Nucleic Acids Research, Vol. 19, No. 21 6027–6031

A random sequencing approach for placing markers on the physical map of *Mycoplasma genitalium*

Scott N.Peterson[1]*, Nara Schramm[2], Ping-chuan Hu[2,3] Kenneth F.Bott[1,2] and Clyde A.Hutchison III[1,2]

[1]Curriculum in Genetics, [2]Department of Microbiology and Immunology and [3]Department of Pediatrics and Infectious Diseases, University of North Carolina at Chapel Hill, NC 27599, USA

Received April 18, 1991; Revised and Accepted October 11, 1991 EMBL accession nos X61510–X61539 (incl.)

ABSTRACT

A physical map of the *Mycoplasma genitalium* genome has been prepared using pulsed-field gel electrophoresis (1). This report details recent efforts made to add markers or specific loci to this map in the absence of any mutants or system of genetic exchange. A total of 44 random clones were partially sequenced. Computer analysis was performed in an attempt to identify homologies with genes already recorded in the DNA sequence database. Clones with a large extent of homology to genes from other microorganisms have been assigned to specific loci on the *M. genitalium* map by hybridization to selected restriction digests. The additional data has facilitated an updated version of the physical map, and verified this random sequencing method as a useful mapping procedure as well as offering new insight into the physiological processes of this fastidious organism.

INTRODUCTION

Mycoplasma genitalium contains a genome of less than 600 kilobase pairs and is the smallest known genome of any free living species (1,2). It has been proposed that the current size of this genome as well as other Mycoplasma species is the result of a large reduction of genetic information, having evolved from gram-positive bacteria with larger genomes (3–7). The number and nature of the genes this organism has maintained throughout this genomic reduction has not yet been characterized. Like other mycoplasma species, *M. genitalium* lacks a cell wall and has a characteristically low G+C content, (32%) (7). This species is believed to contain only one copy of a ribosomal RNA operon (Colman, S.D., Ph.D. thesis, 1990), whereas other Mycoplasmas possess one or two copies (8,9). In species where it has been examined, many isoaccepting tRNA genes are absent (7,10). All mycoplasmas are prevalent parasites of man, animals, arthropods, and plants, most infections causing disease (11). Aside from its significance as a secondary genital pathogen (12), *M. genitalium* represents a useful model system for the determination of the minimal number of genes needed for host independent existence and/or pathogenesis.

Random sequencing has recently been applied to the genome of *Saccharomyces cerevisiae*, as a means of surveying that genome with regard to coding capacity, and the degree to which it has been characterized (Davies, C.J. Ph.D. thesis, 1991). Here we report the use of random sequencing to identify sequences which appear to be homologous to conserved genes in other bacterial species, such clones would serve as useful molecular markers for the *M. genitalium* physical map.

The increased occurrence of antibiotic resistance among mycoplasma species intensifies the need to characterize individual genes of this, and other closely related species (13). The mapping and sequencing of highly conserved genes is an important initial step in the further characterization of this genome. With a physical map of the *M. genitalium* genome already completed (1), the assignment of useful molecular markers to genomic restriction fragments by hybridization is readily performed.

MATERIALS AND METHODS

Clones and sequencing

Genomic *M. genitalium* DNA was digested with either Eco RI or Bam HI, ligated into pUC118, then used to transform competent DH5αF′ cells. Single stranded templates were prepared directly from clonal isolates in microtiter dishes, (14) using the helper phage M13CO7. Sequencing was performed using the Sanger dideoxynucleotide method (15), with the M13 universal primer and DNA polymerase large fragment (BRL). Sequences were electrophoresed on 6% polyacrylamide buffer gradient gels (5× TBE to 0.5× TBE). Sequences were read directly into a portable computer (16) and each sequence was proofread to ensure accuracy. The 44 sequences were then compared to each other, using the Staden programs (17). Unique sequences were concatenated and compared to version 63.0 of the GenBank database using the FASTA algorithms of Pearson et. al. (18) and the Wisconsin GCG computer package (19) running on the UNC VAX 6330 computer system. The GCG program MAP was used to translate the concatenated sequences into amino acid sequence. Open reading frames were used

* To whom correspondence should be addressed

individually to search the PIR database using the program FASTA (17). In each case sequences were determined on only one strand. In some cases sequencing reactions and gels were repeated, to check accuracy or lengthen the reading.

