






A continental-scale survey of *Wolbachia* infections in blue butterflies reveals evidence of interspecific transfer and invasion dynamics

Vivaswat Shastry ^{1,*}, Katherine L. Bell,² C. Alex Buerkle ³, James A. Fordyce ⁴, Matthew L. Forister,² Zachariah Gompert ⁵, Sarah L. Lebeis ⁶, Lauren K. Lucas,⁵ Zach H. Marion,⁷ Chris C. Nice⁸

¹Committee on Genetics, Genomics and Systems Biology, University of Chicago, Chicago, IL 60637, USA,

²Department of Biology, University of Nevada, Reno, NV 89557, USA,

³Department of Botany, University of Wyoming, Laramie, WY 82071, USA,

⁴Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA,

⁵Department of Biology, Utah State University, Logan, UT 84322, USA,

⁶Department of Microbiology & Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA,

⁷Bio-Protection Research Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand,

⁸Department of Biology, Population and Conservation Biology, Texas State University, San Marcos, TX 78666, USA

*Corresponding author: Committee on Genetics, Genomics and Systems Biology, University of Chicago, Chicago, IL 60637, USA. Email: vivaswat@uchicago.edu

Abstract

Infections by maternally inherited bacterial endosymbionts, especially *Wolbachia*, are common in insects and other invertebrates but infection dynamics across species ranges are largely under studied. Specifically, we lack a broad understanding of the origin of *Wolbachia* infections in novel hosts, and the historical and geographical dynamics of infections that are critical for identifying the factors governing their spread. We used Genotype-by-Sequencing data from previous population genomics studies for range-wide surveys of *Wolbachia* presence and genetic diversity in North American butterflies of the genus *Lycaeides*. As few as one sequence read identified by assembly to a *Wolbachia* reference genome provided high accuracy in detecting infections in host butterflies as determined by confirmatory PCR tests, and maximum accuracy was achieved with a threshold of only 5 sequence reads per host individual. Using this threshold, we detected *Wolbachia* in all but 2 of the 107 sampling localities spanning the continent, with infection frequencies within populations ranging from 0% to 100% of individuals, but with most localities having high infection frequencies (mean = 91% infection rate). Three major lineages of *Wolbachia* were identified as separate strains that appear to represent 3 separate invasions of *Lycaeides* butterflies by *Wolbachia*. Overall, we found extensive evidence for acquisition of *Wolbachia* through interspecific transfer between host lineages. Strain *wLycC* was confined to a single butterfly taxon, hybrid lineages derived from it, and closely adjacent populations in other taxa. While the other 2 strains were detected throughout the rest of the continent, strain *wLycB* almost always co-occurred with *wLycA*. Our demographic modeling suggests *wLycB* is a recent invasion. Within strain *wLycA*, the 2 most frequent haplotypes are confined almost exclusively to separate butterfly taxa with haplotype A1 observed largely in *Lycaeides melissa* and haplotype A2 observed most often in *Lycaeides idas* localities, consistent with either cladogenic mode of infection acquisition from a common ancestor or by hybridization and accompanying mutation. More than 1 major *Wolbachia* strain was observed in 15 localities. These results demonstrate the utility of using resequencing data from hosts to quantify *Wolbachia* genetic variation and infection frequency and provide evidence of multiple colonizations of novel hosts through hybridization between butterfly lineages and complex dynamics between *Wolbachia* strains.

Keywords: *Wolbachia*; *Lycaeides*; infection acquisition; geography of infection; GBS data; host parasite interactions

Introduction

The endosymbiotic bacteria in the genus *Wolbachia* (Hertig and Wolbach 1924; Hertig 1936) have been studied for their phenotypic effects on their invertebrate hosts, mostly arthropods and nematodes (Yen and Barr 1971; Charlat et al. 2003; Moran et al. 2008; Werren et al. 2008; Kriesner et al. 2013). The impacts on hosts include extraordinary reproductive manipulation as well as mutualistic interactions (Werren et al. 2008). Reproductive manipulations include cytoplasmic incompatibility (mortality of

host embryos when infected males mate with uninfected females), feminization, sex ratio distortion, and male killing. Mutualistic interactions are observed when *Wolbachia* infections protect hosts from viral attack (Hedges et al. 2008; Teixeira et al. 2008) or facilitate sequestration of vital nutrients (Brownlie et al. 2009; Hosokawa et al. 2010). These interactions have spurred development of evolutionary models to explain the persistence of infection within populations and the spread of infections across populations, lineages, and taxa (Turelli and Hoffmann 1991;

Received: June 06, 2022. Accepted: August 10, 2022

© The Author(s) 2022. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Turelli 1994; Kriesner et al. 2013, 2016). The manipulation of host biology has even been harnessed to control pest insect populations, including insects that vector human diseases (e.g. Kambris et al. 2009; Hoffmann et al. 2011; Iturbe-Ormaetxe et al. 2011; Walker and Moreira 2011; Ross et al. 2019). However, despite nearly a half a century of research on the phenotypic effects of *Wolbachia* on their hosts, we have a relatively poor understanding of (1) how *Wolbachia* infects novel hosts and (2) the historical biogeography of infection dynamics within and among host lineages.

Wolbachia infections occur in more than half of insect species (Werren et al. 1995; Hilgenboecker et al. 2008; Weinert et al. 2015; Zug and Hammerstein 2015; Bailly-Bechet et al. 2017) and acquisition by novel hosts can occur in multiple ways. Cladogenic acquisition, also known as cospeciation or codivergence, occurs when infections are acquired from an ancestral lineage, resulting in sister taxa sharing common *Wolbachia* strains or genotypes (Cooper et al. 2019; Sanaei et al. 2021). Introgressive acquisition occurs through reproductive exchange between lineages (i.e. via hybrid formation) and constitute host shifts (Cooper et al. 2019; Sanaei et al. 2021). Alternatively, horizontal transfer might result from parasitoid (Heath et al. 1999; Stevens et al. 2001; Duron et al. 2010; Gehrler and Vorburger 2012; Gupta et al. 2021) or ectoparasite attack (Hoy and Jeyaprakash 2005; Le Clec'h et al. 2013; Gupta et al. 2021), or possibly through predation and other food sources (Huigens et al. 2000, 2004; Gerth et al. 2013). Evidence for cladogenic acquisition is sparse (Raychoudhury et al. 2009; Gerth et al. 2013; Turelli et al. 2018) and requires phylogenetic information from *Wolbachia* and hosts. Distinguishing between introgressive acquisition and the various pathways of horizontal transfer not involving reproductive interaction can be accomplished by comparisons of divergence times of *Wolbachia* and mitochondrial DNA (mtDNA) (Conner et al. 2017; Turelli et al. 2018).

Once acquired, the dynamics of *Wolbachia* prevalence in a population are presumably governed by the phenotypic effects on hosts, host immune responses to infections, coevolution, and the fidelity of *Wolbachia* transmission from host mother to offspring (Turelli 1994; Weeks et al. 2007; Jaenike 2009; Hoffmann 2020; Sanaei et al. 2021). For example, models of cytoplasmic incompatibility-inducing strains, where infected females experience fitness effects that are dependent on the frequency of infected males predict that prevalence within a host population can exhibit decreases in frequency when infections are rare, but rapidly increase when infection frequency is above an unstable threshold frequency. Such “bistable” infection dynamics could produce variation in prevalence among host populations that depend on initial infection frequencies where some populations have low to nonexistent infections, whereas others might be fixed, or nearly fixed, for the infection (Turelli and Hoffmann 1991; Barton and Turelli 2011; Kriesner et al. 2013). Alternatively, strains with positive, frequency-independent fitness effects on their hosts are predicted to increase in frequency within populations and spread spatially (Barton and Turelli 2011; Kriesner et al. 2013) much like alleles under positive selection. Range-wide surveys of infection prevalence across species ranges provide critical information for understanding the spread and maintenance of infections (Hague et al. 2022). Such data can also be used to estimate the patterns of strain distribution among host species to inform models of evolution and transmission of *Wolbachia* (Kriesner et al. 2013; Turelli et al. 2018; Cooper et al. 2019). Further, temporal sampling could also be incorporated to investigate the factors that govern the spatial spread of infections (Riegler et al. 2005; Kriesner et al. 2013).

Despite a robust theoretical foundation for *Wolbachia* infection acquisition by novel hosts and evolutionary dynamics, detailed range-wide surveys of *Wolbachia* prevalence in natural populations and species have been conducted in a limited number of hosts (e.g. Shoemaker et al. 2000; Shoemaker, Ahrens, et al. 2003; Shoemaker, Keller, et al. 2003; Narita et al. 2006; Baldo et al. 2008; Nunes et al. 2008; Schuler et al. 2016, 2018; Turelli et al. 2018; Hague et al. 2021; Walker et al. 2021) or ecological communities (Gupta et al. 2021). Historically, *Wolbachia* infections have been assayed using PCR-based methods targeting *Wolbachia* 16S rDNA genes or other *Wolbachia*-specific markers such as the Multilocus Strain Typing (MLST) loci (Turelli and Hoffmann 1995; Werren and Windsor 2000; Baldo et al. 2006, 2008). Presence of *Wolbachia* in a host individual is indicated by an amplified band on agarose gels and strain identification can be performed by sequencing the MLST loci (Baldo et al. 2006, 2008). However, PCR-based assays can be time-consuming, especially for surveys of large numbers of individual hosts, and are subject to false-positive errors from contamination of samples and false-negative errors from failed PCR reactions, among other problems. While various methods have been developed to minimize them (e.g. Turelli and Hoffmann 1995; Nice et al. 2009), these errors can present challenges for range-wide surveys of infections in hosts.

A promising and inexpensive approach for such surveys uses *Wolbachia* resequencing data from phylogeographic and population genomics studies of host species (Richardson et al. 2012; Signor 2017; Pascari and Chandler 2018; Hinojosa et al. 2020; Arif et al. 2021). Here, we pursue a bioinformatics approach similar to these previous studies and use a large population genomics data set from multiple species of *Lycaeides* butterflies to highlight the potential of this approach for estimating infection prevalence and to support inferences of modes of acquisition and histories of infections.

Lycaeides butterflies colonized North America through Beringia approximately 2–4 MYA (Gompert, Fordyce, et al. 2008; Vila et al. 2011). This colonization was followed by a period of diversification that included extensive admixture among lineages (Nice et al. 2013; Gompert et al. 2014). We recognize 5 lineages that correspond to nominal species (Fig. 1 and Table 1; Gompert et al. 2014), including: *Lycaeides idas*, which also occurs in the Palearctic, and 4 North American endemics: *L. melissa*, *L. anna*, *L. ricei*, and *L. samuelis*. *Lycaeides samuelis* (formerly *L. melissa samuelis*) is known as the Karner blue butterfly and is a federally listed endangered species (Black and Vaughan 2005; Forister et al. 2011). We also recognize 3 distinct lineages within *L. melissa*: *L. melissa*—East, *L. melissa*—Rockies, and *L. melissa*—West (Chaturvedi et al. 2018). In addition to the nominal species, there are several admixed lineages that we refer to as hybrid lineages. These occur in several mountain ranges of the western United States. Putative ancient hybrid lineages formed from admixture between *L. melissa* and *L. anna* occur in the Sierra Nevada and in the White Mountains of California and Nevada (Gompert, Fordyce, et al. 2006; Nice et al. 2013; Gompert et al. 2014). In the vicinity of Jackson, Wyoming in the Grand Tetons, and Yellowstone area of the Rocky Mountains, we find populations that exhibit admixture between *L. melissa* and *L. idas* that we refer to as the Jackson hybrid lineage (Gompert et al. 2012, 2013, 2014; Chaturvedi et al. 2020). Hybrid lineages in the Warner Mountains of northeastern California, the Jarbidge Range in northern Nevada, and Steens Mountain in southeastern Oregon have complex ancestry potentially from *L. melissa*, *L. anna*, and *L. idas* (Gompert et al. 2014).

PCR-based surveys have demonstrated that North American *Lycaeides* harbor *Wolbachia* infections (Gompert, Forister, et al.

