Pathogens and Disease, 75, 2017, ftx115

doi: 10.1093/femspd/ftx115
Advance Access Publication Date: 1 November 2017
Shortomics

SHORTOMICS

Draft genome sequence of the *Wolbachia* endosymbiont of *Wuchereria bancrofti* *w*Wb

Matthew Chung[1,2,†,#], Scott T. Small[3,4,#], David Serre[1,2], Peter A. Zimmerman[3,#] and Julie C. Dunning Hotopp[1,2,*,‡,#]

[1]Institute for Genome Sciences, University of Maryland, Baltimore, MD 21201, USA, [2]Department of Microbiology and Immunology, University of Maryland, Baltimore, MD 21201, USA, [3]Center for Global Health and Diseases, Case Western Reserve University, Cleveland, OH 44106, USA and [4]Eck Institute for Global Health, University of Notre Dame, South Bend, IN 46556, USA

[*]**Corresponding author:** Institute for Genome Sciences, University of Maryland, 801 W Baltimore St, Baltimore, MD 21201, USA. Tel: 410-706-5673; E-mail: jdhotopp@som.umaryland.edu
[#]These authors contributed equally to this work.
One sentence summary: The *Wuchereria bancrofti Wolbachia* endosymbiont *w*Wb genome provides insight on the *Wolbachia* core genome from causative agents of lymphatic filariasis and the evolutionary relationship between *Wolbachia* endosymbionts in supergroup D.
Editor: David Rasko
[†]Matthew Chung, http://orcid.org/0000-0002-9545-523X
[‡]Julie C. Dunning Hotopp, http://orcid.org/0000-0003-3862-986X

ABSTRACT

The draft genome assembly of the *Wolbachia* endosymbiont of *Wuchereria bancrofti* (*w*Wb) consists of 1060 850 bp in 100 contigs and contains 961 ORFs, with a single copy of the 5S rRNA, 16S rRNA and 23S rRNA and each of the 34 tRNA genes. Phylogenetic core genome analyses show *w*Wb to cluster with other strains in supergroup D of the *Wolbachia* phylogeny, while being most closely related to the *Wolbachia* endosymbiont of *Brugia malayi* strain TRS (*w*Bm). The *w*Wb and *w*Bm genomes share 779 orthologous clusters with *w*Wb having 101 unclustered genes and *w*Bm having 23 unclustered genes. The higher number of unclustered genes in the *w*Wb genome likely reflects the fragmentation of the draft genome.

Keywords: *Wolbachia*; lymphatic filariasis; nematode; endosymbiont; genomics; *Wuchereria bancrofti*

INTRODUCTION

Lymphatic filariasis afflicts ~120 million individuals worldwide. *Wuchereria bancrofti*, *Brugia malayi* and *B. timori* cause human lymphatic filariasis, with *W. bancrofti* being responsible for >90% of cases (WHO 2016). Most filarial nematodes have an obligate bacterial *Wolbachia* endosymbiont that is required for the proper development and reproduction of the nematode (Taylor, Bandi and Hoerauf 2005). Within the group of *Wolbachia* endosymbionts originating from lymphatic filarial worms, the only sequenced, full-length genome is that of the *Wolbachia* endosymbiont of *B. malayi* (*w*Bm) (Foster *et al.* 2005). While there is an existing sequenced genome available for the *Wolbachia* endosymbiont of *W. bancrofti*, that assembly consists of 763 contigs (Desjardins *et al.* 2013), which equates to ~1 gene per contig. Here, we present an independently sequenced and improved draft genome sequence of the *Wolbachia* endosymbiont of *Wuchereria bancrofti* (*w*Wb).

Received: 20 July 2017; **Accepted:** 31 October 2017

MATERIALS AND METHODS

Genome sequencing, assembly and annotation

The *w*Wb sequences used were obtained during whole-genome sequencing of *Wuchereria bancrofti*, taken from Patient 0022 at the sampling location of Tau, Papua New Guinea: GPS coordinates −3.666163, 142.766774 (Small *et al*. 2016). *Wolbachia* contigs were identified by aligning to the *w*Bm genome using MUMmer v3.0 (Kurtz *et al*. 2004; Foster *et al*. 2005), discarding contigs with <80% identity across <50% of their length. Reads mapping to these putative *Wolbachia* contigs were identified using Bowtie 2 (Langmead and Salzberg 2012; Small *et al*. 2016), extracted and used to construct a new *de novo* assembly using SPAdes v3.6.2 (Bankevich *et al*. 2012). The process was repeated iteratively until no further contigs were added to the assembly. The contigs from the *de novo* assembly were then reordered using Mauve (Rissman *et al*. 2009), with the *w*Bm genome as the reference. The final assembly consists of 100 scaffolds >500 bp with a scaffold N50 of 19 998 bp. GLIMMER v3.02 (Delcher *et al*. 2007) and the IGS Prokaryotic Annotation Pipeline (Galens *et al*. 2011) were used to annotate the *w*Wb assembly.

