Phylogenomics of the Reproductive Parasite

*Wolbachia pipientis wMel*: A Streamlined Genome Overrun by Mobile Genetic Elements

Martin Wu1, Ling V. Sun2, Jessica Vamathevan1, Markus Riegler3, Robert Deboy1, Jeremy C. Brownlie3, Elizabeth A. McGraw3, William Martin4, Christian Esser4, Nahal Ahmadinejad4, Christian Wiegand4, Ramana Madupu1, Maureen J. Beanan1, Lauren M. Brinkac1, Sean C. Daugherty1, A. Scott Durkin1, James F. Kolonay1, William C. Nelson1, Yasmin Mohamoud1, Perris Lee1, Kristi Berry1, M. Brook Young1, Teresa Utterback1, Janice Weidman1, William C. Nierman1, Ian T. Paulsen1, Karen E. Nelson1, Hervé Tettelin1, Scott L. O’Neill2,3, Jonathan A. Eisen1*

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 wMel*, 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 wMel genome, in particular phylogenomic comparisons with other intracellular bacteria, has revealed many insights into the biology and evolution of wMel and *Wolbachia* in general. For example, the wMel 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 wMel, 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 wMel 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 wMel to infect the germline of its host, we find no evidence for either recent lateral gene transfer between wMel 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 wMel-*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 transovarially 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 *Anaplasmataceae*, *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
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 wMel, a strain from the A supergroup that naturally infects Drosophila melanogaster (Zhou et al. 1998).

### Results/Discussion

#### Genome Properties

The wMel 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 wMelPop (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>
<thead>
<tr>
<th>Genome size</th>
<th>1,267,782</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted CD5</td>
<td>1,270</td>
</tr>
<tr>
<td>Average gene length</td>
<td>852</td>
</tr>
<tr>
<td>Percent coding</td>
<td>85.4%</td>
</tr>
<tr>
<td>Assigned function</td>
<td>719 (56.6%)</td>
</tr>
<tr>
<td>Conserved hypothetical</td>
<td>123 (9.7%)</td>
</tr>
<tr>
<td>Unknown function</td>
<td>91 (7.2%)</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>337 (26.5%)</td>
</tr>
<tr>
<td>Transfer RNA</td>
<td>34</td>
</tr>
<tr>
<td>Ribosomal RNA</td>
<td>1 each of 5S, 16S, 23S</td>
</tr>
<tr>
<td>Structural RNAs</td>
<td>2</td>
</tr>
<tr>
<td>Prophage</td>
<td>3</td>
</tr>
<tr>
<td>GC content</td>
<td>35.2%</td>
</tr>
</tbody>
</table>

DOI: 10.1371/journal.pbio.0020069.t001

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 χ² 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
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 wMel 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 wMel 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. wKue (Masui et al. 2001) (Figure 2). Thus, we have named them wMel WO-A and WO-B, based upon their location in the genome. wMel WO-B has undergone a major rearrangement and translocation, suggesting it is inactive. Phylogenetic analysis indicates that wMel WO-B is more closely related to the wKue WO than to wMel WO-A (Figure S1). Thus, wMel WO-A likely represents either a separate insertion event in the Wolbachia lineage or a duplication that occurred prior to the separation of the wMel and wKue 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 wMel 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-

