Complete *de novo* assembly of *Wolbachia* endosymbiont of contemporary *Drosophila simulans* using long-read genome sequencing.

Running Title: *De novo* assembly of *Wolbachia* in *Drosophila simulans*

**Authors:**
Jodie Jacobs[a,b], Alexandra Lum[a], Elyse Mina[a], Camryn Morey[a,b],Darren D. Lee[a], Emry Gutierrez[a], Jonah Dionisio[a], Cade Mirchandani[a,b], Luke Sylvester[a], Anne Nakamoto[a,b], Hailey Loucks[a,b], Ciara Wanket[b,c], Ariana Cisneros[a,b], Alessandro Calicchio[a], Alexis N. Enstrom[a], Camille Headrick[a], Faith Okamoto[a,b], Harrison Heath[a,b], Kseniya Malukhina[a], Petria Russell[a], Sagorika Nag[a], Thomas Gillespie[a], William Sobolewski[a], Zia Truong[a], Shelbi L. Russell[#a,b]

[a]Biomolecular Engineering and Bioinformatics Department at the University of California, Santa Cruz
[b]Genomics Institute at the University of California, Santa Cruz
[c]Ecology and Evolutionary Biology Department at the University of California, Santa Cruz

[#]Corresponding Author: Shelbi Russell shelbilrussell@gmail.com

Authorship order was determined based on contributions to data generation and analysis. For authors with equal contributions, ties were resolved alphabetically by first name.

**Abstract**
We present a contemporary high-quality, complete *de novo* assembly of *Wolbachia pipientis* (*w*Ri Merrill 23, OZ411647), an alphaproteobacterial endosymbiont of *Drosophila simulans*. This assembly was generated using long read sequencing of *w*Ri-infected *D. simulans* embryos collected from the Merrill College at the University of California, Santa Cruz in October 2023.

*Wolbachia pipientis* infects diverse arthropods and nematodes, manipulating host phenotypes through cytoplasmic incompatibility (CI), male killing, and fertility rescue [1,2]. The Riverside strain (*w*Ri) was first identified in California *Drosophila simulans* in the 1980s [3] and rapidly spread statewide due to exceptionally strong CI [4]. Despite its significance [5], the only reference genome reflects *w*Ri present in 1984 [3,6]. Here, we present a complete *de novo* *w*Ri genome assembly from contemporary *D. simulans* collected at the UC Santa Cruz Alan Chadwick Garden, located in Merrill College, in October 2023.

To generate a contemporary *w*Ri genome assembly, we collected wild D. simulans flies, established isofemale lines, and sequenced wRi-infected embryos. We established isofemale lines by deploying banana-baited bottles for ~5 days and collecting gravid females onto white food medium. After offspring eclosed, species identity was confirmed by phenotyping males and by PCR using silf-F/R primers to distinguish *D. simulans* from *D. melanogaster [7]* and wsp_1F/592R primers to confirm *w*Ri identity [8]. We extracted DNA from *w*Ri-infected embryos using the Wizard HMW DNA Extraction Kit (Promega) without shearing and prepared libraries with the Native Barcoding Kit V14 (SQK-NBD114-24) without size selection. We sequenced these libraries on the Nanopore MinION Mk1B with a R10 flow cell (FLO-MIN-114) and MinKNOW v23.07.8 with adaptive sampling (fast model) to deplete *D. simulans* reads (GCF_016746395.2) yielding 4.26M reads, (N50 749bp, 20 hours) that were basecalled with Dorado (v0.7.3, hac model, --min-qscore 10). After filtering for host-free reads >3kb (93K reads remaining, N50 10.5kb), we assembled the *w*Ri genome using Flye [9] following Jacobs and Nakamoto *et al.* (2024) [10], yielding a 1.26 Mb circular assembly with 30x coverage. Flye determined circularization (4000bp minimum overlap), the genome was not rotated.

