horne *et al*., 1988). The phage ΦH1 is a plasmid in its lysogenic form, so to this extent it is also a plasmid with known genetic functionality (Schnabel and Zillig, 1984). Different strains of *Halobacterium halobium* contain different variants of pHH1. Rearranged or deleted versions of the plasmid are readily isolated (Pfeifer and Blaseio, 1989). Several of the purple membrane-containing extreme halophiles also contain small, very high copy number plasmids (Hackett and DasSarma, 1989; Hackett *et al*., 1990). The supercoiled fraction obtained by CsCl-ethidium bromide equilibrium ultracentrifugation of *Halobacterium halobium* DNA contains, in addition to the characterized plasmids, a heterogeneous population of "minor ccc DNAs". It is supposed that these arise by intramolecular recombination in the chromosome and it is not known whether they are replicated (Ebert and Goebel, 1985; Pfeifer *et al*., 1982).

Using an *in situ* lysis method, Gutiérrez *et al*. (Gutiérrez *et al*., 1986) observed large plasmids in most halophilic strains tested, ranging in size up to a very approximate 450 kbp, including two plasmids of about 180 and 442 kbp in *Haloferax volcanii* .

Restriction

Type II restriction-modification systems similar to those in eubacteria have been characterized from *Thermoplasma acidophilum* (McConnell *et al.*, 1978), *Methanococcus aeolicus* (Schmid *et al.*, 1984) and *Sulfolobus acidocaldarius* (Prangishvilli *et al.*, 1985). Uncharacterized restriction systems are known to be present in *Halobacterium halobium*, and *Haloferax volcanii* from their resistance to infection and transfection/transformation (Charlebois *et al.*, 1987b; Cline and Doolittle, 1987).

Satellite fraction

As first reported by Joshi *et al.* (1963) in *Halobacterium salinarium* and *H. cutirubrum*, the DNA of halobacteria can be separated into two fractions on the basis of composition. This observation was confirmed and extended by Moore and McCarthy (1969b) who demonstrated a minor more AT-rich fraction in *Halobacterium halobium*, *Halobacterium salinarium*, *H. cutirubrum*, and *Halococcus morrhuae*, while taking pains to exclude the possibility that this was due to a contaminant. Renaturation studies (Moore and McCarthy, 1969a) indicated that the minor fraction was neither a simple repeated sequence nor multiple copies of a small episomal element. The minor fraction is approximately 10 mol % lower in G+C content than the bulk of the DNA and varies from about 10 to roughly 30% of the total in different species. It may be absent from *Halobacterium trapanicum* (Pfeifer *et al.*. 1982). Although Moore and McCarthy concluded that the satellite fraction was not a plasmid, it has since been found that halobacteria do contain plasmids, and that these are more AT-rich than the bulk of the DNA (Weidinger *et al.*, 1979), but also that at least some satellite DNA is interspersed in the genome (Pfeifer and Betlach, 1985).

Repeated sequences and instability

Halobacteria, especially the extreme halophiles, show a variety of visibly different phenotypes when plated from a single colony. Frequencies of 10^{-2} for loss of the gas vacuole and 10^{-4} for loss of bacterioruberin and bacteriorhodopsin have been measured in *Halobacterium halobium* (Pfeifer *et al.*, 1981a). Vacuole loss could be correlated with loss of (Simon, 1978), insertions into (Weidinger *et al.*, 1979), or insertions and rearrangements (Pfeifer *et al.*, 1981a) of the large plasmid pHH1, from which the gas-vacuole gene was eventually cloned (DasSarma *et al.*, 1987). Complex and varied rearrangements of the plasmid also accompanied bacterioruberin and bacteriorhodopsin mutations(Pfeifer *et al.*, 1981b), however and it was clear that an extraordinarily active mechanism for DNA rearrangement was present. Insertion, inversion and deletion variants of the *Halobacterium halobium* phage ΦH1 could also be readily observed (Schnabel *et al.*, 1982).

Supposing that the observed instability might be due to transposable elements, Sapienza and Doolittle (1982a), demonstrated by hybridization that *Halobacterium halobium* contains repeated sequences and then screened randomly cloned EcoRI and EcoRI/BamHI fragments of about 3 kbp length for hybridization to multiple EcoRI fragments(Sapienza and Doolittle, 1982b). 31 of 35 EcoRI/BamHI fragments from *Halobacterium halobium* R1 (a gas vacuole mutant strain) and 27 of 28 EcoRI fragments from NRC1 (wild type) contained repeated sequences. A similar experiment using PstI fragments gave only unique sequences, leading to the conclusion that this organism contains many repeated sequences, perhaps 50 families of 2-20 copies each, which are clustered in that part of the genome frequently cut by EcoRI, but not PstI ("Pst-poor regions"). Some of the repeated elements hybridized with repeated elements in the genome of *Haloferax volcanii*. Comparison of the patterns revealed by hybridization of repeated sequences to separate but not visibly different single colony isolates of *Halobacterium halobium* indicated that these elements are highly mobile. A quantitative study (Sapienza *et al.*, 1982) produced an estimate of >4x10^{-3} per family per generation for changes detectable on Southern transfers probed with repeated sequences. No changes were seen in the two unique sequences tested. The analysis also suggested that the events occur in bursts, where a single isolate was affected by events involving several repeat families.

The most likely explanation for this instability is that the repeated sequences are insertion sequences similar to those which are widely distributed in the eubacteria and in eukaryotes. The first of these characterized was an insertion in the bacteriorhodopsin gene of *Halobacterium halobium* (Simsek *et al.*, 1982). The element is 1118 bp in length, with an 8 bp terminal inverted repeat. Eight base pairs of the target are duplicated The target sequence is itself flanked by a 9 bp inverted repeat which is similar to the terminal repeat. Within the element are found an additional seven inverted repeats of over 8 bp in length There are two overlapping reading frames on opposite strands, of which one could be shown to correspond to a transcript by Northern analysis. The sequence begins with T-G and ends in C-A as do most of the eukaryotic and eubacterial insertion sequences. Twenty-one independent insertions of ISH1 into the bacteriorhodopsin gene have been characterized, all at the same position, in either orientation.

ISH2, the smallest insertion sequence characterized to date, (DasSarma *et al.*, 1983) was also isolated as an insertion in the bacteriorhodopsin gene. It is 520 bp in length and has a 19 bp terminal inverted repeat. Unlike ISH1, the target site duplication varies in length. It has little insertion specificity and has been found at several sites in the bacteriorhodopsin gene and one 102 bp upstream of the gene, which also inactivates the gene.

Five other, less commonly inserted ISH elements have been found in this way: ISH23, ISH24 (Pfeifer *et al.*, 1984), ISH26, 27, and 28 (Pfeifer *et al.*, 1983). ISH23 is very similar in sequence to ISH50 (Xu and Doolittle, 1983), which was isolated from the 50 kbp plasmid in *Halobacterium halobium* strain R1. ISH27 resembles in sequence ISH51 of *Haloferax volcanii* (Hofman *et al.*, 1986; Pfeifer and Blaseio, 1989).

ISH51 is a large family of degenerate repeated sequences in the genome of *Haloferax volcanii*. Different copies (of which there are 20-30) have sequence similarity of 85%, on average, and are present in both high and low-%GC regions of the genome. A spontaneous insertion of ISH51 into the plasmid pHv2 has been detected (Lam and Doolittle, 1989)

The fact that insertions occurring up to 1.4 kbp upstream of the *bop* gene affect its expression suggested that there might be another gene involved in bacteriorhodopsin expression in this region. Sequencing of this region (Pfeifer *et al.*, 1984) revealed a 1118 bp open reading frame in the opposite orientation to the bacteriorhodpsin (*bop*) gene, which is expressed, as judged by Northern transfer analysis, and has been named bacterioopsin related protein (*brp* gene).

Insertions into the DNA of the phage ΦH are also readily isolated (Schnabel, 1984; Schnabel *et al.*, 1982)]. These are due to a second copy of the resident element ISH1.8, of which there are also two copies in the genome of the host. This element does not have terminal inverted repeats, nor does it apparently cause a duplication of the target sequence.

Starting with a chromosomal copy of ISH1, Pfeifer and Betlach, (1985) carried out a cosmid walk in both directions and succeeded in cloning an entire "island" of FII DNA, that is to say a region of more AT-rich sequence embedded in the chromosome. A total of 160 kbp was cloned, of which 70 kbp was FII, as judged by probing isolated fragments to fractionated genomic DNA. One copy of ISH1, two each of ISH2 and 26 and ten or more other, uncharacterized repeated sequences were found in the cloned region, mostly in the FII part.

F. Why choose this species?

While each group of archaebacteria has its own points of interest, the halophilic branch is the most attractive for the development of a genetic system because these organisms are easily manipulated mesophilic aerobes. *H.volcanii* is particularly so because it is capable of growing on a simple minimal medium (Mevarech and Werczberger, 1985), so that there is the possibility of isolating auxotrophic strains and thus identifying genes for many biosynthetic functions. Methods for mating (Mevarech and Werczberger, 1985) and protoplast fusion (Rosenshine *et al.*, 1989) of *Haloferax volcanii* have been developed. The transfection method developed for *Halobacterium halobium* (Cline and Doolittle, 1987) has been extended to *Haloferax volcanii* and used for plasmid transformation

(Charlebois *et al.*, 1987a). Genomic DNA can be used to transform auxotrophic strains (Cline *et al.*, 1989a), and a shuttle vector based on pHv2 and a sequence conferring resistance to mevinolin has been described (Lam and Doolittle, 1989). Cosmids from the present work (Charlebois *et al.*, 1989b) have already been used in pools to transform histidine (Conover and Doolittle, 1990) and tryptophan (Lam *et al.*, 1990) auxotrophs to prototrophy, and thus isolate genes in the pathways in question.

Methods and Materials

I. Strains and culture conditions

Haloferax volcanii DS2 was obtained from the German Collection of Microorganisms (DSM 3757) and the first transfer from the original 3 ml liquid culture was maintained at room temperature in the dark as a stock. From this the strain was entered in the laboratory strain collection as WFD18. *Haloferax volcanii* WFD11 was from the lab strain collection (Charlebois *et al.*, 1987a). It was derived from WFD7, also a derivative of DS2, which was obtained from C. Woese in 1984.

Haloferax mediterranei was from the American Type Culture Collection (ATCC 33500, WFD40). Both *H.volcanii* and *H.mediterranei* were grown on a medium containing, per liter, 125 g NaCl, 45 g $MgCl_2 \cdot 6H_2O$, 10 g $MgSO_2 \cdot 7H_2O$, 10 g KCl, 1.34 g $CaCl_2 \cdot 2H_2O$, 3 g Bacto yeast extract and 5 g Bacto tryptone (Daniels *et al.*, 1984). The salts were autoclaved separately from the yeast extract and tryptone. If agar was used (18 g) it was also autoclaved separately in water.

Halobacterium species GRB (WFD34) was obtained from W. Goebel. *Halobacterium halobium* NRC1 and *H. halobium* R1 were from the laboratory collection (WFD8 and 13) These were grown on medium containing, per liter, 250 g NaCl, 20 g $MgSO_2 \cdot 7H_2O$, 3 g trisodium citrate, 2 g $CaCl_2$ and yeast extract, tryptone and agar as for the *Haloferax* species.

Escherichia coli strains used for cloning are listed and described in table 5. For preparation of plasmids and cosmids, *Escherichia coli* was grown in 2YT (16 g tryptone, 10 g yeast extract, 5 g NaCl per liter of tap water) or terrific broth (Tartof and Hobbs, 1987), which contains, per liter, 12 g Bacto tryptone, 24 g Bacto yeast extract, 4 ml glycerol, 2.31 g KH_2PO_4, and 12.5 g K_2HPO_4. The phosphates are sterilised separately in one tenth of the volume. For preparation of bacteriophage lambda DNA, I used modified L-broth (Helms *et al.*, 1987; Helms *et al.*, 1985), containing, per liter, 10 g tryptone, 5 g yeast extract, 5 g NaCl and 0.25 g $MgCl_2$, brought to pH 7.5 with KOH. For broth for plating cultures and for top agar, I omitted the yeast extract and added 2 g per liter of maltose.

II. Reagents and general methods

Chemicals were reagent grade, with some exceptions noted below. Buffers were prepared wherever possible by mixing calculated amounts of acid and base, rather than by titration. Reaction mix concentrates for digestion or labelling were prepared from stock solutions which had been sterilized by autoclaving where possible, and otherwise by membrane filtration.

Phenol for extraction was equilibrated with water or buffer as indicated, and used without other additives. Chloroform was also used without additives. Absolute ethanol (2 volumes) was used for precipitation of DNA, after the solution had been brought to 2.5 M ammonium acetate using a 10 M stock.

strain	genotype	use	ref	source
DH5α	F$^-$Δ*(lacZYA-argF) hsdR17* (r_{k^-}. m_{k^+}). *endA1. recA1. gyrA96. deoR. supE44*F80d*lacZ*Δ*M15*	general cloning host	1	BRL
ED8767	*supE.supF. hsdS* (r_{k^-} m_{k^-}). *lacY.recA56*	host for cosmids	2	P.F.R. Little
GM48	*thr. leu. thi. lacY. galK. galT. ara. tonA. tsx.dam. dcm. supE44*	non-methylating host for lambda	3	K. Conover
JM101	Δ*lacpro. supE. thi. F' traD36. proAB. lac IqZ*Δ*M15*	host for m13	4	lab coll'n
Q359	*supE. hsd R*(r_{k^-}.m_{k^+}). (P2)	host for lambda cloning	5	lab coll'n

Table 5. *Escherichia coli* strains used for cloning. References are: 1. BRL research products: 2. (Grosveld *et al*., 1981): 3. (Marinus. 1975): 4. (Messing. 1983): 5. (Karn *et al*., 1980)

III. DNA preparation and labelling

A. Halobacterial genomic DNA

A culture, usually in the late logarthmic phase of growth, was pelleted (5 min, 4,000 x g) and resuspended in 5% or more of the culture volume of the supernatant fluid or of half-strength medium salts. Alternatively, a lawn of cells was washed from a plate with 5-10 ml of medium salts. The cells quickly lysed after dilution of the suspension with an equal volume or more of 50 mM EDTA, 50 mM Tris-HCl (pH 8.0), 0.5% sarkosyl (sodium n-lauroyl sarkosine). After an optional treatment for 30-60 minutes at 55° C with 50 µg/ml proteinase K, the preparation was extracted with phenol saturated with 1 M Tris-HCl pH 8.0. When a clear interface was obtained, the solution was extracted with chloroform, and finally the DNA was either spooled from the solution after the addition of 0.6 volumes of isopropanol, rinsed with ethanol and slowly redissolved in TE or extensively dialysed against TE. TE is 10 mM Tris-HCl, 1 mM sodium EDTA, pH 8.0.

B. Preparation of intact DNA

DNA isolated by conventional techniques is not of sufficiently high molecular weight for pulsed-field-gel analysis. Relatively simple techniques based on immobilizing cells in agarose before lysis allow protection of the DNA from mechanical shearing. Detergent, a chelating agent,and an alkaline pH serve to protect the DNA from nucleases and acid depurination.

I prepared halobacterial DNA in either blocks or beads of agarose by preparing a suspension of cells as described above, being especially careful to avoid any lysis before encapsulation of the cells. The suspension was warmed to 55°C and mixed with an equal volume of 10 g/l Seaplaque agarose (FMC) in 1 M NaCl at the same temperature. For agarose blocks, I immediately drew the suspension into a length of silicone tubing (2 mm internal diameter, which has a volume of about 30 µl/cm). When set, the agarose could be cut with a cover slip into convenient pieces. To prepare microbeads, I added a volume of warmed light mineral oil (Fisher) greater than that of the agarose plus cells, such that a head space of 1 cm was left in the tube, and then shook the tube by hand back and forth lengthwise through a 45 degree arc until an oil-continuous emulsion with aqueous globules about 100 µm in diameter was obtained. This can easily be monitored by eye, since 100 µm is about the smallest size of globule which one can readily resolve. The mixture was then chilled on ice with frequent shaking until the agarose was set. After 0.5 h on ice, the bulk of the oil could be removed with a pipette, and the beads were rinsed several times with 1 M NaCl. To do this, the chilled beads were pelleted by a centrifugation of 4 min in a microcentrifuge or in a SS-34 rotor (Sorvall) for the time it takes to reach 10 krpm. Cells were lysed by resuspension of the beads or blocks in 10 volumes of NDS (0.5 M disodium EDTA, 1 M Tris base, 10 g/l sodium N-lauroyl sarcosinate) and warming to 55°C.

In order to remove the bulk of the proteins as quickly as possible, DNA preparations in agarose are usually treated with 100 mg/l or more of proteinase K at 55° C (Schwartz and Cantor, 1984). This inactivates nucleases, and helps remove the many proteins in the typical cell which are insoluble in NDS, which is 1 M in Na^+. Halobacterial cells are not rich in nucleases, and their proteins do not precipitate in NDS, and good results can be obtained without the use of proteinase K. If proteinase K is used, measures must be taken to inactivate any residue, so that it will not interfere with eventual restriction enzyme digestion. Whether proteinase treatment was used or not, bead preparations were incubated overnight at 50-55° C and then washed with TE + 1 M NaCl until no sign of detergent foam could be seen (five or more washes over four or more hours at 0-4 °C). Bead preparations can be extracted several times with chloroform at this stage to inactivate residual proteinasc K. The preparation is then ready to digest after three or more washes in TE. For block preparations, more time must be allowed for diffusion, with the initial incubation in NDS requiring at least 24 h. Inactivation of proteinase K in blocks or beads can also be achieved by incubation in 1 mM PMSF in TE for 30 min at 0°C (Smith *et al.*, 1987a).

C. Digestion of block and bead preparations

The success of digestion of DNA in agarose depends most critically on the quality of agarose used. Seaplaque (FMC) low-melting agarose has been consistently satisfactory over several lots. Seakem

GTG regular-melting agarose has also given good results and is easier to handle, but does not leave the option of melting the preparation at a later stage. Restriction enzymes vary in their sensitivity to inhibition by agarose and cell debris, so for each enzyme used, various buffers, additives, and enzyme concentrations were tested.

Most enzymes gave a good result under the following conditions: Agarose beads or blocks, equilibrated with TE, were incubated with an equal volume of double strength restriction buffer for 0.25 h at room temperature, and then chilled on ice before addition of about ten units of enzyme per lane. After a further 0.5 h at ice temperature, the mixture was incubated for two or more hours at the recommended temperature. Most restriction digests were done in KGB (McClelland *et al.*, 1988), which contains 100 mM potassium glutamate, 25 mM Tris acetate pH 7.6, 10 mM magnesium acetate, and is used at single or double strength depending on the salt requirement of the enzyme. For some enzymes, including XbaI, SpeI, and BglII, bovine serum albumin (100 mg/l) and 2-mercaptoethanol (1 mM) were added.

D. Plasmids, cosmids, lambda, M13

Plasmid and cosmid DNAs from *Escherichia coli* were prepared by the following version of the Birnboim and Doly (1979) method. Cultures were grown with selection (ampicillin or kanamycin at 50 mg/l) to stationary phase. After centrifugation (20 s in a microfuge (11000 x g) or 5000 rpm, 5 min Sorvall SS34 rotor), the pellet was resuspended at no more than 50 g wet weight per liter of autoclaved 25 mM Tris-HCl, 50 mM EDTA, 10 g/l glucose pH 8.0 (for example, cells from 750 ml 2YT or 200 ml terrific broth resuspended in 20 ml). Twice the resuspension volume of 0.2 M NaOH, 10 g/l SDS was added at room temperature and mixed thoroughly by inversion. Once lysis was complete, 1.5 times the resuspension volume of 3 M potassium acetate, 2 M acetic acid was added with vigorous mixing by shaking and vortexing. After chilling on ice (about 10 min depending on the volume) the precipitate was removed by centrifugation (5 min in a microfuge or 10 min at 10,000 rpm, Sorvall SS34 or HB4 rotor). In most cases this supernatant fluid was then extracted twice with an equal volume each time of water-saturated phenol. This is especially useful for obtaining intact plasmid DNA from hosts with wild type *end*A, and also reduces chromosomal DNA contamination. Nucleic acids were then precipitated with 0.54 volume of isopropanol, resuspended in TE, brought to 2.5 M ammonium acetate and chilled to precipitate large RNAs. After centrifugation, the DNA was precipitated again from the supernatant with isopropanol.

For landmark analysis, colonies were picked into 5 ml of terrific broth containing 30 mg/l of kanamycin in a 50 ml disposable centrifuge tube, and incubated shaking on an angle overnight or until saturated. A sample of each culture was used to prepare a 0.2 ml stock in Titertek tubes (1.5 ml polypropylene tubes in racks of the same dimensions as standard microtiter dishes) containing 15% glycerol for frozen storage. The remaining culture was centrifuged, and the pellet suspended in the Tris-EDTA-glucose solution and stored frozen. DNA preparation was done as above, except that the potassium acetate was replaced by 7.5 M ammonium acetate for the neutralization stage, which saves a step by precipitating large RNAs at this stage. The preparations were precipitated once with isopropanol. The most recent version of the method uses custom-built carriers to centrifuge entire racks of Titertek tubes.

E. Labelling methods

Nick translation

I used nick- translation (Rigby *et al.*, 1977) for most routine labelling of double stranded DNA for hybridization, except where fragments were gel purified, in which case I used random priming. A typical reaction contained 200 ng of DNA, 20 µCi [α-^{32}P]dATP at approximately 3000 Ci/mmol, 5 U polymerase I (BMC) and 100 pg DNAse I (Worthington DPFF) in 10 µl of 20 mM each dCTP, dTTP and dGTP, 5 mM DTT, 10 mM $MgCl_2$, 10 mM Tris-HCl (pH 7.5) and 0.1 g/l BSA. Incorporation was monitored by a DE81 assay (Maniatis et al. 1982). I spotted an unknown volume, nominally 0.1 µl, of the reaction on a 5 mm square piece of DE81 paper (Whatman), and measured the total radioactivity of the sample by Cerenkov counting of the dry filter. This gives a counting efficiency approximately the same as is obtained by counting in 10 ml of water if the filter is flat on the bottom of the vial. I then washed the filters (up to ten together) five times in 10 ml per wash of 0.5 M Na_2HPO_4, rinsed in water and then in 95% ethanol, dried and recounted. Background (no filter) counts were typically less than 50 cpm and were not subtracted. Percent incorporation is then the quantity (cpm after washing)(100)/(total counts). Zero-time or no-enzyme controls were less than 1%, while a typical nick-translation gave 50% incorporation after 20 min at room temperature, for a specific activity of 50 Ci/g. I occasionally optimized the DNAse concentration for a certain length of incubation and then measured the incorporation over the course of time, in order to obtain probes of high specific activity. The DNAse stock was diluted in nick translation buffer with 500 g/l glycerol and stored at -20°C .

Random priming

DNA fragments were labelled by random priming in the presence of agarose at room temperature in 50 mM Tris-HCl (pH 8.0), 10 mM $MgCl_2$, 5 mM dithiothreitol, 100 mg/ml BSA, 50 mM each dCTP and α-thio dTTP with 10 mCi each [α-^{32}P]dATP and dGTP and 40 U/ml Klenow polymerase, after boiling with 80 mg/ml random hexamers. These conditions differ from those of Feinberg and Vogelstein (1983) in that incorporation of an α-thio nucleotide is used instead of low pH to suppress the 3' exonuclease activity of the polymerase.

In later work, I labelled agarose-containing DNA samples by nick translation as described above, except that the agarose has to be melted before the components are assembled. A nick translation reaction containing 0.5% agarose proceeds at about half the rate of a comparable one without agarose.

Transcription using Sp6 or T7 RNA polymerase

This method was used to produce end-specific probes from cosmid clones in Lorist M. It has the advantage of allowing such probes to be made in the absence of map data, and of producing probes of high specific activity. Templates were standard alkaline preparations which had not been treated with RNase. A typical reaction contained 1 µl of cosmid stock (ca 100 ng), 20 µCi of [α-^{32}P] GTP at 3000 Ci/mmol (Amersham), 20 U of Sp6 or T7 polymerase (DuPont, NEN) and 10 U RNASin (Promega) in a volume of 10 ml of 40 mM Tris-HCl (pH 7.6), 60 mM spermidine and 0.4 mM each ATP, CTP and TTP. This was incubated for 2 h at 42^{0} C, during which a variable fraction (as much as 80%) of the label was incorporated, as measured by the DE81 assay.

The length of the transcripts is limited by the low GTP concentration (ca 0.7 µM), and the results of genomic probings indicate that the bulk of the probe is under 1 kb in length.

IV. RNA preparation and labelling

A. Isolation

For isolation of soluble RNAs I used a procedure based on that of Von Ehrenstein (1967). A one liter culture of *Haloferax volcanii* was harvested at an OD_{550} of 1.7 and resuspended in 200 mM NaCl, 50 mM Na-Hepes pH 7.0 to a volume of 16 ml. This was quickly homogenized in a motorized Dounce homogenizer with 2 ml of 200 g/l SDS and 2 ml of 0.5 M Na-EDTA pH 7.0, and again after the addition of 20 ml of phenol equilibrated with 10 mM Na-Hepes pH 7.0. After centrifugation, the supernatant fluid was extracted once with chloroform. The aqueous phase was weighed and 0.54 ml of isopropanol added for each gram, yielding a large precipitate of DNA and high molecular weight RNAs, which was removed by centrifugation.

The isopropanol supernatant, containing the crude sRNA, was mixed with about 1.5 times its volume (40 ml) of absolute ethanol and chilled to -20°C. After centrifugation, the well drained pellet was suspended in 5 ml of 1 mM NaCl, 10 mM Na-Hepes pH 7.0 and chilled on ice. After centrifugation to remove the small pellet of insoluble material, the sRNA was precipitated from the supernatant fluid with the addition of 5 ml of isopropanol. After centrifugation, the pellet was dissolved in 0.5 ml of 0.2 M NaCl, 10 mM Na-Hepes pH 7.0, phenol extracted and precipitated with 1 volume of isopropanol. The pellet was dissolved in 100 ml of 0.5 M Tris-HCl (pH 8.8) and incubated for 1 h at 37°C in order to deaminoacylate the tRNAs.

then ethanol precipitated, dissolved in 300 μl of 10 mM Na-Hepes pH 7.0 and stored at -20°C after the addition of two volumes of ethanol. The yield was approximately 12 mg of soluble RNA based on an extinction of 25 $l \cdot g^{-1} cm^{-1}$ at 260 nm. The isopropanol precipitable fraction yielded about 20 mg of DNA and 200 mg of salt-precipitable RNAs. The sRNA fraction contains 5 and 7S RNA as well as tRNA, as demonstrated by denaturing 8% polyacrylamide gel electrophoresis.

B. Labelling

pCp ([5'-^{32}P]cytidine 3',5' bisphosphate) for 3'-end labelling of RNA with T4 RNA ligase (Peattie, 1979) was prepared from [γ-^{32}P]ATP (crude, >3000 Ci/mmol, from ICN Biomedical) and the nucleoside 3' monophosphate using polynucleotide kinase, 3' phosphatase free (Boehringer Mannheim). The 10 μl reaction contained 50 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 5 mM dithiothreitol (DTT), 3 mM spermidine-HCl, 0.15 mM Cp, 17 U kinase, and 1 mCi (6 μl) of [γ-32P]ATP. Progress of the reaction was monitored using thin layer chromatography on polyethyleneimine cellulose (Macherey and Nagel) in 30 g/l ammonium bicarbonate. The mixture was boiled and used without further purification.

The labelling reaction (Bonen, 1980; Peattie, 1979) contained 6.5 μg (about 250 pmol) RNA, 400 μCi (125 pmol) pCp, 1.5 μmol ATP, and 10 U of bacteriophage T4 RNA ligase (Pharmacia) in 20 μl of 50 mM Hepes-NaOH, pH 7.5, 3.3 mM DTT, 15 mM $MgCl_2$, 100 g/l dimethylsulfoxide, 10 mg/l BSA (acetylated, Promega), and was incubated for 18 h on ice. The products were ethanol precipitated, dissolved in 98% formamide, 1 g/l each bromphenol blue and xylene cyanol, and separated on a 0.5 mm thick denaturing acrylamide gel. In some hybridization experiments the entire reaction mixture was used without purification.