DNA sequencing reads for BioProject PRJNA275548 have been deposited at NCBI SRA: SRP056161. The whole-genome shotgun project for *w*Wb has been deposited at DDBJ/ENA/GenBank under the accession NJBR00000000. The version described in this paper is version NJBR02000000. The corresponding whole-genome shotgun project for *W. bancrofti* is available at DDBJ/EMBL/GenBank under the accession LAQH00000000.

Phylogenetic and comparative genomic analyses

Mugsy v1.2 (Angiuoli and Salzberg 2011) and MOTHUR v1.22 (Schloss *et al*. 2009) were used to construct a core genome alignment of *w*Wb and 13 other *Wolbachia* strains, spanning members from five of the *Wolbachia* supergroups (Table S1, Supporting Information) (Wu *et al*. 2004; Foster *et al*. 2005; Klasson *et al*. 2008, 2009; Darby *et al*. 2012; Comandatore *et al*. 2013; Ellegaard *et al*. 2013; Nikoh *et al*. 2014; Sutton *et al*. 2014; Cotton *et al*. 2016; Lindsey *et al*. 2016). RAxML v7.3 (Stamatakis 2006) was used to construct a maximum-likelihood phylogenetic tree (bootstrap number = 1000, substitution mode = GTRGAMMA, default for all other settings) from the core genome alignment. Similarly, a core genome alignment was constructed with *w*Wb and its closest related *Wolbachia* strain *w*Bm. NUCmer v3.06 and MUMmerplot v3.5 (Kurtz *et al*. 2004) were used to produce and visualize a synteny plot between *w*Wb and *w*Bm, respectively. Although the *w*Wb contigs were ordered and oriented to the wBm assembly, the mummer plot enables us to visualize any chromosomal rearrangements within a contig. However, rearrangements within the 100 gaps between the contigs cannot be visualized. For the comparative genome analyses, Mugsy clusters (Angiuoli *et al*. 2011) were used to assign all proteins from wWb and wBm to orthologous clusters. A Sybil instance (Riley *et al*. 2012) was used to identify shared and unique genes between the two strains, along with pseudogenes in *w*Wb.

Identification of lower confidence positions in *w*Wb

Putative low-confidence positions in the *w*Wb assembly were assessed using high-sequencing depth and high-sequence variation. To measure sequencing depth, reads were aligned to the *w*Wb genome with Bowtie2 (Langmead and Salzberg 2012), PCR duplicates were removed with PicardTools (http://broadinstitute.github.io/picard), and depth was measured using the depth function of SAMtools v1.1 (Li *et al*. 2009). Regions with elevated sequencing depth were defined as all ≥50 bp stretches with a sequencing depth of ≥4 median absolute deviations from the major mode of the sequencing depth (43.72×) (Fig. S1, Supporting Information). To validate this LGT cutoff method, the sequencing depth thresholds derived from previous studies (Geniez *et al*. 2012; Ioannidis *et al*. 2013; Dunning Hotopp Slatko and Foster 2017) were re-examined using this method and were found to be within 1% of the original published cutoff (Fig. S2a and b, Supporting Information).

To assess regions with high-sequence variation, 5423 variant positions were identified as having ≥20× sequencing depth with at least one alternative base call that consisted of >5% of the reads at the position. In order to find regions of the genome with higher sequence variation, the percentage of variant positions in 50-bp sliding windows was calculated throughout the entire assembly. Regions with high-sequence variation were defined as 50-bp windows with >12.73% (4× average absolute deviations) variant positions (Fig. S3, Supporting Information). ANNOVAR (Wang, Li and Hakonarson 2010) was used to assess possible frameshifts caused by variant base calls in the low-confidence regions.