To polish the assembly, we generated Illumina short-read whole-genome sequencing data from whole *w*Ri-infected *D. simulans* flies (Merrill 23 stocks). Illumina libraries were prepared using a Tn5-based tagmentation protocol [11]. The Tn5 enzyme (Tn5R27S,E54K,L372P) was obtained from QB3 MacroLab (University of California, Berkeley). Tagmentation simultaneously fragmented genomic DNA and ligated adapter sequences using custom oligos (Tn5-A: FC-121-1030, Tn5-B: FC-121-1031, Tn5-rev; IDT). Libraries were indexed and amplified using Nextera XT Index Kit v2 (i5/i7 adapters, Illumina) with KAPA HiFi (Roche) and sequenced on a NovaSeqX Plus (2×150 bp). Reads were trimmed to remove sequencing adapters and low quality regions using fastp v1.0.1 [11] before polishing the assembly with Pilon [12] v1.24 (849K reads). We assessed the quality of the polished assembly with BUSCO [13] (v5.7.0, rickettsiales_odb10), achieving a 99.2% completeness, annotated the assembly with Prokka [14] (v1.1.1, kingdom:bacteria) (Table 1) and calculated and visualized GC content and GC skew with Proksee [15] v1.1.2 (Figure 1). Default parameters were used unless otherwise specified.

**Data Availability**

Raw sequencing reads, ERX15644803 (Illumina) and ERX15644805 (ONT) and the assembled genome (OZ411647) are available under BioProject accession number PRJEB107302.
Assembly pipeline:
https://github.com/jodiejacobs/Jacobs_et_al_2026_de_novo_wRi_merrill_23_assembly.