V. Electrophoresis

A. Polyacrylamide

I used a home made vertical gel apparatus of the Studier type, but in which the top chamber extends the entire 20 cm length of the plate. One plate was permanently attached to the upper chamber so that there was no need of a gasket. This back plate was treated with dimethyldichlorosilane (50 ml/l in chloroform). Spacers and combs (0.6 or 1.5 mm thick) were handmade from delrin sheet. For purification of labelled tRNA the gel contained 50% urea, 7.8% acrylamide, 0.2% N, N'-methylenebisacrylamide, 0.1% N,N,N',N'-tetramethylethylenediamine, 0.1% ammonium persulfate, 50 mM Tris base, 50 mM boric acid, and 1 mM Na_2EDTA. The 20 cm long gel was electrophoresed in 50 mM Tris-borate, 1 mM EDTA for 30-45 min at 0.7-1 kV. tRNAs were eluted by diffusion from gel slices directly into hybridization solution, which was filtered through a 0.45 µm membrane filter before hybridization. For separation of small DNAs, I used the same conditions but without urea. DNA fragments for use as probes were eluted by soaking the crushed gel slice in 2.5 M ammonium acetate overnight, and precipitating the DNA from the supernatant fluid with ethanol.

B Regular agarose

I used simple tanks with platinum electrodes to perform horizontal agarose gel electrophoresis using wicks. The gels, most commonly of 1% agarose in 1 x TEAS (see below) were 3 mm thick, 20 cm long and were typically electrophoresed at 40-50 V for 16-24 h. The composition of 1 x TEAS is 50 mM Tris base, 20 mM sodium acetate, 30 mM acetic acid, 20 mM NaCl, 2 mM Na_2EDTA. The gels used for landmark analysis were 20 cm wide, 25 cm long, and 5mm thick and were run submerged. The well-forming combs used for junction-hunting and landmark analysis had a tooth spacing of 4.5 mm center to center to allow loading with a multichannel pipette.

C. Pulsed Field Gels

All of the work presented here was done with CHEF (Chu *et al.*, 1986) electrophoresis, though in earlier work I used both OFAGE (Carle and Olson, 1984) and FIGE (Carle *et al.*, 1986). The general technique is similar for all three. The original CHEF design had each of the four electrodes in an array powered from one pole of a four-pole-double-throw relay. This was done in order that the electrodes would be isolated from one another while inactive. I chose instead to isolate the electrodes from each other with diodes (figure !). This allows a two-pole switch to be used, just as in other PFG applications. The switch used for most of the work was a high voltage vacuum relay (K-12, Kilovac Co., San Diego), as recommended by Schwartz and Cantor (1984), but equally good results were obtained with common 1 A 110 V ac relays. The values of the resistors of the contour clamp (470 Ω) and the geometry of the electrodes were those described (Chu *et al.*, 1986). High quality, 2 W flameproof resistors were necessary-some resistors of the same rating deteriorate quickly from the sustained high current and dampness. The 110 V coil of the relay could be controlled by a variety of timing devices, of which the most convenient was a Commodore 64 computer, which costs about the same as a GraLab interval timer and only slightly more than an industrial recycle timing module. The computer has a port with eight channels which can be directly controlled from the built-in programming language (Basic). A very simple Basic program can be used to control the switching through one channel, and the power supply through another. The interface consists of a DP1210 solid state relay which, activated by the logic-level signal from the computer, can control 110 V ac devices drawing up to one Ampere. I used one such solid state relay to shut off the power supply at the end of a preset run time, and another to control the switching interval.

I used a 10 kV Savant power supply for most experiments, with the onboard timer replaced with a relay controlled by the computer. The applied voltage was 220 V, giving a measured field

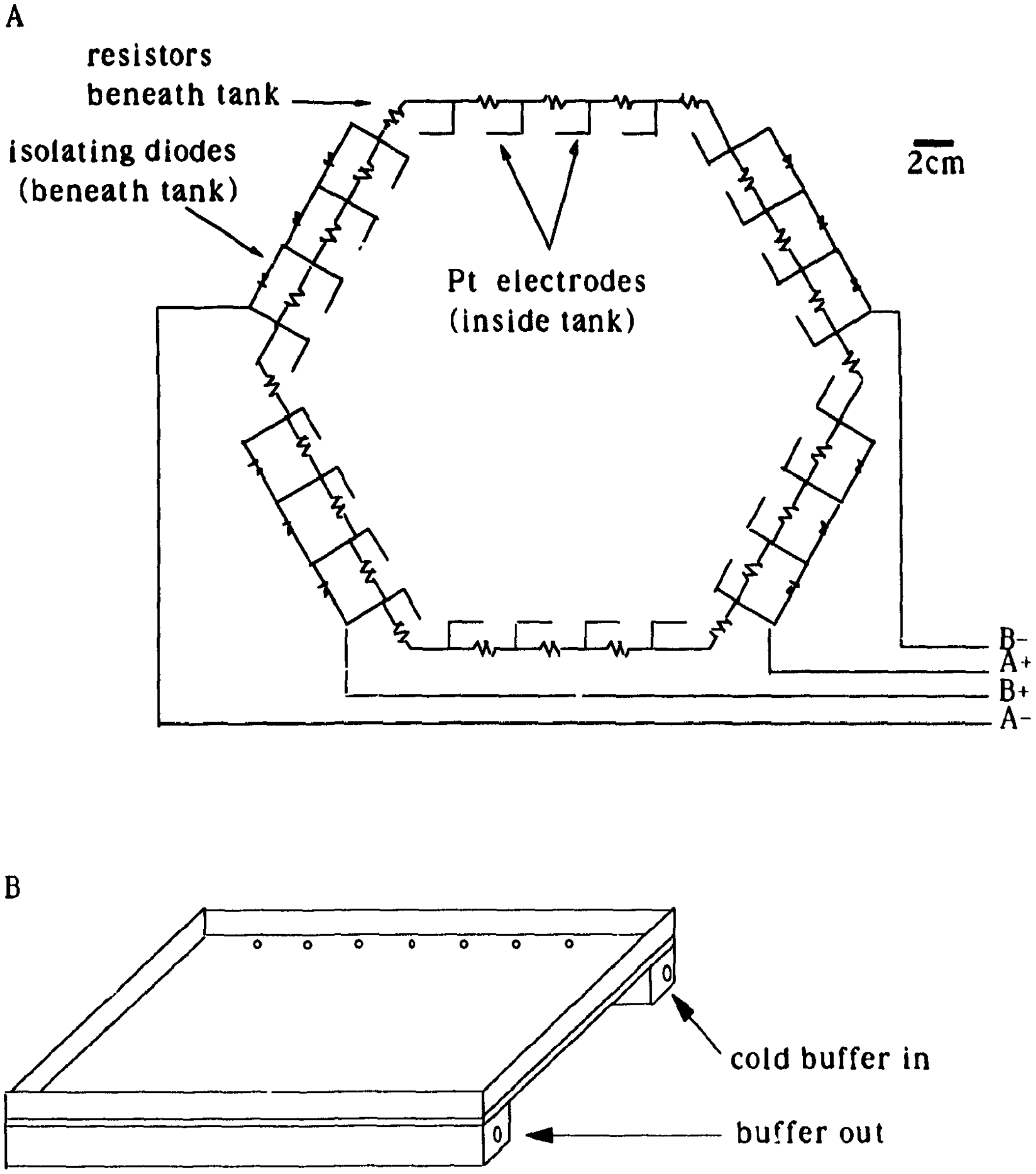


Figure 1. Drawing of the CHEF electrophoresis tank.;

A. Wiring diagram. The electrodes are shown to scale and are inside the box. All diodes are 1N4007, resistors 470Ω, 2W. Diodes and resistors are below the box.

B. The tank, constructed of 6mm acrylic sheet. Overall dimensions are 39 x 39 x 8.3 cm. All plumbing is of 6 mm inside diameter vinyl or polyethylene tubing.

strength of 6.5 V/cm in the area of the gel. A monstrous power supply such as the Savant is not necessary, but not all power supplies rated to produce the required voltage and current (about 0.2 A) can actually put out this amount of power over a long period. The gel tank proper is an elaborated submarine cell (figure 1), with provision for the circulation of the buffer (45 mM Tris base, 45 mM boric acid, 1 mM Na_2 EDTA, typically 1 cm depth) through an external cooling coil at a high rate in order to maintain an even and constant temperature, which is essential for good resolution. In early work I used a Masterflex (Cole Parmer 7553-20,7017-20) peristaltic pump, as recommended by Carle (1984), to circulate the buffer. Peristaltic pumps have the advantage of being electrically isolated from the buffer, but even with frequent tubing changes, occasional runs are lost because of tubing or head failure. On the recommendation of K. Conover, low cost gear pumps (Cole Parmer 7012-00) were used. These can be relied on to run continuously for about five months. The buffer was circulated through about 3 m of 1/4 x 3/8 inch polyethylene tubing (Plastics Maritime) stuffed into a 50% ethylene glycol bath maintained at about -1° C . This resulted in a buffer temperature of 11° C during electrophoresis.

The gel tank was assembled from 6 mm thick acrylic sheet solvent welded with chloroform. Tubing ports were fashioned from acrylic tubing, also solvent welded. Electrodes were 1.5 cm lengths of 0.2 mm platinum wire soldered to pieces of 14 gauge copper wire passing through 1.5 mm holes in the box bottom. The holes were sealed with silicone bathtub sealant in such a way that the copper and solder were not in contact with the buffer. The free ends of the electrodes were tacked down with more sealant, leaving 1 cm of electrode exposed. The 0.2 mm platinum wire is the thinnest practical for this application, since the electrodes erode and require periodic replacement.

Gels were cast on 15 cm square glass plates 1 mm thick. Two pieces of filter paper were glued near the edges to anchor the gel to the plate. The well-forming comb (1 mm thick) was placed 2.5 cm from one edge of the plate. Liquid, melted agarose or solid agarose samples were placed in the wells before the gel was put into the buffer. Some gels were first conventionally electrophoresed with wicks at a relatively low voltage before CHEF electrophoresis . This improved band straightness and sharpness by minimizing the effect of inhomogeneous conditions in and near the wells. I stained gels in the circulating buffer, using about 0.2 mg/l of ethidium bromide for an hour, and destained them in water.

VI. Junction hunting

A. Library construction and maintenance of clones

The objective of this search was to find clones of *Haloferax volcanii* DNA containing cleavable XbaI or SpeI sites for use as hybridization probes to prove the contiguity of large restriction fragments.

I prepared a library of *Haloferax volcanii* SalI partial digest fragments, starting with DNA in agarose microbeads. Three of a series of partial digests were pooled and electrophoresed on a 0.5% low-melting-point agarose CHEF gel with a switching time of 0.5 s. A 15-23 kbp slice was cut out, melted and diluted with 0.1 M NaCl before phenol extraction. The vector, lambda EMBL3 (Frischauf *et al.*, 1987), was prepared by ligation of the lambda ends and digestion with SalI before purification of the arms in the same way as the insert. Approximately equal quantities of the two samples (3 µg) were ligated together and packaged with a commercial packaging mix (Promega) and plated on *Escherichia coli* Q359. The yield was about 7.5 x 10^4 plaques from packaging one tenth of the ligation mixture.

Plaques were picked from the initial plating and patched onto a lawn of *Escherichia coli* GM48. Plates oveı laid with ɪnoculated top-agar were prepared and stabbed about twenty times with the broad end of a sterile Pasteur pipette. These prepared plates could be stored for a few days at 4°C before use. Each plaque was transferred to one of the prepunched circles with a toothpick. After 8 h incubation at 37°C, the plugs were transferred to Titertek tubes containing 0.5 ml of 50 mM Tris-HCl (pH 7.5), 100 mM NaCl, 10 mM $MgCl_2$, 0.1 g/l gelatin, with the aid of sterile toothpicks. These stocks were kept at 4°C.

B. Growth of Clones and DNA preparation

Pools of eight lambda clones were grown on plates and DNA purified exactly as described by Helms *et al.* (1985), using GM48 as the host. The purification involves eluting the phage from the plate by overlaying with 10 mM Tris-HCl (pH 8.0), passing the eluate over a 2 ml column of DEAE-cellulose, washing the column with 10 mM Tris-HCl (pH 8.0), 10 mM magnesium acetate, and 60 mM sodium acetate, and finally eluting intact phage with 10 mM Tris-HCl (pH 8.0), 50 mM magnesium acetate. DNA is then extracted from the phage by treatment with Proteinase K and SDS, the SDS precipitated with potassium acetate, and the DNA precipitated with isopropanol.

C. Assay for restriction sites

Assay of the clones for restriction sites was done on one quarter of the DNA produced by the plate lysate. The 10 µl reaction mixture contained, in addition to the DNA, 10 mM Tris-HCl pH 7.4, 10 mM $MgCl_2$, 50 mM NaCl, 6 mM 2-mercaptoethanol, 15 µM α-thio dTTP, 0.1 µCi [α-^{32}P] dCTP, 0.5 U Klenow polymerase, 2.5 U XbaI. After 0.5 h at 37 ^{o}C, 5 µl of 98% formamide containing 0.1 g/l bromophenol blue was added and after 5 more minutes at 37^{o}C, the samples were loaded on a 1% agarose gel. A specially made comb with 4.5 mm spacing between teeth was used to allow loading with a multichannel pipettor, whose tips are 9 mm apart. The gel was dried and autoradiographed under the same plastic wrap under which it was run to avoid contamination with the radioactive electrophoresis buffer.

Size estimates from photographs or autoradiograms of CHEF gels were made with the aid of a digitizing tablet and the program MacDigisizer by K. Conover. Whenever possible, marker lanes on both sides of the lanes of interest were used. The image was aligned with the digitizer in such a way that the equivalent marker bands were nearly at the same latitude, and both sets digitized. The program fits a polynomial of order chosen by the operator, by a least-squares method. The objective is to make a near-linear fit, *i.e.*, the higher-order coefficients are small. Provided there are enough points, CHEF gels consistently fit a sixth order polynomial. Each set of run conditions gives regions of higher and lower resolution, with inflection points, so individual size determinations have to be evaluated critically. The best size estimates are produced by measuring size under a variety of conditions.

VII. Blotting and hybridization

A. Southern transfer

DNA was transferred from agarose gels to Genescreen or Genescreen Plus (DuPont) membranes by capillary transfer in one or two directions with 0.4 M NaOH (Reed and Mann, 1985). DNA was fragmented before transfer, where necessary, by a treatment with 0.25 M HCl at room temperature. In an optimization experiment, the strongest hybridization signals were obtained from pulsed- field gels treated for 20 min (this depends on the thickness of the gel and its buffering capacity). Conventional gels (fragments of less than 20 kbp) gave the strongest signals when untreated. For bidirectional transfer. the gel was soaked in 0.4 M NaOH for 30 min. The transfer was arranged as shown in Maniatis *et al.* (1982). except that the membrane was wet in water. the transfer solution was 0.4 M NaOH. no weight other than a piece of glass was used. and the transfer was left for as little as 4 h. After transfer the membrane was rinsed in 50 mM sodium phosphate pH 7.2 and (in the case of Genescreen) treated with UV light for 2.5 min. The UV source was a 30W germicidal bulb at a distance of 10 cm.

B. Dot blotting

Since they were not intended for quantitative analysis. dot blots were prepared by pipetting a volume of less than 10 µl of DNA dissolved in 0.4 M NaOH directly onto dry nylon membrane. This method is less laborious than applying samples with a manifold. and a more compact spacing can be used. The filter was subsequently treated with 50 mM sodium phosphate buffer pH 7.2 to neutralize the alkali and irradiated while damp with UV light in the same way as Southern blots.

In order to facilitate the production of many copies of a dot blot of the minimal set. I devised a mimeography-like procedure. In this method. the samples of DNA in 0.4 M NaOH were applied to small pieces of Whatman #1 filter paper which had previously been bonded to a sheet of glass with Parafilm M. The parafilm adheres well to clean class, and the dry paper adheres to the parafilm when pressure is applied. Once the samples are applied, impressions can be taken on pieces of nylon membrane by placing them on the stencil, overlaying with a piece of filter paper moistened with 0.4 M NaOH and rolling with an 18 mm test tube. Up to twelve similar filters have been made this way, with only one round of pipetting required.

C. Hybridization

All hybridizations were done in Ziploc freezer bags. These polyethylene bags are heavy enough to remain flat and rarely leak. I routinely used 25 µl/cm^2 of hybridisation solution containing up to 100 µg/l of labelled nucleic acids and any quantity of unincorporated nucleotides, which were not routinely separated from the probe. Filters were sometimes stacked in the same bag, provided they were first individually dipped into the probe solution. Bags were stacked in a flat-bottomed plastic box and covered with a layer of wet paper towels which served to keep the area of the filters flat and bubble-free. The hybridization reactions were usually allowed to proceed for 18 to 20 h in an air incubator.

Washing of blots was done in Frig-O-Seal containers, which have flat bottoms and high sides. The solution was first heated to the stated temperature and, after the blots were added, maintained in a shaking waterbath. Many filters were washed in the same box.

All autoradiography was on Kodak X-O-Mat AR5 film with one DuPont Cronex Lightning-Plus intensifying screen at -70°C. Filters were wrapped in plastic wrap and exposed while damp in order to facilitate stripping of probes. Exposure of filters in the damp state also has consequences for the analysis of the result, since Genescreen membrane shrinks approximately 10% upon drying. Wherever possible size determinations were done with size markers seen on the same exposure, but since this was not always convenient, I was as consistent as possible in the manner of preparing the blots for autoradiography.

I stripped probes from filters by washing them for 0.5 to 1 h in a 1:1 mixture of technical grade formamide and 50 mM sodium phosphate, pH 7.2 (this and other phosphate buffer concentrations are given here with respect to sodium), 1% SDS. Each filter was marked with a unique identifying number in ballpoint pen, and each hybridization result was referenced to a chronological record, so that for any hybridization the previous use of the filter could be retrieved to check for bands showing through. Performance varied, but certain filters have been used more than 25 times. Most filters can be used at least 10 times.

RNA probes

Hybridization of DNA on filters with labelled tRNA w is done in a solution of 0.6 M NaCl, 70 mM trisodium citrate ("4 x SSC"),10 g/l SDS, 25 mM NaH_2PO_4, 500 ml/l formamide, pH approximately 7.2, at 42°C. Washing was in 500 ml/l formamide, 50 mM sodium phosphate pH 7.2, 1 g/l SDS at 45° C (once, 0.5 h), followed by 4x SSC,10 g/l SDS at 42° C (twice, 0.5 h each).

Hybridization with in vitro transcripts was initially done in the same solution as above, with the addition of 100 g/l each of fish sperm DNA and yeast tRNA, after at least an overnight prehybridization. In later work I used 0.5 M sodium phosphate pH 7.2, 70 g/l SDS (Church and Gilbert, 1985) with carrier RNA and DNA as above, and washed the filters in 50 mM sodium phosphate, 10 g/l SDS at 65-68° C in three or more half hour changes. Exposure times varied from 2 h to one week. Occasional very bad background was treated by incubating the blots in 50 mM sodium phosphate pH 7.2 containing 10 mg/l RNAse A for 0.5 h at 37°C . Before reuse, such filters were incubated in washing solution containing 1 mg/l proteinase K.

DNA probes

DNA-DNA hybridization were done in a solution of 0.5 M sodium phosphate, pH 7.2, 70 g/l SDS, at 65-68 °C for probes from *Haloferax volcanii*, 42 or 45 °C for probes from other species.

Labelling reactions were stopped by the addition of .05-.1 of the reaction volume of 10 M NaOH. After at least 5 min at room temperature, the denatured probe was diluted in hybridization solution and added to the bag containing the filter. No prehybridization was necessary as long as the filter was completely wet. Washing was as described above for Sp6 or T7 polymerase. Low-stringency hybridizations were washed in the same washing solution at a series of temperatures, starting at the hybridization temperature and increasing in 3-5°C steps until the signal disappeared.

VIII. Malachite green bisacrylamide

Malachite green bisacrylamide is an adsorbent for double stranded nucleic acids, specific for AT-base pairs, and is a convenient way of separating DNA on the basis of composition. Chromatography was as described by Bünemann and Müller (1978), using a 0.5 x 8 cm column. The solvent was 10 mM sodium phosphate (in this case the concentration is with respect to phosphate), pH 6.0, 1 mM EDTA (PE) and elution was with a gradient of 0-1 M sodium perchlorate in the same buffer. A 5 M sodium perchlorate stock was filtered and adjusted to neutral pH before being diluted into PE. All buffers were filtered and degassed before use. DNA samples to be applied to the column were passed over a 2 ml Sephadex G-50 column in PE. The gradient was formed by diluting 10 ml of PE with 1 M sodium perchlorate in PE at constant volume. Although in principle this kind of gradient mixing produces a convex gradient, over the range used it approximates linearity. A three channel peristaltic pump was used to pump the salt solution in and the mixed gradient out of the mixing chamber in lockstep. The flowrate was approximately 0.2 ml/min and 0.5 ml fractions were taken in decapped 1.5 ml polypropylene microcentrifuge tubes. Fractions were surveyed by A_{260}, scintillation counting of the entire fraction or loading of samples of alternate fractions on an agarose gel, as necessary, and then selected fractions were passed over 2 ml Sephadex columns in water before being ethanol precipitated with 20 µg of mussel glycogen (Boehringer) as carrier.

The malachite green bisacrylamide was regenerated by passing two column volumes of 2 M sodium perchlorate in PE through the column, followed by washing with PE.

Results

I. Karyotype

An estimate of the size of the genome of *Haloferax volcanii* (total complexity, including plasmids) was produced by adding the sizes of as many of the fragments produced by complete BamHI digestion of the DNA as possible, as detailed in table 6. Representative gels are presented in figure 2. The size estimate will be low to an extent depending on the size of the small fraction of the genome contained in BamHI fragments of less than 10 kbp. Another source of error is that in the range of 10-40 kbp, discrimination between single and multiple fragments is imperfect. Finally, while the length of the bacteriophage lambda DNA multimers used as markers is known with the greatest possible precision, it is possible that a structural peculiarity of lambda DNA could cause a systematic error in sizing. Restriction fragments of lambda DNA containing *att*P migrate anomalously on conventional agarose gels (Ross and Landy, 1982), as do fragments containing the lambda origin of replication (Zahn and Blattner, 1985). The latter region has been shown to contain a fixed bend which is inherent in the sequence. Sizes of the fragments total 3.5 x10^6 bp.

Haloferax volcanii is reported to contain three plasmids. The smallest, pHv2 (Pfeifer *et al*., 1981a) has been completely sequenced (Charlebois *et al*., 1987b). The other two are pHv1 of about 90 kbp

n	size,kbp	n	size,kbp
5	463.4	3	48.84
5	369.5	2	42.83
6	195.0	2	42.97
3	176.8	2	33.75
3	123.3	2	32.23
3	119.2	2	30.76
3	118.8	3	31.30
3	114.0	3	27.26
3	114.0	2	27.35
3	110.0	2	24.20
3	99.04	1	24.85
3	98.98	1	23.45
3	92.59	1	22.40
3	88.85	1	21.62
3	86.65	1	18.41
3	79.16	1	16.61
3	73.08	1	15.18
3	70.09	1	14.19
3	66.88	1	13.92
3	62.59	1	13.15
3	62.37	1	11.65
3	58.66	1	11.17
4	54.65	1	10.62
3	52.25		

Table 6. Sizes of the BamHI fragments of Haloferax volcanii DNA. larger than 10 kbp. The sum of these sizes is 3.5 mbp. n is the number of determinations averaged to give the size reported.

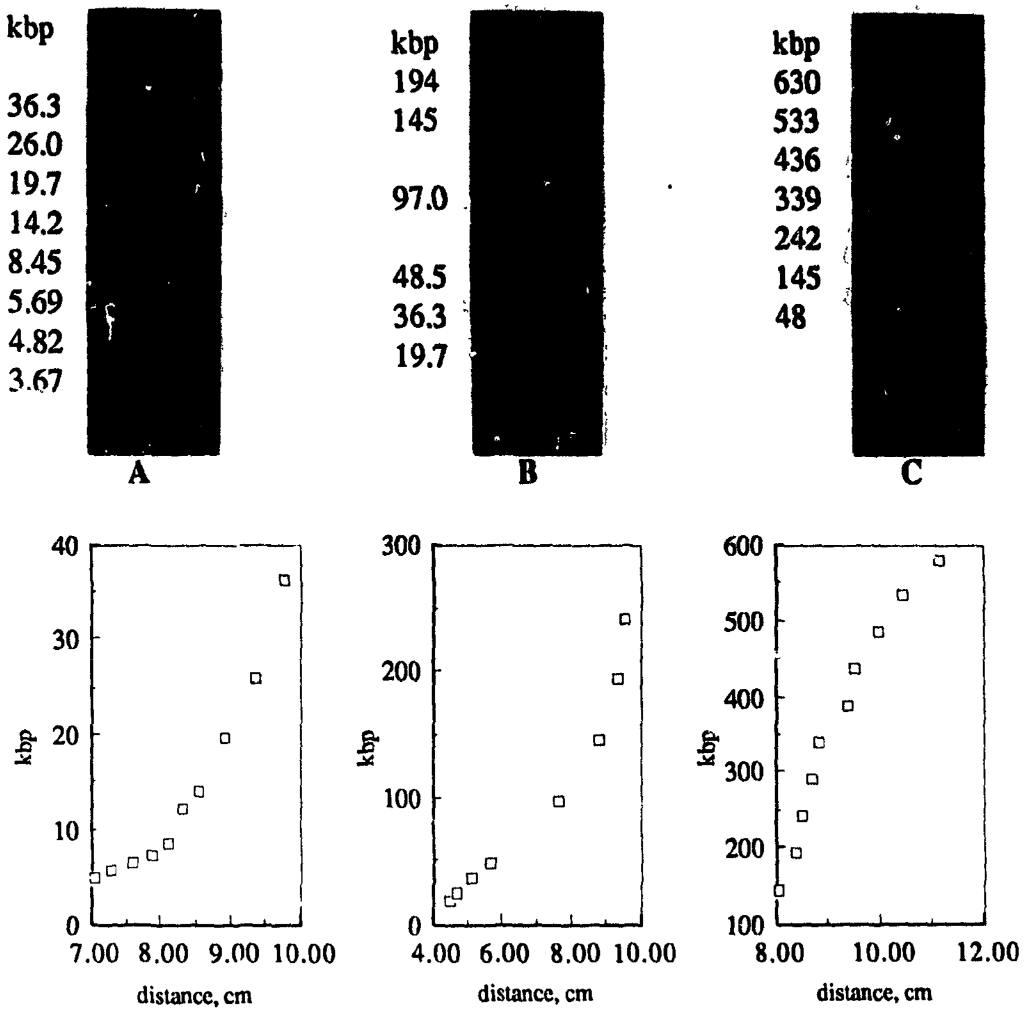


Figure 2. Representative separations of BamHI digests of *Haloferax volcanii* DNA by CHEF electrophoresis. The left hand lane in each case contains multimers and restriction digests of bacteriophage lambda DNA.

Electrophoresis conditions were:

A. 40s, run for 24 h;

B. 5 s for 18 h and then 10 s for 12 h;

C. 1.5 s for 15 h.

(Pfeifer *et al.*, 1981a) and a megaplasmid of about 430kbp (Gutiérrez *et al.*, 1986). Compete cloning of these will be described later. Their migration on CHEF gels is shown in figure 3. Distance of migration is approximately independent of pulse frequency, consistent with these being supercoiled circles (Hightower *et al.*, 1989). Much longer runs (not shown) of undigested DNA of *Haloferax volcanii* show no evidence of chromosomal DNA leaving the wells, consistent with the chromosome also being circular.

