3100–3104 *Nucleic Acids Research, 2000, Vol. 28, No. 16*

Mapping analysis of the *Xylella fastidiosa* genome

Marcus Frohme[1,*], Anamaria A. Camargo[2], Steffen Heber[1,3], Claudia Czink[1], Andrew J. D. Simpson[2], Jörg D. Hoheisel[1] and Anete Pereira de Souza[4]

[1]Functional Genome Analysis, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 506, D-69120 Heidelberg, Germany, [2]Cancer Genetics, Ludwig Institute for Cancer Research, São Paulo, Brazil, [3]Theoretical Bioinformatics, Deutsches Krebsforschungszentrum, Heidelberg, Germany and [4]Genética e Evolução, Centro de Biologia Molecular e Engenharia Genética, UNICAMP, Campinas, Brazil

Received April 27, 2000; Revised and Accepted June 23, 2000

ABSTRACT

A cosmid library was made of the 2.7 Mb genome of the Gram-negative plant pathogenic bacterium *Xylella fastidiosa* and analysed by hybridisation mapping. Clones taken from the library as well as genomic restriction fragments of rarely cutting enzymes were used as probes. The latter served as a backbone for ordering the initial map contigs and thus facilitated gap closure. Also, the co-linearity of the cosmid map, and thus the eventual sequence, could be confirmed by this process. A subset of the eventual clone coverage was distributed to the Brazilian *X.fastidiosa* sequencing network. Data from this effort confirmed more quantitatively initial results from the hybridisation mapping that the redundancy of clone coverage ranged between 0 and 45-fold across the genome, while the average was 15-fold by experimental design. Reasons for this not unexpected fluctuation and the actual gaps are being discussed, as is the use of this effect for functional studies.

INTRODUCTION

As the target of Brazil's initial genome project, the bacterium *Xylella fastidiosa* (1) was chosen because of its economic and biological importance, being the first plant pathogen to be sequenced completely (2). *Xylella fastidiosa* is associated with citrus variegated chlorosis (CVC) among other diseases (3). Currently, CVC is the most severe and economically damaging plant disease in Brazil's citriculture (4), affecting mainly sweet oranges. *Xylella fastidiosa* is a Gram-negative, rod-shaped bacterium with a wrinkled cell wall. It is present in the xylem of the infected plant and was detected not only in sweet oranges but also other plants of economic importance, such as grapevine, alfalfa, almond, peach and coffee (5). It is thought to act as a pathogen by clogging the plant's vascular system, thus inducing water stress and nutritional disorders (6). The resulting symptoms manifest themselves in the fruit's appearance and chlorotic spots on the leaves. The bacterium's special habitat of low nutrition made the culturing and diagnostic requirements needed to fulfil Koch's postulates a real challenge to microbiologists (7) and different genotypes of the bacterium were identified (8,9). In natural habitats, *X.fastidiosa* is transmitted by different insect vectors, mainly sharpshooters (4). Thus, vector control is a means of containing the disease. Also, some orange varieties exhibit a resistance to CVC.

To understand and then tackle the pathogenic characteristics of *X.fastidiosa*, the molecular basis of the infection mechanisms and host–pathogen interactions needs to be uncovered. Unravelling the genomic sequence and the sequence of a so far uncharacterised megaplasmid in their entirety is a first important step towards this end. Since *X.fastidiosa* is the first plant pathogen to be fully sequenced, a large benefit for the whole field of disease research in plants can be expected. An initial step towards the sequencing was the provision of an ordered cosmid library covering the genome. Hybridisation mapping has been proven to be a powerful tool to this end (10–12). Even pure shotgun sequencing strategies actually rely on map information obtained on larger genomic fragments for contig arrangement and confirmation of co-linearity (13–15). Moreover, the existence of a map provides an overview on the extent of critical regions in the genome, such as repeats, extrachromosomal DNA or regions of instability. Even gaps in the map highlight regions difficult or impossible to be cloned in *Escherichia coli* and therefore allow early anticipation of different sequencing strategies for such areas.