Mapping

Exponential *M. genitalium* cultures, (approximately 1×10^9 cells) grown in Hayflick's medium (20) were used for genomic DNA preparations. DNA was prepared as described previously (1) from cell cultures and fixed in low-gelling temperature agarose (InCert, FMC BioProducts) at a concentration of $\sim2\mu g/100\ \mu l$ agarose block. Approximately $1\mu g$ of DNA was used for each restriction enzyme digest. Agarose blocks were equilibrated in KGB buffer (21) with 40 units of either Apa I, Mlu I, Sma I, or Xho I and incubated overnight at the appropriate temperature.

The digested genomic DNA was electrophoresed at 7V/cm in 0.5× TBE, in 1% agarose using a contour-clamped homogeneous electric field (CHEF) device (22) with pulse times of 10 sec. for 24 hrs., and 5 sec. for 24 hrs. DNAs in gels were stained with ethidium bromide ($0.5\mu g$/ml) then nicked by U.V. treatment for

Table 1. A summary of sequence analysis.

Accession number	Contig	Length	Homologous gene	Homologous sequence file
X61510	1	238	ORF	
X61511	2	232	ORF	
X61512	3	220	(ORF) 35−220	
X61513	4	209	ORF	
X61514	5*	301	*uvr*A (*E. coli*)	ECOUVRAA
X61515	6$^+$	230	23S rRNA gene (*M. luteus*)	MLURN23S
X61516	7$^+$	237	*rrn*B (*B. subtilis*) 23S 5'end	BACRGRRNB
X61517	8	207	ORF	
X61518	10	102	ORF	
X61519	12	284	ORF	
X61520	13	258	ORF	
X61521	14*	269	*lep*A (*E. coli*)	ECOLEP
X61522	15	216	ORF (two open)	
X61523	16	252	ORF	
X61524	17	255	ORF	
X61525	18	192	ORF	
X61526	19$^+$	270	*rrn*B (*B. subtilis*)	BACRGRRNB
X61527	20*$^{\#}$	404	MgPa operon (*M. genitalium*)	MYCMGP
X61528	22*	231	RNA pol β' subunit (*E. coli*)	ECORPLRPO
X61529	24	275	(ORF) 26−275	
X61530	27	217	ORF	
X61531	31	211	(ORF) 1−202	
X61532	32	262	ORF	
X61533	34*	394	*gyr*A (*B. subtilis*)	BACORIC
X61534	35*	295	RNA pol β' subunit (*E. coli*) (1−284)	ECORPLRPO
X61535	36*	306	RNA pol σ subunit (*M. xanthus*) (1−303)	MXARPOD
X61536	40	262	ORF	
X61537	41	238	(ORF) 1−190	
X61538	42*	148	Glu-tRNA synthetase (*E. coli*)	ECOGLTX
X61539	44	127	ORF	

Listed are the 30 unique contigs with accession numbers. + = rRNA homology, * = potential protein coding genes, identifiable by homology to other species, # = homology to *Mycoplasma genitalium* adhesin operon and repetitive sequence. (ORF) = contigs that did not have an open reading frame throughout the gel reading. The length of individual contigs is given in nucleotides. All sequence information was determined for only one strand.

Table 2. Percentage identity and similarity on the nucleotide and amino acid level.