II. Top down mapping

In order to identify enzymes which might be useful for top-down mapping, digests of *Haloferax volcanii* DNA with a variety of restriction enzymes were screened on regular agarose gels such as that shown in figure 4. Promising digests could then be repeated on intact DNA and separated on pulsed field gels. The most promising enzymes identified were XbaI and SpeI. Both of these have recognition sequences containing the tetranucleotide core CTAG, which for unknown reasons is rare in many different DNAs (McClelland *et al.*, 1987). In *Haloferax volcanii*, some XbaI sites, such as one known from the sequencing of ISH51 (Hofman *et al.*, 1986) are protected from cleavage, presumably by modification of one of the bases. Some other sites are not modified, because digestion produces a distinct and reproducible pattern of bands, different from undigested or mock-digested DNA. Digests of *Haloferax volcanii* DNA with XbaI and SpeI are shown in figure 5. Both XbaI and SpeI produce large

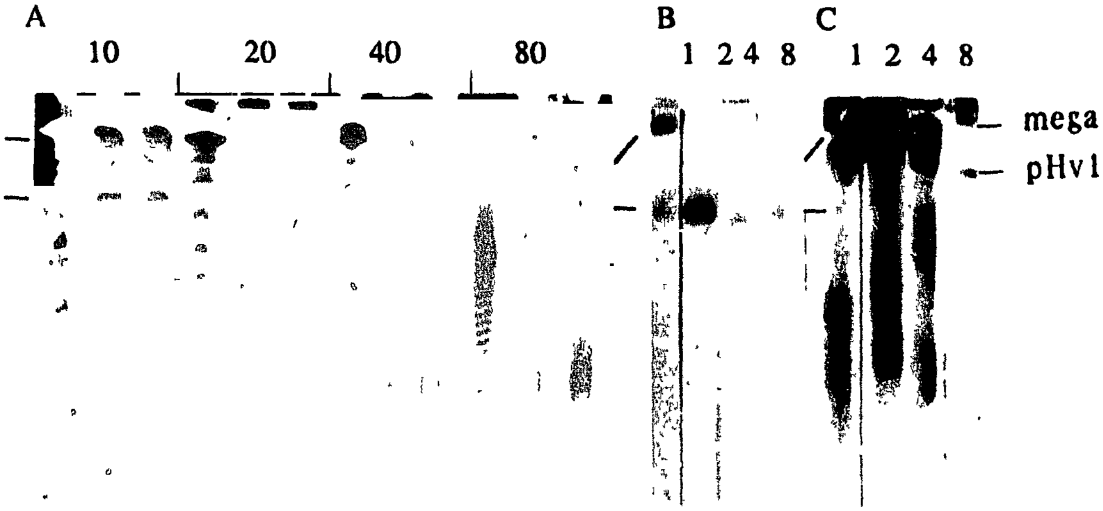


Figure 3. Plasmids of *Haloferax volcanii.*

A. a composite of four CHEF gels similarly run for 15h at switching times of 10, 20, 40, :nd 80 s. The leftmost lane of each gel is loaded with linear multimers of bacteriophage lambda DNA. The rest are loaded with undigested *Haloferax volcanii* DNA.

B. strips of Southern blots of the gels shown in A. probed with the the single copy 10.4kbp fragment of pHv1. The strip of the 10 s switching time gel is labelled 1; 20s,.2; 40 s, 4; 80 s, 8.

C. as in B., but pr·bed with cosmid B42 (megaplasmid)

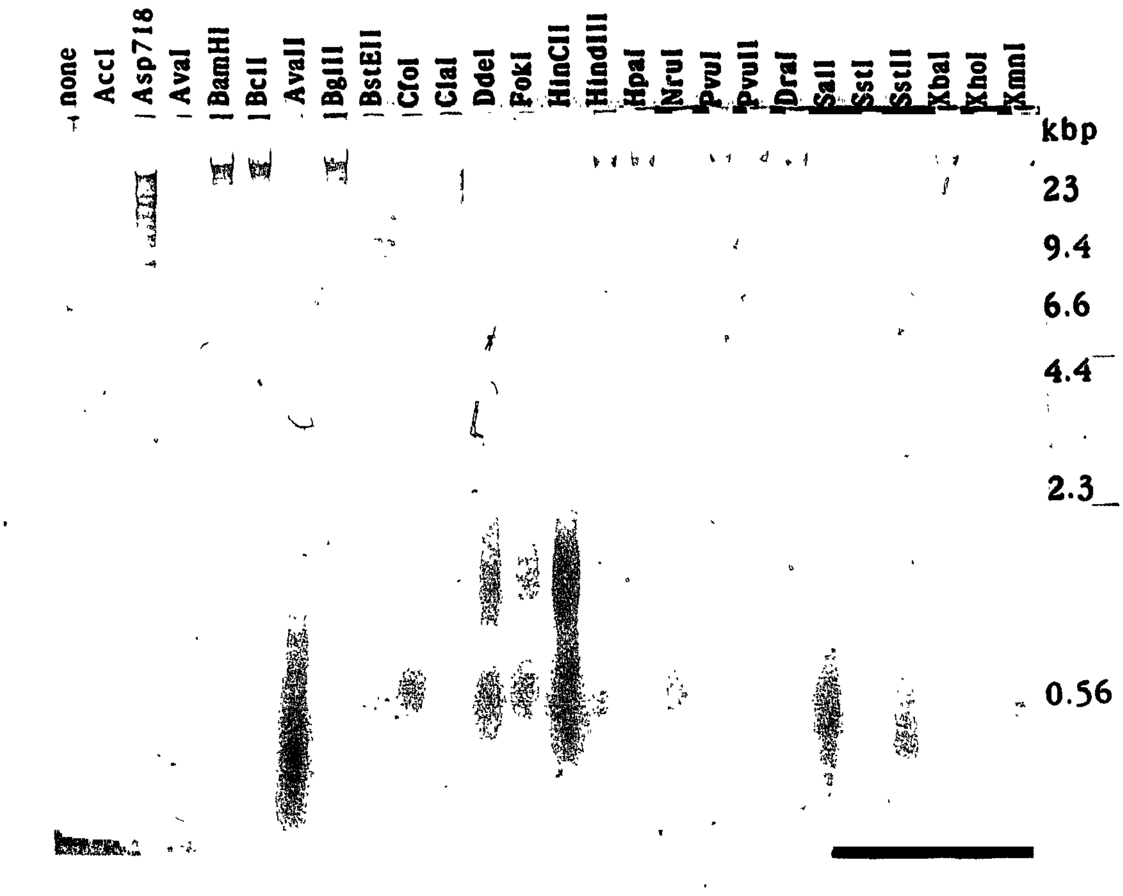


Figure 4. Digestion of *Haloferax volcanii* DNA with a variety of restriction enzymes, performed to identify those potentially useful for top down analysis.

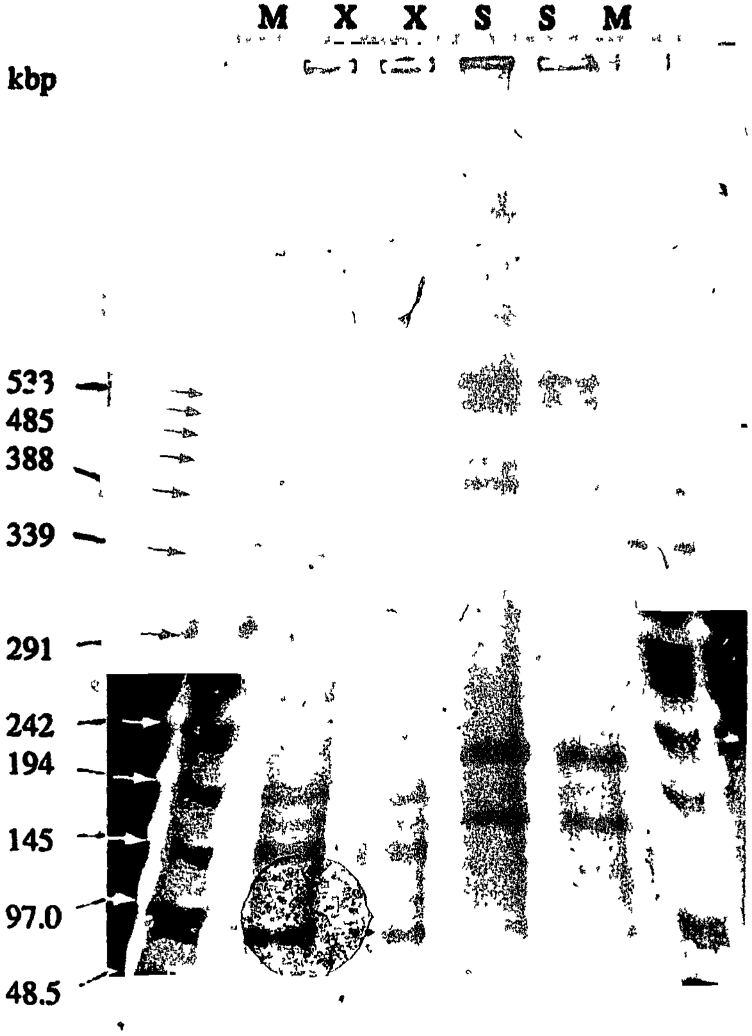


Figure 5. Separation of digests of *Haloferax volcanii* DNA by CHEF electrophoresis.

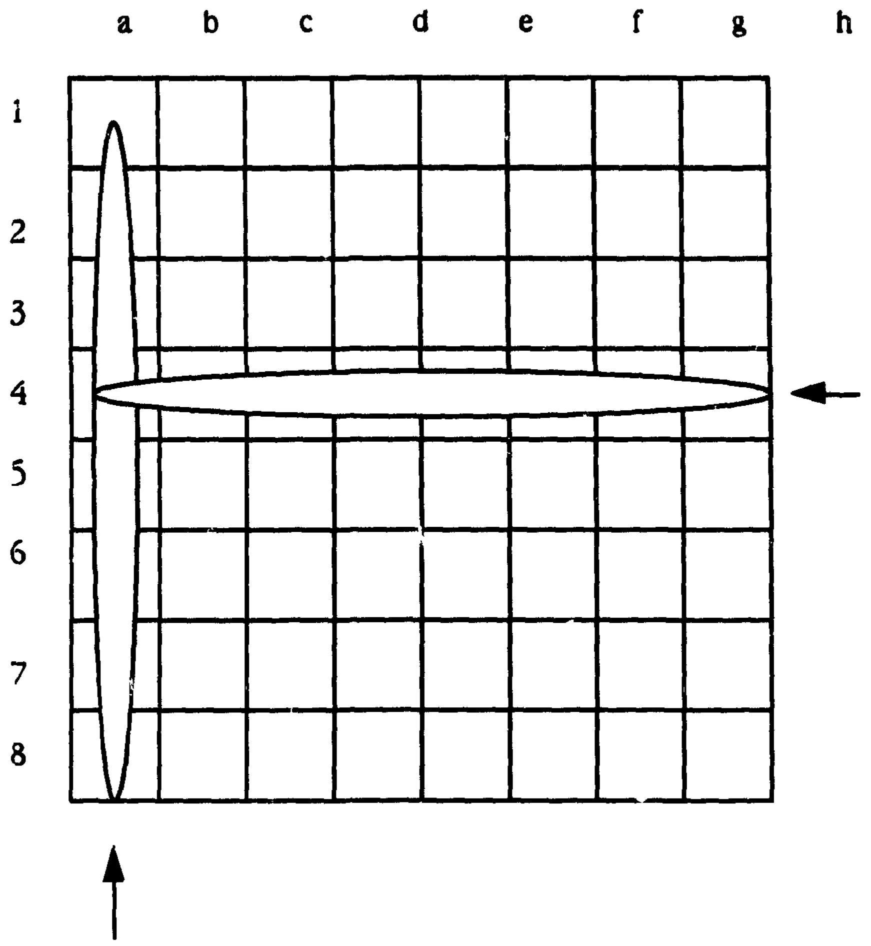
X, XbaI. S, SpeI. M, multimers of lambda DNA.

Switching was at 80 s intervals for 24 h.

fragments and are thus promising for top-down analysis, but there is a background of minor digestion products.

In order to find out more about XbaI and SpeI sites, and to produce a top-down map if it were feasible, I devised a method of searching for junction clones based on digestion and labelling of pools of lambda clones. The junction hunt also served as a useful pilot for later work involving parallel handling of many clones. I prepared a partial SalI digest library of *Haloferax volcanii* DNA in the lambda replacement vector EMBL3 (Frischauf *et al.*, 1987) and picked 1000 plaques. These were transferred to fresh plates and amplified as patches (20 per plate). Phage was eluted from agar plugs cut from the plates and samples of these stocks were pooled in 8x8 arrays, *ie*, in such a way that each clone was represented twice (figure 6). The pooling allowed the hunt to be done in a direct but not impossibly laborious way.

I grew each pool as a plate lysate on a *dam*$^-$ host, since Dam methylation would protect some of the XbaI sites. I prepared DNA by a very streamlined chromatographic method which relies on specific elution of phage from DEAE-cellulose by Mg^{++} (Helms *et al.*, 1985). Each DNA was digested with XbaI or SpeI, and the ends specifically labelled by fill-in with [α-^{32}P]dCTP. Digestion and labelling were done concurrently in the wells of a microtiter dish and the reagents assembled and the products loaded on a gel with the aid of an eight channel pipettor. Cutting of a site in the insert of a clone will result in two fragments, the smaller of which will not be larger than 24 kbp, the combined length of the insert and the short arm of



pool 4 contains clones a4, b4...h4

pool A contains clones a1, a2...a8

Figure 6. Scheme for pooling of clones for junction hunting. With the 64 clones arranged in a grid as shown above, each row and each column is a pool (16 pools from 64 clones). If pool a and pool 4 have an XbaI digestion product in common, for example, it can be assigned to the clone a4, which the two pools have in common.

the vector. All sites should thus produce bands distinguishable from uncut clones. A particular site is recognizable by the sizes of fragments it produces, and the occurrence of the same band in two intersecting pools identifies a site in a particular clone. A set of pools digested with XbaI is shown in figure 7. The two ' ıallest bands (quite faint) produced by pool 6, for example, match two bands from pool D, identifying clone 6D as an XbaI site-containing clone. In figure 8 it can be seen that the genomic sequence corresponding to clone 6D contains a digestible XbaI site. Assuming an average insert size of 15 kbp, the 64 clones analysed in the figure represent 960 kbp, or 0.27 genome equivalent of DNA, and 20 XbaI sites are detectable. There are thus on the order of 75 sites in the unmodified DNA, assuming no drastic cloning bias. The SpeI site frequency is similar.

Each positive (site-containing) clone could then be individually grown and used as a probe on a Southern transfer of SalI, SalI+ SpeI and SalI+ XbaI digested genomic DNA in order to determine whether that site could be cut. Most of the sites found were very slightly digested, while a few were mostly cut (figure 8).

The high frequency of XbaI and SpeI sites in unmodified DNA and their varying degree of digestibility indicated that top-down mapping using these enzymes would be difficult. It appears that the strong bands seen in the XbaI and SpeI digests could be due to partial digestion if there are regions where the sites are clustered. Several clones were detected, for each enzyme, which contained two or more sites.

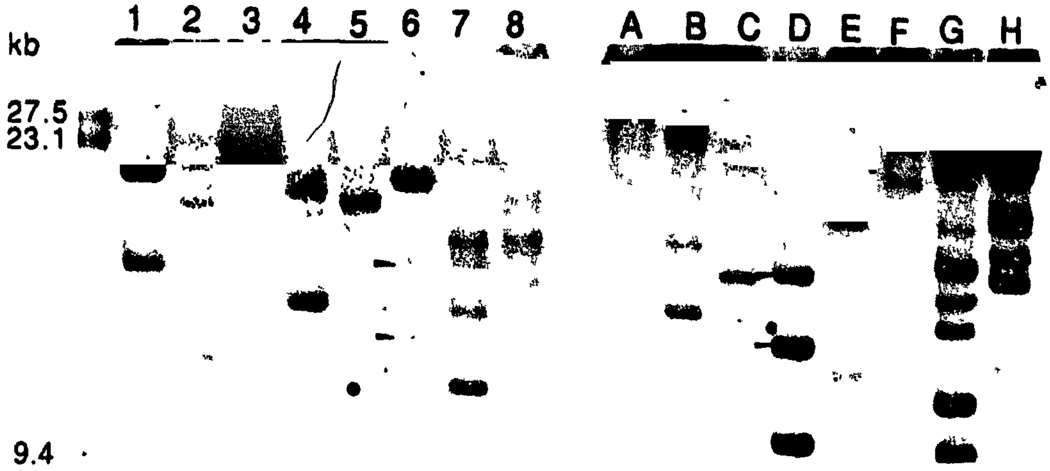


Figure 7. Screening of pooled clones for XbaI junctions.

Each lane is an XbaI digest of pools of a clones, end-labelled and separated on a 1% agarose gel (long run). Each band is one of the products of XbaI digestion of a clone in the pool. Some nonspecific !abelling of the much greater quantity of undigested cones is also visible. Bands which mach between a number and a letter number pool indicate a particular site-containing clone, such as the ones marked with arrowheads above which identify clone 6D, also featured in figure 8.

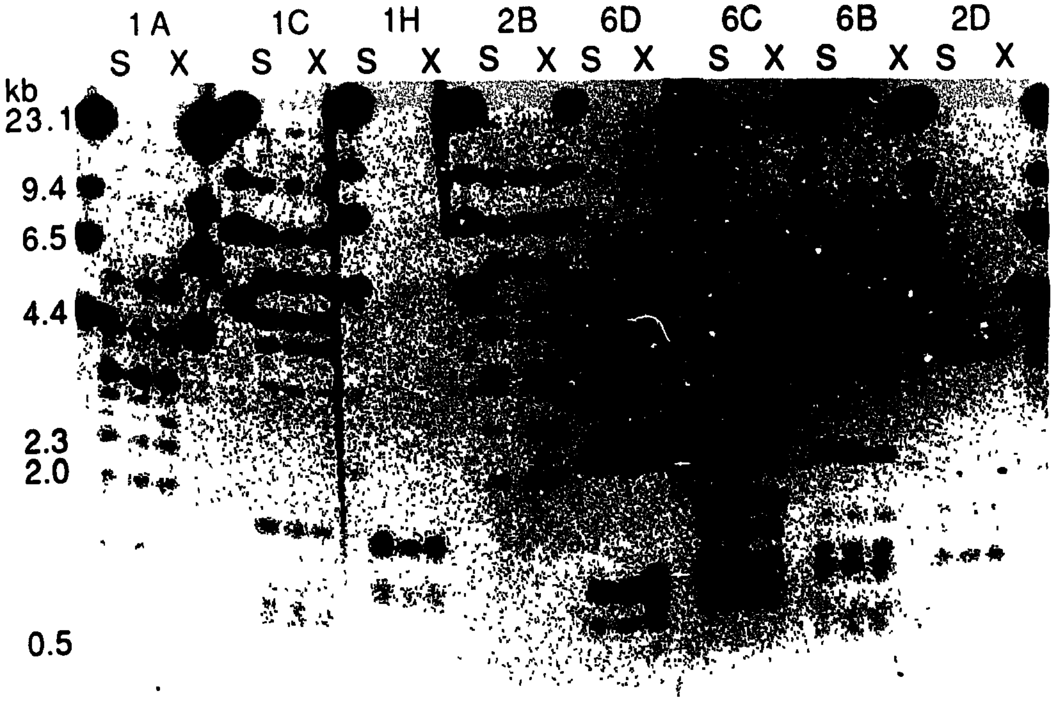


Figure 8. Southern hybridizations, with individual junction clone candidates from figure 7 used to probe digests of *Haloferax volcanii* DNA. Each group of three lanes plus markers (lambda-HindIII) is a separate filter, probed with a nick-translated lambda clone. The three lanes all contain SalI-digested genomic DNA. Those lanes labelled S are also digested with SpeI; those marked X with XbaI.

III. Bottom up mapping

At about this time, it was becoming clear that bottom-up mapping approaches are feasible and have many advantages. A novel mapping strategy was devised by R.L. Charlebois (Charlebois *et al.*, 1989b). The objective was to keep the number of clones to be analysed to a minimum. Simulation experiments indicated that the size of the clones and the length of overlap which can be detected are important parameters. To maximize insert length, a cosmid cloning vector was chosen. It is also technically easier to maintain cosmid clones and to obtain DNA from them than from the other large-capacity cloning vehicles then available, the lambda replacement vectors. The cosmid vector Lorist X (Gibson *et al.*, 1987), used for all of this work, has several advantages. The origin of replication is from bacteriophage lambda, and has a copy number independent of the size of the clone. The insertion site is flanked by transcription terminators in order to minimize effects of transcription from promoters in the insert. The vector contains no pBR322-derived sequences, which is an advantage when screening with plasmid-derived hybridization probes. The insertion site is flanked with promoters, on one side for bacteriophage T7 RNA polymerase, and on the other Sp6. In order to facilitate cloning, a derivative of Lorist X, called Lorist M, was prepared by R. Charlebois by inserting the polylinker sequence from the BamHI site to the HindIII site of m13UM21 (Charlebois *et al.*, 1989a). After some initial trials with DH5, we used the *Escherichia coli* host ED8767 recommended by Little.

The overlap-detection strategy was to use relatively rare restriction sites as landmarks. Each clone is digested with the infrequent-cutting enzyme and the cloning enzyme, and this double digest is compared with the cloning enzyme digest. The presence of a landmark site in a fragment of DNA, and the sizes of the products of its scission, together are a highly distinctive signature of the fragment. Two clones sharing as few as one landmark-containing fragment can be unambiguously identified as overlapping. In the ideal case, every fragment produced by the cloning enzyme would have a landmark, and every overlap would be detected.

We prepared a second cosmid library using partial digestion with MluI, which produces fragments of an average size of roughly 4 kbp, so each cosmid was expected to contain ten MluI fragments. Ten landmark enzymes were then chosen, which were each to have sites on average once in 40 kbp, or one per clone. Analysis of each cosmid requires eleven digests, and the average MluI fragment would be cut once. Cutting frequencies were surveyed using CHEF electrophoresis, as shown in figure 9. The enzymes chosen for landmark analysis also had to fulfill criteria of cost and reliability.

A. First round of landmark analysis

The first round of cosmid preparation and landmark analysis was carried out by R.L. Charlebois and J.D. Hofman, and consisted of 550 cosmids, numbered 1...550. These were first digested with MluI and their electrophoretic patterns compared. Exact duplicates were

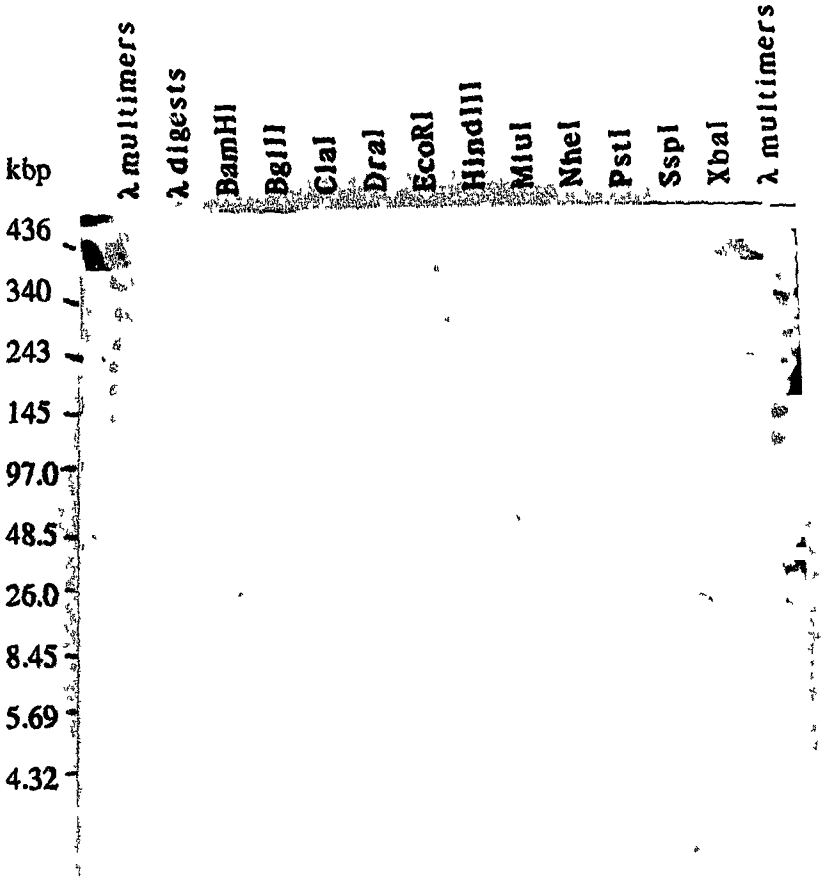


Figure 9. Estimation of fragment size distributions of *Haloferax volcanii* DNA digested with the ten landmark enzymes. Run conditions were 1 s for 24 h followed by 10 s for 8 h.

unexpectedly frequent, perhaps having arisen by growth of the host during the free-expression period allowed before plating. Duplicate clones were eliminated, leaving 319 different cosmids.

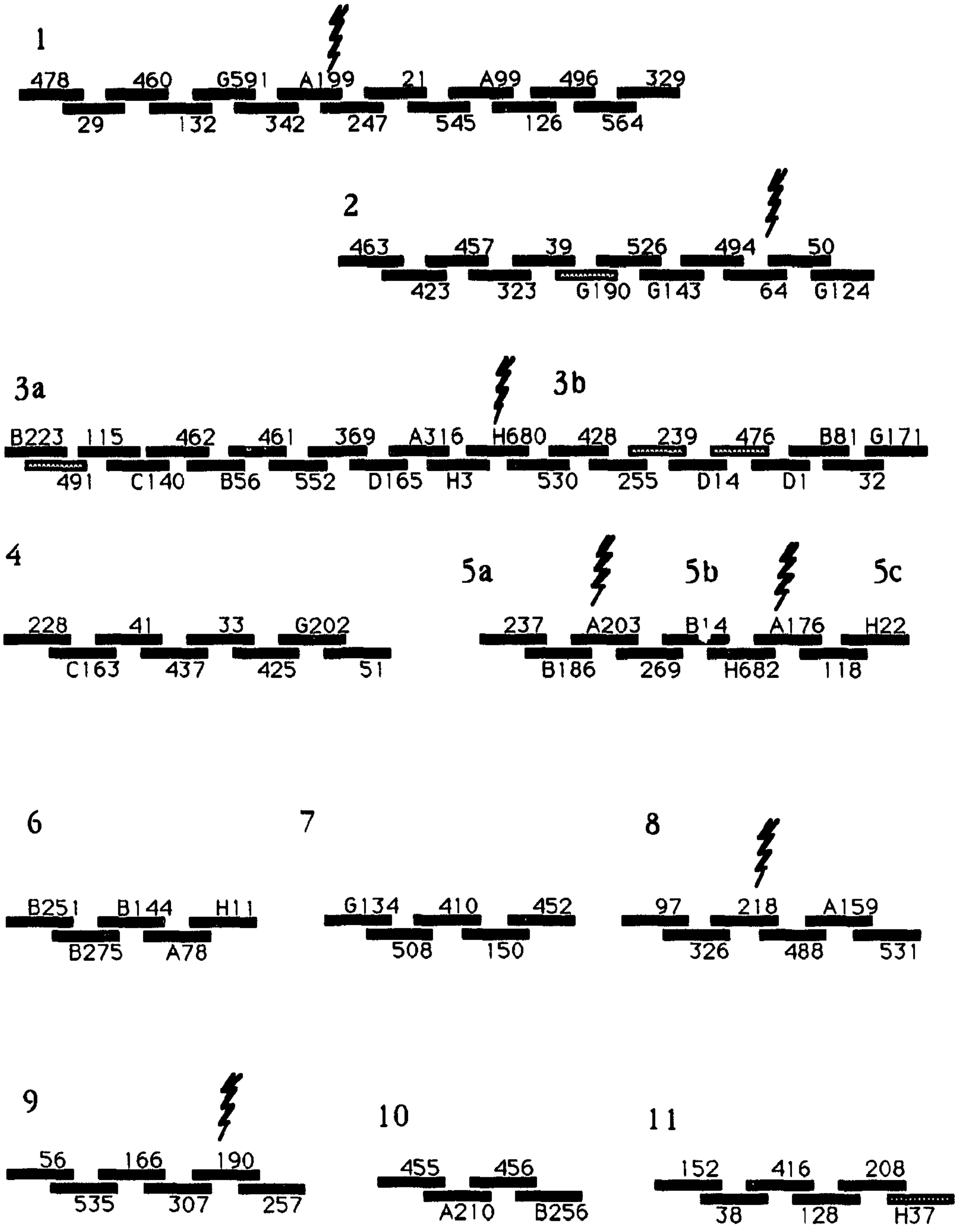
The 319 cosmids could be joined into 59 islands by analysis of the landmarks. Fragment sizes were measured with the aid of a digitizing tablet and fragment sizes and possible assignments of double digest fragments to single digest bands were generated by computer. These were then assembled by hand. Compared to the analysis of Lander and Waterman (1988), the number of gaps was roughly twice that expected for 10-20% overlap. A likely reason would be a bias in cloning.

