ELSEVIER

International Journal for Parasitology xx (2002) xxx–xxx

INTERNATIONAL Journal for PARASITOLOGY

www.parasitology-online.com

Sequencing and analysis of a 63 kb bacterial artificial chromosome insert from the *Wolbachia* endosymbiont of the human filarial parasite *Brugia malayi*☆

Jennifer Ware[a,*], Laurie Moran[a], Jeremy Foster[a], Janos Posfai[a], Tamas Vincze[a], David Guiliano[b], Mark Blaxter[b], Jonathan Eisen[c], Barton Slatko[a]

[a]*New England Biolabs, Inc., 32 Tozer Road, Beverly, MA 01915, USA*
[b]*Institute of Cell, Animal and Population Biology, King's Buildings, University of Edinburgh, Edinburgh EH9 3JT, UK*
[c]*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA*

Received 4 October 2001; received in revised form 19 November 2001; accepted 19 November 2001

Abstract

Wolbachia endosymbiotic bacteria are widespread in filarial nematodes and are directly involved in the immune response of the host. In addition, antibiotics which disrupt *Wolbachia* interfere with filarial nematode development thus, *Wolbachia* provide an excellent target for control of filariasis. A 63.1 kb bacterial artificial chromosome insert, from the *Wolbachia* endosymbiont of the human filarial parasite *Brugia malayi,* has been sequenced using the New England Biolabs Inc. Genome Priming System™ transposition kit in conjunction with primer walking methods. The bacterial artificial chromosome insert contains approximately 57 potential ORFs which have been compared by individual protein BLAST analysis with the 35 published complete microbial genomes in the Comprehensive Microbial Resource database at The Institute for Genomic Research and in the NCBI GenBank database, as well as to data from 22 incomplete genomes from the DOE Joint Genome Institute. Twenty five of the putative ORFs have significant similarity to genes from the α-proteobacteria *Rickettsia prowazekii,* the most closely related completed genome, as well as to the newly sequenced α-proteobacteria endosymbiont *Sinorhizobium meliloti*. The bacterial artificial chromosome insert sequence however has little conserved synteny with the *R. prowazekii* and *S. meliloti* genomes. Significant sequence similarity was also found in comparisons with the currently available sequence data from the *Wolbachia* endosymbiont of *Drosophila melanogaster*. Analysis of this bacterial artificial chromosome insert provides useful gene density and comparative genomic data that will contribute to whole genome sequencing of *Wolbachia* from the *B. malayi* host. This will also lead to a better understanding of the interactions between the endosymbiont and its host and will offer novel approaches and drug targets for elimination of filarial disease.

Keywords: DNA sequencing; *Wolbachia*; BAC; *Brugia malayi*; Filarial parasite; Transposon

1. Introduction

The World Health Organization (TDR/UNDP/World Bank) has sponsored an international collaboration of seven endemic and non-endemic laboratories to implement a program of gene discovery, genome mapping, and post-genomic analysis of the human filarial parasite *Brugia malayi* (Blaxter et al., 1999; Filarial Genome Project, 1999). Over 120 million people are infected by lymphatic filarial nematodes, and over 1 billion people are at risk from *B. malayi* and related filarial parasites (Ottesen and Ramachandran, 1995). cDNA libraries representing various stages of the life cycle and large insert genomic DNA libraries have been created and serve as the core of the gene discovery and genome mapping initiatives (Williams et al., 1999; http://nema.cap.ed.ac.uk/fgn/filgen1.html). These resources have also enabled the positive identification and preliminary characterisation of the genome of an obligate endosymbiotic bacteria, *Wolbachia* (Bourtzis and Braig, 1999; Williams et al., 2000), and several genes useful for phylogenetic studies, including *fts*Z, *wsp,* and 16S rDNA have been identified (Bandi et al., 1998; Hoerauf et al., 1999; Bazzocchi et al., 2000a; Taylor et al., 1999). In addition to their presence in the majority of assayed filarial nematodes, *Wolbachia* are maternally inherited endosymbionts present in numerous arthropod species (up to 76%) (Jeyaprakash and Hoy, 2000). They have been shown to be

☆ Sequence submitted to GenBank: accession # AF373870.

* Corresponding author. Tel.: +1-978-927-5054; fax: +1-978-921-1350.
E-mail address: ware@neb.com (J. Ware).

0020-7519/02/$20.00 © 2002 Published by Elsevier Science Ltd. on behalf of Australian Society for Parasitology.
PII: S0020-7519(01)00367-8

the causative agents of a wide variety of evolutionarily interesting phenomena such as cytoplasmic incompatability, feminisation and parthenogenesis in their hosts (Bourtzis and Braig, 1999; Werren, 1997).

Wolbachia appear to be directly involved in the inflammation and pathogenesis induced by filarial parasites through their lipopolysaccharides (Brattig et al., 2000; Taylor et al., 2000). *Wolbachia* encoded molecules are also recognised by host antibodies, including *Wolbachia* surface protein (*wsp*) (Punkosdy et al., 2001; Bazzocchi et al., 2000b). Anti-rickettsial antibiotic treatments of *Onchocerca volvulus* (and several related filarial parasites) appear to reduce the fertility and fecundity of the worms (Bandi et al., 1999a; Hoerauf et al., 1999, 2000a,b). Recent human trials indicate that the endosymbionts of these filarial parasites are viable drug targets for treatment of filarial disease, as reduction of *Wolbachia* leads to a sustained reduction in microfilaridermia following ivermectin treatment (Hoerauf et al., 2001).