RESULTS AND DISCUSSION

This *w*Wb draft genome consists of 1060 850 bp with an average G + C content of 34.3% in 100 contigs (maximum length = 59 950 bp; average length = 10 609 bp; average sequencing depth = 35×, major mode sequencing depth = 20×). Using a 582 455-bp core genome alignment from the *w*Wb genome and 13 other *Wolbachia* genomes, with members from 5 of the *Wolbachia* supergroups (Table S1), we created a maximum-likelihood phylogenetic tree (Fig. 1A) that places *w*Wb within *Wolbachia* supergroup D subset, with *w*Ls and *w*Bm, while being most closely related to *w*Bm.

Additionally, a core genome alignment between only *w*Wb and *w*Bm reveals a sequence identity of 96.9%, differing by 31 907 SNPs in the 1046 453 bp genome shared between them, which comprises 98.2% and 96.9% of their total genome lengths, respectively. Synteny between the 100 contigs of wWb and the *w*Bm genome was assessed using NUCmer v3.06 and visualized with MUMmerplot v3.5 (Fig. 1B) (Delcher *et al*. 2002). The synteny plot shows that the *w*Wb contigs are largely syntenic to the *w*Bm genome. However, given that *Wolbachia* endosymbionts have one circular chromosome and the assembly has 100 gaps, there is the potential for synteny changes in these gaps. Furthermore, synteny changes are more likely to occur between similar sequences in a genome, such as duplicated genes, which can result in gaps in the assembly. Therefore, synteny analysis in any draft genome has limitations.

The *w*Wb genome contains 961 ORFs, one copy of each of the 5S rRNA, 16S rRNA and 23S rRNA as well as one copy of each of the 34 tRNA genes. Given that GLIMMER is known to inflate the number of small ORFs in a genome, we removed all ORFs <60 aa and all ORFs coding for hypothetical proteins <100 aa with no ortholog in *w*Bm (Skovgaard *et al*. 2001). Using Sybil (Riley *et al*. 2012) to visualize and interrogate orthologous proteins between *w*Wb and *w*Bm, the two genomes were found to share 779 orthologous clusters, with *w*Wb having 101 unclustered genes and *w*Bm having 23 unclustered genes. While 20 of the 23 unclustered genes in *w*Bm were identified as hypothetical proteins, the other 3 genes were found to code for a RadC-like DNA repair