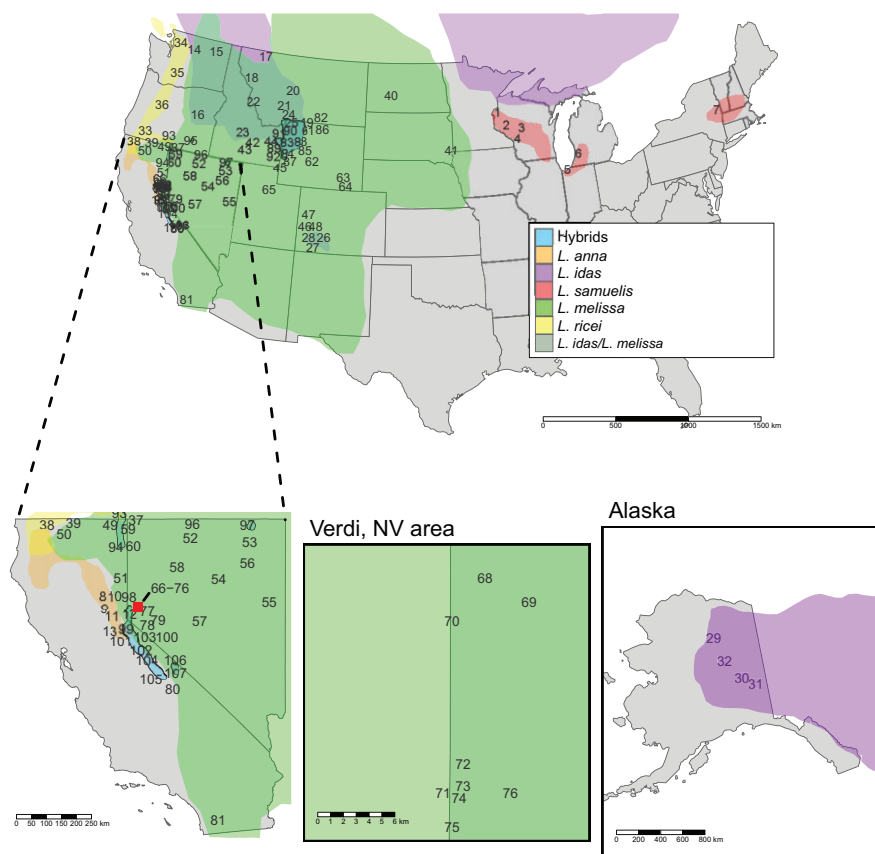


Fig. 1. Range maps of the 5 nominal species of *Lycaeides* in the United States with the 107 sampled locations plotted as site numbers corresponding to Table 1. The dense sampling in the southwestern United States is expanded in the lower left. The inset square indicates the Verdi, Nevada sampling area, including sites 66–76, and is also expanded (bottom, middle). Sample locations in Alaska are illustrated in the map on the lower right.

2008; Nice et al. 2009). For example, populations of *L. samuelis*, Karner blue butterflies, in the western portion of their range in Wisconsin were found to be nearly entirely infected (near 100% prevalence). These populations also possessed a mitochondrial haplotype identical to a haplotype found in *L. melissa* (the proposed source of the infection), but distinct from haplotypes found in the eastern portion of their range (east of Lake Michigan) (Nice et al. 2009). However, surveys of *Wolbachia* in *Lycaeides* have been limited in terms of geography and butterfly taxonomy. Here, we expand our survey to provide a continent-wide view of *Wolbachia* diversity using sequence reads from population genomics studies of *Lycaeides*.

The data considered here are Genotyping-by-Sequencing (GBS) sequence reads from 2,377 butterflies of the genus *Lycaeides* from 107 localities in North America sampled from 1996 to 2018 (Table 1 and Supplementary Table 1). These data were generated for several projects investigating patterns of differentiation and admixture across North America (Gompert et al. 2014), genomic changes during shifts to novel host plants (Gompert et al. 2015; Chaturvedi et al. 2018), and comparisons of genomic architecture between ancient hybrid lineages and a contemporary hybrid zone (Chaturvedi et al. 2020). A chromosome-level reference genome for *L. melissa* has been assembled to facilitate comparative genomic studies (Chaturvedi et al. 2018, 2020).

We used *Wolbachia* sequence reads found among GBS reads from North American *Lycaeides* butterflies to address the following questions: (1) how does variation in detection thresholds of sequence depth and sequence length influence *Wolbachia* infection frequency estimation? (2) how does the frequency of

infection (prevalence) vary across host populations, lineages, and geography? and (3) how are *Wolbachia* genotypes and groups of genotypes (which we equate to strain types) distributed across geography and host taxonomy? We use the answers to these questions to construct hypotheses about the history and biogeography of infections in *Lycaeides*. We also discuss the opportunities and limitations to using *Wolbachia* reads present in resequencing data as an inexpensive tool for understanding *Wolbachia* dynamics in natural populations. We also argue that similar data from other host taxa might contribute to the growing understanding of the evolution and history of *Wolbachia*-host interactions.

Materials and methods

Sequencing of *Lycaeides* individuals

We extracted genomic DNA, generated GBS libraries, and sequenced these libraries following the methods described in Parchman et al. (2012), Gompert et al. (2014), and Chaturvedi et al. (2018). Briefly, genomic DNA was extracted from thoracic tissue for all specimens and purified using Qiagen's DNeasy Blood and Tissue kit (Qiagen Inc.). Genomic DNA was digested with restriction enzymes EcoRI and MseI. Adaptors, including a unique 8- to 10-bp sequence barcode and the Illumina primer sequences, were ligated to the fragmented DNA with T4 DNA ligase. Adaptor and primer sequences are provided in Supplementary Table 2. We then PCR-amplified the fragment libraries with standard Illumina PCR primers. Amplified libraries were then pooled and size selected (300–450 bp) with a BluePippin. The GBS libraries were sequenced across several lanes of Illumina HiSeq 2500 or

Table 1. Sample information for 107 *Lycaeides* butterfly collection localities.

No.	Locality	Nominal species	n	No. infected	Data	Wolbachia haplotype
1	Fish Lake	<i>L. samuelis</i>	20	20	Present	A1(14)
2	Eau Claire	<i>L. samuelis</i>	22	21	Present	A1(2)A8(1)
3	Black River	<i>L. samuelis</i>	17	17	Present	A1(14)
4	Fort McCoy	<i>L. samuelis</i>	23	23	Present	A1(20)A10(1)
5	Indiana Dunes	<i>L. samuelis</i>	21	1	Present	
6	Allegan	<i>L. samuelis</i>	30	0	Present	
7	Saratoga Spr.s	<i>L. samuelis</i>	27	0	Present	
8	Fall Cr	<i>L. anna</i>	20	20	G	C1(11)
9	Yuba Gap	<i>L. anna</i>	20	20	G	C1(14)J1(1)
10	Castle Pk	<i>L. anna</i>	18	16	G	C1(9)
11	Donner Pass	<i>L. anna</i>	18	17	G	C1(4)
12	Marlette Lk	<i>L. anna</i>	19	19	Present	C1(9)
13	Leek Spr.s	<i>L. anna</i>	20	20	G	C1(18)
14	Cottonwood	<i>L. idas</i>	25	25	Present	A2(24)
15	White Mt.	<i>L. idas</i>	24	24	Present	A2(15)
16	StrawB Mt.s	<i>L. idas</i>	20	20	G	A2(17)
17	Siyeh Cr	<i>L. idas</i>	20	20	G	A2(14)
18	Soldier Cr	<i>L. idas</i>	20	19	G	A2(12)
19	Tibbs Butte	<i>L. idas</i>	20	20	G	A2(17)
20	King's Hill	<i>L. idas</i>	18	18	G	A2(12)
21	Garnet Pk	<i>L. idas</i>	20	19	G	A2(5)A12(1)B1(2)
22	Shook Mtn	<i>L. idas</i>	28	28	Present	A2(13)A13(4)A15(1)
23	Wolftone Rd	<i>L. idas</i>	4	4	Present	A2(3)A13(1)
24	Bunsen Pk	<i>L. idas</i>	20	19	G	A2(11)
25	Hayden V	<i>L. idas</i>	22	22	G	A2(11)B1(1)
26	Animas RH	<i>L. idas</i>	13	13	G	A1(6)A2(2)
27	Red Mt. P	<i>L. idas</i>	4	4	G	A1(1)A2(1)
28	Tomboy Rd	<i>L. idas</i>	24	24	G	A1(12)
29	Nolan Rd	<i>L. idas</i>	8	8	Present	
30	Spruce Barley	<i>L. idas</i>	20	20	Present	A2(1)B1(1)
31	Tok	<i>L. idas</i>	14	14	Present	A2(2)
32	Tolovana Cr	<i>L. idas</i>	9	9	Present	A2(1)A13(1)A15(1)
33	Soda Mt.	<i>L. ricei</i>	20	19	G	A2(12)
34	Rainy Pass	<i>L. ricei</i>	20	20	Present	A2(12)A17(3)
35	Chinook Pass	<i>L. ricei</i>	25	25	Present	A2(17)
36	Big Lk	<i>L. ricei</i>	20	20	G	A2(10)A3(1)A4(1)B1(5)
37	Cave Lk	<i>L. ricei</i>	24	24	G	A1(1)A2(20)
38	Marble Mts.	<i>L. ricei</i>	12	7	G	C1(5)G1(1)G2(1)
39	Shovel Cr	<i>L. ricei</i>	21	20	G	C1(15)C3(1)
40	Beulah	<i>L. melissa</i> —East	10	10	Present	A1(1)
41	Brandon	<i>L. melissa</i> —East	20	18	C	A1(3)
42	Silver Cr	<i>L. melissa</i> —East	6	6	Present	
43	Richfield	<i>L. melissa</i> —East	6	5	Present	A1(2)
44	Victor	<i>L. melissa</i> —East	20	20	G	A1(11)
45	Cokeville	<i>L. melissa</i> —East	10	10	G	A1(4)
46	Montrose	<i>L. melissa</i> —East	20	20	G	A1(9)A16(1)
47	De Beque	<i>L. melissa</i> —East	20	19	G	A1(5)
48	Cimarron	<i>L. melissa</i> —East	6	6	Present	A1(1)A7(1)
49	Goose Lk	<i>L. melissa</i> —East	20	20	G	A1(7)
50	Montague	<i>L. melissa</i> —East	19	19	G	A1(17)
51	Susanville	<i>L. melissa</i> —East	10	10	Present	A1(6)
52	Abel Cr	<i>L. melissa</i> —East	19	19	C	A1(1)
53	Deeth	<i>L. melissa</i> —East	20	20	G	A1(8)
54	Mill Cr	<i>L. melissa</i> —East	24	24	Present	A1(14)
55	East Cr CG	<i>L. melissa</i> —East	25	25	Present	A1(8)
56	Lamoille	<i>L. melissa</i> —East	20	19	G	A1(10)B1(2)
57	Ophir City	<i>L. melissa</i> —East	19	19	G	A1(8)
58	Star Cr	<i>L. melissa</i> —East	16	16	G	A1(6)
59	Upper Alkali	<i>L. melissa</i> —East	20	19	C	A1(6)A18(2)
60	Surprise V	<i>L. melissa</i> —East	20	20	G	A1(13)
61	Cody	<i>L. melissa</i> —Rockies	23	22	G	A1(11)A2(1)
62	Lander	<i>L. melissa</i> —Rockies	24	23	G	A1(4)
63	Wheatland	<i>L. melissa</i> —Rockies	16	16	Present	A1(9)A6(1)A19(2)I1(1)
64	Yellow Pine CG	<i>L. melissa</i> —Rockies	20	20	G	A1(9)A2(1)
65	Albion Meadow	<i>L. melissa</i> —Rockies	46	46	G	A2(40)
66	Lake Davis	<i>L. melissa</i> —West	4	4	Present	A1(2)
67	Sierravalley	<i>L. melissa</i> —West	20	20	Present	A1(2)
68	White Lk	<i>L. melissa</i> —West	27	27	Present	A1(15)A6(4)A11(1)A19(1) B8(1)B44(1)
69	Silver Lk	<i>L. melissa</i> —West	18	17	G	A1(5)B10(7)
70	Girl Farm	<i>L. melissa</i> —West	24	24	Present	A1(3)A6(1)A11(2)B5(1)

(continued)

Table 1. (continued)

