Phylogenomics of the Reproductive Parasite *Wolbachia pipientis w*Mel: A Streamlined Genome Overrun by Mobile Genetic Elements

Martin Wu[1], Ling V. Sun[2], Jessica Vamathevan[1], Markus Riegler[3], Robert Deboy[1], Jeremy C. Brownlie[3], Elizabeth A. McGraw[3], William Martin[4], Christian Esser[4], Nahal Ahmadinejad[4], Christian Wiegand[4], Ramana Madupu[1], Maureen J. Beanan[1], Lauren M. Brinkac[1], Sean C. Daugherty[1], A. Scott Durkin[1], James F. Kolonay[1], William C. Nelson[1], Yasmin Mohamoud[1], Perris Lee[1], Kristi Berry[1], M. Brook Young[1], Teresa Utterback[1], Janice Weidman[1], William C. Nierman[1], Ian T. Paulsen[1], Karen E. Nelson[1], Hervé Tettelin[1], Scott L. O'Neill[2,3], Jonathan A. Eisen[1*]

1 The Institute for Genomic Research, Rockville, Maryland, United States of America, **2** Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut, United States of America, **3** Department of Zoology and Entomology, School of Life Sciences, The University of Queensland, St Lucia, Queensland, Australia, **4** Institut für Botanik III, Heinrich-Heine Universität, Düsseldorf, Germany

The complete sequence of the 1,267,782 bp genome of *Wolbachia pipientis w*Mel, an obligate intracellular bacteria of *Drosophila melanogaster*, has been determined. *Wolbachia*, which are found in a variety of invertebrate species, are of great interest due to their diverse interactions with different hosts, which range from many forms of reproductive parasitism to mutualistic symbioses. Analysis of the *w*Mel genome, in particular phylogenomic comparisons with other intracellular bacteria, has revealed many insights into the biology and evolution of *w*Mel and *Wolbachia* in general. For example, the *w*Mel genome is unique among sequenced obligate intracellular species in both being highly streamlined and containing very high levels of repetitive DNA and mobile DNA elements. This observation, coupled with multiple evolutionary reconstructions, suggests that natural selection is somewhat inefficient in *w*Mel, most likely owing to the occurrence of repeated population bottlenecks. Genome analysis predicts many metabolic differences with the closely related *Rickettsia* species, including the presence of intact glycolysis and purine synthesis, which may compensate for an inability to obtain ATP directly from its host, as *Rickettsia* can. Other discoveries include the apparent inability of *w*Mel to synthesize lipopolysaccharide and the presence of the most genes encoding proteins with ankyrin repeat domains of any prokaryotic genome yet sequenced. Despite the ability of *w*Mel to infect the germline of its host, we find no evidence for either recent lateral gene transfer between *w*Mel and *D. melanogaster* or older transfers between *Wolbachia* and any host. Evolutionary analysis further supports the hypothesis that mitochondria share a common ancestor with the α-Proteobacteria, but shows little support for the grouping of mitochondria with species in the order Rickettsiales. With the availability of the complete genomes of both species and excellent genetic tools for the host, the *w*Mel–*D. melanogaster* symbiosis is now an ideal system for studying the biology and evolution of *Wolbachia* infections.

Introduction

Wolbachia are intracellular gram-negative bacteria that are found in association with a variety of invertebrate species, including insects, mites, spiders, terrestrial crustaceans, and nematodes. *Wolbachia* are transovarialy transmitted from females to their offspring and are extremely widespread, having been found to infect 20%–75% of invertebrate species sampled (Jeyaprakash and Hoy 2000; Werren and Windsor 2000). *Wolbachia* are members of the Rickettsiales order of the α-subdivision of the Proteobacteria phyla and belong to the Anaplasmataceae family, with members of the genera *Anaplasma*, *Ehrlichia*, *Cowdria*, and *Neorickettsia* (Dumler et al. 2001). Six major clades (A–F) of *Wolbachia* have been identified to date (Lo et al. 2002): A, B, E, and F have been reported from insects, arachnids, and crustaceans; C and D from filarial nematodes.

Wolbachia–host interactions are complex and range from mutualistic to pathogenic, depending on the combination of host and *Wolbachia* involved. Most striking are the various forms of "reproductive parasitism" that serve to alter host reproduction in order to enhance the transmission of this maternally inherited agent. These include parthenogenesis (infected females reproducing in the absence of mating to produce infected female offspring), feminization (infected males being converted into functional phenotypic females), male-killing (infected male embryos being selectively killed), and cytoplasmic incompatibility (in its simplest form, the

Received November 19, 2003; Accepted January 6, 2004; Published March 16, 2004
DOI: 10.1371/journal.pbio.0020069

Abbreviations: CDS, coding sequence; ENc, effective number of codons; IS, insertion sequence; LPS, lipopolysaccharide; RT, reverse transcription; TIGR, The Institute for Genomic Research

Academic Editor: Nancy A. Moran, University of Arizona

* To whom correspondence should be addressed. E-mail: jeisen@tigr.org

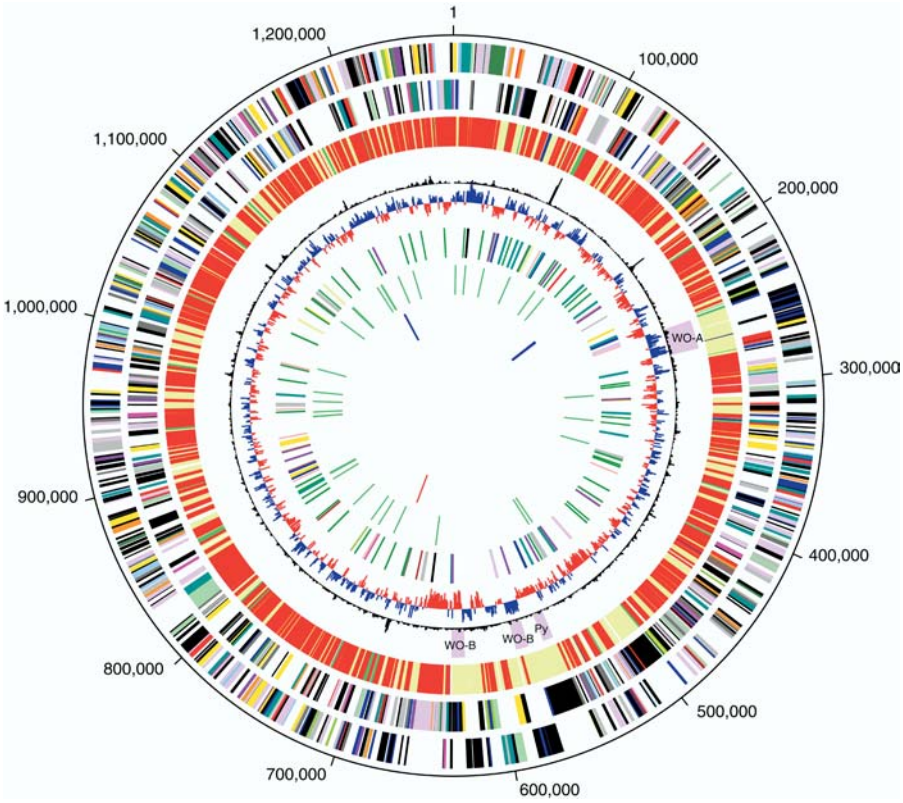

Figure 1. Circular Map of the Genome and Genome Features

Circles correspond to the following: (1) forward strand genes; (2) reverse strand genes, (3) in red, genes with likely orthologs in both *R. conorii* and *R. prowazekii*; in blue, genes with likely orthologs in *R. prowazekii*, but absent from *R. conorii*; in green, genes with likely orthologs in *R. conorii* but absent from *R. prowazekii*; in yellow, genes without orthologs in either *Rickettsia* (Table S3); (4) plot is of χ^2 analysis of nucleotide composition; phage regions are in pink; (5) plot of GC skew (G–C)/(G+C); (6) repeats over 200 bp in length, colored by category; (7) in green, transfer RNAs; (8) in blue, ribosomal RNAs; in red, structural RNA.

DOI: 10.1371/journal.pbio.0020069.g001

developmental arrest of offspring of uninfected females when mated to infected males) (O'Neill et al. 1997a).

Wolbachia have been hypothesized to play a role in host speciation through the reproductive isolation they generate in infected hosts (Werren 1998). They also provide an intriguing array of evolutionary solutions to the genetic conflict that arises from their uniparental inheritance. These solutions represent alternatives to classical mutualism and are often of more benefit to the symbiont than the host that is infected (Werren and O'Neill 1997). From an applied perspective, it has been proposed that *Wolbachia* could be utilized to either suppress pest insect populations or sweep desirable traits into pest populations (e.g., the inability to transmit disease-causing pathogens) (Sinkins and O'Neill 2000). Moreover, they may provide a new approach to the control of human and animal filariasis. Since the nematode worms that cause filariasis have an obligate symbiosis with mutualistic *Wolbachia*, treatment of filariasis with simple antibiotics that target *Wolbachia* has been shown to eliminate microfilaria production as well as ultimately killing the adult worm (Taylor et al. 2000; Taylor and Hoerauf 2001).

Despite their common occurrence and major effects on host biology, little is currently known about the molecular mechanisms that mediate the interactions between *Wolbachia* and their invertebrate hosts. This is partly due to the difficulty of working with an obligate intracellular organism that is difficult to culture and hard to obtain in quantity. Here we report the completion and analysis of the genome sequence of *Wolbachia pipientis* *w*Mel, a strain from the A supergroup that naturally infects *Drosophila melanogaster* (Zhou et al. 1998).

Table 1. *w*Mel Genome Features

Genome size	1,267,782
Predicted CDS	1,270
Average gene length	852
Percent coding	85.4%
Assigned function	719 (56.6%)
Conserved hypothetical	123 (9.7%)
Unknown function	91 (7.2%)
Hypothetical	337 (26.5%)
Transfer RNA	34
Ribosomal RNA	1 each of 5S, 16S, 23S
Structural RNAs	2
Prophage	3
GC content	35.2%

DOI: 10.1371/journal.pbio.0020069.t001

Results/Discussion

Genome Properties

The *w*Mel genome is determined to be a single circular molecule of 1,267,782 bp with a G+C content of 35.2%. This assembly is very similar to the genetic and physical map of the closely related strain *w*MelPop (Sun et al., 2003). The genome does not exhibit the GC skew pattern typical of some prokaryotic genomes (Figure 1) that have two major shifts, one near the origin and one near the terminus of replication. Therefore, identification of a putative origin of replication and the assignment of basepair 1 were based on the location

Table 2. *w*Mel DNA Repeats of Greater than 200 bp

Repeat Class	Size (Median)	Copies	Protein Motifs/ Families	IS Family	Possible Terminal Inverted Repeat Sequence
1	1512	3	Transposase	IS4	5′-ATACGCGTCAAGTTAAG-3′
2	360	12	—	New	5′-GGCTTTGTTGCATCGCTA-3′
3	858	9	Transposase	IS492/IS110	5′-GGCTTTGTTGCAT-3′
4	1404.5	4	Conserved hypothetical, phage terminase	New	5′-ATACCGCGAWTSAWTCGCGGTAT-3′
5	1212	15	Transposase	IS3	5′-TGACCTTACCCAGAAAAAGTGGAGAGAAAG-3′
6	948	13	Transposase	IS5	5′-AGAGGTTGTCCGGAAACAAGTAAA-3′
7	2405.5	8	RT/maturase	—	
8	468	45	—	—	
9	817	3	Conserved hypothetical, transposase	ISBt12	
10	238	2	ExoD	—	
11	225	2	RT/maturase	—	
12	1263	4	Transposase	???	
13	572.5	2	Transposase	???	
14	433	2	Ankyrin	—	
15	201	2	—	—	
16	1400	6	RT/maturase	—	
17	721	2	Transposase	IS630	
18	1191.5	2	EF-Tu	—	
19	230	2	Hypothetical	—	

DOI: 10.1371/journal.pbio.0020069.t002

of the *dnaA* gene. Major features of the genome and of the annotation are summarized in Table 1 and Figure 1.