CLONE	HOMOLOGOUS SEQUENCE	% IDENTITY NUCLEIC ACID	% IDENTITY AMINO ACID	% SIMILARITY AMINO ACID
5	*uvr*A	66%	54%	88%
6	23s rRNA	78%		
7	*rrn*B	64%		
14	*lep*A	45%	43%	86%
19	*rrn*B	76%		
20	MgPa	100%	99%	100%
22	*pol* β'	40%	46%	88%
34	*gyr*A	52%	48%	89%
35	*pol* β'	57%	61%	90%
36	*pol* σ	51%	72%	93%
42	*glt*X	65%	55%	91%

Conservative amino acid changes were determined by the programs FASTA (16).

5 min. on a transilluminator (Fotodyne) and transferred to HybondN$^+$ filters according to the manufacturers instructions. DNA was fixed using an alkali treatment procedure (Amersham).

Probes for hybridization were prepared using a random primer labeling kit (Boehringer Mannheim), and ^{32}P-dCTP (3000 Ci/mM, New England Nuclear). Southern hybridizations (23) were performed at 65°C. Filters were washed in 2×SSC, 0.1% SDS once at room temperature for 10 min., and once at 65°C for 10 min. This washing series was repeated using 1×SSC, 0.1% SDS and finally 0.2×SSC, 0.1% SDS. Probed filters were placed on Kodak X-OMAT AR X-ray film at −70°C with two intensifying screens.

Clones were mapped by analyzing their hybridization to the known restriction fragments of the *M. genitalium* genome (1) from the filters of CHEF gels and comparison with the physical map, as well as to a new series of fragments obtained by Mlu I digestion.

RESULTS

Random sequencing and analysis of *M. genitalium* clones

Mycoplasma genitalium DNA was digested to completion with Eco RI or Bam HI in order to make two genomic libraries in the vector pUC118. White colonies were chosen at random and grown in microtiter dishes and sequencing reactions were performed on each. Nucleotide sequence was read in a single orientation from every clone. Gel reading errors were minimized by proofreading all sequences. Sequences were obtained from 44 clones. These were compared to each other using the Staden programs for shotgun sequencing projects (17). One clone produced only vector sequence. The remaining 43 sequences fell into 30 contigs. 11 contigs consisted of duplicate isolates of the same sequence, and one sequence was represented in triplicate. This overabundance of some clones appears to be related to the higher efficiency in cloning small DNA fragments. One contig consisted of overlapping sequences from the two ends of a 404 bp. Bam HI fragment cloned in opposite orientations. The remaining 17 contigs each consisted of one unique sequence. Duplicated sequences allowed us an independent means of assessing the quality of our sequence data, which we judge to be greater than 99%. This was determined by dividing the number of total discrepancies by the length of the total overlapping or redundant sequence information. Sequence data was then reexamined in regions where such inconsistencies existed to aid in the determination of the most probable final sequence.

The 30 contigs were concatenated into a single sequence file and used to search the DNA sequence database, using FASTA, (18) in the GCG program package. This analysis allowed the identification of three contigs which were homologous to rRNA sequences, and one sequence determined to be that of the cloning vector. The remaining 26 contigs were then translated using a translation code modified for Mycoplasmas, such that UGA encodes tryptophan rather than serving as a stop codon (24). Long open reading frames (ORFs) were found in each of these sequences. In all cases except six, the open reading frame extended throughout the length of the gel reading. In cases where stop codons were encountered, they were present either at the beginning or end of gel readings where sequencing information is less reliable. Contig 15 contained two open reading frames (see Table 1).

Each ORF was used individually to search the PIR protein sequence database, using FASTA (18). Seven of the 27 ORF's had sequences predicted to encode proteins with homology to previously characterized bacterial genes. Additional sequencing has been performed on the *gyr*A, *lep*A, RNA polymerase σ