No.	Locality	Nominal species	n	No. infected	Data	Wolbachia haplotype
71	Verdi Crystal	<i>L. melissa</i> —West	73	68	C	B7(1)B10(1)B18(1)B23(2) B24(1)B25(1)B26(1)B27(1) B28(1)D2(1)E1(1) A1(14)A6(2)A11(1)A19(1) B5(2)B10(1)B23(1)B29(1) B30(2)B31(1)B32(1)B33(1) B34(1)H1(1)
72	Verdi classic	<i>L. melissa</i> —West	26	25	Present	A1(2)A19(1)B5(1)B28(1) B35(1)B36(1)B37(1)B38(1) B39(1)B40(1)
73	Verdi tracks	<i>L. melissa</i> —West	20	16	Present	B10(1)B11(1)B18(1)B22(1) B24(2)B33(1)B41(1)B42(1) B43(1)H2(1)
74	Verdi hwy	<i>L. melissa</i> —West	11	11	Present	A1(1)A19(2)B23(1)B37(1)
75	Qui	<i>L. melissa</i> —West	18	16	Present	A6(1)B2(1)B3(1)B4(2) B5(1)B6(1)B7(2)B8(1) B9(1)B10(1)B11(1)
76	Deer Mt Road	<i>L. melissa</i> —West	27	23	Present	B4(2)B7(1)B12(1)B13(1) B14(1)B15(1)B16(1)B17(1) B18(1)B19(1)B20(1)B21(1) B22(1)
77	Washoe Lk	<i>L. melissa</i> —West	20	18	G	A1(2)B10(1)
78	Gardnerville	<i>L. melissa</i> —West	18	17	G	B10(6)F1(1)
79	Red Earth	<i>L. melissa</i> —West	20	20	G	A1(8)
80	Bishop	<i>L. melissa</i> —West	20	20	G	A1(11)
81	Trout Pond	<i>L. melissa</i> —West	13	13	C	A1(4)
82	Big Ice	Hybrid	18	18	G	A2(11)
83	Blacktail Butte	Hybrid	46	45	G	A2(32)
84	Bull Cr	Hybrid	46	45	G	A2(27)
85	Dubois	Hybrid	41	41	G	A1(1)A2(29)
86	Hunt Mt.	Hybrid	30	30	G	A2(24)
87	Periodic Spr	Hybrid	20	20	G	A2(28)
88	Pinnacles Butte	Hybrid	20	19	G	A2(17)
89	Rendezvous Mt	Hybrid	32	32	G	A2(28)
90	Riddle Lk	Hybrid	30	28	G	A2(22)
91	Sheffield Cr	Hybrid	26	26	G	A2(22)
92	Swift Cr	Hybrid	4	3	G	A2(2)
93	Buck Mt	Hybrid	44	44	G	A2(28)A5(1)
94	Eagle Pk	Hybrid	40	40	G	A2(32)A9(1)
95	Steens Mt	Hybrid	13	11	G	A2(5)
96	Hinkley	Hybrid	26	26	Present	A2(21)A13(1)A14(2)
97	Jarbridge	Hybrid	42	40	Present	A2(30)A11(1)A13(5)A14(2) A15(1)
98	Mt Rose	Hybrid	52	8	G	
99	Carson Pass	Hybrid	50	32	G	C1(20)
100	Corey Pk	Hybrid	8	8	G	
101	Sonora Pass	Hybrid	44	33	G	C1(15)
102	Lake Emma	Hybrid	33	17	G	C1(8)
103	Sweetwater	Hybrid	23	13	G	C1(10)
104	Tioga Crest	Hybrid	38	21	G	C1(5)
105	South Fork	Hybrid	14	5	G	C1(5)
106	County Line	Hybrid	40	35	G	B1(1)B10(6)C1(18)D1(1)
107	Reed Flat	Hybrid	9	8	G	C1(4)C2(1)

Locality numbers, locality names, nominal species designations (see text for details), number of individuals sampled, number of infected individuals detected, data source and *Wolbachia* haplotypes (Fig. 4) are provided. Infected individuals were identified using a threshold of a minimum of 5 sequence reads of at least 80 bp in length. The data source column indicates previously published sequence data (G = Gompert et al. 2014, C = Chaturvedi et al. 2018 or sequence data presented here for the first time (present)).

HiSeq 4000 (100 bp, single-end reads) by the Genome Sequencing and Analysis Facility at the University of Texas (Austin, TX).

Obtaining *Wolbachia* sequence reads from host GBS reads

Though we knew from PCR-based surveys that *Lycaeides* butterflies harbor *Wolbachia* (Gompert, Fordyce, et al. 2008; Nice et al. 2009), we had very little information about strain types or even the diversity of strains that might be encountered in a broader survey. Preliminary assemblies of sequence reads from *Lycaeides* to different publicly available *Wolbachia* genomes revealed variation in number of assembled reads across localities and reference

genomes (data not presented). We interpreted this as a possible indication that butterflies from different localities, or taxa, harbored a diversity of *Wolbachia* strains and that different reference genomes might yield better assemblies for some strains and therefore also some host localities or taxa.

Given this possibility, we explored several assembly strategies for creating reference genomes in silico. Our first assembly was performed by concatenating 3 *Wolbachia* reference genomes representing each of supergroups A, B, and F (Ramírez-Puebla et al. 2015). Reference genomes are available for supergroups A–F. *Wolbachia* supergroups A and B are commonly found in insects. Supergroup F is found in both insects and nematodes, while,

among supergroups with representative *Wolbachia* reference genomes, supergroups C, D, and E are found exclusively in nematodes (Ramírez-Puebla *et al.* 2015). Thus, this concatenated genome we constructed represented the most likely supergroups that might be observed in *Lycaeides* butterflies. The 3 representative reference genomes came from *Wolbachia* in *Drosophila melanogaster* (*wMel*, supergroup A; Wu *et al.* 2004), *Aedes albopictus* (*wAlbB*, supergroup B; Mavingui *et al.* 2012), and *Cimex lectularius* (*wCle*, supergroup F; Nikoh *et al.* 2014).

Because nuclear integration of *Wolbachia* genes into the host genome (e.g. Nikoh *et al.* 2008; Choi *et al.* 2015) is possible, we mapped putative *Wolbachia* reads identified by assembly to the concatenated reference genomes described above to the *L. melissa* reference genome (Chaturvedi *et al.* 2018) (details in Supplementary Table 3). Upon querying the location of the mapped *Wolbachia* reads in the host genome, we found that all the reads mapped to one of the smaller scaffolds (Scaffold 1260, 1.62 Mb) out of the 1,651 scaffolds in the *L. melissa* genome, and not from any of the larger chromosomal level (23 autosomes and Z sex chromosome) scaffolds. Based on the length (similar to size of other *Wolbachia* genomes) and mapping metrics of this region, we believe this scaffold to be the genome of *Wolbachia* infecting the host butterfly individual used in the genome assembly. We then pursued a second assembly in which we used this Scaffold 1260 as a species-specific reference *Wolbachia* genome for further analysis.

Lastly, we used a pan-genome approach (Tettelin *et al.* 2005; Vernikos *et al.* 2015) to build a reference representing the super set of genes from the above *Wolbachia* reference genomes. Here we used the supergroup A, B, and F genomes described above plus the Scaffold 1260 from the *Lycaeides* reference genome. The pan-genome was constructed by first annotating the representative *Wolbachia* genomes using *prokka* (version 1.14.6, Seemann 2014) to convert the *fasta* files to *gff* format (using the 'Moderate' parameters in <https://github.com/tseemann/prokka>, last accessed 2022 Jul 13), and then combining the files to produce a reference pan-genome using the *Roary* (version 3.13.0, Page *et al.* 2015) and GNU *Parallel* softwares (Tange 2011). Genes with paralogs and genes with less than 98% BLASTp percentage identity with each other were removed from the pan-genome. Finally, the pan-genome contained 11, 114, and 4,547 genes (total: 4,672 genes) present in 3, 2 and 1 of the 4 constituent genomes, respectively. The pan-genome was also considerably larger at approximately 3.25 Mb (as expected), compared to Scaffold 1260 (1.62 Mb) and the concatenated reference (2.8 Mb).

In all 3 iterations (i.e. using the concatenated reference, Scaffold 1260 from the *Lycaeides* reference genome, or the pan-genome), reads were aligned using *bowtie2* software (all aligned reads reported using `-a -a1 -no-una1`, version 2.3.4.2, Langmead and Salzberg 2012). The mapped reads from each of the 2,377 individuals were output as *sam* files to allow for easy parsing and analysis downstream. Multiple previous studies (e.g. Richardson *et al.* 2012; Signor 2017; Scholz *et al.* 2020) show that an approach similar to the above is effective in not only retrieving large amounts of endosymbiont genomic data from host reads but also conducting population-level analyses on the extracted endosymbiont lineages. There were minor differences in the metrics of the intermediate bioinformatics analyses (e.g. number of reads, etc., listed in Supplementary Table 3) depending on the reference genome used, but we found very similar results in detecting infected individuals (see below) and in the final construction of the haplotype networks, identification of strains and in the geographic patterns of genetic variation (see below) across these

assemblies. As a result, our subsequent analyses were based on the pan-genome reference assembly as this contains a super set of our genes from all assemblies. Details about the analyses with the other reference genomes are presented in the Supplementary Material (for instance, gene annotations for the pseudo-haplotype in the Scaffold 1260 reference presented in Supplementary Table 4).

Detecting infection from the mapped *Wolbachia* reads

We first quantified the number of mapped reads and the length of mapped reads in an individual's *sam* file from the pan-genome assembly as 2 metrics for detecting infected individuals. We then examined how various minimum thresholds of these 2 metrics affected the classification of host individuals as infected and compared the results to a previous PCR-based study of a subset of the current individuals (128 out of 2,377) from Nice *et al.* (2009). To collect these metrics, we used *samtools view -F 2432* (skipping secondary alignments, version 1.12, Li *et al.* 2009) to determine the length (in base pairs, bp) and number of reads of each unique individual alignment for reads filtered to have a mapping quality of greater than 20 (less than a 1% chance of error, as is standard in typical pipelines). A similar type of bioinformatics approach has been previously used successfully by Pascar and Chandler (2018) to detect *Wolbachia* infection in various arthropod species. We deviate from previous purely bioinformatics studies by choosing a more appropriate threshold (for *Lycaeides*) for infection that maximizes concordance of infection status with results from the previous PCR-based amplification study in these same butterfly species.

Quantifying genetic diversity in *Wolbachia* strains

The individual *sam* files were each compressed into *bam* files (using *samtools*) for more efficient downstream analyses. We then performed variant calling and genotyping on the sorted and indexed *bam* files from the previous step using the *bcftools mpileup* command (skipping indels) followed by the *bcftools call* and *view* commands (version 1.9, Li 2011) to produce a raw *vcf* file across all 2,377 individuals using the pan-genome reference genome described above. We ignored indels, assumed a ploidal level of 1 (haploid) and retained only bi-allelic sites (`-ploidy 1 -variants-only -m2 -M2 -v snps`). The choice to employ a haploid model was based on preliminary analyses from a diploid model. Given the existence of more than 1 major *Wolbachia* strain and sympatry among strains in some instances (see Results), it is possible that individual host butterflies could contain multiple infections (i.e. a single individual hosting 2 or more *Wolbachia* strains). However, models for variant calling with higher ploidy (for instance, a diploid model that might be more appropriate for multiple infections) compromised our ability to call variants as haplotypes because phasing of alleles at multiple sites was not possible. Therefore, we employed the haploid model to produce useful haplotypic data. This undoubtedly prevented discovery of additional haplotypes in individuals with multiple infections, but did produce population genetic data for those individuals with single infections. The raw *vcf* file was then filtered to only keep sites with a maximum missingness of 25% using *vcftools* (version 0.1.14, Danecek *et al.* 2011). The final *vcf* file contained 115 SNPs and 2,377 individuals in total, as a result of our conservative filtering.

The alleles in each individual from this *vcf* file could now be regarded as representing *Wolbachia* haplotypes. However, to minimize uncertainty in the haplotypic data, we again filtered the

data by retaining individuals with no missing data across variant sites (i.e. individuals with no missing data had at least 1 read of mapping quality greater than 20 of either the reference or the alternative allele at every site). We retained 1,277 individuals (out of 2,113 infected individuals, see Results) with haplotypes of length 115 bp.

We clustered the individual haplotypes using a statistical parsimony network approach (Templeton et al. 1992; Crandall et al. 2000) using the `haplotypes` (version 1.1.2, Aktas 2020) package in R with a parsimony threshold of 95%. All analyses in R were performed on version 4.0.3 (R Core Team 2021). As a complementary approach, we performed a principal coordinates analysis (PCoA) on the matrix of pairwise sequence distances calculated with the `haplotypes` package and using the `prcomp` function in R. Based on these analyses, we identified 3 major groups of haplotypes that we consider as distinct *Wolbachia* strains (see Results). Strain types, and haplotypes within strains, were then mapped onto the geographical and taxonomic distributions of the host butterflies. Thus, the distribution of *Wolbachia* strains and haplotypes (choropleth maps produced using `tmap v3.3-1`, Tennekes 2018) were examined in the context of the biogeography of their hosts and used to construct hypotheses about the origin and dynamics of infection within *Lycaeides* butterflies.

Reconstructing demographic history of *Wolbachia* strains

Lastly, we investigated the demographic history, specifically, changes in effective population size through time, for each of the 3 major strains (see Results) to understand *Wolbachia* population dynamics. We created a NEXUS file of all haplotypes from each of the 3 major strains (see Results) and used BEAST v2.6.3 to estimate Bayesian Skyline Plots (Drummond et al. 2005). This method fits a piece-wise linear function to the estimated population size as calculated from coalescent rates across the sequence. A single long chain, total of 75 million steps with a burn-in of 50 million steps, thinned every 50,000 steps for *wLycA* and a total of 50 million steps with a burn-in of 10 million steps, thinned every 50,000 steps for *wLycB* and *wLycC*, due to the large number of individuals and parameters in *wLycA*, was run. We ran a coalescent Bayesian skyline analysis with a HKY site model (Hasegawa et al. 1985) with a strict clock and a uniform prior on the clock rate. The full settings in the BEAUTi files are presented in the Supplementary Material. Convergence to a posterior distribution was assessed based on visualizations of the trace plots and calculation of effective sample sizes (ESS) of the posterior distribution for each network using Tracer v1.7.1 (Rambaut et al. 2018), which was also used to obtain uncorrelated parameter estimates from the sampling distribution.