Repetitive and Mobile DNA

The most striking feature of the *w*Mel genome is the presence of very large amounts of repetitive DNA and DNA corresponding to mobile genetic elements, which is unique for an intracellular species. In total, 714 repeats of greater than 50 bp in length, which can be divided into 158 distinct families (Table S1), were identified. Most of the repeats are present in only two copies in the genome, although 39 are present in three or more copies, with the most abundant repeat being found in 89 copies. We focused our analysis on the 138 repeats of greater than 200 bp (Table 2). These were divided into 19 families based upon sequence similarity to each other. These repeats were found to make up 14.2 % of the *w*Mel genome. Of these repeat families, 15 correspond to likely mobile elements, including seven types of insertion sequence (IS) elements, four likely retrotransposons, and four families without detectible similarity to known elements but with many hallmarks of mobile elements (flanked by inverted repeats, present in multiple copies) (Table 2). One of these new elements (repeat family 8) is present in 45 copies in the genome. It is likely that many of these elements are not able to autonomously transpose since many of the transposase genes are apparently inactivated by mutations or the insertion of other transposons (Table S2). However, some are apparently recently active since there are transposons inserted into at least nine genes (Table S2), and the copy number of some repeats appears to be variable between *Wolbachia* strains (M. Riegler et al., personal communication). Thus, many of these repetitive elements may be useful markers for strain discrimination. In addition, the mobile elements likely contribute to generating the diversity of phenotypically distinct *Wolbachia* strains (e.g., mod$^-$ strains [McGraw et al. 2001]) by altering or disrupting gene function (Table S2).

Three prophage elements are present in the genome. One is a small pyocin-like element made up of nine genes (WD00565–WD00575). The other two are closely related to and exhibit extensive gene order conservation with the WO phage described from *Wolbachia* sp. *w*Kue (Masui et al. 2001) (Figure 2). Thus, we have named them *w*Mel WO-A and WO-B, based upon their location in the genome. *w*Mel WO-B has undergone a major rearrangement and translocation, suggesting it is inactive. Phylogenetic analysis indicates that *w*Mel WO-B is more closely related to the *w*Kue WO than to *w*Mel WO-A (Figure S1). Thus, *w*Mel WO-A likely represents either a separate insertion event in the *Wolbachia* lineage or a duplication that occurred prior to the separation of the *w*Mel and *w*Kue lineages. Phylogenetic analysis also confirms the proposed mosaic nature of the WO phage (Masui et al. 2001), with one block being closely related to lambdoid phage and another to P2 phage (data not shown).

Genome Structure: Rearrangements, Duplications, and Deletions

The irregular pattern of GC skew in *w*Mel is likely due in part to intragenomic rearrangements associated with the many DNA repeat elements. Comparison with a large contig from a *Wolbachia* species that infects *Brugia malayi* is consistent with this (Ware et al. 2002) (Figure 3). While only trans-

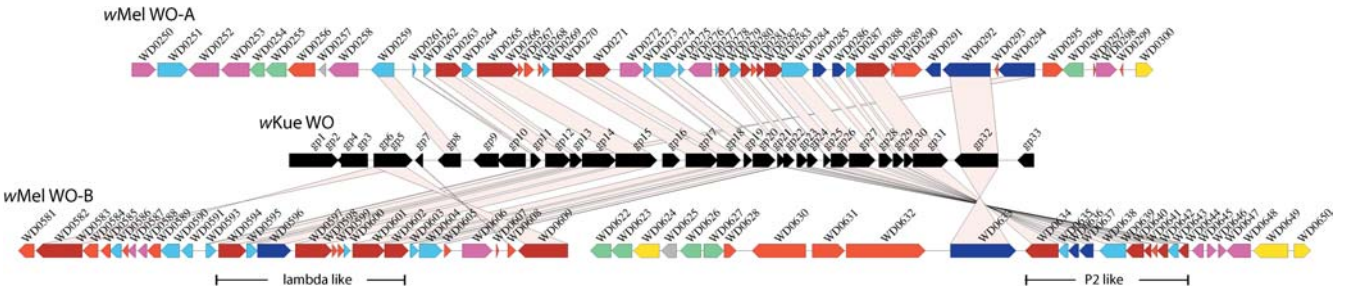

Figure 2. Phage Alignments and Neighboring Genes

Conserved gene order between the WO phage in *Wolbachia* sp. *w*Kue and prophage regions of *w*Mel. Putative proteins in *w*Kue (Masui et al. 2001) were searched using TBLASTN against the *w*Mel genome. Matches with an *E*-value of less than $1e^{-15}$ are linked by connecting lines. CDSs are colored as follows: brown, phage structural or replication genes; light blue, conserved hypotheticals; red, hypotheticals; magenta, transposases or reverse transcriptases; blue, ankyrin repeat genes; light gray, *radC*; light green, paralogous genes; gold, others. The regions surrounding the phage are shown because they have some unusual features relative to the rest of the genome. For example, WO-A and WO-B are each flanked on one side by clusters of genes in two paralogous families that are distantly related to phage repressors. In each of these clusters, a homolog of the *radC* gene is found. A third *radC* homolog (WD1093) in the genome is also flanked by a member of one of these gene families (WD1095). While the connection between *radC* and the phage is unclear, the multiple copies of the *radC* gene and the members of these paralogous families may have contributed to the phage rearrangements described above.

DOI: 10.1371/journal.pbio.0020069.g002

locations are seen in this plot, genetic comparisons reveal that inversions also occur between strains (Sun et al., 2003), which is consistent with previous studies of prokaryotic genomes that have found that the most common large-scale rearrangements are inversions that are symmetric around the origin of DNA replication (Eisen et al. 2000). The occurrence of frequent rearrangement events during *Wolbachia* evolution is supported by the absence of any large-scale conserved gene order with *Rickettsia* genomes. The rearrangements in *Wolbachia* likely correspond with the introduction and massive expansion of the repeat element families that could serve as sites for intragenomic recombination, as has been shown to occur for some other bacterial species (Parkhill et al. 2003). The rearrangements in *w*Mel may have fitness consequences since several classes of genes often found in clusters are generally scattered throughout the *w*Mel genome (e.g., ABC transporter subunits, Sec secretion genes, rRNA genes, F-type ATPase genes).

Although the common ancestor of *Wolbachia* and *Rickettsia* likely already had a reduced, streamlined genome, *w*Mel has lost additional genes since that time (Table S3). Many of these recent losses are of genes involved in cell envelope biogenesis in other species, including most of the machinery for producing lipopolysaccharide (LPS) components and the alanine racemase that supplies D-alanine for cell wall synthesis. In addition, some other genes that may have once been involved in this process are present in the genome, but defective (e.g., mannose-1-phosphate guanylyltransferase, which is split into two coding sequences [CDSs], WD1224 and WD1227, by an IS5 element) and are likely in the process of being eliminated. The loss of cell envelope biogenesis genes has also occurred during the evolution of the *Buchnera* endosymbionts of aphids (Shigenobu et al. 2000; Moran and Mira 2001). Thus, *w*Mel and *Buchnera* have lost some of the same genes separately during their reductive evolution. Such convergence means that attempts to use gene content to infer evolutionary relatedness needs to be interpreted with caution. In addition, since *Anaplasma* and *Ehrlichia* also

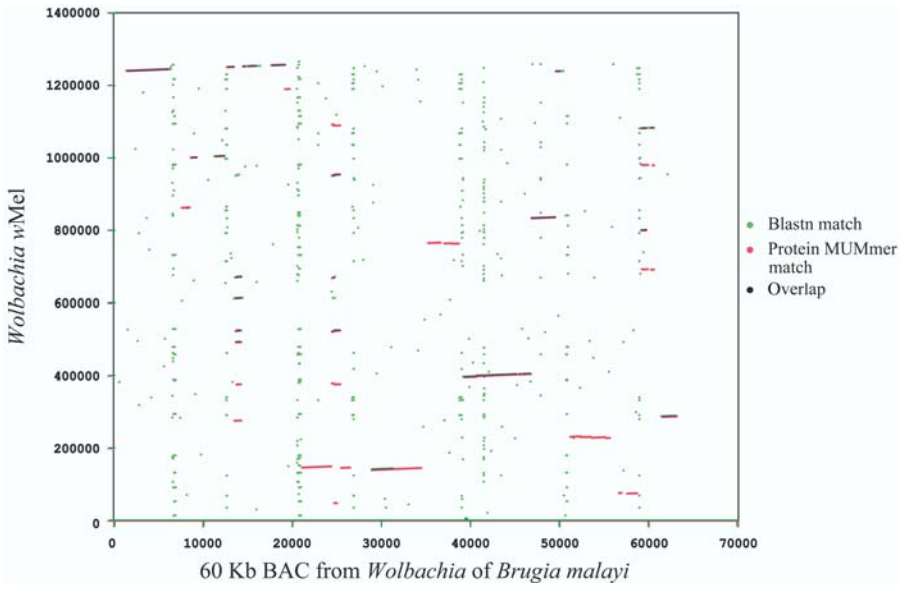

Figure 3. Alignment of *w*Mel with a 60 kbp Region of the *Wolbachia* from *B. malayi*

The figure shows BLASTN matches (green) and whole-proteome alignments (red) that were generated using the "promer" option of the MUMmer software (Delcher et al. 1999). The *B. malayi* region is from a BAC clone (Ware et al. 2002). Note the regions of alignment broken up by many rearrangements and the presence of repetitive sequences at the regions of the breaks.

DOI: 10.1371/journal.pbio.0020069.g003

apparently lack genes for LPS production (Lin and Rikihisha 2003), it is likely that the common ancestor of *Wolbachia*, *Ehrlichia*, and *Anaplasma* was unable to synthesize LPS. Thus, the reports that *Wolbachia*-derived LPS-like compounds is involved in the immunopathology of filarial nematode disease in mammals (Taylor 2002) either indicate that these *Wolbachia* have acquired genes for LPS synthesis or that the reported LPS-like compounds are not homologous to LPS.

Despite evident genome reduction in *w*Mel and in contrast to most small-genomed intracellular species, gene duplication appears to have continued, as over 50 gene families have apparently expanded in the *w*Mel lineage relative to that of all other species (Table S4). Many of the pairs of duplicated genes are encoded next to each other in the genome, suggesting that they arose by tandem duplication events and may simply reflect transient duplications in evolution (deletion is common when there are tandem arrays of genes). Many others are components of mobile genetic elements, indicating that these elements have expanded significantly after entering the *Wolbachia* evolutionary lineage. Other duplications that could contribute to the unique biological properties of *w*Mel include that of the mismatch repair gene *mutL* (see below) and that of many hypothetical and conserved hypothetical proteins.

One duplication of particular interest is that of *wsp*, which is a standard gene for strain identification and phylogenetic reconstruction in *Wolbachia* (Zhou et al. 1998). In addition to the previously described *wsp* (WD0159), *w*Mel encodes two *wsp* paralogs (WD0009 and WD0489), which we designate as *wspB* and *wspC*, respectively. While these paralogs are highly divergent from *wsp* (protein identities of 19.7% and 23.5%, respectively) and do not amplify using the standard *wsp* PCR primers (Braig et al. 1998; Zhou et al. 1998), their presence could lead to some confusion in classification and identification of *Wolbachia* strains. This has apparently occurred in one study of *Wolbachia* strain *w*KueYO, for which the reported *wsp* gene (gbAB045235) is actually an ortholog of *wspB* (99.8% sequence identity and located at the end of the *virB* operon [Masui et al. 2000]) and not an ortholog of the *wsp* gene. Considering that the *wsp* gene has been extremely informative for discriminating between strains of *Wolbachia*, we designed PCR primers to the *w*Mel *wspB* gene to amplify and then sequence the orthologs from the related *w*Ri and *w*AlbB *Wolbachia* strains from *Drosophila simulans* and *Aedes albopictus*, respectively, as well as the *Wolbachia* strain that infects the filarial nematode *Dirofilaria immitis* to determine the potential utility of this locus for strain discrimination. A comparison of genetic distances between the *wsp* and *wspB* genes for these different taxa indicates that overall the *wspB* gene appears to be evolving at a faster rate than *wsp* and, as such, may be a useful additional marker for discriminating between closely related *Wolbachia* strains (Table S5).

