Presence of Extensive *Wolbachia* Symbiont Insertions Discovered in the Genome of Its Host *Glossina morsitans morsitans*

Corey Brelsfoard1,2,3, George Tsiamis3,4, Marco Falchetto4, Ludvik M. Gomulski4, Erich Telleria1, Uzma Alam1, Vangelis Doudoumis3, Francesca Scolari4, Joshua B. Benoit1,5, Martin Swain6, Peter Takac7, Anna R. Malacrinda4, Kostas Bourtzis3,8,9, Serap Aksoy1*

1 Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, United States of America, 2 Department of Natural Sciences, St. Catharine College, St. Catharine, Kentucky, United States of America, 3 Department of Environmental and Natural Resources Management, University of Patras, Agrinio, Greece, 4 Dipartimento di Biologia e Biotecnologie, Università di Pavia, Pavia, Italia, 5 Department of Biological Sciences, McMicken College of Arts and Sciences, University of Cincinnati, Cincinnati, Ohio, United States of America, 6 Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Penglais, Aberystwyth, Ceredigion, United Kingdom, 7 Institute of Zoology, Section of Molecular and Applied Zoology, Slovak Academy of Science, Bratislava, Slovakia, 8 Biomedical Sciences Research Center Al. Fleming, Vari, Greece, 9 Insect Pest Control Laboratory, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, Vienna, Austria

Abstract

Tsetse flies (*Glossina* spp.) are the cyclical vectors of *Trypanosoma* spp., which are unicellular parasites responsible for multiple diseases, including nagana in livestock and sleeping sickness in humans in Africa. *Glossina* species, including *Glossina morsitans morsitans* (Gmm), for which the Whole Genome Sequence (WGS) is now available, have established symbiotic associations with three endosymbionts: *Wigglesworthia glossinidia*, *Sodalis glossinidius* and *Wolbachia pipientis* (*Wolbachia*). The presence of *Wolbachia* in both natural and laboratory populations of *Glossina* species, including the presence of horizontal gene transfer (HGT) events in a laboratory colony of Gmm, has already been shown. We herein report on the draft genome sequence of the cytoplasmic *Wolbachia* endosymbiont (cytWol) associated with Gmm. By in silico and molecular and cytogenetic analysis, we discovered and validated the presence of multiple insertions of *Wolbachia* (chrWol) in the host Gmm genome. We identified at least two large insertions of chrWol, 527,507 and 484,123 bp in size, from Gmm WGS data. Southern hybridizations confirmed the presence of *Wolbachia* insertions in Gmm genome, and FISH revealed multiple insertions located on the two sex chromosomes (X and Y), as well as on the supernumerary B-chromosomes. We compare the chrWol insertions to the cytWol draft genome in an attempt to clarify the evolutionary history of the HGT events. We discuss our findings in light of the evolution of *Wolbachia* infections in the tsetse fly and their potential impacts on the control of tsetse populations and trypanosomiasis.

Citation: Brelsfoard C, Tsiamis G, Falchetto M, Gomulski LM, Telleria E, et al. (2014) Presence of Extensive Wolbachia Symbiont Insertions Discovered in the Genome of Its Host Glossina morsitans morsitans. PLoS Negl Trop Dis 8(4): e2728. doi:10.1371/journal.pntd.0002728

Editor: Jesus G. Valenzuela, National Institute of Allergy and Infectious Diseases, United States of America

Received August 9, 2013; Accepted January 20, 2014; Published April 24, 2014

Copyright: © 2014 Brelsfoard et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study received support from NIH AI068932 and Ambrose Monell Foundation awarded to SA as well as from the European Community’s Seventh Framework Programme CSA-SA, REGPOT-2008-2 under Grant Agreement 245746 and the Greek Secretariat for Research and Technology (GSRT) under Grant Agreement 12SLO_ET29_1178 to KB. We are grateful to FAO/IAEA Coordinated Research Program “Improving SIT for Tsetse Flies through Research on their Symbionts and Pathogens” for support of this study. VD, GT, SA and KB also acknowledge support from EU COST Action FA0701 “Arthropod Symbiosis: From Framework Programme CSA-SA_ REGPOT-2008-2 under Grant Agreement 245746 and the Greek Secretariat for Research and Technology (GSRT) under Grant Agreement 12SLO_ET29_1178 to KB. We are grateful to FAO/IAEA Coordinated Research Program “Improving SIT for Tsetse Flies through Research on their Symbionts and Pathogens” for support of this study. VD, GT, SA and KB also acknowledge support from EU COST Action FA0701 “Arthropod Symbiosis: From Fundamental Studies to Pest and Disease Management.” This investigation also received financial support from UNICEF/UNDP/World Bank/WHO Special Program for Research and Training in Tropical Diseases (TDR) project ID ABO132 to ARM. Gmm Whole Genome Sequence was obtained at the Welcome Trust Sanger Institute, UK under the leadership of The International Glossina Genome Initiative (IGGI) consortium. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

1 E-mail: K.Bourtzis@iae.org (KB); serap.aksoy@yale.edu (SA)

9 These authors contributed equally to this work.

Introduction

The genus *Wolbachia* encompasses intracellular maternally inherited Gram-negative bacteria estimated to infect over 40% of insect species, in addition to filarial nematodes, crustaceans, and arachnids [1,2]. *Wolbachia* interactions with its host can have diverse outcomes that range from mutualistic to pathogenic or reproductive parasitism [3]. In arthropods, *Wolbachia* alterations to host reproduction include parthenogenesis induction, male killing, feminization of genetic males, and cytoplasmic incompatibility (CI) [1,4]. In its simplest form, CI occurs when a *Wolbachia* infected male mates with an uninfected female, causing developmental arrest of the embryo. In contrast, *Wolbachia* infected females can mate with either an uninfected male or a male infected with the same *Wolbachia* strain, and produce viable *Wolbachia* infected offspring. It has been suggested that the reproductive advantage afforded by the *Wolbachia* induced CI mechanism may permit the rapid spread of desirable host phenotypes into natural populations as a novel disease control approach [4–7].

A number of *Wolbachia* whole genome sequence (WGS) data are available to date and at least ten more genomes are currently being sequenced from a diverse set of hosts [8–15]. The majority of the
Wolbachia strains have genomes that range from 1.08 to 1.7Mb in size [12]. Although most Rickettsiales have small genomes, Wolbachia sets a different pace by carrying an extremely high number of mobile and repetitive elements [4,16,17]. In addition, a number of Ecdysozoan genomes have been reported to contain chromosomal insertions originating from Wolbachia, including the mosquito Aedes aegypti [18,19], the longhorn beetle Monochamus alternatus [20], filarial nematodes of the genera Onchocerca, Brugia, and Drosophila [21,22], parasitoid wasps of the genus Nasonia [22], the fruit fly Drosophila ananassae [22], the pea aphid Acyrthosiphon pisum [23], and the bean beetle Callosobruchus chinensis [23,24]. Horizontal gene transfer (HGT) events in prokaryotes are rather common, and represent a way for bacteria to acquire novel features that enable them to adapt to different environments and to reorganize their genome [25–27]. In unicellular eukaryotes, gene transfer events are also relatively common [28]. Since many unicellular eukaryotes are phagotrophic on bacteria and other micro-organisms, they are constantly exposed to prokaryotic DNA, which may predispose them to incorporate foreign genetic material into their genomes [29]. By contrast, in multi-cellular organisms HGTs are rare [30]. It is likely that the localization of Wolbachia within the host germ-line cells [31] may have enabled the transfer of its genetic material to the host chromosomes.

Tsetse flies are the exclusive vectors of Human African Trypanosomiasis (HAT), also known as sleeping sickness, and of the livestock disease Nagana in sub-Saharan Africa. These diseases are caused by different members of the kinetoplastid protozoan parasites, Trypanosoma spp. The World Health Organization (WHO) has estimated that 60 million people in Africa live in tsetse infested areas, and are at risk of contracting sleeping sickness [32]. Disease control in the mammalian host is complicated due to the lack of vaccines, cheap and effective therapeutic treatments, and simple accurate diagnostic tools [33,34].