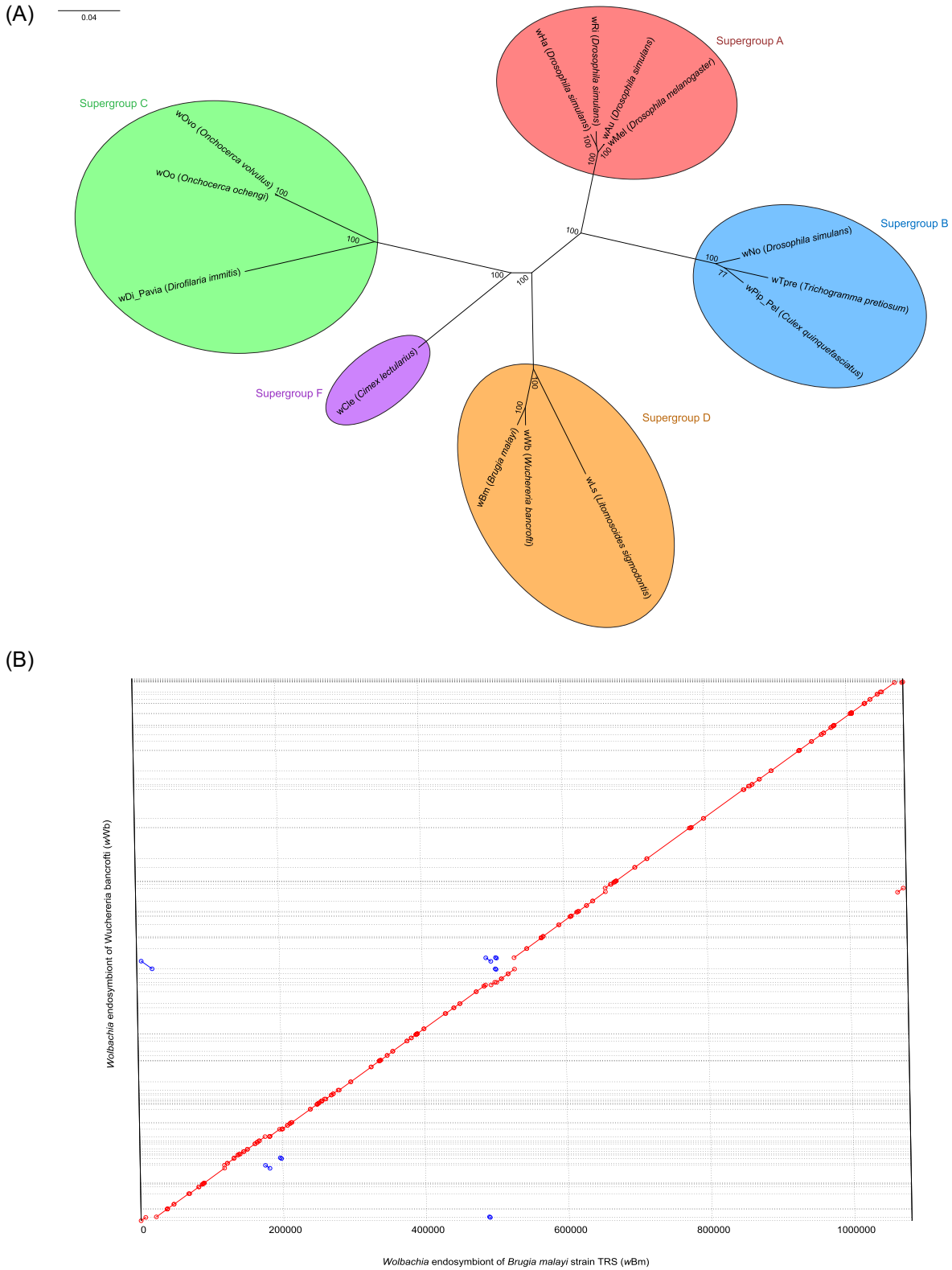

Figure 1. Phylogeny and synteny. (**A**) A RAxML maximum-likelihood phylogenetic tree of 14 *Wolbachia* genomes was constructed based on a 582 455-bp core genome alignment using 1000 bootstraps. The five *Wolbachia* supergroups present in the core genome alignment are denoted by the circles (red, supergroup A; blue, supergroup B; green, supergroup C; orange, supergroup D; and violet, supergroup F). The *w*Wb genome clusters with the genomes of other strains of supergroup D, *w*Bm and *w*Ls, while being most closely related to *w*Bm. (**B**) Synteny between *w*Wb and *w*Bm was compared using NUCMER. Red lines with a slope of 1 are indicative of conserved regions between the two strains, while blue lines with a slope of −1 are indicative of inverted conserved regions. The black dotted horizontal lines represent the boundaries of each of the 100 contigs of wWb. The contigs of *w*Wb cover the entirety of the *w*Bm genome apart from the 100 small breaks between the *w*Wb contigs. While only four small inversions were identified, it is important to consider that more such inversions may occur in the physical gaps between the 100 contigs.

