Gene regulation and genome evolution in *Wolbachia pipientis*

Preston Basting
October 2019

Major Professor: Casey Bergman

Committee Member: Jan Mrazek
Committee Member: Ying Xu
Committee Member: Sidney Kushner

Important information for the Committee:

Instructions for written exam evaluation:
Each committee member has two weeks to review and submit a grade of pass or fail via email to the student and the Graduate Program Administrator (iobgradadmin@UGA.EDU). To pass the written portion and go on to the oral portion, the student must receive no more than one dissenting (failing grade) vote. If a committee member does not provide a grade two weeks after submission of the written exam, the grade will be marked as a pass for that committee member. The written exam takes the form of an NIH grant proposal; if you are not familiar with the IOB written exam requirements, see Appendix B in the most recent graduate student handbook (https://iob.uga.edu/graduate-program/graduate-handbooks/).

Instructions for oral exam:
The oral exam will last at least two hours, but not longer than three hours. The student will prepare a presentation of no more than 20 slides that are intended to serve as a framework of the discussion of the proposed research. The student's presentation should last for approximately 20-25 minutes without interruptions, followed by questions from the advisory committee and other faculty present. Questions during the exam will test both general and specific knowledge related to the student's proposed research as described in their presentation and written proposal. A member of the student's committee, other than the advisor, will serve as chair of the exam. The advisor is not allowed to answer questions for the student, and will not participate in the discussion unless granted permission by the exam chair.

Specific Aims

Wolbachia pipientis, a bacterial endosymbiont that infects up to 40% of arthropod species and some nematode species, has become a promising tool for biocontrol of disease vectors and invasive insect species [1, 2]. *Wolbachia* are capable of inducing a wide range of effects on hosts to promote transmission to the host offspring. Many strains manipulate host reproduction by inducing feminization, male killing, parthenogenesis, or cytoplasmic incompatibility (CI) to increase the proportion of infected females in the population and promote *Wolbachia* spread to offspring [3]. Some *Wolbachia* strains are also capable of reducing viral replication in hosts, effectively improving host viral resistance [4]. These interactions provide multiple paths in which *Wolbachia* can be used to control insect populations, such as use as a biopesticide for invasive insect populations or increasing insect resistance to viral diseases to decrease their spread to humans. While these host phenotypes have many potential applications, the underlying mechanisms by which they are induced are poorly understood.

Unique features of the *Wolbachia* genome likely point to mechanisms involved in symbiont-host interactions. Previous studies have demonstrated that *Wolbachia* genomes commonly exhibit an overabundance of repetitive DNA relative to other endosymbionts, which typically have streamlined genomes. Repetitive DNA is a major driver of genome evolution and specific repetitive elements, such as ankyrin repeats, are known to be involved in host-pathogen interactions [5]. Further characterization of repetitive DNA in *Wolbachia* is needed to understand why they are preserved in the genome.

Wolbachia are capable of dynamic gene expression in response to changes in host sex and tissue type, suggesting that the dynamically expressed genes play a role in symbiont-host interaction [6]. The mechanisms that underlie *Wolbachia* gene regulation have been poorly characterized to date. Many *Wolbachia* strains lack operon models, and the current operon models have not been experimentally verified. There are also few studies in *Wolbachia* that examine alternative mechanisms of gene regulation such as small RNAs and antisense transcription. In this study, I seek to improve characterization of *Wolbachia* genome features such as repetitive elements, operon models, non-coding RNAs, and antisense transcription to provide an enhanced genomic framework for future studies into *Wolbachia*-host interactions.
I will achieve this objective through the completion of the following specific aims:

Aim 1: Validate and improve *Wolbachia* operon predictions. I hypothesize that current gene and operon model predictions do not accurately represent the larger *Wolbachia* transcriptional context. I will validate and improve operon predictions by using RNAseq data to reconstruct the transcripts expressed by *Wolbachia* and comparing these transcripts to current sequence-based operon predictions. Highly-supported operons in *Wolbachia* will provide information about operon structure that will be leveraged to accurately predict operons in previously uncharacterized *Wolbachia* strains lacking RNAseq data. Accurate operon models are essential to understanding *Wolbachia* gene regulation and genome evolution and can also be utilized to improve the annotation of genes with no known function.
Aim 2: Explore patterns in *Wolbachia* repetitive DNA evolution. I hypothesize that *Wolbachia* strains maintain abundant repetitive DNA in the genome as they provide a evolutionary benefit and possibly play a role in interactions with hosts. To test this, I will use repeat annotation tools to identify and classify repetitive DNA in complete genome assemblies of *Wolbachia* strains as well as in WGS data for the *Wolbachia* strain wMel. This will reveal how different classes of repetitive DNA, such as TEs, Group II introns, and ankyrin repeats, are changing in the short-term and long-term. Implications about the function of repetitive DNA can be revealed through evolutionary patterns in repetitive DNA content across the *Wolbachia* phylogeny.
Aim 3: Determine the biological significance of short RNAs in *Wolbachia* I hypothesize that antisense transcription is involved in gene regulation in *Wolbachia* through targeted degradation of double-stranded RNA and that short RNAs (<50 nt) are a byproduct of this process. Preliminary evidence suggests that antisense transcription is pervasive in the *Wolbachia* genome and that distinctive patterns of short RNA formation exists in *Wolbachia*. I will test my hypothesis by identifying regions of antisense transcription and determine if short RNA formation is correlated with antisense transcription or other genomic features, including repeats. Short RNAs could represent the products of double-stranded RNA degradation by RNase III, which could be a mechanism by which *Wolbachia* regulate gene expression. As *Wolbachia* have few transcription factors, antisense transcription could play a role in dynamically regulating genes important for endosymbiont-host interactions.

Significance

Wolbachia-host interactions are a popular area of research due to the potential applications in control and manipulation of insect populations. Studies are underway to infect *Aedes* mosquito populations with strains of *Wolbachia* that increase viral resistance [2, 7, 8, 9]. These studies aim to reduce the transmission of viral diseases like Dengue fever in humans by manipulating the insect vectors. *Wolbachia* are also capable of inducing a diverse range of reproductive manipulations on their hosts [3]. Researchers have also proposed taking advantage of *Wolbachia*-induced CI to inhibit reproduction in insect populations that may act as disease vectors or are harmful to crops [3, 10]. As the molecular mechanisms involved in these host manipulations are largely uncharacterized, efforts to use *Wolbachia* as a tool for biocontrol would benefit greatly from an improved understanding of *Wolbachia* genomic features involved in host-symbiont interactions.

A vital element of characterizing gene regulation in *Wolbachia* is accurate operon models. Accurate operon models are essential to characterization of regulatory networks and could help determine pathways involved in interactions with hosts. The *Wolbachia* genes *cifA* and *cifB* have been previously identified as candidate effectors of CI [11]. Thus, regulation of these genes is important to understanding how *Wolbachia* induces CI in hosts. The genes *cifA* and *cifB* are proposed to be components of a toxin-antitoxin operon, though conflicting experimental evidence and a lack of highly supported operon models make this difficult to confidently establish [11, 12]. This demonstrates that there is a need for accurate operon predictions, and that operon models can be leveraged to characterize pathways involved in host-*Wolbachia* interactions. By utilizing the extensive RNAseq data available for the *Wolbachia* strain *w*Mel, I will be able to establish operons supported by experimental data which can also be used to improve operon predictions of other *Wolbachia* strains.

Wolbachia genomes are reduced in size (1.08Mb-1.7Mb) as a consequence of adaptation to the host environment [3, 13, 14]. However, their genomes are not as streamlined as those of other obligate endosymbionts such as *Buchnera* and *Candidatus* species [3]. A common feature of *Wolbachia* genomes is an enrichment of repetitive DNA, which is atypical of most obligate endosymbionts [13, 15, 16, 17]. For example, 14% of the genome of the *Wolbachia* strain *w*Mel is repetitive, much of which consists of mobile genetic elements and tandem repeats [13]. The maintenance of repeats in an otherwise streamlined genome suggests that repetitive DNA may have a biological role in *Wolbachia* that supersedes the drive to lose genetic redundancy, as seen in other obligate endosymbionts. Previous researchers have proposed a few mechanisms by which this repetitive DNA could be beneficial to *Wolbachia*, such as facilitating genome evolution and interactions with the host via ankyrin repeats in *Wolbachia* proteins [5, 16]. However, the type and extent of repetitive DNA varies among *Wolbachia* genomes [5, 13, 14, 16]. It is also unknown how widespread this phenomenon is among *Wolbachia* strains, and whether or not it is specific to certain host groups. In order to draw accurate conclusions about the biological role of this repeat enrichment, context is needed to understand how and why repetitive element content varies among *Wolbachia* strains and how much repeat fluctuation is occurring in the short term.

Likely a consequence of genome reduction, few transcription factors have been discovered in the *Wolbachia* genome [13]. Despite this, *Wolbachia* are capable of dynamic gene regulation in response to changes in host sex and tissue [6]. As the predominant host phenotype manipulations induced by *Wolbachia* are sex or tissue specific, further characterization of *Wolbachia* gene regulation could reveal mechanisms involved in *Wolbachia*-host interactions. Little is known about how *Wolbachia* regulate gene expression. In other bacteria such as *E. coli*, cis- and trans-encoded antisense RNAs are highly prevalent and act as a mechanism to regulate gene expression [18, 19]. While some trans-encoded small RNAs have been discovered in *Wolbachia* much more work is needed to understand the extent of antisense RNAs and their role in gene regulation [20]. Preliminary evidence suggests that cis-antisense transcripts are extensively prevalent in the *Wolbachia* transcriptome. As cis-antisense RNAs have been demonstrated to impact transcript abundance in other bacteria, it is possible that the extensive antisense transcription is involved in regulating gene expression in *Wolbachia* [21, 22].