MATERIALS AND METHODS

DNA preparation

For DNA preparation, *X.fastidiosa* strain 9a5c was used. This strain was derived from isolate 8.1b (7) which had been obtained from a sweet orange tree affected by CVC in the State of São Paulo. Bacteria were grown in agitated medium as described (16) at 28°C for 94 h; every 24 h the culture was diluted 1:10 with fresh medium.

Genomic DNA was isolated by phenol/chloroform extraction for cloning or as intact molecules after immobilisation in agarose plugs for pulsed field gel electrophoresis. For phenol/chloroform extraction, the bacterial pellet of a 500 ml saturated culture was taken up in 5 ml 200 mM NaCl, 10 mM Tris–HCl, pH 7.2, 100 mM EDTA containing 10 µg/ml proteinase K and incubated at 55°C overnight. Protein contaminants were

*To whom correspondence should be addressed. Tel: +49 6221 42 4678; Fax: +49 6221 42 4687; Email: m.frohme@dkfz-heidelberg.de

removed by three subsequent phenol/chloroform extractions followed by ethanol precipitation in the presence of 0.3 M NaAc, pH 5.2. For plug preparation, a standard protocol was used (17).

Production of the cosmid library and clone filters

Production procedures were as described in detail elsewhere (12) except for the following modifications. An aliquot of 17.5 µg *X.fastidiosa* DNA was partially digested with 10 U *Mbo*I (NEB, Schwalbach, Germany) in 100 µl for 2.5 min. Genomic DNA (4 µg) was ligated to 2 µg vector arms of cosmid Lawrist-4, packaged and transfected into *E.coli* DH5α; 1056 individual clones, equivalent to a 15-fold genome coverage, were picked in microtitre dishes for storage. Filter spotting was carried out with a commercial robotic device (BioGrid; BioRobotics, Cambridge, UK).

Pulsed field gel electrophoresis

Agarose plugs were digested with 40–100 U of restriction enzyme in 200 µl for 10 h under conditions as recommended by the manufacturers. Electrophoresis was in 1% agarose gels (Seakem LE agarose; FMC, Rockland, ME) on a Chef-II system (Bio-Rad, München, Germany) for 45 h at 5 V/cm with switch time ramping from 10 to 40 s. After ethidium bromide staining, individual bands were cut out under long wavelength UV light and equilibrated in water prior to labelling. The megaplasmid was isolated from an electrophoresis of undigested *X.fastidiosa* DNA.

Probe preparation and hybridisation

Usually, cosmid DNA was isolated by a fast protocol: 1 ml overnight cultures were pelleted and taken up in ~100 µl medium. After subsequent addition of 300 µl 10 mM Tris, 1 mM EDTA, 0.1 M NaOH, 0.5% SDS and 150 µl 3 M NaOAc, pH 5.2, the sample was centrifuged at high speed for 5 min. The supernatant was directly mixed with 1 ml ice-cold ethanol and precipitated for 15 min at –80°C. After another centrifugation, the pellet was taken up in 50 µl 10 mM Tris–HCl, pH 8.5. A tenth of such a preparation, or a band cut from a pulsed field gel, was labelled by random hexamer priming (18). All subsequent processes concerning hybridisation and detection followed standard protocols (12).

Cosmid end-sequencing

Cosmid DNA from 5 ml overnight cultures was extracted using commercial kits (Promega, Mannheim, Germany). An aliquot of 1 µg DNA was used for sequencing reactions with the Big Dye Terminator Kit (PE Biosystems, Foster City, CA) using the primers GAAATTAATACGACTCACTATAGGGAGACC and CCTTATGTATCATACACATACGATTTAGG. After initial denaturation at 96°C, 25 cycles of 30 s at 96°C, 15 s at 50°C and 4 min at 60°C were carried out. Samples were analysed on ABI Prism 377 sequencers.