Inefficiency of Selection in *w*Mel

The fraction of the genome that is repetitive DNA and the fraction that corresponds to mobile genetic elements are among the highest for any prokaryotic genome. This is particularly striking compared to the genomes of other obligate intracellular species such as *Buchnera*, *Rickettsia*, *Chlamydia*, and *Wigglesworthia*, that all have very low levels of repetitive DNA and mobile elements. The recently sequenced genomes of the intracellular pathogen *Coxiella burnetti* (Seshadri et al. 2003) has both a streamlined genome and moderate amounts of repetitive DNA, although much less than *w*Mel. The paucity of repetitive DNA in these and other intracellular species is thought to be due to a combination of lack of exposure to other species, thereby limiting introduction of mobile elements, and genome streamlining (Mira et al. 2001; Moran and Mira 2001; Frank et al. 2002). We examined the *w*Mel genome to try to understand the origin of the repetitive and mobile DNA and to explain why such repetitive/mobile DNA is present in *w*Mel, but not other streamlined intracellular species.

We propose that the mobile DNA in *w*Mel was acquired some time after the separation of the *Wolbachia* and *Rickettsia* lineages but before the radiation of the *Wolbachia* group. The acquisition of these elements after the separation of the *Wolbachia* and *Rickettsia* lineages is suggested by the fact that most do not have any obvious homologous sequences in the genomes of other α-Proteobacteria, including the closely related *Rickettsia* spp. Additional evidence for some acquisition of foreign DNA after the *Wolbachia–Rickettsia* split comes from phylogenetic analysis of those genes present in *w*Mel, but not in the two sequenced rickettsial genomes (see Table S3; unpublished data). The acquisition prior to the radiation of *Wolbachia* is suggested by two lines of evidence. First, many of the elements are found in the genome of the distantly related *Wolbachia* of the nematode *B. malayi* (see Figure 3; unpublished data). In addition, genome analysis reveals that these elements do not have significantly anomalous nucleotide composition or codon usage compared to the rest of the genome. In fact, there are only four regions of the genome with significantly anomalous composition, comprising in total only approximately 17 kbp of DNA (Table 3). The lack of anomalous composition suggests either that any foreign DNA in *w*Mel was acquired long enough ago to allow it to "ameliorate" and become compositionally similar to endogenous *Wolbachia* DNA (Lawrence and Ochman 1997, 1998) or that any foreign DNA that is present was acquired from organisms with similar composition to endogenous *w*Mel genes. Owing to their potential effects on genome evolution (insertional mutagenesis, catalyzing genome rearrangements), we propose that the acquisition and maintenance of these repetitive and mobile elements by *w*Mel have played a key role in shaping the evolution of *Wolbachia*.

It is likely that much of the mobile/repetitive DNA was introduced via phage, given that three prophage elements are present; experimental studies have shown active phage in some *Wolbachia* (Masui et al. 2001) and *Wolbachia* superinfections occur in many hosts (e.g., Jamnongluk et al. 2002), which would allow phage to move between strains. Whatever the mechanism of introduction, the persistence of the repetitive elements in *w*Mel in the face of apparently strong pressures for streamlining is intriguing. One expla-nation is that *w*Mel may be getting a steady infusion of mobile elements from other *Wolbachia* strains to counteract the elimination of elements by selection for genome streamlining. This would explain the absence of anomalous nucleotide composition of the elements. However, we believe that a major contributing factor to the presence of all the repetitive/mobile DNA in *w*Mel is that *w*Mel and possibly *Wolbachia* in general have general inefficiency of natural selection relative to other species. This inefficiency would limit the ability to eliminate repetitive DNA. A general inefficiency of natural selection

Table 3. Regions of Anomalous Nucleotide Composition in the *wMel* Genome

Region	Chromosome Location	GC (%)	CDS Number	Putative Function
1	1079500–1084000	41.55	WD1129	NADH-ubiquinone oxidoreductase, putative
			WD1130	Hypothetical protein
			WD1131	Conserved hypothetical protein, degenerate
			WD1132	Conserved hypothetical protein
			WD1133	DNA topoisomerase I
2	182500–186000	46	WD0199	Hypothetical protein
			WD0200	Hypothetical protein
3	690500–695000	42.55	WD0717	Conserved hypothetical protein
			WD0718	Conserved hypothetical protein, authentic point mutation
			WD0719	Penicillin-binding protein
4	87500–92000	40.8	WD0095	D-alanine-D-alanine ligase
			WD0096	Cell division protein FtsQ, putative
			WD0097	Hypothetical protein
			WD0098	D-alanyl-D-alanine carboxy peptidase
			WD0099	Multidrug resistance protein

DOI: 10.1371/journal.pbio.0020069.t003

(especially purifying selection) has been suggested previously for intracellular bacteria, based in part on observations that these bacteria have higher evolutionary rates than free-living bacteria (e.g., Moran 1996). We also find a higher evolutionary rate for *w*Mel than that of the closely related intracellular *Rickettsia*, which themselves have higher rates than free-living α-Proteobacteria (Figure 4). Additionally, codon bias in *w*Mel appears to be driven more by mutation or drift than selection (Figure S2), as has been reported for *Buchnera* species and was suggested to be due to inefficient purifying selection (Wernegreen and Moran 1999). Such inefficiencies of natural selection are generally due to an increase in the relative contribution of genetic drift and mutation as compared to natural selection (Eiglmeier et al. 2001; Lawrence 2001; Parkhill et al. 2001). Below we discuss different possible explanations for the inefficiency of selection in *w*Mel, especially in comparison to other intracellular bacteria.

Low rates of recombination, such as occur in centromeres and the human Y chromosome, can lead to inefficient selection because of the linkage among genes. This has been suggested to be occurring in *Buchnera* species because these species do not encode homologs of RecA, which is the key protein in homologous recombination in most species (Shigenobu et al. 2000). The absence of recombination in *Buchnera* is supported by the lack of genome rearrangements in their recent evolution (Tamas et al. 2002). Additionally, there is apparently little or no gene flow into *Buchnera* strains. In contrast, *w*Mel encodes the necessary machinery for recombination, including RecA (Table S6), and has experienced both extensive intragenomic homologous recombination and introduction of foreign DNA. Therefore, the

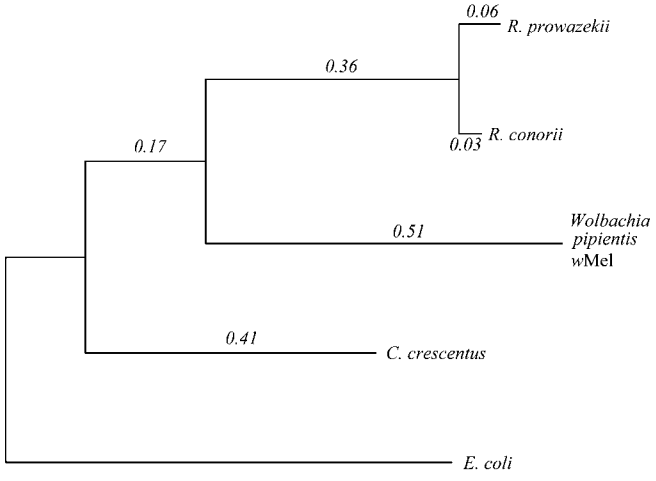

Figure 4. Long Evolutionary Branches in *w*Mel

Maximum-likelihood phylogenetic tree constructed on concatenated protein sequences of 285 orthologs shared among *w*Mel, *R. prowazekii*, *R. conorii*, *C. crescentus*, and *E. coli*. The location of the most recent common ancestor of the α-Proteobacteria (*Caulobacter*, *Rickettsia*, *Wolbachia*) is defined by the outgroup *E. coli*. The unit of branch length is the number of changes per amino acid. Overall, the amino acid substitution rate in the *w*Mel lineage is about 63% higher than that of *C. crescentus*, a free-living α-Proteobacteria. *w*Mel has evolved at a slightly higher rate than the *Rickettssia* spp., close relatives that are also obligate intracellular bacteria that have undergone accelerated evolution themselves. This higher rate is likely in part to be due to an increase in the rate of slightly deleterious mutations, although we have not ruled out the possibility of G+C content effects on the branch lengths.

DOI: 10.1371/journal.pbio.0020069.g004

unusual genome features of *w*Mel are unlikely to be due to low levels of recombination.

Another possible explanation for inefficient selection is high mutation rates. It has been suggested that the higher evolutionary rates in intracellular bacteria are the result of high mutation rates that are in turn due to the loss of genes for DNA repair processes (e.g., Itoh et al. 2002). This is likely not the case in *w*Mel since its genome encodes proteins corresponding to a broad suite of DNA repair pathways including mismatch repair, nucleotide excision repair, base excision repair, and homologous recombination (Table S6). The only noteworthy DNA repair gene absent from *w*Mel and present in the more slowly evolving *Rickettsia* is *mfd,* which is involved in targeting DNA repair to the transcribed strand of actively transcribing genes in other species (Selby et al. 1991). However, this absence is unlikely to contribute significantly to the increased evolutionary rate in *w*Mel, since defects in *mfd* do not lead to large increases in mutation rates in other species (Witkin 1994). The presence of mismatch repair genes (homologs of *mutS* and *mutL*) in *w*Mel is particularly relevant since this pathway is one of the key steps in regulating mutation rates in other species. In fact, *w*Mel is the first bacterial species to be found with two *mutL* homologs. Overall, examination of the predicted DNA repair capabilities of bacteria (Eisen and Hanawalt 1999) suggests that the connection between evolutionary rates in intracellular species and the loss of DNA repair processes is spurious. While many intracellular species have lost DNA repair genes in their recent evolution, different species have lost different genes and some, such as *w*Mel and *Buchnera* spp., have kept the genes that likely regulate mutation rates. In addition, some free-living species without high evolutionary rates have lost some of the same pathways lost in intracellular species, while many free-living species have lost key pathways resulting in high mutation rates (e.g., *Helicobacter pylori* has apparently lost mismatch repair [Eisen1997, Eisen 1998b; Bjorkholm et al. 2001]). Given that intracellular species tend to have small genomes and have lost genes from every type of biological process, it is not surprising that many of them have lost DNA repair genes as well.

We believe that the most likely explanations for the inefficiency of selection in *w*Mel involve population-size related factors, such as genetic drift and the occurrence of population bottlenecks. Such factors have also been shown to likely explain the high evolutionary rates in other intracellular species (Moran 1996; Moran and Mira 2001; van Ham et al. 2003). *Wolbachia* likely experience frequent population bottlenecks both during transovarial transmission (Boyle et al. 1993) and during cytoplasmic incompatibility mediated sweeps through host populations. The extent of these bottlenecks may be greater than in other intracellular bacteria, which would explain why *w*Mel has both more repetitive and mobile DNA than other such species and a higher evolutionary rate than even the related *Rickettsia spp.* Additional genome sequences from other *Wolbachia* will reveal whether this is a feature of all *Wolbachia* or only certain strains.