A second cosmid library was made, again using MluI partial digests, this time fractionated by two rounds of CHEF electrophoresis to produce a highly purified 40-50 kbp size range of fragments. In addition to resolution of larger DNAs, purification by CHEF electrophoresis had the advantage of allowing the desired size range to be concentrated into a narrow zone in the second round. Packaging was with a commercial extract free of EcoK restriction activity. EcoK activity was cited as a probable reason for the strong bias seen in the cosmid libraries used in the *Caenorhabditis* mapping project (Coulson *et al.*, 1986). A combination of cosmid walking and random cosmid landmarking in the new library involving 500 cosmids brought the contig count to 25, with no singletons. The cosmids added to the set at this stage are designated by a letter (A...H) and a number. A set of cosmids intended to cover the contigs with minimal overlap was chosen and large preparations of DNA and long-term frozen stocks of the cosmid-bearing *Escherichia coli* strains were prepared. A schematic representation of the 25 contigs is presented in figure 10.

From here, we proceeded in two directions at once. R. Charlebois began to prepare restriction maps of the minimal set of cosmids, and I began to link up the contigs as described in the next section. Six of the ten landmark enzymes were used for restriction mapping of the cosmids. EcoRI and ClaI were not used because of their high site frequency, and NheI and XbaI because the sites for these enzymes are modified in *Haloferax volcanii*, so that map information from these cannot be compared with or applied to genomic DNA. Lists of the single and double digests necessary for mapping were generated from the landmark data. Cosmids were also checked for sites in small MluI fragments, which could have been missed by landmark analysis. We wrote a computer program for deriving cosmid maps from single and double digests, which turned out to be similar to that of Bellon (1988).

B. Linkup

We imagined the cloning to be nearly finished, and it seemed to be time to take advantage of top-down methods to link up the cosmids. In the following discussion, 'joining" or "closing" means cloning of missing sequences; "linking" means determining relative positions of the contigs in the absence of joining clones. I set out to to link the 25 contigs by hybridization of end probes to Southern transfers of CHEF gels. The gaps were expected to be small, but large fragments


1
478
460
G591
A199
21
A99
496
329
29
132
342
247
545
126
564
2
463
457
39
526
494
50
423
323
G190
G143
64
G124
3a
3b
B223
115
462
461
369
A316
H680
428
239
476
B81
G171
491
C140
B56
552
D165
H3
530
255
D14
D1
32
4
228
41
33
G202
C163
437
425
51
5a
5b
5c
237
A203
B14
A176
H22
B186
269
H682
118
6
B251
B144
H11
B275
A78
7
G134
410
452
508
150
8
97
218
A159
326
488
531
9
56
166
190
535
307
257
10
455
456
A210
B256
11
152
416
208
38
128
H37

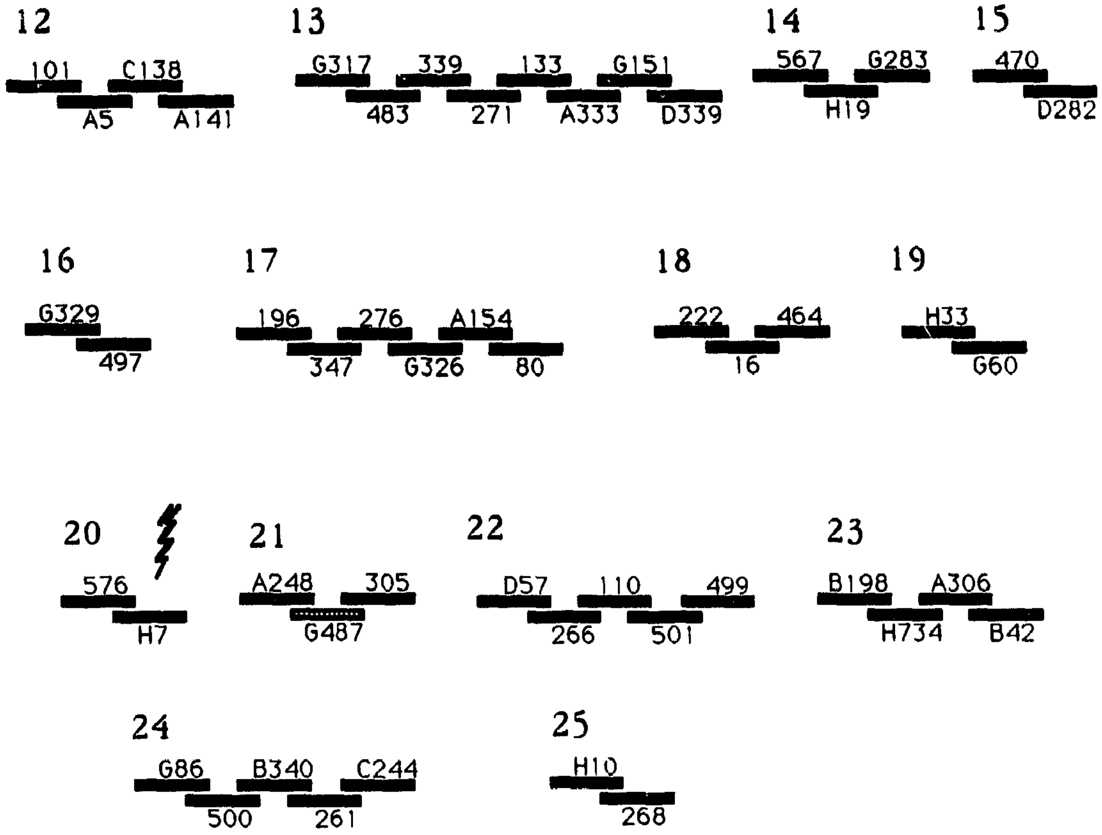


Figure 10. Schematic representation of the map at the 25-contig stage. Each bar represents a cosmid. Length and degree of overlap are not to scale. The degree of overlap between cosmids was not yet determined. Clones which proved to be redundant are shaded and problem clones are indicated with thunderbolts.

are nonetheless the most informative. To link two contigs, it is necessary to find fragments which span the gap, *i.e.*, which are produced by enzymes for which there are no sites in the gap. A low probability of having sites in a given interval is synonymous with infrequent cutting, so I chose to use the least frequent-cutting of the mapping enzymes, BamHI, BglII and DraI. In the first stage, the linking strategy can be seen as producing a fingerprint. Ends of contigs from the same region of the genome should hybridize with the same large restriction fragments.

As a fingerprint, the sizes of BamHI, BglII, and DraI fragments are quite distinctive. Numbers of large fragments in the genome are low, on the order of 100 for each of these enzymes, and they cover a wide range of sizes. The linking strategy can take advantage of much more information than a simple fingerprint match, however. Once the ends of the contigs have been restriction mapped, the size of the fragment measured from the CHEF Southern transfer can be incorporated into the map. Ends which match by a fingerprint criterion can thus be further tested for production of a consistent map, and an estimate of the size of the gap can be produced.

The Sp6 and T7 promoters flanking the insert in the Lorist M vector allow the production of short, end-specific probes from cosmids without mapping of the clone. Short probes are advantageous because they avoid, to a large extent, the problem of hybridization with repeated sequences. The length of the transcript was limited by the low concentration of GTP used. The products of the reaction form a broad smear when separated electrophoretically, but experience with probing of genomic Southern blots has indicated that fragments more than 1 kbp from the end of the insert are rarely seen.

Analysis of the linking data was not as simple as may at first be supposed. This is not least because there are 1225 pairwise comparisons of 50 ends, but also because the data set could not be treated as complete–many of the cosmid maps within the contigs were not yet complete, some fragments were difficult to detect on Southern blots, and finally, there was no assurance that for a given end a counterpart could be found, since some of the gaps could have been large. Blots were reused many times, but for practical reasons, not all of the 50 ends being considered could be probed to the same blot. Rather than directly comparing autoradiograms, I measured the sizes of the fragments against markers made visible by hybridization with nick-translated lambda DNA and made a sorted list of fragment sizes for each enzyme. The gels used in this first round were run under conditions designed to display as wide as possible a range of fragment sizes, but not all sizes could be accurately measured in this single determination. Because the quality of the size information at this stage was variable, I considered potential matches using a very generous 20% error criterion. Potentially matching fragments were then evaluated by checking if the fragments produced by the other two enzymes would also match, and by checking for a consistent map and gap size estimate, to the extent possible with the restriction map data available at the time.

Probings with pairs of ends which were good candidates for linkage were then compared directly. If the two hybridizations had not been done on similar or identical blots, one of the two was repeated, so that identity of the fragments could be checked by superposition of autoradiograms. An example of such a link is presented in figure 11. In the first two panels are the initial matching blots, on which it can be seen that the BamHI and BglII fragments match. The comparison is repeated in the second set of panels at a longer switching time to confirm that the large BglII fragments are identical. They are, and the (accidental) partial digestion products also match. The BamHI fragment (105 kbp) minus the distance from the end of 463 to the first BamHI site (56.7 kbp) and the distance from the end of B42 to the first BamHI site (35.0 kbp) gives an estimate of 13 kbp for the gap. Similarly, tne Bgl fragment predicts that the gap is 1 kbp, and the fact that the DraI fragments do not match predicts that there is a Dra I site in the gap. Cosmid B42 hybridizes with two DraI fragments because of the DraI site in the cosmid. A second round of landmark analysis of random clones (discussed later) has subsequently closed this gap with 12 kbp of new sequence which contains two DraI sites. A brief history of each link is given in table 7.

The largest BamHI and BglII fragments provided their own specific problems. These were not separated on the gels used in the first round, so in a second round I used longer switching time gels, which allowed the individual larger bands to be identified. A given end could thus be assigned with certainty to, for example, the largest BamHI fragment, but the assignment was not very specific because in the BamHI case the largest fragment covers more than 10% of the

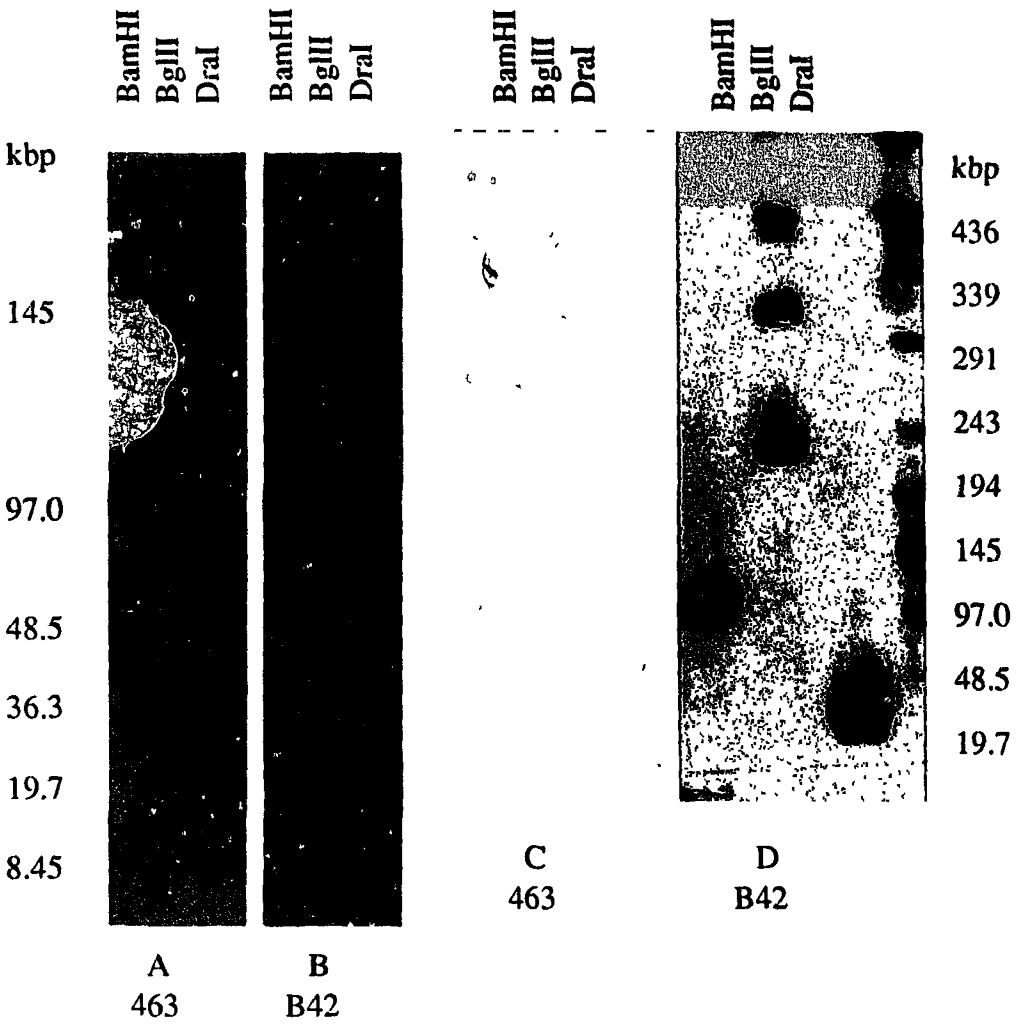


Figure 11. Sequential probings with cosmids 463 and B42 of two CHEF gels of digests of *Haloferax volcanii* DNA. Electrophoretic conditions were 5 s for 18 h and then 10 s for 12 h for panels A and B, and 20 s for 24 h for panels C and D. In panel A the probe was a T7 RNA polymerase transcript, in the rest, nick-translated whole cosmids.

con	cosmid	con	cosmid	comments
1	478	19	G60	G60 replaced and extended by 1A7, closing gap
1	329	6	B251	B251 replaced by 2E10, gap C
2	463	23	B42	closed by 1F3. see figure 10
2	G124	14	B198	overlap not detected by landmarks
3	B223	14	567	closed by 2H1
3	H3	17		still homeless. Most likely goes with 11A7
3	530	17	196	530 replaced by 2D7, closing gap
3	G411	9	257	closed by 5E1
4	228	5	237	228 replaced and extended by 4E5, gap G
4	51	15	470	closed by 10D2
5	H22	20	427	H7 replaced with 427
6	H11	22	499	closed by 2C7
7	G134	21	305	305 extended by 2B1, G134 replaced with 11E9, closed
7	452	5b	118	overlap detected by a new clone (walking)
8	97	12	101	97 replaced by 5G7 which links to 56,101 overlaps 516
8	531	11	208	after 218 discarded, G203 linked to 11B11 which extends 152.
8	G203	11	152	152 extended by 11B11, ends overlap
9	56	8	5G7	gap F. See figure 13
10	455	21	A248	overlap undetected by landmarks
10	B256	25	H10/268	previously undetected overlap with 268
11	208	18	464	208 extended by 3B10. gap A
12	A141	5a	B186	overlap missed by landmarks
13	G317	5b	4A5	H682 extended by 4A5, linked to G317, gap H
13	D339			still homeless. probably goes with G86
14	G283	5b	269	overlap undetected by landmarks
15	D282			extended by 11A7, but still homeless, probably goes with H3
16	G329	18	222	G329 replaced with 11C2, closing gap
16	497	24	C244	overlap undetected by landmarks
17	80	20	576	576 replaced by 1B5. gap D
19	H33	25	268/H10	6E3 closes gap with H10
22	D57	8	531	overlap undetected by landmarks
24	G86			still homeless, probably goes with D339

Table 7. Summary of links. Con, contig numbers which correspond to those in figure 10. These numbers correspond with those in figure 9. Gaps A through H are marked on figure 9 and are mentioned in table 8. "Homeless" means not linked to another contig. "Closed" means that clones have been found to span the gap.

chromosome and contains 6 of the contigs shown in figure 10. This left only BglII and DraI with the potential of being informative in this subset of the genome, and more information was necessary. For those ends inhabiting the largest BamHI fragment and the largest BglII fragment, I undertook a third round of probing in which these ends were sequentially hybridized to the same blot, on which genomic digests using all six mapping enzymes were separated such as shown in figure 12, which shows the linking of cosmids 305 and G134, whose ends are two of the 12 contig ends from figure 10 in the largest BglII fragment. In addition to the BglII fragment, the DraI fragments match. Cosmid 305 hybridizes to two DraI fragments because it contains a DraI site. The stronger of the two DraI bands is a 37 kbp internal fragment. The more weakly hybridiz-ing fragment is identical in size to that of cosmid G134. HindIII, PstI and SspI fragments also match. These fragments predict a gap of 15-20 kbp with one or more BamHI sites, with the exception of PstI, which predicts 4 kbp. The shortness of the PstI fragment indicates that there is a PstI site in the gap, and the fragments are coincidentally similar in size. The gap was subsequently closed by chromosome walking (discussed below), with 20.2 kbp of new sequence containing a BamHI and a PstI site.

Much of the restriction mapping of the contigs was completed at the end of the first round, and the second and third rounds could be carried out largely with nick-translated whole cosmids, which is much less technically demanding than the RNA end-probes. With knowledge of the map, internal fragments can be identified, and

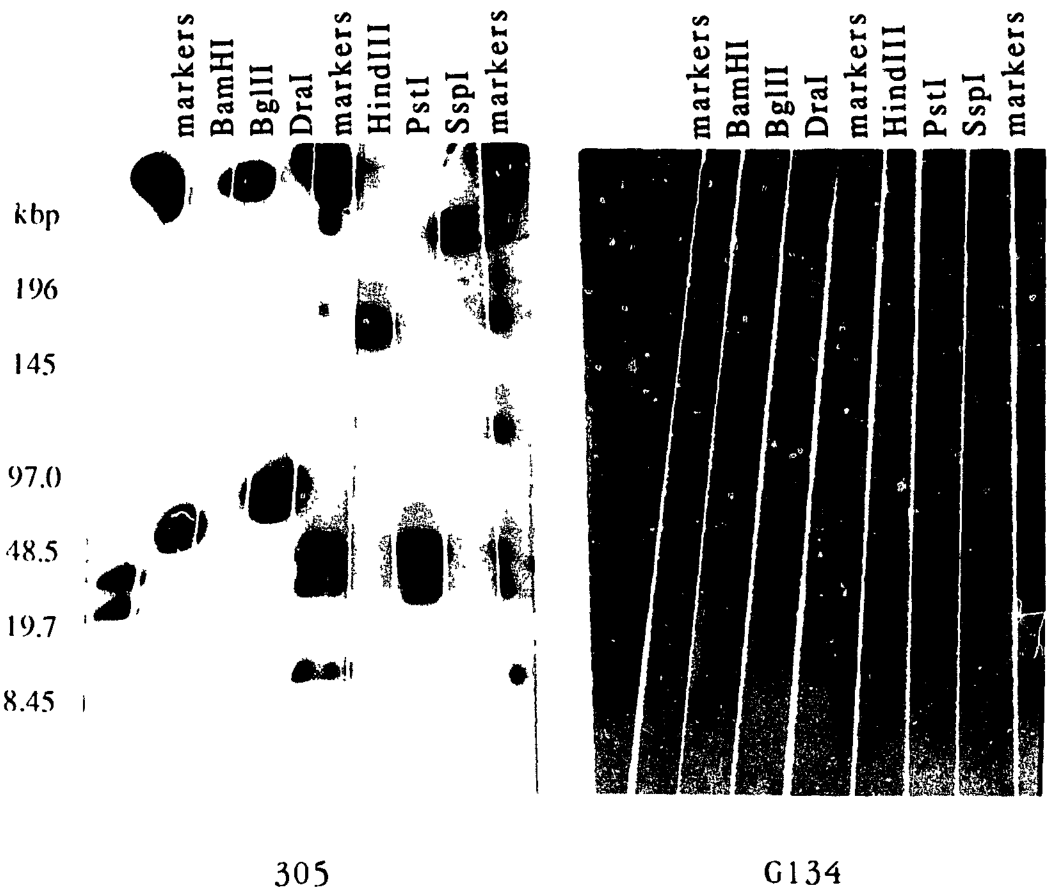


Figure 12. Two probings of a single blot of a CHEF gel of *Haloferax volcanii* DNA with cosmids 305 and G134, to demonstrate that these two contig ends are linked. In both cases, the entire cosmid was used as a probe. The lanes of the gel have been separated for ease of interpretation. BglII, DraI (fainter of the two bands in G134), HindIII, PstI and SspI fragments are of matching size. Electrophoretic conditions were 10 s switching for 24 h.

preliminary information was already available on the terminal fragments from the first round of probings. In those cases where the cosmid sequences were not unique. RNA end probes or gel-purified DNA fragments were used as probes.

Merging of the restriction maps of cosmids to make maps of contigs was in general a simple process. For those overlaps not including sites for the mapping enzymes. we used MluI partial digests and probing with vector sequences to determine the degree of overlap. The clones were first linearized. usually with DraI. which has sites in the vector. Since DraI sites are rare in the inserts. in most cases this allowed us to read enough of the sequence of MluI fragments from the end of a clone to match with the order similarly determined in the overlapping clone. The sizes of the MluI fragments had been accurately determined during landmark analysis. and could be added to get the size of the overlap. Partial digestion using other enzymes was also used to construct the restriction maps of some of the most site-rich areas.

C. Problem regions

Restriction mapping of the cosmids revealed that some of the cosmids in the original minimal set were redundant (*i.e.*. contained no sequences not also present in the neighboring clones). as might have been expected. These. shaded in figure 10. were dropped from the set. In other cases neighboring cosmids did not overlap because the choice of the minimal set had been too parsimonious. This was the case with cosmids A199-247 and 50-64 (marked with thunderbolts in figure 10), and was remedied by going back to the full set of clones for cosmids 449 and 509 respectively.

A small number of the cosmids proved to be difficult to propagate. The original preparations of DNA of cosmid C163 and its neighbor, 41, both contained substoichiometric fragments, which were lost upon repropagation. These fragments were deleted in a portion of the original culture. Similarly, cosmid 190, its replacement 280, and cosmid 218 could not be recovered as pure, full sized cosmids from glycerol stocks or by transformation of *Escherichia coli* with cosmid DNA, even though the original small scale DNA preparations had been unremarkable. In each of these cases I used genomic digests probed with the cosmids to check that the restriction maps of the cosmids match the corresponding genomic sequences. C163, 41 and 280 do match and have been retained as part of the minimal set, but 218 does not.

In addition to being difficult to propagate, cosmid 218 is an example of the final class of problem clones, which also includes H7, A176, A203 and H680. Restriction maps of these clones do not correspond to the genomic DNA. With the exception of cosmid 218, this class of problem clones was detected because their restriction maps conflicted with those of other clones of the same region. Cosmid 218 (contig 8, figure 10) was first suspected because the cosmids on either side of it hybridize with different BamHI fragments, but there is no BamHI site in any of the three cosmids. Furthermore, the cosmids on the 531 side hybridize with the largest BamHI fragment, which already had two known ends, in cosmids 497 and 499. The clones could have arisen by the ligation together of unrelated fragments during cloning, or by rearrangement of the DNA either in the original culture or in *Escherichia coli* after cloning. Rearrangement events in the original culture have not been completely excluded, but in each of these cases the genomic DNA did not show evidence of containing the alternate arrangement as a minor constituent (ie. the pattern corresponding to the problem clone was not visible as faint bands). Other evidence (see Stability, below) also reassures me that genomic rearrangements are infrequent in *Haloferax volcanii*. Cosmids 218, H7, A176, A203 and H680 were discarded. H7 was replaced with the slightly shorter cosmid 427, slightly shortening the contig. The other four created eight new ends to be linked.

D. Second round of random landmark analysis

We decided that the potential for joining contigs by analysis of random clones was still not exhausted. The additional copies of already-cloned regions would also be useful for finding and resolving problem regions. We thus picked 400 additional clones from an amplified sample of the second cosmid library and analysed them by a modified landmark method. Only two landmark enzymes were used in the first instance, allowing sequences which had already been cloned to be identified quickly. Clones containing informative sequences could then be analysed with the remaining enzymes. This closed four gaps, of which I had already linked three, and extenoed a further four ends. The clones introduced at this stage are numbered 1A1...4H8.

E. Walking

As a last round of acquisition of new clones, I used RNA end probes as described above to walk from the remaining unlinked ends. In designing the walk, we profited from the experience of the first chromosome walking undertaken by J.D. Hofman and R.L. Charlebois, which suffered from a high proportion of false positives. Instead of colonies, we spotted partially purified cosmid DNA on filters. The DNAs were prepared and spotted as pools of eight, by the same scheme described for my junction hunt. Pooling has several advantages: it reduced the labor required to produce the filte.:s, it allowed 768 cosmids to be spotted on a single filter without technical difficulty, and each positive dot had to correspond to a signal in an intersecting pool, which increased the confidence in positive signals. Another important technical point was that when the quantity of target DNA is high, such as it is in this case, and the concentration of the RNA probe is also high, the probe will hybridize with all DNA present, complementary or not, according to mass. This effect (which is not seen to the same extent with DNA probes) is not effectively suppressed by the use of 100 mg/l each of unlabelled, unrelated DNA and yeast RNA. It may be due the short polylinker fragment which is transcribed, and present in all of the cosmids. Spurious hybridization was avoided by limiting the activity of labelled RNA to 2 µCi per 3 ml hybridization (which at 3000Ci/mmol and 70%G+C is about 0.2 ng/ml). Two examples of these blots are shown in figure 13. Because some of the still unlinked ends were in regions rich in sites for the mapping enzymes (discussed below), and thus particularly difficult to link, we did a second round of walking with selected ends on the same filters. Candidates were found for nearly all of the ends probed, and seven gaps were closed, of which I had previously linked five.

Many of the links established by Southern hybridization have subsequently been confirmed by cloning of the missing DNA (closure, see table 7). Eight pairs of contigs proved to overlap. The overlaps either do not contain landmarks or were missed because of imperfect landmark assignment. One case in which the linkup conflicted with gap closure was resolved when a problem clone was detected and eliminated. The link was correct in the sense that the ends were in the same region of the chromosome. I had linked cosmids 101 and 97, but when cosmid 218 had been discarded, and one of the resulting broken ends extended with cosmid 516, 516 was found to overlap with 101. Cosmid 97 does thus belong near 101, but I had placed the fragment 516...101 backwards.