Results

Genotype-by-sequencing data for *Lycaeides* individuals

For the 2,377 *Lycaeides* individuals sequenced, a total of 3,727,714,988 sequence reads were generated (mean = 1,568,244 per individual, median = 1,363,955 per individual). From the *Wolbachia* mapping protocol described in the Materials and Methods section and on filtering for reads with mapping quality (MAPQ) greater than 20, we obtained approximately 8.75 million reads spread across all individuals, with a median of approximately 3,500 mapped reads per individual and more than 90% of the reads having lengths greater than 80 bp. The total *Wolbachia* reads comprise approximately 0.2% of all sequence reads. The

distribution of mapped read lengths is shown in Supplementary Fig. 1.

Detecting infection from the mapped *Wolbachia* reads

We set our detection threshold for infection in individual butterflies at a minimum of 5 reads of >80 bp (with the maximal length being 87 bp). We found that results from this threshold matched very well with results from a previous PCR amplification study (Nice et al. 2009), with a 96.9% accuracy rate (i.e. concordance with PCR-based results, Fig. 2). A threshold read length of 80 bp was chosen since this was very close to the largest possible read from an individual, and would act as a stringent threshold for infection detection. We also found that > 90% of mapped reads had lengths greater than 80 bp (see Supplementary Fig. 1). Similarly, we chose 5 reads as our threshold sequence depth because this threshold minimized error with comparison to PCR tests. We found that by increasing the threshold number of mapped reads, we were increasing our false negative rate (FNR) for classification by declaring putatively infected individuals (based on PCR tests) as being uninfected. This type of approach results in a sharp drop in the accuracy as we increase the threshold beyond a read depth of 600 since fewer individuals are classified as being infected (due to the stringent threshold) and therefore, increase the false negative error in our comparison. The 5 reads threshold provided a good balance between the false positive and false negative error rates (Fig. 2). However, we note that the PCR-based amplification studies are also prone to inaccuracies that could affect our accuracy estimates.

The numbers of infected individuals were not substantially changed by varying the minimum number of reads required to diagnose infected individuals for most localities. The exceptions where prevalence did vary with different thresholds were localities for Karner Blue butterflies in the eastern portion of their range (Indiana Dunes (5), Allegan (6), Saratoga (7)) and several of the Sierra Nevada hybrid lineage localities (98–99, 101–105). (Note: when referring to specific localities, we include the site number(s) from Table 1 and Fig. 1 in parentheses following the locality names.) In these localities, raising the minimum number of reads substantially reduced the number of infected individuals detected. Supplementary Table 5 presents the numbers of infected individuals using thresholds of a minimum of 1×, 5×, and 20× reads, and Supplementary Fig. 2 provides a detailed examination of the relationship between minimum number of reads and read lengths on the percentage of infected individuals detected across all individuals.

Based on the threshold of a minimum of 5 reads of at least 80 bp (Supplementary Fig. 2), we found that a majority of *Lycaeides* localities had infection frequencies that exceeded 90% of individuals, with 85 of the 107 sampled localities showing greater than 90% (with 64 localities having infection frequencies of 100%) (Fig. 3 and Table 1). In populations where we observed variation for infection (i.e. infection frequencies not 0 or 1), 90.6% of females and 86.4% of males were infected (population treated as a random effect, $\chi^2 = 4.62$, $df = 1$, P -value = 0.032). At the species or lineage level, most infections rates are greater than 94% (Table 2). The exceptions included *L. samuelis* localities in the eastern portion of their range (5–7) (infection rates: 0–0.5%), 1 population of *L. ricei* from the Marble Mts. in California (38) (infection rate: 58%), a small number of *L. melissa* populations, mostly in the western Great Basin (43, 73, 75, 76) (infection rates: 80–88%), 1 population of the hybrid lineages in the Jackson area (at Swift Creek (92) [infection rate: 75%]), in the Sierra Nevada (98,

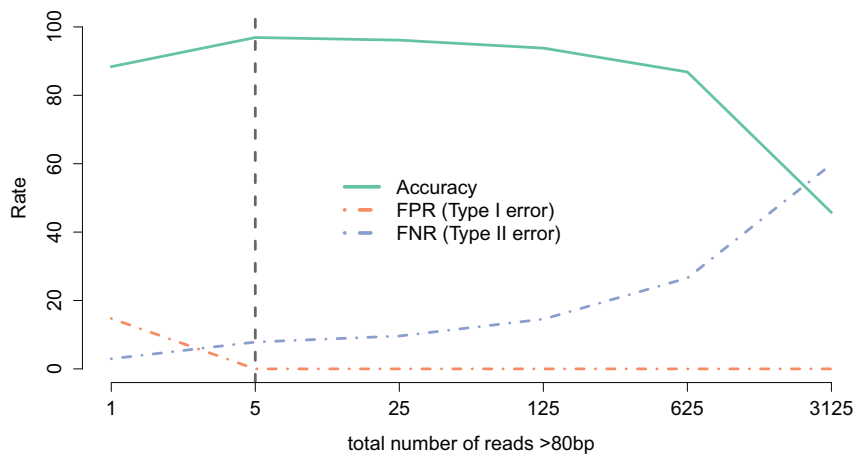


Fig. 2. Accuracy and error rates of comparing bioinformatics results to previous PCR-based studies for detecting putative *Wolbachia* infections in the genome for 129 individuals (shown here for a threshold of varying number of reads of length greater than 80 bp). We used a threshold read depth of 5 for classifying an individual as infected, as it had the highest accuracy of 96.9% correspondence with the PCR-based results, while still maintaining a low FNR (classifying an individual as not being infected when the individual is inferred to be infected from PCR-based analysis). FPRs (compared to PCR-based results) were generally low. Note that the X-axis is on a log-scale.

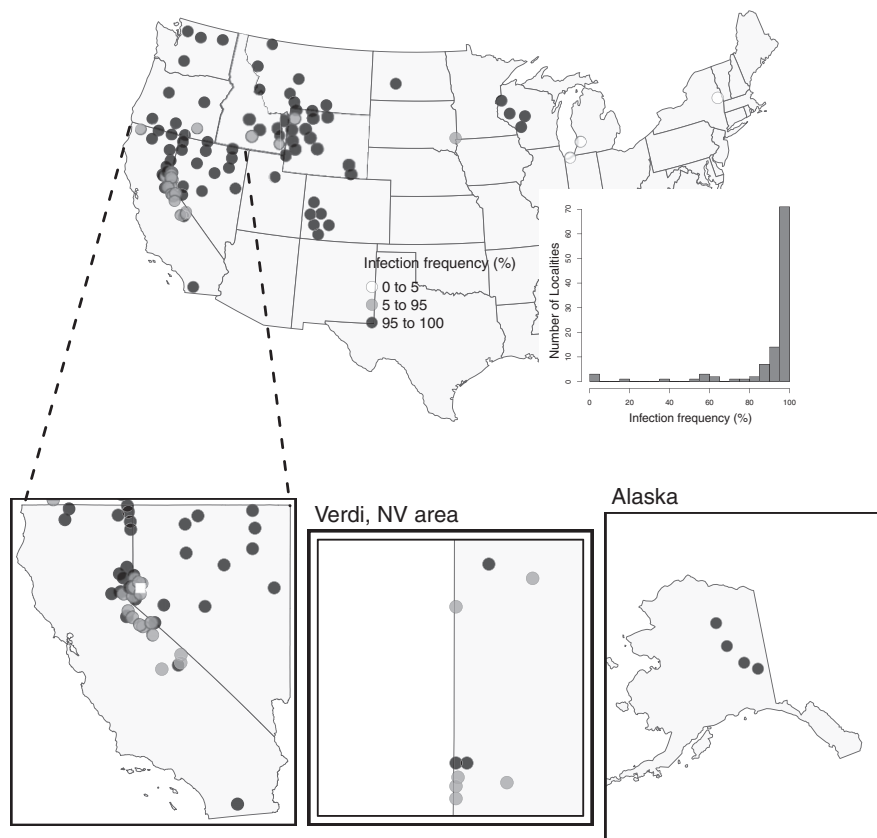


Fig. 3. Bubble plots indicating the proportion of infected individuals in a population across the 107 sampled locations. Most populations in the western United States are mostly or wholly infected (> 95%), while the *L. samuelis* populations in the east show low to no infection (< 5%). Inset plots zoomed into regions of interest for visibility. The white square indicates the Verdi, NV sampling area, and is expanded (bottom, middle). Inset plot is a histogram of infection frequencies across 107 sampling localities using a threshold of a minimum of 5 sequence reads of at least 80 bp.

99, 101–105) (infection rates: 15–75%) and in the White Mountains (106, 107) (infection rates: 87–89%) (Table 1).

Quantifying genetic diversity in *Wolbachia* strains

The filtered *vcf* file with 115 variable sites and 1,277 individuals was used for population genetic analyses. Based on a haplotype

network analysis with 95% statistical parsimony and PCoA of pairwise distances among haplotypes, we found that 1,267 out of 1,277 genotyped individuals carried *Wolbachia* haplotypes from 1 of 3 major haplotype networks (Fig. 4 and Supplementary Fig. 3) that correspond to 3 clusters of haplotypes in our ordination of haplotypes (Fig. 5). We consider these networks as distinct

Table 2. Infection frequencies and strain distributions for *Lycaeides* butterfly species or lineages.

Species/lineage	Locality	n	Infection rate	Dominant strain	Minor strains
<i>L. samuelis</i>	1–6	160	0.51	A1	
<i>L. anna</i>	7–13	115	0.97	C	
<i>L. idas</i>	14–32	333	0.99	A2	A1, B
<i>L. ricei</i>	33–39	142	0.95	A2	A1, C
<i>L. melissa</i> —East	40–60	350	0.98	A1	
<i>L. melissa</i> —Rockies	61–65	129	0.98	A1	A2
<i>L. melissa</i> —West	66–81	359	0.94	A1	B
Jackson Hybrid	82–92	313	0.98	A2	A1
Warner Hybrid	93–95	97	0.97	A2	
Sierra/Whites Hybrid	98–107	311	0.58	C	B

Locality numbers correspond with Fig. 1 and Table 1. Dominant strains are the most frequently observed major strains in each lineage (i.e. *wLycA*, *wLycB* or *wLycC*). Minor strains are less frequently observed strains that are often dominant in other lineages and most likely occur in the focal lineage via interspecific transfer. The major subdivisions of strain *wLycA*, A1 and A2 are treated as strains in this accounting.

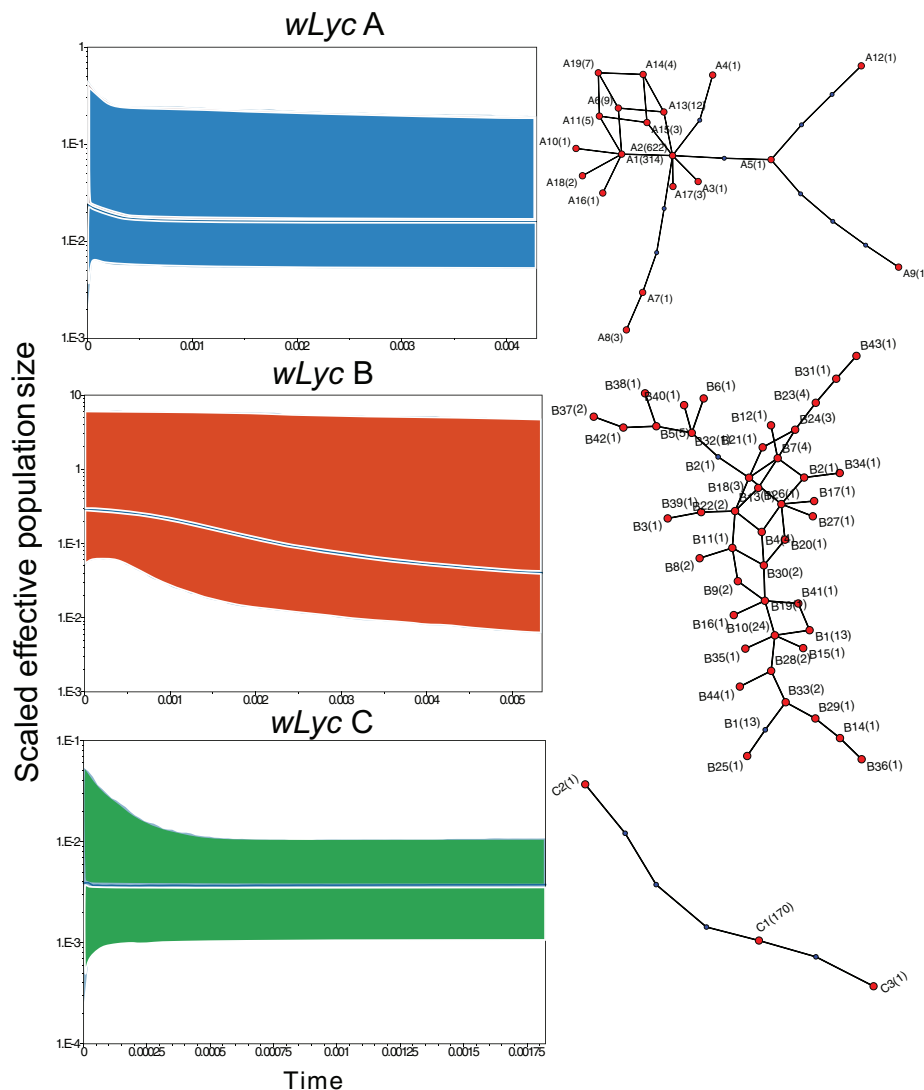


Fig. 4. Demographic histories (left) and haplotype networks (right) for each major strain (*wLycA*, *wLycB*, *wLycC*). Population sizes were estimated using BEAST 2 (Bouckaert et al. 2014). The median mutation-scaled effective population size (dashed line) and 95% credible interval (central posterior density, shaded region) for each strain is presented over time (measured in substitution rate). For simplicity, we assume equal substitution rates across strains to aid interpretation. Ninety-five percent parsimony networks show observed haplotypes in red and inferred haplotypes in blue with numbers of individuals observed possessing each haplotype in parentheses. Haplotypes are 115 bp in length.