<table>
<thead>
<tr>
<th>Repeat Class</th>
<th>Size (Median)</th>
<th>Copies</th>
<th>Protein Motifs/Families</th>
<th>IS Family</th>
<th>Possible Terminal Inverted Repeat Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1512</td>
<td>3</td>
<td>Transposase</td>
<td>IS4</td>
<td>5’-ATACCGGTCAGTTAAG-3’</td>
</tr>
<tr>
<td>2</td>
<td>360</td>
<td>12</td>
<td>—</td>
<td>New</td>
<td>5’-GGCTTTGGTCATCGTA-3’</td>
</tr>
<tr>
<td>3</td>
<td>858</td>
<td>9</td>
<td>Transposase</td>
<td>IS492/IS110</td>
<td>5’-GGCTTTGGTCAT-3’</td>
</tr>
<tr>
<td>4</td>
<td>1404.5</td>
<td>4</td>
<td>Conserved hypothetical, phage terminase</td>
<td>New</td>
<td>5’-ATACCGGAWTSAWTCGCGGTAT-3’</td>
</tr>
<tr>
<td>5</td>
<td>1212</td>
<td>15</td>
<td>Transposase</td>
<td>IS3</td>
<td>5’-TGACCTTTCCAGAAAAAGTGGGAGAAAG-3’</td>
</tr>
<tr>
<td>6</td>
<td>948</td>
<td>13</td>
<td>Transposase</td>
<td>IS5</td>
<td>—</td>
</tr>
<tr>
<td>7</td>
<td>2405.5</td>
<td>8</td>
<td>RT/maturase</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>468</td>
<td>45</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>9</td>
<td>817</td>
<td>3</td>
<td>Conserved hypothetical, transposase</td>
<td>IS8t12</td>
<td>—</td>
</tr>
<tr>
<td>10</td>
<td>238</td>
<td>2</td>
<td>ExoD</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>11</td>
<td>225</td>
<td>2</td>
<td>RT/maturase</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>12</td>
<td>1263</td>
<td>4</td>
<td>Transposase</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>13</td>
<td>572.5</td>
<td>2</td>
<td>Transposase</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>14</td>
<td>433</td>
<td>2</td>
<td>Ankyrin</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>15</td>
<td>201</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>16</td>
<td>1400</td>
<td>6</td>
<td>RT/maturase</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>17</td>
<td>721</td>
<td>2</td>
<td>Transposase</td>
<td>IS630</td>
<td>—</td>
</tr>
<tr>
<td>18</td>
<td>1191.5</td>
<td>2</td>
<td>EF-Tu</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>19</td>
<td>230</td>
<td>2</td>
<td>Hypothetical</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

DOI: 10.1371/journal.pbio.0020069.t002
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 wMel may have fitness consequences since several classes of genes often found in clusters are generally scattered throughout the wMel genome (e.g., ABC transporter subunits, Sec secretion genes, tRNA genes, F-type ATPase genes).

Although the common ancestor of Wolbachia and Rickettsia likely already had a reduced, streamlined genome, wMel 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, wMel 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 2. Phage Alignments and Neighboring Genes
Conserved gene order between the WO phage in Wolbachia sp. wKue and prophage regions of wMel. Putative proteins in wKue (Masui et al. 2001) were searched using TBLASTN against the wMel 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

Figure 3. Alignment of wMel 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 wMel 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 wMel 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 wMel 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), wMel 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 wKueYO, for which the reported wsp gene (gbAB045235) is actually an ortholog of wspB (99.8% sequence identity and located at the end of the vir 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 wMel wspB gene to amplify and then sequence the orthologs from the related wKi and wAlbB 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 wMel

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 wMel. 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 wMel genome to try to understand the origin of the repetitive and mobile DNA and to explain why such repetitive/mobile DNA is present in wMel, but not other streamlined intracellular species.

We propose that the mobile DNA in wMel 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 wMel, 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 Brugia 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 wMel was acquired long enough ago to allow it to “amplify” 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 wMel 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 wMel 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 wMel in the face of apparently strong pressures for streamlining is intriguing. One explanation is that wMel 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 wMel is that wMel 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
(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 wMel than that of the closely related intracellular Rickettsia, which themselves have higher rates than free-living α-Proteobacteria (Figure 4). Additionally, codon bias in wMel 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 wMel, 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, wMel 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

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

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

DOI: 10.1371/journal.pbio.0020069.t003

**Figure 4. Long Evolutionary Branches in wMel**

Maximum-likelihood phylogenetic tree constructed on concatenated protein sequences of 285 orthologs shared among wMel, 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 wMel lineage is about 63% higher than that of C. crescentus, a free-living α-Proteobacteria. wMel has evolved at a slightly higher rate than the Rickettsia 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 αMel 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 αMel 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 αMel 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 αMel, 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 αMel is particularly relevant since this pathway is one of the key steps in regulating mutation rates in other species. In fact, αMel 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 ωMel and Bachnera 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 [Eisen 1997, 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 ωMel 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 αMel 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 αMel, 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 unifying 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...
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 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