Mitochondrial Evolution

There is a general consensus in the evolutionary biology literature that the mitochondria evolved from bacteria in the α-subgroup of the Proteobacteria phyla (e.g., Lang et al. 1999). Analysis of complete mitochondrial and bacterial genomes has very strongly supported this hypothesis (Andersson et al. 1998, 2003; Muller and Martin 1999; Ogata et al. 2001). However, the exact position of the mitochondria within the α-Proteobacteria is still debated. Many studies have placed them in or near the Rickettsiales order (Viale and Arakaki 1994; Gupta 1995; Sicheritz-Ponten et al. 1998; Lang et al. 1999; Bazinet and Rollins 2003). Some studies have further suggested that mitochondria are a sister taxa to the *Rickettsia* genus within the Rickettsiaceae family and thus more closely related to *Rickettsia* spp. than to species in the Anaplasmataceae family such as *Wolbachia* (Karlin and Brocchieri 2000; Emelyanov 2001a, 2001b, 2003a, 2003b).

In our analysis of complete genomes, including that of *w*Mel, the first non-*Rickettsia* member of the Rickettsiales order to have its genome completed, we find support for a grouping of *Wolbachia* and *Rickettsia* to the exclusion of the mitochondria, but not for placing the mitochondria within the Rickettsiales order (Figure 5A and 5B; Table S7; Table S8). Specifically, phylogenetic trees of a concatenated alignment of 32 proteins show strong support with all methods (see Table S7) for common branching of: (i) mitochondria, (ii) *Rickettsia* with *Wolbachia*, (iii) the free-living α-Proteobacteria, and (iv) mitochondria within α-Proteobacteria. Since amino acid content bias was very severe in these datasets, protein LogDet analyses, which can correct for the bias, were also performed. In LogDet analyses of the concatenated protein alignment, both including and excluding highly biased positions, mitochondria usually branched basal to the *Wolbachia–Rickettsia* clade, but never specifically with *Rickettsia* (see Table S7). In addition, in phylogenetic studies of individual genes, there was no consistent phylogenetic position of mitochondrial proteins with any particular species or group within the α-Proteobacteria (see Table S8), although support for a specific branch uniting the two *Rickettsia* species with *Wolbachia* was quite strong. Eight of the proteins from mitochondrial genomes (YejW, SecY, Rps8, Rps2, Rps10, RpoA, Rpl15, Rpl32) do not even branch within the α-Proteobacteria, although these genes almost certainly were encoded in the ancestral mitochondrial genome (Lang et al. 1997).

This analysis of mitochondrial and α-Proteobacterial genes reinforces the view that ancient protein phylogenies are inherently prone to error, most likely because current models of phylogenetic inference do not accurately reflect the true evolutionary processes underlying the differences observed in contemporary amino acid sequences (Penny et al. 2001). These conflicting results regarding the precise position of mitochondria within the α-Proteobacteria can be seen in the high amount of networking in the Neighbor-Net graph of the analyses of the concatenated alignment shown in Figure 5. An important complication in studies of mitochondrial evolution lies in identifying "α-Proteobacterial" genes for comparison (Martin 1999). For example, in our analyses, proteins from *Magnetococcus* branched with other α-Proteobacterial homologs in only 17 of the 49 proteins studied, and in five cases they assumed a position basal to α-, β-, and γ-Proteobacterial homologs.

Host–Symbiont Gene Transfers

Many genes that were once encoded in mitochondrial genomes have been transferred into the host nuclear

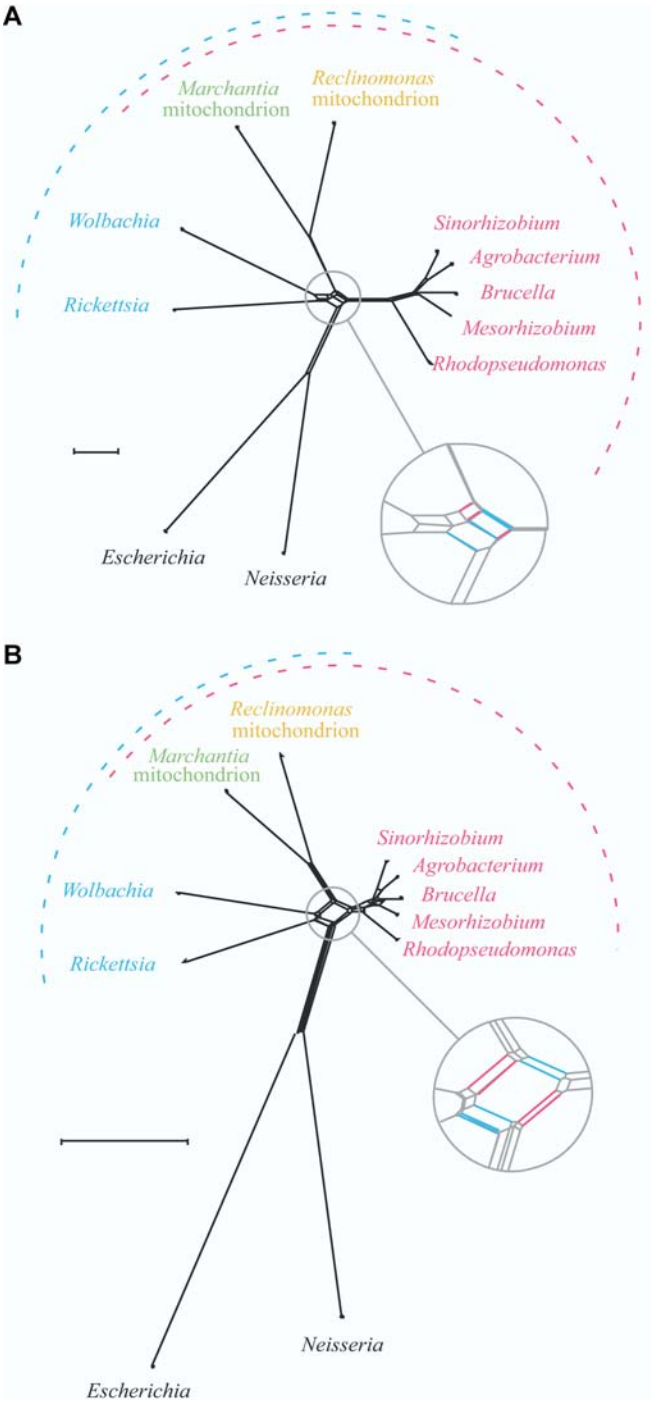

Figure 5. Mitochondrial Evolution Using Concatenated Alignments

Networks of protein LogDet distances for an alignment of 32 proteins constructed with Neighbor-Net (Bryant and Moulton 2003). The scale bar indicates 0.1 substitutions per site. Enlargements at lower right show the component of shared similarity between mitochondrial-encoded proteins and (i) their homologs from intracellular endosymbionts (red) as well as (ii) their homologs from free-living α-Proteobacteria (blue). (A) Result using 6,776 gap-free sites per genome (heavily biased in amino acid composition). (B) Result using 3,100 sites after exclusion of highly variable positions (data not biased in amino acid composition at $p = 0.95$). All data and alignments are available upon request. Results of phylogenetic analyses are summarized in Table S7. Since amino acid content bias was very severe in these datasets, protein LogDet analyses were also preformed. In neighbor-joining, parsimony, and maximum-likelihood trees generated from alignments both including and excluding highly biased positions (6,776 and 3,100 gap-free amino acid sites per genome, respectively), mitochondria usually branched basal to the *Wolbachia*–*Rickettsia* clade, but never specifically with *Rickettsia* (Table S7).

DOI: 10.1371/journal.pbio.0020069.g005

genomes. Searching for such genes has been complicated by the fact that many of the transfer events happened early in eukaryotic evolution and that there are frequently extreme amino acid and nucleotide composition biases in mitochondrial genomes (see above). We used the *w*Mel genome to search for additional possible mitochondrial-derived genes in eukaryotic nuclear genomes. Specifically, we constructed phylogenetic trees for *w*Mel genes that are not in either *Rickettsia* genomes. Five new eukaryotic genes of possible mitochondrial origin were identified: three genes involved in de novo nucleotide biosynthesis (*purD*, *purM*, *pyrD*) and two conserved hypothetical proteins (WD1005, WD0724). The α-Proteobacterial origin of these genes suggests that at least some of the genes of the de novo nucleotide synthesis pathway in eukaryotes might have been laterally acquired from bacteria via the mitochondria. The presence of such genes in other Proteobacteria suggests that their absence from *Rickettsia* is due to gene loss (Gray et al. 2001). This finding supports the need for additional α-Proteobacterial genomes to identify mitochondrion-derived genes in eukaryotes.

While organelle to nuclear gene transfers are generally accepted, there is a great deal of controversy over whether other gene transfers have occurred from bacteria into animals. In particular, claims of transfer from bacteria into the human genome (Lander et al. 2001) were later shown to be false (Roelofs and Van Haastert 2001; Salzberg et al. 2001; Stanhope et al. 2001). *Wolbachia* are excellent candidates for such transfer events since they live inside the germ cells, which would allow lateral transfers to the host to be transmitted to subsequent host generations. Consistent with this, a recent study has shown some evidence for the presence of *Wolbachia*-like genes in a beetle genome (Kondo et al. 2002). The symbiosis between *w*Mel and *D. melanogaster* provides an ideal case to search for such transfers since we have the complete genomes of both the host and symbiont. Using BLASTN searches and MUMmer alignments, we did not find any examples of highly similar stretches of DNA shared between the two species. In addition, protein-level searches and phylogenetic trees did not identify any specific relationships between *w*Mel and *D. melanogaster* for any genes. Thus, at least for this host–symbiont association, we do not find any likely cases of recent gene exchange, with genes being maintained in both host and symbiont. In addition, in our phylogenetic analyses, we did not find any examples of *w*Mel proteins branching specifically with proteins from any invertebrate to the exclusion of other eukaryotes. Therefore, at least for the genes in *w*Mel, we do not find evidence for transfer of *Wolbachia* genes into any invertebrate genome.

Metabolism and Transport

*w*Mel is predicted to have very limited capabilities for membrane transport, for substrate utilization, and for the biosynthesis of metabolic intermediates (Figure S3), similar to what has been seen in other intracellular symbionts and pathogens (Paulsen et al. 2000). Almost all of the identifiable uptake systems for organic nutrients in *w*Mel are for amino

acids, including predicted transporters for proline, asparate/glutamate, and alanine. This pattern of transporters, coupled with the presence of pathways for the metabolism of the amino acids cysteine, glutamate, glutamine, proline, serine, and threonine, suggests that *w*Mel may obtain much of its energy from amino acids. These amino acids could also serve as material for the production of other amino acids. In contrast, carbohydrate metabolism in *w*Mel appears to be limited. The only pathways that appear to be complete are the tricarboxylic acid cycle, the nonoxidative pentose phosphate pathway, and glycolysis, starting with fructose-1,6-biphosphate. The limited carbohydrate metabolism is consistent with the presence of only one sugar phosphate transporter. *w*Mel can also apparently transport a range of inorganic ions, although two of these systems, for potassium uptake and sodium ion/proton exchange, are frameshifted. In the latter case, two other sodium ion/proton exchangers may be able to compensate for this defect.

Many of the predicted metabolic properties of *w*Mel, such as the focus on amino acid transport and the presence of limited carbohydrate metabolism, are similar to those found in *Rickettsia*. A major difference with the *Rickettsia* spp. is the absence of the ADP–ATP exchanger protein in *w*Mel. In *Rickettsia* this protein is used to import ATP from the host, thus allowing these species to be direct energy scavengers (Andersson et al. 1998). This likely explains the presence of glycolysis in *w*Mel but not *Rickettsia*. An inability to obtain ATP from its host also helps explain the presence of pathways for the synthesis of the purines AMP, IMP, XMP, and GMP in *w*Mel but not *Rickettsia*. Other pathways present in *w*Mel but not *Rickettsia* include threonine degradation (described above), riboflavin biosynthesis, pyrimidine metabolism (i.e., from PRPP to UMP), and chelated iron uptake (using a single ABC transporter). The two *Rickettsia* species have a relatively large complement of predicted transporters for osmoprotectants, such as proline and glycine betaine, whereas *w*Mel possesses only two of these systems.