Wolbachia strain types with individual haplotypes within networks representing mutational variation within strains (referred to hereafter as *wLycA*, *wLycB*, and *wLycC*) (Table 1 and

Supplementary Table 1). Each of these strains included between 3 and 44 distinct but closely related haplotypes (Fig. 4). The strains were substantially divergent from one another with mean

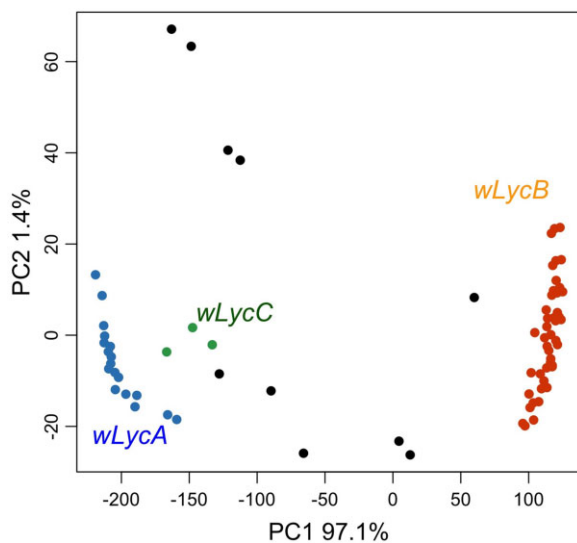


Fig. 5. Plot of PCoA of *Wolbachia* haplotypes (76 in total) based on uncorrected pairwise distances among haplotypes. Colored dots represent 115 bp haplotypes in the 3 major strains (blue: *wLycA*, orange: *wLycB*, green: *wLycC*). Strain *wLycA* was found mostly in the *L. melissa*, *L. idas*, and *L. samuelis* populations continent-wide. Strain *wLycB* was mostly found in the *L. melissa* populations in the western Great Basin. Strain *wLycC* was found exclusively in the *L. anna* populations and in the hybrids between *L. melissa* and *L. anna*. Black dots represent haplotypes found as singletons and not considered part of the 3 main strains (see Table 1 and Supplementary Fig. 3).

Table 3. Sequence divergence (across 115 variable sites) within and among the 3 major strains presented as uncorrected percent sequence divergence ($p \times 100$) and (standard deviations).

	<i>wLycA</i>	<i>wLycB</i>	<i>wLycC</i>
<i>wLycA</i>	3.3% (1.8%)		
<i>wLycB</i>	37.4% (2.8%)	3.9% (1.7%)	
<i>wLycC</i>	11.4% (2.2%)	32.1% (2.4%)	3.5% (1.7%)

pairwise divergences between strains ranging from 11.4% to 37.4% (Table 3). In addition, the diversity of *Wolbachia* strains and haplotypes within butterfly populations varied widely (Supplementary Table 1). Butterfly sampling localities ranged from localities that contained a single *Wolbachia* haplotype to localities with a maximum of 15 haplotypes (at Girl Farm (70)). The highest strain diversities were observed in the western Great Basin *L. melissa* populations and in some of the localities of hybrid lineages of *Lycaeides* (Table 1 and Supplementary Table 1 and Fig. 4).

Wolbachia strain *wLycA* was observed in 992 individuals and was the most frequent strain. Among the 19 haplotypes within strain *wLycA*, haplotypes A1 and A2 were observed in 936 individuals (94% of individuals with *wLycA* haplotypes). Though these 2 haplotypes were differentiated by a single mutational step (Fig. 4), they were mostly observed in different butterfly taxa. The A1 haplotype was found almost exclusively in *L. melissa*, while A2 was limited to *L. idas* (Tables 1 and 2). The exceptions include all 3 disjunct *L. idas* localities sampled in Colorado (26–28) where A1 was observed; A1 was also observed in the 4 *L. samuelis* localities sampled in Wisconsin (1–4) (results that match earlier PCR-based surveys; Nice et al. 2009); A2 was observed in the *L. melissa* population at Albion Meadows (65), in the Jackson hybrids (82–92), and hybrid lineages in the Warner Mountains (93–94), Jarbidge

Mountains (97), and at Steens Mountain (95) (notably not in the hybrid lineages in the Sierra Nevada and White Mountains in California and Nevada (98–107) for which *L. anna* is the maternal parent) (Table 2). Both A1 and A2 were also observed in the contemporary hybrid zone between *L. melissa* and *L. idas* at Dubois (85) (Chaturvedi et al. 2020), in the *L. ricei* population at Cave Lake (37), and in two *L. melissa* localities in the Rockies (61, 64) (Supplementary Fig. 5).

Strain *wLycB* was observed in 103 individuals and included 44 haplotypes. Strain *wLycB* haplotypes occurred most frequently in the populations of *L. melissa* in the western Great Basin (68–78) (“*L. melissa*—West” in Table 1 and Supplementary Table 1). Haplotypes B1 and B10 were the most common *wLycB* haplotypes in these western populations. The other haplotypes occur in low frequency in these *L. melissa* West populations and at County Line (106), part of the hybrid lineage in the White Mountains, and in 2 *L. idas* populations (21, 25), 1 population of *L. melissa* East (56), the Big Lake (33) population of *L. ricei*, and 3 populations of *L. idas* (Spruce Barley (30), Garnet Peak (21), and Hayden Valley (25); Supplementary Fig. 6).

Strain *wLycC* was observed in 172 individuals and included 3 haplotypes. Strain *wLycC* haplotypes were confined to *L. anna* populations (8–13) and hybrids in the Sierra Nevada and White Mountains (99, 101–107). These hybrids have mixed ancestry from both *L. melissa* and *L. anna* and the latter is presumed to be the maternal lineage based on patterns of mtDNA variation (Gompert, Fordyce, et al. 2006). The 2 exceptions for the distribution of *wLycC* haplotypes was their presence in the Shovel Creek, CA (39) and Marble Mt.s (38) *L. ricei* populations, which are the southern-most sampled *L. ricei* localities and adjacent to the range of *L. anna* (Supplementary Fig. 7).

For all 3 major strains, we found distinct right-skewed frequency distributions with 1–6 haplotypes observed in the majority of individuals and the remaining haplotypes were found in relatively few individuals, often spread over extensive areas (Fig. 4 and Table 1). The remaining 10 *Lycaeides* individuals that did not possess *Wolbachia* haplotypes from strains *wLycA*, *wLycB*, or *wLycC* contained very rare haplotypes assigned to 7 rare strains (*wLycD*–*wLycJ*) that were observed as singletons, 5 (D2, E1, F1, H1, and H2) in 4 localities in the western Great Basin (70, 71, 73, 78) (*L. melissa* West), 1 (D1) in the County Line (106) hybrid population, 2 (G1 and G2) in the Marble Mountains (38), 1 (I1) at Wheatland (*L. melissa* Rockies, 63), and 1 (J1) at Yuba Gap (*L. anna*, 9) (Table 1 and Supplementary Fig. 2).

Reconstructing demographic history of *Wolbachia* strains

Based on our analysis of demographic history across the haplotypes within each of the 3 major strains, we find different patterns for each strain in the past (Fig. 4). We found well-mixed trace plots for all 3 strains and ESS values of about 200 for strain *wLycA* and *wLycC*, and about 400 for strain *wLycB* (all 3 above the recommended threshold for independent samples from the BEAST2 manual). Strain *wLycA* (which contains mostly the *L. melissa* and *L. idas* individuals) shows a constant scaled population size of 0.02 stretching into the very distant past. Strain *wLycB* (which includes individuals from the western Great Basin (68–78), the hybrid lineages in the White Mountains in California (106), and Jackson, Wyoming area (83–84)) seems to have existed at much higher population sizes ($\sim 2.5\times$ population size of strain *wLycA*) in the distant past, but has experienced a growth phase starting 0.003 time units in the past and has grown up to $\sim 4\times$ its previous size since then. Strain *wLycC* (which is observed in the *L. anna* individuals and adjacent localities)

has a very small and constant population size (roughly $0.1\times$ of the other 2 strains) stretching into the distant past. Based on the tree event times presented in [Supplementary Fig. 8](#), we observe that both *wLycA* and *wLycB* strains have undergone population size changes in the recent past whereas strain *wLycC* shows the highest spike at time 0, indicating that population size has been relatively constant over previous time periods. The time units are measured in substitutions and we assume equal rates across the strains to aid in interpretation.

Discussion

We used a bioinformatics approach for detecting *Wolbachia* infection from GBS reads of 2,377 *Lycaeides* butterflies and validated the results from this approach by comparison with PCR-based analyses of a small subset of the host individuals ([Nice et al. 2009](#)). Using a threshold of a minimum of 5 reads of at least 80bp, we found that most individuals were infected (2,117 out of 2,377 surveyed) and 105 out of 107 localities contained infected individuals. Infection prevalences within locality samples ranged from 0% to 100% of individuals infected with a mean infection prevalence per locality of 91% infected individuals. Population genetic analyses of *Wolbachia* haplotype data provided relatively detailed phylogeographic information on 3 major *Wolbachia* strains that infect *Lycaeides* butterflies in North America. Examination of the geographic and host-taxonomic distributions of *Wolbachia* strains revealed extensive sharing of strains between populations and lineages of *Lycaeides* which represents evidence for introgressive acquisition ([Tables 1 and 2](#)). Coalescent-based demographic inferences also provided evidence that 1 of the major strains has had a recent and dramatic increase in effective population size and might currently be invading and possibly displacing another strain.

Varying the threshold minimum sequence length had little effect on detecting infected individuals because the vast majority of sequence reads were greater than 80bp in length ([Supplementary Fig. 1](#)). While the threshold of a minimum of 5 (5) reads provided the greatest accuracy (based on comparisons to PCR surveys), varying the minimum number of reads threshold had a limited impact on estimated infection frequencies except in 10 localities (5–7, 98–99, 101–105, see [Supplementary Tables 1 and 5](#)). In these localities, increasing the minimum reads threshold substantially reduced our estimate of prevalence of infected individuals. Three of these localities occur in the eastern portion of the range of the Karner Blue butterfly (*L. samuelis*) (5–7), but Karner blue populations in the western portion of the range (1–4) do not exhibit the same reduction in estimated prevalence with increasing minimum reads threshold. Similarly, the other localities that show the decline in numbers of infected individuals with increasing minimum reads threshold occur in the hybrid lineage of *Lycaeides* in the Sierra Nevada (98–99, 101–105), yet other hybrid lineages do not show a similar pattern. It is not immediately obvious why these localities differed in their apparent sensitivity to the minimum reads threshold. The overall number of sequence reads per individual could affect the probability of detection, but while the eastern Karner localities have lower median number of reads compared to the total set of 2,377 individuals, the Sierran hybrid populations have more reads per individual than the overall median (median number of sequence reads: eastern Karners: 1,078,622, Sierran hybrids: 1,810,680, overall: 1,359,589). Alternatively, it is possible that there is variation in *Wolbachia* densities within individuals among localities that influences detection probability ([Unckless et al. 2009](#); [Hague et al. 2022](#); [Shropshire et al. 2021](#)). While we cannot explain this observation

at present, it suggests that variation in *Wolbachia* infection densities in host tissues might be an important consideration when mining resequencing data for evidence of endosymbiont infection. Variation among host taxa might require careful inspection of these thresholds. In the absence of corroborating PCR-based data on infection status, we recommend examining a range of thresholds to understand how these affect the probability of detection. It is also possible that more sophisticated statistical modeling that accounts for uncertainty created by variation in numbers of sequence reads, and possibly variation in *Wolbachia* densities, could improve the probability of detecting infections.

Population genetic analyses of *Wolbachia* infections in the *Lycaeides* system facilitated inference of infection history. We do not know where or how the 3 major *Wolbachia* strains (*wLycA*, *wLycB*, and *wLycC*) were ultimately acquired by *Lycaeides* in North America. Analysis of *Wolbachia* infections from *Lycaeides* from Europe and Asia, or from associated parasites or parasitoids, might shed light on the origins of North American infections. However, our survey of geographic patterns of population genetic variation in combination with inference of demographic histories of the 3 major strains suggest that transmission of infection within North American *Lycaeides* butterflies occurred primarily through introgressive acquisition. We provide an overview of these patterns.