To identify potential novel drug targets we are sequencing the *Wolbachia* genome from *B. malayi*. The genome is estimated to be 1.1 Mb in size (Sun et al., 2001). Physical mapping of the *Wolbachia* genome using bacterial artificial chromosome (BAC) clones is in progress (Guiliano et al., 1999; Williams et al., 2000). Once the physical map is completed, a minimum tiling path of BACs will be used as a template for sequencing of the *Wolbachia* endosymbiont genome from *B. malayi* (Bandi et al., 1999b; Slatko et al., 1999). Purification of *Wolbachia* DNA from *B. malayi*, via pulsed field gel electrophoresis (PFGE) methods will complement these efforts to reconstruct the genome. In parallel with these efforts, purification of the *Wolbachia* endosymbiont from the parasitic hosts *Dirofilaria immitis* and *O. volvulus* is also underway. Sequencing of the *Wolbachia* endosymbiont from *Drosophila melanogaster* is currently being completed at The Institute for Genomic Research (TIGR; http://www.tigr.org). The mapping and sequencing of *Wolbachia* from nematodes will offer novel approaches and targets for drug discovery.

DNA sequencing was initiated on a BAC clone with a 63 kb insert containing part of the *groEL* gene to provide significant preliminary information about *Wolbachia* from *B. malayi*, and to implement a novel DNA sequencing approach for large DNA constructs using transposon DNA sequencing (GPS™-1 Kit, New England Biolabs, Inc.).

2. Methods

Mapping of the *Wolbachia* genome is being done by a BAC approach in the pBeloBAC11 vector, gridded on high density filter arrays (Incyte Genomics), in preparation for full genomic sequencing. The library was made from *Hin*d III partially digested genomic DNA from *B. malayi*, and about 0.5–1% of the library derived from the endosymbiont (Williams et al., 2000). Physical mapping of the *Wolbachia* endosymbiont genome has provided sets of continuous overlapping segments composing over 90% of the endosymbiont genome.

BAC DNA (BMBAC03K15: pBeloBAC11 plus 63 kb insert) was prepared for the transposition reaction using standard alkaline lysis from a 3 ml culture with slight modifications. A general alkaline lysis procedure (Sambrook et al., 1989) was followed using Qiagen solutions P1, P2, and P3 at 4°C for each step, but column purification was not performed which distinguishes this from most current BAC purification protocols. DNA was ethanol precipitated, dried in a vacuum concentrator and resuspended in 20 μl of distilled water to a final concentration of ~100 ng/μl.

2.1. Genome priming system (GPS™-1) reactions

The GPS™-1 Genome Priming System (New England Biolabs, Inc.) provides a simple in vitro method for generating a population of DNA sequencing templates with randomly interspersed primer-binding sites. This system is useful for projects requiring multiple sequencing reactions to complete the entire sequence, providing a faster alternative to primer walking, random subcloning and nested deletion methods. Transprimer insertions show essentially random site preference and only one insertion occurs per target DNA molecule, due to 'target immunity'. Therefore, the in vitro reaction produces a selected population of DNA molecules each containing a transprimer element at a different position. Unique priming sites on both ends of the transprimer element, together with supplied primers (Primer N and Primer S), allow DNA sequence to be obtained from both strands of the target DNA at the position of the insertion.

For the BAC insert GPS reaction, transprimer donor pGPS1.1, specifying kanamycin resistance, was used since the BAC vector contains chloramphenicol resistance. The GPS reaction was optimised with differing amounts of BAC DNA and best results occurred when 5 μl (200–500 ng) of BAC DNA, 2 μl 10 × GPS buffer, 1 μl pGPS1.1 donor (20 ng) and 10 μl dH_2O were utilised in a total reaction volume of 18 μl. The standard GPS protocol was followed (GPS™-1 Instruction Manual, New England Biolabs Inc.; see also http://www.neb.com/neb/products/seq_and_label/seq_frame.html) with the exception that the reaction temperature was 30°C, rather than 37°C.

For electroporation, 1 μl of undiluted GPS reaction was added to 25 μl DH10B ElectroMax™ cells (Life Technologies). Electroporation was performed at 100 Ohms, 2.5 kV, and 25 μF using a Gene Pulser (BioRad Laboratories). One millilitre of LB was then added to the electroporated cells and they were grown for 1 h at 30°C before being plated on small (9 cm diameter) LB agar plates supplemented with kanamycin (25 μg/ml) and chloramphenicol (15 μg/ml) to select for BACs containing the transposon insertions. An aliquot was also grown on LB agar plates containing only chloramphenicol, to verify BAC electroporants. One hundred to 200 μl of the cell growth provided optimal

colony density, (~200 colonies per plate), with several plates used per electroporation. Individual colonies were streaked out on master plates on the correct antibiotic(s) for overnight incubation at 37°C, and then stored at 4°C. These were subsequently used for BAC DNA preparations for sequencing.

Control plasmid (Litmus 28, New England Biolabs) GPS reactions were performed in parallel with the BAC reactions to verify reaction procedure and solutions. The plasmid control was plated on LB agar plates supplemented with ampicillin (100 μg/ml) and kanamycin (25 μg/ml) to select for plasmids containing transposon insertions. Aliquots were also plated on LB plates supplemented with ampicillin to check for BAC viability.