Tsetse flies also harbor multiple symbiotic microbes, which display different levels of integration with their host. The obligate mutualist genus Wigglesworthia provides dietary supplements to support host fecundity and is also necessary during larval development for the adult immune maturation processes [35–38]. The facultative symbiont genus Sodalis is present in some individuals in natural populations and may play a role in tsetse’s trypanosome transmission ability (vector competence) [38,39]. The ability to cultivate Sodalis in vitro and transform and repopulate tsetse with modified Sodalis has led to a potential paratransgenic control strategy to modify tsetse’s vector competence by expressing trypanocidal molecules in recSodalis [40–44]. Natural populations of many tsetse species also harbor a third symbiont, which belongs to the genus Wolbachia. Recent surveys indicate that Wolbachia infection prevalence in natural populations of different tsetse species can vary considerably, with some populations having near 100% infection prevalence [41,45]. We recently demonstrated that Wolbachia infections in Glossina morsitans morsitans (Gmm) induce CI in the laboratory and confer a reproductive advantage to infected females [41]. Further modeling of CI demonstrated the potential use of Wolbachia to drive a desirable host phenotype into a natural tsetse population [41,46]. Thus, it is suggested that tsetse carrying modified Sodalis expressing antiparasite molecules in their midgut can be used to replace their wild parasite-susceptible counterparts through Wolbachia-mediated CI. One population control method that has been successful for tsetse, and currently being implemented in Africa, is the sterile insect technique (SIT), where males rendered sterile through irradiation are released to mate with wild females and suppress their fecundity [41,47]. A promising alternative/complementary approach to SIT could be the use of the incompatible insect technique (IIT), which relies on Wolbachia-induced sterility in the released males instead of irradiation [48,49].

In this paper, which is being submitted as a satellite to the manuscript describing the WGS of the tsetse species Gmm, we report on the draft genome sequence of its associated cytoplasmic Wolbachia endosymbiont (cytWob). Moreover, we mined the WGS of Gmm and report on the presence of multiple extensive chromosomal insertions of Wolbachia (chrWob) in the host genome. These results confirmed our previous PCR-amplification based data suggesting the presence of HGT event(s) between Wolbachia and Gmm [45]. The HGT events were validated by Southern blot and Fluorescent in situ Hybridization (FISH) analyses on Gmm chromosomes. We compared the chrWob insertions discovered in the assembled Gmm genome to cytWob to understand the evolution of HGT events, and discuss our findings in light of the evolution of Wolbachia infections in tsetse. Finally, we analyzed the presence of Wolbachia HGT events in several Gmm natural populations, and discuss the potential to harness Wolbachia effects for the control of tsetse-transmitted diseases.

Materials and Methods

Cytoplasmic Wolbachia source DNA and sequencing

For the genome sequencing of the naturally infected Wolbachia strain of G. m. morsitans (wGmm), approximately 250 ovaries were dissected from adult females from the Gmm colony maintained in the Yale University insectary. DNA was prepared using Qiagen DNeasy kit (Qiagen, Inc., Valencia, CA). The complete genome sequence was determined using whole-genome shotgun pyrosequencing using the Roche 454 GS sequencer FLX Titanium system (454 Life Sciences, Branford, CT, USA).

In order to improve the wGmm draft genome, Illumina read libraries from the tsetse genome assembly were used. These were obtained from: (a) a pool of five tsetse flies, and (b) the first larval progeny of tetracycline-treated female. Two sets of Illumina reads were used: a PCR-free small fragment (~300 bp) library and Hi-Seq mate-pair libraries with an insert of approximately 1.6 kb.

Cytoplasmic Wolbachia assembly and annotation

The tsetse ovary DNA used for wGmm sequencing contained a mixture of host genetic material, as well as cytoplasmic (cyt) and...
chromosomal (chr) Wolbachia DNA. A customized informatics pipeline was developed to computationally distinguish between sequence reads. An initial assembly was performed using MIRA [50]. First, all host sequences were removed by mapping the 454 reads to the Wolbachia reference genomes (aMel, aRi, aPip and aBm). The filtered sequence reads contained chromosomal and cytoplasmic reads. The chromosomal reads were further removed using MIRA by mapping the filtered sequences to the chromosomal Wolbachia contigs (99% cut-off). The same procedure was followed for the Illumina data. The resulting 454 and Illumina reads were de novo assembled using MIRA. This initial assembly was subsequently improved using approaches described in the PAGIT protocol [51]. In brief, the contigs were aligned to the aMel genome using ABACAS [52], creating one large scaffold that consisted of the contigs successfully mapped to the aMel genome and a set of contigs that did not map. An attempt was made to close the gaps in the large scaffold using IMAGE [53] with the PCR-free small fragment library. After gap closing, the large scaffold was reduced once more to a set of contigs by breaking it around any of the unclosed gaps. This is because there are usually many genome rearrangements between different Wolbachia strains, and we would therefore expect a number of rearrangements to exist between the aMel and aGmm genomes. Breaking the scaffold makes allowance for these gaps. Finally, scaffolding was then performed on this reduced set of contigs using SCARPA [54] with the Hi-Seq mate-pair libraries. The statistics for the assembly at each stage of the process are given in Table S1. The genome was annotated with XBASE and RAST [55,56], followed by manual curation. Putative protein-encoding genes were identified using GLIMMER [57] and tRNA by trNAScan-SE [50]. Predicted proteins were examined to detect frame-shifts or premature stop codons to identify pseudogenes using ARTEMIS [59]. Those for which the frame-shift or premature stops were of high quality by examining re-mapped reads in these regions were annotated as “authentic” mutations. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AWUH00000000. The version described in this paper is version AWUH01000000.

Chromosomal Wolbachia assembly and annotation

The Sanger and 454 reads used in the tsetse genome assembly were obtained from flies treated with tetracycline as described previously [41]; therefore, these reads did not contain cytWol sequences. As mentioned above, Wolbachia specific sequences were filtered out from WGS reads of each sequencing technology with MIRA [50] using the complete genomes of aMel (AE017190), aRi (CP001391), and aBm (AE017321) as reference sequences. We obtained 5,306 (Sanger), and 10,978 (454) Wolbachia-specific sequences respectively. All the filtered putative Wolbachia-specific sequences were further examined using blast and a custom made Wolbachia database.

CytWol-specific sequences were assembled with MIRA and AMOS [50,60] using as a reference sequence the aGmm draft genome. The statistics for the two cytWol assemblies are as follows: N50 2970, mean contig length 1261.97, longest contig 15053, total length 527,504 bp for insertion A, while for insertion B N50 2791, mean contig length 1092.92, and total length 484,123. Genes were identified with Glimmer [61], followed by a round of manual curation using Blastn [62] and MegaBlast [62] against the non-redundant and custom made Wolbachia databases. The predicted CDSs were translated and used to search the NCBI non-redundant database, KEGG, and COG databases. The tRNAscan-SE tool [58] was used to identify tRNA genes.

Phylogenetic analyses

Phylogenetic analyses were performed using Maximum-Likelihood (ML) and Neighbor-Joining (NJ) estimation for a concatenated set of six phage genes from aGmm, aRi, aMel and aPip. The genes used for the phylogeny included HI97 family phage major capsid protein (aGmm_0802, WD_0458, WRi_002750, WP0102), phage integrase family site-specific recombinase (aGmm_0004, WD_1148, WRi_009900, WP0900), phage SPO1 DNase-related protein (aGmm_0674, WD_0164, WRi_000900, WP0922), prophage LambdaW5 baseplate assembly protein W (aGmm_0971, WD_0640, WRi_005490, WP0303), prophage LambdaW1, baseplate assembly protein Ψ (aGmm_0970, WD_0639, WRi_010130, WP0302) and a prophage LambdaW1, site-specific recombinase resolase protein (aGmm_0960, WD_0634, WRi_005490, WP0342).

In addition, a concatenated set of ten genes (DNA-directed RNA polymerase, DNA polymerase III (alpha subunit), DNA gyrase B, translation elongation factor G, aspartyl-tRNA synthetase, CTP synthase, glutamyl-tRNA synthetase, GTP-binding protein, cell division protein FtsZ, fructose-bisphosphate aldolase) from the identified Gmm chromosomal insertions, aGmm, aRi, aMel, aPip, and aBm were used. All sequences were aligned using MUSCLE [63] and ClustalW [64] as implemented in Geneious 5.4 [65], and adjusted manually. ML and NJ trees were constructed using MEGA 5.0 [66] with gamma distributed rates with 1000 bootstrap replications and the method of Tamura-Nei as genetic distance model [67].