The comparison of demographic histories of each strain, as coalescent effective population sizes ($N_e\mu$), is facilitated by previous evidence for constant *Wolbachia* substitution rates over long time-scales ([Cooper et al. 2019](#)). The demographic history of strain *wLycA* reveals a relatively constant population size over time, and the geographic and taxonomic distribution of strain *wLycA* haplotypes is possibly consistent with either a cladogenic mode or an introgressive mode of acquisition. The 2 most frequent haplotypes in *wLycA* (A1 and A2) exhibit just 1 mutational difference ([Fig. 4](#)), yet A1 is largely confined to *L. melissa* individuals and A2 is found almost exclusively in *L. idas* individuals ([Table 1 and Fig. 6](#)). Exceptions to this pattern include hybrid lineages with either *L. melissa* or *L. idas* ancestry, or ancestry from both species (i.e. in the Jackson, Wyoming area (82–92), the contemporary hybrid zone between *L. melissa* and *L. idas* at Dubois (85), or localities at or near range boundaries, such as Cave Lake (37)). The confinement of these haplotypes largely within 2 *Lycaeides* species seems compatible with the hypothesis of cladogenic acquisition in the ancient past through a common ancestral lineage of *L. idas* and *L. melissa*, followed by independent divergence of the 2 lineages. Alternatively, the distribution of haplotypes A1 and A2 might be consistent with introgression from 1 of the species into the other accompanied by mutation. Further, the exceptions to the distributional pattern (e.g. hybrid lineages and a hybrid zone) appear to be examples of introgressive acquisition of strain *wLycA* haplotypes outside of *L. melissa* and *L. idas*. Thus, there is perhaps more support for introgressive acquisition of *wLycA* haplotypes, though cladogenic acquisition cannot be ruled out. Evidence for multiple modes of *Wolbachia* transmission in natural populations is also found in the *Drosophila* ([Cooper et al. 2019](#)) and *Nasonia* ([Raychoudhury et al. 2010](#)) species complexes.

A similar demographic history of constant population size over time is seen in strain *wLycC*, though the estimated population size of *wLycC* is very much smaller than the other strains ([Fig. 4](#)). Strain *wLycC* haplotypes are confined to *L. anna* populations and the hybrid lineages in the Sierra Nevada and White Mountains (98–107) for which *L. anna* is the presumed maternal lineage ([Table 1 and Fig. 6](#)) ([Gompert, Fordyce, et al. 2006](#)). The exceptions include 2 localities where strain *wLycC* haplotypes were observed, both of which lie on the boundary between the

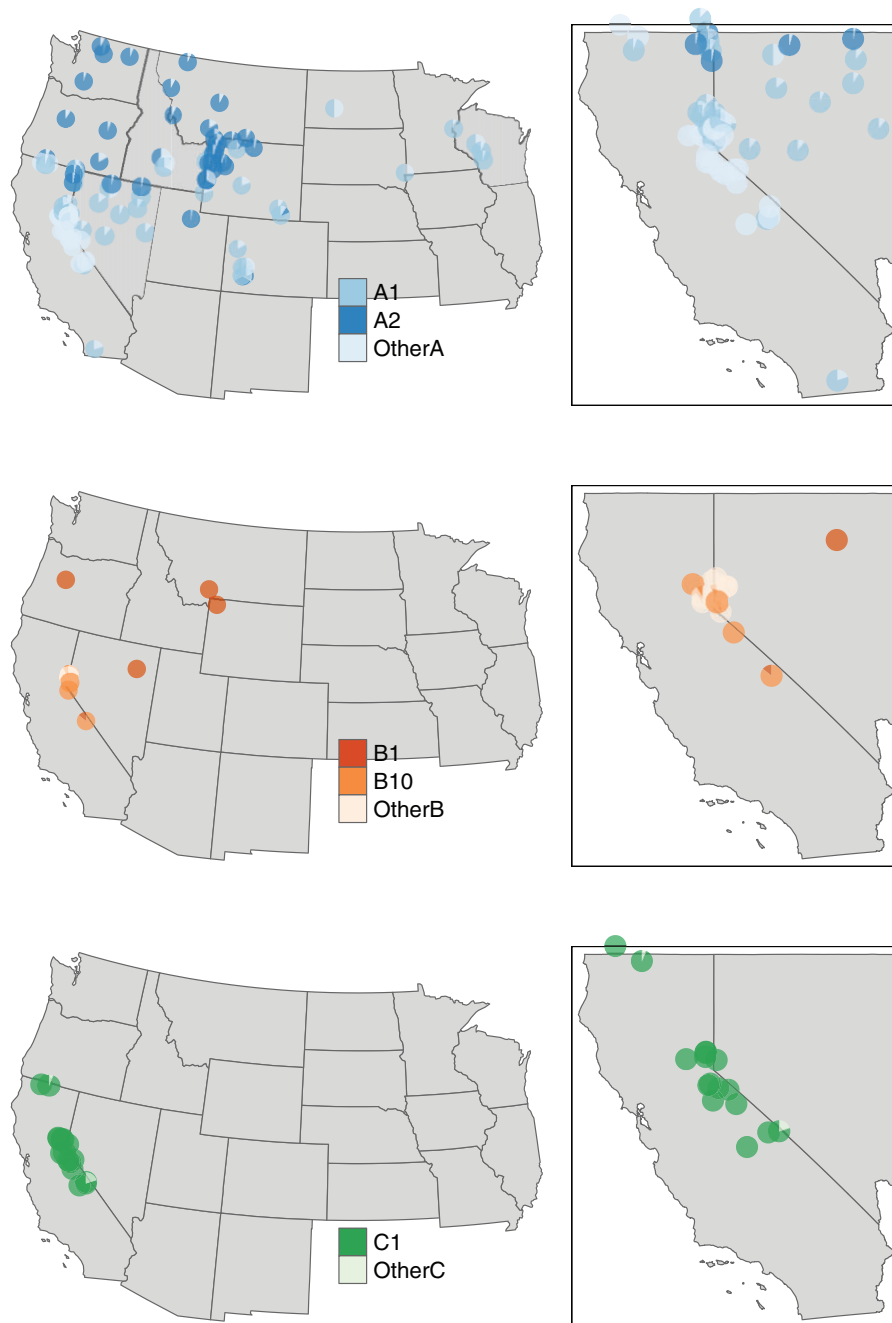


Fig. 6. Pie charts showing the distribution of haplotypes from all 3 strains (row-wise: *wLycA*, *wLycB*, *wLycC*). Haplotypes A1 and A2 are present in 90% of individuals infected with strain *wLycA*. The label “OtherA” corresponds to rare haplotypes in *wLycA* (A3–A19). Haplotypes B5, B9 and B10 make up 78% of all infections in the *wLycB* strain. Haplotype C1 makes up for 98% of all infections in the *wLycC* strain, and all other *wLycC* haplotypes are found in localities that also include haplotype C1. Pies are only shown if a given haplotype is present in the population.

ranges of *L. anna* and *L. ricei* at the Marble Mountains (38) and Shovel Creek (39). Thus, as with strain *wLycA* haplotypes, *wLycC* haplotypes appear to have spread to a limited extent outside of a *Lycaeides* species range via introgression among lineages, specifically in this case from *L. anna* to nearby populations of *L. ricei*.

The phylogeography of strain *wLycB* is different compared to the other 2 strains. This is the least frequently observed strain over all and the majority of *wLycB* haplotypes were observed in the western Great Basin in populations of *L. melissa* (68–78) (Table 1 and Fig. 6). In these locations, multiple *wLycB* haplotypes

are commonly observed along with *wLycA* haplotypes. In fact, *wLycB* haplotypes were observed without accompanying A haplotypes in only 3 locations (Verdi Tracks (73), Deer Mt. Road (76), and Gardnerville (78)). However, *wLycB* haplotypes were observed in other widely distributed places and other *Lycaeides* taxa including: *L. idas* in Alaska (30) and Montana (21), *L. ricei* at Cave Lake (37), *L. melissa* in central Nevada (56), in the Jackson hybrid lineage (84), the hybrid lineage in the White Mountains of California (106), and the putative hybrid lineage at Hinkley in northern Nevada (96). The relative rarity of this strain, its recent

population expansion (Fig. 4), coupled with its presence almost exclusively with *wLycA* across different host species points to an introgressive mode of acquisition. Strain *wLycB* haplotypes appear to be invading localities that already contain infections of strain *wLycA*. Such a mode of acquisition will lead to the presence of multiple *Wolbachia* infections or haplotypes from different strains segregating in the same population and hence, an enriched genetic diversity of *Wolbachia* in these populations (Supplementary Fig. 4 and Table 1). The concentration of strain *wLycB* haplotypes in *Lycaeides* localities in the western Great Basin, and the resulting high haplotype diversity there, suggests that this area is where the invasion of strain *wLycB* began. There is weak evidence for the hypothesis that strain *wLycB* is invading from 1 locality, Verdi Crystal (71), that was sampled over multiple years and strain *wLycB* appears to have increased in frequency from 2011 to 2018 (Supplementary Table 6). The studies from which these GBS data were obtained were not designed to assay *Wolbachia* or for temporal comparisons, and we lack statistical power to fully test this hypothesis without further sampling. The host butterflies at these localities have colonized alfalfa (*Medicago sativa*) relatively recently (Forister, Philbin, et al. 2020; Forister, Yoon, et al. 2020), probably as 1 of 3 or more independent colonizations of alfalfa (Chaturvedi et al. 2018), and probably within the last 200 years (400–600 butterfly generations; Chaturvedi et al. 2018; Forister, Philbin, et al. 2020; Forister, Yoon, et al. 2020). So, it is possible that we are tracking the effect of host population expansion in the demographic history of strain *wLycB* as it is impossible to disentangle the 2 histories without more information on host demography and quantification of *Wolbachia* titer levels. Thus, it seems that the invasion of a novel *Wolbachia* strain is occurring while the butterfly host is switching to a novel host plant. Whether there is any connection between these parallel host switches is an open question.

At a continental scale, the nominal species or lineages of *Lycaeides* butterflies each contained a dominant (most frequently observed) major strain (Table 2). Some lineages shared major strains. For example, *L. melissa* and *L. samuelis* shared strain *wLycA* (specifically haplotype A1). This pattern is consistent with interspecific transfer from *L. melissa* to *L. samuelis* (Gompert, Nice, et al. 2006; Nice et al. 2009). Beyond their specific dominant strains, most lineages also contained other “minor” strains that were dominant in other lineages but observed at lower frequency in the focal lineage (Table 2). These minor strains were commonly observed at range margins and are consistent with limited interspecific transfer. Hybrid lineages were observed to be infected by major strains associated with their putative maternal parental lineage. The Sierra/Whites hybrid lineage is infected with *wLycC* as is the inferred maternal parent *L. anna*. Similarly, the Jackson hybrid lineage is infected with *wLycA*, specifically haplotype A2, as is its maternal parental lineage *L. idas* (Table 2). Taken together, these observations illustrate considerable interspecific transfer of *Wolbachia* strains among host lineages.

The distribution of *Wolbachia* strains in *Lycaeides* butterflies is paralleled by geographical patterns of mtDNA variation observed in previous studies of these butterflies (Nice et al. 2005; Gompert, Fordyce, et al. 2008; Gompert, Forister, et al. 2008). Because *Wolbachia* infections and mtDNA are maternally inherited, they are commonly observed to be in linkage disequilibrium (Turelli and Hoffmann 1991; Turelli et al. 1992; Jiggins 2003; Hurst and Jiggins 2005). Direct comparisons are not possible because even where sampling localities overlap with the current study, those older studies of mtDNA variation used different individuals that were not sequenced for this study. Nevertheless, the presence of

3 major *Wolbachia* strains discovered here parallels the 3 major mtDNA lineages discovered in *Lycaeides*. For example, using mitochondrial sequences of the cytochrome oxidase I (COI) and cytochrome oxidase II (COII) genes, Gompert, Forister, et al. (2008) found 3 mitochondrial lineages. One lineage (lineage III from Gompert, Forister, et al. [2008]) was widely distributed across space and butterfly taxonomy that corresponds to the distribution of *wLycA* here. Another mtDNA lineage (lineage II) co-occurred with the first lineage and was detected in populations of *L. melissa* from the western Great Basin and from the hybrid population in the White Mt.s (County Line, 106), corresponding to the distribution of *wLycB*. Lastly, the third mtDNA lineage (lineage I) was observed in *L. anna*, hybrid lineages derived from *L. anna* in the Sierra Nevada and adjacent *L. ricei* localities, corresponding to *wLycC*. The close geographical correspondence of major *Wolbachia* lineages observed here and previous mtDNA haplotype distributions suggest that the expected disequilibrium between *Wolbachia* strains and mtDNA can be detected using GBS data.