2.2. Sequencing reactions

R.E.A.L. Preps (R.E.A.L.™ Prep 96 Plasmid Kit Manual, Qiagen) were used to prepare BAC DNA from the colonies resulting from the GPS procedure. The 96-well format uses a standard alkaline lysis procedure with a block purification step (Qiagen manual) and provides sufficient DNA for gel quantitation and for sequencing each GPS colony in both directions from the transposon.

To sequence the BAC templates obtained from the GPS reaction, a modified dye-terminator cycle sequencing protocol was used (Boysen et al., 1997). The procedure doubled the standard cycle sequencing dye-terminator reaction from 20 to 40 μl, and used 1 μg BAC DNA, 16 μl premix (PRISM™ Ready Reaction DyeDeoxy™ Terminator Cycle Sequencing Kit with AmpliTaq® DNA Polymerase FS) (Applied Biosystems), 50 pmol primer, and water to a final volume of 40 μl.

Thermal cycling was performed using an MWG Primus 66 plus cycler (MWG-Biotech Inc), or an MJ Research 96 well plate cycler-PTC-200 DNA Engine. The samples were first denatured at 96°C for 4 min and cycled for 35 cycles of 10 s at 96°C, 5 s at 50°C and 4 min at 60°C. After cycling, the reactions were purified using Centrisep® Spin Columns (Princeton Separations) and dried in a vacuum concentrator. The samples were then resuspended in loading dye and electrophoresed on a Model 373 automated DNA Sequencer (Applied Biosystems) using a 34 cm well to read plate with 5.75% Long Ranger® Singel® pack acrylamide gels (BMA). Accutrac™ loading dye (Commonwealth Technologies) was used to aid in sample tracking.

2.3. Sequence analysis

The sequencing results were analysed using ABI programs (Factura™, Editview™ and AutoAssembler™) which allowed the automated removal of ambiguous sequence calls and vector (transposon) sequences and allowed the assembly of the 63.1 kb BAC insert. The GPS™-1 system provided a quick and simple method to obtain about half of the total BAC sequence (see results, below). Primer walking was utilised to join the large contigs, which were formed using the GPS™-1 system. Oligonucleotides were designed to be 24–25 bases in length and were synthesised at the 40 nmole scale using an Applied Biosystems 392 or 394 synthesiser (NEB Organic Synthesis Division). They were used as unpurified (crude) syntheses and were diluted in water to a final concentration of 50 pm/μl for the sequencing reactions.

3. Results and discussion

3.1. GPS analysis

The GPS has historically been used to sequence plasmid DNA and was successfully implemented, in this case, to sequence a 63 kb BAC insert from the *Wolbachia* endosymbiont of *B. malayi*. In the process of developing this method, it was important to determine the optimal number of sequences needed to attain significant coverage of the template without duplication of effort (i.e. production of excessive overlap sequence). We tracked the data by observing increasing BAC coverage as we added random sequences to the contig sets.

For this project, the first 60 GPS transpositions (each sequenced in both direction for a total of 120 reactions) produced 12 contigs between 1 and 4 kb in size (average sequence read length was 500 bases). Twelve sequences (10%) were too ambiguous to assemble (i.e. failed reactions). Seven sequences (~6%) mapped to the BAC vector, agreeing with the relative sizes of the BAC vector and insert. No transprimer insertions were observed at the exact same nucleotide positions.

Additional transpositions from a second electroporation, did not significantly increase the number and size of the contigs. Again, there was no example of the same nucleotide insertion position of the transprimer. The data are suggestive of some regional, but not precise, nucleotide clustering of insertions. The location of the transprimer insertions does not appear to be based on any difference in A + T content at the exact, or local nucleotide sequence. Our results may be specific to this BAC but it appears that for a minimum project start point, the number of transprimer electroporants to be sequenced should roughly equal the kilobase length of the BAC insert (i.e. for a 60 kb BAC, pick 60 GPS electroporants to be sequenced in both directions from the transprimer). After initial assembly, a decision can be made as to whether to use more transprimer electroporants as templates or to initiate a complementary sequencing strategy, such as primer walking.

For this project, we continued sequencing transprimer products in order to follow the distribution of insert positions (data not presented), but effectively finished the sequence with the primer walking strategy from the ends of the GPS-generated sequences. The final sequence was determined with 2–5-fold redundancy sequencing in both directions.