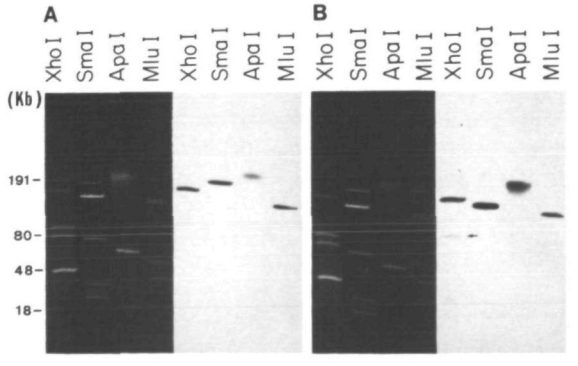

Figure 1. Southern analysis of CHEF filters. **A:** Ethidium stained genomic DNA digested with Xho I, Sma I, Apa I, and Mlu I, then probed with the clone with homology to *uvr*A. **B:** Ethidium stained genomic DNA digested with Xho I, Sma I, Apa I and Mlu I, then probed with the clone with homology to gyrase A. Size standards are in kilobase pairs. Size estimates of bands produced by each digestion are given in (1,2).

Table 3. A summary of results obtained by probing CHEF filters.

CLONE	*Xho I*	*Sma I*	*Apa I*	*Mlu I*
*uvr*A	X2	S1	A1	M2/M3
*lep*A	X4	S3	A1	M4
	X7			
pol β'	X5	S1	A1	M2/M3
*gyr*A	X2	S2	A1	M2/M3
	X3			
pol σ	X1	S5	A2	M1

It was not possible to resolve the bands M2 from M3 unambiguously in Mlu I digests under our electrophoresis conditions and in this case are referred to as M2/M3. X=Xho I, S=Sma I. A=Apa I, M=Mlu I.

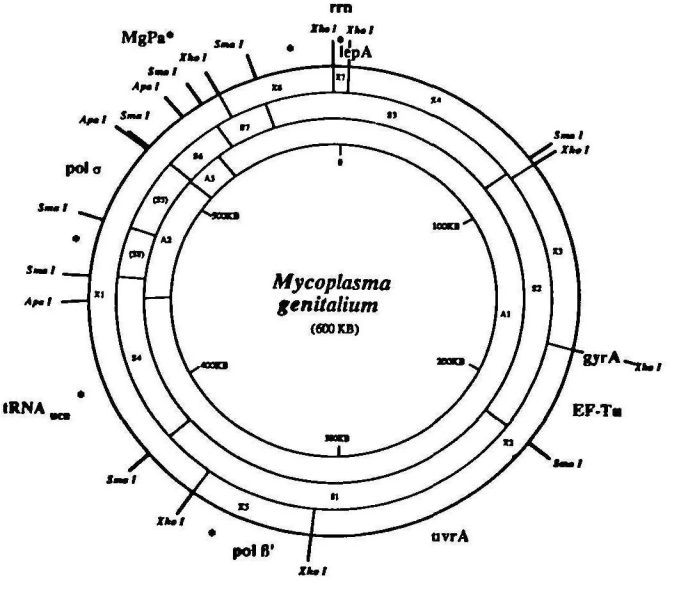

Figure 2. The *M. genitalium* physical map with new markers. * indicate the positions of repeated units of the MgPa operon. Markers representing MgPa, tRNA-UCA, EF-Tu, and rrn were mapped previously (1).

subunit and *uvr*A derivatives; in all cases the significant degree of sequence relatedness is maintained. (Bott and Sancar, personal communications). An amino acid comparison of the seven sequences with homologies to other known genes in the database (see Table 2.) shows a range of identity from 41%-72%.

Adding markers to the *M. genitalium* map

High molecular weight DNA, prepared in low melting agarose, was digested separately with four different restriction enzymes: Xho I, Sma I, Apa I, and Mlu I. Fragments generated from these digests were separated using CHEF electrophoresis. Southern analyses (see Figure 1.) were performed using probes from the clones homologous to DNA gyrase subunit A, (*gyr*A), *lep*A, a bacterial leader peptidase, RNA polymerase s subunit, RNA polymerase β' subunit, and *uvr*A, an excision repair protein. Clones encoding the rRNA determinant, and the MgPa adhesin operon had been mapped previously and so were not repeated here (1). Analysis of each allowed us to assign these sequences to a specific locus on the *M. genitalium* map (see Table 3). Two Xho I bands hybridized to both the *gyr*A and *lep*A homology clones. This was explained, and confirmed experimentally by the presence of an Xho I site within the original clones. This enabled an exact placement of these two markers. By including Mlu I digests in our Southern analyses we have been able to make associations of Mlu I fragments to different positions on the physical map, thus advancing our efforts to improve the *M. genitalium* restriction map (see Figure 2.).