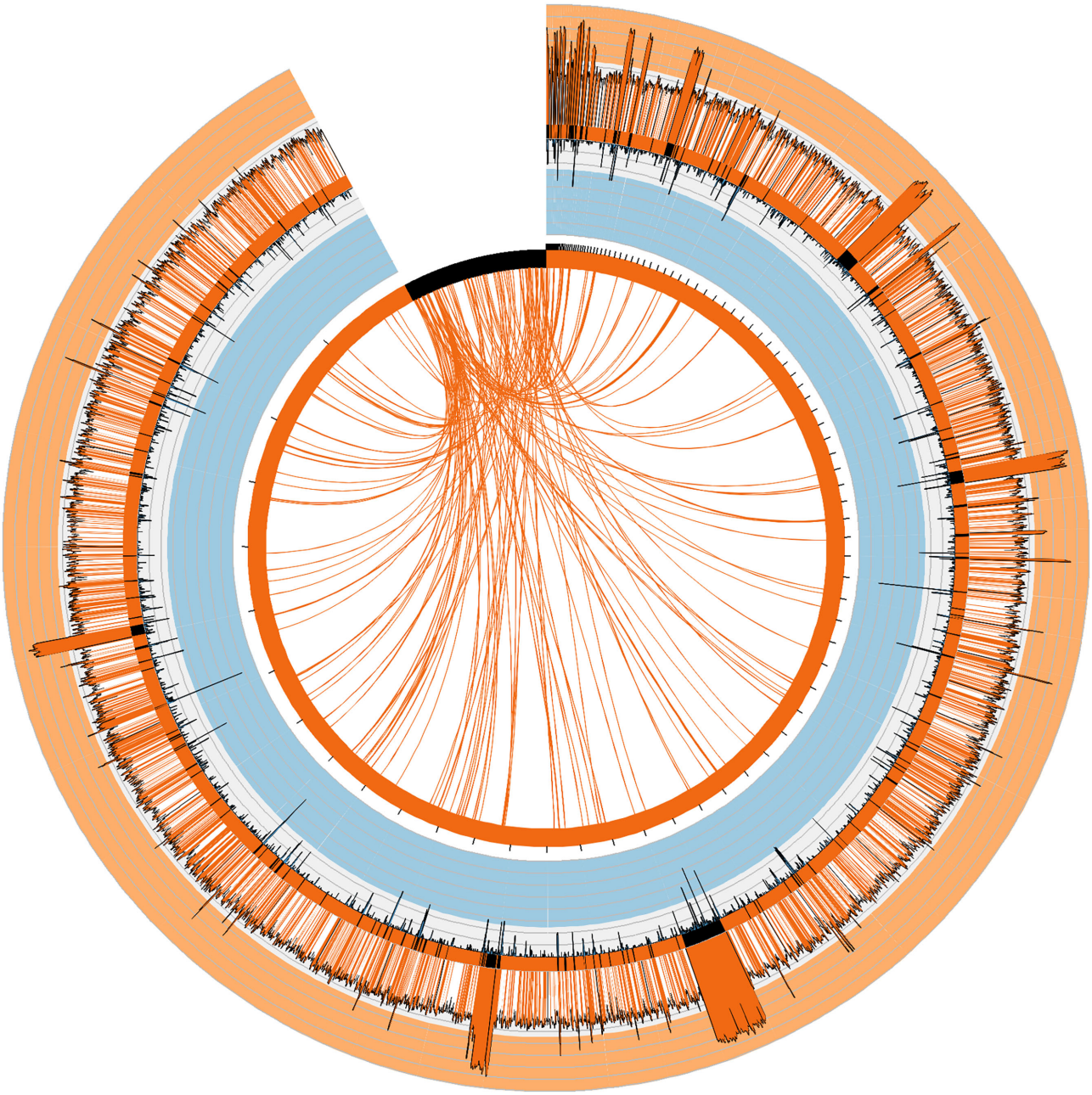

Figure 2. Circos plot of NUCmer linkages between *W. bancrofti* and *w*Wb, *w*Wb sequencing depth, and *w*Wb SNPs and indels. The innermost ring illustrates the concatenated *w*Wb contigs delineated by tick marks (orange) alongside the concatenated *W. bancrofti* contigs (black). The *W. bancrofti* contigs are scaled to 1/1000 the size of *w*Wb contigs and are not delineated by tick marks for visualization purposes, given that there are 5105 *W. bancrofti* contigs. The orange links between the wWb and *W. bancrofti* genomes are indicative of genomic positions present in both the nematode and *Wolbachia* assemblies as determined using MUMmer. The second track, counting outward from the center, contains an inward-facing histogram that indicates the percentage of variant positions in 100 bp bins (blue). Areas with histogram bars that reach the light blue background are indicative of windows with a percentage of variant positions >4 average absolute deviations from the major mode (>12.73%). The third track, flanked by the two histograms, indicates low-confidence regions in the *w*Wb genome, with black indicating regions that fulfill any of our low-confidence criteria and orange indicating normal regions. The fourth track, and outermost track, shows an outward-facing, log2-transformed sequencing depth histogram in 100 bp bins. All positions with <20× sequencing depth are depicted in white, while positions with ≥20× sequencing depth are depicted in orange. All histogram bins that have >43.72× sequencing depth (4× median absolute deviations from the major mode sequencing depth) are indicated by the light-orange background.

protein, an ankyrin repeat-containing protein and elongation factor Tu. All three of these latter genes are duplicated in wBm. Since the *w*Wb genome is incomplete and the library insert size is less than the length of these genes, the assembly is likely to have collapsed in these regions with identical genes being assembled together in one contig instead of separately. Therefore, these genes should not be considered unique to *w*Bm, thus highlighting one of the many nuances of orthologous gene predictions in draft genomes. In the *w*Wb genome, we identified 10 unique ORFs that coded for proteins ≥200 aa, including a bacterial type II and III secretion system protein, 3-dehydroquinate synthase and a pyridoxamine 5′-phosphate synthase. However,

differences in the annotation methods for the *w*Bm and *w*Wb genomes could negatively impact the calculation of orthologs between the two organisms. Additionally, the *w*Wb genome has numerous pseudogenes that will need to be assessed in future research; these could be of interest or could be an artifact in the assembly from inclusion of reads from *Wolbachia*-*Wuchereria* lateral gene transfers (LGTs), a *Wolbachia* sequencing dilemma (Dunning Hotopp *et al.* 2017).