Innovation

My work will determine the role of short RNAs and antisense transcription, a potential gene regulation system that has not been characterized in *Wolbachia* to date. I will also apply a number of innovative bioinformatics approaches to improve operon predictions and repetitive DNA annotations in *Wolbachia* strains. Specifically, this

proposal seeks to achieve these advancements:

- Developing a new method for predicting operons in *Wolbachia* based on features from RNAseq-supported transcriptional units. (Aim 1)
- Establish a pipeline of repeat annotation tools to automate accurate prediction of repetitive elements in *Wolbachia* genome assemblies. (Aim 2)
- Exploring patterns of antisense transcription and short RNAs in *Wolbachia* to determine if these patterns are indicative of a gene regulation mechanism previously uncharacterized in this organism. (Aim 3)

Approach

Aim 1: Validate and improve *Wolbachia* operon predictions

Introduction:

The objective of this aim is to improve operon models for *Wolbachia* strains by evaluating and updating current operon predictions using comparative genomics and RNAseq data. Many *Wolbachia* strains currently lack operon predictions and I hypothesize that current operon predictions are inaccurate as they are based on classifiers trained using operon features from free-living bacteria. My approach will be to use RNAseq data from *Wolbachia* strain *w*Mel to identify co-expressed genes with transcriptional unit predictions and transcript assembly which will represent experimentally validated operons. I will also use comparative genomics to confirm operon predictions by examining gene neighborhood conservation across *Wolbachia* genomes. Highly supported *Wolbachia* operons will then be used to establish parameters for prediction of operons in *Wolbachia* strains lacking operon models. The rationale being that improving operon models will be useful for functional studies of *Wolbachia* and will aid in the classification of hypothetical proteins, which are abundant in current *Wolbachia* gene annotations.

Justification and Feasibility:

Operons are sets of genes that have been predicted or experimentally verified to be transcribed into a single mRNA molecule. In bacteria, genes belonging to the same operon are typically regulated by a common promoter, thus their regulation is linked. Typically, genes from the same operon function within the same biological pathway. This organization acts as a simplified mechanism to co-regulate expression of genes within metabolic or functional pathways. Determining operon structures can provide a wide-range of information about gene regulatory networks in an organism and can provide clues as to the function of previously uncharacterized genes. This has lead to a need for methods to accurately predict operons.

Microbial operons can often be predicted based upon the gene organization in the genome. Co-transcribed gene pairs often have reduced intergenic distances, are not separated by promoters or terminators, and are located adjacent on the same strand [23, 24]. As genes from the same operon are typically functionally related, their organization is often evolutionarily conserved. The conservation of gene neighborhoods when comparing across many genomes is another source of evidence of operon structure.

There are a number of existing tools that attempt to classify gene-pairs into operons based upon features such as gene-pair organization and conservation [23, 24, 25, 26]. As few organisms have extensive experimentally-verified operon information, these tools are typically optimized for prediction of operons in well-studied organisms like *E. coli*, and *B. subtilis*. Consequently, it is difficult to establish how well these tools generalize to microbes with significantly different genome organization patterns and much fewer experimentally-confirmed operons, such as with obligate endosymbionts. Also, the results of these different tools can conflict with each other, making it challenging to determine which operon predictions are most reflective of the actual co-transcribed genes.

Many approaches can be applied to experimentally verify operons, such as RT-PCR, primer extension, and Northern blotting [27]. As these methods are time consuming and expensive when performed on a large scale, they are not ideal for identification or verification of all operons in a genome. Advances in RNA sequencing have made this a more attractive method for characterizing operons across a genome than traditional experimental approaches. A number of methods have been developed to utilize RNAseq data to identify operons based on gene expression patterns [28, 29]. These tools compare the expression across neighboring genes and expression in the intergenic region between genes to identify co-transcribed genes. While expression patterns from RNAseq can be highly predictive of co-transcribed genes, transcription of operons is dynamic and can change depending on the environment. Methods that rely solely on RNAseq data for predictions can only predict the operons expressed in the testing condition that was sequenced.

Table 1: *Wolbachia* operon predictions available in operon databases. The first column is the *Wolbachia* strain designation. The second column is the host that the *Wolbachia* strain infects. The third column is the GenBank Assembly accession. Columns four through seven are operon databases. Dots indicate that operon predictions for that particular *Wolbachia* strain can be found in the database.

Strain	Host	Genbank Accession	DOOR	ProOpDB	MicrobesOnline	OperonDB
wRi	*Drosophila simulans*	GCA_000022285.1	•	•	•	•
wAlbB	*Aedes albopictus*	GCA_004171285.1				
wNo	*Drosophila simulans*	GCA_000376585.1		•		
wHa	*Drosophila simulans*	GCA_000376605.1		•		
wAu	*Drosophila simulans*	GCA_000953315.1				
Cameroon	*Onchocerca volvulus*	GCA_000530755.1				
wMel	*Drosophila melanogaster*	GCA_000008025.1	•	•	•	•
TRS	*Brugia malayi*	GCA_000008385.1	•	•	•	•
wCle	*Cimex lectularius*	GCA_000829315.1				
wOo	*Onchocerca ochengi*	GCA_000306885.1	•	•		
Berlin	*Folsomia candida*	GCA_001931755.2				
wPip	*Culex quinquefasciatus*	GCA_000073005.1	•	•	•	•
China1	*Bemisia tabaci*	GCA_003999585.1				
wInc_Cu	*Drosophila incompta*	GCA_001758565.1				
wInc_SM	*Drosophila incompta*	GCA_001758585.1				
wTpre	*Trichogramma pretiosum*	GCA_001439985.1				
wPpe	*Pratylenchus penetrans*	GCA_001752665.1				
wPip2	*Culex quinquefasciatus*	GCA_000156735.1			•	
wMelPop	*Drosophila melanogaster*	GCA_000475015.1				
wRec	*Drosophila recens*	GCA_000742435.1				
TSC140300811.24	*Drosophila willistoni*	GCA_000153585.1			•	
wFex	*Formica exsecta*	GCA_003704235.1				
wAus	*Plutella Australiana*	GCA_002318985.1				
wStri	*Laodelphax striatella*	GCA_001637495.1				
wWb	*Wuchereria bancrofti*	GCA_002204235.2				
wPip_Mol	*Culex molestus*	GCA_000723225.2				
wACP3	*Diaphorina citri*	GCA_000331595.1				
wSpc	*Drosophila subpulchrella*	GCA_002300525.1				
valsugana	*Drosophila suzukii*	GCA_000333795.2				
Ob_Wba	*Operophtera brumata*	GCA_001266585.1				
wAna_india	*Drosophila ananassae*	GCA_003671365.1			•	
wUni	*Muscidifurax uniraptor*	GCA_001983635.1				
FL2016	*Aedes albopictus*	GCA_002379145.1				
wVitA	*Nasonia vitripennis*	GCA_001983615.1				
wBol1-b	*Hypolimnas bolina*	GCA_000333775.1				
wDacA	*Dactylopius coccus*	GCA_001648025.1				
wNfla	*Nomada flava*	GCA_001675695.1				
HN2016	*Aedes albopictus*	GCA_002374845.1				
wNleu	*Nomada leucophthalma*	GCA_001675715.1				
wNpa	*Nomada panzeri*	GCA_001675775.1				
wDacB	*Dactylopius coccus*	GCA_001648015.1				
wGmm	*Glossina morsitans*	GCA_000689175.1				
wNfe	*Nomada ferruginata*	GCA_001675785.1				
wCon	*Cylisticus convexus*	GCA_003344345.1				
wBTv1	*Bemisia tabaci*	GCA_900097055.1				
Cameroon2	*Onchocerca volvulus*	GCA_000338375.1				
wVitB	*Nasonia vitripennis*	GCA_000204545.1				

Wolbachia strains are not well represented in the operon predictions found on popular operon databases. Most of the *Wolbachia* strain assemblies available in NCBI do not have operon predictions available on DOOR, ProOpDB, MicrobesOnline or OperonDB (Table 1)[30, 31, 32, 23]. Even when operon predictions are available, there are significant differences seen between operon prediction methods. For example, 77% of operon predictions for *w*Mel from DOOR match operon predictions from ProOpDB, while only 8% of DOOR predictions for *w*Mel match predictions from OperonDB. Little effort has been made to comprehensively verify existing operon predictions for *Wolbachia*. Together, the discordance between the available operon predictions and the lack of verified operons makes it difficult to determine which operon models are the most accurate. This indicates a need for verification of operon predictions so that appropriate operon prediction methods can be applied to *Wolbachia* strains currently lacking operon models.