![Figure 5. Mitochondrial Evolution Using Concatenated Alignments](http://biology.plosjournals.org/figure/05)

**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
acids, including predicted transporters for proline, asparagine, 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 \( \omega \)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 \( \omega \)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. \( \omega \)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 \( \omega \)Mel, such as the focus on amino acid transport and the presence of limited carbohydrate metabolism, are similar to those found in \textit{Rickettsia}. A major difference with the \textit{Rickettsia} spp. is the absence of the ADP–ATP exchanger protein in \( \omega \)Mel. In \textit{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 \( \omega \)Mel but not \textit{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 \( \omega \)Mel but not \textit{Rickettsia}. Other pathways present in \( \omega \)Mel but not \textit{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 \textit{Rickettsia} species have a relatively large complement of predicted transporters for osmoprotectants, such as proline and glycine betaine, whereas \( \omega \)Mel possesses only two of these systems.

**Regulatory Responses**

The \( \omega \)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 \textit{ctrA}, a transcription regulator for two component systems in other \( \alpha \)-Proteobacteria (WD0732), and two \( \sigma \) 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 \textit{Wolbachia} infect host cells and by which they cause the diverse phenotypic effects on host reproduction and fitness are poorly understood, and the \( \omega \)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 \( \textit{vir} \) genes previously described from \textit{Wolbachia \( wKueYO \) (Masui et al. 2001), an additional four are present in \( \omega \)Mel. Of the nine \( \omega \)Mel \( \textit{vir} \) ORFs, eight are arranged into two separate operons. Similar to the single operon identified in \( wTai \) and \( wKueYO \), the \( \omega \)Mel \( \textit{virB}^{-} \) operons (\( \textit{virB8}, \textit{virB9}, \textit{virB10}, \textit{virB11}, \text{and} \textit{virB4} \) CDSs are adjacent to \( \textit{ wspB} \), forming a 7 kb operon (WD0004–WD0009). The second operon contains \( \textit{virB3}, \textit{virB4}, \text{and} \textit{virB6} \) as well as four additional non-\( \textit{vir} \) CDSs, including three putative membrane-spanning proteins, that form part of a 15.7 kb operon (WD0859–WD0853). Examination of the \textit{Rickettsia conorii} genome shows a similar organization (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 \( \textit{ wspB} \) and WD0853–WD0856 are each expressed as part of the two \( \textit{vir} \) operons and further indicate that these additional encoded proteins are novel components of the \textit{Wolbachia} Type IV secretion system (Figure 6B).

In addition to the two major \( \textit{vir} \) clusters, a paralog of \( \textit{virB8} \) (WD0817) is also present in the \( \omega \)Mel genome. WD0818 is quite divergent from \( \textit{virB8} \) 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 \( \omega \)Mel-infected \textit{Drosophila} (data not shown). PCR primers were designed to all CDSs of the \( \omega \)Mel Type IV secretion system and used to successfully amplify orthologs from the divergent \textit{Wolbachia} strains \( wRi \) and \( wAlbB \) (data not shown). We were able to detect orthologs to all of the \( \omega \)Mel Type IV secretion system components as well as most of the adjacent non-\( \textit{vir} \) CDSs, suggesting that this system is conserved across a range of \( \alpha \)- and \( \beta \)-group \textit{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 \textit{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 \( \omega \)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 FtsZ, and D-aryl-D-alanine carboxypeptidase) 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 polysaccharide. Other regions with anomalous nucleotide composition include the \( \textit{argB} \) gene, which may have been transferred from \( \textit{Rickettsia} \) to \( \omega \)Mel, and the \( \textit{trpC} \) gene, which may have been transferred from \( \textit{Escherichia coli} \) to \( \omega \)Mel.
Wolbachia contain ankyrin repeats (Elfring et al. 1997); and (iv) the activity of cell-cycle-regulating proteins in host cells and may be involved in host cell-cycle regulation of phylogeny relative to other genes in the genome (i.e., the genes in this region have normal \( \omega \)Mel nucleotide composition and branch in phylogenetic trees with genes from other \( \alpha \)-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. \( \omega \)Mel has 25 ankyrin repeat-containing genes, the most currently described for a prokaryote, with \( \omega \)Mel being next with 13. This is particularly striking given \( \omega \)Mel’s relatively small genome size. The functions of the ankyrin repeat-containing proteins in \( \omega \)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 \( \omega \)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 \( \omega \)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 \( \omega \)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 \( \omega \)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. \( \omega \)Mel DNA was obtained from \( D. melanogaster yw^{1118} \) flies that naturally carry the \( \omega \)Mel infection. \( \omega \)Mel...
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 (GGGCGGCCG), 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 0.3 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 × 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. wMel 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 wMel (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 wMel assembly. Primers were designed to match these assemblies and PCR attempted from total DNA of wMel 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 (Volovsky 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 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). Phylogenetic 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).