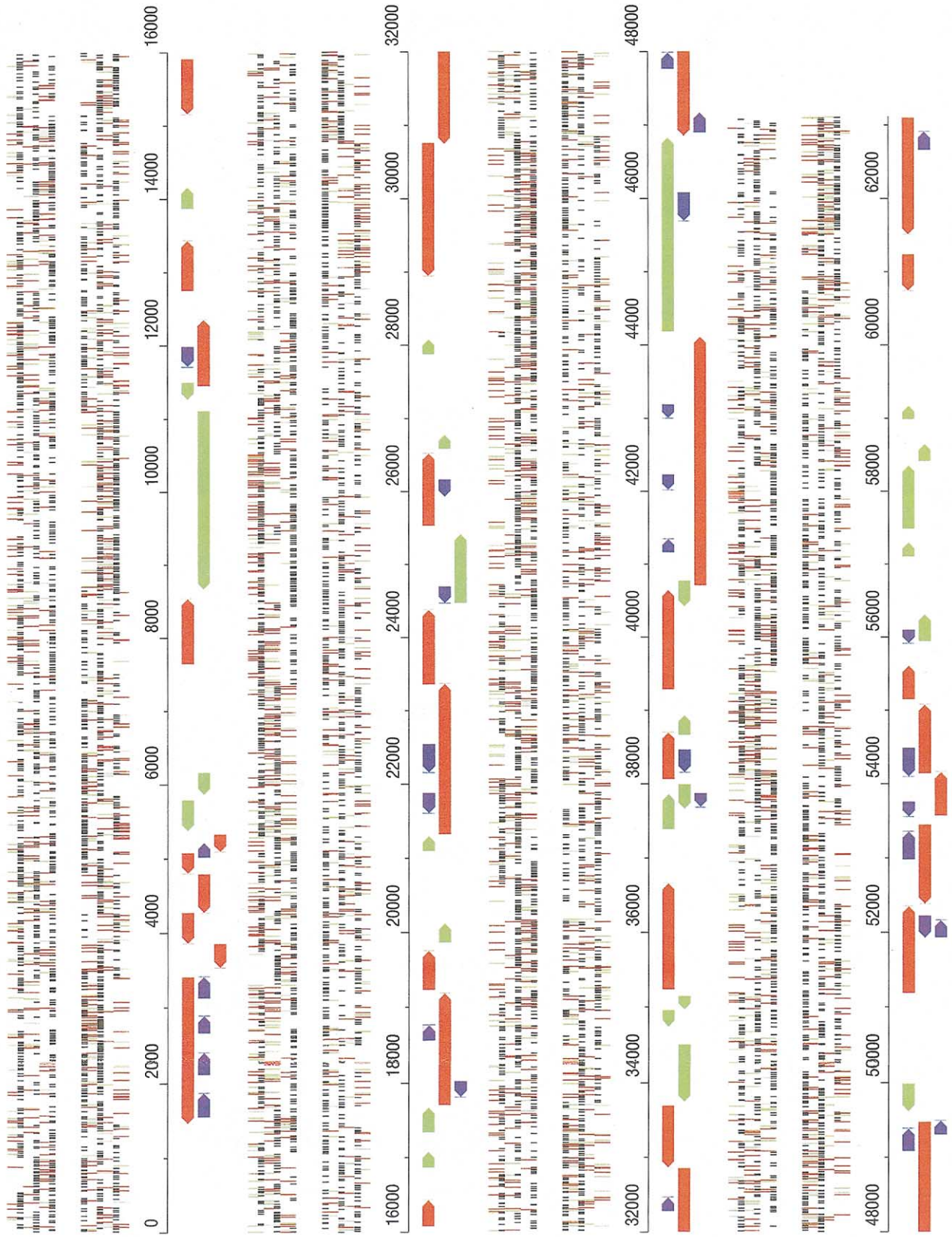
Comparison of the BAC ORF's to the 35 finished microbial genomes

ORF	Rickettsia prowazekii	Caulobacter crescentus	Xylella fastidiosa	Neisseria meningitidis B	Neisseria meningitidis A	Vibrio cholerae	Haemophilus influenzae	Escherichia coli	Buchnera aphidicola	Pseudomonas aeruginosa	Helicobacter pylori 26695	Helicobacter pylori strain J99	Campylobacter jejuni	Treponema pallidum	Borrelia burgdorferi	Chlamydia trachomatis	Chlamydophila pneumoniae J138	Chlamydia pneumoniae	Chlamydophila pneumoniae AR39	Synechocystis	Bacillus subtilis	Mycoplasma genitalium	Ureaplasma urealyticum	Mycoplasma pneumoniae	Mycobacterium tuberculosis	Deinococcus radiodurans	Thermotoga maritima	Aquifex aeolicus	Methanococcus jannaschii	Methanobacterium thermoautotrophicum	Archaeoglobus fulgidus	Pyrococcus horikoshii	Pyrococcus abyssi	Aeropyrum pernix	Sinorhizobium meliloti chrom	Sinorhizobium meliloti A	Sinorhizobium meliloti B	Gene hit
c1453-3402	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	-	-	-	+	-	-	RNA polymerase sigma factor (RPOD-sigma 70)
c3525-3848	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	-	+	+	-	+	+	-	-	-	-	unknown
c3851-4264	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	-	-	-	+	+	-	+	-	-	unknown
c4260-4784	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	+	+	-	+	-	-	unknown
c4789-5073	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	-	+	-	-	hypothetical protein
c5083-5322	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	cation transport system protein
7646-8539	-	+	+	+	+	+	-	+	+	+	+	+	-	+	-	-	-	-	+	+	-	-	-	-	+	+	-	+	+	+	+	+	+	+	+	-	-	aspartate carbamoyltransferase
11450-12358	-	+	+	+	+	-	-	+	-	+	-	-	+	-	+	-	-	-	-	+	+	-	-	-	+	+	+	+	+	+	+	+	+	-	+	-	-	cobalt-zinc-cadm
12749-13435	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	+	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-	-	2-oxoglutarate dehydrogenase E1 component
c15147-15902	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	unknown
16081-16434	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	+	-	-	-	-	-	-	496aa long hypothetical protein
17705-19189	+	+	-	+	+	+	-	+	-	+	+	+	+	-	-	+	+	+	+	+	-	-	-	-	+	-	+	+	+	+	+	+	+	-	+	+	-	conserved hypothetical protein
19233-19757	+	-	-	-	-	-	-	-	-	+	+	+	-	-	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	unknown
21325-23370	+	+	+	+	+	+	-	-	+	+	+	+	+	+	-	-	-	-	-	+	+	-	-	-	+	+	+	+	+	+	-	+	+	-	+	+	-	NADH dehydrogenase I chain G
23357-24379	+	+	+	+	+	+	-	-	+	+	+	+	+	+	-	-	-	-	-	+	-	-	-	-	+	+	+	-	-	+	-	+	+	+	+	-	-	NADH dehydrogenase I chain H
25532-26524	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	-	-	+	+	+	+	-	-	-	+	+	+	-	+	+	+	-	+	+	-	-	-	putative delta aminolevulinic acid dehydratase
c28937-30754	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	-	-	+	-	-	excinuclease ABC subunit
c30742-32847	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	+	-	-	glycyl-tRNAsynthetase beta chain
c32854-33690	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	-	-	-	-	-	+	+	-	-	-	-	-	-	+	-	-	glycyl-tRNAsynthetase alpha chain
35235-36659	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	isocitrate dehydrogenase (NADP)
38064-38699	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	-	-	endonuclease III
39287-40648	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	-	-	unknown
40710-44117	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	probable isoleucyltRNA synthetase
c46846-49470	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	alanyl-tRNA synthetase
51186-52349	+	+	-	+	+	+	-	-	+	-	-	+	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-	+	+	-	-	bicyclomycin resistance protein
c52372-53448	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-	peptide chain release factor
53571-54161	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	+	+	+	+	-	-	+	+	+	-	-	-	-	-	+	-	-	unknown
54139-55077	-	+	-	+	+	-	-	+	+	+	-	-	+	-	-	-	-	-	-	+	+	-	-	-	+	+	+	+	+	-	-	-	+	-	-	acetylglutamate kinase
55147-55614	+	+	+	+	+	-	+	+	+	+	+	-	+	-	+	+	+	+	-	+	+	-	-	-	+	-	+	+	-	-	-	-	-	+	-	-	-	deoxyuridine 5'triphosphate nucleotidohydrolase
c60735-61235	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	thiodisulfide interchange
c61503-63095	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	-	-	-	+	+	+	+	+	60KD chaperonin/groEL