Bioinformatics tools

The cosmid map was analysed by an established programme package (19). Cosmid end-sequences were analysed using a Blast programme version available at the home page of the *X.fastidiosa* Sequencing Project (www.ibi.dcc.unicamp.br). Calculations of clone coverage were based on the Matlab programme package.

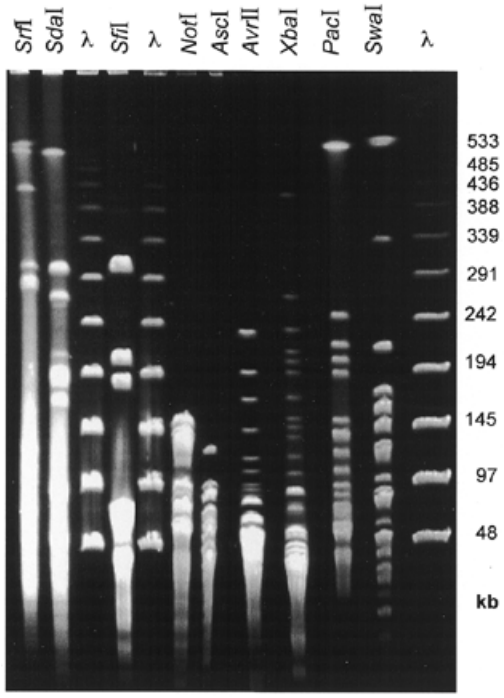

Figure 1. Pulsed field gel electrophoresis of *X.fastidiosa* DNA. Genomic DNA was cut with several rarely cutting enzymes. λ concatemers (Roche, Mannheim, Germany) were used as molecular weight markers; fragment sizes are indicated.

RESULTS

While the genome size of *X.fastidiosa* had been estimated to be some 2.1 Mb prior to sequencing, its actual length turned out to be 2.7 Mb (20). Therefore, the cosmid library of 1056 clones represented only about 15-fold coverage of the genome instead of the 20-fold coverage initially aimed at. Nevertheless, redundancy was well above the 10-fold level known to be essential to cover with high probability the entire genome (10). For the mapping process, the selection of clones for hybridisation was done in two stages. First, in a regime of sampling without replacement, individual clones were hybridised to the entire library until all clones had been positive in at least one experiment. Then, clones that were apparently located at contig ends were picked for gap closure. Also, genomic fragments were generated by cleaving intact chromosomal DNA with rarely cutting enzymes (Fig. 1) and used as probes in order to arrange the cosmid contigs in correct succession. The biggest fragments used were longer than 600 kb whereas the smallest was of ~80 kb. Apart from octamer-recognising enzymes, the six-cutters *Xba*I and *Avr*II were found to cleave *X.fastidiosa* DNA infrequently. Locating the clone contigs on the basis of the restriction fragments greatly facilitated and directed gap closure. In a few cases, bands in electrophoresis patterns contained more than one fragment (e.g. the 140 kb band of the *Pac*I digest in Fig. 1), a fact which had to be reflected in the eventual map.

Overall, data from 300 hybridisations (270 individual cosmids, 28 genomic restriction fragments, the megaplasmid