Currently, there are 11 complete assemblies of *Wolbachia* strains from a diverse range of arthropod and nematode hosts. Preliminary evidence suggests that the genomes of these strains are highly rearranged, with large differences in gene order seen even between closely related strains. When comparing the genomes of the strain *w*Mel and Berlin, there are extensive genomic rearrangements, causing their synteny plots to look scattered

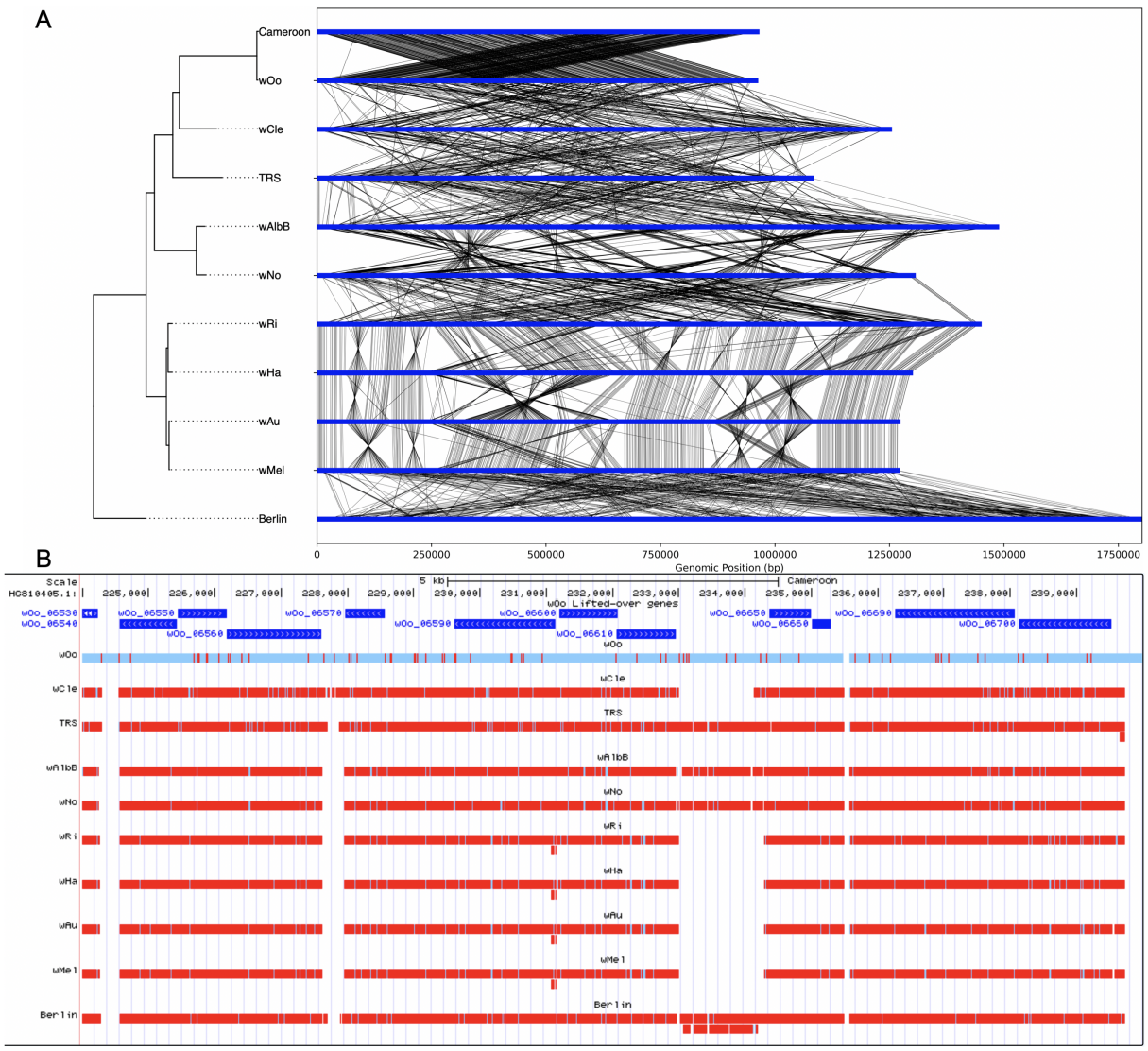


Figure 1: ***Wolbachia*** **genomes show conserved gene neighborhoods despite extensive structural rearrangements.** (A) Synteny plots of complete *Wolbachia* assemblies arranged by phylogenetic relationships. Horizontal blue bars represent the strain genome and small black lines represent links between homologous genes. Extensive genomic rearrangements are seen even between closely related strains. (B) Comparative assembly hub showing *Wolbachia* strain genome assemblies aligned to the Cameroon strain assembly. Red/light blue tracks represent contiguous regions of homology for each strain. White gaps between homologous regions represent breaks in the contiguity. Dark blue track shows *Wolbachia* gene model. Contiguous regions of homology show patterns of conservation of operon structures across *Wolbachia* genomes.

(Fig. 1A). Even with these extensive rearrangements, patterns of conserved gene-pairs and gene-neighborhoods can be seen in assembly alignments. Indeed, alignment of the *w*Mel and Berlin strains to a common reference (Cameroon) reveals that numerous gene neighborhoods are consistently contiguous in both strains despite their vastly differing gene structure (Fig. 1B). As gene-pairs belonging to the same operon are often conserved, it is likely that I can use comparative genomics with the complete *Wolbachia* genome assemblies to determine conserved neighborhoods which can support operon predictions.

Alongside conserved neighborhoods, RNAseq data can also be used in *Wolbachia* to establish operon models. The modENCODE RNAseq dataset for the *Wolbachia*-infected *D. melanogaster* sub-strain BDSC ISO1 [33, 34, 35] contains an abundance of *Wolbachia* reads which have been used previously to study *Wolbachia* gene expression [6]. This dataset contains RNAseq samples from 30 time points across the *D. melanogaster* life cycle, 24 of which have two biological replicates. The BDSC ISO1 *Wolbachia* infection has been identified as a variant of *w*Mel [6]. This extensive dataset, in conjunction with operon prediction tools that utilize expression information, should provide ample evidence to confidently identify co-transcribed genes in the *Wolbachia* strain *w*Mel.

Unfortunately, most *Wolbachia* strains do not have RNAseq data available, so RNAseq-based operon predictions cannot be applied broadly in *Wolbachia* without the generation of much more RNAseq datasets. As this would be an expensive and time-consuming process, I instead propose that the RNAseq data and conserved neighborhood information can be used to establish highly supported operons in the strain *w*Mel, which can be leveraged to predict operons in other *Wolbachia* strains. Genomic features of these highly supported operons, can be extracted and used to train an operon classifier specifically for *Wolbachia* genomes. This classifier can then be used to predict operons in *Wolbachia* strains where external expression data like RNAseq is not available. I expect the *w*Mel-trained classifier to perform better than current operon classifiers that are trained on data from distantly related species like *E. coli*, and *B. subtilis*.

Research Design:

Activity 1-1: Use modENCODE RNAseq data to predict operon structure in *w*Mel. To establish highly supported operon models for the *Wolbachia* strain *w*Mel, I will use the modENCODE RNAseq data from the *w*Mel-infected *D. melanogaster*strain ISO1 [33, 34, 35] to predict operons based on gene expression patterns. As the modENCODE data constrains 30 different samples, 24 of which have a biological replicate, I will predict operons for each sample and compare the operon prediction models to each other to find operons that are consistently predicted across most of the samples. I will perform these operon predictions with multiple RNAseq-based operon prediction methods like Rockhopper and SeqTU [28, 29] and compare operon predictions between methods. Operons that are consistently predicted across samples and with multiple prediction tools will be considered highly supported, while operons that are inconsistently predicted across samples and between methods will be considered low support operons.

Activity 1-2: Determine conserved gene neighborhoods in *Wolbachia* strain genome assemblies. To further support *w*Mel operon predictions from RNAseq-based methods, I will use conserved gene neighborhoods to provide an additional source of support for predicted operons. First I will determine the location of homologous genes across each of the 11 complete *Wolbachia* assemblies. This will be accomplished by creating all-by-all assembly alignments using the Cactus genome aligner from the Comparative Genomics Toolkit [36]. These all-by-all genome alignments will then be used to transfer gene annotations across assemblies using HALtools from the Comparative Genomics Toolkit [37]. This will provide us with the location of homologous genes found in all assemblies. With this information, I will then calculate the frequency at which gene-pairs occur adjacent to one another in the same strand across all of the complete assemblies. Adjacent gene pairs found frequently in the *Wolbachia* genomes can be used as additional support for *w*Mel operons predicted using RNAseq-based methods (Activity 1-1).

Activity 1-3: Use predicted *w*Mel operon features to predict operons in *Wolbachia* genomes. After determining highly supported operons in *w*Mel using the modENCODE RNAseq data (Activity 1-1), I will use this information to train a machine-learning classifier to predict operons in other *Wolbachia* assemblies that lack RNAseq data. I will generate a positive training set using gene-pairs from *w*Mel that are consistently predicted to be co-transcribed in the RNAseq samples. This positive set will include features such as the intergenic distance, conservation of gene-pairs among *Wolbachia* assemblies (Activity 1-2) and the presence of intergenic promoters/terminators. A negative set will also be established using the same features for gene-pairs known to not belong to the same operon such as genes occurring on opposite strands. These positive and negative datasets will then be used to train machine learning classifiers implemented in the Python package Scikit-learn [38]. To determine the appropriate classifier to use for *Wolbachia* operon classification, I will use 10-fold cross validation with the positive and negative data from *w*Mel to evaluate the performance of decision trees, support-vector machines, and multi-layer perceptrons. Once an appropriate classifier and optimal parameters have been identified, I will generate testing sets for each of the *Wolbachia* genome assemblies containing the same features as the training set. The classifier will then be trained on the training set and then make operon predictions for gene-pairs for all *Wolbachia* assemblies in Genbank.

Expected Outcome:

I expect to be able to leverage the extensive modENCODE RNAseq dataset for ISO1 to make highly supported operon predictions in the *Wolbachia* strain *w*Mel. I also expect to be able to utilize these operon predictions as well as conserved gene neighborhood information to construct an operon classification method that can be applied to the *Wolbachia* strains that currently lack operon predictions. After the completion of this aim I should have experimentally supported operon models for *w*Mel and accurate operon predictions for the remaining *Wolbachia* assemblies. This method should also be applicable to future *Wolbachia* assemblies that are produced.