Due to the widespread occurrence of *Wolbachia*-nematode LGT events and the possibility of collapsed repeats in the assembly, we sought to identify lower confidence regions in the *w*Wb genome, where the sequence supports some sequence variation based on three criteria: sequencing depth, sequence variation and the presence of the sequence in the *W. bancrofti* assembly indicative of a putative LGT. Regions with abnormally high-sequencing depth were defined as ≥50 bp stretches with a sequencing depth of ≥4 median absolute deviation from the major mode of the sequencing depth (43.72×), while regions with high-sequence variation were defined as 50-bp windows with ≥12.73% (4× average absolute deviations from the major mode) variant positions. A total of 75 702 and 3144 positions were identified using these criteria respectively, and an additional 26 832 positions were identified as shared between the *w*Wb and *W. bancrofti* assemblies. Integrating all three criteria, a total of 92 821 low-confidence genome sequence positions (8.75% of the *w*Wb genome) spanning 69 contigs were identified, with 12 119 positions being supported by two criteria and 738 positions being supported by all three (Fig. 2, Table S2 and Fig. S4, Supporting Information). Such regions could indicate (i) *Wolbachia*-*Wuchereria* LGT, (ii) collapsed repeats, (iii) population-level variation in the endosymbiont since multiple nematodes were sequenced or (iv) some combination thereof.

To determine whether or not alternative base calls in low-confidence regions could have possibly altered the consensus base call in the *w*Wb assembly, we sought to identified 1974 variant positions with ≥20× sequencing depth and <90% of reads supporting the consensus base call. Of these positions, alternative base calls with >5% read support were identified and analyzed with ANNOVAR (Wang *et al.* 2010) to determine whether these alternative base calls resulted in the possibility of a frameshifted gene call. Using this method, alternative base calls can be differentiated from sequencing errors since this requires at least 1 read to support the alternative base call. Within these 1974 positions, 2234 variant calls were identified with 1891 being SNPs (993 transitions and 898 transversions) and 343 being indels/substitutions. A total of 1335 variants were found in genic regions, with 182 of the variants having the potential to generate a frameshift within gene calls (stop gains, stop losses, frameshift insertions and frameshift deletions) across 67 genes (Table S3). We also identified 3449 variant positions located outside of the low-confidence regions. Despite our ability to identify these variants, we have no means of determining the source of the sequence variation.

SUMMARY

The sequencing and characterization of the *w*Wb genome adds more insight on the evolutionary relationships between the different *Wolbachia* supergroups, specifically supergroup D. The addition of another supergroup D *Wolbachia* genome should aid in future studies delineating core *Wolbachia* supergroup D genome characteristics. However, we continue to demonstrate that the presence of LGTs in the nematode genome has the potential to confound the accurate sequencing of *Wolbachia* endosymbiont genomes.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSPD* online.

FUNDING

This work was funded by the National Institutes of Health (AI103263) to PAZ and the National Institute of Allergy and Infectious Diseases (U19AI110820) to JCDH. STS received support through a T32 Postdoctoral Fellowship in Geographic Medicine and Infectious Disease (AI007024).

AUTHORS' CONTRIBUTIONS

STS, DS and PAZ conceived the study. STS sequenced and assembled the genome, and edited the manuscript. MC annotated the genome, conducted analyses and generated figures. MC and JCDH drafted the manuscript. All authors read and approved the final manuscript.

Conflict of Interest. None declared.

REFERENCES

Angiuoli SV, Dunning Hotopp JC, Salzberg SL *et al.* Improving pan-genome annotation using whole genome multiple alignment. *BMC Bioinformatics* 2011;**12**:272.

Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 2011;**27**:334–42.

Bankevich A, Nurk S, Antipov D *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77.