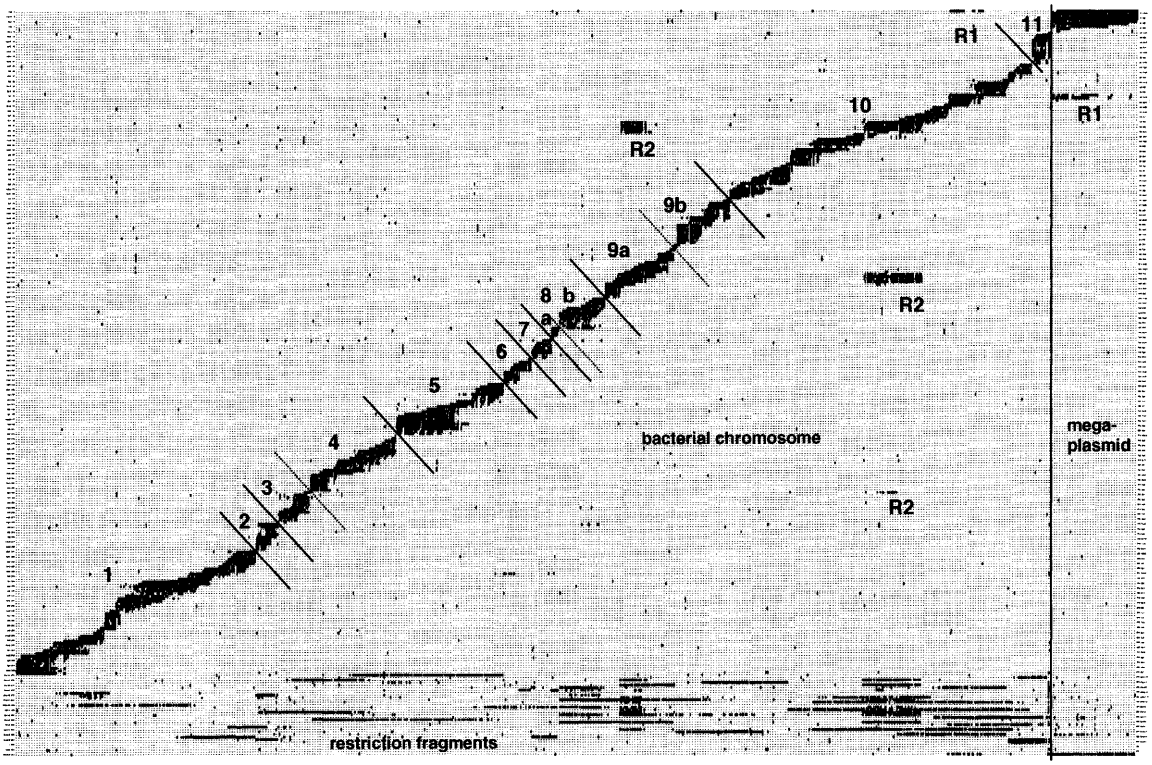

Figure 2. Physical map of the *X.fastidiosa* genome. Hybridisation probes are represented by horizontal lines, clones by vertical lines. A positive hybridisation signal is represented by a black spot at the crossover point. The entire data set is presented, including all false positive or negative results. Clone contigs are numbered and gaps are indicated. Scattered lines indicate positions where a contig was thought to be continuous from hybridisation data while a small gap was identified by sequencing (see main text). On the right, the cosmids representing the megaplasmid are shown. R1 and R2 mark large repeat regions. At the bottom, the results obtained from hybridising genomic restriction fragments are shown; they were not taken into account for the actual clone ordering process and, thus, served as an independent control for co-linearity of the cosmid order.

and an rDNA probe) were used for assembly of the hybridisation map (Fig. 2). From the data, no clone of apparently chimeric character could be detected. However, even if such clones had been present, they would not have disturbed the overall map quality, since redundant mapping information generated by more than one probe was always required for the ordering process. Some cross-hybridisation between part of the megaplasmid and the bacterial chromosome was identified (R1) and could be attributed to a repeat of some 9 kb during the sequencing phase. Another major repeat sequence, this time intra-chromosomal, was also found during mapping (R2), which is present mainly at 2 positions in the genome and clearly visible by strings of positive signals located on either side of the main diagonal.