Potential Problems and Alternative Strategies:

As transcription in microbes can be dynamic, transcriptional units may change as the environmental conditions change. The modENCODE RNAseq data samples were taken across the life cycle of the host [33, 34], which implies that the environmental conditions of the *Wolbachia* infection were likely changing with each sample. It is possible that there will be little agreement in the predicted operons for *w*Mel across the modENCODE RNAseq samples due to changes in expressed transcriptional units or imperfections in the available RNAseq-based classification methods. Alternatively, I could accomplished this same aim by generating SMRT-cappable-seq datasets from *Wolbachia* transcripts. This is a long-read approach that utilizes PacBio SMRT sequencing to sequence full-length transcripts [**?**]. This would produce a dataset where reads are capable of spanning multiple genes, thus making it possible to make definitive calls about whether or not gene-pairs are co-transcribed. This method would be an alternative to Activity 1-1, allowing me to establish highly supported operons for use as a positive training set for *Wolbachia* operon prediction (Activity 1-3).

Aim 2: Explore patterns in *Wolbachia* repetitive DNA evolution

Introduction:

The objective of this aim is to determine how repeat abundance is evolving short-term and long-term in *Wolbachia* to better understand why *Wolbachia* genomes are enriched with repetitive DNA. As *Wolbachia* genomes have been found to be abundant in repeat content relative to other obligate endosymbionts, I hypothesize that *Wolbachia* have maintained repetitive DNA due to an evolutionary benefit, potentially related to interactions with hosts. My approach to testing this hypothesis is to quantify the abundance of different classes of repeats in complete *Wolbachia* genome assemblies using repeat annotation software. I will then compare changes in repeat content across the *Wolbachia* phylogeny to see if repeat content is conserved or specific to *Wolbachia* clades or supergroups. I will also look at recent changes in repeat abundance by quantifying repetitive DNA in *w*Mel population resequencing data. The rationale for this aim is that understanding the evolution of repeats in *Wolbachia* should provide context about the unusual overabundance of repetitive DNA.

Justification and Feasibility:

There are a number of shared genomic features that are characteristic of bacterial endosymbiont genomes. Typically, bacterial endosymbionts have A+T rich genomes, exhibit rapid sequence evolution, and have highly reduced genomes [17, 39]. As these bacteria live in a strictly intracellular environment, they exhibit relaxed selection due to the stability of their environments and large population bottlenecks when being transmitted to host offspring. This relaxed selection results in the accumulation of mildly deleterious mutations, genome rearrangements, and deletions [17, 39]. These mutations result in the gradual loss of nonessential DNA. As a consequence of this genome reduction, repetitive DNA elements such as insertion sequences are often completely lacking from endosymbiont genomes.

Wolbachia genomes share many of the features typical of bacterial endosymbionts such as a high A+T bias and reduced genome size. However, unlike most other bacterial endosymbionts, *Wolbachia* genomes have been shown to maintain a high proportion of repetitive DNA [13, 14, 40]. This is particularly evident in arthropod infecting strains, such as wRi and *w*Mel whose genomes can be over 20% repetitive DNA. Most of the repetitive DNA found in these arthropod-infecting *Wolbachia* can be attributed to enrichment of particular repeat classes, including ankyrin repeats, insertion sequences, and group II introns [5, 16, 17]. The biological significance for the unusual enrichment of these repeat classes in *Wolbachia* genomes has yet to be resolved.

The abundance of ankyrin repeat domains in *Wolbachia* proteins is of particular interest as these are only rarely found in bacteria and are much more prevalent in eukaryotes and viruses [5]. Ankyrin repeats are found in 4% of the genes in the *Wolbachia* strains *w*Mel, wRi and wPip. Ankyrin repeat domains are tandem arrays of 33-residue motifs that typically fold into structures involved in protein-to-protein interactions [41]. In some intracellular pathogens, ankyrin repeats have been shown to be secreted into host cells through type IV secretion systems and interact with host factors [42]. This, along with their unusual abundance in *Wolbachia*, suggests that ankyrin repeats may play a role in host-symbiont interactions. To date, no one has been able to establish a connection between *Wolbachia* ankyrin repeats and host phenotypes, so the role of ankyrin repeats in *Wolbachia* is still unknown.

The overabundance of transposable elements in many *Wolbachia* strains is also of interest as endosymbionts typically lack mobile DNA. Some *Wolbachia* strains have insertion sequence (IS) copies accounting for 11%

of their genomes, which is much larger than what is typically found in prokaryote genomes (<3%) [16]. Some *Wolbachia* strains are also enriched for group II introns, which are self-splicing mobile elements similar to nonlong terminal repeat retrotransposons. Previous work by Leclercq et al. [17] surveyed 961 bacterial genomes and found that the *Wolbachia* strains *w*Mel, wRi, and wPel were among the top 5% of genomes based upon group II intron content. They also found that only ~12% of endosymbiont genomes contain any group II introns, so the extensive enrichment of group II introns in *Wolbachia* strains is particularly unusual.

Transposable elements are major drivers of genome evolution due to their high recombinogenic potential and ability to generate insertion mutations. In particular, group II introns, when abundant, have high potential for enabling recombination events between copies due to their large size (~2kb). Indeed, when comparing the genomes between *w*Mel and wRi, Leclercq et al. [17] found that 26% of the genomic rearrangements occurred in regions that corresponded to group II introns, and 50% corresponded to IS elements. This suggests that these repetitive elements play a major role in genome evolution of *Wolbachia*.

Previous work exploring repeat evolution in *Wolbachia* was limited by the available complete assemblies at the time. Most of these studies only focused on 3-4 *Wolbachia* strains, as the other available *Wolbachia* assemblies were incomplete and likely inaccurately represented the repeat content of the associated strain. Currently, there are eleven complete *Wolbachia* genome assemblies available (Figure 1A). Unlike the strains used in previous studies, the available complete genomes are diverse and have representatives for six of the major *Wolbachia* phylogenetic supergroups. This allows me to expand upon previous *Wolbachia* repeat studies and potentially identify supergroup specific repeat patterns.

While comparisons of complete *Wolbachia* assemblies will be useful for study of long-term repeat evolution, whole genome resequencing data of multiple *Wolbachia* strains from the same host species can be used to explore repeat evolution in the short-term. Previous work by Richardson et al. [43] found that 179 *D. melanogaster* strains from the *Drosophila* Genetic Reference Panel (DGRP) and the *Drosophila* Population Genomics Project (DPGP) had identifiable infections of *w*Mel in their whole-genome resequencing data. They used this *w*Mel resequencing data to build a phylogeny of the *Wolbachia* infections in these *D. melanogaster* strains. Work by Early et al. [44] sequenced 65 *D. melanogaster* strains from the Drosophila Global Diversity project that were infected with *w*Mel, adding to the available whole-genome resequencing data available for this strain. Furthermore, additional *w*Mel resequencing data can be discovered from *D. melanogaster* samples produced from the *Drosophila* Genome Nexus panel [45] as well as from the many *D. melanogaster* samples available on NCBI. Using this *w*Mel resequencing data I can determine the abundance of repetitive elements in these samples and compare these values to a phylogeny. This will allow me to identify patterns of repeat evolution occurring in the *w*Mel strain.

Currently, there is no consistently used method for identifying repeats in *Wolbachia*. When surveying publications, repeats are identified using different methodologies for each of the current complete *Wolbachia* assemblies. A number of assemblies only have repeat predictions for specific classes such as insertion sequences, but lack predictions for other classes of interest. Even with the assemblies with extensive repeat predictions, the repeat identification process involves manual curation and correction by the researcher which introduces an element of subjectivity to the process which cannot be replicated in other assemblies. Also, my preliminary results indicate that the repeat predictions vary significantly between repeat annotation tools, meaning that current repeat predictions that use different methodologies may not be directly comparable. In order to compare repeat patterns among the *Wolbachia* assemblies, a single workflow is needed to produced accurate and comparable repeat predictions. In this aim, I will evaluate repeat annotation tools to determine the best methods for annotating repeats in *Wolbachia*. These tools will be applied to all complete *Wolbachia* assemblies to produce repeat annotations that are comparable across genomes.

Research Design:

Activity 2-1: Evaluate repeat annotation tools for accuracy in *Wolbachia* genomes. In order to compare repeat content across *Wolbachia* strains, I will need to determine which repeat annotation tools work best at identifying repeats in *Wolbachia* genomes. Manually curated repeat annotations have been established for the *Wolbachia* strain *w*Mel by previous work [13, 16]. I will run various repeat annotation tools such as Repeatoire, RepeatMasker, RepeatScout, ISEScan, and ISFinder to determine which tools best replicate the manually curated *w*Mel results from previous work [46, 47, 48, 49, 50]. It is likely that multiple tools will have to be employed together to accurately predict repeats as some tools perform better at annotating insertion sequences, while others are designed for identification of simple sequence repeats. I will develop a pipeline that runs complementary

repeat annotation tools and perform filtering to remove redundant predictions. This will provide a method of repeat annotation for *Wolbachia* assemblies that is consistent and scalable to larger number of assemblies without manual curation.

Activity 2-2: Compare repeats across complete *Wolbachia* assemblies. After determining the most effective tools for repeat annotation in *w*Mel, I will apply these tools to the other complete *Wolbachia* assemblies to produce repeat annotations. From these repeat annotations I will calculate the relative repeat abundance in each genome. I will also calculate the abundance of specific repeat classes such as ankyrin repeats and transposable elements. For transposable elements, I will calculate the number of full-length elements as well as the number of truncated elements. I will then compare these abundances to the *Wolbachia* strain phylogeny to determine if repeat features are consistent across strains or if there are clade specific patterns. Clade specific patterns should provide insight into the biological cause of the maintained repeats.