Fig. 1. Open reading frame analysis of the 63 kb *Wolbachia* BAC from *Brugia malayi* BLASTed against the 35 finished microbial genomes. Pluses represent a significant gene match, while minuses represent no database match to the organism with BLAST criteria of a score of 55, or a probability at or below $1e^{-8}$. Those insignificant and overlapping ORF's have been removed from analysis and considered artifacts of the larger, significant ORF's. NCBI finished genomes: *Rickettsia prowazekii*, *Caulobacter crescentus*, *Xylella fastidiosa*, *Neisseria meningitidis* serogroup B strain MC58, *Neisseria meningitidis* serogroup A strain Z2491, *Vibrio cholerae*, *Haemophilus influenzae* Rd, *Escherichia coli* K-12 MG1655, *Buchnera aphidicola*, *Pseudomonas aeruginosa* PA01, *Helicobacter pylori* 26695, *Helicobacter pylori* strain J99, *Campylobacter jejuni*, *Treponema pallidum*, *Borrelia burgdorferi*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae* J138, *Chlamydia pneumoniae*, *Chlamydophila pneumoniae* AR39, *Synechocystis* PCC6803, *Bacillus subtilis*, *Mycoplasma genitalium* G37, *Ureaplasma urealyticum*, *Mycoplasma pneumoniae* M129, *Mycobacterium tuberculosis*, *Deinococcus radiodurans*, *Thermotoga maritima*, *Aquifex aeolicus*, *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum* delta H, *Archaeoglobus fulgidus*, *Pyrococcus horikoshii* OT3, *Pyrococcus abyssi*, *Aeropyrum pernix* K1. From Galibert et al. (2001): *Sinorhizobium meliloti*: chromosome, *Sinorhizobium melitoti*: pSymA, *Sinorhizobium melitoti*: pSymB.

3.2. Open reading frame analysis

The BMBAC03K15 insert was determined to be 63,116 nucleotides in length with an overall A + T content of 66%. The insert does not appear to show a difference in A + T content between coding and non-coding regions, as defined by the open reading frames. Using computer programs designed at NEB, the sequences and positions of potential ORFs were identified and analysed. These programs included a frameshift analyser to help ensure that no DNA sequencing mistakes had occurred in the data set. An ORF is arbitrarily defined as a sequence region with an ATG or GTG start, with a minimum of 60 amino acids in length before a termination codon. In order to help identify each ORF, we used NCBI BLAST programs to compare each potential ORF with protein sequences (BLASTX), nucleic acid sequences (BLASTN) and EST sequences (dbEST, BLASTN) (NCBI; http://www.ncbi.nlm.nih.gov). An automated script performed BLASTN, BLASTX, and dbestBLAST searches with each ORF. We also performed BLASTX only searches against each of the 35 individual microbial genomes whose sequences were completed (available at the TIGR Compre-


0
2000
4000
6000
8000
10000
12000
14000
16000
16000
18000
20000
22000
24000
26000
28000
30000
32000
32000
34000
36000
38000
40000
42000
44000
46000
48000
48000
50000
52000
54000
56000
58000
60000
62000

hensive Microbial Resource, http://www.tigr.org and NCBI GenBank, http://www.ncbi.nlm.nih.gov) (Fig. 1), as well as searches against 22 genomes that were in progress at the DOE Joint Genome Institute (DOE/JGI, http://www.jgi.doe.gov) (data not presented). Comparisons between the currently incomplete data set of *Wolbachia* from *Drosophila* (in progress at TIGR) and the sequence from the BMBAC03K15 *Wolbachia* BAC insert were also performed. All these results were utilised to assign putative identifications to ORFs where appropriate data were available based on a minimum individual database BLAST score of 55 or a probability at or below $1e^{-8}$. These sensitive parameters allow for all potential database matches to be recognised, as each individual organism database was BLASTed separately against the ORFs.