Southern blot hybridization analyses

To determine the number of chromosomal insertions, genomic DNA from tetracycline-treated Gmm females and normal Gmm individuals were restricted with HindIII endonuclease, electrophoresed on 1% agarose gel in 1x TBE buffer, and transferred to a positively charged nylon membrane according to Southern protocol [68]. The membrane was hybridized at 55°C with 350 ng of a 569 bp probe corresponding to part of the wsp gene labeled with the Gene Images Alkphos Direct labeling system (GE Healthcare, Little Chalfont, UK) using the random primer method following manufacturer protocols. Signal detection was performed using CDP-star followed by exposure to autoradiographic film (X-OMAT AR, Kodak). The absence of cytWol from the tetracycline treated Gmm DNA was confirmed by a PCR assay, which resulted in only a single 16S rRNA amplification product originating from the chromosomal insertions [45].

FISH chromosomal preparations and hybridization

Mitotic chromosome spreads were obtained from freshly deposited larvae from the Slovakia Academy of Sciences Institute of Zoology tsetse laboratory Gmm strain. Briefly, larval nerve ganglia were incubated on a slide in 100 μl 1% sodium citrate for 10 min at room temperature, and sodium citrate was replaced with methanol-acetic acid (3:1 solution) for 4 min. The tissue was disrupted by pipetting in 100 μl 60% acetic acid for fixation and dropped onto clean slides heated on a hot plate at 70°C until acetic acid evaporation. After dehydration in 80% ethanol, slides were stored at −20°C for at least 2 weeks.

For in situ hybridization experiments, multiple probes specific for Wolbachia 16S rRNA, fthA and wsp genes were amplified from the Slovakian strain DNA [45,69]. To generate the labeled probes, 1 μg of DNA resuspended in 16 μl ddH2O was denatured by boiling for 10 min. 4 μl of labeling mix (Biotin High Prime kit; Roche, Basel, Switzerland) were added and the reaction was incubated overnight at 37°C. After the reaction was stopped, ddH2O (5 μl), 20×SSC buffer (25 μl) and formamide (50 μl) were added and 25 μl of denatured probe was placed on each pre-treated slide. The hybridization was performed at 37°C overnight.
in a humid chamber and detection of hybridization signals was performed using the Vectastain ABC elite kit (Vector Laboratories, Burlingame, CA, USA) and Alexa Fluor 594 Tyramide (Invitrogen). Chromosomes were DAPI stained and the slides were mounted using the VECTASHIELD mounting medium (Vector Laboratories). Chromosomes were screened under an epi-fluorescence Zeiss Axioplan microscope and images were captured using an Olympus DP70 digital camera. For the localization of signals on mitotic chromosomes the karyotype description of Willhoeft [69,70] was adopted.

Analysis of HGT fragments in Gmm genome via PCR and sequencing

Natural samples of Gmm used to examine HGT fragments originated from four populations collected in Zambia, Zimbabwe and Tanzania (Table 1). DNA was isolated from adult flies stored in EtOH using the Qiagen DNeasy kit (Qiagen, Valencia, CA) and stored at −20°C. The aposymbiotic (Wolbachia-free) Gmm line [41] was used as a control. For detection of Wolbachia, a PCR assay that amplified a 438 bp 16S rRNA fragment was used with the specific primer set wspF and wspFr [71]. For input DNA control, a 377 bp fragment of the mitochondrial 12S rRNA gene was amplified with the primer set 12SCFR and 12SCRR [72]. The PCR amplification protocol was 10 min at 95°C, 35 cycles of 30 sec at 95°C, 30 sec at 54°C and 1 min at 72°C, and 10 min at 72°C.

The identification of the Wolbachia strain infections was based on MLST (gbfB, coxA, hcpA, fhpA and fsiZ) and wsp-based genotyping approaches [45,69]. PCR reactions were performed using the following program: 5 min of denaturation at 95°C, followed by 35 cycles of 30 sec at 95°C, 30 sec at the appropriate temperature for each primer pair (52°C for fsiZ, 54°C for gbfB, 55°C for coxA, 56°C for hcpA, 58°C for fhpA and wsp) and 1 min at 72°C. All reactions were followed by a final extension step of 10 min at 72°C. Both strands of the products were sequenced using the respective primers. In addition, PCR products of 16S rRNA, wsp and MLST genes from the Gmm populations analyzed were cloned in pGEM-T Easy Vector System, and PCR products from several clones generated by the primers T7 and SP6 were sequenced in both directions using the BigDye Terminator v3.1 Cycle Sequencing Kit (PE Applied Biosystems). All Wolbachia gene sequences were manually edited with SeqManII by DNAStar and aligned using MUSCLE [63], as implemented in Geneious 5.4 [65], and adjusted manually.

Recovery of Wolbachia reads from RNA-seq data sets

To determine if genes from the chromosomal insertions were present in RNA-seq data sets, we utilized mapping of Illumina datasets from other studies, that included transcriptome reads from somatic tissues [73–75]. Reads were mapped to the chromosomal insertions using CLC Genomics Workbench (CLC Bio, Cambridge, MA) allowing no mismatches per read, a maximum of 10 hits per read and 80% of the gene must match at 95%. Predicted open reading frames (ORF) from the insertions were extracted and the following criteria were utilized to determine the possibility of expression: 1) at least 25 reads were recovered from the ORF and 2) those represented had coverage of over 85% of the ORF. This filtering approach excluded genes with a high number of mapped reads that were only present in small limited sections of the ORFs. These sections with high read numbers mapping but low coverage could be where sequence similarity between Gmm, Wigglesworthia or Sodalis is high enough to yield mapping to the chromosomal insertions.

### Table 1. Identity of analyzed G. m. morsitans samples.

| Code | Country (Area, Collection Date) | Number of analyzed individuals | cytWol present |
|------|---------------------------------|--------------------------------|-----------------|
| 12.3A | Zambia (MFWE, Eastern Zambia, 2007) | 43 | 41 |
| 30.9D | Zimbabwe (Rukomeshi, 2006) | 50 | 42 |
| 32.3D | Zimbabwe (Makuti, 2006) | 32 | 28 |
| 34.7G | Tanzania (Ruma, 2005) | 44 | 35 |

**Results**

**Cyttoplasmic wGmm genome features**

The draft genome of cyttoplasmic wGmm contains 201 contigs of 1,019,687 bp, comprised of 800 putative functional coding sequences (CDS) and 16 pseudogenes (Figure 1 and Table 2). The GC content of wGmm is 35.2%, in the range observed for the other sequenced Wolbachia genomes (Table 2). Although, the wGmm genome is not complete, based on comparison of the identified contigs, it is most similar to the two Wolbachia strains associated with Drosophila melanogaster and D. simulans, wMel and wRi, respectively (Table S2). It is more distantly related to the genomes of the Wolbachia strains associated with Culex pipiens and Brugia malayi, wPip and wBm, respectively (Table S2). The majority of the regions and genes missing from the wGmm genome relative to the wMel and wRi genomes encode phage, ankyrin and hypothetical proteins (Tables S3 and S4).

**Repetitive and mobile DNA.** One interesting feature of Wolbachia genomes is the presence of high numbers of genes encoding proteins containing ankyrin repeat domains (ANK), which are thought to play an important role in host-symbiont interactions, the establishment of symbiosis and the induction of reproductive phenotypes [76]. In comparison to the closely related wMel and wRi genomes, which contain 23 and 35 such genes respectively, the draft genome of wGmm has only 10 genes encoding proteins with one or more ANK repeat domains exhibiting the highest sequence identity with wMel, wRi, and wPip (Table 3).

An additional feature of the Wolbachia genomes is the presence of a high number of repeat sequences, IS elements and prophages. However, the draft wGmm genome contains a much reduced number of repeat elements, 1.2% compared to 8.9% in wMel and 22.1% in wRi, respectively (Table 2). This could be due to assembly issues in the draft wGmm assembly i.e. collapsed or unassembled repeats. The wGmm contains only 10 IS elements made up of the following families: IS3, IS5, and ISwPi6 (Table 2). Only 14 phage related genes (partial or putatively protein encoding genes) were discovered in the wGmm genome, a relatively small number when compared with wMel, wRi, and wPip. Phylogenetic analysis based on six concatenated phage genes suggested that the wGmm phage genes are more closely related to the wMel and wRi than the wPip corresponding phage (Figure S1).

**General comparison with other Wolbachia genomes.** Comparisons of wGmm, wMel, wRi, and wBm suggest that a high degree of rearrangement has occurred in the multiple genomes. There are many blocks of genes that share co-linearity with wRi, wMel and wBm. While several of the genomes have undergone extensive rearrangements, the co-linear blocks are most likely maintained due to their important biological functions and co-transcription. An example that has already been discussed.
in the literature [10,77] is the type IV secretion system (T4SS), for which the gene order function is also conserved in wGmm (Figure S2).