I will also examine other repeat features that might reveal the dynamics of repeat evolution in *Wolbachia*. I will compare nucleotide divergence of IS elements and group II introns among strains. To do so, I would produce alignments of IS elements and group II intron genes using MUSCLE [51]. This alignment would then be used to build a phylogenetic tree of the elements using RAxML [52]. I will then compare the IS and group II intron phylogeny to the strain phylogeny. Repeat phylogenies that are discordant from the strain phylogenies would be evidence of horizontal transfer of repeats across strains.

As repetitive DNA are often targets of ectopic recombination, they often induce genomic rearrangements. I will determine how often repeats are found at the flanks of inversions and deletions to see how much they are contributing to the extensive structural rearrangements seen among *Wolbachia* strains (Figure 1A). I will do this by identifying likely sites of genomic rearrangements by comparing assemblies using MUMmer [53]. I will then determine the frequency that these identified rearrangements are flanked by predicted repetitive DNA. This will reveal the frequency that repeats are involved in genomic rearrangements as well as which classes of repeats are most involved.

Activity 2-3: Compare repeat content across *w*Mel infections in *D. melanogaster* WGS samples. To examine repeat content changes occurring within the *Wolbachia* strain *w*Mel, I will use the WGS samples available for this strain identified in Richardson et al. [43], Early et al. [44] as well as infected *D. melanogaster* WGS samples from the Drosophila Genome Nexus and NCBI [45]. To determine which *D. melanogaster* samples are infected from the Drosophila Genome Nexus and NCBI, I will map the WGS data to the *w*Mel reference genome (NC_002978.6), count the mapped reads, and, using an appropriate threshold, determine which samples are infected based on percentage of mapped reads to *w*Mel. Together, the data from Richardson et al. [43], Early et al. [44], Drosophila Genome Nexus [45] and NCBI, will then be aligned to the *w*Mel reference genome (NC_002978.6) and mutations will be called using the Breseq pipeline [54]. Mutations that result in changes in repeat content such as deletions, IS element insertions, and duplications will be identified for each sample. Evidence of Repeat-mediated genomic rearrangements will also be recorded for each sample. The SNPs found in each sample will also be used to build a phylogeny using RAxML [52]. The strain differences will then be compared to the phylogeny to determine the rate of repeat content changes in *w*Mel. This will reveal how active mobile genetic elements are in this genome, as well as determine if the overall repeat content is conserved or in flux across the *w*Mel phylogeny.

Expected Outcome:

With the completion of this aim, I expect that we will be able to expand upon previous knowledge of repeat evolution in *Wolbachia* and gain greater insights into why repeats are overrepresented in these genomes. I expect to be able to leverage the complete genome assemblies to determine how repeat content is conserved across *Wolbachia* strains which will reveal if repeat abundance is clade specific or characteristic of all *Wolbachia* strains. I also expect to be able to utilize the extensive WGS data for *w*Mel to compare repeat content changes to the dated phylogeny to determine the rate of repeat content changes over time. Together this should reveal the short-term and long-term evolutionary dynamics of repeats in *Wolbachia*.

Potential Problems and Alternative Strategies:

It is possible that current repeat annotation tools cannot accurately identify repeats for use in this aim. Manual curation, such as removal of false positive data or altering intervals of defined repeat elements, can be used to improve predictions. BLAST can also be used for larger repetitive elements to determine the proper start and end positions based on annotations from other bacteria. While significantly more time-consuming, the small size and relatively few complete genome assemblies make manual curation a plausible solution to inaccurate repeat

annotation results from repeat annotation tools.

It is also possible that we may not be able to accurately determine all repeat content changes using the *w*Mel WGS data. It can be difficult to determine genomic rearrangements using short read sequencing information, particularly when repetitive regions are involved. If this is the case, we can focus on changes in easier to identify repeats such as IS element insertions as these changes show identifiable patterns in short read sequencing data and a number of tools exist for identifying transposable element insertions [55].

Aim 3: Determine the biological significance of short RNAs in *Wolbachia*

Introduction:

The objective of this aim is to determine if pervasive antisense transcription and patterns of short RNAs in *Wolbachia* are evidence of a mechanism of gene regulation. We discovered short RNA mapping patterns in Wolbachia that is consistent across replicates. It's possible these short RNA patterns are evidence of gene regulation through mRNA degradation. I hypothesize that antisense transcription results in the formation of double stranded RNA products that are targets for degradation by RNase III which acts as a mechanism of gene regulation as characterized in *Staphylococcus aureus* by Lasa et al. [21]. My approach will be to use stranded RNAseq data to identify regions of antisense transcription in the *Wolbachia* genome. Short RNAseq data from *Drosophila melanogaster* infected with *Wolbachia* can then be mapped to the *Wolbachia* genome to characterize patterns of short RNAs and determine their co-occurrence with antisense transcription. The rationale for this aim is that understanding gene regulation in *Wolbachia* is essential to determining mechanisms of host-*Wolbachia* interactions. *Wolbachia* are capable of dynamic gene expression in different host sexes and tissues despite limited transcription factors, suggesting that other mechanisms of gene regulation, such as antisense RNAs, are active.

Justification and Feasibility:

Advances in the application of high-throughput RNAseq analysis has revealed that overlapping transcription is pervasive in bacteria [22]. Overlapping transcription occurs when RNA transcripts are produced that have complementary regions. This can occur via transcription of cis- or trans- non-coding RNAs or by mRNA that overlap on the 3' or 5' UTRs. Genes encoded in the middle of an operon on the opposite strand can also produce complementary transcripts. Studies on cis-antisense RNAs have shown that their promoters are similar to their complementary transcript, which suggests that the sense and antisense transcripts are regulated by the same mechanism [22]. There are a number of proposed mechanisms by which antisense transcripts could impact their sense counterparts. Overlapping transcripts could result in inhibited translation of the sense transcript if the pairing occurs near the ribosomal binding site [19, 22]. Some studies have shown that overlapping transcription can result in an increase in the protein translated from the sense mRNA, possibly through protection of the sense transcript from degradation [22]. Cis-antisense RNAs could also result in transcriptional interference, where collisions between RNA polymerase complexes from complementary strands could inhibit the production of mRNA, reducing the expression of the involved genes. Other studies provide evidence that the binding of sense and antisense transcripts results in targeted degradation by RNases [21].

In a study by Lasa et al. [21], sequencing of long and short RNA fractions from *S. aureus* revealed that the prevalent antisense transcripts in this organism were being digested by RNase III. This RNase III mediated degradation produced short RNA byproducts (19-22 nt) that were enriched in regions that exhibited antisense transcription. They found this pattern to also be present in the low-GC gram-positive bacteria *Enterococcus faecalis*, *Listeria monocytogenes*, and *Bacillus subtilis*, suggesting this mechanism of gene regulation may be more widespread than previously characterized. The authors suggest that this could be a mechanism to control transcriptional noise or to fine-tune expression of the sense transcript [21]. Lasa et al. [22] propose that this process could limit 'leaky' transcription, where antisense RNAs create a threshold that sense transcripts must exceed before proteins are translated.

While studies have revealed potential functions of overlapping transcripts, the biological role of the short RNA fragments derived from the RNase III-mediated degradation has not been established. These short RNAs may just be stable non-functional products of double stranded RNA degradation. Alternatively, as these short RNAs resemble microRNAs in eukaryotes, they may play a similar role in bacteria. Prokaryotes lack argonaute-like proteins but that is not sufficient to rule out that these short RNAs are not functional. Lasa et al. [22] hypothesizes that the YbeY protein, which shares structural homology with the MID domain of the eukaryotic

Argonaute protein, may interact with short RNAs to contribute to post-transcriptional regulation. Much more study of these short RNAs is needed to make any definitive claims about their biological role.