There remain eight gaps in the cloned sequence, across which I have mapped all but two. Data on the six remaining linked gaps are presented in table 8. A gap size estimate can be obtained for each fragment which spans the gap by subtracting from the size of the fragment the size of the cloned part. These estimates are quite

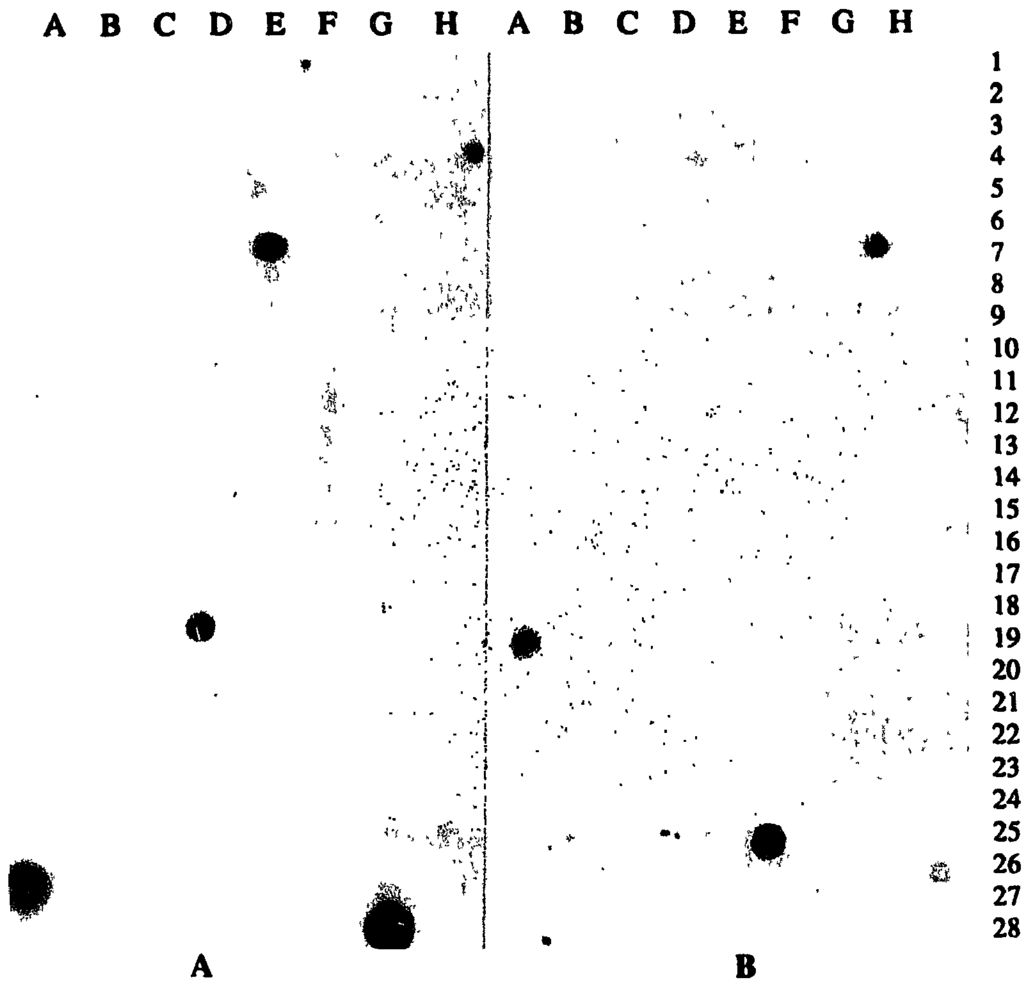


Figure 13. Examples of chromosome walking. Each spot is a pool of eight clones, in a scheme like that of figure 6.

a. the probe was prepared by T7 RNA polymerase transcription of cosmid 228. The two darker spots identify 4E5, which closes the gap with cosmid 237

b. T7 transcript from cosmid D282. Pools 7G and 19a identify cosmid 11A7 which extends from D282

	cos	enzyme	dist	frg	gap	frg	dist	cos
A	464	BamHI	131.7	463	<160	463	172.8+	3B10
		BglII	7.6	18	2.5	15	6.4	
		DraI	45.6	100	9	100	45.5	
C	2E10	BamHI	17.7	23	(22)	30.6	14.2	329
		BglII	27.7	61	10	63.8	35.2	
		DraI	101.5	122	0	121.5	21.3	
D	1B5	BamHI	66.5	89	6	81	12.6	80
		BglII	51.4	46.4	-11	58	11.9	
		DraI	75.2	147	12	149	61.0	
F	56	BamHI	67.9	68	(1.8)	210	208.3	5G7
		BglII	87.9	86	(2.8)	16	5.3	
		DraI	155.0	152	-15	155	14.3	
		HindIII	63.4	69	(1.5)	90	88.5	
		PstI	15.9	16	(2.4)	40	37.7	
		SspI	65.4	62	(-5.2)	20	21.8	
G	237	BamHI	30.9	136	10	136	94.7	4E5
		BglII	35.7	84.2	0	84.6	48.8	
		DraI	50.9	75.5	3	75.9	28.0	
H	4A5	BamHI	17.8	139	12	151	115.5	G317
		BglII	32.8	93	-5	86	62.2	
		DraI	34.5	115	0	100	73.3	

Table 8. Estimates of the sizes of the gaps Gap B was eliminated after the letters were assigned. Gap E is not included because it is the result of the loss of cosmid 280 and has been completely mapped. Gap estimates given in brackets are calculated on the basis of the assumption of a single site in the gap. Cos. cosmid number. Dist. length in kbp of cloned sequence beyond the last site for the enzyme in question. Frg. size in kbp of the restriction fragment detected by hybridization with an end probe. Gap. deduced size of the uncloned sequence separating the two ends.

rough, because they are generally a small difference of two much larger numbers.

Although a gap (gap B) is shown in figure 15, cosmids 11B11 and G203 overlap, a fact realised after the figure was drawn.

All but one of the links rely on more than one restriction fragment spanning the gap. The exception is that of cosmid 56 with 5G7, which is shown in figure 14. The DraI fragment that spans the gap is identifiable as a single band, and the sizes of the terminal fragments produced by the other five mapping enzymes are consistent with one or more sites in the gap.

The two gaps which I have still not spanned are probably deep rather than wide. By this I mean that they are not necessarily large gaps, but they are in a region of the genome rich in both sites for the mapping enzymes and in insertion sequences (discussed later), which makes them difficult to link. 11A7 in particular has two copies of ISH57 at its unlinked end and sites for all of the mapping enzymes within 10 kbp of the end, making it impossible to obtain information on fragments extending from the end. The fragments would likely be short, and so would only be informative if the gap were small. No mapping enzyme fragment extends more than 40 kbp from any of the other three ends. Another restriction enzyme, partial digests, or a completely different approach will be necessary to make the final links.

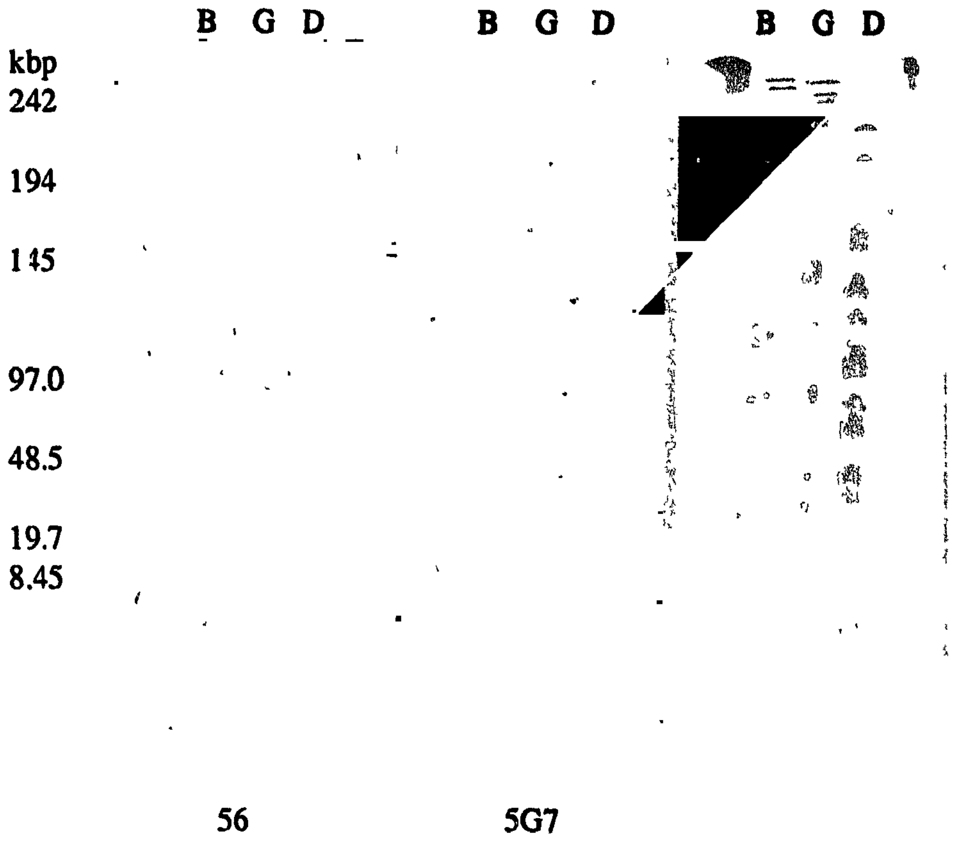


Figure 14. Probings of identical blots with purified fragments of cosmid 56 and cosmid 5G7. The ethidium-bromide-stained gel is shown on the right. Electrophoretic conditions were 10 s switching for 22 h.

The matching DraI fragments are marked with arrowheads.

F. Verification

There is no independent map of the *Haloferax volcanii* genome against which to check this one, and the possibility remains that a chimaeric clone or a false overlap has caused us to link regions which do not belong together. Several aspects of the data give us some assurance that this is not so. The vast majority of the sequences mapped have been cloned more than once. Since ligation of unrelated sequences in cloning is a random event, one is unlikely to isolate the same chimaera twice. The internal consistency of the map also gives assurance of its correctness. When the map is nearly complete, nearly any spurious linkage will produce a topologically unlikely result, such as a fork in the map. The combination of consistency and multiple cloning of most of the genome makes me confident that the map is correct. Nevertheless, some regions are singly cloned, and a check of the long-range continuity of the map is desirable.

I devised a scheme to check the overall correctness of the linkage using a set of overlapping restriction fragments predicted by the map, which can be checked by Southern hybridization. A manageable number of these "leapfrog" fragments of up to 200 kbp in size would survey the map with reasonable sensitivity. Small discrepancies in the map, on the scale of one or a few kbp, will be missed by this analysis. None of the conclusions of this work is sensitive to such discrepancies. Most of the verification data was available as a side product of the linking hybridizations. Use of the linking data for verification is not logically circular for two reasons. Most of the links which the linking hybridizations established were subsequently closed. In the rest, except in the case of the 5G7-56 link, more than one fragment spans the gap, so that the gap can be estimated independent of the leapfrog fragment. The linking hybridizations in which whole cosmids were used as probes give information on internal fragments as well as the terminal ones which were the original objective. Many of the singly cloned sequences were also used as probes in order to check all of their fragments. Our original verification scheme was to do a complete set of hybridizations with the cosmids containing singly cloned fragments. Fragments spanning several cosmids allow one such hybridization to give confirmatory information on more than the individual clone, however, and so the original scheme would give a highly redundant check of the sparsely cloned regions. The leapfrog strategy presented itself as a more efficient alternative, with a more global coverage.

The leapfrogging fragments, by being present and approximately the expected size, help confirm the long-range continuity of the map. Many of the fragments, in particular the longer ones, have been checked at several points along their length, giving further confidence. A first round of leapfrog analysis has been done using the data available from the linking hybridizations and a number of hybridizations designed to check sparsely cloned regions. The fragments are indicated in figure 15 and their observed (from CHEF gel Southern transfers) and expected (from the restriction maps of cosmids) sizes are listed in table 9. The overall agreement is good.

with a discrepancy (absolute value) averaging 8.1%. Most of the variability is likely to be from the CHEF determinations, since sizing of fragments from Southern transfers is necessarily less precise than direct measurement from a gel. The largest discrepancies are found in one-off determinations, as might be expected. Almost all of the CHEF size estimates are lower than the estimates from the cloned DNA, suggesting a systematic error in either the sizing of the MluI fragments of the clones by conventional electrophoresis, or of the larger fragments by CHEF electrophoresis. The latter is the more likel} possibility, because the system is less well characterized, and few independent molecular weight standards are available. Most of the CHEF size estimates are lower than the estimates from cloned DNA, suggesting a systematic error. My genome size estimate of 3500 kbp (figure 2) is also lower (by 12%), than the total length of DNA cloned so far, 3947 kbp.

Some of the fragments (numbers 33, 42, 50) end in gaps, and we can thus only put limits on the size expected. Fragment 33, for example, should (and does) fit in gap F.

Several fragments have large discrepancies, which could either indicate problems or simply poor sizing. Fragment 17 is a one-off determination in a region which has been multiply cloned except for part of cosmid A199, and is probably a poor size determination. Fragment 32 is an extra fragment, completely within the satisfactorily verified fragment 30. It may indicate an error in the BglII mapping data, but the linkage of these sequences is correct.

fr	exp	obs	n	% diff	gap	fr	exp	obs	n	%diff	gap
1	28.7	26.7	1	7.0		30	378.3	374.7	2	1.0	E=6.5
2	47.9	46.5	3	2.9		31	235.6	224.2	3	4.8	
3	169.7	164.8	1	2.9		32	111.5	152.4	4	-29.2	E=6 5
4	157.7	158.7	8	-0.6		33	87.9+	97	1		
5	69.1	75.6	2	-9.4		34	149.9	168.1	4	-9.2	F=20
6	93.7	90.4	2	3.5	A=2.5	35	31.3	27.1	1	13.4	
7	217	231.4	6	10		36	137.2	113.4	2	17.3	
8	147.7	158.2	3	-7.1		37	39.5	29.6	1	25.1	
9	73.7	71.4	2	3.1		38	97.4	88.9	1	8.7	G=0
10	99.2	98.5	5	0.7		39	125.6	123.6	2	1.6	G=0
11	154.2	152.7	4	1.0		40	139.0	135.5	6	2.5	
12	126.7	139.2	1	-9.9		41	53.6	59.4	7	-10.8	
13	108.1	121.3	2	-12.1	C=2	42	77.6+	118.9	7		
14	64.9	62	3	4.5	C=2	43	65.3	48.1	1	26.3	
15	54.0	54.5	2	-0.9		44	126.9	118.9	7	6.3	
16	313.0	265.5	1	15.2		45	41.2	41.2	4	0.0	
17	133.0	105.7	1	20.5		46	34.8	36.5	1	-4.9	
18	117.6	110.2	4	6.3		47	64.5	63.2	1	2.0	
19	82.5	79.0	6	4.2		48	14.1	13.5	1	4.3	
20	83.7	79.6	3	4.9		49	107.8	100	1	7.2	H=0
21	462.7	387.3	1	16.3		50	113.9+	125.3	2		
22	96.2	93.6	4	2.7		51	204.7	196	4	4.2	
23	101.6	91.4	3	10.0		52	76.6	68.5	3	10.6	
24	128.7	118.8	1	7.7		53	27	36	1	-33.3	
25	145.3	143.1	3	1.5		54	236	237.5	4	-0.6	
26	71.9	66.3	1	7.8							
27	104.4	97.3	1	6.8							
28	87.1	81.9	5	6.0	D=8						
29	147.7	146.7	5	0.7							

Table 9. Verification. The numbers in the first column correspond to those given in figure 14. The "expected" size is derived from the restriction maps of the cosmids. The observed size is measured from one or more Southern transfers of *Haloferax volcanii* DNA, the number of determinations is shown under n. Fragment 21 is the largest BglII fragment and its size was measured from an ethidium bromide-stained gel. The %diff is the quantity 100(exp-obs)/exp. Where the expected size includes a gap, the gap estimate used is listed in the last column.

Fragment 37 is intended to verify the small space between fragments 36 and 38. Fragment 36 is also shorter than expectation, so there is the possibility of a problem in this region. B186 contains fragments which have only been cloned once. Fragment 43 is a one-off determination and may be a poor size estimate. This region is the most poorly verified, and its richness in restriction sites and repeated sequences (discussed later) make it difficult to deal with. A restriction enzyme which cuts this region infrequently would be very useful for further checking of continuity and linking of the remaining ends. Fragment 53 is designed to check the junction between fragments 52 and 54 on the 442 kbp plasmid. Because the size doesn't agree well, it does not exclude some irregularity at the junction, but it would have to have affected both cosmids B198 and H734. Confidence in the overall continuity of the map is thus high, but the leapfrog verification has indicated the need for more data on the areas of cosmids B186, H3, and B198.

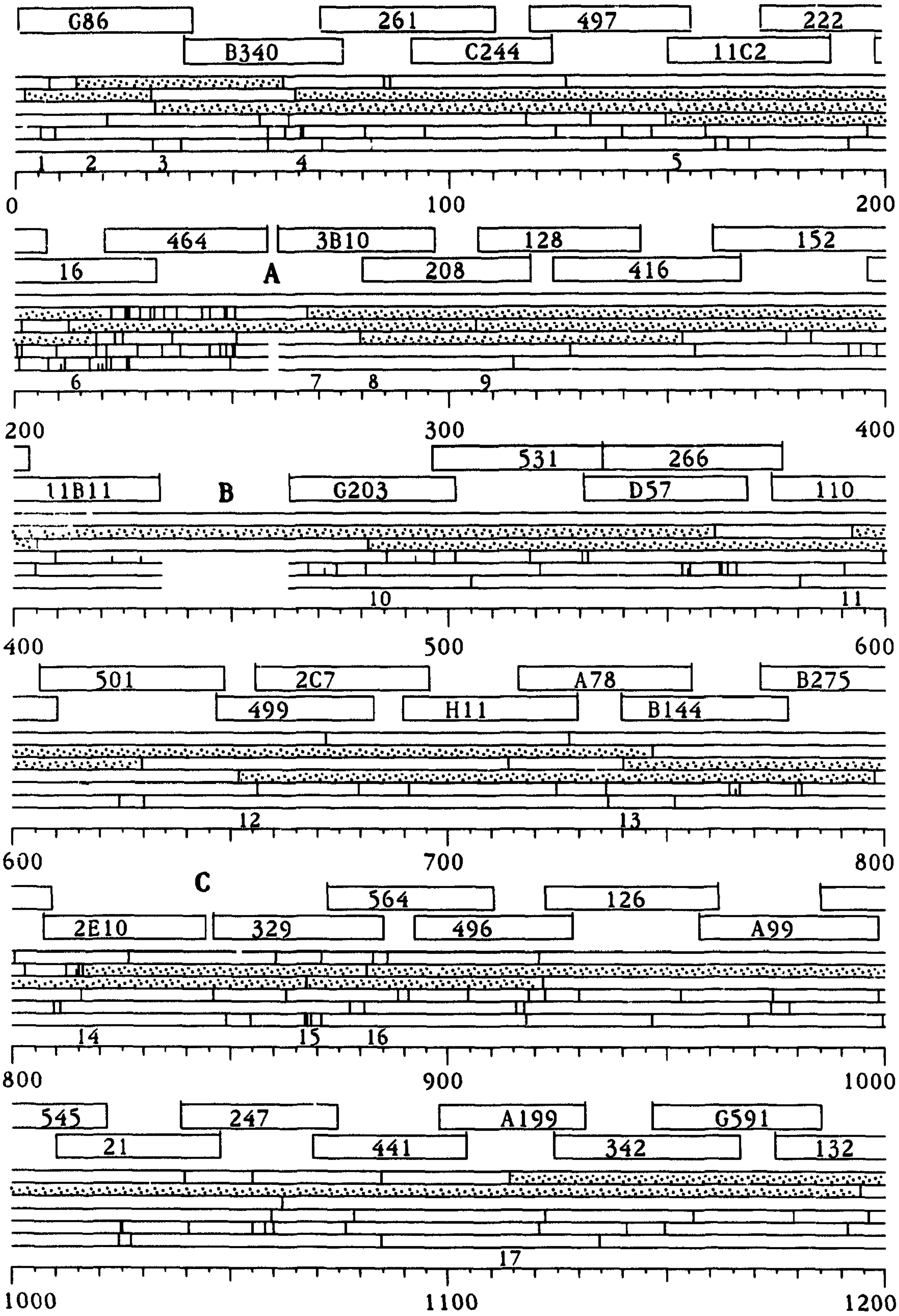
G. The big map

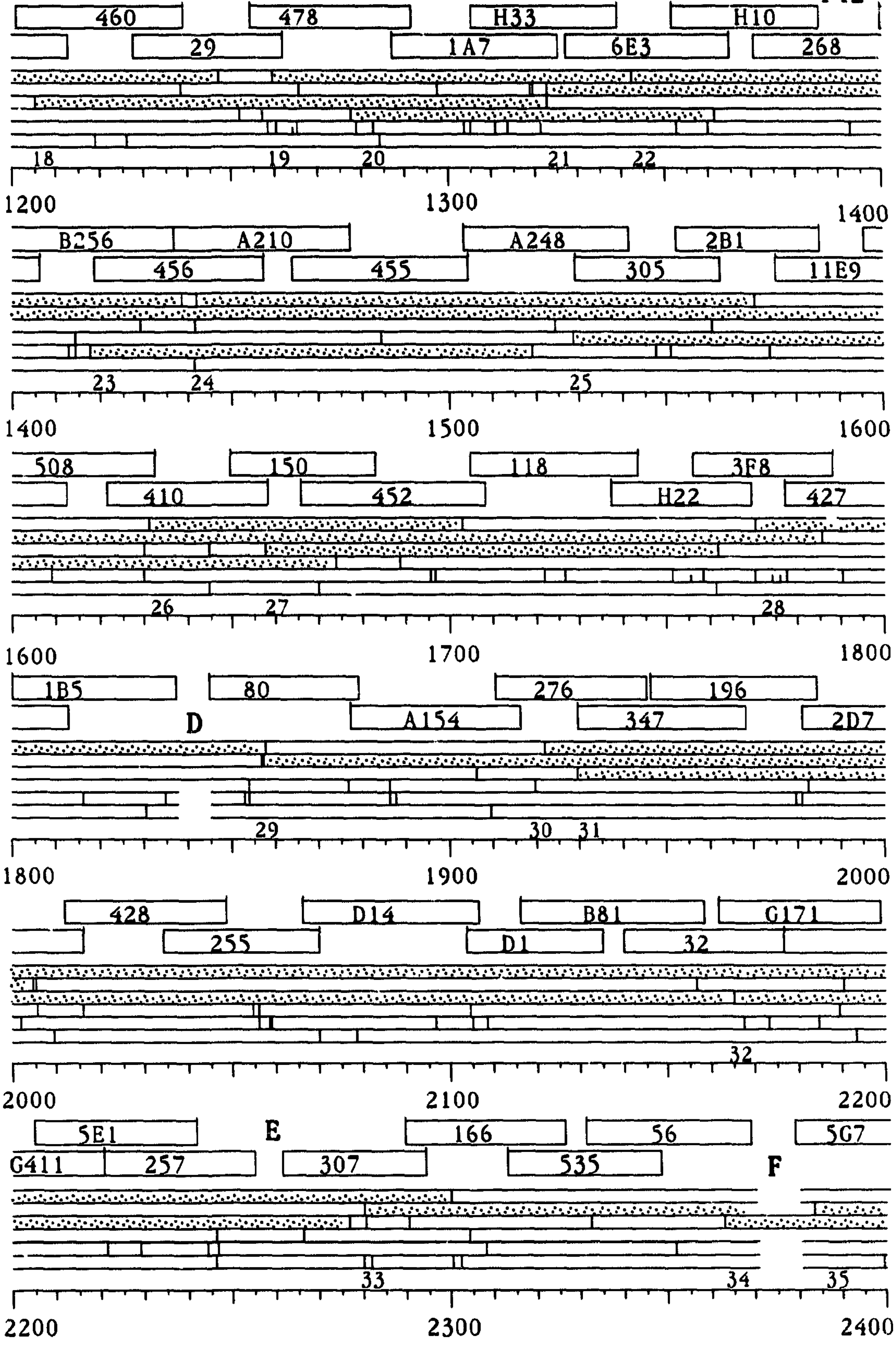
The nearly complete restriction map of the entire *Haloferax volcanii* genome is presented in figure 15. Figure 15 is a summary of all of the work described above. The verified fragments enumerated in table 7 are indicated by shading, and the numbers are given below

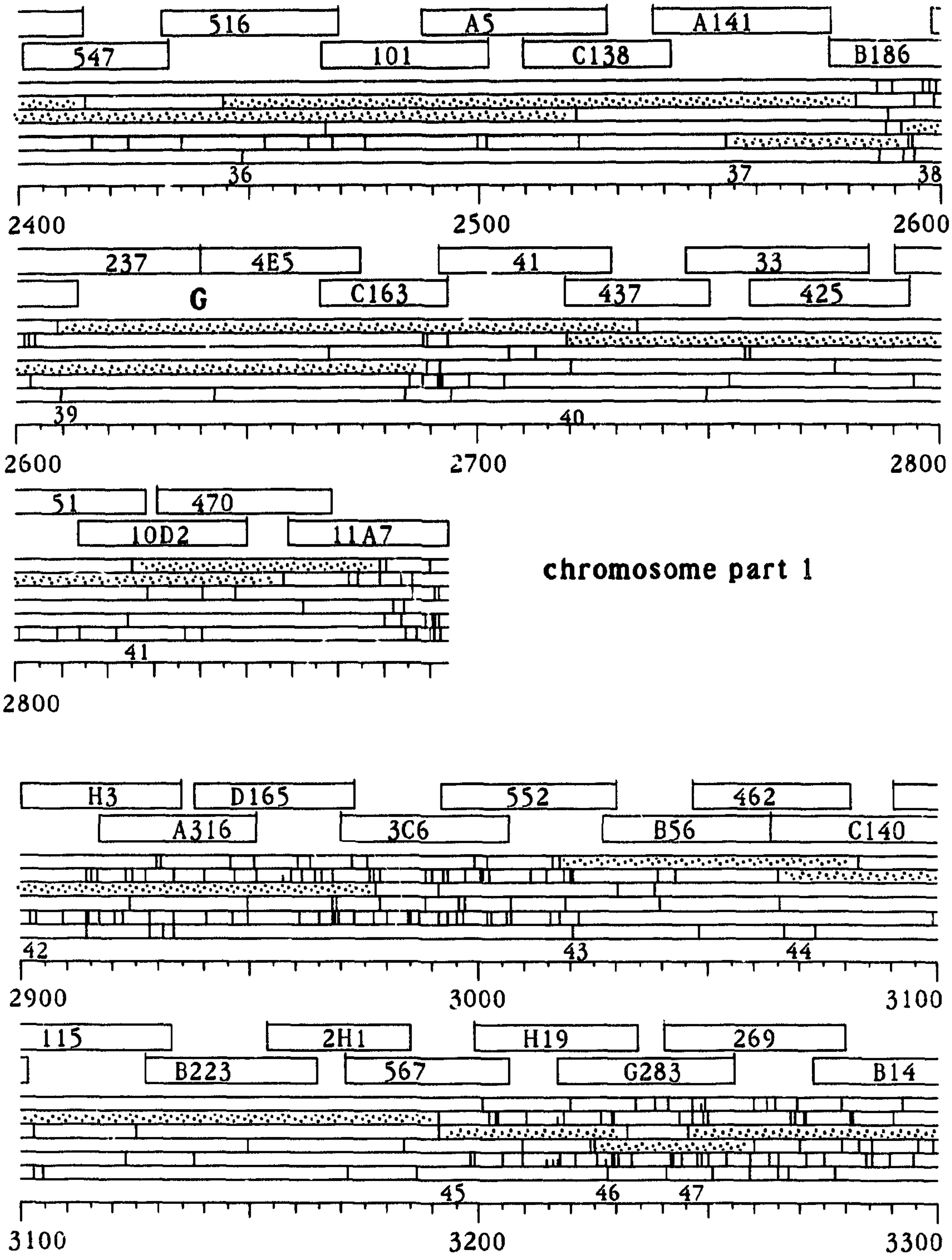
Figure 15. (Next four pages) Nearly complete restriction map of the *Haloferax volcanii* chromosome. Each cosmid is represented by a box with the orientation indicated by a tick at the HindIII end. Each restriction site is represented with a tick. From top to bottom, the enzymes are:

BamHI
BglII
HindIII
PstI
SspI

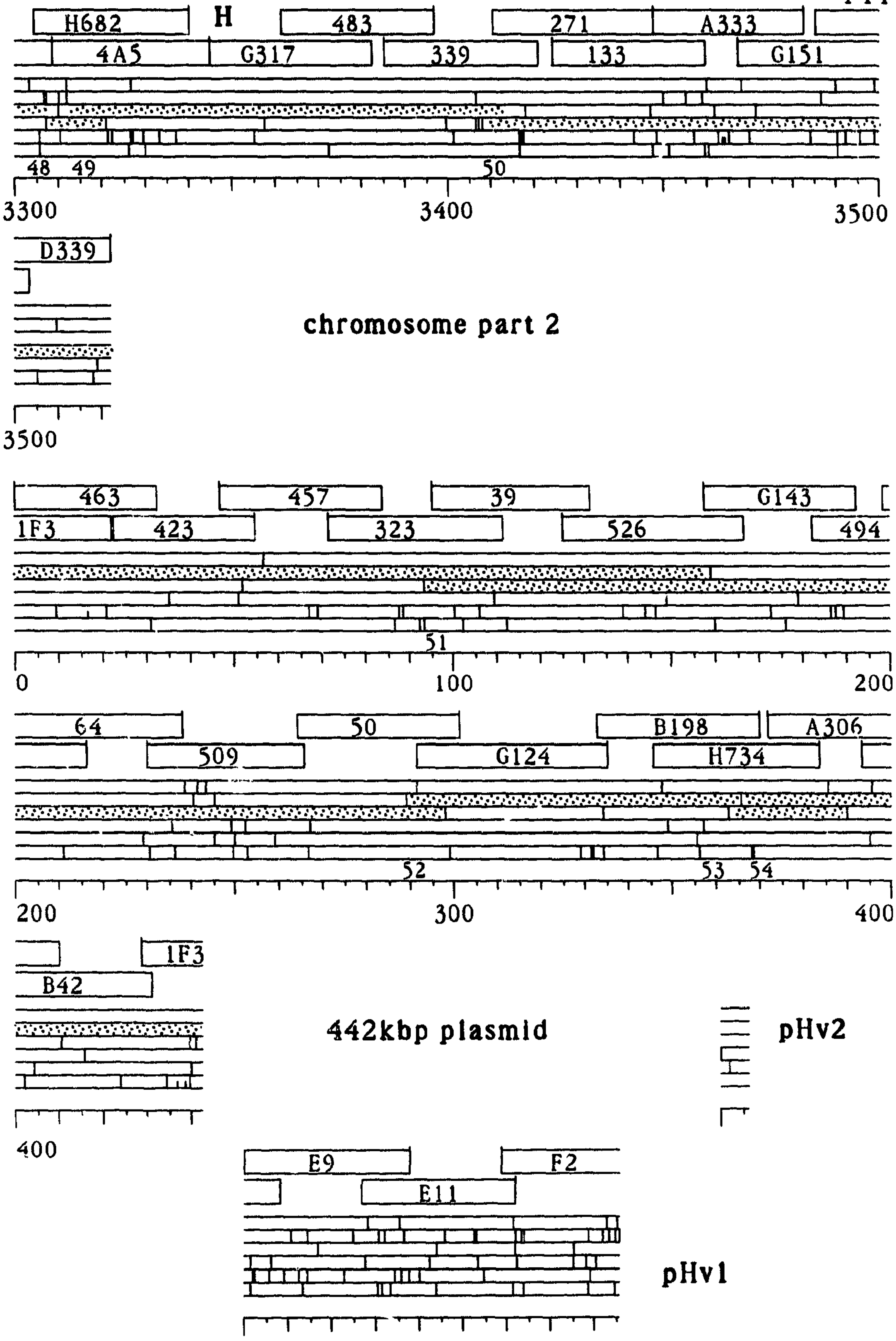
At the bottom is a cumulative kilobase scale. The map is drawn as if the two map fragments are joined in the most likely way, with H3 connected to 11A7. Stippling indicates fragments mentioned in table 9. The numbers immediately above the kilobasepair scale correspond to those in table 9.