In some map regions, seemingly superfluous results from very similar probes can be seen. This redundancy of information was caused by hybridisations done for the extension of contig ends after the first mapping phase. Any pair of such areas which is linked by relatively few probes indicates a successful gap closure. The number of gaps could finally be reduced to 11. By comparing the map with the eventually completed sequence, it was found that 92% of the genome was covered by cosmids, while 8% was missing in the library despite its high degree of coverage. Gap sizes varied between a few hundred base pairs and more than the average cosmid insert length of 40 kb. Actually, there were 13 gaps in the map rather than the 11 actually seen. There were two isolated cosmids containing an almost identical insert which eventually could be located between contigs 1 and 2. In both directions there was a stretch of >20 kb of DNA missing to the next cosmid. Because of the two clones isolation and thus the lack of real mapping information, they were erroneously placed in the map, being connected to another contig by a false positive hybridisation signal. Furthermore, the sequence analysis showed that a small overlap of contigs 9a and 9b is not real but only incidentally connected neighbouring clone contigs. Eventually, all gaps were covered for sequencing by PCR, identification of relevant λ clones from another library, sequence walking approaches and shotgun data.

Overall, the order of the cosmids determined by hybridisation was in very good correlation with their order as determined by locating 186 clones accurately after end-sequencing (data not shown). Local disorders were due to regionally high redundancy in the hybridisation map. In a few cases, individual clones containing repeat sequences were placed at the corresponding, homologous region instead of their real map position. However, a cross-hybridisation event slightly disturbed the map in only one region, by incorrectly connecting contigs 8a and 8b (Fig. 2). The eventual order of the contigs as determined by sequencing was found to be 6–8a–7–8b–9.

The coverage of the cosmid library turned out to fluctuate widely across the genome (Fig. 3). The distribution of *Mbo*I sites along the genome does not vary significantly enough, however, to serve as a reasonable explanation for the variation

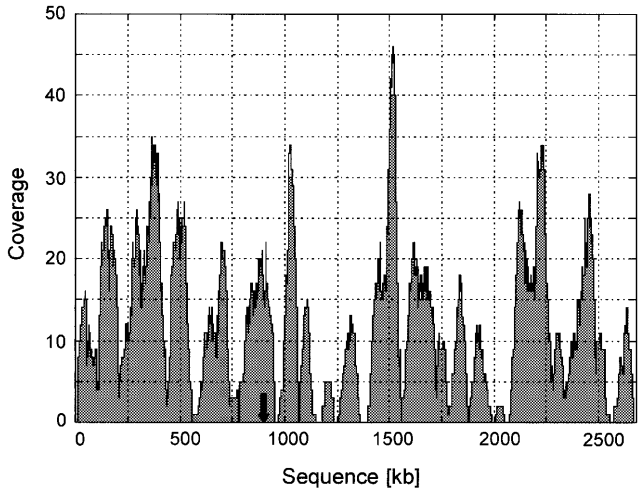

Figure 3. Fluctuation of clone coverage. Based on the final genome sequence, the localisation of 75% of all cosmids was accurately determined. From this data, variation in coverage across the genome was calculated and found to range between 0 and 45. The arrow indicates the putative origin of replication.

in coverage, let alone the existence of gaps. Therefore, it is rather likely that biological reasons, either a direct incompatibility of *X.fastidiosa* DNA fragments and the *E.coli* host or amplification of such sequences by the cosmid vector, were responsible for the effect.

DISCUSSION

For the construction of a cosmid map of the *X.fastidiosa* genome, a hybridisation-based approach was taken. Various repeat regions could be identified very early during the project. The hybridisation of large genomic restriction fragments was found especially essential for resolving the discontinuity caused by the repetitive sequences; they also helped in both the closure of gaps and the correct localisation of remaining ones. The cosmid map served as the basis for a complete sequence analysis of the genome described elsewhere (20).