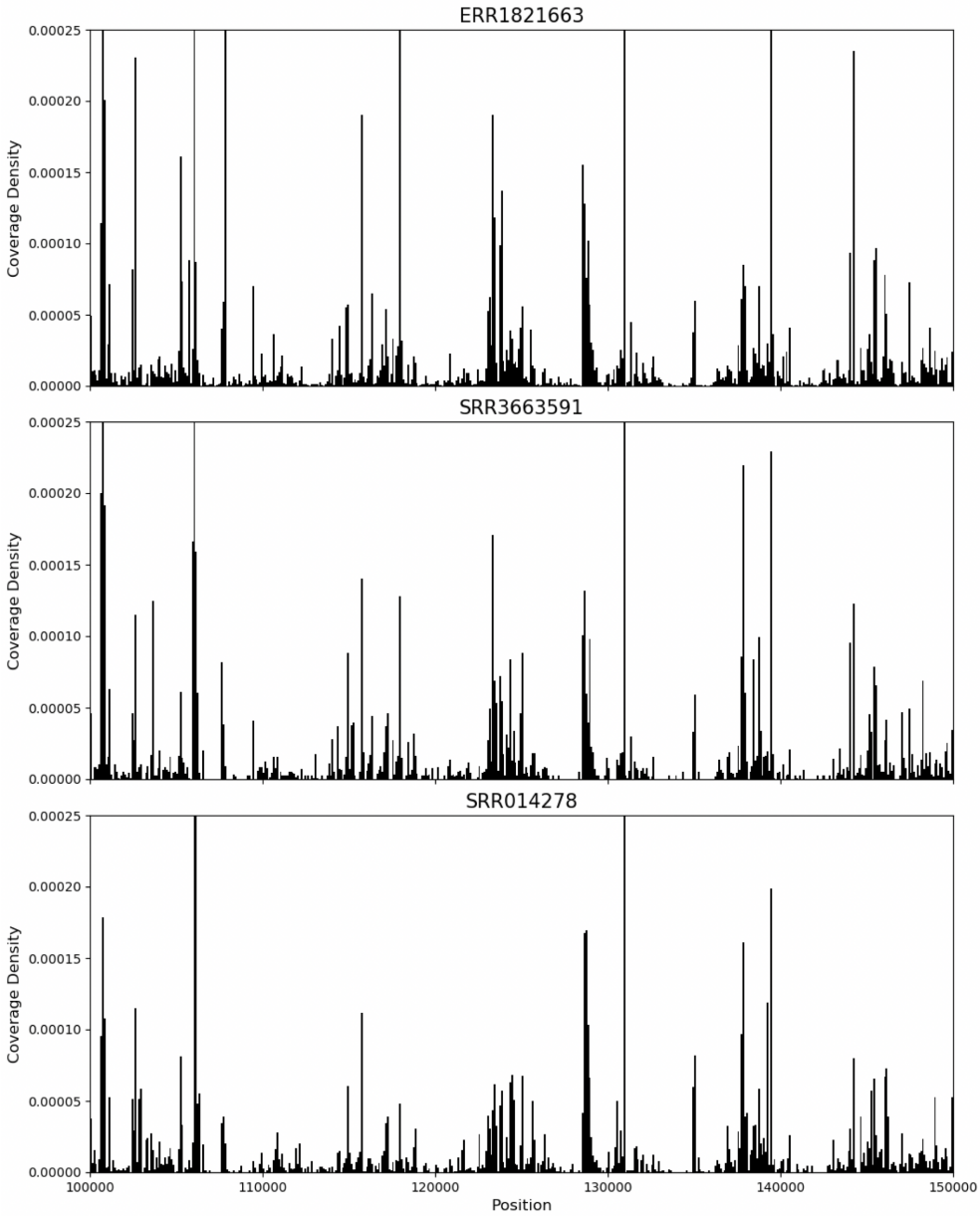


Figure 2: Short RNAs mapped to the *w*Mel genome show distinct coverage patterns that are consistent across multiple samples. Short RNA samples were mapped to hologenome consisting of *D. melanogaster* Release 6 and *Wolbachia w*Mel. Shown is the short RNA coverage across a 50kb region of the *w*Mel genome for three separate short RNA samples produced for separate projects in different lab groups. Peaks in short RNA coverage are remarkably similar between samples

As small RNA gene regulation is heavily studied in eukaryotes like *D. melanogaster*, high-throughput sequencing of short (< 50 nt) size-selected RNA libraries are readily available. Preliminary evidence shows that many of the *D. melanogaster* strains used in these experiments are infected with *Wolbachia*. When mapping these short RNAseq samples to a hologenome of *D. melanogaster* ISO1 and *Wolbachia w*Mel, I found that a significant number (2-4%) of the short RNA reads map to the *w*Mel genome. When looking at the read coverage across the *w*Mel genome, distinct patterns of short RNA mapping are observed that do not appear to be random. Indeed, I found that the distinct pattern in coverage profiles are consistent across samples, even samples produced separately in independent studies (Figure 2). These conserved coverage profiles are intriguing as they show that stable short RNAs are potentially being extensively produced in the *Wolbachia* genome, and that certain genomic regions are 'hotspots' for short RNA mapping. Determining why certain regions are enriched for short RNAs could provide clues as to the origin and mechanism behind the formation of these *Wolbachia* short RNAs.

As previous studies have shown that short RNAs in bacteria are derived from antisense transcripts, it would be interesting to see if the 'hotspots' are linked to overlapping transcripts, which would suggest the existence of a gene regulation mechanism based on antisense transcription.

My preliminary scans of publicly available *D. melanogaster* short RNA sequencing data shows that many of these samples contain a significant fraction that maps to the *w*Mel genome. I propose that this short RNAseq data coupled with the modENCODE RNAseq data for *w*Mel can be used to derive insights into the origin and function of the *Wolbachia* short RNAs. I will attempt to determine if these reads are truly originating from *Wolbachia* genomes and whether or not the conserved coverage patterns seen across samples are biologically significant. The preliminary evidence hints at the possibility of a gene expression mechanism previously uncharacterized in *Wolbachia*.

Research Design:

Activity 3-1: Search NCBI SRA for *Wolbachia*-infected *D. melanogaster* short RNA projects. Preliminary searches of the NCBI database has uncovered a number of *D. melanogaster* short RNA sequencing samples that contain reads that map to *w*Mel. Following up on this, I will expand my search of NCBI to determine which *D. melanogaster* short RNAseq samples have *Wolbachia* reads. *D. melanogaster* short RNAseq samples will be downloaded and reads will be mapped to a hologenome of *D. melanogaster* and *w*Mel using Bowtie2 [56]. The percent of reads that map to *w*Mel will be logged and the distribution of the mapped read lengths will be plotted. These metrics will be used to identify samples that are enriched for *w*Mel reads of length 19-22 nt. Preliminary results suggest that short RNA samples with over 2% reads mapping to the *w*Mel genome represent infected samples. Samples that exceed this threshold will then be used as replicate data for further activities in this aim.

Activity 3-2: Identify regions of antisense transcription and ncRNAs in *w*Mel. As short RNAs in other bacteria have been demonstrated to be the degradation products of double stranded transcripts [21], I want to determine if this is true for *w*Mel. Using the 30 RNAseq samples from the modENCODE *D. melanogaster* ISO1 developmental time course [33, 34, 35], I can confidently identify consistent regions of antisense transcription. These RNAseq samples were produced using a Illumina TruSeq Stranded Total RNA kit [35], thus strand-specific expression information can be derived from these samples [6], which indicates regions of antisense transcription can be identified. These RNAseq samples will be mapped to a hologenome of *D. melanogaster* Release 6 and the *Wolbachia w*Mel using BWA [57]. Coverage of reads mapping to *w*Mel will be calculated for each strand using BEDtools [58]. Intervals where transcription is identified in both strands of *w*Mel will be identified.

Short RNAs could also be produced from the binding of trans-ncRNAs to transcripts which would be missed from the previous antisense transcription analysis. The RNAseq-based operon identification tool Rockhopper can also predict novel ncRNAs from RNAseq data [28]. I will use Rockhopper with the modENCODE RNAseq data to identify novel ncRNAs in the *w*Mel genome. Previously identified small ncRNAs [20] combined with ncRNA predictions can then be used to determine if ncRNAs are enriched for short RNA mapping. These ncRNAs and antisense transcription intervals will then be used in Activity 3-3 to determine if they are enriched for short RNAs.

Activity 3-3: Determine which features of *w*Mel genome are hotspots for short RNA mapping. Coverage of short RNAs in *w*Mel shows that distinct regions of the genome are enriched for short RNA mapping (Figure 2). Determining the biological significance of the enriched regions will provide insight into the origin or role of these short RNAs. To start, I will determine if short RNAs abundance is related to gene transcription. Using the modENCODE RNAseq data, I will determine which regions of the genome are actively transcribed. I will then use BEDtools shuffle [58] to simulate short RNA mapping in the *Wolbachia* genome. I will then calculate the expected fraction of short RNAs mapping to transcribed regions. After this I will calculate the abundance of short RNAs mapping to transcribed regions for the experimental samples and compare their distribution to the simulated value. If significantly more short RNAs are found in transcribed regions than expected from the simulation, then it is likely that short RNAs are derived from or act on transcripts expressed by *Wolbachia*.

To determine if short RNA patterns coincide with regions of antisense transcription, I will use the antisense transcription intervals identified in Activity 3-2. To determine if the short RNAs mapping to regions of antisense is enriched, I will simulate short RNA random mapping using BEDtools shuffle [58] and compare the simulated abundance of short RNAs in antisense regions to those found in the experimental samples. Finding that the short RNAs mapping to antisense transcribed regions are significantly more abundant than what was found with simulation would suggest that short RNAs are derived from antisense transcripts.

To determine if ncRNAs coincide with regions of antisense transcription, I will use the ncRNAs identified via Rockhopper as well as previously identified ncRNAs described in Activity 3-2. Simulation of random short

RNA mapping with BEDtools shuffle as described previously will be used to establish the expected abundance of short RNAs mapping to ncRNA regions. The abundance of short RNAs mapping to ncRNA regions will then be calculated for each *Wolbachia*-infected short RNAseq sample and compared to the simulated value. If the experimental abundance is significantly greater than expected based on simulation, then it is likely that ncRNAs are related to the formation of short RNAs.

It is also possible that the conserved patterns in short RNA coverage (Figure 2) are an artifact of read multimapping in repetitive regions of the genome. Reads mapping tools are unable to determine the exact location of origin for reads that originate entirely from a repetitive region, thus they are often assigned two locations. This makes repeat regions appear to have a higher coverage that is proportional to the number of repeats found in the genome. Its possible that this could cause the conserved coverage patterns seen in Figure 2. To determine if this is the case, I will use the repeat annotations produced in **Aim 2** to determine the abundance of each repeat element. The coverage at each position will then be divided by the how repetitive that region is in the genome. The repeat-normalized coverage will then be compared across short RNA samples to see if the conserved peaks of short RNA mapping are still observed.