Our survey of *Wolbachia* infection frequencies and genetic variation using GBS data from host *Lycaeides* butterflies suggests that this approach could be applied in other systems. Given the quantity of resequencing data generated recently, it might be possible to rapidly survey *Wolbachia* and other endosymbiont infections in a wide variety of host organisms and answer broad questions about the history, geography and mode of acquisition of infections. However, resequencing methods do not specifically target *Wolbachia* genomes and there exist several limitations. The sequence reads from *Lycaeides* GBS data did not map to any of the multilocus sequence typing (MLST) loci (Baldo et al. 2006, 2008) and it seems unlikely that GBS data in general will overlap the MLST loci. Thus, it will be impossible to identify conventionally designated strains (or possibly even *Wolbachia* supergroups) and connect studies phylogenetically from surveys of GBS data without further sequencing. Additionally, the stochasticity inherent in the methods for resequencing data, combined with the sparseness of endosymbiont sequence reads from host organisms, presents some challenges. Stochasticity arising from library preparation and from the sequencing of these multiplexed genomic libraries, among other possible sources of stochasticity, creates variation in sequence depth across fragments and individuals. This variation can contribute to false negatives for infection detection. Given variation in sequencing effort across studies, we note that the threshold for infection detection (here we used a minimum of 5 sequence reads) will need to be carefully examined for each investigation. False positives from GBS data seem less likely than false negatives compared to PCR-based methods for infection detection, though contamination of samples is an important consideration for both PCR-based and GBS survey methods. The usefulness of resequencing data for population genetics investigations of endosymbionts will be facilitated by the development of new methods for detecting infection and for genotyping that can, for example, more fully account for uncertainty and accommodate the possibility of multiple infections within individuals. Despite these limitations, the use of resequencing data can cheaply and relatively easily facilitate surveys of endosymbiont infection and population genetics.

Data availability

DNA-sequence data have been deposited in the NCBI SRA with accession codes PRJNA246037, PRJNA577236, PRJNA432816, and PRJNA862870. Scripts for analysis are uploaded to https://github.com/VivaswatS/wolbachia_lycaeides.git.

Supplemental material is available at G3 online.

Acknowledgments

We thank J. Ott and M. Turelli for critical comments and discussion. We thank N. Anthony, G. Gelembiuk, D. Raterman, C. O'Brien, and M. Amaral for collections of *Lycaeides samuelis* under USFWS permit PRT842392. J. Oliver, K. Prudic, J. Jahner, and A. Epeset helped with collections. We also thank Dr Hidenori Tachida and 2 anonymous reviewers for help in improving the manuscript. We also thank members of the University of Chicago population genetics community for providing recommendations on appropriate bioinformatics tools to use. Computing was performed in the Teton Computing Environment at the Advanced Research Computing Center (University of Wyoming, <https://doi.org/10.15786/M2FY47>) and at the LEAP High Performance Computing Cluster at Texas State University.

Funding

This work was supported by National Science Foundation grant DEB-1638793 to MLF, DEB-1638768 to ZG, DEB-1638773 to CCN, DEB-1638922 to JAF, and DEB-1638602 to CAB; MLF was additionally supported by a Trevor James McMinn professorship.

Conflicts of interest

None declared.

Author contributions

VS and CCN conceived and designed the study. All authors collected data. VS analyzed the data. VS and CCN drafted the initial version of the manuscript. All authors contributed to later versions of the manuscript.

Literature cited

- Aktas C. haplotypes: Manipulating DNA Sequences and Estimating Unambiguous Haplotype Network with Statistical Parsimony. R Package Version 1.1.2; 2020. (<https://cran.r-project.org/web/packages/haplotypes/index.html>)
- Arif S, Gerth M, Hone-Millard WG, Nunes MD, Dapporto L, Shreeve TG. Evidence for multiple colonisations and *Wolbachia* infections shaping the genetic structure of the widespread butterfly *Polyommatus icarus* in the British Isles. *Mol Ecol*. 2021;30(20):5196–5213.
- Bailly-Bechet M, Martins-Simões P, SzölloSI GJ, Mialdea G, Sagot M-F, Charlat S. How long does *Wolbachia* remain on board? *Mol Biol Evol*. 2017;34(5):1183–1193.
- Baldo L, Ayoub NA, Hayashi CY, Russell JA, Stahlhut JK, Werren JH. Insight into the routes of *Wolbachia* invasion: high levels of horizontal transfer in the spider genus *Agelenopsis* revealed by *Wolbachia* strain and mitochondrial DNA diversity. *Mol Ecol*. 2008;17(2):557–569.
- Baldo L, Dunning Hotopp JC, Jolley KA, Bordenstein SR, Biber SA, Choudhury RR, Hayashi C, Maiden MCJ, Tettelin H, Werren JH. Multilocus sequence typing system for the endosymbiont *Wolbachia pipientis*. *Appl Environ Microbiol*. 2006;72(11):7098–7110.
- Barton NH, Turelli M. Spatial waves of advance with bistable dynamics: cytoplasmic and genetic analogues of Allee effects. *Am Nat*. 2011;178(3):E48–E75.
- Black S, Vaughan D. Species profile: *Lycaeides melissa samuelis*. In: Black SH, Laursen J, Shepherd M (editors). *Red List of Pollinator Insects of North America*. Portland (OR): The Xerces Society for Invertebrate Conservation; 2005. pp. 300–301.
- Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 2014;10(4):e1003537.
- Brownlie JC, Cass BN, Riegler M, Witsenburg JJ, Iturbe-Ormaetxe I, McGraw EA, O'Neill SL. Evidence for metabolic provisioning by a common invertebrate endosymbiont, *Wolbachia pipientis*, during periods of nutritional stress. *PLoS Pathog*. 2009;5(4):e1000368.
- Charlat S, Hurst GD, Merçot H. Evolutionary consequences of *Wolbachia* infections. *Trends Genet*. 2003;19(4):217–223.
- Chaturvedi S, Lucas LK, Buerkle CA, Fordyce JA, Forister ML, Nice CC, Gompert Z. Recent hybrids recapitulate ancient hybrid outcomes. *Nat Commun*. 2020;11(1):1–15.
- Chaturvedi S, Lucas LK, Nice CC, Fordyce JA, Forister ML, Gompert Z. The predictability of genomic changes underlying a recent host shift in *Melissa* blue butterflies. *Mol Ecol*. 2018;27(12):2651–2666.
- Choi JY, Bunnell JE, Aquadro CF. Population genomics of infectious and integrated *Wolbachia pipientis* genomes in *Drosophila ananassae*. *Genome Biol Evol*. 2015;7(8):2362–2382.
- Conner WR, Blaxter ML, Anfora G, Ometto L, Rota-Stabelli O, Turelli M. Genome comparisons indicate recent transfer of wRi-like *Wolbachia* between sister species *Drosophila suzukii* and *D. subpulchrella*. *Ecol Evol*. 2017;7(22):9391–9404.
- Cooper BS, Vanderpool D, Conner WR, Matute DR, Turelli M. *Wolbachia* acquisition by *Drosophila yakuba*-clade hosts and transfer of incompatibility loci between distantly related *Wolbachia*. *Genetics*. 2019;212(4):1399–1419.
- Crandall M, Clement M, Posada D. TCS: a computer program to estimate gene genealogies. *Mol Ecol*. 2000;9(10):1657–1660.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al.; 1000 Genomes Project Analysis Group. The variant call format and vcfTools. *Bioinformatics*. 2011;27(15):2156–2158.
- Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. 2005;22(5):1185–1192.
- Duron O, Wilkes TE, Hurst GD. Interspecific transmission of a male-killing bacterium on an ecological timescale. *Ecol Lett*. 2010;13(9):1139–1148.
- Forister ML, Gompert Z, Fordyce JA, Nice CC. After 60 years, an answer to the question: what is the karner blue butterfly? *Biol Lett*. 2011;7(3):399–402.
- Forister ML, Philbin CS, Marion ZH, Buerkle CA, Dodson CD, Fordyce JA, Forister GW, Lebeis SL, Lucas LK, Nice CC, et al. Predicting patch occupancy reveals the complexity of host range expansion. *Sci Adv*. 2020;6(48):eabc6852.
- Forister ML, Yoon SA, Philbin CS, Dodson CD, Hart B, Harrison JG, Shelef O, Fordyce JA, Marion ZH, Nice CC, et al. Caterpillars on a phytochemical landscape: the case of alfalfa and the *Melissa* blue butterfly. *Ecol Evol*. 2020;10(10):4362–4374.
- Gehrler L, Vorburger C. Parasitoids as vectors of facultative bacterial endosymbionts in aphids. *Biol Lett*. 2012;8(4):613–615.
- Gerth M, Röthe J, Bleidorn C. Tracing horizontal *Wolbachia* movements among bees (*Anthophila*): a combined approach using multilocus sequence typing data and host phylogeny. *Mol Ecol*. 2013;22(24):6149–6162.

- Gompert Z, Fordyce JA, Forister ML, Nice CC. Recent colonization and radiation of North American *Lycaeides* (*Plebejus*) inferred from mtDNA. *Mol Phylogenet Evol.* 2008;48(2):481–490.
- Gompert Z, Fordyce JA, Forister ML, Shapiro AM, Nice CC. Homoploid hybrid speciation in an extreme habitat. *Science.* 2006;314(5807):1923–1925.
- Gompert Z, Forister ML, Fordyce JA, Nice CC. Widespread mitochondrial discordance with evidence for introgressive hybridization and selective sweeps in *Lycaeides*. *Mol Ecol.* 2008;17(24):5231–5244.
- Gompert Z, Jahner JP, Scholl CF, Wilson JS, Lucas LK, Soria-Carrasco V, Fordyce JA, Nice CC, Buerkle CA, Forister ML, et al. The evolution of novel host use is unlikely to be constrained by trade-offs or a lack of genetic variation. *Mol Ecol.* 2015;24(11):2777–2793.
- Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC. Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Mol Ecol.* 2014;23(18):4555–4573.
- Gompert Z, Lucas LK, Nice CC, Fordyce JA, Alex Buerkle C, Forister ML. Geographically multifarious phenotypic divergence during speciation. *Ecol Evol.* 2013;3(3):595–613.
- Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution.* 2012;66(7):2167–2181.
- Gompert Z, Nice CC, Fordyce JA, Forister ML, Shapiro AM. Identifying units for conservation using molecular systematics: the cautionary tale of the Karner blue butterfly. *Mol Ecol.* 2006;15(7):1759–1768.
- Gupta M, Kaur R, Gupta A, Raychoudhury R. Are ecological communities the seat of endosymbiont horizontal transfer and diversification? A case study with soil arthropod community. *Authorea Preprints, Ecol Evol.* 2021;11(21):14490–14508.
- Hague MTJ, Shropshire JD, Caldwell CN, Statz JP, Stanek KA, Conner WR, Cooper BS. Temperature effects on cellular host-microbe interactions explain continent-wide endosymbiont prevalences. *Curr Biol.* 2022;32(4):878–888.
- Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 1985;22(2):160–174.
- Heath BD, Butcher RD, Whitfield WG, Hubbard SF. Horizontal transfer of *Wolbachia* between phylogenetically distant insect species by a naturally occurring mechanism. *Curr Biol.* 1999;9(6):313–316.
- Hedges LM, Brownlie JC, O'Neill SL, Johnson KN. *Wolbachia* and virus protection in insects. *Science.* 2008;322(5902):702–702.
- Hertig M. The rickettsia, *Wolbachia pipiensis* (gen. et sp. n.) and associated inclusions of the mosquito, *Culex pipiens*. *Parasitology.* 1936;28(4):453–486.
- Hertig M, Wolbach SB. Studies on rickettsia-like micro-organisms in insects. *J Med Res.* 1924;44(3):329–374.7.
- Hilgenboecker K, Hammerstein P, Schlattmann P, Telschow A, Werren JH. How many species are infected with *Wolbachia*? – a statistical analysis of current data. *FEMS Microbiol Lett.* 2008;281(2):215–220.
- Hinojosa JC, Koubínová D, Dincă V, Hernández-Roldán J, Munguira ML, García-Barros E, Vila M, Alvarez N, Mutanen M, Vila R. Rapid colour shift by reproductive character displacement in cupido butterflies. *Mol Ecol.* 2020;29(24):4942–4955.
- Hoffmann A. *Wolbachia*. *Curr Biol.* 2020;30(19):R1113–R1114.
- Hoffmann AA, Montgomery BL, Popovici J, Iturbe-Ormaetxe I, Johnson PH, Muzzi F, Greenfield M, Durkan M, Leong YS, Dong Y, et al. Successful establishment of *Wolbachia* in *Aedes* populations to suppress dengue transmission. *Nature.* 2011;476(7361):454–457.
- Hosokawa T, Koga R, Kikuchi Y, Meng XY, Fukatsu T. *Wolbachia* as a bacteriocyte-associated nutritional mutualist. *Proc Natl Acad Sci USA.* 2010;107(2):769–774.
- Hoy MA, Jeyaparakash A. Microbial diversity in the predatory mite *Metaseiulus occidentalis* (acari: Phytoseiidae) and its prey, *Tetranychus urticae* (acari: Tetranychidae). *Biol Control.* 2005;32(3):427–441.
- Huigens M, De Almeida R, Boons P, Luck R, Stouthamer R. Natural interspecific and intraspecific horizontal transfer of parthenogenesis-inducing *Wolbachia* in *Trichogramma* wasps. *Proc Biol Sci.* 2004;271(1538):509–515.
- Huigens M, Luck R, Klaassen R, Maas M, Timmermans M, Stouthamer R. Infectious parthenogenesis. *Nature.* 2000;405(6783):178–179.
- Hurst GD, Jiggins FM. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proc Biol Sci.* 2005;272(1572):1525–1534.
- Iturbe-Ormaetxe I, Walker T, O'Neill SL. *Wolbachia* and the biological control of mosquito-borne disease. *EMBO Rep.* 2011;12(6):508–518.
- Jaenike J. Coupled population dynamics of endosymbionts within and between hosts. *Oikos.* 2009;118(3):353–362.
- Jiggins FM. Male-killing *Wolbachia* and mitochondrial DNA: selective sweeps, hybrid introgression and parasite population dynamics. *Genetics.* 2003;164(1):5–12.
- Kambris Z, Cook PE, Phuc HK, Sinkins SP. Immune activation by life-shortening *Wolbachia* and reduced filarial competence in mosquitoes. *Science.* 2009;326(5949):134–136.
- Kriesner P, Conner WR, Weeks AR, Turelli M, Hoffmann AA. Persistence of a *Wolbachia* infection frequency cline in *Drosophila melanogaster* and the possible role of reproductive dormancy. *Evolution.* 2016;70(5):979–997.
- Kriesner P, Hoffmann AA, Lee SF, Turelli M, Weeks AR. Rapid sequential spread of two *Wolbachia* variants in *Drosophila simulans*. *PLoS Pathog.* 2013;9(9):e1003607.
- Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods.* 2012;9(4):357–359.
- Le Clec'h W, Chevalier FD, Genty L, Bertaux J, Bouchon D, Sicard M. Cannibalism and predation as paths for horizontal passage of *Wolbachia* between terrestrial isopods. *PLoS One.* 2013;8(4):e60232.
- Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics.* 2011;27(21):2987–2993.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics.* 2009;25(16):2078–2079.
- Mavingui P, Moro CV, Tran-Van V. 2012. Whole-genome sequence of *Wolbachia* strain wAlbB, an endosymbiont of tiger mosquito vector *Aedes albopictus*. *J Bacteriol.* 2012;194(7):1840–1840.
- Moran NA, McCutcheon JP, Nakabachi A. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet.* 2008;42:165–190.
- Narita S, Nomura M, Kato Y, Fukatsu T. Genetic structure of sibling butterfly species affected by *Wolbachia* infection sweep: evolutionary and biogeographical implications. *Mol Ecol.* 2006;15(4):1095–1108.
- Nice CC, Anthony N, Gelembiuk G, Raterman D, Ffrench-Constant R. The history and geography of diversification within the butterfly genus *Lycaeides* in North America. *Mol Ecol.* 2005;14(6):1741–1754.