**Chromosomal Wolbachia features**

Both PCR-based evidence from Wolbachia infected tsetse flies, and analysis of the Gmm annotated genome data indicated the presence of Wolbachia gene fragments inserted in the host genome. We mined the final assembly of the Gmm host genome and were able to identify 261 contigs that carried chrWol DNA sequences. Based on nucleotide diversity, close examination of the 261 contigs indicated that these represented at least three different events, which we refer to as insertions A, B and C. Manual editing and implementation of the AMOS snps script enabled the separation of the contigs into different insertions, with insertions A and B being the largest in size. Figure 2 shows the mapping of these two insertions on the wGmm reference genome. The observed pattern suggests that at least two large Wolbachia genome segments of 527,507 and 484,123 bps have been integrated into the Gmm chromosomes indicating that at least 51.7% and 47.5% of the draft Wolbachia genome were transferred to the host nuclear genome. Sequence analysis of insertion A predicted 197 putative functional coding sequences, 148 pseudogenes, and 15 tRNAs. Remnants of 163 pseudogenes were discovered that are greater than 100 bp in size and that have either partially been integrated into the host genome, or only represent part of the pseudogene. For insertion B, sequencing analysis revealed the presence of 159 putative functional coding sequences, 148 pseudogenes and 13 tRNAs. In insertion B, 157 remnants of pseudogenes were also identified. Thus, on average more than 60% of the genes transferred to the tsetse nuclear genome have been pseudogenized.

The average length of the putative functional coding sequences is slightly smaller than wMel, wRi and the cytoplasmic wGmm at 690 bp for insertion A and 677 bp for insertion B (Table S4). The GC% content for insertion A and B is 35.1%. Comparison between the chromosomal insertions A and B and the cytoplasmic wGmm draft genome using Blastn and lastz indicated that: (a) the two insertions are very similar to each other (Figure S4) and (b) at least four genes, three hypothetical proteins and hemK are present in the chromosomal insertions but not in the cytoplasmic Wolbachia genome. The sequence identity between chromosomal and cytoplasmic genes and phylogenetic analysis based on ten concatenated genes clearly suggests that the chromosomal insertions A and B are closely related to the cytoplasmic wGmm genome (Table 5 and Figure S3). In more detail, comparison of the sequence identity in eleven chromosomal genes indicates that the

Figure 1. Circular map of the cytoplasmic wGmm genome. The outermost circle represents the scale in Kbp. Contigs of the draft genome are presented as boxes and they have been randomly ordered into a single circle for presentation purposes. In the second and fourth circle CDS in the two strands are presented. In the third and fifth circle the position of the tRNAs are presented. In the sixth and seventh circle CDSs identified for the wGmm genome are colored according to the Clusters of Orthologous Groups (COG) categories and represented as lines and boxes. doi:10.1371/journal.pntd.0002728.g001
majority of them exhibit a high sequence identity with the wGmm sequences (Table 5). The third Wolbachia HGT segment, insertion C, is only 2,089 bp in size and sequence analysis predicted the presence of only six pseudogenes.

A number of different types of mutations were identified in insertions A and B present in the host nuclear genome, and these shed light on the pseudogenization process. Our analysis suggests that more than 80% of the mutations that accumulated in the putative functional coding sequences represent single nucleotide polymorphisms (SNPs) (Figure 3). The majority of the genes that have been pseudogenized accumulated mutations that consist of nucleotide polymorphisms with deletions (NPD) and NPs. In both insertions, genes that have been pseudogenized contain mutations that combine NPs and deletions (NPDs) are more than those pseudogenized by NPs (Figure 3). In addition, we identified two additional types of mutations, NPs with insertions and NPs with deletions, associated with both chromosomal Wolbachia insertions but to a much lesser degree. A list of partial and full genes corresponding to the chrWol insertions is available in Tables S6, S7 and S8.

Expression of chromosomal sequences
Based on our results, there were very few ORFs that met our criteria for expression from chromosomal insertions. In general, there were multiple ORFs that had high number of mapped reads (>100), but in nearly all cases the coverage of the mapping was below 30% indicating that these may represent reads from another symbiont or tsetse transcripts. Results were similar for the three transcriptomes analyzed from heads, salivary glands and the bacteriome. However, three putative ORFs satisfied our criteria: \textit{serB}, \textit{ccmB} and a degenerate transposase located at both insertions (102636-102894 for insertion A and 97255-97523 for insertion B).

| Locus | Annotation | Size (bp) | Contig number | Present (highest homology first) |
|-------|------------|----------|---------------|---------------------------------|
| wGmm 0356 | ankyrin repeat-containing prophage LambdaW1 | 1968 | 42 | wPip, wMel, wRi |
| wGmm 0372 | ankyrin repeat-containing protein | 1245 | 44 | wMel, wRi |
| wGmm 0536 | ankyrin repeat containing protein | 867 | 72 | wMel, wRi |
| wGmm 0561 | ankyrin repeat-containing protein | 255 | 76 | wRi, wMel |
| wGmm 0573 | ankyrin repeat-containing protein | 2082 | 78 | wRi, wMel |
| wGmm 0691 | ankyrin repeat-containing protein | 807 | 99 | wMel, wRi |
| wGmm 0870 | ankyrin repeat-containing protein | 1191 | 139 | wMel, wRi, wPip |
| wGmm 0967 | ankyrin repeat containing protein | 462 | 164 | wPip, wMel, wRi |
| wGmm 0968 | ankyrin repeat-containing prophage LambdaW1 | 750 | 164 | wPip, wRi, wMel |
| wGmm 1071 | ankyrin repeat-containing protein | 1113 | 176 | wRi, wMel |

Table 3. Ankyrin-domain containing proteins encoded by the wGmm genome.

Table 2. wGmm genome features and comparisons with other sequenced Wolbachia genomes.
These analyses suggest that most of the genes present in the chromosomal insertions are likely not expressed, but the few specific genes we identified may have low levels of expression. Further studies will be necessary to validate their expression.

**Southern blot analysis**

Hybridization of the wsp probe to Gmm female DNA restricted with the HindIII enzyme produced five bands of about 1200, 1600, 2150, 2600 and 2700 bp (Figure 4, lanes 1 and 3). DNA from tetracycline-treated females (cytWol-free) had a similar profile, except that the 2700 bp band, corresponding to the expected cytWol wsp fragment, was absent (lane 2). Untreated male DNA displayed an additional band of 1500 bp, indicating the presence of insertions on the Y chromosome (Lane 4). This banding pattern suggests the presence of at least five independent wsp chromosomal insertions, including one on the Y chromosome, supporting the in silico analyses.

### Table 4. IS-elements identified in the wGmm genome.

| Locus | Family   | contig |
|-------|----------|--------|
| wGmm 0025 | IS5 | 002 |
| wGmm 0026 | IS5 | 002 |
| wGmm 0027 | IS5 | 002 |
| wGmm 0028 | IS5 | 002 |
| wGmm 0191 | IS3 | 014 |
| wGmm 0310 | IS5 | 031 |
| wGmm 0628 | ISWPi6 | 082 |
| wGmm 1201 | Transposase Tn5 related | 0187 |
| wGmm 1200 | Transposase Tn5 related | 0187 |
| wGmm 0961 | Putative uncharacterized | 0163 |

doi:10.1371/journal.pntd.0002728.t004

---

Figure 2. Circular map of the wGmm genome (red) and the chromosomal insertions in Gmm genome (orange and light yellow). The outermost circle represents the scale in Kbp. Contigs comprising the two identified chromosomal insertions are presented as boxes while the position of the third insertion relatively to the wGmm contigs is presented as blue boxes just below the outmost circle. In the second circle CDSs identified for the wGmm genome are colored according to the Clusters of Orthologous Groups (COG) categories and represented as lines and boxes. Genes identified in the insertions (orange and light yellow) of wGmm in the tse-tse fly genome are represented as lines and boxes. Pseudogenes are presented in red while in green coding DNA. Circles three presents the position of ankyrins (blue), prophages (yellow) and transposons (orange), in wGmm, and the chromosomal insertions. Finally, the wGmm genome and the two insertions in the tse-tse fly genome are arranged around the circle, with bands connecting regions of homology. Blue ribbons are composed of synteny regions, identified using Mauve and Mummer 3.0, between wGmm genome and the first set of identified insertions in the tse-tse genome. Orange ribbons are composed of synteny regions, identified using Mauve and Mummer 3.0, between wGmm genome and the second set of identified insertions in the tse-tse genome.