Regulatory Responses

The *w*Mel genome is predicted to encode few proteins for regulatory responses. Three genes encoding two-component system subunits are present: two sensor histidine kinases (WD1216 and WD1284) and one response regulator (WD0221). Only six strong candidates for transcription regulators were identified: a homolog of arginine repressors (WD0453), two members of the TenA family of transcription activator proteins (WD0139 and WD0140), a homolog of *ctrA*, a transcription regulator for two component systems in other α-Proteobacteria (WD0732), and two σ factors (RpoH/WD1064 and RpoD/WD1298). There are also seven members of one paralogous family of proteins that are distantly related to phage repressors (see above), although if they have any role in transcription, it is likely only for phage genes. Such a limited repertoire of regulatory systems has also been reported in other endosymbionts and has been explained by the apparent highly predictable and stable environment in which these species live (Andersson et al. 1998; Read et al. 2000; Shigenobu et al. 2000; Moran and Mira 2001; Akman et al. 2002; Seshadri et al. 2003).

Host–Symbiont Interactions

The mechanisms by which *Wolbachia* infect host cells and by which they cause the diverse phenotypic effects on host reproduction and fitness are poorly understood, and the *w*Mel genome helps identify potential contributing factors. A complete Type IV secretion system, portions of which have been reported in earlier studies, is present. The complete genome sequence shows that in addition to the five *vir* genes previously described from *Wolbachia w*KueYO (Masui et al. 2001), an additional four are present in *w*Mel. Of the nine *w*Mel *vir* ORFs, eight are arranged into two separate operons. Similar to the single operon identified in *w*Tai and *w*KueYO, the *w*Mel *virB8*, *virB9*, *virB10*, *virB11*, and *virD4* CDSs are adjacent to *wspB*, forming a 7 kb operon (WD0004–WD0009). The second operon contains *virB3*, *virB4*, and *virB6* as well as four additional non-*vir* CDSs, including three putative membrane-spanning proteins, that form part of a 15.7 kb operon (WD0859–WD0853). Examination of the *Rickettsia conorii* genome shows a similar orga-nization (Figure 6A). The observed conserved gene order for these genes between these two genomes suggests that the putative membrane-spanning proteins could form a novel and, possibly, integral part of a functioning Type IV secretion system within these bacteria. Moreover, reverse transcription (RT)-PCRs have confirmed that *wspB* and WD0853–WD0856 are each expressed as part of the two *vir* operons and further indicate that these additional encoded proteins are novel components of the *Wolbachia* Type IV secretion system (Figure 6B).

In addition to the two major *vir* clusters, a paralog of *virB8* (WD0817) is also present in the *w*Mel genome. WD0818 is quite divergent from *virB8* and, as such, does not appear to have resulted from a recent gene duplication event. RT-PCR experiments have failed to show expression of this CDS in *w*Mel-infected *Drosophila* (data not shown). PCR primers were designed to all CDSs of the *w*Mel Type IV secretion system and used to successfully amplify orthologs from the divergent *Wolbachia* strains *w*Ri and *w*AlbB (data not shown). We were able to detect orthologs to all of the *w*Mel Type IV secretion system components as well as most of the adjacent non-*vir* CDSs, suggesting that this system is conserved across a range of A- and B-group *Wolbachia*. An increasing body of evidence has highlighted the importance of Type IV secretion systems for the successful infection, invasion, and persistence of intracellular bacteria within their hosts (Christie 2001; Sexton and Vogel 2002). It is likely that the Type IV system in *Wolbachia* plays a role in the establishment and maintenance of infection and possibly in the generation of reproductive phenotypes.

Genes involved in pathogenicity in bacteria have been found to be frequently associated with regions of anomalous nucleotide composition, possibly owing to transfer from other species or insertion into the genome from plasmids or phage. In the four such regions in *w*Mel (see above; see Table 3), some additional candidates for pathogenicity-related activities are present including a putative penicillin-binding protein (WD0719), genes predicted to be involved in cell wall synthesis (WD0095–WD0098, including D-alanine-D-alanine ligase, a putative FtsQ, and D-alanyl-D-alanine carboxy peptidase) and a multidrug resistance protein (WD0099). In addition, we have identified a cluster of genes in one of the phage regions that may also have some role in host–symbiont interactions. This cluster (WD0611–WD0621) is embedded within the WO-B phage region of the genome (see Figure 2) and contains many genes that encode proteins with putative roles in the synthesis and degradation of surface polysac-

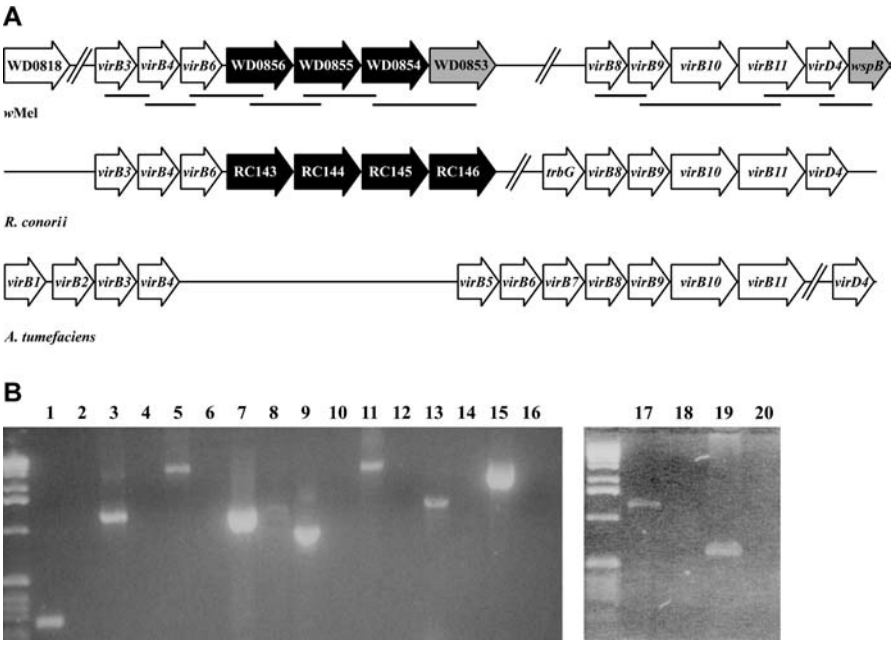

Figure 6. Genomic Organization and expression of Type IV Secretion Operons in *w*Mel

(A) Organization of the nine *vir*-like CDSs (white arrows) and five adjacent CDSs that encode for either putative membrane-spanning proteins (black arrows) or non-*vir* CDSs (gray arrows) of *w*Mel, *R. conorii*, and *A. tumefaciens*. Solid horizontal lines denote RT experiments that have confirmed that adjacent CDSs are expressed as part of a polycistronic transcript. Results of these RT-PCR experiments are presented in (B). Lane 1, *virB3-virB4*; lane 2, RT control; lane 3, *virB6*-WD0856; lane 4, RT control; lane 5, WD0856-WD0855; lane 6, RT control; lane 7, WD0854-WD0853; lane 8, RT control; lane 9, *virB8-virB9*; lane 10, RT control; lane 11, *virB9-virB11*; lane 12, RT control; lane 13, *virB11-virD4*; lane 14, RT control; lane 15, *virD4-wspB*; lane 16, RT control; lane 17, *virB4-virB6*; lane 18, RT control; lane 19, WD0855-WD0854; lane 20, RT control. Only PCRs that contain reverse transcriptase amplified the desired products. PCR primer sequences are listed in Table S9.

DOI: 10.1371/journal.pbio.0020069.g006

charides, including a UDP-glucose 6-dehydrogenase (WD0620). Since this cluster appears to be normal in terms of phylogeny relative to other genes in the genome (i.e., the genes in this region have normal *w*Mel nucleotide composition and branch in phylogenetic trees with genes from other α-Proteobacteria), it is not likely to have been acquired from other species. However, it is possible that these genes can be transferred among *Wolbachia* strains via the phage, which in turn could lead to some variation in host–symbiont interactions between *Wolbachia* strains.

Of particular interest for host-interaction functions are the large number of genes that encode proteins that contain ankyrin repeats (Table 4). Ankyrin repeats, a tandem motif of around 33 amino acids, are found mainly in eukaryotic proteins, where they are known to mediate protein–protein interactions (Caturegli et al. 2000). While they have been found in bacteria before, they are usually present in only a few copies per species. *w*Mel has 23 ankyrin repeat-containing genes, the most currently described for a prokaryote, with *C. burnetti* being next with 13. This is particularly striking given *w*Mel's relatively small genome size. The functions of the ankyrin repeat-containing proteins in *w*Mel are difficult to predict since most have no sequence similarity outside the ankyrin domains to any proteins of known function. Many lines of evidence suggest that the *w*Mel ankyrin domain proteins are involved in regulating host cell-cycle or cell division or interacting with the host cytoskeleton: (i) many ankyrin-containing proteins in eukaryotes are thought to be involved in linking membrane proteins to the cytoskeleton (Hryniewicz-Jankowska et al. 2002); (ii) an ankyrin-repeat protein of *Ehrlichia phagocytophila* binds condensed chromatin of host cells and may be involved in host cell-cycle regulation (Caturegli et al. 2000); (iii) some of the proteins that modify the activity of cell-cycle-regulating proteins in *D. melanogaster* contain ankyrin repeats (Elfring et al. 1997); and (iv) the *Wolbachia* strain that infects the wasp *Nasonia vitripennis* induces cytoplasmic incompatibility, likely by interacting with these same cell-cycle proteins (Tram and Sullivan 2002). Of the ankyrin-containing proteins in *w*Mel, those worth exploring in more detail include the several that are predicted to be surface targeted or secreted (Table 4) and thus could be targeted to the host nucleus. It is also possible that some of the other ankyrin-containing proteins are secreted via the Type IV secretion system in a targeting signal independent pathway. We call particular attention to three of the ankyrin-containing proteins (WD0285, WD0636, and WD0637), which are among the very few genes, other than those encoding components of the translation apparatus, that have significantly biased codon usage relative to what is expected based on GC content, suggesting they may be highly expressed.

Conclusions

Analysis of the *w*Mel genome reveals that it is unique among sequenced genomes of intracellular organisms in that it is both streamlined and massively infected with mobile genetic elements. The persistence of these elements in the genome for apparently long periods of time suggests that *w*Mel is inefficient at getting rid of them, likely a result of experiencing severe population bottlenecks during every cycle of transovarial transmission as well as during sweeps through host populations. Integration of evolutionary reconstructions and genome analysis (phylogenomics) has provided insights into the biology of *Wolbachia*, helped identify genes that likely play roles in the unusual effects *Wolbachia* have on their host, and revealed many new details about the evolution of *Wolbachia* and mitochondria. Perhaps most importantly, future studies of *Wolbachia* will benefit both from this genome sequence and from the ability to study host–symbiont interactions in a host (*D. melanogaster*) well-suited for experimental studies.