- Nice CC, Gompert Z, Fordyce JA, Forister ML, Lucas LK, Buerkle CA. Hybrid speciation and independent evolution in lineages of alpine butterflies. *Evolution*. 2013;67(4):1055–1068.
- Nice CC, Gompert Z, Forister ML, Fordyce JA. An unseen foe in arthropod conservation efforts: the case of *Wolbachia* infections in the Karner blue butterfly. *Biol Conserv*. 2009;142(12):3137–3146.
- Nikoh N, Hosokawa T, Moriyama M, Oshima K, Hattori M, Fukatsu T. Evolutionary origin of insect–*Wolbachia* nutritional mutualism. *Proc Natl Acad Sci USA*. 2014;111(28):10257–10262.
- Nikoh N, Tanaka K, Shibata F, Kondo N, Hizume M, Shimada M, Fukatsu T. *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. *Genome Res*. 2008;18(2):272–280.
- Nunes MD, Nolte V, Schlötterer C. Nonrandom *Wolbachia* infection status of *Drosophila melanogaster* strains with different mtDNA haplotypes. *Mol Biol Evol*. 2008;25(11):2493–2498.
- Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691–3693.
- Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA. Genome-wide association genetics of an adaptive trait in lodgepole pine. *Mol Ecol*. 2012;21(12):2991–3005.
- Pascari J, Chandler CH. A bioinformatics approach to identifying *Wolbachia* infections in arthropods. *PeerJ*. 2018;6:e5486.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna (Austria): R Foundation for Statistical Computing; 2021.
- Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst Biol*. 2018;67(5):901–904.
- Ramírez-Puebla ST, Servín-Garcidueñas LE, Ormeño-Orrillo E, Vera-Ponce de León A, Rosenblueth M, Delaye L, Martínez J, Martínez-Romero E. Species in *Wolbachia*? Proposal for the designation of ‘*candidatus Wolbachia bourtzisii*’, ‘*candidatus Wolbachia onchocercicola*’, ‘*candidatus Wolbachia blaxteri*’, ‘*candidatus Wolbachia brugii*’, ‘*candidatus Wolbachia taylora*’, ‘*candidatus Wolbachia collembolicola*’ and ‘*candidatus Wolbachia multihospitum*’ for the different species within *Wolbachia* supergroups. *Syst Appl Microbiol*. 2015;38(6):390–399.
- Raychoudhury R, Baldo L, Oliveira DC, Werren JH. Modes of acquisition of *Wolbachia*: horizontal transfer, hybrid introgression, and codivergence in the *Nasonia* species complex. *Evolution*. 2009;63(1):165–183.
- Raychoudhury R, Grillenberger BK, Gadau J, Bijlsma R, van de Zande L, Werren JH, Beukeboom LW. Phylogeography of *Nasonia vitripennis* (Hymenoptera) indicates a mitochondrial–*Wolbachia* sweep in North America. *Heredity*. 2010;104(3):318–326.
- Richardson MF, Weinert LA, Welch JJ, Linheiro RS, Magwire MM, Jiggins FM, Bergman CM. Population genomics of the *Wolbachia* endosymbiont in *Drosophila melanogaster*. *PLoS Genet*. 2012;8(12):e1003129.
- Riegler M, Sidhu M, Miller WJ, O’Neill SL. Evidence for a global *Wolbachia* replacement in *Drosophila melanogaster*. *Curr Biol*. 2005;15(15):1428–1433.
- Ross PA, Turelli M, Hoffmann AA. Evolutionary ecology of *Wolbachia* releases for disease control. *Annu Rev Genet*. 2019;53:93–116.
- Sanaei E, Charlat S, Engelstädter J. *Wolbachia* host shifts: routes, mechanisms, constraints and evolutionary consequences. *Biol Rev Camb Philos Soc*. 2021;96(2):433–453.
- Scholz M, Albanese D, Tuohy K, Donati C, Segata N, Rota-Stabelli O. Large scale genome reconstructions illuminate *Wolbachia* evolution. *Nat Commun*. 2020;11(1):1–11.
- Schuler H, Egan SP, Hood GR, Busbee RW, Driscoe AL, Ott JR. Diversity and distribution of *Wolbachia* in relation to geography, host plant affiliation and life cycle of a heterogonic gall wasp. *BMC Evol Biol*. 2018;18(1):1–15.
- Schuler H, Köppler K, Daxböck-Horvath S, Rasool B, Krumböck S, Schwarz D, Hoffmeister TS, Schlick-Steiner BC, Steiner FM, Telschow A, et al. The hitchhiker’s guide to Europe: the infection dynamics of an ongoing *Wolbachia* invasion and mitochondrial selective sweep in *Rhagoletis cerasi*. *Mol Ecol*. 2016;25(7):1595–1609.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068–2069.
- Shoemaker DD, Ahrens M, Sheill L, Mescher M, Keller L, Ross KG. Distribution and prevalence of *Wolbachia* infections in native populations of the fire ant *Solenopsis invicta* (Hymenoptera: Formicidae). *Environ Entomol*. 2003;32(6):1329–1336.
- Shoemaker DD, Keller G, Ross KG. Effects of *Wolbachia* on mtDNA variation in two fire ant species. *Mol Ecol*. 2003;12(7):1757–1771.
- Shoemaker DD, Ross KG, Keller L, Vargo E, Werren JH. *Wolbachia* infections in native and introduced populations of fire ants (*Solenopsis* spp.). *Insect Mol Biol*. 2000;9(6):661–673.
- Shropshire JD, Hamant E, Cooper BS. Male age and *Wolbachia* dynamics: investigating how fast and why bacterial densities and cytoplasmic incompatibility strengths vary. *mBio*. 2021;12(6):e0299821.
- Signor S. Population genomics of *Wolbachia* and mtDNA in *Drosophila simulans* from California. *Sci Rep*. 2017;7(1):1–11.
- Stevens L, Giordano R, Fialho RF. Male-killing, nematode infections, bacteriophage infection, and virulence of cytoplasmic bacteria in the genus *Wolbachia*. *Annu Rev Ecol Syst*. 2001;32(1):519–545.
- Tange O. GNU parallel – the command-line power tool. *USENIX Mag*. 2011;36:42–47.
- Teixeira L, Ferreira Á, Ashburner M. The bacterial symbiont *Wolbachia* induces resistance to RNA viral infections in *Drosophila melanogaster*. *PLoS Biol*. 2008;6(12):e1000002.
- Templeton AR, Crandall KA, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*. 1992;132(2):619–633.
- Tennekes M. tmap: thematic maps in R. *J Stat Softw*. 2018;84(6):1–39.
- Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci USA*. 2005;102(39):13950–13955.
- Turelli M. Evolution of incompatibility-inducing microbes and their hosts. *Evolution*. 1994;48(5):1500–1513.
- Turelli M, Cooper BS, Richardson KM, Ginsberg PS, Peckenpaugh B, Antelope CX, Kim KJ, May MR, Abrieux A, Wilson DA, et al. Rapid global spread of wRi-like *Wolbachia* across multiple *Drosophila*. *Curr Biol*. 2018;28(6):963–971.
- Turelli M, Hoffmann AA. Rapid spread of an inherited incompatibility factor in California *Drosophila*. *Nature*. 1991;353(6343):440–442.
- Turelli M, Hoffmann AA. Cytoplasmic incompatibility in *Drosophila simulans*: dynamics and parameter estimates from natural populations. *Genetics*. 1995;140(4):1319–1338.
- Turelli M, Hoffmann A, McKechnie SW. Dynamics of cytoplasmic incompatibility and mtDNA variation in natural *Drosophila simulans* populations. *Genetics*. 1992;132(3):713–723.
- Unckless RL, Boelio LM, Herren JK, Jaenike J. *Wolbachia* as populations within individual insects: causes and consequences of density variation in natural populations. *Proc Biol Sci*. 2009;276(1668):2805–2811.

- Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol.* 2015;23:148–154.
- Vila R, Bell CD, Macniven R, Goldman-Huertas B, Ree RH, Marshall CR, Bálint Z, Johnson K, Benyamini D, Pierce NE, et al. Phylogeny and palaeoecology of *Polyommatus* blue butterflies show Beringia was a climate-regulated gateway to the New World. *Proc Biol Sci.* 2011;278(1719):2737–2744.
- Walker T, Moreira LA. Can *Wolbachia* be used to control malaria? *Mem Inst Oswaldo Cruz.* 2011;106(Suppl 1):212–217.
- Walker T, Quek S, Jeffries CL, Bandibabone J, Dhokiya V, Bamou R, Kristan M, Messenger LA, Gidley A, Hornett EA, et al. Stable high-density and maternally inherited *Wolbachia* infections in *Anopheles moucheti* and *Anopheles demeilloni* mosquitoes. *Curr Biol.* 2021;31(11):2310–2320.
- Weeks AR, Turelli M, Harcombe WR, Reynolds KT, Hoffmann AA. From parasite to mutualist: rapid evolution of *Wolbachia* in natural populations of *Drosophila*. *PLoS Biol.* 2007;5(5):e114.
- Weinert LA, Araujo-Jnr EV, Ahmed MZ, Welch JJ. The incidence of bacterial endosymbionts in terrestrial arthropods. *Proc Biol Sci.* 2015;282(1807):20150249.
- Werren JH, Baldo L, Clark ME. *Wolbachia*: master manipulators of invertebrate biology. *Nat Rev Microbiol.* 2008;6(10):741–751.
- Werren JH, Windsor DM. *Wolbachia* infection frequencies in insects: evidence of a global equilibrium? *Proc Biol Sci.* 2000;267(1450):1277–1285.
- Werren JH, Zhang W, Guo LR. Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proc Biol Sci.* 1995;261(1360):55–63.
- Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, McGraw EA, Martin W, Esser C, Ahmadinejad N, et al. Phylogenomics of the reproductive parasite *Wolbachia pipientis wMel*: a streamlined genome overrun by mobile genetic elements. *PLoS Biol.* 2004;2(3):e69.
- Yen JH, Barr AR. New hypothesis of the cause of cytoplasmic incompatibility in *Culex pipiens* L. *Nature.* 1971;232(5313):657–658.
- Zug R, Hammerstein P. Bad guys turned nice? A critical assessment of *Wolbachia* mutualisms in arthropod hosts. *Biol Rev Camb Philos Soc.* 2015;90(1):89–111.

Communicating editor: H. Tachida