doi:10.1371/journal.pntd.0002728.g002
**Table 5.** Sequence identity between chromosomal and cytoplasmic genes.

| Insertion set A (coding genes) | wGmm | wRi | wMel | wPip |
|-------------------------------|------|-----|------|------|
| Type IV secretion protein virB10 | 98.3 | 87.1 | 95.9 | 78.1 |
| Type IV secretion protein virB11 | 99.2 | 96.6 | 98.2 | 88.0 |
| prophage LambdaW1 | 98.5 | 80.7 | 87.5 | 91.4 |
| ankyrin repeat-containing gene | 97.9 | 67 | 80.7 | 67.8 |
| prophage LambdaW5 site-specific recombinase resolvase | 99.5 | 84.9 | 87.5 | 92.5 |
| Insertion set B (coding genes) | wGmm | wRi | wMel | wPip |
| Type IV secretion protein virB8 | 99.1 | 93.2 | 98.7 | 88.4 |
| Type IV secretion protein virB10 | 99.1 | 83.5 | 98.5 | 77.1 |
| Type IV secretion protein virB11 | 99.7 | 98.3 | 98 | 87.7 |
| Type IV secretion protein virB6 | 99 | 98.7 | 95.8 | 90 |
| ankyrin repeat-containing prophage LambdaW1 | 99.6 | 88.1 | 88.5 | 90.5 |
| prophage LambdaW5 | 99.7 | 81.2 | 81.2 | 97.9 |

doi:10.1371/journal.pntd.0002728.t005

**chrWol insertions as determined by FISH**

To determine the location of *Wolbachia* insertions on *Gmm* chromosomes, we performed FISH analyses on mitotic spreads using *wsp*, *16S rRNA* and *fbpA* specific probes. The *Gmm* mitotic complement, comprising the supernumerary dispensable chromosomes (B chr) [78] is depicted in Figure 5, where the AT-rich heterochromatic nature of Y and B chromosomes is indicated by the strong DAPI-staining. The two autosomes, L1 and L2, as well as the X chromosome, appear to contain heterochromatic regions on both sides of the centromere. FISH results indicate that the *Wolbachia* genes *16S*, *fbpA*, and *wsp* consistently display a biased location on the distal part of the X, Y and B chromosomal arm. Although tyramide labeling generates strong and site-specific signals, it is difficult to detect the presence of multiple insertions on one chromosome if these events are localized in close proximity. The *16S* rRNA signal detected on the short arm of the X chromosome appears to be particularly strong and diffused, and may thus represent more than one insertion event in that region.
HGT events in natural populations of *Gmm*

Our previous characterization of the laboratory *Gmm* strain by *Wolbachia*-specific 16S rRNA-based PCR screening, the *wsp*-based and the MLST typing system revealed several HGT events [45]. Our results presented above indicate that these transfer events are in fact more extensive than previously considered. We next investigated the presence of HGT events in natural populations of *Gmm* originating from Zambia, Tanzania and Zimbabwe. We detected the pseudogenized fragment of the 16S rRNA gene carrying a deletion of 142 bp (Figure 6), similar to that we described in *Gmm* colony DNA prepared from the tetracycline-treated (cytWol-free) samples [41,45]. We observed a similar phenomenon for *fbpA*, where a pseudogenized gene fragment could be amplified containing two deletions of 47 and 9 bp from the same four natural populations, as well as from the cytWol-free

![Figure 4. Southern blot analysis showing wsp hybridizing fragments. DNA cleaved with HindIII restriction enzyme from normal females (lanes 1 and 3), tetracycline-treated females (lane 2) and normal males (lane 4) are shown. The sizes of the hybridizing bands are shown. The ~2700 bp band (indicated by arrows) may represent the cytoplasmic wsp containing fragment (lanes 1 and 3) which is absent in the tetracycline treated females (lane 2). The ~1500 bp band present only in the male (lane 4) is indicated by an arrow. doi:10.1371/journal.pntd.0002728.g004](image)

![Figure 5. Fluorescent in situ hybridisation (FISH) on Gmm mitotic chromosomes. The chromosomes are numbered as described by Willhoeft (1997) [70]. A–B. Banding pattern of (DAPI)-stained chromosome spreads (A–B). The DAPI positive regions indicate the heterochromatic patterns. B-chromosomes vary in number. FISH on female and male chromosomes with *fbpA* probe (C–D), 16S rDNA probe (E–F), and wsp probe on chromosomes from a male individual (G). doi:10.1371/journal.pntd.0002728.g005](image)
prevalence of Wolbachia varied in the different populations and was not fixed (Table 1).

**Discussion**

Here we report on a newly sequenced genome of cytoplasmic Wolbachia strain associated with the tsetse fly *G. m. morsitans*. Previous studies have shown that *w*Gmm belongs to Wolbachia supergroup A [79] and functional investigations have demonstrated that this Wolbachia strain can induce strong CI in the *Gmm* laboratory line [41]. Our comparative analysis confirms that *w*Gmm belongs to Wolbachia supergroup A, and is most similar to *w*Mel, based on the extensive synteny between their genomes. We also show evidence for extensive chromosomal insertions of *w*Gmm in the host genome: with at least two large insertions of 527,507 and 484,123 bp identified from WGS data. Southern blot hybridizations confirmed the presence of Wolbachia insertions in the *Gmm* genome, and FISH revealed their biased location on the two sex chromosomes (X and Y), as well as on the supernumerary B-chromosomes.

The genome sequence of the cytoplasmic *w*Gmm strain, when compared to Wolbachia genomes from other ecdysozoans, revealed the following striking features: (a) genome size comparable to that of the *w*Bm infecting the filarial nematode *B. malayi*, (b) genome size significantly smaller from all other insect Wolbachia strains and particularly from the *w*Pip infecting the mosquito *Caélex pipiens*, (c) reduced number of repetitive sequences including ISs, mobile II introns and phages, and (d) absence of functional phage copies. It is worth noting that the genome reduction has not affected the stable symbiotic association, including the expression of strong CI phenomena, as has been documented *in vitro* [41,80].

Previous research has demonstrated that Wolbachia genomes undergo frequent rearrangements and rapid evolution due to the high number of transposable elements and repeat regions, which can provide sites for homologous recombination [18,81,82]. The rearrangements in Wolbachia may have arisen from the introduction and 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 [27,82,83].

Phylogenetic analysis suggests that the phage of cytWol (*w*Gmm-WO) and the phage regions present on the two main chromosomal insertions are closely related, implying that the chromosomal phage sequences most likely originated from the cytoplasmic Wolbachia phage. However, it appears that the *w*Gmm phage copies are more closely related to the *w*Mel and *w*Ri than the *w*Pip phages. Given that the Wolbachia prophages can laterally transfer between Wolbachia strains shaping the bacterial genome evolution [84–88], the origin of the *w*Gmm phage copies remains an open question.

Of particular interest for host-symbiont interactions are the number of genes that encode proteins that contain ankyrin repeat domains. The ankyrin repeat domain (ANK)-containing proteins, tandem motifs of around 33 amino acids that are involved in protein-protein interactions, are mainly found in eukaryotes and viruses [89]. In eukaryotes, ANK proteins are known to participate in diverse pathways affecting the structure and function of cells regulating host cell cycle or cell division or interacting with the host cytoskeleton [89–91]. In addition, they have been shown to act as T4SS effectors participating in host-pathogen interactions [92]. For example, in the intracellular pathogen *Anaplasma phagocytophilum*, AnkA, which is secreted through T4SS, interacts with the host chromatin and regulates gene transcription, while in *Legionella pneumophila*, the AnkX protein prevents microtubule-dependent endocytic maturation of pathogen-occupied vacuoles [92]. While ANK proteins have been reported from bacteria, they are usually present in only a few copies per species [93]. *w*Gmm

---

**Figure 6.** 16S rRNA, fbpA and wsp gene fragments from tsetse Wolbachia chromosomal insertions sequenced from *Gmm* laboratory and natural populations aligned with the corresponding regions of *w*Gmm. Black lines represent the deleted region. The numbers show the positions before and after the deletions in respect to the *w*Mel genome and the right-left white arrows below the number indicate the size of deletion in base pairs.

doi:10.1371/journal.pntd.0002728.g006
has 10 putative ANK proteins, comparable to the number reported for other insect Wolbachia strains (23 in uMel, 35 in uRi and 60 in uPip). ANK proteins have been considered to play an important role in host-Wolbachia interactions, including the establishment of symbiosis. However, their role in the induction of reproductive abnormalities such as CI has not been confirmed as yet [76,94,95].