Comandatore F, Sassera D, Montagna M *et al.* Phylogenomics and analysis of shared genes suggest a single transition to mutualism in *Wolbachia* of nematodes. *Genome Biol Evol* 2013;**5**:1668–74.

Cotton JA, Bennuru S, Grote A *et al.* The genome of Onchocerca volvulus, agent of river blindness. *Nat Microbiol* 2016;**2**:16216.

Darby AC, Armstrong SD, Bah GS *et al.* Analysis of gene expression from the *Wolbachia* genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis. *Genome Res* 2012;**22**:2467–77.

Delcher AL, Bratke KA, Powers EC *et al.* Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 2007;**23**:673–9.

Delcher AL, Phillippy A, Carlton J *et al.* Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 2002;**30**:2478–83.

Desjardins CA, Cerqueira GC, Goldberg JM *et al.* Genomics of Loa loa, a *Wolbachia*-free filarial parasite of humans. *Nat Genet* 2013;**45**:495–500.

Dunning Hotopp JC, Slatko BE, Foster JM. Targeted enrichment and sequencing of recent endosymbiont-host lateral gene transfers. *Sci Rep* 2017;**7**:857.

Ellegaard KM, Klasson L, Naslund K *et al.* Comparative genomics of *Wolbachia* and the bacterial species concept. *PLoS Genet* 2013;**9**:e1003381.

Foster J, Ganatra M, Kamal I *et al.* The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol* 2005;**3**:e121.

Galens K, Orvis J, Daugherty S *et al*. The IGS standard operating procedure for automated prokaryotic annotation. *Stand Genomic Sci* 2011;**4**:244–51.

Geniez S, Foster JM, Kumar S *et al*. Targeted genome enrichment for efficient purification of endosymbiont DNA from host DNA. *Symbiosis* 2012;**58**:201–7.

Ioannidis P, Johnston KL, Riley DR *et al*. Extensively duplicated and transcriptionally active recent lateral gene transfer from a bacterial *Wolbachia* endosymbiont to its host filarial nematode *Brugia malayi*. *BMC Genomics* 2013;**14**:639.

Klasson L, Walker T, Sebaihia M *et al*. Genome evolution of *Wolbachia* strain wPip from the Culex pipiens group. *Mol Biol Evol* 2008;**25**:1877–87.

Klasson L, Westberg J, Sapountzis P *et al*. The mosaic genome structure of the *Wolbachia* wRi strain infecting *Drosophila simulans*. *P Natl Acad Sci U S A* 2009;**106**:5725–30.

Kurtz S, Phillippy A, Delcher AL *et al*. Versatile and open software for comparing large genomes. *Genome Biol* 2004;**5**:R12.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.

Li H, Handsaker B, Wysoker A *et al*. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**: 2078–9.

Lindsey AR, Werren JH, Richards S *et al*. Comparative genomics of a parthenogenesis-inducing *Wolbachia* symbiont. *G3 (Bethesda)* 2016;**6**:2113–23.

Nikoh N, Hosokawa T, Moriyama M *et al*. Evolutionary origin of insect-*Wolbachia* nutritional mutualism. *P Natl Acad Sci USA* 2014;**111**:10257–62.

Riley DR, Angiuoli SV, Crabtree J *et al*. Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics* 2012;**28**:160–6.

Rissman AI, Mau B, Biehl BS *et al*. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 2009; **25**:2071–3.

Schloss PD, Westcott SL, Ryabin T *et al*. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microb* 2009;**75**: 7537–41.

Skovgaard M, Jensen LJ, Brunak S *et al*. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* 2001;**17**:425–8.

Small ST, Reimer LJ, Tisch DJ *et al*. Population genomics of the filarial nematode parasite *Wuchereria bancrofti* from mosquitoes. *Mol Ecol* 2016;**25**:1465–77.

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006;**22**:2688–90.

Sutton ER, Harris SR, Parkhill J *et al*. Comparative genome analysis of *Wolbachia* strain wAu. *BMC Genomics* 2014;**15**:928.

Taylor MJ, Bandi C, Hoerauf A. *Wolbachia* bacterial endosymbionts of filarial nematodes. *Adv Parasitol* 2005;**60**: 245–84.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.

WHO. Global programme to eliminate lymphatic filariasis: progress report, 2015. *Wkly Epidemiol Rec* 2016;**91**: 441–55.

Wu M, Sun LV, Vamathevan J *et al*. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. *PLoS Biol* 2004;**2**:E69.