Materials and Methods

Purification/source of DNA. *w*Mel DNA was obtained from *D. melanogaster* yw^{67c23} flies that naturally carry the *w*Mel infection. *w*Mel

Table 4. Ankyrin-Domain Containing Proteins Encoded by the *w*Mel Genome

Locus	Annotation	Number of Ankyrin Repeats	Signal Peptide	Predicted to Be Highly Expressed
WD0035	Ankyrin repeat domain protein	6		
WD0073	Ankyrin repeat domain protein	5		
WD0147	Ankyrin repeat domain protein	11		
WD0191	Ankyrin repeat domain protein	1		
WD0285	Prophage λW1, ankyrin repeat domain protein	3		Y
WD0286	Prophage λW1, ankyrin repeat domain protein	3		
WD0291	Prophage λW1, ankyrin repeat domain protein	5		
WD0292	Prophage λW1, ankyrin repeat domain protein	2		
WD0294	Ankyrin repeat domain protein	9		
WD0385	Ankyrin repeat domain protein	11		
WD0438	Ankyrin repeat domain protein	2		
WD0441	Ankyrin repeat domain protein	1	Y	
WD0498	Ankyrin repeat domain protein	9		
WD0514	Ankyrin repeat domain protein	6		
WD0550	Ankyrin repeat domain protein	6		
WD0566	Ankyrin repeat domain protein	2		
WD0596	Prophage λW4, ankyrin repeat domain protein	9		
WD0633	Prophage λW5, ankyrin repeat domain protein	4		
WD0636	Prophage λW5, ankyrin repeat domain protein	2		Y
WD0637	Prophage λW5, ankyrin repeat domain protein	3		Y
WD0754	Ankyrin repeat domain protein	2		
WD0766	Ankyrin repeat domain protein	8		
WD1213	Ankyrin repeat domain protein, putative	1	Y	

DOI: 10.1371/journal.pbio.0020069.t004

was purified from young adult flies on pulsed-field gels as described previously (Sun et al. 2001). Plugs were digested with the restriction enzyme AscI (GG^CGCGCC), which cuts the bacterial chromosome twice (Sun et al. 2001), aiding in the entry of the DNA into agarose gels. After electrophoresis, the resulting two bands were recovered from the gel and stored in 0.5 M EDTA (pH 8.0). DNA was extracted from the gel slices by first washing in TE (Tris–HCl and EDTA) buffer six times for 30 min each to dilute EDTA followed by two 1-h washes in β-agarase buffer (New England Biolabs, Beverly, Massachusetts, United States). Buffer was then removed and the blocks melted at 70°C for 7 min. The molten agarose was cooled to 40°C and then incubated in β-agarase (1 U/100 μl of molten agarose) for 1 h. The digest was cooled to 4°C for 1 h and then centrifuged at $4{,}100 \times g_{\max}$ for 30 min at 4°C to remove undigested agarose. The supernatant was concentrated on a Centricon YM-100 microconcentrator (Millipore, Bedford, Massachusetts, United States) after prerinsing with 70% ethanol followed by TE buffer and, after concentration, rinsed with TE. The retentate was incubated with proteinase K at 56°C for 2 h and then stored at 4°C. *w*Mel DNA for gap closure was prepared from approximately 1,000 *Drosophila* adults using the Holmes–Bonner urea/phenol:chloroform protocol (Holmes and Bonner 1973) to prepare total fly DNA.

Library construction/sequencing/closure. The complete genome sequence was determined using the whole-genome shotgun method (Venter et al. 1996). For the random shotgun-sequencing phase, libraries of average size 1.5–2.0 kb and 4.0–8.0 kb were used. After assembly using the TIGR Assembler (Sutton et al. 1995), there were 78 contigs greater than 5000 bp, 186 contigs greater than 3000 bp, and 373 contigs greater than 1500 bp. This number of contigs was unusually high for a 1.27 Mb genome. An initial screen using BLASTN searches against the nonredundant database in GenBank and the Berkeley *Drosophila* Genome Project site (http://www.fruitfly.org/blast/) showed that 3,912 of the 10,642 contigs were likely contaminants from the *Drosophila* genome. To aid in closure, the assemblies were rerun with all sequences of likely host origin excluded. Closure, which was made very difficult by the presence of a large amount of repetitive DNA (see below), was done using a mix of primer walking, generation, and sequencing of transposon-tagged libraries of large insert clones and multiplex PCR (Tettelin et al. 1999). The final sequence showed little evidence for polymorphism within the population of *Wolbachia* DNA. In addition, to obtain sequence across the AscI-cut sites, PCR was performed on undigested DNA. It is important to point out that the reason significant host contamination does not significantly affect symbiont genome assembly is that most of the *Drosophila* contigs were small due to the approximately 100-fold difference in genome sizes between host (approximately 180 Mb) and *w*Mel (1.2 Mb).

Since it has been suggested that *Wolbachia* and their hosts may undergo lateral gene transfer events (Kondo et al. 2002), genome assemblies were rerun using all of the shotgun and closure reads without excluding any sequences that appeared to be of host origin. Only five assemblies were found to match both the *D. melanogaster* genome and the *w*Mel assembly. Primers were designed to match these assemblies and PCR attempted from total DNA of *w*Mel infected *D. melanogaster*. In each case, PCR was unsuccessful, and we therefore presume that these assemblies are the result of chimeric cloning artifacts. The complete sequence has been given GenBank accession ID AE017196 and is available at http://www.tigr.org/tdb.

Repeats. Repeats were identified using RepeatFinder (Volfovsky et al. 2001), which makes use of the REPuter algorithm (Kurtz and Schleiermacher 1999) to find maximal-length repeats. Some manual curation and BLASTN and BLASTX searches were used to divide repeat families into different classes.

Annotation. Identification of putative protein-encoding genes and annotation of the genome was done as described previously (Eisen et al. 2002). An initial set of ORFs likely to encode proteins (CDS) was identified with GLIMMER (Salzberg et al. 1998). Putative proteins encoded by the CDS were examined to identify frameshifts or premature stop codons compared to other species. The sequence traces for each were reexamined and, for some, new sequences were generated. Those for which the frameshift or premature stops were of high quality were annotated as "authentic" mutations. Functional assignment, identification of membrane-spanning domains, determination of paralogous gene families, and identification of regions of

unusual nucleotide composition were performed as described previously (Tettelin et al. 2001). Phylogenomic analysis (Eisen 1998a; Eisen and Fraser 2003) was used to aid in functional predictions. Alignments and phylogenetic trees were generated as described (Salzberg et al. 2001).

Comparative genomics. All putative *w*Mel proteins were searched using BLASTP against the predicted proteomes of published complete organismal genomes and a set of complete plastid, mitochondrial, plasmid, and viral genomes. The results of these searches were used (i) to analyze the phylogenetic profile (Pellegrini et al. 1999; Eisen and Wu 2002), (ii) to identify putative lineage-specific duplications (those proteins with a top *E*-value score to another protein from *w*Mel), and (iii) to determine the presence of homologs in different species. Orthologs between the *w*Mel genome and that of the two *Rickettsia* species were identified by requiring mutual best-hit relationships among all possible pairwise BLASTP comparisons, with some manual correction. Those genes present in both *Rickettsia* genomes as well as other bacterial species, but not *w*Mel, were considered to have been lost in the *w*Mel branch (see Table S3). Genes present in only one or two of the three species were considered candidates for gene loss or lateral transfer and were also used to identify possible biological differences between these species (see Table S3). For the *w*Mel genes not in the *Rickettsia* genomes, proteins were searched with BLASTP against the TIGR NRAA database. Protein sequences of their homologs were aligned with CLUSTALW and manually curated. Neighbor-joining trees were constructed using the PHYLIP package.

Phylogenetic analysis of mitochondrial proteins. For phylogenetic analysis, the set of all 38 proteins encoded in both the *Marchantia polymorpha* and *Reclinomonas americana* (Lang et al. 1997) mitochondrial genomes were collected. *Acanthamoeba castellanii* was excluded due to high divergence and extremely long evolutionary branches. Six genes were excluded from further analysis because they were too poorly conserved for alignment and phylogenetic analysis (*nad7*, *rps10*, *sdh3*, *sdh4*, *tatC*, and *yejV*), leaving 32 genes for investigation: *atp6*, *atp9*, *atpA*, *cob*, *cox1*, *cox2*, *cox3*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad9*, *rpl16*, *rpl2*, *rpl5*, *rpl6*, *rps1*, *rps11*, *rps12*, *rps13*, *rps14*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *yejR*, and *yejU*. Using FASTA with the mitochondrial proteins as a query, homologs were identified from the genomes of seven α-Proteobacteria: two intracellular symbionts (*W. pipientis* *w*Mel and *Rickettsia prowazekii*) and five free-living forms (*Sinorhozobium loti*, *Agrobacterium tumefaciens*, *Brucella melitensis*, *Mesorhizobium loti*, and *Rhodopseudomonas* sp.). *Escherichia coli* and *Neisseria meningitidis* were used as outgroups. *Caulobacter crescentus* was excluded from analysis because homologs of some of the 32 genes were not found in the current annotation. In the event that more than one homolog was identified per genome, the one with the greatest sequence identity to the mitochondrial query was retrieved. Proteins were aligned using CLUSTALW (Thompson et al. 1994) and concatenated. To reduce the influence of poorly aligned regions, all sites that contained a gap at any position were excluded from analysis, leaving 6,776 positions per genome for analysis. The data contained extreme amino acid bias: all sequences failed the χ^2 test at $p = 0.95$ for deviation from amino acid frequency distribution assumed under either the JTT or mtREV24 models as determined with PUZZLE (Strimmer and von Haeseler 1996). When the data were iteratively purged of highly variable sites using the method described (Hansmann and Martin 2000), amino acid composition gradually came into better agreement with acid frequency distribution assumed by the model. The longest dataset in which all sequences passed the χ^2 test at $p = 0.95$ consisted of the 3,100 least polymorphic sites. PROTML (Adachi and Hasegawa 1996) analyses of the 3,100-site data using the JTT model detected mitochondria as sisters of the five free-living α-Proteobacteria with low (72%) support, whereas PUZZLE, using the same data, detected mitochondria as sisters of the two intracellular symbionts, also with low (85%) support. This suggested the presence of conflicting signal in the less-biased subset of the data. Therefore, protein log determinants (LogDet) were used to infer distances from the 6,776-site data, since the method can correct for amino acid bias (Lockhart et al. 1994), and Neighbor-Net (Bryant and Moulton 2003) was used to display the resulting matrix, because it can detect and display conflicting signal. The result (see Figure 5A) shows both signals. In no analysis was a sister relationship between *Rickettsia* and mitochondria detected.

For analyses of individual genes, the 63 proteins encoded in the *Reclinomonas* mitochondrial genome were compared with FASTA to the proteins from 49 sequenced eubacterial genomes, which included the α-Proteobacteria shown in Figure 5, *R. conorii*, and *Magnetococcus* MC1, one of the more divergent α-Proteobacteria. Of those proteins, 50 had sufficiently well-conserved homologs to perform phylogenetic analyses. Homologs were aligned and subjected to phylogenetic analysis with PROTML (Adachi and Hasegawa 1996).

Analysis of *wspB* sequences. To compare *wspB* sequences from different *Wolbachia* strains, PCR was done on total DNA extracted from the following sources: *w*Ri was obtained from infected adult *D. simulans*, Riverside strain; *w*AlbB was obtained from the infected Aa23 cell line (O'Neill et al. 1997b), and *D. immitis Wolbachia* was extracted from adult worm tissue. DNA extraction and PCR were done as previously described (Zhou et al. 1998) with *wspB*-specific primers (*wspB*-F, 5′-TTTGCAAGTGAAACAGAAGG and *wspB*-R, 5′-GCTTTGCTGGCAAAATGG). PCR products were cloned into pGem-T vector (Promega, Madison, Wisconsin, United States) as previously described (Zhou et al. 1998) and sequenced (Genbank accession numbers AJ580921–AJ508923). These sequences were compared to previously sequenced *wsp* genes for the same *Wolbachia* strains (Genbank accession numbers AF020070, AF020059, and AJ252062). The four partial *wsp* sequences were aligned using CLUSTALV (Higgins et al. 1992) based on the amino acid translation of each gene and similarly with the *wspB* sequences. Genetic distances were calculated using the Kimura 2 parameter method and are reported in Table S5.