Several studies clearly suggest that the occurrence of HGT events in host-Wolbachia symbiotic associations is more widespread than previously thought [18–24]. Our results provide evidence of extensive HGT events between Wolbachia and tsetse genome, and further advance our knowledge on HGT during their co-evolution. From in situ hybridization results, it appears that at least three Wolbachia genes, 16S rRNA, fhtA, and wsp are located on X, Y and multiple B supernumerary chromosomes. Under the canonical model of sex chromosome evolution, X and Y are believed to have originated from an autosome pair via a three-step process beginning with the acquisition of one or more sex-determining genes [96–99]. X and Y are thought to have diverged due to sexually antagonist selection [100,101]. The suppression of recombination between the two sex chromosomes would be favored by chromosomal inversions and other genetic changes [102–105]. As the X became progressively haploid in males (hemizygous), selection may have favored increased transcription of X-linked genes in males through dosage compensation mechanisms [106,107]. In the later stages, lack of recombination between X and Y allowed for genetic degeneration of the Y, which is usually heterochromatic, accumulating large amounts of repetitive DNA [104,108,109]. Due to the highly repetitive nature, the accumulation of Wolbachia sequences may not be deleterious for Y functionality, and thus the inserted sequences are not eliminated. The presence of Wolbachia HGT events on the B chromosomes may reflect the common evolutionary origin of B and Y chromosomes. Indeed, in Glossina species homology between the supernumerary and sex chromosomes has been reported, suggesting the formation of B via Y chromosome duplication and subsequent accumulation of repetitive DNA sequences [110]. However, Carvalho and colleagues (2009) [97] do not exclude the alternative evolutionary scenario of Y originating from B.

The localization of the Wolbachia inserts in heterochromatin regions might protect them against the negative selection that would otherwise arise if they were inserted into functional genes, as occurs for transposable elements [111]. However, the heterochromatic location of the inserts may not necessarily imply loss of function, especially for those that are inserted in the facultative heterochromain. It has been suggested for other insects [23] that Wolbachia genes transferred to host chromosomes are structurally disrupted, and functionally impaired via pseudogenization. Through the acquisition of point mutations, insertions and/or deletions, these inserts may be destined to become junk DNA in the insect genome [18]. It has been reported that some horizontally transferred genes can be transcribed in the insect hosts. In the case of the pea aphid Acrithosiphon pisum [112], and the mosquito Aeles aegypt, the transferred genes have been found to be transcriptionally active in the salivary glands [18,19]. In the tripartite mealybug symbiosis, at least twenty-two highly expressed genes have been identified from multiple diverse bacteria [113]. In addition, almost 2% of the Wolbachia genes that were transferred to the second chromosome of D. ananassae are transcribed [22]. In the case of the nematode Onchocerca flexuosa, which does not carry a cytoplasmic Wolbachia infection, Wolbachia-like DNA sequences have been identified in the nuclear genome [114]. Despite the fact that several of these sequences are degenerate, many are expressed at both the RNA and protein levels [115]. The only case of Wolbachia genes transferred to the X chromosome has been reported in the adzuki bean beetle C. chennosi, where the insertion was presumably transcriptionally inactive [23,24]. The present study showed that only a few specific genes may be expressed at low levels from chrWol, however, further studies are required to confirm potential expression of these and other genes in a temporal and spatial manner. Given the biological interdependence between insect hosts and bacterial symbionts, transfer of symbiont genes of functional (possibly regulatory) relevance may be beneficial for the host. Thus, it is of importance to clarify the potential functional role(s) these inserted sequences may play on host Gmm physiology. In addition, whether Wolbachia fragments in the Glossina genome may be on an evolutionary trajectory of degradation and loss [18] needs to be verified, especially given the large size of the inserts we detected, which may indicate a relatively recent origin for these events.

The origin of horizontal transfer of Wolbachia genes in Gmm is of evolutionary significance. The phylogenetic analysis presented in Figure S3 shows a long branch from uGmm and short distance between insertion A and insertion B, which strongly support a single transfer event. Also, the genetic distance between several genes present in the cytWol and their homologues in the chrWol insertions is minimal, thus making it difficult to assess the history of the insertion events. While speculative, it is most likely that the common ancestor for the two chromosomal insertions we detect is the uGmm cytoplasmic strain (Table 5).

It is thought that Wolbachia induced CI can promote reproductive isolation in host insects that can potentially lead to speciation [1,116]. While the genetic mechanism and specific genes involved in CI are currently unknown, if genes involved in CI integrated into the host chromosome were functional, this could result in reproductive isolation and speciation. Unpredictable rates of CI expression could complicate Wolbachia-based strategies for tsetse control, if genes involved in the CI mechanism are expressed from chromosomal loci. The results presented here could be used as part of future research to test this hypothesis in tsetse, once the molecular mechanism behind CI has been further defined.

Our analysis with Gmm individuals from natural populations indicates the presence of the chromosomal insertions in the field populations as well. Interestingly not all individuals in the field carried the cytoplasmic infections, despite the presence of chromosomal insertions. We can speculate that maternal transmission of Wolbachia may be less than perfect in the field, resulting in individuals with no infections. In addition, Wolbachia densities have been shown to vary as a function of host age [117,118], but the field samples could not be scored for relative age. Alternatively, recent studies have identified low-density infections in several tsetse flies including subspecies of G. morsitans [45,116,119], which could not be detected using the PCR conditions that were employed in this study. Studies that determine infection prevalence or infection densities in natural populations could be compromised if chromosomal sequences are mistaken for cytoplasmic infections. The results raise the question of whether HGT events as shown here are common in other species of tsetse flies, and ongoing WGS of other tsetse species will provide important insights. Future work should focus on determining the prevalence and ancestry of the chromosomal insertions in tsetse.

List of genes, locus tags and GI numbers

uMel genome (AE017196), uRi genome (CP001391), uPip genome (AM999887), uBn genome (AE017321), DNA-directed RNA polymerase (WD_0024/GI: 4209679, WR_000230/GI: 225591874, WP0554/GI:190357240, Whm0647/GI:58419220),
DNA polymerase III (alpha subunit) (WD_0780/GI:42410358, WRi_006220/GI:225592574, WP0650/GI:190357396, Wbm0499/GI:58419072), DNA gyrase B (WD_0112/GI:42409735, WRi_001420/GI:225591969, WP1103/GI:190357759, Wbm0364/GI:42409671), translation elongation factor G (WD_0016/GI:42409671, WRi_000410/GI:225591866, WP0362/GI:190357248, Wbm0944/GI:58418919), aspartyl-tRNA synthetase (GI:190357396, WP0637/GI:190357763), CTP synthase (WD_0413/GI:42410026, WRi_002800/GI:225592127, WP0387/GI:190357090, Wbm0012/GI:58418598), 0468/GI:190357644), phage SPO1 DNA polymerase-related protein (WD_1148/GI:42410690, WRi_000140/GI:225591866, WP0562/GI:190357336, Wbm0499/GI:58418745), fructose-bisphosphate aldolase (WD_1238/GI:190357884, Wbm0169/GI:58418745), glutamyl-tRNA(Glu) amidotransferase B (WD_0723/GI:42410305, WRi_002750/GI:225592077, WP0387/GI:190357759), GTP-binding protein (WD_0198/GI:42410645, WRi_001270/GI:225592935, WP0891/GI:190357560, Wbm0032/GI:58418609), cell division protein FtsZ (WD_0723/GI:42410305, WRi_007520/GI:225592482, WP0377/GI:190357261, Wbm0602/GI:58419175), fructose-bisphosphate aldolase (WD_1238/GI:42410776, WRi_001270/GI:225592879, WP1061/GI:190357738, Wbm0097/GI:58418674), HK97 family phage major capsid protein (WD_0458/GI:42410067, WRi_002750/GI:225592127, WP0102/GI:190356840), phage integrase family site-specific recombinase (WD_1148/GI:42410690, WRi_009900/GI:225592678, WP0980/GI:190357644), phage SP01 DNA polymerase-related protein (WD_0164/GI:42409803, WRi_000900/GI:225591926, WP0922/GI:190357589), prophage LambdaW5 baseplate assembly protein W (WD_0640/GI:42410229, WRi_005480/GI:225592306, WP0303/GI:190357018), prophage LambdaW1, baseplate assembly protein J (WD_0639/GI:42410228, WRi_010130/GI:225592699, WP0302/GI:190357017) and a prophage LambdaW1 site-specific recombinase resolvase protein (WD_0634/GI:42410223, WRi_005400/GI:225592900, WP0342/GI:190357056).