DISCUSSION

The relatively uncharacterized nature, and small size of the *Mycoplasma genitalium* genome coupled with a lack of phenotypically distingiushable mutants and genetic transfer procedures make a random sequencing approach an excellent method for identifying chromosomal loci. This procedure identified potentially interesting avenues for future experimentation while providing additional information for the physical map of the genome. Functional studies, or more extensive sequence analysis will need to be performed to determine if the clones that were sequenced actually represent homologous genes to those sequences that the database predicted. Regardless, these clones are useful molecular markers which will assist the further characterization of this Mycoplasma genome and allow more direct comparisons of this physical map to other Mycoplasma species.

One particularly surprising result was the discovery of a clone with homology to an *E. coli* repair gene, *uvr*A. It has been reported that *Mycoplasma gallisepticum* is deficient for repair functions (25). Mycoplasma species with significantly larger genomes do posess the ability to repair U.V. damage (26,27). It had been tempting to speculate that during the reduction of genome size in *M. gallisepticum*, genes responsible for U.V. repair were deleted. This is now difficult to reconcile with the presence of such genes in *M. genitalium* which has an even smaller genome. The significance of this finding is not clear at this time, and it has not yet been confirmed that this gene is functional, however it does provide additional means by which various phylogenetic relationships can be tested.

It is also worth noting that no homologies to sequences previously characterized in other species encoding metabolic or biosynthetic pathway genes have yet been found. This is not necessarily surprising in light of the large number of nutritional requirements needed in culture media to support Mycoplasmal growth, but it does seem clear that more than just the genetic information encoding copies of ribosomal operons, tRNA's, and cell wall components have been lost during genome reduction as was previously suggested (27). Broadening these results will allow speculation as to the nature of the genes that were deleted as the genome size decreased through evolution, and which genes are an absolute requirement for life and possibly pathogenicity. It does seem apparent from this data that *M. genitalium* does make efficient use of DNA, as shown by the high fraction of clones containing open reading frames.

Genes encoding various epitopes of the P1 adhesin protein represent a relatively large proportion of the *M. pneumoniae* genome, perhaps as much as 6% (28–30). Although the analogous *M. genitalium* MgPa operon has been cloned and sequenced (31), clones representing that operon do not seem to be as abundant in randomly selected clones. This prompts the speculation that *M. genitalium* may not have as large a representation of truncated adhesin gene epitopes in its genome.

A continued study following this procedure on an expanded scale would possibly allow some useful conclusions to be drawn about the types of genes Mycoplasma evolution has maintained, their relatedness to other bacterial species and a better insight into the physiology of this extremely fastidious organism. Additionally it is possible that genes could be identified that would help our understanding of Mycoplasma pathogenicity and how to best combat it.

ACKNOWLEDGEMENTS

We would like to thank Camella Bailey, Theresa Irish, Elisa Fox, Andrew Sparks, and David Hsu, for their help and for sharing unpublished data for this manuscript, and Chris Davies for supplying vector preparations, and technical advice. S.N.P. would also like to extend special thanks to Dr. John Lucchesi for all of his support. This work was supported in part by a genetics training grant to K.F.B.: T32 GM 07092, also by an NIH grant GM 21313, and AI08998 to C.A.H.