G86
261
497
222
B340
C244
11C2
1
2
3
4
5
0
100
200
464
3B10
128
152
16
A
208
416
6
7
8
9
200
300
400
531
266
11B11
B
G203
D57
110
10
11
400
500
600
501
2C7
A78
B275
499
H11
B144
12
13
600
700
800
C
564
126
2E10
329
496
A99
14
15
16
800
900
1000
545
247
A199
G591
21
441
342
132
17
1000
1100
1200


460
478
H33
H10
29
1A7
6E3
268
18
19
20
21
22
1200
1300
1400
B256
A210
A248
2B1
456
455
305
11E9
23
24
25
1400
1500
1600
508
150
118
3F8
410
452
H22
427
26
27
28
1600
1700
1800
1B5
80
276
196
D
A154
347
2D7
29
30
31
1800
1900
2000
428
D14
B81
G171
255
D1
32
32
2000
2100
2200
5E1
E
166
56
5G7
G411
257
307
535
F
33
34
35
2200
2300
2400



chromosome part 1


H682
H
483
271
A333
4A5
G317
339
133
G151
48
49
50
3300
3400
3500
D339
chromosome part 2
3500
463
457
39
G143
1F3
423
323
526
494
51
0
100
200
64
50
B198
A306
509
G124
H734
52
53
54
200
300
400
1F3
B42
442kbp plasmid
pHv2
400
E9
F2
E11
pHv1

the kilobasepair scale. The gaps A...H are also indicated below the kbp scale. The two fragments of the map of the chromosome are assembled with 11A7 adjacent to H3. This is the way they are most likely to fit together, since the region near the end of 11A7 resembles H3, with a high density of sites for the mapping enzymes.

IV. Stability

The *Haloferax volcanii* strain used for the bottom-up map was received from the German Strain Collection (DSM) in 1987, and the culture from which DNA was prepared for the cosmid libraries was derived from it through a minimal number of passages. Such a short culture history would have been important if the frequency of rearrangement in this genome were even remotely approaching that of *Halobacterium halobium* (Sapienza *et al*., 1982). High frequencies of rearrangement were not expected in this species, but the high number of copies of at least one insertion sequence element (ISH51 (Hofman *et al*., 1986)), of which one transposition has been serendipitously observed (Lam and Doolittle, 1989b), indicate that it is a possibility.

The top-down mapping work was done with an *Haloferax volcanii* stock which had been cultivated in this laboratory since 1983. This was obtained from the laboratory of C. Woese, and also derives from the original Dead Sea isolate DS2, deposited by Helge Larsen in the National Collection of Marine Bacteria (NCMB, Scotland) in 1975. It was thus possible to compare strains with many years of separate cultivation in order to test for gross rearrangements and for transpositions of ISH51. The histories of these strains are summarized in figure 16. The strain WFD11, which has been cured of the plasmid pHv2 by a treatment with ethidium bromide, is also included in the comparison. WFD11 is used as the host of the *Escherichia coli-Haloferax volcanii* shuttle vector (Lam and Doolittle, 1989b).

Comparison of restriction patterns of these strains using BamHI, BglII and DraI will detect gross rearrangements that may have happened since the common origin of these strains. This comparison is shown in figure 17. Very little difference is apparent. A 130 kbp BamHI fragment of WFD18 is replaced by a 150 kbp fragment, and a 160 kbp BglII fragment is present in WFD 7 and 11, apparently at the expense of the second-largest BglII. These are consistent with a single event in the 1000 kbp region (figure 15).

The restriction patterns obtained using with more frequent cutting enzymes would be a more sensitive indicator of small-scale rearrangements, if one could compare them in the same detail as those discussed above. The number of fragments precludes fragment-by-fragment comparison, but the subset of fragments most likely to be involved in rearrangements can be viewed by hybridization with the most common insertion sequence element, ISH51. Such a hybridization is presented in figure 18. Once again, there is at least one difference, but the overall pattern is remarkably similar. This is the opposite of the result seen in *Halobacterium halobium* (Sapienza *et al*., 1982). That species contains an IS element closely related to ISH51, as well as many other IS-elements.

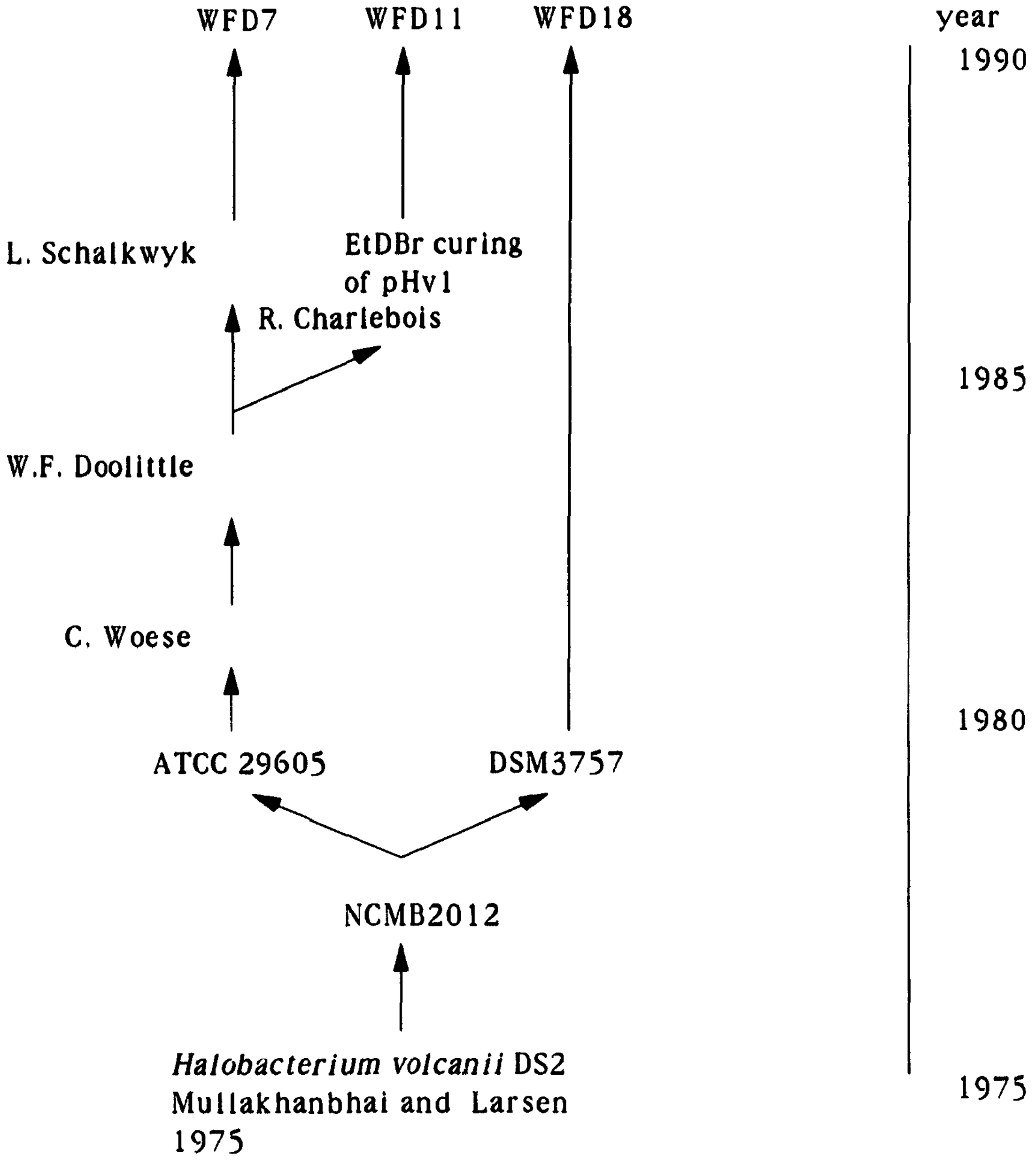


Figure 16. History of the strains WFD7, 11, and 18. WFD 7 and 18 have been cultivated separately at least since 1983. WFD 7 and 11 were established as frozen stocks in 1988, as was WFD18, but for these experiments the latter was obtained from the original DSM3757 stock, which had been kept at room temperature in the dark since it was received in 1987.

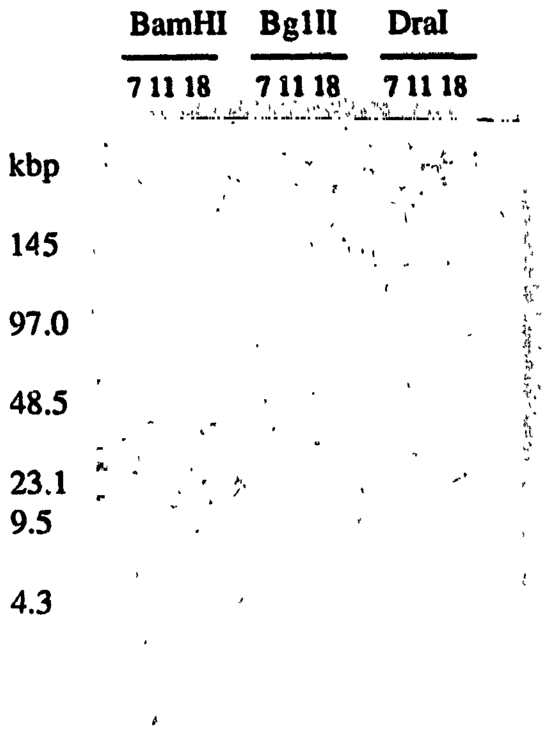


Figure 17. Comparison of BamH1, BglII and DraI digests of the DNA of WFD7, 11 and 18, by CHEF electrophoresis (5 second switching, 18 h).

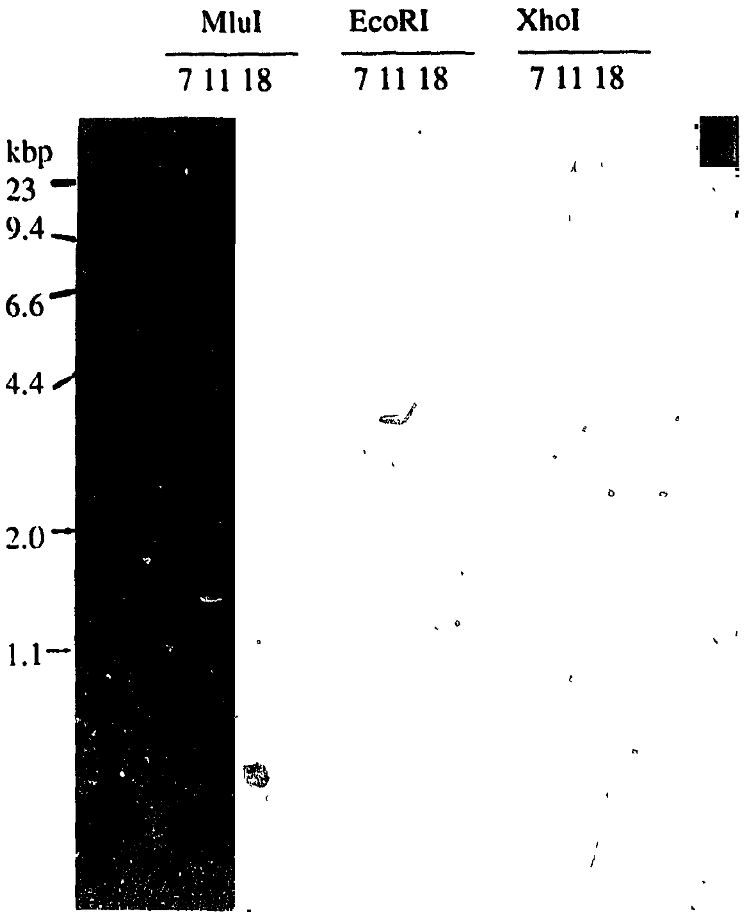


Figure 18. Southern transfer of the DNA of *Haloferax volcanii* WFD7, 11, and 18 digested with MluI, EcoRI and XhoI, and probed with the ISH51-containing clone p7x4.1.

V. Bulk markers

When this work was begun, the only genes that had been cloned from *Haloferax volcanii* were for ribosomal (Gupta *et al.*, 1983) and transfer RNAs (Daniels *et al.*, 1985b). There are now several protein coding genes, but the total is still too small to allow conclusions to be made about overall genome organization. I expected that the genome would be segregated into regions rich in insertion sequence elements (similar to the AT-rich island in *Halobacterium halobium* characterized by Pfeifer and Betlach [1985]) and more stable regions containing most of the genes (similar to the Pst-poor regions in *Halobacterium halobium* [Sapienza and Doolittle, 1982b]). This hypothesis could be tested by mapping the locations of two classes of markers which could be harvested in bulk, IS elements and tRNA genes.

The most common repeated sequence in *Haloferax volcanii* DNA is ISH51. Two copies of this element have been sequenced (Hofman *et al.*, 1986), revealing a structure similar to that of IS elements in eubacteria. I probed two sets of six large Southern blots containing MluI digests of the minimal set of cosmids with the "left" and "right' halves of ISH51 (as defined by Hofman *et al.* [1986]). One of the six gels and the two hybridizations of it are shown in figures 19, 20 and 21. Most (probably all) copies of ISH51 have one or more MluI sites, so that in almost all cases this allows the number of ISH51 elements at a locus to be determined. Our original plan was to produce a

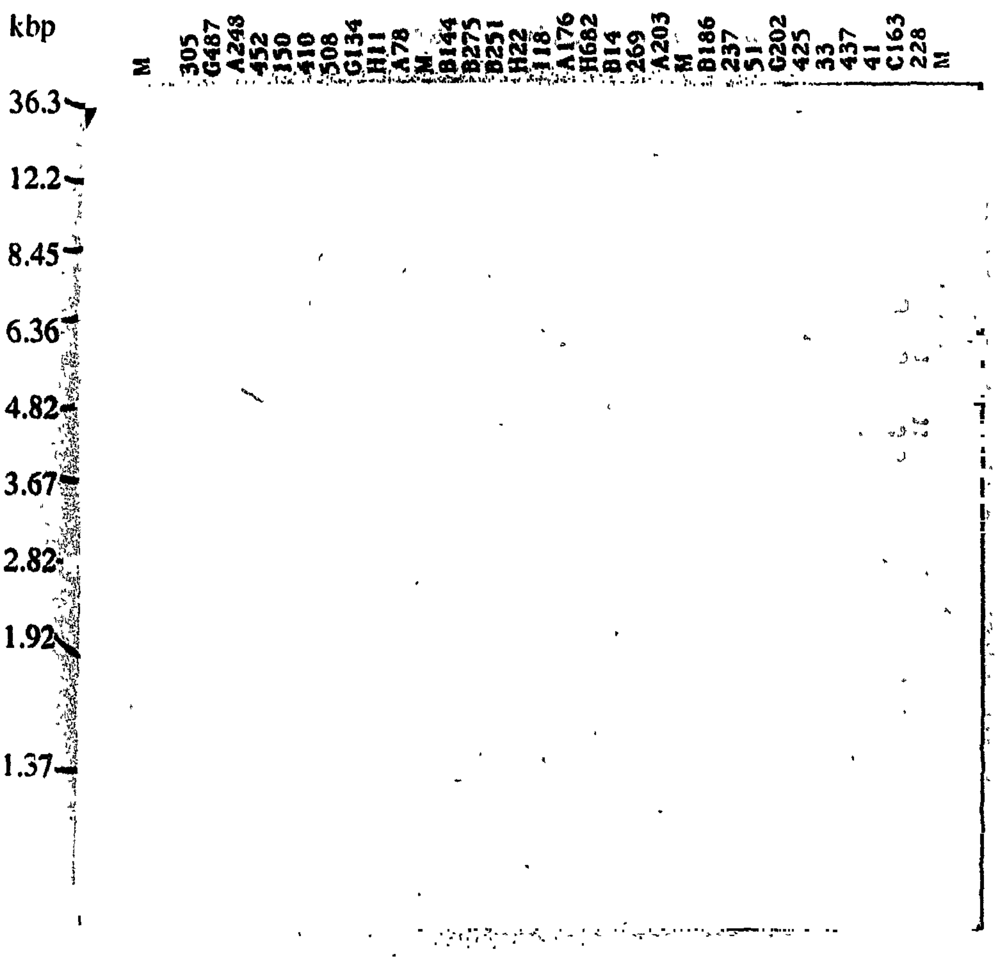


Figure 19. Photograph of the agarose gel used in figures 20 to 23. This is one of six gels containing the minimal set of cosmids, digested with MluI. The cosmids are named at the top. M: molecular weight markers. The outside lanes contain MluI-digested *Haloferax volcanii* DNA.

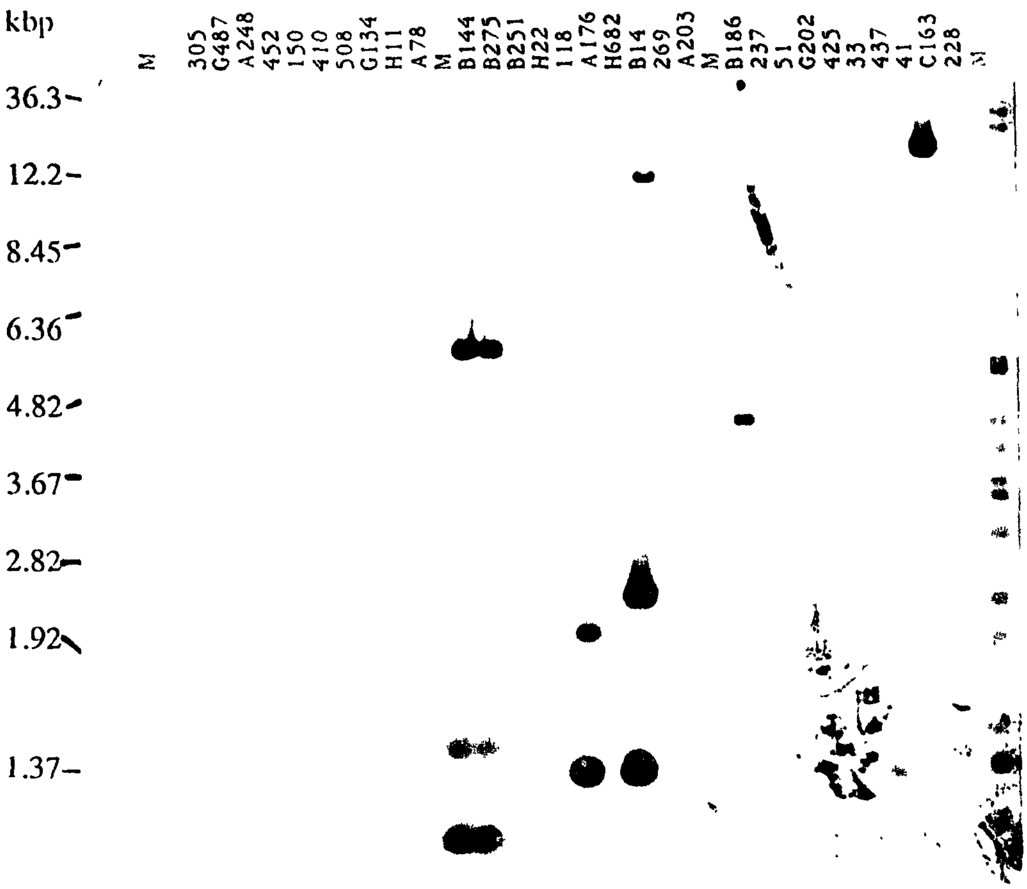


Figure 20. Mapping of ISH51 elements, part I. A blot of the gel shown in figure 19 was probed with the "left half" of ISH51.

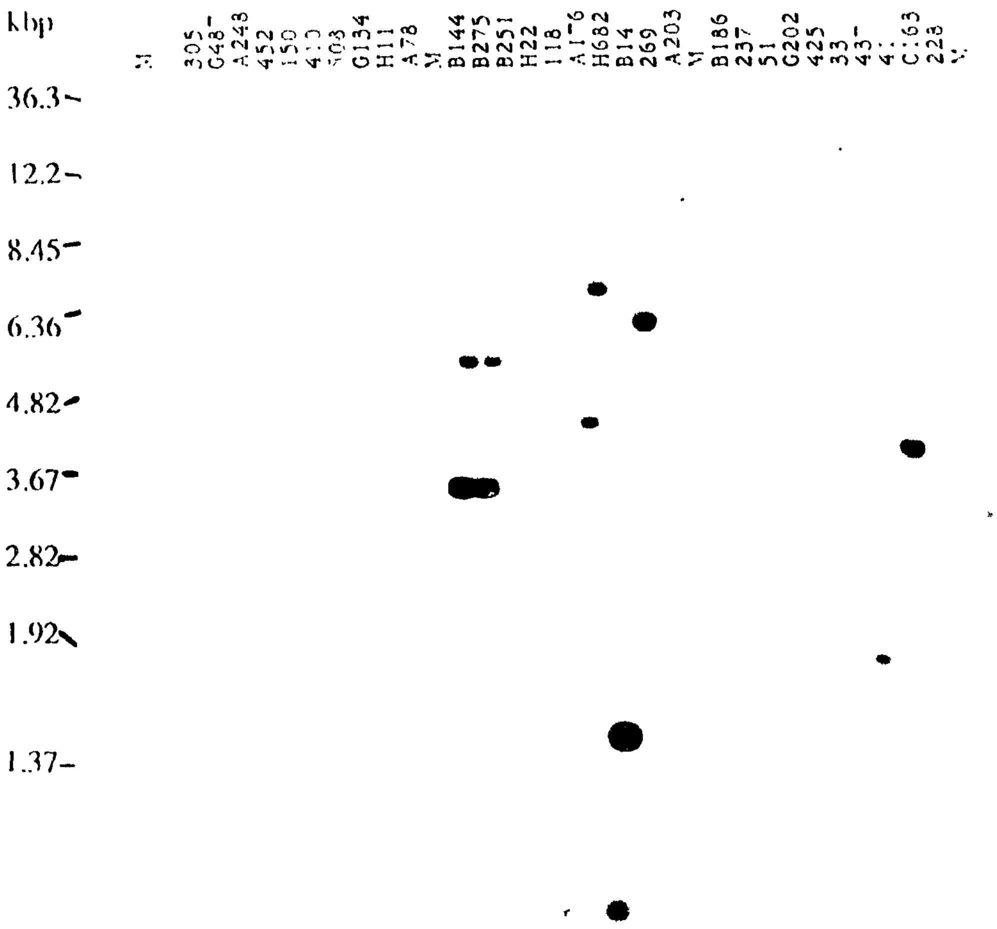


Figure 21. Mapping of ISH51 elements, part II. A blot of the gel shown in figure 19 was probed with the "right half" of ISH51.

complete MluI map, which would have allowed extremely precise placement of the ISH1 elements with the MluI data. We have abandoned that goal, but the MluI data still locates the ISH51 copies to intervals defined by the overlaps between cosmids, which are about 10 kbp or 0.25% of the map. The ISH51 data are given in table 10 and locations of ISH51 elements are plotted on figures 25 and 40.

Nine tRNA genes have been cloned and sequenced to date (Daniels *et al.*, 1986; Datta *et al.*, 1989, R. Gupta pers. comm.) There must be at least another 22 tRNA genes to make a complete set. tRNAs can be labelled to a high specific activity by ligation of [5'-^{32}P] cytidine 5'-3' bisphosphate to the 3' end with bacteriophage T4 RNA ligase. The total tRNA can be labelled *in vitro* in this manner, isolated and used as a mixed probe. I used such a probe mixture on the same blots used above for the mapping of IS elements. Positive fragments must contain either tRNA genes or pseudogenes (of which there is an example known in *Haloferax volcanii*, (Daniels *et al.*, 1986). The probe is contaminated with some degraded 7S and ribosomal RNA fragments, but the locations of the genes for these have been separately determined (see next section), and so signals from these RNAs will not be confused with tRNA genes. The tRNA population of *Haloferax volcanii* has been extensively characterized (Gupta, 1984; Gupta, 1985; Gupta, 1986) with every abundant RNA in the 4S region having been characterized. All of these proved to be tRNAs, so that I am quite sure that the probe is specific for tRNA and tRNA-related sequences.

clone	ISH51L fragments	ISH51R fragments	other
1B5	3075	3075,1394,sm	
3B10	1157, 712	9187	
3C6	3133, 1752, 1206	3133,1752,1206,2664,1513,955	
4E5		2451	
4A5	1159/1162, 997, 812, sm	6479,1548,1159	
11A7	6866,4903,692	6866,4903,692	
11B1	5352, sm	3320	
41	3865, sm	1857, sm	D
51	13215?		
56		6118	D
64	2783		
16	3040		
208	1136, 2sm	1136, 2sm	
237	2446		D
233	2079		
269	1388, sm	5671, 1388	D,E
271			D
456	6451	1540	
461		2333/2436,1834	D, E
462	848, 655		
509	2766, sm		D
530			D,E
552	2400, 1714		
564		4566	D, E
567	sm		
576			D
222	3056	7259	
488	5357	3292	D
496		4564	
A159	5357	3292	
A210	6454	1542/1543	

table continues overleaf

clone	ISH51L fragments	ISH51R fragments	other
A316	2313, 2sm	3165, 10539, 2313, 2sm	7bx
A333	2082, 1759, 689	6870, 1104/1099	D
B14	2582, 1371(2x)		D, E
B56	2322/2422, 852?, 666?		
B198	2872(820)		
B251	3184, 846, 2sm	4885, 2sm	E
B275	3186, 852, 2sm	4896, 2sm	E
C163	6639, sm	3838/3836, sm	
D165	2610	2610	7bx
G60		934/935	
G151	1772		
G202	13215		
G317	916, sm	1578	
H3	4761, 2311, 2sm	4761,4077,10603,10431,2311,2sm	
H680	4644, 2311	4644, 4080, 1066	
H632	1206, 791, 680?	6505, 3944	
H734	2869, 809		

Table 10. Locations of ISH51 and other repeated elements. MluI fragments hybridizing with each half of ISH51 are listed. D and E are less abundant repeated sequences first found in pHv1. 7bx is the location of two copies of ISH51 which have been sequenced (Hofman *et al.* 1985). The left and right probes for ISH51 were, respectively, 717 and 745 bp fragments from a XhoI/EcoRI digest of p7x4.1 (Hofman *et al.* 1986).