Supporting Information

Figure S1 Maximum Likelihood phylogeny based on phage concatenated genes (5,912 bp). The topology resulting from the Neighbor-Joining method was identical. Strains are characterized by the names of their host species. ML bootstrap values based on 1000 replicates are given.

Figure S2 Circular map of the Type IV genes present in wGmm, wMel, wRi, and wPip. The outermost circle represents the scale in Kbp. In the second circle Type IV genes are color-coded based on their homology. Regions of homology are connected with bands. Blue ribbons are composed of synteny regions identified using MAUVE and Mummer 3.0 between wGmm and insertion A, B and C insertion A and insertion B. For plots (A) and (B) the size of the wGmm genome is given in x-axis, while for plot (C) the size of insertion A. High-scoring segment pairs are presented as blue dots while low-scoring segment pairs are presented as red dots.

Table S1 Assembly statistics after each major stage of the assembly process of wGmm draft genome.

Table S2 Number of unique genes present in wGmm compared with the genomes of wMel, wRi, wPip and wBm.

Table S3 Missing regions and genes from the wGmm genome in respect to wMel. Regions have been identified after alignment of the two genomes with MAUVE using the default settings of the program. Gaps in the genomes were identified using Geneious.

Table S4 Missing regions and genes from the wGmm genome in respect to wRi. Alignment of the two genomes was performed with MAUVE using the default settings of the program. Gaps in the genomes were identified using Geneious v. 5.4.

Table S5 Features and comparisons of the wGmm genome and chromosomal insertions with other sequenced Wolbachia genomes. Alignment of the two genomes was performed with MAUVE using the default settings of the program. Gaps in the genomes were identified using Geneious v. 5.4.

Table S6 Description of the first set of Wolbachia inserted regions into the G. m. morsitans chromosomes.

Table S7 Description of the second set of Wolbachia inserted regions into the G. m. morsitans chromosomes.

Table S8 Description of the third set of Wolbachia inserted regions into the G. m. morsitans chromosomes.

Author Contributions

Conceived and designed the experiments: KB SA. Performed the experiments: CB GT MF VD MS ARM KB SA. Analyzed the data: CB GT MF VD MS ARM KB SA. Contributed reagents/materials/analysis tools: CB GT MF VD MS ARM KB SA. Wrote the paper: CB GT ARM KB SA.

References

1. Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate biology. Nat Rev Microbiol 6: 741–751.
2. Zug R, Hammerstein P (2012) Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. PloS One 7: e30344.
3. Saridaki A, Bourtsis K (2010) Wolbachia: more than just a bug in insects genitals. Curr Opin Microbiol 13: 67–72.
4. Hoffmann AA, Hercus M, Dagher H (1998) Population Dynamics of the Wolbachia infection causing cytoplasmic incompatibility in Drosophila melanogaster. Genetics 148: 221–231.
22. Hotopp JC, Clark ME, Oliveira DCSG, Foster JM, Fischer P, et al. (2008) The Wolbachia genome of *Bactrocera* malayi: endosymbiont evolution within a human pathogenic nematode. PLoS Biol 5: e121.

23. Serbus LR, Sullivan W (2007) A cellular basis for horizontal gene transfer in *Wolbachia* pipientis. Mol Microbiol 63: 454–457.

24. Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T (2002) Genome fragment atlas of human African trypanosomiasis. Int J Health Geogr 8: 15.

25. Hoffmann AA, Montgomery BL, Popovici J, Iturbe-Ormaetxe I, Johnson PH, et al. (2007) Wolbachia endosymbionts in the host African tsetse fly. Proc Biol Sci 274: 3791–3798.

26. Anderson JO (2005) Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62: 1679–1694.

27. Doolittle WF (1998) You are what you eat: a gene transfer ratchet could explain the long-term evolutionary dynamics of bacterial insertion sequences: insights from *Wolbachia* endosymbionts. Genome Biol and Evol 5: 3173–3186.

28. Andersson JO (2009) Longicorn beetle that vectors pinewood nematode carries many transferred endosymbiont genes. Genome Res 18: 272–280.

29. Vreysen MJ SM, Sall B, Bouyer J (2013) Phylogenomics and analysis of shared genes suggest a single transition to parasitism and cytoplasmic incompatibility on host population size in natural and manipulated *Wolbachia* populations. Nat Commun 4: 1575.

30. Serbus LR, Sullivan W (2007) A cellular basis for horizontal gene transfer in *Wolbachia pipientis*: a McL: a streamlined genome overrun by mobile genetic elements. PLoS Biol 5: e59.

31. Kurland C (2000) Something for everyone. Horizontal gene transfer in *Wolbachia* and the bacterial species concept. PLoS Genetics 9: e1003381.

32. Klassen L, Klassen L, Nikshchik B, Oura K, Anderson S (2013) Comparative genomics of *Wolbachia* and the bacterial species concept. PLoS Genetics 9: e1003381.

33. Simarro PP, Diarra A, Ruiz Postigo J, Franco J, Jannin J (2011) The human African trypanosomiasis control and surveillance programme of the World Health Organization. PLoS Negl Trop Dis 5: e1000619.

34. Anderson JO (2005) Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62: 1679–1694.

35. Farkou O, Najoska F, Mbida Mbida JA, Njichooungu GR, Djengua HN, et al. (2010) Tripartite interactions between tsetse flies, *Sodalis glossinaridis* and trypanosomes—an epidemiological approach in two historical human African trypanosomiasis foci in Cameroon. Infect Genet Evol 10: 115–121.

36. Weiss B, Aksoy S (2011) Microbiome influence on insect host vector competence. Trends Parasitol 27: 514–522.

37. Weiss B, Wang J, Aksoy S (2011) Tsetse immune system maturation requires the presence of obligate symbionts in larvae. PLoS Biol 9: e1000619.

38. Weiss B, Maltz MA, Aksoy S (2012) Obligate symbionts activate immune system development in the *Anopheles* mosquito. J Immunol 188: 3395–3402.

39. Maltz MA, Weiss BL, O'Neill SL, Mason P, Mendelo L, Woese CR, et al. (1993) Genetic transformation and phylogeny of bacterial symbionts from tsetse. Insect Mol Biol 1: 123–131.

40. Aikawa T, Anbutsu H, Nikoh N, Kikuchi T, Shibata F, et al. (2009) Longicorn beetle that vectors pinewood nematode carries many transferred endosymbiont genes. Genome Res 18: 272–280.

41. Alam U, Medlock J, Brelsfoard CL, Pais R, Lohs C, et al. (2011) Tsetse fly genome structure of the *Brugia* melanoi ds strain infecting *Melophagus ovinus*. Nat Genet 43: 757–760.

42. Klassen L, Klassen L, Nikshchik B, Oura K, Anderson S (2013) Comparative genomics of *Wolbachia* and the bacterial species concept. PLoS Genetics 9: e1003381.

43. Beard CB, O'Neill SL, Mason P, Mendelo L, Woese CR, et al. (1993) Genetic transformation and phylogeny of bacterial symbionts from tsetse. Insect Mol Biol 1: 123–131.

44. Maltz MA, Weiss BL, O'Neill SL, Mason P, Mendelo L, Woese CR, et al. (1993) Genetic transformation and phylogeny of bacterial symbionts from tsetse. Insect Mol Biol 1: 123–131.

45. Maltz MA, Weiss BL, O'Neill SL, Mason P, Mendelo L, Woese CR, et al. (1993) Genetic transformation and phylogeny of bacterial symbionts from tsetse. Insect Mol Biol 1: 123–131.

46. Zabalou S, Riegler M, Theodorakopoulou M, Stauffer C, Savakis C, et al. (2008) The comparison of bacterial symbiont genomes of nematodes. Genome Biol Evol 5: 1668–1674.

47. Vreysen MJ SM, Sall B, Bouyer J (2013) Phylogenomics and analysis of shared genes suggest a single transition to parasitism and cytoplasmic incompatibility on host population size in natural and manipulated *Wolbachia* populations. Nat Commun 4: 1575.