Those ORFs with no significant BLAST hit to any of the 58 genomes and which overlapped another ORF of significant size and positive ID, were removed from further ORF analysis (but are present in the overall ORF figure). While it is possible that these are actually coding sequences, for subsequent analysis these ORFs were considered artifacts encompassed within the larger ORF sequences. It is also important to note that those ORFs that hit only one (or two if *Wolbachia* from *Drosophila* was hit as well) of the 58 genomes were further examined and those few that were found to have minimal identity matches (<30%) and/or only to short stretches of sequence were considered as ORFs with no database hits.

Overall, there appear to be 57 potential ORFs within the 63 kb insert, with a frequency of just under 1 ORF/kb (Fig. 2). Among these 57 ORFs, 31 (54%) show a significant sequence similarity to the databases of completed microbial genomes. Of these 31, however, 10 (17.5% of the 57 ORFs) do not have a known function and show similarities only to hypothetical proteins or proteins of unknown function among the annotated microbial genomes.

Of the remaining putatively identified ORFs in the BMBAC03K15 insert, 26/57 (46%) do not show significant similarity to any sequences in the completed microbial genomes databases. This provides possible evidence of *Wolbachia* genome specific ORFs that have no matches to any of the published sequences.

Comparison to the α-proteobacteria *Rickettsia prowazekii* (Andersson et al., 1998), the most closely related, fully sequenced microbial genome, provides 25 matches. All the ORFs shared by the *Wolbachia* from *Brugia* and *R. prowazekii*, are also shared with other organisms in the microbial genome data set. There are six ORFs in the *Wolbachia* data set which are not present in *R. prowazekii* but which are present in other microbial organisms. Analysis by comparison to only the *R. prowazekii* genome would have missed these similarities. The newly sequenced α-proteobacteria *Sinorhizobium meliloti,* which is also an endosymbiont (symbiont of alfalfa), provides 25 gene hit matches to the *Wolbachia* ORFs (Galibert et al., 2001). There are also 22 putative ORF matches to genes from the closely related α-proteobacteria *Caulobacter crescentus* (Nierman et al., 2001).

Although there is significant gene similarity, there is no long-range synteny with the *R. prowazekii* genome (data not presented). The gross gene order has not been conserved between the *Wolbachia* from *Brugia* and the *Rickettsial* genome. Large scale synteny conservation with *S. meliloti* is also not observed, but there are multiple small regions of three–four genes showing conservation of gene order (data not presented). The organisation of *C. crescentus* also does not show synteny in comparison to the *Wolbachia* from *Brugia* ORFs. A similar result has been observed in comparison to an additional endosymbiont genome, that of the gamma-proteobacteria *Buchnera aphidocola* (Shigenobu et al., 2000) (data not presented).

Comparison of the ORFs to the 22 DOE/JGI unfinished genomes provides results consistent with the data from the 35 finished genomes (data not presented). While the unfinished genomes have not been annotated, the ORFs with significant BLAST similarity scores to the 35 finished genomes also have consistent, significant matches to the unfinished genomes as well.

Preliminary sequence data are available for the *Wolbachia* genome from *D. melanogaster,* currently being sequenced at TIGR. Comparisons between the incomplete data set and the sequence from this *Wolbachia* BAC insert indicate strong similarity and some conservation of synteny (data not presented) between them. Of the potential *Wolbachia* ORFs from *Brugia* 41/57 are present in the *Wolbachia* from *Drosophila* data set (72%) and have significantly strong matches. Of these, 15 ORFs are shared by only these two *Wolbachia* species and have no similarity to any other microbial sequences, which could indicate these are *Wolbachia* specific. An independent set of 275 sequences has been generated from the *Wolbachia* from *B. malayi* during the creation of a BAC physical map. These sequences were also compared with the *Wolbachia* from *D. melanogaster* and the results confirm the percent identity found here. Once the *Wolbachia* from *Drosophila* genome sequencing and annotation have been completed, more precise information will be available.

As of now, there are 16/57 potential ORFs in the *Wolbachia* from *Brugia* BAC insert not present in the *Wolbachia*

Fig. 2. A total of 63.1 kb BAC insert sequence. Open reading frames (ORFs) are designated as heavy arrows pointing in the direction of transcription, and are staggered. The heavy red arrows indicate those ORFs that have positive gene hits to the databases, while the heavy green arrows indicate those ORFs that do not hit anything in the databases. Those heavy arrows shaded in blue are the ORFs that are not considered coding and have been removed from analysis. The top set of three stripes represents the forward strand and the bottom set represents the complementary strand. Positions at stop codons are indicated as thin black lines. Red and green thin stripe marks indicate ATG and GTG starts, respectively.

from *Drosophila* data set. Ten of these 16 potential ORFs fail to show similarity to any other currently sequenced (or unfinished DOE/JGI genomes) microbial organism. These might be ORFs specific to the *Wolbachia* endosymbiont in *Brugia* only. Further analysis of these ORFs (which range from 200 to 340 bp) will be necessary to verify whether these are gene coding regions.