A hybridization of total tRNA with a dot blot of the first cosmid bank had made it clear that tRNA genes are thinly sown in the genome, since about half of the cosmids are positive. Hybridization of tRNAs with the MluI Southern blots bore this out (figures 19, 22, and 23). Forty five positive bands of widely varying intensity were identified, including bands corresponding to all of the previously cloned tRNA genes from this species except for tRNA^{met} (elongator) (Datta *et al.*, 1989). The fragments are listed in table 11 and shown on the summary map figure 39. These fragments may contain more than one tRNA gene, but most of the variation in intensity is due to the different abundances of different tRNA species. I probed separately with two gel-purified fractions corresponding roughly to class I and II tRNAs. The sensitivity of the hybridization to Southern blots is better than that of dot blot hybridizations because a similar or larger (molar) amount of the target sequence is present in the smaller area of a band, and is encumbered with less irrelevant DNA, so the background is low. Nevertheless, it is likely that some of the genes for less-abundant tRNA species will have been missed. Some genes may also reside on MluI fragments smaller than the lower limit of these gels, which is about 300 bp.

VI. Satellite fraction

Inspection of the tRNA and ISH51 results indicates that the tRNA genes and IS elements do indeed reside in different neighborhoods. It seemed likely that these correspond to the fractions of

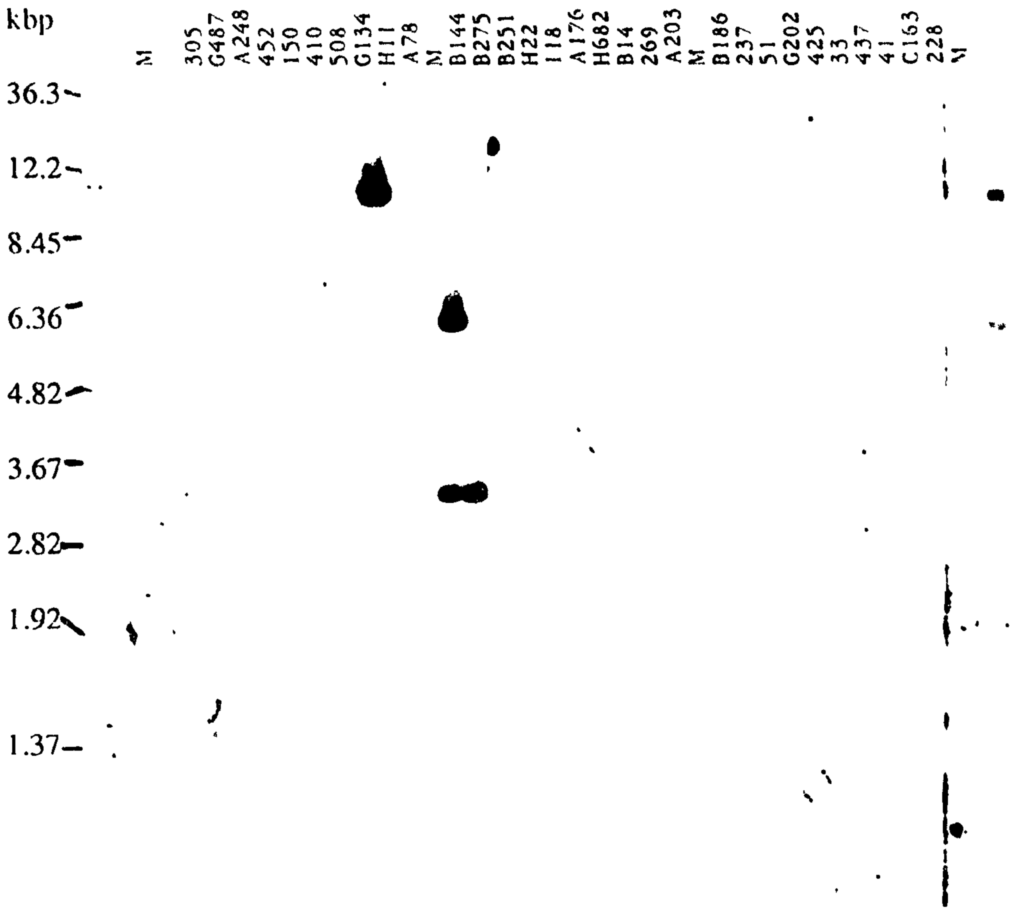


Figure 22. Mapping of anonymous tRNA genes, part I. A blot of the gel shown in figure 19 was hybridized with the faster-migrating tRNAs.

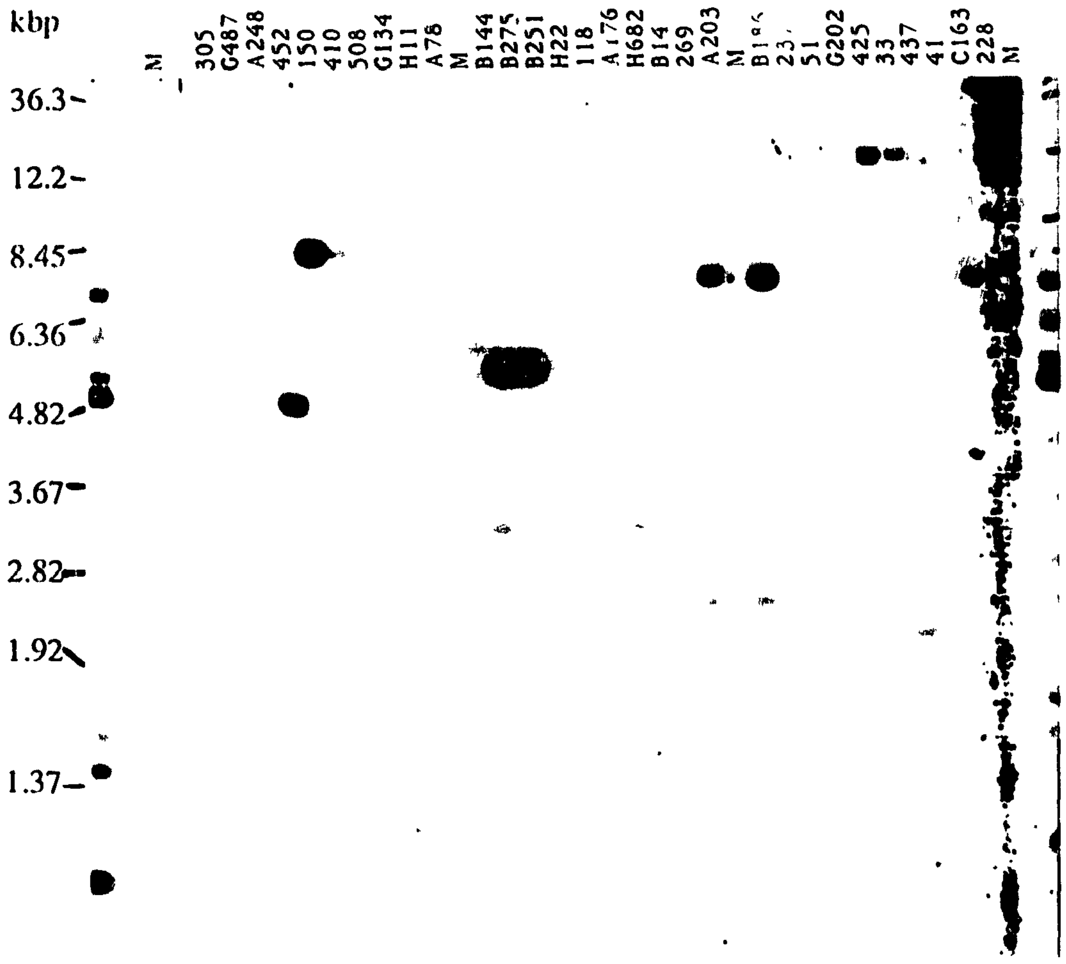


Figure 23. Mapping of anonymous tRNA genes part, II. A blot of the gel shown in figure 19 was probed with the slower-migrating tRNAs.

cos	tRNA fragments	other	cos	tRNA fragments	other
1A7	15622, 3727, 947		425	12714	
2B1	5472		427	2932	
2D7	9606		437	2049	tRNA thr,gly
2E10	5436,4528,2603		452	4549	*trp*B
3C6	1513955		456	2317	
3F8	2937		460	6499, 5264	
4B10	7107, 5623		461		*sod*B
6B1	2909/2915		470	6239	
10E8	6632,2912, 1781		488	12276	
11B1	11812		496	4564, 5554	rrnB
11C2	4603, 2553		508	7718	tRNAimet
21	5496	*arg*	531	5583	tRNAmet 1264
29	6510, 5299		564	4566, 5595?	*sod*A
32	5910	*fol*A, tRNAtrp	A159	12276	
33	12853	tRNA ser	A203	6635, 2198	*rrn*A
38	1136/1124		B144	5304, 2905	*rpl*K
41	1978	D,tRNAthr, gly	B186	6628, 2196	*rrn*A
51	6358		B251	4885	
110		*mev*	B256	2497	
150	4549		B275	4896, 2913	
166	4761		C138	2356, 7579	tRNAlys
208	1407		C163	6639	
266	3151		D57	3145	
268	1360		D282	6219	
276		7S	G60	935	*rpo-tuf*
307	4779		G143	2382	
347		7S	G171	5812	*his*C, *dhf*, tRNAtrp
410	7681	tRNAimet	G411		*his*C
416	1124/1145		H11	9495	
			H37	1456	

Table 11. Locations of tRNA and other genes. The size of the MluI fragment inhabited by each of the putative tRNA genes is given in basepairs. The sizes are from the landmark data and are given to the nearest base pair in order to make them convenient, unique identifiers for the fragments.

differing G+C content which the DNA of *Haloferax volcanii* contains, like that of *Halobacterium halobium* (Pfeifer *et al.*, 1982). By analogy with *Halobacterium halobium*, the less GC-rich fraction (FII) was expected to be composeo of plasmids and regions interspersed in the genome (Pfeifer and Betlach, 1985). The compositions of the two *Haloferax volcanii* fractions are 66.5 mol% G+C (FI) and 55.3 mol% G+C (FII) (Ross and Grant, 1985).

The plasmid pHv1 was cloned using a crude alkaline (Birnboim and Doly, 1979) preparation of pHv1 DNA to probe the second cosmid library. Among the positive cosmids were found three which cover the entire sequence. On restriction mapping, these were found to be exceptionally rich in sites for the ten landmark enzymes (figure 15, figure 24). Other clones have been identified which are similarly rich in sites (compare, for example cosmid 269, figure 14). The genome as a whole contains fewer sites for these enzymes than might be expected from base composition (especially BamHI, BglII, HindIII, and PstI), but in this desert, oases of site-rich sequence can be found (Charlebois *et al.*, 1989b).

A plot of site frequency over the length of the maps of the chromosome and the two plasmids is shown in figure 25. Three well-defined site-rich regions can be seen, in which sites for the six enzymes are more than twice as frequent as in the genome as a whole. Because the sites being counted are infrequent, the frequency is calculated over a 50 kbp window, which is large with respect to the size of the oases and results in the pointed appearance of the peaks. While most of the lumpiness in the remainder of the

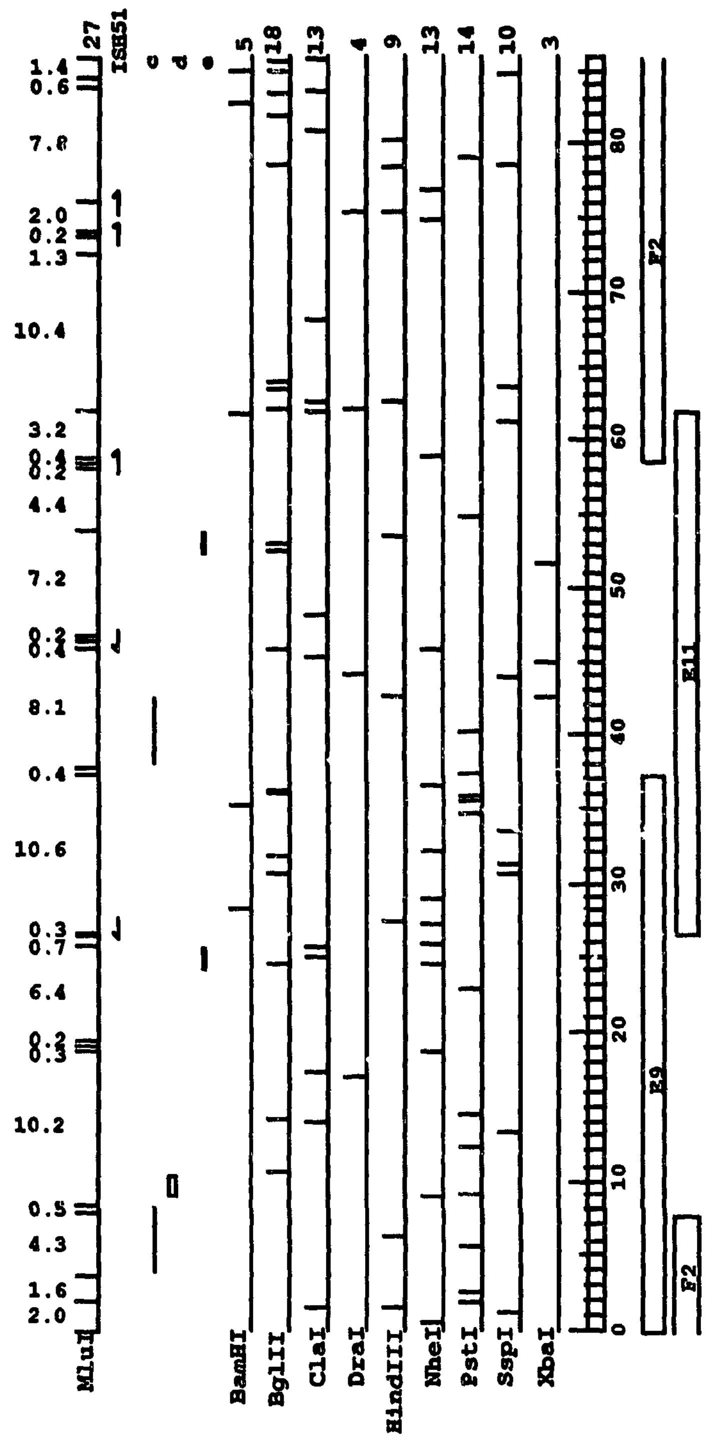


Figure 24. Detailed restriction map of the plasmid pHv1 and its IS elements.

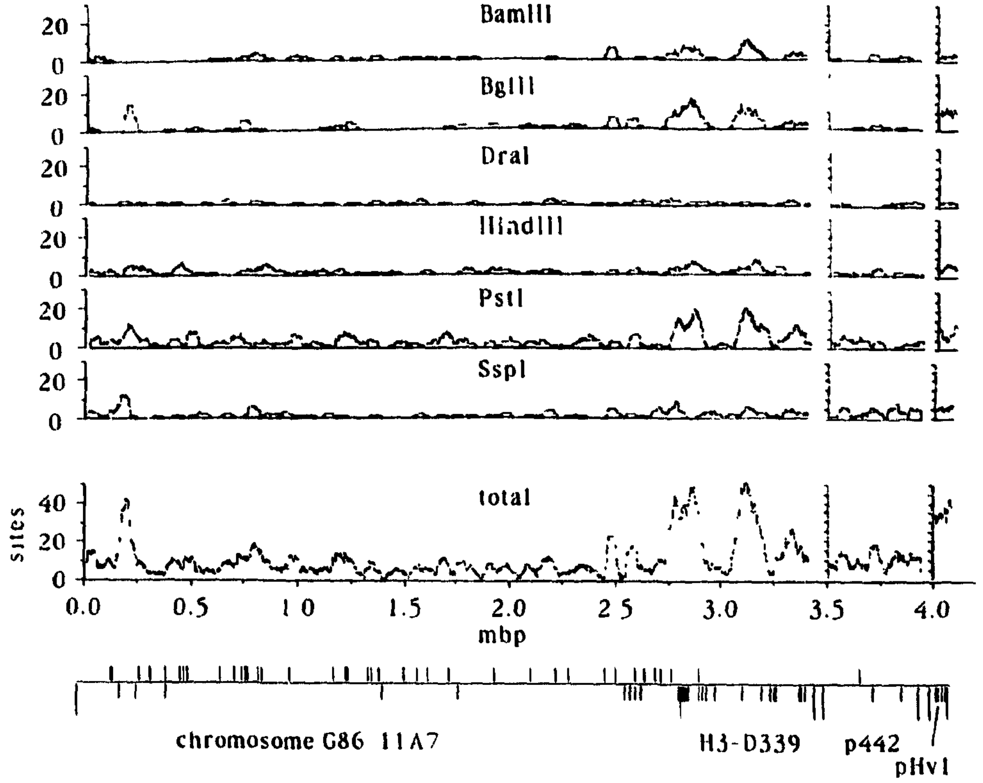


Figure 25. Plots of restriction site frequency. The number of sites for each enzyme was counted in a 50kbp window which was stepped along the chromosome 5kbp at a time. The contigs are assembled with the gaps set to 1kbp and cosmid H3 arbitrarily joined to cosmid 11A7. On the right are shown the megaplasmid and pHv1. At the bottom are plotted the positions of tRNA genes (top ticks) and ISH51 (bottom ticks). The long ticks on the bottom indicate the ends of the chromosomal map fragments and plasmid maps.

chromosome is nothing more than one might expect from a stochastic process, some of the features may be due to particular local constraints on nucleotide and oligonucleotide composition. For example, the blips in site frequency at 0.6 and 2.5 mbp correspond to the two *rrn* operons and their flanking sequences.

Considering the enzymes separately, one can see that the distribution of DraI and SspI sites is much more uniform than the rest. These two enzymes differ from the others in having recognition sequences containing only A and T, which must contribute to their rarity. DraI and SspI would each cut random sequence of 66% G+C at intervals of about 6^6bp or 50 kbp, which is approximately what is seen.

To determine the composition of pHv1, I prepared a mixture of the three pHv1 clones, digested with MluI and end-labelled with ^{32}P, and mixed this with about 100 fold the quantity of MluI-digested *Haloferax volcanii* DNA. The chromatography of this mixture on malachite-green bisacrylamide is shown in figure 26. The ^{32}P-labelled peak corresponds with the FII shoulder in the genomic DNA, demonstrating that the overall composition of pHv1 is FII. The agarose gel of the fractions from this column shows that the compositions of the fragments differ somewhat. The smallest fragments are probably more AT-rich than their position implies, because malachite-green bisacrylamide chromatography has a small size-bias.

In order to determine whether FII DNA is site-rich, I prepared randomly sheared, ^{32}P-labelled DNA and separated FI and FII fractions (figure 27A) These fractions, along with a sample of the

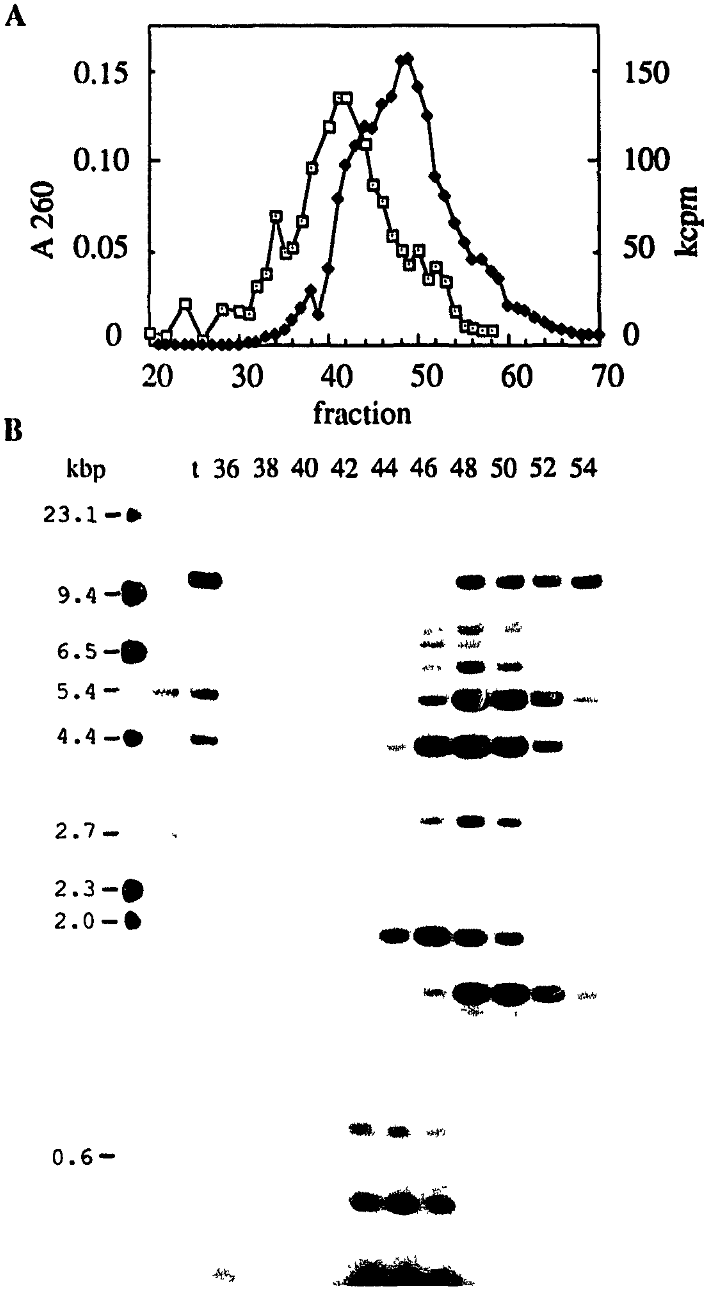


Figure 26. Malachite green bisacrylamide chromatography of pHv1. A. MluI-digested *Haloferax volcanii* DNA and a much smaller quantity of ^{32}P-labelled, MluI-digested pHv1 DNA derived from cosmids separated on the column. Open squares, absorbance; closed, radioactivity. B. An autoradiogram of a 1% agarose gel on which samples of the fractions have been separated.

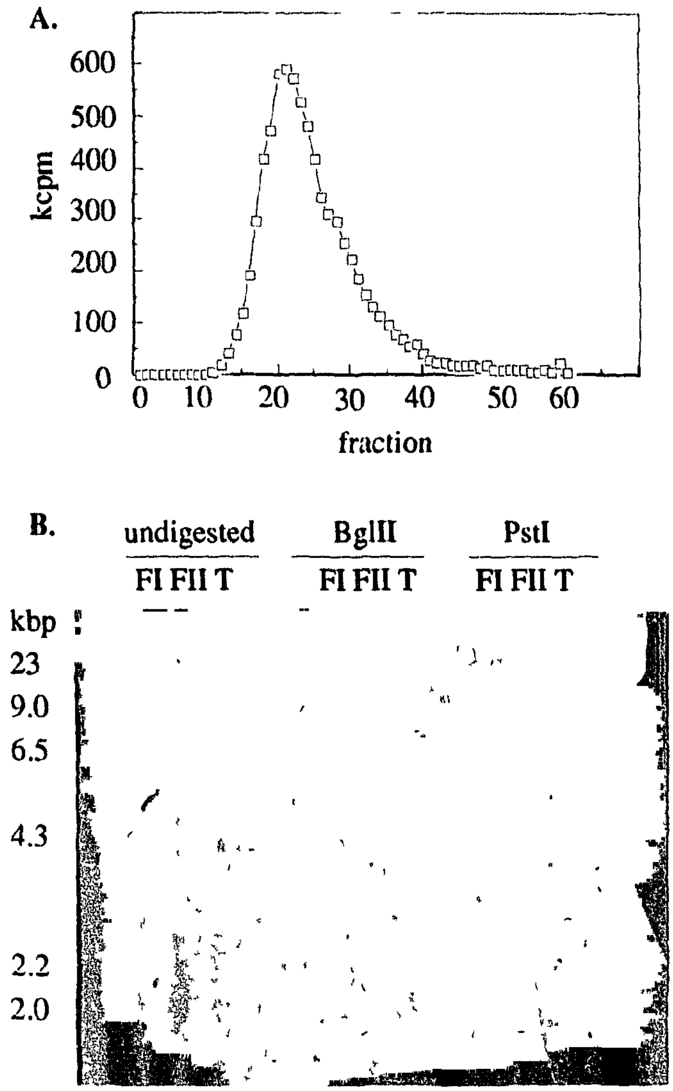


Figure 27. Digestion of nick-translated FI and FII DNA with infrequent-cutting restriction enzymes.

A. Malachite-green-bisacrylamide chromatography of nick-translated DNA.

B. Fractions 20 (FI) and 28 (FII) from above and the starting material (T) digested as indicated and separated by 1% agarose gel electrophoresis.

starting material, were digested with BglII or PstI. An autoradiogram of these samples separated on a 1% agarose gel is shown in figure 27B. The trend of more extensive digestion of FII than FI is visible, even though by this point the DNA had been through many manipulations and was not of very high molecular weight before digestion. The correlation between FII and oasis sequences can be explored further by probing fractionated DNA with representative probes. tRNAs and rRNA hybridize only with FI DNA, as expected (figure 28A and C),and ISH51 hybridizes with both (figure 28B). Hybridization of an ISH51 probe to a Southern transfer of MluI-digested fractions of *Haloferax volcanii* DNA (figure 29 panel 2) shows that different copies of ISH51 are present in the two fractions, though most of the copies are in FII DNA. Each ISH51 location that can be recognized by its MluI fragment size can thus be identified as FI or FII sequence.

pHv1 seemed to be a likely home for insertion sequences, so I searched it for repeated sequences. I did this by isolating each of the MluI fragments of the clones of pHv1 for use as a probe on a Southern transfer~~s~~ of MluI-digested *Haloferax volcanii* DNA. The result is shown in figure 30, in which each lane is a separate probing. It is striking how few of the fragments are unique. In part this is due to the fact that most ISH51 copies have an MluI site, so that most ISH51 elements (all of the ones in pHv1) contaminate two fragments. In all, five different patterns can be discerned in figure 30, which are listed in the legend and designated A through E. A and B are the two halves of ISH51, as could be recognized by comparison with the

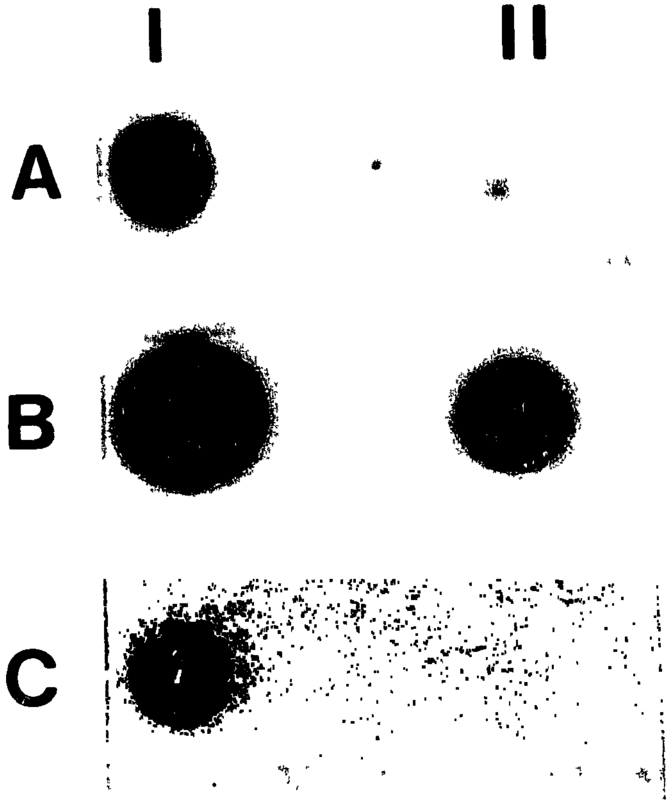


Figure 28. Dot blots of malachite green bisacrylamide-fractionated *Haloferax volcanii* DNA probed with

a) total end-labelled *Haloferax volcanii* tRNA

b) ISH51

c) pvt6 (5S RNA and $tRNA^{cys}$)

The fractions I (FI) and II (FII) were from a separation like that shown in figure 27, except that the starting material was not labelled.