REFERENCES

1. Colman, S.D., Hu, P.-C., Litaker, W., and Bott, K.F. (1990) Mol. Micro. 4, 683–687
2. Su, C.J., and Baseman, J.B. (1990) J. Bact. 172, 4705–4707
3. Maniloff, J. (1983) Ann. Rev. Microbiol. 37, 477–499
4. Woese, C.R., Stackebrandt, E., and Ludwig, W. (1985) J. Mol. Evol. 21, 305–316
5. Weisburg, W.G., Tully, J.G., Rose, D.L., Petzel, J.P., Oyaizu, H., Yang, D., Mandelco, L., Sechrest, J., Lawrence, T.G., Van Etten, J., Maniloff, J., and Woese, C.R. (1989) J. Bacteriol. 171, 6455–6467
6. Rogers, M.J., Simmons, J., Walker, R.T., Weisburg, W.G., Woese, C.R., Tanner, R.S., Robinson, I.M., Stahl, D.A., Olsen, G., Leach, R.H., and Maniloff, J. (1990) Proc. Natl. Acad. Sci. USA 82, 1160–1164
7. Razin, S. (1985) Microbiol. Rev. 49, 419–455
8. Sawada, M., Osawa, S., Kobayashi, H., Hori, H., and Muto, A. (1981) Mol. Gen. Genet. 182, 502–504
9. Amikam, D., Glaser, G., and Razin, S. (1984) J. Bacteriol. 158, 376–378
10. Muto, A., Andachi, Y., Yuzawa, H., Yamao, F., and Osawa, S. (1990) Nuc. Acids Res. 18, 5037–5043
11. Tully, J.G., Taylor-Robinson, D., Cole, R.M., and Rose, D.L. (1981) Lancet i: 1288–1291
12. Tully, J.G., Taylor-Robinson, D., Rose, D.L., Cole, R.M., and Bove, J.M. (1983) Int. J. Syst. Bacteriol. 33, 387–396
13. Kenny, G.E., Hooton, T.M., Roberts, M.C., Cartwright, F.D., and Hoyt, J. (1989) Antimicrobial Agents and Chemotherapy 33, 103–107
14. Hutchison, C.A. III, Swanstrom, R., and Loeb, D.D. (1991) Meth. in Enzym. 202, 356–390

15. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA. 74, 5463–5467
16. Hutchison, C.A. III (1986) Nuc. Acids Res. 14, 1917
17. Staden, R. (1982) Nuc. Acids Res. 10, 4731–4751
18. Pearson, W.R., and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA. 2444–2448
19. Devereux, J., Haeberli, P., and Smithies, O. (1984) Nuc. Acids Res. 12, 387–395
20. Hayflick, L. (1965) Texas Rep. Biol. Med. 23, 285–303
21. McClelland, M., Hanish, J., Nelson, M., and Patel, Y. (1988) Nuc. Acid Res. 16, 364
22. Chu, G., Vollrath, D., and Davis, R.W. (1986) Science 234, 1582–1585
23. Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) Molecular Cloning: A Labroratory Manual, 2nd edition. Cold Spring Harbor University Press, Cold Spring Harbor.
24. Inamine, J.M., Ho, K.-C., Loechel, S., Hu, P.-C. (1990) J. Bacteriol. 172, 504–506
25. Ghosh, A., Das, J., and Maniloff, J. (1977) J. Mol. Biol. 116, 337–344
26. Das, J., Nowak, J.A., and Maniloff, J. (1977) J. Bacteriol. 129, 1424–1427
27. Aoki, S., Ito, S., and Watanabe, T. (1979) Microbiol. Immunol. 23, 147–158
28. Neimark, H. (1983) The Yale Journal of Biology and Medicine 56, 377–383
29. Wenzel R., and Herrmann R. (1988) Nuc. Acids Res. 16, 8337–8350
30. Colman, S.D., Hu, P -C., and Bott, K.F. (1990) Gene 87, 91–96
31. Ruland, K., Wenzel, R., and Herrmann, R. (1990) Nuc. Acids Res. 18, 6311–6317
32. Inamine, J.M., Loechel, S., Collier, A.M., Barile, M.F., and Hu, P.-c. (1989) Gene 82, 259–267