The strong fluctuation in coverage could not be explained by the distribution of restriction sites used for generating the cloning inserts. Although a variation in cleavage activity at different sites is known for *Mbo*I, their frequency is so high and relatively uniform along the entire chromosome that such an effect is unlikely to lead to the variance in coverage seen. A more likely explanation is the occurrence of sequences which are either unstable in *E.coli* or inhibit bacterial growth. This is known for rDNA sequences, which are generally difficult or impossible to clone. One of the gaps covers the entirety of one of the two rDNA operons of *X.fastidiosa*; a second gap reaches into the second operon. Thus, there is no full representation of the rDNA sequence in any of the cosmids.

Two other gaps span a large number of genes coding for putative ribosomal proteins, which, when expressed in large amounts from a high copy number cosmid, might compete with *E.coli* ribosomal proteins for ribosome assembly. Less functional ribosomes could result in a reduced growth rate of the cells. In the *Mycoplasma pneumoniae* sequencing project (21) the only gap in the cosmid coverage was related to a ribosomal protein (R.Herrmann, personal communication). A close spatial proximity of several such genes seems to be necessary, however, since single ribosomal protein genes could be found in the *X.fastidiosa* map, even in genomic regions that were highly represented in the cosmid library. All eight subunits of ATP synthetase, which are also known to be recalcitrant to cloning (H.U.Schairer, personal communication), were present in one gap and genes for putative phage proteins occurred in another two.

For the remaining six gaps, there are various possible explanations why the relevant DNA could not be cloned. A detailed analysis of the open reading frames identified in the regions after completion of the sequence revealed genes from very different functional groups and it remains speculative whether a single putative protease or nuclease gene was responsible, considering that this type of gene could also be found elsewhere in regions frequently represented in the cosmid library. Nevertheless, some functional clusters of genes in the gap regions are conceivable contenders: one gap covers a large number of putative genes of the dTDP-rhamnose biosynthesis pathway, of RNA processing proteins and of proteins known to be active in microbial cation efflux systems. In another case, a remarkable group of genes for membrane transporters and other membrane proteins were found, as well as open reading frames with homology to transcriptional regulators and some mobile genetic elements. A third gap occurred within a region covering coding sequences with structural function in the synthesis of bacterial pili. Also, a cluster of gene homologues related to colicin V secretion coincided with the position of a gap. Finally, we found two putative transcriptional regulators together with some other metabolic genes in the gap regions. Apart from the genes described, there were numerous ORFs of conserved sequence or unknown function within the gap regions, which could obviously not be excluded from being responsible for the unclonability.

While the absence of particular genomic regions in the clone library was a nuisance during sequencing, the effect offers an opportunity for functional analysis. Fragments of defined length and gene content from one genome could be transferred to another organism. The occurrence of lethality, besides reduced or improved cell growth, could provide important information on the activity of the cloned genes. The process could be combined with chip-based detection, similar to a system set-up for deletion mutants of yeast (22), thus permitting the analysis of progression of the respective cells even in a complex mixture of library clones. Such functional analyses are planned to be performed upon availability of a *X.fastidiosa*-specific plasmid vector, whose creation is well progressed. The respective promoter regions still need to be determined more accurately and (deletion) constructs have to be made in some instances. Then, however, we expect to learn significant new facts about functional aspects from a sequence exchange between *X.fastidiosa* and *E.coli* as well as other bacteria.

Additional information on this project is available at www.ibi.dcc.unicamp.br and www.dkfz-heidelberg.de/funct_genome/index.html

ACKNOWLEDGEMENTS

The *X.fastidiosa* strain 9a5c was kindly provided by Joseph-Marie Bové and Monique Garnier of the Centre de Recherches de Bordeaux. We are indebted to Marcos A. Marchado (Cordeiropolis, Brazil) for culturing the bacterium and the provision of genomic DNA. The fast cosmid miniprep was adapted from a protocol kindly provided by Klaus Hellmuth (Berlin). Work was financially supported by fellowship no. 98/01770-2 from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Finally, we wish to thank Andre Goffeau and Steve Oliver, members of the scientific steering committee of the Brazilian *X.fastidiosa* Sequencing Project, for initiating this collaborative effort.