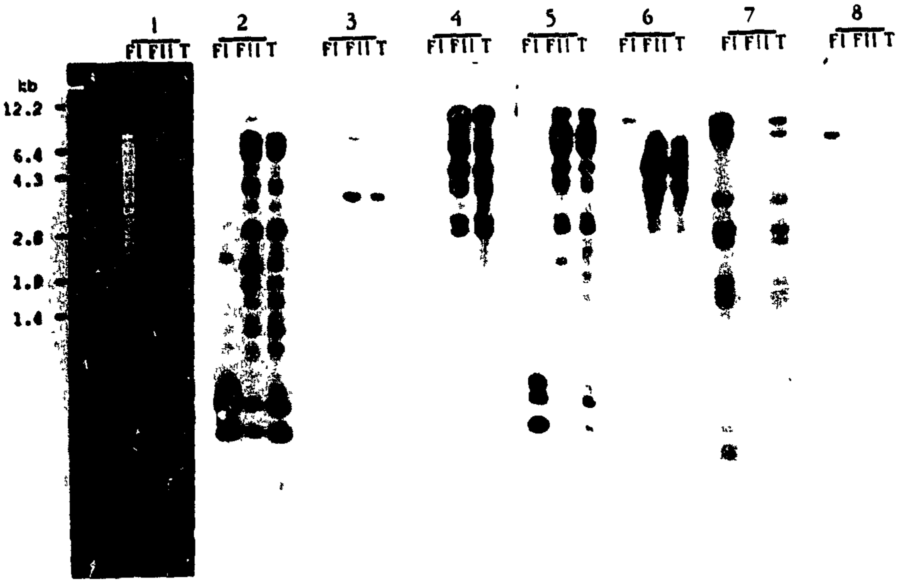


Figure 29. Hybridization of repeated and unique sequences with a Southern blot of MluI-digested FI, FII, and Total DNA (FI, FII, and T).

Panel 1. The ethidium bromide-stained gel
Panels 2 through 8, Southern transfer probed as follows:
2. ISH51
3. Element C
4. Element D
5. Element D, hybridized at a lower stringency
6. Element E
7. Cosmid 50
8. Cosmid 276

patterns produced by probings with subfragments of p7x4.1 (Hofman *et al.*, 1986). For each of the remaining repeated sequences I sought to isolate a copy from another context in order to be able to map the location of the plasmid-borne sequence in detail.

In the case of ISH51, a probe (p7x4.1) was already available (Hofman *et al.*, 1986). The positions of the ISH51 elements can also be deduced directly from figure 30, given the MluI map. Each of the copies of ISH51 has one or more MluI sites within it and the hybridization patterns produced by the two halves can be distinguished, so the orientation can also be deduced. The locations of the five ISH51 elements in pHv1 are given in figure 24.

No chromosomal copy of element C could be found. The two strong signals in the C pattern are both fragments of pHv1. Element D is the second most abundant repeated element after ISH51. The ISH51 pattern can be faintly seen in hybridizations using D elements from several locations as probes, and at low stringency, ISH51 and D hybridize to each other. D is thus likely to be a relative of ISH51. Eight strong signals are seen in the D pattern. I isolated a 3.4 kbp MluI fragment from cosmid 271, which corresponds to the smallest of the strongly bybridizing fragments seen in strip 12 of figure 30. Using the 3.4 kbp fragment as a probe, I was able to delimit the position of the D element on pHv1 as shown in figure 24. This kind of detailed mapping is too arduous to be done on the entire genome, so I used the entire 10.2 kbp MluI fragment of pHv1 as a probe on a dot blot of the minimal set of cosmids (figure 32). The blot used does not include clones of pHv1. The signals obtained vary in intensity.

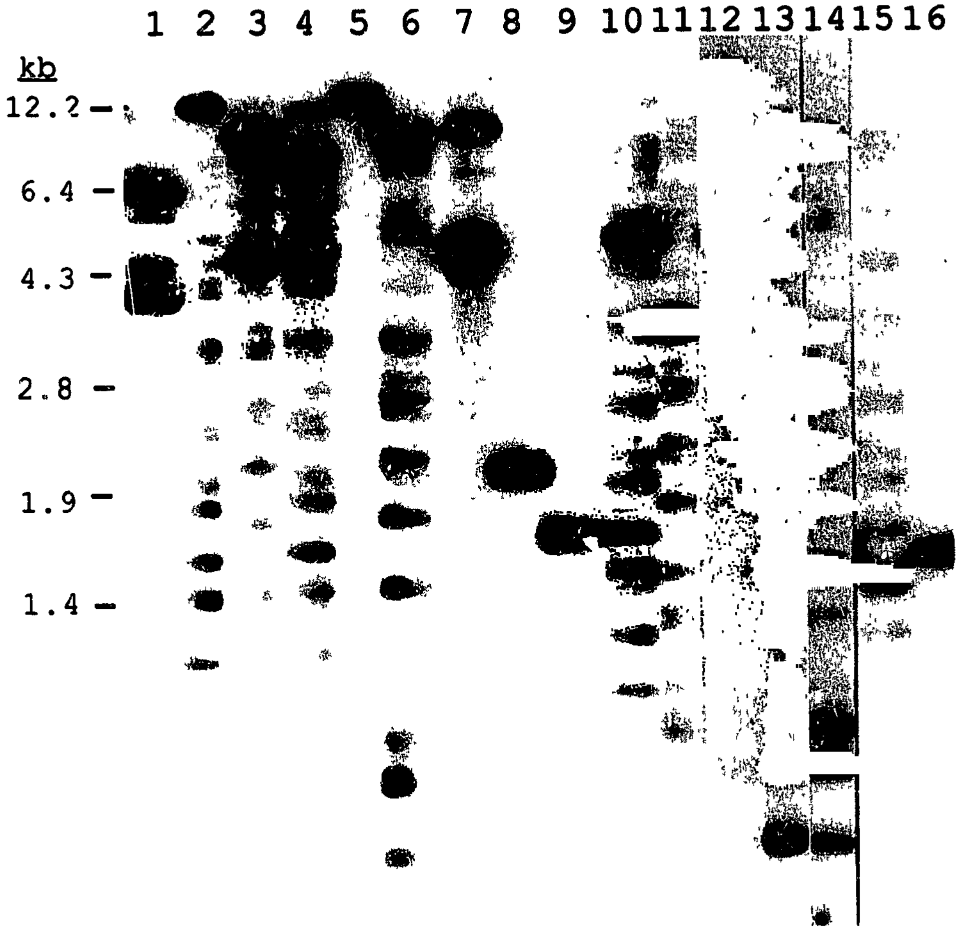


Figure 30. Search of pHv1 for repeated sequences. Each lane is a separate hybridization of an MluI fragment of pHv1 to a strip of a Southern transfer of MluI-digested *Haloferax volcanii* DNA. The fragments are:

strip	size	pattern	strip	size	pattern
1	6.4kbp	E	9	1.6	U
2	10.6	B	10	4.4	B
3	8.1	A, C	11	3.2	A
4	7.2	B, E	12	10.2	D
5	10.4	U	13	2.0	A
7	4.3	C	15	1.3	B
8	2.0	U	16	1.4	U

and the strongest are presented in table 10 and shown on figure 40. The strongly-hybridizing copies of element D are all in FII DNA (figure 29). The D element, like ISH51, hybridizes with the DNA of *Halobacteriu.r halobium* , but not that of *Haloferax mediterranei* or *Halobacterium* species GRB (figure 31). This suggests that ISH51 and D diverged before ISH27 (as the *Halobacterium halobium* element similar to ISH51 is known) and D were separated. The distribution of these two elements is remarkable because GRB and *Haloferax mediterranei* are thought to be more closely related to *Halobacterium halobium* and *Haloferax volcanii*, respectively than the latter two are to each other.

Element E is not present in other halobacterial species (figure 31). I used two subclones from the genome, corresponding to strong and weak bands of the E pattern, to map the copies on pHv1. Both gave the same result, which is diagrammed in figure 24. Hybridization of the more strongly hybridizing subclone to a dot blot of the minimal set allowed the mapping of 9 copies (figure 32), listed in table 10 and shown in figure 40. All but one of the most strongly hybridizing copies of element E are in FII DNA (figure 29).

As an example of a unique, non-oasis sequence, I probed the Southern blot of malachite green fractions with the entire cosmid 276. It is entirely FI DNA, as expected. The megaplasmid is not an oasis (figure 25), though one might have expected it to be FII DNA. A sample of the megaplasmid, cosmid B42, contains both FI and FII fragments (Figure 29, panels 7 and 8).

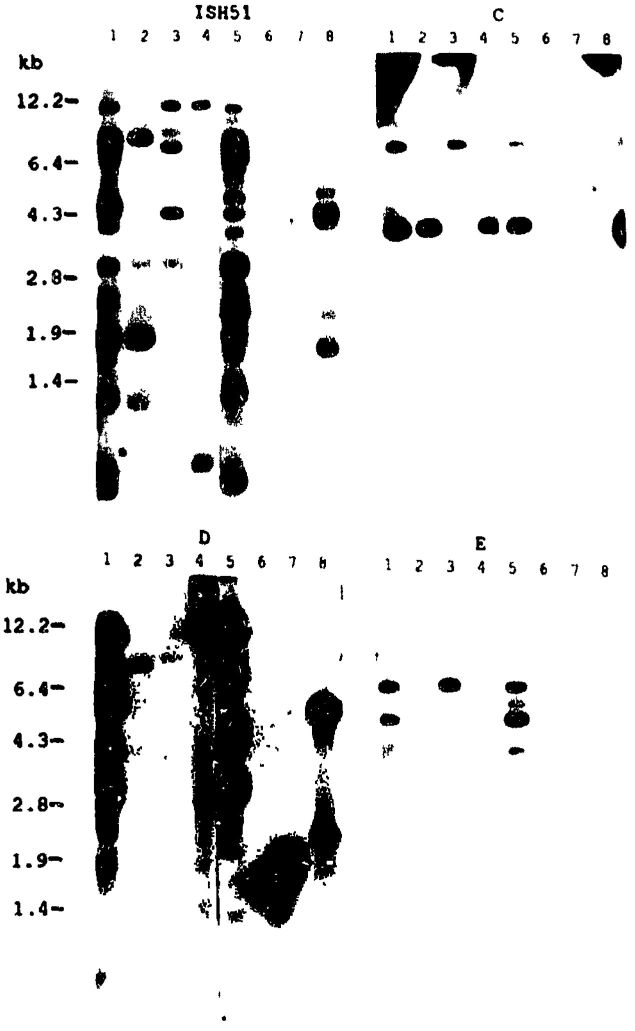


Figure 31. Hybridization of repeated sequence elements to DNAs of several halobacteria. All of the DNAs are digested with MluI.

Lanes: 1 Crude preparation of pHv1 from *Haloferax volcanii* WFD11
2. Cosmid F2 (pHv1)
3. Cosmid E11 (pHv1)
4. Cosmid E9 (pHv1)
5. *Haloferax volcanii*
6. *Haloferax mediterranei*
7 *Halobacterium* GRB
8. *Halobacterium halobium* NRC-1

The probe is mentioned at the top of each panel. ISH51-p7x4.1, C-4.3 kbp MluI fragment of pHv1, D-2.2kbp MluI-BglII fragment from pHv1, E-p13.2.

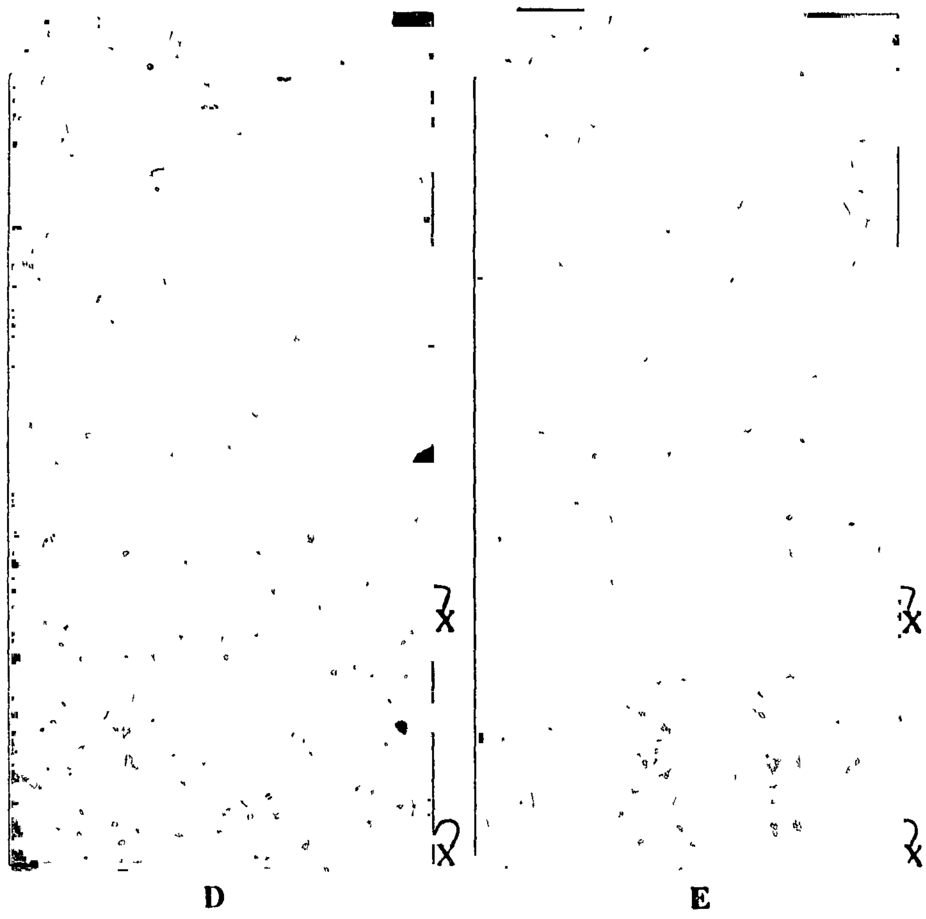


Figure 32. Dot-blot hybridization of repeats D and E to cosmid DNAs. The results are summarized in table 10. These are blots 5 and 4 of a series of twelve impressions of a 10 by 20 array of paper "sponges" containing cosmid DNA (described under Methods and Materials). At the bottom right of each filter are two alignment marks (X) and round spots encoding the filter number in *Haloferax volcanii* DNA.

VII. Cloned markers

A number of markers can now be added to the map by hybridization of cloned genes from *Haloferax volcanii* or other species to the minimal set of cosmids. A list of probes and the locations deduced from them is given in table 12. The nine tRNA genes cloned and sequenced so far from *Haloferax volcanii* were mapped by hybridization with dot blots of cosmid DNA, as in the examples in figure 33.

The tRNAcys clone pVt6, which also contains part of a 5S rRNA gene, also allowed mapping of the two ribosomal RNA operons (Daniels *et al.*, 1985b) whose exact location and orientation could then be recognized from their restriction sites (Charlebois *et al.*, 1989a)

The gene for the gyrase B subunit was mapped using a probe from the species *Haloferax* phenon K, closely related to *Haloferax volcanii* (Holmes and Dyall-Smith 1990). The gene has been used in the construction of a *Haloferax-Escherichia coli* shuttle vector, along with the origin of replication of the plasmid pHK2. Both *gyr*B and pHK2 sequences hybridize with *Haloferax volcanii* chromosomal locations at high stringency (figure 34).

It is also possible to transform auxotrophic *Haloferax volcanii* strains with genomic DNA (Cline *et al.*, 1989b), cosmid DNA or DNA fragments (Conover and Doolittle, 1990). This will allow the mapping of the genes for many anabolic enzymes. So far, this approach has been used for the isolation, mapping, and characterization of the *his*C

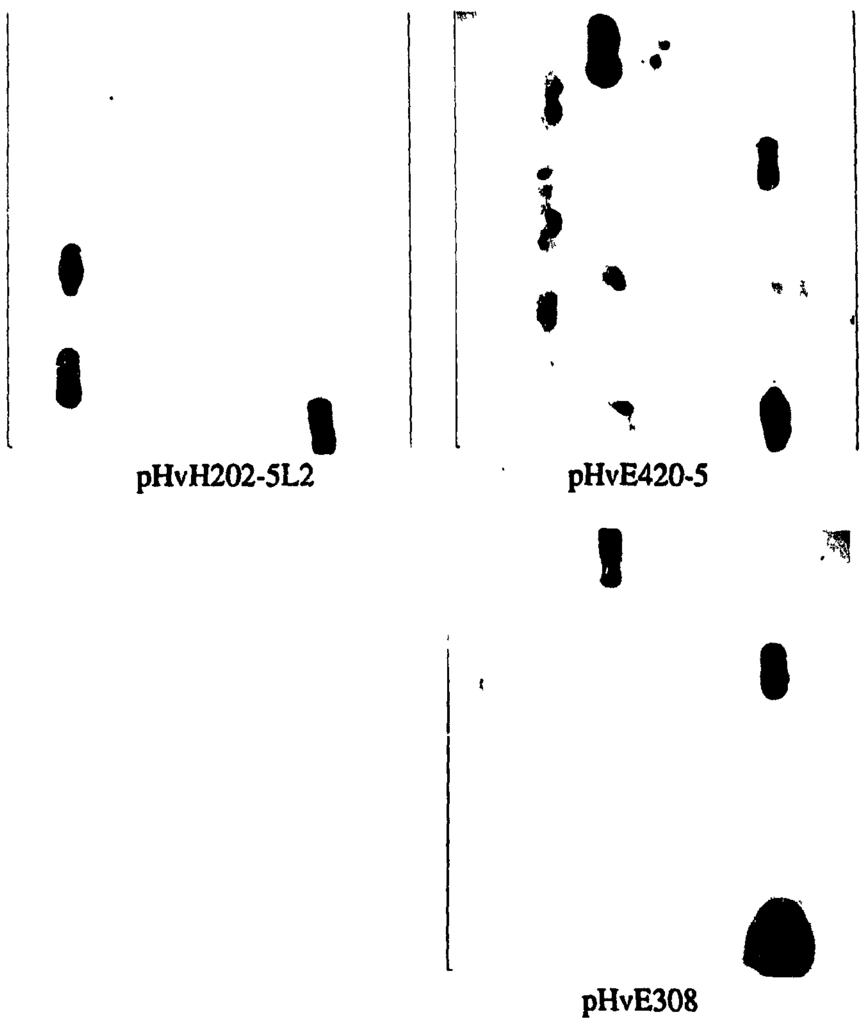


Figure 33. Dot-blot hybridization of cloned tRNA genes to cosmid DNAs. In each case the probe is the entire plasmid. The spot in the lower right of each blot is a pUC19 DNA control. Each plasmid hybridizes with two cosmids which overlap. ThetRNA^{met} (pHvH202-5L2) maps to the overlap of cosmids 531 and D57, and the other two, tRNA^{thr}(GGU) and tRNA^{gly}(CCC) both map to the overlap of cosmids 41 and 437.

PROBE	GENE	SOURCE	LOCATION
pVt1	tRNA imet	C. Daniels	410, 508
pVt2	tRNAval GAC	C. Daniels	488
pVt6	tRNAcys+5S	C. Daniels	496, A203, B186
pVt9	tRNA trp CCA	C. Daniels	32
pVt29	tRNAlys CTT	C. Daniels	C138
pVt38	tRNAser CGA	C. Daniels	33
pHvH202-5L2	tRNA emet	R. Gupta	531, D57
pHvE420-5	tRNA thrGGU	R. Gupta	437, 41
pHvE308	tRNA glyCCC	R. Gupta	437, 41
7S RNA	7S RNA	purified RNA	276,347
m13-DHFR	*fol*A	M. Mevarech	G171, 32
p7x4.1	ISH51	J. Hofman	D165, A316
pLW99	*rpl*L (*H.cut*)	P. P. Dennis	B144
p1kb1	*sod*A,B (*H. cut*)	P. P. Dennis	461,
m13Ba3.5	ORF75 (*H. halo*)	H. Leffers	G60
m13SE1.8	ORF75 (*H. halo*)	H. Leffers	G60
m13HB3.2	ORF75,B",B' (*H. halo*)	H. Leffers	G60
m13Bg2.3	B",A (*H. halo*)	H. Leffers	G60
m13Sst2.0	A,C (*H. halo*)	H. Leffers	G60
m13Sst2.4	r-proteins(*H. halo*)	H. Leffers	G60
m13SB4.5	Eftu-EfG(*H. halo*)	H. Leffers	G60
pMDS11	*gyr*B	M. Dyall-Smith	547,326

Table 12. Clones used for mapping markers. Genes are denoted by the names of their homologues from *Escherichia coli* (italic) or by their product. Those from other species are indicated with (*H. halo*) for *Halobacterium halobium* or (*H. cut*) for *Halobacterium cutirubrum.*

(Conover and Doolittle, 1990) , *trp*C,B, and A genes (Lam *et al.*, 1990) and a cluster of genes involved in arginine synthesis (K. Conover, pers. comm.) The genes are named to correspond with their homologues in *Escherichia coli*.

Probes derived from other species can also be used in low-stringency hybridization experiments to locate genes. I have mapped several protein-coding genes from the extreme halophiles (*Halobacterium* spp.), from which more genes have been cloned so far than from *Haloferax volcanii*. The probes and their sources are listed in table 12. Mapping of the superoxide dismutase (SOD) gene is shown in figure 35. Two independent parts of the *Hf. volcanii* genome hybridize with the *sod* gene from *Halobacterium cutirubrum*. In addition to its *sod* gene, *H.cutirubrum* has a second, related sequence elsewhere in the genome which is probably a pseudogene (May and Dennis, 1989). Hybridizations using the *rpl*12 gene from *Halobacterium cutirubrum* are shown in figure 36. The *sod* and *rpl* genes are being characterized in the P.Dennis laboratory.

The RNA polymerase operon from *Halobacterium halobium* (Leffers *et al.*, 1989), appears to be a similarly constituted operon in *Hf. volcanii*. Probes from all regions of the operon, except the upstream ORF, hybridize with cosmid G60 (figure 37). Homologues of the 75 codon open reading frame (of unknown function) are found in this operon in diverse archaebacteria (Leffers *et al.*, 1989). Hybridization of ORF75 probes to Southern blots of*Haloferax volcanii* genomic DNA and cosmid G60 indicate that it is present in *Haloferax volcanii* as well.

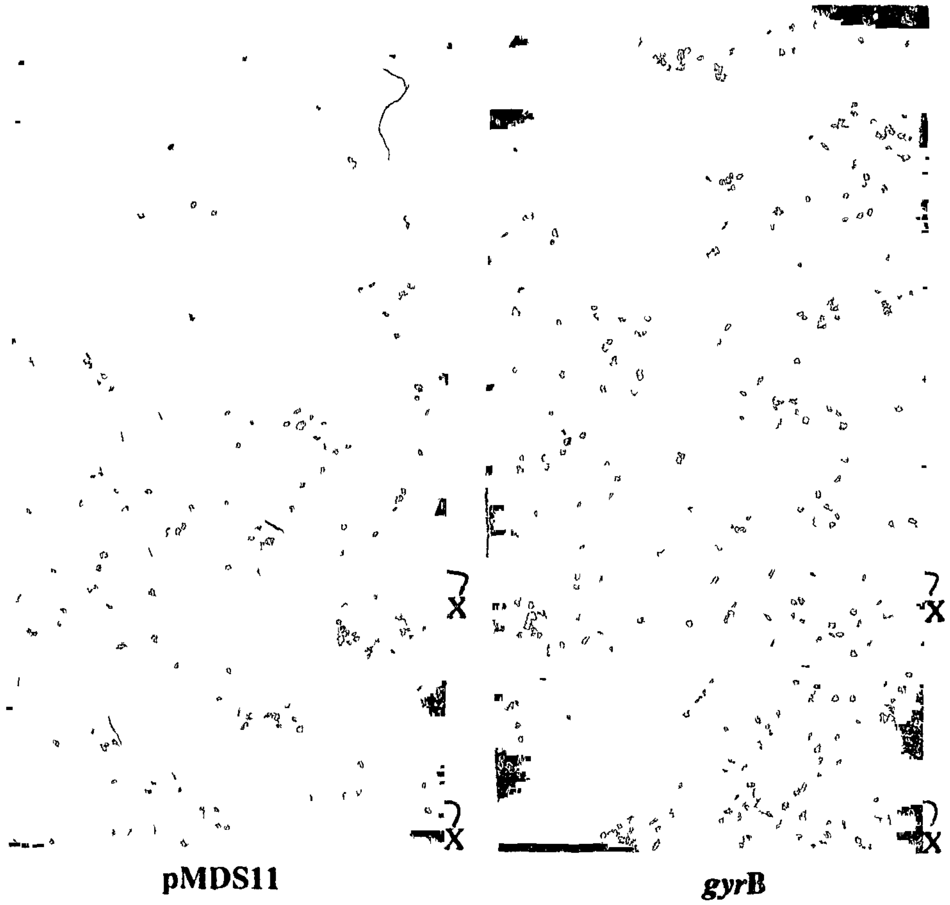


Figure 34. Dot-blot hybridization of pMDS11 and the isolated *gyrB* gene to cosmid DNAs. The filters are of the same series used in figure 32. The two strong spots seen on both are the overlapping cosmids 547 and 326. The latter is no longer in the minimal set. The entire plasmid hybridizes strongly to two additional cosmids, 509 and 307.

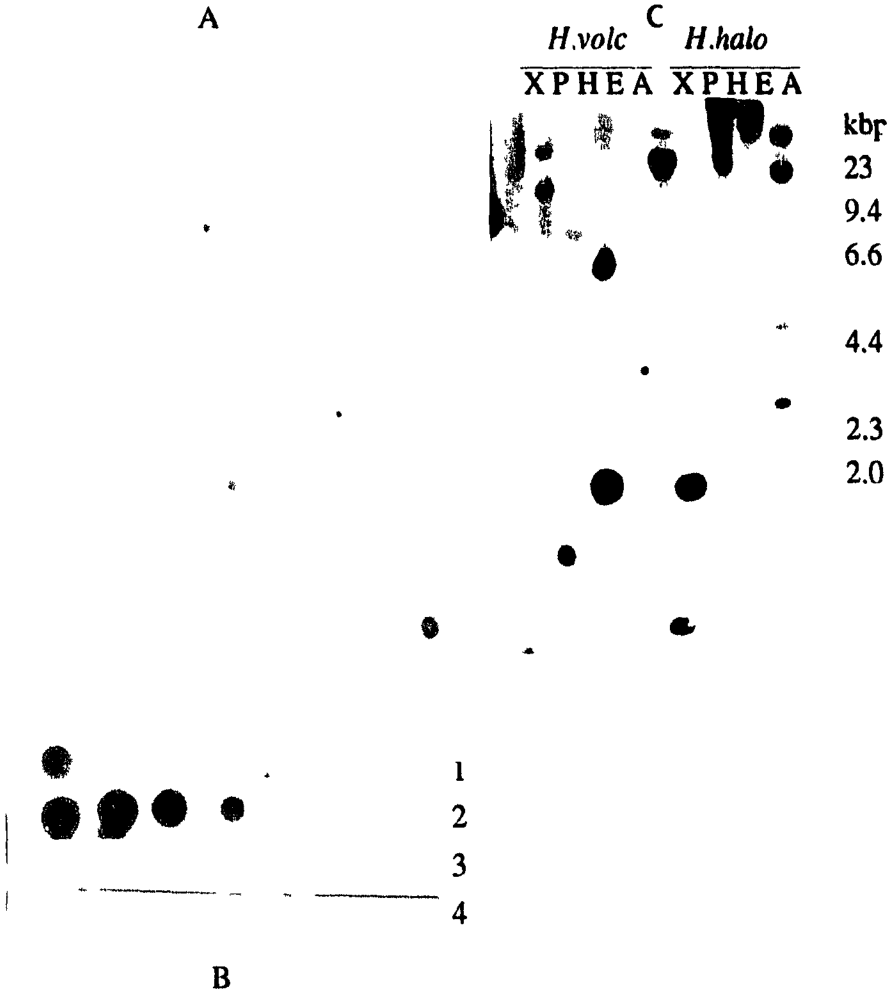


Figure 35. Mapping of the genes for superoxide dismutase. The probe was the purified insert of plkbl, which contains the *sod* gene from *Halobacterium cutirubrum*.

A. dot blot of cosmid DNAs

B. dot blots of tenfold dilutions of

1. *Haloferax volcanii* DNA 10μg
2. *Halobacterium halobium* DNA 10μg
3. pUC18 DNA 0.1μg
4. Herring DNA 10μg

C. Southern transfer of total DNA of *Haloferax volcanii* and *Halobacterium halobium*. A, Asp718; E, EcoRI; H, HindIII; P, Pst I; X, XhoI