4. Conclusion

The complete sequencing of a 63.1 kb insert from the *Wolbachia* endosymbiont of *B. malayi* has provided significant comparative genomic information that will aid in identification of potential novel targets for drug discovery and treatment of filarial disease. This has been done using a new and simple method for large clone sequencing with the NEB GPS™-1 kit, in conjunction with primer walking. Using the GPS™-1 kit and a minimum start point of one transposition per 1 kb of BAC sequence, significant sequence data provides a framework to continue on with more electroporants, ending with a finishing strategy, such as primer walking.

The sequencing of the BMBAC03K15 63.1 kb insert has provided important data, such as gene density, A + T content and gene similarity data to other microbial organisms and will contribute to completion of a physical map and ultimately the complete genome sequence of the *Wolbachia* from *B. malayi*. Other techniques such as microarray and/or proteomics approaches will provide verification of the ORFs discovered in this preliminary analysis. This information will allow more identification of commonality between *Wolbachia* in *B. malayi* and *Wolbachia* in *D. melanogaster* as well as similarity to other published and unfinished microbial genomes. These data will not only provide targets for the rational design of *Wolbachia* specific drugs but also enable more detailed molecular analysis of host/endosymbiont interactions in various organisms of interest. It will also allow us to take advantage of, and further enhance, studies of host antibody responses and anti-*Wolbachia* treatments that will reduce worm burdens and help to eliminate filarial disease.

Acknowledgements

Funding for this project has been provided by internal funding from New England Biolabs, Inc., and a World Health Organization (TDR/UNDP/World Bank) grant. We wish to thank Fiona Stewart for all of her help and suggestions for optimising the GPS conditions. We also wish to thank Donald Comb for support of this project.

References

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C., Podowski, R.M., Naslund, A.K., Eriksson, A.S., Winkler, H.H., Kurland, C.G., 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396, 133–40.

Bandi, C., Anderson, T.J.C., Genchi, C., Blaxter, M.L., 1998. Phylogeny of *Wolbachia* bacteria in filarial nematodes. Proc. R. Soc. Lond. B265, 2407–13.

Bandi, C., McCall, J.W., Genchi, C., Corona, S., Venco, L., Sacci, L., 1999a. Effects of tetracycline on the filarial worms *Brugia pahangi* and *Dirofilaria immitis* and their bacterial endosymbionts, *Wolbachia*. Int. J. Parasitol. 29, 357–64.

Bandi, C., Slatko, B., O'Neill, S., 1999b. *Wolbachia* genomes and the many faces of symbiosis. Parasitol. Today 15, 428–9.

Bazzocchi, C., Jamnongluk, W., O'Neill, S.L., Anderson, T.J., Genchi, C., Bandi, C., 2000a. *wsp* gene sequences from the *Wolbachia* of filarial nematodes. Curr. Microbiol. 41, 96–100.

Bazzocchi, C., Ceciliani, F., McCall, J.W., Ricci, I., Genchi, C., Bandi, C., 2000b. Antigenic role of the endosymbionts of filarial nematodes: IgG response against the *Wolbachia* surface protein in cats infected with *Dirofilaria immitis*. Proc. R. Soc. Lond. B267, 2511–6.

The Filarial Genome Project, Blaxter, M.L., Aslett, M., Daub, J., Guiliano, D., 1999. Parasitic helminth genomics. Parasitology 118, 339–51.

Bourtzis, K., Braig, H.R., 1999. The many faces of Wolbachia. rickettsiae and rickettsial diseases at the turn of the third millennium. In: Raoult, D., Brouqui, P. (Eds.). Elsevier, Paris, pp. 199–219.

Boysen, C., Simon, M., Hood, L., 1997. Fluorescence-based sequencing directly from bacterial and P1-derived artificial chromosomes. Biotechniques 23, 978–82.

Brattig, N.W., Rathjens, U., Ernst, M., Geisinger, F., Renz, A., Tischendorf, F.W., 2000. Lipopolysaccharide-like molecules derived from *Wolbachia* endobacteria of the filaria *Onchocerca volvulus* are candidate mediators in the sequence of inflammatory and anti-inflammatory responses of human monocytes. Microbes Infect. 2, 1147–57.

Filarial Genome Project, 1999. Deep within the filarial genome: an update on the progress of the filarial genome project. Parasitol. Today 15, 219–24.

Galibert, F., Finan, T.M., Long, S.R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M.J., Becker, A., Boistard, P., Bothe, G., Boutry, M., Bowser, L., Buhrmester, J., Cadieu, E., Capela, D., Chain, P., Cowie, A., Davis, R.W., Dreano, S., Federspiel, N.A., Fisher, R.F., Gloux, S., Godrie, T., Goffeau, A., Golding, B., Gouzy, J., Gurjal, M., Hernandez-Lucas, I., Hong, A., Huizar, L., Hyman, R.W., Jones, T., Kahn, D., Kahn, M.L., Kalman, S., Keating, D.H., Kiss, E., Komp, C., Lelaure, V., Masuy, D., Palm, C., Peck, M.C., Pohl, T.M., Portetelle, D., Purnelle, B., Ramsperger, U., Surzycki, R., Thebault, P., Vandenbol, M., Vorholter, F.J., Weidner, S., Wells, D.H., Wong, K., Yeh, K.C., Batut, J., 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science 293, 668–72.