Type IV secretion system. To determine whether the *vir*-like CDSs, as well as adjacent ORFs, were actively expressed within *w*Mel as two polycistronic operons, RT-PCR was used. Total RNA was isolated from infected *D. melanogaster* yw^{67c23} adults using Trizol reagent (Invitrogen, Carlsbad, California, United States) and cDNA synthesized using SuperScript III RT (Invitrogen) using primers *wspB*R, WD0817R, WD0853R, and WD0852R. RNA isolation and RT were done according to manufacturer's protocols, with the exception that suggested initial incubation of RNA template and primers at 65°C for 5 min and final heat denaturation of RT-enzyme at 70°C for 15 min were not done. PCR was done using r*Taq* (Takara, Kyoto, Japan), and several primer sets were used to amplify regions spanning adjacent CDSs for most of the two operons. For operon virB3-WD0853, the following primers were used: (*virB3-virB4*)F, (*virB3-virB4*)R, (*virB6*-WD0856)F, (*virB6*-WD0856)R, (WD0856-WD0855)F, (WD0856-WD0855)R, (WD0854-WD0853)F, (WD0854-WD0853)R. For operon *virB8-wspB*, the following primers were used: (*virB8-virB9*)F, (*virB8-virB9*)R, (*virB9-virB11*)F, (*virB9-virB11*)R, (*virB11-virD4*)F, (*virB11-virD4*)R, (*virD4-wspB*)F, and (*virD4-wspB*)R. The coexpression of *virB4* and *virB6*, as well as WD0855 and WD0854, was confirmed within the putative *virB3*-WD0853 operon using nested PCR with the following primers: (*virB4-virB6*)F1, (*virB4-virB6*)R1, (*virB4-virB6*)F2, (*virB4-virB6*)R2, (WD0855-WD0854)F1, (WD0855-WD0854)R1, (WD0855-WD0854)F2, and (WD0855-WD0854)R2. All ORFs within the putative *virB8-wspB* operon were shown to be coexpressed and are thus considered to be a genuine operon. All products were amplified only from RT-positive reactions (see Figure 6). Primer sequences are given in Table S9.

Supporting Information

Figure S1. Phage Trees

Phylogenetic tree showing the relationship between WO-A and WO-B phage from *w*Mel with reported phage from *w*Kue and *w*Tai. The tree was generated from a CLUSTALW multiple sequence alignment (Thompson et al. 1994) using the PROTDIST and NEIGHBOR programs of PHYLIP (Felsenstein 1989).

Found at DOI: 10.1371/journal.pbio.0020069.sg001 (60 KB PDF).

Figure S2. Plot of the Effective Number of Codons against GC Content at the Third Codon Position (GC3)

Proteins with fewer than 100 residues are excluded from this analysis because their effective number of codon (ENc) values are unreliable. The curve shows the expected ENc values if codon usage bias is caused by GC variation alone. Colors: yellow, hypothetical; purple, mobile element; blue, others. Most of the variation in codon bias can be traced to variation in GC, indicating that the mutation forces dominate the *w*Mel codon usage. Multivariate analysis of codon usage was performed using the CODONW package (available from http://www.molbiol.ox.ac.uk/cu/codonW.html).

Found at DOI: 10.1371/journal.pbio.0020069.sg002 (289 KB PDF).

Figure S3. Predicted Metabolism and Transport in *w*Mel

Overview of the predicted metabolism (energy production and organic compounds) and transport in *w*Mel. Transporters are grouped by predicted substrate specificity: inorganic cations (green), inorganic anions (pink), carbohydrates (yellow), and amino acids/

peptides/amines/purines and pyrimidines (red). Transporters in the drug-efflux family (labeled as "drugs") and those of unknown specificity are colored black. Arrows indicate the direction of transport. Energy-coupling mechanisms are also shown: solutes transported by channel proteins (double-headed arrow); secondary transporters (two-arrowed lines, indicating both the solute and the coupling ion); ATP-driven transporters (ATP hydrolysis reaction); unknown energy-coupling mechanism (single arrow). Transporter predictions are based upon a phylogenetic classification of transporter proteins (Paulsen et al. 1998).

Found at DOI: 10.1371/journal.pbio.0020069.sg003 (167 KB PDF).

Table S1. Repeats of Greater Than 50 bp in the *w*Mel Genome (with Coordinates)

Found at DOI: 10.1371/journal.pbio.0020069.st001 (649 KB DOC).

Table S2. Inactivated Genes in the *w*Mel Genome

Found at DOI: 10.1371/journal.pbio.0020069.st002 (147 KB DOC).

Table S3. Ortholog Comparison with *Rickettsia* spp.

Found at DOI: 10.1371/journal.pbio.0020069.st003 (718 KB XLS).

Table S4. Putative Lineage-Specific Gene Duplications in *w*Mel

Found at DOI: 10.1371/journal.pbio.0020069.st004 (116 KB DOC).

Table S5. Genetic Distances as Calculated for Alignments of *wsp* and *wspB* Gene Sequences from the Same *Wolbachia* Strains

Found at DOI: 10.1371/journal.pbio.0020069.st005 (24 KB DOC).

Table S6. Putative DNA Repair and Recombination Genes in the *w*Mel Genome

Found at DOI: 10.1371/journal.pbio.0020069.st006 (26 KB DOC).

Table S7. Phylogenetic Results for Concatenated Data of 32 Mitochondrial Proteins

Found at DOI: 10.1371/journal.pbio.0020069.st007 (29 KB DOC).

Table S8. Individual Phylogenetic Results for *Reclinomonas* Mitochondrial DNA-Encoded Proteins

Found at DOI: 10.1371/journal.pbio.0020069.st008 (91 KB DOC).

Table S9. PCR Primers

Found at DOI: 10.1371/journal.pbio.0020069.st009 (28 KB DOC).

Accession Numbers

The complete sequence for *w*Mel has been given GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) accession ID number AE017196 and is available through the TIGR Comprehensive Microbial Resource at http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=dmg

The GenBank accession numbers for other sequences discussed in this paper are AF020059 (*Wolbachia* sp. *w*AlbB outer surface protein precursor *wsp* gene), AF020070 (*Wolbachia* sp. *w*Ri outer surface protein precursor *wsp* gene), AJ252062 (*Wolbachia* endosymbiont of *D. immitis* sp. gene for surface protein), AJ580921 (*Wolbachia* endosymbiont of *D. immitis* partial *wspB* gene for *Wolbachia* surface protein B), AJ580922 (*Wolbachia* endosymbiont of *A. albopictus* partial *wspB* gene for *Wolbachia* surface protein B), and AJ580923 (*Wolbachia* endosymbiont of *D. simulans* partial *wspB* gene for *Wolbachia* surface protein B).

Acknowledgments

We acknowledge Barton Slatko, Jeremy Foster, New England Biolabs, and Mark Blaxter for helping inspire this project; Rehka Seshadri for help in examining pathogenicity factors and reading the manuscript; Derek Fouts for examination of group II introns; Susan Lo, Michael Heaney, Vadim Sapiro, and Billy Lee for IT support; Maria-Ines Benito, Naomi Ward, Michael Eisen, Howard Ochman, and Vincent Daubin for helpful discussions; Steven Salzberg and Mihai Pop for help in comparing *w*Mel with the *D. melanogaster* genome; Elodie Ghedin for access to the *B. malayi Wolbachia* sequence data; Maria Ermolaeva for assistance with analysis of operons; Dan Haft for designing protein family hidden Markov models for annotation; Owen White for general bioinformatics support; four anonymous reviewers for very helpful comments and suggestions; and Claire M. Fraser for continuing support of TIGR's scientific research. This project was supported by grant UO1-AI47409–01 to Scott O'Neill and Jonathan A. Eisen from the National Institutes of Allergy and Infectious Diseases.

Conflicts of interest. The authors have declared that no conflicts of interest exist.

Author contributions. M. Wu contributed ideas and analysis in all aspects of the work. L. Sun performed purification of *w*Mel DNA for initial libraries and closure. J. Vamathevan was the closure team leader, performed sequence assembly and analysis, and screened contigs against the *Drosophila* genome. M. Riegler performed validation of assembly against the physical map and confirmation of rearrangements by long PCR and analysis of repeat regions. R. Deboy was the annotation leader and managed the annotation, ORF management, and frameshifts. J. C. Brownlie performed analysis of Type IV secretion systems. E. A. McGraw performed validation of assembly against physical map and confirmation of rearrangements by long PCR and analysis of *wsp* paralogs. W. Martin, C. Esser, N. Ahmadinejad, and C. Wiegand performed the mitochondrial evolution analysis. R. Madupu, M. J. Beanan, L. M. Brinkac, S. C. Daugherty, A. S. Durkin, J. F. Kolonay, and W. C. Nelson performed genome annotation. Y. Mohamoud, P. Lee, and K. Berry performed the closure experiments (closed sequencing gaps, multiplex PCR, resolution of small repeats, coverage reactions, contig editing, resolution of large repeats by transposon and primer walking). M. B. Young was the shotgun sequencing leader. T. Utterback and J. Weidman performed shotgun sequencing and frameshift checking; Utterback also worked on the assembly. W. C. Nierman handled the library construction. I. T. Paulsen performed transporter analysis. K. E. Nelson performed metabolism analysis. H. Tettelin analyzed genome properties, repeats, and membrane proteins. S. L. O'Neill and J. A. Eisen supplied ideas, coordination, and analysis; Eisen is the corresponding author. ■

References

Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42: 459–468.

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, et al. (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, W*igglesworthia glossinidia*. Nat Genet 32: 402–407.

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, et al. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396: 133–140.

Andersson SG, Karlberg O, Canback B, Kurland CG (2003) On the origin of mitochondria: A genomics perspective. Philos Trans R Soc Lond B Biol Sci 358: 165–167.

Bazinet C, Rollins JE (2003) *Rickettsia-like* mitochondrial motility in *Drosophila* spermiogenesis. Evol Dev 5: 379–385.

Bjorkholm B, Sjolund M, Falk PG, Berg OG, Engstrand L, et al. (2001) Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. Proc Natl Acad Sci U S A 98: 14607–14612.

Boyle L, O'Neill SL, Robertson HM, Karr TL (1993) Interspecific and intraspecific horizontal transfer of *Wolbachia* in *Drosophila*. Science 260: 1796–1799.

Braig HR, Zhou W, Dobson SL, O'Neill SL (1998) Cloning and characterization of a gene encoding the major surface protein of the bacterial endosymbiont *Wolbachia pipientis*. J Bacteriol 180: 2373–2378.

Bryant D, Moulton V (2003) Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 20: Dec 5 [Epub ahead of print].

Caturegli P, Asanovich KM, Walls JJ, Bakken JS, Madigan JE, et al. (2000) *ankA*: An *Ehrlichia phagocytophila* group gene encoding a cytoplasmic protein antigen with ankyrin repeats. Infect Immun 68: 5277–5283.

Christie PJ (2001) Type IV secretion: Intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. Mol Microbiol 40: 294–305.

Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, et al. (1999) Alignment of whole genomes. Nucleic Acids Res 27: 2369–2376.

Dumler SJ, Barbet AF, Bekker CPJ, Dasch GA, Palmer GH, et al. (2001) Reorganization of genera in the families Rickettsiaceae and Anaplasmataceae in the order Rickettsiales: Unification of some species of *Ehrlichia* with *Anaplasma*, *Cowdria* with *Ehrlichia* and *Ehrlichia* with *Neorickettsia*—Descriptions of six new species combinations and designation of Ehrlichiaqui and "HGE agent" as subjective synonyms of *Ehrlichia phagocytophila*. Intl J System Evol Microbiol 51: 2145–2165.

Eiglmeier K, Parkhill J, Honore N, Garnier T, Tekaia F, et al. (2001) The decaying genome of *Mycobacterium leprae*. Lepr Rev 72: 387–398.

Eisen JA (1997) Gastrogenomic delights: A movable feast. Nat Med 3: 1076–1078.

Eisen JA (1998a) A phylogenomic study of the MutS family of proteins. Nucleic Acids Res 26: 4291–4300.