48. Zabalou S, Riegler M, Theodorakopoulou M, Stauffer C, Savakis C, et al. (2008) The comparison of bacterial symbiont genomes of nematodes. Genome Biol Evol 5: 1668–1674.

49. Brelsfoard CL, Dobson SL (2009) *Bactrocera* malayi: endosymbionts. Genome Biol and Evol 3: 1175–1186.

50. Chevreux B, Pfisterer T, Drescher B, Driesel A, Muler W, et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14: 1147–1159.

51. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M, et al. (2012) A post-assembly genome-improvement toolkit (PAGITT) to obtain annotated genomes from contigs. Nat Protoc 7: 1260–1284.

52. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) *ABACAS*: algorithm-based automatic configuration of assembled sequences. Bioinformatics 25: 1968–1969.

53. Swain MT, Tsai T, Assefa SA, Newbold C, Berriman M, et al. (2012) A post-assembly genome-improvement toolkit (PAGITT) to obtain annotated genomes from contigs. Nat Protoc 7: 1260–1284.

54. Swain MT, Tsai T, Assefa SA, Newbold C, Berriman M, et al. (2012) A post-assembly genome-improvement toolkit (PAGITT) to obtain annotated genomes from contigs. Nat Protoc 7: 1260–1284.

55. Swain MT, Tsai T, Assefa SA, Newbold C, Berriman M, et al. (2012) A post-assembly genome-improvement toolkit (PAGITT) to obtain annotated genomes from contigs. Nat Protoc 7: 1260–1284.

56. Swain MT, Assefa SA, Newbold C, Berriman M, et al. (2012) A post-assembly genome-improvement toolkit (PAGITT) to obtain annotated genomes from contigs. Nat Protoc 7: 1260–1284.

57. Delcher AL, Harmon D, Kaulf S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4641–4644.

58. Schattner P, Brooks A, Lowe T (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs: Nucleic Acids Res 33: W696–698.

59. Rutherford K, Parkhill J, Crook J, Hormann T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.

60. Karpf P, Karpf P, Karpf P, Karpf P, Karpf P, et al. (2000) *Tetrahymena* genome sequence: the assembly. Science 287: 203–214.

61. Ouyang Z, Zhu H, Wang J, She Z-S (2004) Multivariate entropy distance mapping and assembly of short reads to eliminate gaps. Genome Biol 11: R41.

62. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for multiple sequence alignment. J Comput Biol 7: 203–214.

63. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.

64. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through the incorporation of position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
66. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA6: molecular evolutionary genetic analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.

67. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.

68. Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98: 503–517.

69. Balmand S, Lohs C, Aksoy S, Heddi A (2013) Tissue distribution and expression, and motif variability of ankyrin domain genes in arthropods. J Mol Evol 51: 491–497.

70. Willhoth U (1997) Fluorescence in situ hybridization of ribosomal DNA to mitotic chromosomes of tsetse flies (Diptera: Glossinae: Glossina). Chromosoma 95: 262–267.

71. Werren JH, Windsor DM (2000) Wolbachia infection frequencies in insects: evidence of a global equilibrium? Proc Biol Sci 267: 1277–1283.

72. Hanner R, Fugate M (1997) Branchiocephid phylogenetic reconstruction from 12S DNA sequence data. J Crustacean Biol 17: 174–183.

73. Benoît JB, Attardo G, Michalkova V, Krause TB, Bohova J, et al. (2014) A novel highly divergent protein family from a viviparous insect identified by RNA-seq analysis: a potential target for tsetse fly-selective arbuscularists. PLoS Genetics 10: Doi: pgen.1003874.

74. Telleria EL, Benoit JB, Zhao X, Savage AF, Regmi S, et al. (2013) Insights into Wolbachia transmission routes for the tsetse fly endosymbionts. J Invertebr Pathol 112: 101–109.

75. Benoit JB, Attardo G, Michalkova V, Krause TB, Bohova J, et al. (2013) A novel highly divergent protein family from a viviparous insect identified by RNA-seq analysis: a potential target for tsetse fly-selective arbuscularists. PLoS Genetics 10: Doi: pgen.1003874.

76. Iturbe-Ormaetxe I, Burke GR, Riegler M, O’Neill SL (2005) Distribution, prevalence of Wolbachia pipientis and expression of cytoplasmic incompatibility in field populations of Culex pipiens. J Bacteriol 187: 5136–5145.

77. Kløsvik L, Walker T, Sebähi M, Sanders MJ, Quail MA, et al. (2008) Genome evolution of Wolbachia strain aPip from the Culex pipiens group. Molecular Biology and Evolution 25: 1873–1880.

78. Southern DIaP, P.E. (1973) Chromosome relationships and meiotic mechanisms of certain mosquitoid groups tsetse flies and their hybrids. Chromosoma 44: 319–335.

79. Cheng Q, Ruel TD, Zhou W, Moloo SK, Majiwa P, et al. (2000) Tissue distribution and prevalence of Wolbachia infections in tsetse flies, Glossina spp. Med Vet Ent 14: 44–50.

80. Brownlie JC, O’Neill SL (2005) Wolbachia endosymbiont of Brugia malayi, the endosymbiont causing sexual alterations in species of Glossina (Tsetse): a two-step model for the origin of superparasitism. Trends Ecol Evol 20: 578–585.

81. Hryniwicz-Jankowska A, Czogalla A, Bok E, Sikorsk A (2002) Ankyrins and differentiation in recently evolved neo-sex chromosomes. Heredity 89: 578–585.

82. Johnson NA, Lachance J (2012) The genetics of sex chromosomes: evolution and implications for hybrid incompatibility. Ann N Y Acad Sci 1256: E1–22.

83. Carvalho AB KL, Clark AG (2009) Origin and evolution of Y chromosomes: Drosophila tales. Trends Genet 25: 270–277.

84. Rice W (1996) Evolution of the Y sex chromosome in animals. BioScience 46: 331–343.

85. Bachdort D (2006) A dynamic view of sex chromosome evolution. Genet Dev 16: 578–585.

86. Wu CI, NE. (2003) Sexual antagonism and X inactivation—the SAXI hypothesis. Trends Genet 19: 243–247.

87. Rice VR (1984) Sex chromosomes and the evolution of sexual dimorphism. Evolution 38: 733–742.

88. Bull J (1983) Evolution of sex determining mechanisms. Menlo Park (CA): Benjamin Cummings.

89. Charlesworth D, Charlesworth B (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity 95: 118–128.

90. Natri H, Shikano T, Merila J (2013) Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. Mol Biol Evol 30: 1111–1114.

91. Charlesworth B (1996) The evolution of chromosomal sex determination and dosage compensation. Curr Biol 6: 149–162.

92. Marín I, Siegal M, Baker B (2000) The evolution of dosage-compensation mechanisms. BioEssays 22: 1166–1170.

93. McNulty SN, Fischer K, Curtis KC, Weil GJ, Brattig NW, et al. (2013) Horizontal Gene Transfer Events in Tse-tse Flies. J Bacteriol 189: 4442–4448.

94. McNulty SN, Foster JM, Mitreve M, Dunnng Hotopp JC, Martin J, et al. (2010) Endosymbiont DNA in endobacteria-free filarial nematodes indicates recent horizontal genetic transfer. PLoS One 5: e11029.

95. McNulty SN, Fischer K, Curtis KC, Weil GJ, Brattig NW, et al. (2013) Localization of Wolbachia-like gene transcripts and peptides in adult Oroch碰ica reticulata worms indicates tissue specific expression. Parasit Vectors 6: 2.

96. Brucker RM, Bordenstein SR (2012) Speciation by symbiosis. Trends Ecol Evol 27: 443–451.

97. Duron O, Fort P, Weill M (2007) Influence of aging on cytoplasmic incompatibility, sperm modification and Wolbachia density in Culex pipiens mosquitos. Heredity 98: 367–374.

98. Kitaayaopong P, Mongkalagonsorn P, Baimai V, O’Neill SL (2002) Host age effect and expression of cytoplasmic incompatibility in field populations of Wolbachia-superinfected Anopheles albopictus. Heredity 88: 270–274.

99. Symula RE, Alam U, Breidbart C, Wu Y, Echodu R, et al. (2013) Wolbachia association with the tsetse fly, Glossina morsitans morsitans, reveals high levels of genetic diversity and complex evolutionary dynamics. BMC Evol Biol 13: 51.