Guiliano, D., Ganatra, M., Ware, J., Parrot, J., Daub, J., Moran, L., Brennecke, H., Foster, J.M., Supali, T., Blaxter, M.L., Scott, A.L., Williams, S.A., Slatko, B.E., 1999. Chemiluminescent detection of sequential DNA hybridisations to high density filter arrayed cDNA libraries: a subtraction method for novel gene discovery. Biotechniques 27, 146–52.

Hoerauf, A., Nissen-Pahle, K., Schmetz, C., Henkle-Duhrsen, K., Blaxter, M.L., Buttner, D., Alquaoud, K.M., Lucius, R., Fleisher, B., 1999. Tetracycline therapy targets intracellular bacteria in the filarial nematode *Litomosoides sigmodontis* and results in filarial infertility. J. Clin. Invest. 103, 11–18.

Hoerauf, A., Volkmann, L., Hamelmann, C., Adjei, O., Autenrieth, I.B., Fleischer, B., Buttner, D.W., 2000a. Endosymbiotic bacteria in worms as targets for a novel chemotherapy in filariasis. Lancet 355, 1242–3.

Hoerauf, A., Volkmann, L., Nissen-Paehle, K., Schmetz, C., Autenrieth, I., Buttner, D.W., Fleisher, B., 2000b. Targeting of *Wolbachia* endobacteria in *Litomosoides sigmodontis*: comparison of tetracyclines with chloramphenicol, macrolides and ciprofloxacin. Trop. Med. Int. Health 5, 275–9.

Hoerauf, A., Mand, S., Adjei, O., Fleischer, B., Buttner, D.W., 2001. Deple-

tion of *Wolbachia* endobacteria in *Onchocerca volvulus* by doxycycline and microfilaridermia after ivermectin treatment. Lancet 357, 1415–6.

Jeyaprakash, A., Hoy, M.A., 2000. Long PCR improves *Wolbachia* DNA amplification: *wsp* sequences found in 76% of sixty-three arthropod species. Ins. Mol. Bio. 9, 393–405.

Nierman, W.C., Feldblyum, T.V., Laub, M.T., Paulsen, I.T., Nelson, K.E., Eisen, J., Heidelberg, J.F., Alley, M.R., Ohta, N., Maddock, J.R., Potocka, I., Nelson, W.C., Newton, A., Stephens, C., Phadke, N.D., Ely, B., DeBoy, R.T., Dodson, R.J., Durkin, A.S., Gwinn, M.L., Haft, D.H., Kolonay, J.F., Smit, J., Craven, M.B., Khouri, H., Shetty, J., Berry, K., Utterback, T., Tran, K., Wolf, A., Vamathevan, J., Ermolaeva, M., White, O., Salzberg, S.L., Venter, J.C., Shapiro, L., Fraser, C.M., 2001. Complete genome sequence of *Caulobacter crescentus*. Proc. Natl. Acad. Sci. 98, 4136–41.

Ottesen, E.A., Ramachandran, C.D., 1995. Lymphatic filariasis infection and disease: control strategies. Parasitol. Today 11, 129–31.

Punkosdy, G.A., Dennis, V.A., Lasater, B.L., Tzertzinis, G., Foster, J.M., Lammie, P.J., 2001. Detection of serum IgG antibodies specific for *Wolbachia* surface protein in rhesus monkeys infected with *Brugia malayi*. J. Infect. Dis. 184, 385–9.

Sambrook, J., Fritsh, E.F., Maniatis, T., 1989. Molecular Cloning: a Laboratory Manual, . 2nd EditionCold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., Ishikawa, H., 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407, 81–86.

The *Wolbachia* Genome Consortium, Slatko, B.E., O'Neill, S.L., Scott, A.L., Werren, J.L., Blaxter, M.L., 1999. The *Wolbachia* genome consortium. Microbial Comparative Genomics 4, 161–5.

Sun, L.V., Foster, J.M., Tzertzinis, G., Ono, M., Bandi, C., Slatko, B.E., O'Neill, S.L., 2001. Determination of *Wolbachia* genome size by pulsed-field gel electrophoresis. J. Bacteriol. 183, 2219–25.

Taylor, M.J., Bilo, K., Cross, H.F., Archer, J.P., Underwood, A.P., 1999. 16S rDNA phylogeny and ultrastructure characterisation of *Wolbachia* intracellular bacteria of the filarial nematodes *Brugia malayi*, *B. pahangi* and *Wucheria bancrofti*. Exp. Parasitol. 91, 356–61.

Taylor, M.J., Cross, H.F., Bilo, K., 2000. Inflammatory responses induced by the filarial nematode *Brugia malayi* are mediated by lipopolysaccharide-like activity from endosymbiotic *Wolbachia* bacteria. J. Exp. Med. 191, 1429–36.

Werren, J.H., 1997. Biology of *Wolbachia*. Annu. Rev. Entemol. 42, 587–609.

The Filarial Genome Project and D.A. Johnston and the Schistosome Genome Project, Williams, S.A., 1999. Helminth genome analysis: the current status of the filarial and schistosome genome projects. Parasitology 118, S19–38.

Williams, S.A., Lizotte-Waniewski, M., Foster, J., Guiliano, D., Daub, J., Scott, A., Slatko, B., Blaxter, M., 2000. The filarial genome project: analysis of the nuclear, mitochondrial and endosymbiont genomes of *Brugia malayi*. Int. J. Parasitol. 30, 411–9.