Eisen JA (1998b) Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8: 163–167.
Eisen JA, Fraser CM (2003) Phylogenomics: Intersection of evolution and genomics. Science 300: 1706–1707.
Eisen JA, Hanawalt PC (1999) A phylogenomic study of DNA repair genes, proteins, and processes. Mutat Res 435: 171–213.
Eisen JA, Wu M (2002) Phylogenetic analysis and gene functional predictions: Phylogenomics in action. Theor Popul Biol 61: 481–487.
Eisen JA, Heidelberg JF, White O, Salzberg SL (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 1: RESEARCH0011: 1–9.
Eisen JA, Nelson KE, Paulsen IT, Heidelberg JF, Wu M, et al. (2002) The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. Proc Natl Acad Sci U S A 99: 9509–9514.
Elfring LK, Axton JM, Fenger DD, Page AW, Carminati JL, et al. (1997) *Drosophila* PLUTONIUM protein is a specialized cell cycle regulator required at the onset of embryogenesis. Mol Biol Cell 8: 583–593.
Emelyanov VV (2001a) Evolutionary relationship of Rickettsiae and mitochondria. FEBS Lett 501: 11–18.
Emelyanov VV (2001b) Rickettsiaceae, *Rickettsia*-like endosymbionts, and the origin of mitochondria. Biosci Rep 21: 1–17.
Emelyanov VV (2003a) Mitochondrial connection to the origin of the eukaryotic cell. Eur J Biochem 270: 1599–1618.
Emelyanov VV (2003b) Phylogenetic affinity of a *Giardia lamblia* cysteine desulfurase conforms to canonical pattern of mitochondrial ancestry. FEMS Microbiol Lett 226: 257–266.
Felsenstein J (1989) PHYLIP—Phylogeny inference package (version 3.2). Cladistics 5: 164–166.
Frank AC, Amiri H, Andersson SG (2002) Genome deterioration: Loss of repeated sequences and accumulation of junk DNA. Genetica 115: 1–12.
Gray MW, Burger G, Lang BF (2001) The origin and early evolution of mitochondria. Genome Biol 2: REVIEWS1018.
Gupta RS (1995) Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. Mol Microbiol 15: 1–11.
Hansmann S, Martin W (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: Influence of excluding poorly alignable sites from analysis. Int J Syst Evol Microbiol 50: 1655–1663.
Higgins D, Bleasby A, Fuchs R (1992) ClustalV: Improved software for multiple sequence alignment. Comput Appl Biosci 8: 189–191.
Holmes DS, Bonner J (1973) Preparation, molecular weight, base composition, and secondary structure of giant nuclear ribonucleic acid. Biochemistry 12: 2330–2338.
Hryniewicz-Jankowska A, Czogalla A, Bok E, Sikorsk AF (2002) Ankyrins, multifunctional proteins involved in many cellular pathways. Folia Histochem Cytobiol 40: 239–249.
Itoh T, Martin W, Nei M (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc Natl Acad Sci U S A 99: 12944–12948.
Jamnongluk W, Kittayapong P, Baimai V, O'Neill SL (2002) *Wolbachia* infections of tephritid fruit flies: Molecular evidence for five distinct strains in a single host species. Curr Microbiol 45: 255–260.
Jeyaprakash A, Hoy MA (2000) Long PCR improves *Wolbachia* DNA amplification: *wsp* sequences found in 76% of sixty-three arthropod species. Insect Mol Biol 9: 393–405.
Karlin S, Brocchieri L (2000) Heat shock protein 60 sequence comparisons: Duplications, lateral transfer, and mitochondrial evolution. Proc Natl Acad Sci U S A 97: 11348–11353.
Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T (2002) Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci U S A 99: 14280–14285.
Kurtz S, Schleiermacher C (1999) REPuter: Fast computation of maximal repeats in complete genomes. Bioinformatics 15: 426–427.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, et al. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387: 493–497.
Lang BF, Seif E, Gray MW, O'Kelly CJ, Burger G (1999) A comparative genomics approach to the evolution of eukaryotes and their mitochondria. J Eukaryot Microbiol 46: 320–326.
Lawrence JG (2001) Catalyzing bacterial speciation: Correlating lateral transfer with genetic headroom. Syst Biol 50: 479–496.
Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: Rates of change and exchange. J Mol Evol 44: 383–397.
Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A 95: 9413–9417.
Lin M, Rikihisha Y (2003) *Ehrlichia chaffeensis* and *Anaplasma phagocytophilum* lack genes for lipid A biosynthesis and incorporate cholesterol for their survival. Infect Immun 71: 5324–5331.
Lo N, Casiraghi M, Salati E, Bazzocchi C, Bandi C (2002) How many *Wolbachia* supergroups exist? Mol Biol Evol 19: 341–346.
Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic evolutionary model. Mol Biol Evol 11: 605–612.
Martin W (1999) Mosaic bacterial chromosomes: A challenge en route to a tree of genomes. Bioessays 21: 99–104.
Masui S, Sasaki T, Ishikawa H (2000) Genes for the type IV secretion system in an intracellular symbiont, *Wolbachia*, a causative agent of various sexual alterations in arthropods. J Bacteriol 182(22): 6529–6531.
Masui S, Kuroiwa H, Sasaki T, Inui M, Kuroiwa T, et al. (2001) Bacteriophage WO and virus-like particles in *Wolbachia*, an endosymbiont of arthropods. Biochem Biophys Res Commun 283: 1099–1104.
McGraw EA, Merritt DJ, Droller JN, O'Neill SL (2001) *Wolbachia*-mediated sperm modification is dependent on the host genotype in *Drosophila*. Proc R Soc Lond B Biol Sci 268: 2565–2570.
Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17: 589–596.
Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93: 2873–2878.
Moran NA, Mira A (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol 2: RESEARCH0054.
Muller M, Martin W (1999) The genome of *Rickettsia prowazekii* and some thoughts on the origin of mitochondria and hydrogenosomes. Bioessays 21: 377–381.
O'Neill SL, Hoffmann AA, Werren JH, editors (1997a) Influential passengers: Inherited microorganisms and arthropod reproduction. Oxford: Oxford University Press. 228 p.
O'Neill SL, Pettigrew MM, Sinkins SP, Braig HR, Andreadis TG, et al. (1997b) *In vitro* cultivation of *Wolbachia pipientis* in an *Aedes albopictus* cell line. Insect Mol Biol 6: 33–39.
Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, et al. (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293: 2093–2098.
Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, et al. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413: 523–527.
Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet 35: 32–40.
Paulsen IT, Sliwinski MK, Saier MH Jr (1998) Microbial genome analyses: Global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. J Mol Biol 277: 573–592.
Paulsen IT, Nguyen L, Sliwinski MK, Rabus R, Saier MH Jr (2000) Microbial genome analyses: Comparative transport capabilities in eighteen prokaryotes. J Mol Biol 301: 75–100.
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
Penny D, McComish BJ, Charleston MA, Hendy MD (2001) Mathematical elegance with biochemical realism: The covarion model of molecular evolution. J Mol Evol 53: 711–723.
Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, et al. (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 28: 1397–1406.
Roelofs J, Van Haastert PJ (2001) Genes lost during evolution. Nature 411: 1013–1014.
Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26: 544–548.
Salzberg SL, White O, Peterson J, Eisen JA (2001) Microbial genes in the human genome: Lateral transfer or gene loss? Science 292: 1903–1906.
Selby CP, Witkin EM, Sancar A (1991) *Escherichia coli mfd* mutant deficient in "mutation frequency decline" lacks strand-specific repair: *In vitro* complementation with purified coupling factor. Proc Natl Acad Sci U S A 88: 11574–11578.
Seshadri R, Paulsen IT, Eisen JA, Read TD, Nelson KE, et al. (2003) Complete genome sequence of the Q-fever pathogen *Coxiella burnetii*. Proc Natl Acad Sci U S A 100: 5455–5460.
Sexton JA, Vogel JP (2002) Type IVB secretion by intracellular pathogens. Traffic 3: 178–185.
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.
Sicheritz-Ponten T, Kurland CG, Andersson SG (1998) A phylogenetic analysis of the cytochrome b and cytochrome c oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. Biochim Biophys Acta 1365: 545–551.
Sinkins SP, O'Neill SL (2000) *Wolbachia* as a vehicle to modify insect populations. In: James AA, editor. Insect transgenesis: Methods and applications. Boca Raton, Florida: CRC Press. pp. 271–288.
Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, et al. (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411: 940–944.
Strimmer K, von Haeseler A (1996) Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13: 964–969.
Sun LV, Foster JM, Tzertzinis G, Ono M, Bandi C, et al. (2001) Determination of *Wolbachia* genome size by pulsed-field gel electrophoresis. J Bacteriol 183: 2219–2225.

Sun LV, Riegler M, O'Neill SL (2003) Development of a physical and genetic map of the virulent Wolbachia strain *w*MelPop. J Bacteriol 185: 7077–7084.
Sutton G, White O, Adams M, Kerlavage A (1995) TIGR assembler: A new tool for assembling large shotgun sequencing projects. Genome Sci Tech 1: 9–19.
Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, et al. (2002) 50 million years of genomic stasis in endosymbiotic bacteria. Science 296: 2376–2379.
Taylor MJ (2002) A new insight into the pathogenesis of filarial disease. Curr Mol Med 2: 299–302.
Taylor MJ, Hoerauf A (2001) A new approach to the treatment of filariasis. Curr Opin Infect Dis 14: 727–731.
Taylor MJ, Bandi C, Hoerauf AM, Lazdins J (2000) *Wolbachia* bacteria of filarial nematodes: A target for control? Parasitol Today 16: 179–180.
Tettelin H, Radune D, Kasif S, Khouri H, Salzberg SL (1999) Optimized multiplex PCR: Efficiently closing a whole-genome shotgun sequencing project. Genomics 62: 500–507.
Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, et al. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 293: 498–506.
Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
Tram U, Sullivan W (2002) Role of delayed nuclear envelope breakdown and mitosis in *Wolbachia*-induced cytoplasmic incompatibility. Science 296: 1124–1126.
van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, et al. (2003) Reductive genome evolution in *Buchnera aphidicola*. Proc Natl Acad Sci U S A 100: 581–586.
Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381: 364–366.
Viale AM, Arakaki AK (1994) The chaperone connection to the origins of the eukaryotic organelles. FEBS Lett 341: 146–151.
Volfovsky N, Haas BJ, Salzberg SL (2001) A clustering method for repeat analysis in DNA sequences. Genome Biol 2: RESEARCH0027.
Ware J, Moran L, Foster J, Posfai J, Vincze T, et al. (2002) Sequencing and analysis of a 63 kb bacterial artificial chromosome insert from the *Wolbachia* endosymbiont of the human filarial parasite *Brugia malayi*. Int J Parasitol 32: 159–166.
Wernegreen J, Moran NA (1999) Evidence for genetic drift in endosymbionts (*Buchnera*): Analyses of protein-coding genes. Mol. Biol. Evol. 16: 83–97.
Werren JH (1998) *Wolbachia* and speciation. In: Berlocher SH, editor. Endless forms: Species and speciation. New York: Oxford University Press. pp. 245–260.
Werren JH, O'Neill SL (1997) The evolution of heritable symbionts. In: O'Neill SL, Hoffmann AA, Werren JH, editors. Influential passengers: Inherited microorganisms and arthropod reproduction. Oxford: Oxford University Press. pp. 1–41.
Werren JH, Windsor DM (2000) *Wolbachia* infection frequencies in insects: Evidence of a global equilibrium? Proc R Soc Lond B Biol Sci 267: 1277–1285.
Witkin EM (1994) Mutation frequency decline revisited. Bioessays 16: 437–444.
Zhou W, Rousset F, O'Neill SL (1998) Phylogeny and PCR-based classification of *Wolbachia* strains using *wsp* gene sequences. Proc R Soc Lond B Biol Sci 265: 509–515.