Expected Outcome:

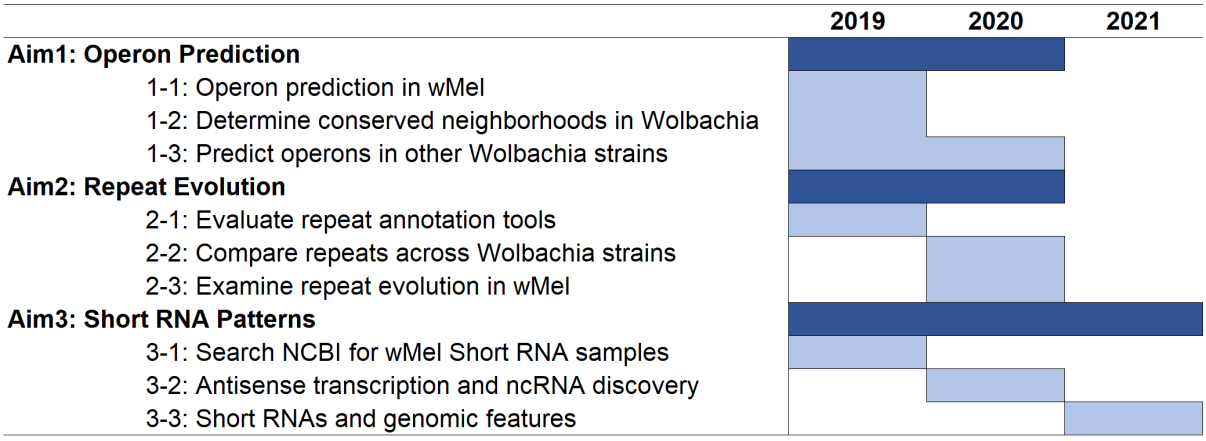
I expect to determine if the short RNA patterns seen in *Wolbachia* are biologically relevant or if they are artifacts of multimapping. If the short RNAs patterns are still seen after controlling for repeat abundance, I expect to determine if these short RNAs correlate with antisense transcripts such as those formed from overlapping transcription or ncRNAs. Enrichment of short RNAs related to antisense transcripts would suggests a system similar to that described in Lasa et al. [21] is occurring in *Wolbachia*.

Potential Problems and Alternative Strategies:

It is possible that conserved patterns seen in Figure 2 are artifacts of multimapping caused by repeats in the genome. If this is the case, it does not necessarily rule out that these short RNAs are present and potentially relevant to gene regulatory processes. Activities 3-2 and 3-3 can still be performed using repeat-normalized short RNA coverage to counteract the inflation of coverage caused by multimapping reads.

It is also possible that we do not find enrichment of short RNAs with transcription, antisense transcription, or ncRNAs. This would indicate that the short RNAs are not related to these processes. Other features of the genome can also be explored such as specific repeat classes to see if sort RNAs are enriched. If we cannot find any link between short RNAs and biological processes, ruling out that these short RNAs are evidence of a gene regulation mechanism is still a significant result and would suggest looking elsewhere for gene regulatory mechanisms involved in *Wolbachia* dynamic expression.

Timeline



References

[1] Roman Zug and Peter Hammerstein. Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. *PLOS ONE*, 7(6):e38544, 2012.

[2] A. A. Hoffmann, B. L. Montgomery, J. Popovici, I. Iturbe-Ormaetxe, P. H. Johnson, F. Muzzi, M. Greenfield, M. Durkan, Y. S. Leong, Y. Dong, H. Cook, J. Axford, A. G. Callahan, N. Kenny, C. Omodei, E. A. McGraw, P. A. Ryan, S. A. Ritchie, M. Turelli, and S. L. O'Neill. Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. *Nature*, 476(7361):454–7, 2011. Hoffmann, A A Montgomery, B L Popovici, J Iturbe-Ormaetxe, I Johnson, P H Muzzi, F Greenfield, M Durkan, M Leong, Y S Dong, Y Cook, H Axford, J Callahan, A G Kenny, N Omodei, C McGraw, E A Ryan, P A Ritchie, S A Turelli, M O'Neill, S L Evaluation Studies Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England Nature Nature. 2011 Aug 24;476(7361):454-7. doi: 10.1038/nature10356.

[3] John H. Werren, Laura Baldo, and Michael E. Clark. Wolbachia: master manipulators of invertebrate biology. *Nat Rev Microbiol*, 6(10):741–751, 2008.

[4] Ewa Chrostek, Marta S. P. Marialva, Sara S. Esteves, Lucy A. Weinert, Julien Martinez, Francis M. Jiggins, and Luis Teixeira. Wolbachia variants induce differential protection to viruses in Drosophila melanogaster: a phenotypic and phylogenomic analysis. *PLOS Genet*, 9(12):e1003896, 2013.

[5] Stefanos Siozios, Panagiotis Ioannidis, Lisa Klasson, Siv G. E. Andersson, Henk R. Braig, and Kostas Bourtzis. The diversity and evolution of Wolbachia ankyrin repeat domain genes. *PLOS ONE*, 8(2):e55390, February 2013.

[6] Florence Gutzwiller, Catarina R. Carmo, Danny E. Miller, Danny W. Rice, Irene L. G. Newton, R. Scott Hawley, Luis Teixeira, and Casey M. Bergman. Dynamics of Wolbachia pipientis gene expression across the Drosophila melanogaster life cycle. *G3*, 5(12):2843–2856, December 2015.

[7] Luciano A. Moreira, Iñaki Iturbe-Ormaetxe, Jason A. Jeffery, Guangjin Lu, Alyssa T. Pyke, Lauren M. Hedges, Bruno C. Rocha, Sonja Hall-Mendelin, Andrew Day, Markus Riegler, Leon E. Hugo, Karyn N. Johnson, Brian H. Kay, Elizabeth A. McGraw, Andrew F. van den Hurk, Peter A. Ryan, and Scott L. O'Neill. A Wolbachia symbiont in Aedes aegypti limits infection with dengue, Chikungunya, and Plasmodium. *Cell*, 139(7):1268–1278, December 2009.

[8] T. Walker, P. H. Johnson, L. A. Moreira, I. Iturbe-Ormaetxe, F. D. Frentiu, C. J. McMeniman, Y. S. Leong, Y. Dong, J. Axford, P. Kriesner, A. L. Lloyd, S. A. Ritchie, S. L. O'Neill, and A. A. Hoffmann. The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations. *Nature*, 476(7361):450–3, 2011. Walker, T Johnson, P H Moreira, L A Iturbe-Ormaetxe, I Frentiu, F D McMeniman, C J Leong, Y S Dong, Y Axford, J Kriesner, P Lloyd, A L Ritchie, S A O'Neill, S L Hoffmann, A A Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England Nature Nature. 2011 Aug 24;476(7361):450-3. doi: 10.1038/nature10355.

[9] M. S. Blagrove, C. Arias-Goeta, A. B. Failloux, and S. P. Sinkins. Wolbachia strain wMel induces cytoplasmic incompatibility and blocks dengue transmission in Aedes albopictus. *Proc Natl Acad Sci USA*, 109(1):255–60, 2012. Blagrove, Marcus S C Arias-Goeta, Camilo Failloux, Anna-Bella Sinkins, Steven P Biotechnology and Biological Sciences Research Council/United Kingdom Wellcome Trust/United Kingdom Research Support, Non-U.S. Gov't United States Proceedings of the National Academy of Sciences of the United States of America Proc Natl Acad Sci U S A. 2012 Jan 3;109(1):255-60. Epub 2011 Nov 28.

[10] Sofia Zabalou, Markus Riegler, Marianna Theodorakopoulou, Christian Stauffer, Charalambos Savakis, and Kostas Bourtzis. Wolbachia-induced cytoplasmic incompatibility as a means for insect pest population control. *Proc Natl Acad Sci USA*, 101(42):15042–15045, October 2004.

[11] Amelia R I Lindsey, Danny W Rice, Sarah R Bordenstein, Andrew W Brooks, Seth R Bordenstein, and Irene L G Newton. Evolutionary Genetics of Cytoplasmic Incompatibility Genes cifA and cifB in Prophage WO of Wolbachia. *Genome Biol Evol*, 10(2):434–451, January 2018.

[12] John F. Beckmann and Ann M. Fallon. Detection of the Wolbachia protein WPIP0282 in mosquito spermathecae: implications for cytoplasmic incompatibility. *Insect Biochem Mol Biol*, 43(9):867–878, September 2013.

[13] Martin Wu, Ling V. Sun, Jessica Vamathevan, Markus Riegler, Robert Deboy, Jeremy C. Brownlie, Elizabeth A. McGraw, William Martin, Christian Esser, Nahal Ahmadinejad, Christian Wiegand, Ramana Madupu, Maureen J. Beanan, Lauren M. Brinkac, Sean C. Daugherty, A. Scott Durkin, James F. Kolonay, William C. Nelson, Yasmin Mohamoud, Perris Lee, Kristi Berry, M. Brook Young, Teresa Utterback, Janice Weidman, William C. Nierman, Ian T. Paulsen, Karen E. Nelson, Hervé Tettelin, Scott L. O'Neill, and Jonathan A. Eisen. Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements. *PLOS Biol*, 2(3):E69, March 2004.

[14] Jeremy Foster, Mehul Ganatra, Ibrahim Kamal, Jennifer Ware, Kira Makarova, Natalia Ivanova, Anamitra Bhattacharyya, Vinayak Kapatral, Sanjay Kumar, Janos Posfai, Tamas Vincze, Jessica Ingram, Laurie Moran, Alla Lapidus, Marina Omelchenko, Nikos Kyrpides, Elodie Ghedin, Shiliang Wang, Eugene Goltsman, Victor Joukov, Olga Ostrovskaya, Kiryl Tsukerman, Mikhail Mazur, Donald Comb, Eugene Koonin, and Barton Slatko. The Wolbachia genome of Brugia malayi: endosymbiont evolution within a human pathogenic nematode. *PLOS Biol*, 3(4):e121, April 2005.

[15] Nancy A. Moran and Gordon R. Plague. Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev*, 14(6):627–633, December 2004.

[16] Nicolas Cerveau, Sébastien Leclercq, Elodie Leroy, Didier Bouchon, and Richard Cordaux. Short- and long-term evolutionary dynamics of bacterial insertion sequences: insights from Wolbachia endosymbionts. *Genome Biol Evol*, 3:1175–1186, 2011.