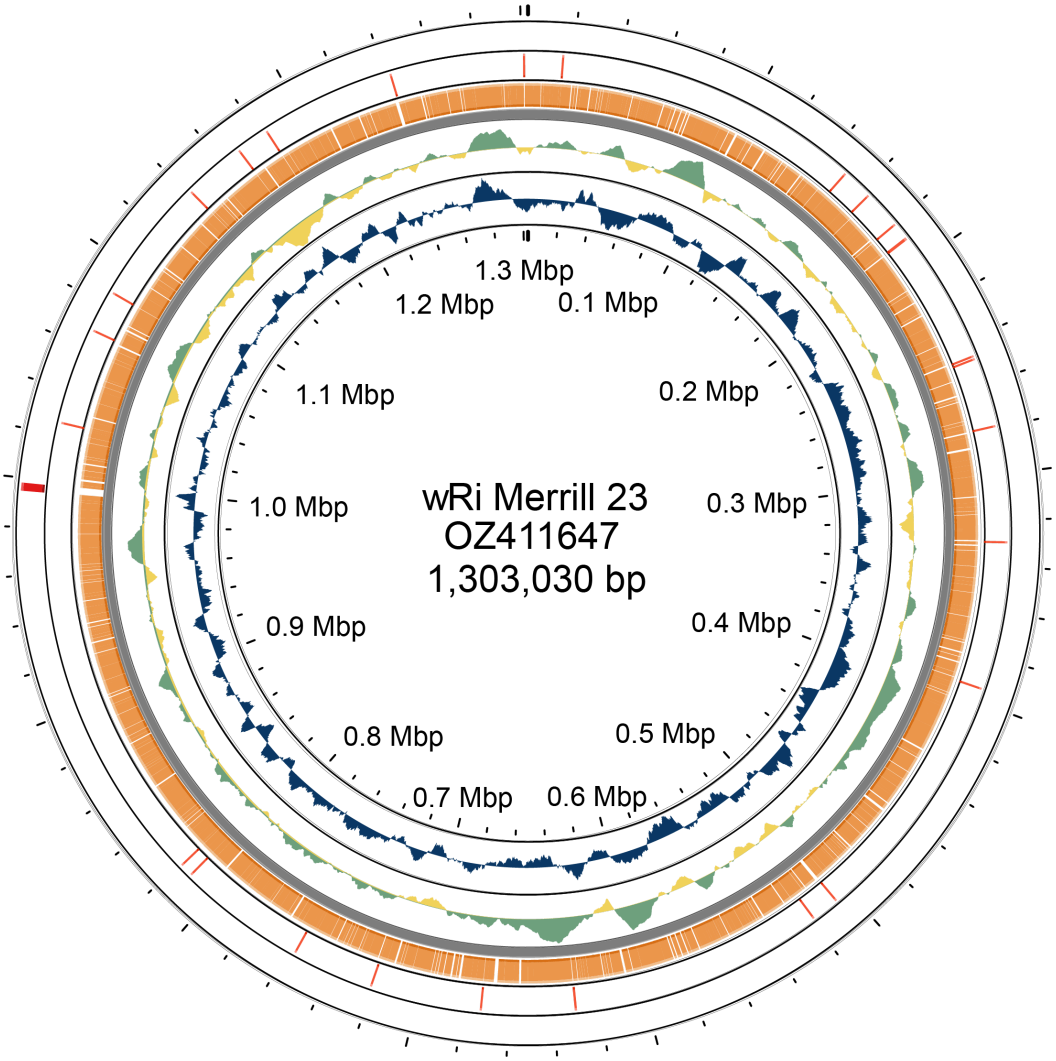


**Figure 1. *Wolbachia* *w*Ri genome map.** Concentric circles show (outer to inner): rRNA genes (red), tRNA genes (red), coding sequences (orange), GC skew (green/yellow for high/low), and GC content (blue), with GC metrics plotted as deviations from genome-wide average.

| ***w*Ri Merrill 23 Annotation summary** | |
|---|---|
| Annotation pipeline | Prokka v1.1.1 |
| Annotation method | kingdom:bacteria |

| | |
|---|---|
| Length (bp) | 1,303,030 |
| GC Content | 35.21% |
| Genes (total) | 1,288 |
| CDSs (total) | 1,258 |
| Genes (RNA) | 30 |
| rRNAs | 1, 1, 1 (5S, 16S, 23S) |
| tRNAs | 28 |
| ncRNAs | 0 |
| Pseudogenes (total) | 0 |

Table 1. Annotation summary statistics.

**Acknowledgements:**
The authors acknowledge the University of California Santa Cruz Genomics Institute for providing computational resources, including the Phoenix computational cluster, and support for this project. The authors thank Rion Parsons for his support and the University of California Santa Cruz for use of the Hummingbird computational cluster. The authors thank James Letchinger for the use of his computer for nanopore sequencing. The authors thank the University of California Santa Cruz Baskin Engineering Lab Support team for providing laboratory space and support. Funding for this project was provided by NIH (T32 HG012344) awarded to JJ, CW, HL, AN, CS, and AC, NIH awards to SLR (R00GM135583, R35GM157189), and the NSF-GRFP awarded to AN.

**References**

1. Russell SL, Castillo JR. Trends in symbiont-induced host cellular differentiation. Results Probl Cell Differ. 2020;69: 137–176.

2. Russell SL, Castillo JR, Sullivan WT. Wolbachia endosymbionts manipulate the self-renewal and differentiation of germline stem cells to reinforce fertility of their fruit fly host. PLoS Biol. 2023;21: e3002335.

3. Hoffmann AA, Turelli M, Harshman LG. Factors affecting the distribution of cytoplasmic incompatibility in Drosophila simulans. Genetics. 1990;126: 933–948.

4. Turelli M, Hoffmann AA. Cytoplasmic incompatibility in Drosophila simulans: dynamics and parameter estimates from natural populations. Genetics. 1995;140: 1319–1338.

5. Carrington LB, Lipkowitz JR, Hoffmann AA, Turelli M. A re-examination of Wolbachia-induced cytoplasmic incompatibility in California Drosophila simulans. PLoS One. 2011;6: e22565.

6. Klasson L, Westberg J, Sapountzis P, Näslund K, Lutnaes Y, Darby AC, et al. The mosaic genome structure of the Wolbachia wRi strain infecting Drosophila simulans. Proc Natl Acad Sci U S A. 2009;106: 5725–5730.

7. Faria VG, Sucena É. From nature to the lab: Establishing Drosophila resources for evolutionary genetics. Front Ecol Evol. 2017;5. doi:10.3389/fevo.2017.00061

8. Casper-Lindley C, Kimura S, Saxton DS, Essaw Y, Simpson I, Tan V, et al. Rapid fluorescence-based screening for Wolbachia endosymbionts in Drosophila germ line and somatic tissues. Appl Environ Microbiol. 2011;77: 4788–4794.

9. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37: 540–546.

10. Jacobs J, Nakamoto A, Mastoras M, Loucks H, Mirchandani C, Karim L, et al. Complete de novo assembly of Wolbachia endosymbiont of Drosophila willistoni using long-read genome sequencing. Sci Rep. 2024;14: 17770.

11. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890.

12. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9: e112963.

13. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38: 4647–4654.

14. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069.

15. Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen C-Y, et al. Proksee: in-depth characterization and visualization of bacterial genomes. Nucleic Acids Res. 2023;51: W484–W492.