Conserved genomics. All putative wMel proteins were searched using BLASTP against the predicted proteomes of published complete organosomal genomes and a set of complete plastid, mitochondrial, plasmid, and viral genomes. The results of these searches were used (i) to identify the phylogenetic profile (Pellegrini et al. 2000; Eisen and Wu 2009b), (ii) to identify putative lineage-specific duplications (those proteins with a top E-value score to another protein from wMel), and (iii) to determine the presence of homologs in different species. Orthologs between the wMel genome and that of the two Rickettsia species were identified by requiring mutual best-hit relationships among all possible pairwise BLASTP comparisons, using the method described (Hansmann and Martin 2000), amino acid composition gradually came into better agreement with acid content 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-joining trees were generated using the PHYLIP package (Thompson et al. 1994).

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 similarity. PCR was successfully performed on extremely long fragments were excluded from further analysis because they were too poorly conserved for alignment and phylogenetic analysis (nad7, rps10, sdh3, sdi4, tatC, and yefY), leaving 32 genes for investigation: atp6, atp8, atp10, cob, cox1, cox2, nad1, nad2, nad3, nad4l, nad5, nad6, nad9, rps16, rps2, rpl5, rpl7, rps11, rps12, rps13, rps14, rps19, rps2, rps3, rps4, rps7, rps8, yefR, and yefU. Using FASTA with the mitochondrial genomics database, sequences of their homologs were aligned with CLUSTALW and manually curated. Neighbor-joining trees were constructed using the PHYLIP package (Felsenstein 1989).

Supporting Information

Figure S1. Phage Trees

Phylogenetic tree showing the relationship between WO-A and WO-B phage from wMel with reported phage from wKue and wTai. 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.s001 (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 wMel codon usage. Multivariate analysis of codon usage was performed using the CODONW package (available from http://molly.ox.ac.uk/codonw.html). Found at DOI: 10.1371/journal.pbio.0020069.s002 (289 KB PDF).

Figure S3. Predicted Metabolism and Transport in wMel

Overview of the predicted metabolism (energy production and organic compounds) and transport in wMel. Transporters are grouped by predicted substrate specificity: inorganic cations (green), inorganic anions (pink), carbohydrates (yellow), and amino acids/
peptidyltetrahydrofolate reductase (NADPH) and 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-darrow 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).

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

Table S2. Inactivated Genes in the uMel Genome

Table S3. Ortholog Comparison with Rickettsia spp.

Table S4. Putative Lineage-Specific Gene Duplications in uMel

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

Table S6. Putative DNA Repair and Recombination Genes in the uMel Genome

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

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

Table S9. PCR Primers

Accession Numbers

The complete sequence for uMel 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.pl?dmg=glm

The GenBank accession numbers for other sequences discussed in this paper are AF200959 (Wolbachia sp. aAlB outer surface protein precursor wsp gene), AF200970 (Wolbachia sp. aRi outer surface protein precursor wsp gene), AJ252062 (Wolbachia endosymbiont of D. immitis sp. gene for surface protein), AY380921 (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; Rehia 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 uMel with the D. melanogaster genome; Elodie Ghedin for access to the B. malayi Wolbachia sequence data; Maria Ermolava 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 help with the construction of TIGR’s sections, 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. Nieman handled the library construction. L. 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


Sequence of Wolbachia pipientis wMel

PLOS Biology | http://biology.plosjournals.org

March 2004 | Volume 2 | Issue 3 | Page 0339