REFERENCES

1. Wells,J.M., Raju,B.C., Hung,H.-Y., Weisburg,W.G., Mandelco-Paul,L. and Brenner,D.J. (1987) *Int. J. Syst. Bacteriol.*, **37**, 136–143.
2. Simpson,A.J. and Perez,J.F. (1998) *Nature Biotechnol.*, **16**, 795–796.
3. Rossetti,V., Garnier,M., Bové,J.M., Beretta,M.-J.-G., Teixeira,A.R.R., Quaggio,J.A. and de Negri,J.D. (1990) *C. R. Acad. Sci. Série III*, **310**, 345–349.
4. Donadia,L.C. and Moreira,C.S. (1998) *Citrus Variagated Chlorosis*. Fundecitrus, Bebedouro, Brazil.
5. Purcell,A.H. and Hopkins,D.L. (1996) *Annu. Rev. Phytopathol.*, **34**, 131–151.
6. Hopkins,D.L. (1989) *Annu. Rev. Phytopathol.*, **27**, 271–290.
7. Chang,C.J., Garnier,M., Zreik,L., Rossetti,V. and Bove,J.M. (1993) *Curr. Microbiol.*, **27**, 137–142.
8. Beretta,M.J.G., Barthe,G.A., Ceccardi,T.L., Lee,R.F. and Derrick,K.S. (1997) *Plant Disease*, **81**, 1196–1199.
9. Rosato,Y.B., Neto,J.R., Miranda,V.S., Carlos,E.F. and Manfio,G.P. (1998) *System. Appl. Microbiol.*, **21**, 593–598.
10. Hoheisel,J.D., Maier,E., Mott,R., McCarthy,L., Grigoriev,A.V., Schalkwyk,L.C., Nizetic,D., Francis,F. and Lehrach,H. (1993) *Cell*, **73**, 109–120.
11. Scholler,P., Karger,A.E., Meier-Ewert,S., Lehrach,H., Delius,H. and Hoheisel,J.D. (1995) *Nucleic Acids Res.*, **23**, 3842–3849.
12. Hoheisel,J.D., Maier,E., Mott,R. and Lehrach,H. (1995) In Birren,B. and Lai,E. (eds), *Analysis of Non-mammalian Genomes—A Practical Guide*. Academic Press, San Diego, CA, pp. 319–346.
13. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) *Science*, **269**, 496–512.
14. White,O., Eisen,J.A., Heidelberg,J.F., Hickey,E.K., Peterson,J.D., Dodson,R.J., Haft,D.H., Gwinn,M.L., Nelson,W.C. *et al.* (1999) *Science*, **286**, 1571–1577.
15. Lin,J., Qi,R., Aston,C., Jing,J., Anantharaman,T.S., Mishra,B., White,O., Daly,M.J., Minton,K.W., Venter,J.C. and Schwartz,D.C. (1999) *Science*, **285**, 1558–1562.
16. Davis,M.J., French,W.J. and Schaad,N.W. (1981) *Curr. Microbiol.*, **6**, 309–314.
17. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
18. Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.*, **132**, 6–13.
19. Mott,R., Grigoriev,A., Maier,E., Hoheisel,J.D. and Lehrach,H. (1993) *Nucleic Acids Res.*, **21**, 1965–1974.
20. Simpson,A.J.G., Reinach,F.C., Arruda,P., Abreu,F.A., Acencio,M., Alvarenga,R., Alves,L.M.C., Araya,J.E., Baia,G.S., Barros,M.H., Batista,C.S., Bonaccorsi,E.D., Bordin,S., Bové,J., Briones,M.R.S., Bueno,M.R.P., Camargo,A.A. *et al.* (2000) *Nature*, **406**, 151.
21. Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.-C. and Herrmann,R. (1997) *Nucleic Acids Res.*, **24**, 4420–4449.
22. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R. *et al.* (1999) *Science*, **285**, 901–906.