[17] Sébastien Leclercq, Isabelle Giraud, and Richard Cordaux. Remarkable abundance and evolution of mobile group II introns in Wolbachia bacterial endosymbionts. *Mol Biol Evol*, 28(1):685–697, January 2011.

[18] Jens Georg and Wolfgang R. Hess. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev*, 75(2):286–300, June 2011.

[19] Susan Gottesman and Gisela Storz. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol*, 3(12), December 2011.

[20] Megan Woolfit, Manjula Algama, Jonathan M. Keith, Elizabeth A. McGraw, and Jean Popovici. Discovery of putative small non-coding RNAs from the obligate intracellular bacterium Wolbachia pipientis. *PLOS ONE*, 10(3):e0118595, March 2015.

[21] Iñigo Lasa, Alejandro Toledo-Arana, Alexander Dobin, Maite Villanueva, Igor Ruiz de los Mozos, Marta Vergara-Irigaray, Víctor Segura, Delphine Fagegaltier, José R. Penadés, Jaione Valle, Cristina Solano, and Thomas R. Gingeras. Genome-wide antisense transcription drives mRNA processing in bacteria. *Proc Natl Acad Sci USA*, 108(50):20172–20177, December 2011.

[22] Iñigo Lasa, Alejandro Toledo-Arana, and Thomas R. Gingeras. An effort to make sense of antisense transcription in bacteria. *RNA Biol*, 9(8):1039–1044, August 2012.

[23] Mihaela Pertea, Kunmi Ayanbule, Megan Smedinghoff, and Steven L. Salzberg. OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res*, 37(suppl 1):D479–D482, January 2009.

[24] Phuongan Dam, Victor Olman, Kyle Harris, Zhengchang Su, and Ying Xu. Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res*, 35(1):288–298, 2007.

[25] Morgan N. Price, Katherine H. Huang, Eric J. Alm, and Adam P. Arkin. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, 33(3):880–892, 2005.

[26] Blanca Taboada, Cristina Verde, and Enrique Merino. High accuracy operon prediction method based on STRING database scores. *Nucleic Acids Res*, 38(12):e130, July 2010.

[27] Brian Tjaden. A computational system for identifying operons based on RNA-seq data. *Methods*, April 2019.

[28] Ryan McClure, Divya Balasubramanian, Yan Sun, Maksym Bobrovskyy, Paul Sumby, Caroline A. Genco, Carin K. Vanderpool, and Brian Tjaden. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res*, 41(14):e140–e140, August 2013.

[29] Wen-Chi Chou, Qin Ma, Shihui Yang, Sha Cao, Dawn M. Klingeman, Steven D. Brown, and Ying Xu. Analysis of strand-specific RNA-seq data using machine learning reveals the structures of transcription units in Clostridium thermocellum. *Nucleic Acids Res*, 43(10):e67–e67, May 2015.

[30] Fenglou Mao, Phuongan Dam, Jacky Chou, Victor Olman, and Ying Xu. DOOR: a database for prokaryotic operons. *Nucleic Acids Res*, 37(Database issue):D459–463, January 2009.

[31] Blanca Taboada, Ricardo Ciria, Cristian E. Martinez-Guerrero, and Enrique Merino. ProOpDB: prokaryotic operon database. *Nucleic Acids Res*, 40(D1):D627–D631, January 2012.

[32] Paramvir S. Dehal, Marcin P. Joachimiak, Morgan N. Price, John T. Bates, Jason K. Baumohl, Dylan Chivian, Greg D. Friedland, Katherine H. Huang, Keith Keller, Pavel S. Novichkov, Inna L. Dubchak, Eric J. Alm, and Adam P. Arkin. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res*, 38(Database issue):D396–400, January 2010.

[33] Brenton R. Graveley, Angela N. Brooks, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri, Marijke J. van Baren, Nathan Boley, Benjamin W. Booth, James B. Brown, Lucy Cherbas, Carrie A. Davis, Alex Dobin, Renhua Li, Wei Lin, John H. Malone, Nicolas R. Mattiuzzo, David Miller, David Sturgill, Brian B. Tuch, Chris Zaleski, Dayu Zhang, Marco Blanchette, Sandrine Dudoit, Brian Eads, Richard E. Green, Ann Hammonds, Lichun Jiang, Phil Kapranov, Laura Langton, Norbert Perrimon, Jeremy E. Sandler, Kenneth H. Wan, Aarron Willingham, Yu Zhang, Yi Zou, Justen Andrews, Peter J. Bickel, Steven E. Brenner, Michael R. Brent, Peter Cherbas, Thomas R. Gingeras, Roger A. Hoskins, Thomas C. Kaufman, Brian Oliver, and Susan E. Celniker. The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339):473–479, March 2011.

[34] James B. Brown, Nathan Boley, Robert Eisman, Gemma E. May, Marcus H. Stoiber, Michael O. Duff, Ben W. Booth, Jiayu Wen, Soo Park, Ana Maria Suzuki, Kenneth H. Wan, Charles Yu, Dayu Zhang, Joseph W. Carlson, Lucy Cherbas, Brian D. Eads, David Miller, Keithanne Mockaitis, Johnny Roberts, Carrie A. Davis, Erwin Frise, Ann S. Hammonds, Sara Olson, Sol Shenker, David Sturgill, Anastasia A. Samsonova, Richard Weiszmann, Garret Robinson, Juan Hernandez, Justen Andrews, Peter J. Bickel, Piero Carninci, Peter Cherbas, Thomas R. Gingeras, Roger A. Hoskins, Thomas C. Kaufman, Eric C. Lai, Brian Oliver, Norbert Perrimon, Brenton R. Graveley, and Susan E. Celniker. Diversity and dynamics of the Drosophila transcriptome. *Nature*, 512(7515):393–399, August 2014.

[35] Michael O. Duff, Sara Olson, Xintao Wei, Sandra C. Garrett, Ahmad Osman, Mohan Bolisetty, Alex Plocik, Susan E. Celniker, and Brenton R. Graveley. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*, 521(7552):376–379, May 2015.

[36] Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res*, 21(9):1512–1528, September 2011.

[37] Glenn Hickey, Benedict Paten, Dent Earl, Daniel Zerbino, and David Haussler. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, 29(10):1341–1342, May 2013.

[38] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, 12:2825–2830, November 2011.

[39] Nancy A. Moran, John P. McCutcheon, and Atsushi Nakabachi. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet*, 42:165–190, 2008.

[40] Lisa Klasson, Joakim Westberg, Panagiotis Sapountzis, Kristina Näslund, Ylva Lutnaes, Alistair C. Darby, Zoe Veneti, Lanming Chen, Henk R. Braig, Roger Garrett, Kostas Bourtzis, and Siv G. Andersson. The mosaic genome structure of the Wolbachia wRi strain infecting Drosophila simulans. *Proc Natl Acad Sci USA*, 106(14):5725–5730, April 2009.

[41] Souhaila Al-Khodor, Christopher T. Price, Awdhesh Kalia, and Yousef Abu Kwaik. Ankyrin-repeat containing proteins of microbes: a conserved structure with functional diversity. *Trends Microbiol*, 18(3):132–139, March 2010.

[42] Xiaoxiao Pan, Anja Lührmann, Ayano Satoh, Michelle A. Laskowski-Arce, and Craig R. Roy. Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. *Science*, 320(5883):1651–1654, June 2008.

[43] Mark F. Richardson, Lucy A. Weinert, John J. Welch, Raquel S. Linheiro, Michael M. Magwire, Francis M. Jiggins, and Casey M. Bergman. Population genomics of the Wolbachia endosymbiont in Drosophila melanogaster. *PLOS Genet*, 8(12):e1003129, 2012.

[44] Angela M. Early and Andrew G. Clark. Monophyly of Wolbachia pipientis genomes within Drosophila melanogaster: geographic structuring, titre variation and host effects across five populations. *Mol Ecol*, 22(23):5765–5778, 2013.

[45] Justin B. Lack, Charis M. Cardeno, Marc W. Crepeau, William Taylor, Russell B. Corbett-Detig, Kristian A. Stevens, Charles H. Langley, and John E. Pool. The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics*, 199(4):1229–1241, April 2015.

[46] Todd J. Treangen, Aaron E. Darling, Guillaume Achaz, Mark A. Ragan, Xavier Messeguer, and Eduardo P. C. Rocha. A novel heuristic for local multiple alignment of interspersed DNA repeats. *IEEE/ACM Trans Comput Biol Bioinform*, 6(2):180–189, June 2009.

[47] A.F.A. Smit, R. Hubley, and P. Green. RepeatMasker, 2013.

[48] A. L. Price, N. C. Jones, and P. A. Pevzner. De novo identification of repeat families in large genomes. *Bioinformatics*, 21 Suppl 1:i351–i358, 2005. 1367-4803 Journal Article.

[49] Zhiqun Xie, Haixu Tang, and John Hancock. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*, 33(21):3340–3347, November 2017.

[50] P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, and M. Chandler. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res*, 34(Database issue):D32–6, 2006. Siguier, P Perochon, J Lestrade, L Mahillon, J Chandler, M Research Support, Non-U.S. Gov't England Nucleic acids research Nucleic Acids Res. 2006 Jan 1;34(Database issue):D32-6.

[51] Robert C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, August 2004.

[52] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, May 2014.

[53] Guillaume Marçais, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. MUMmer4: A fast and versatile genome alignment system. *PLOS Comput Biol*, 14(1):e1005944, January 2018.

[54] Daniel E. Deatherage and Jeffrey E. Barrick. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol*, 1151:165–188, 2014.

[55] Michael G. Nelson, Raquel S. Linheiro, and Casey M. Bergman. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3*, 7:2749–2762, August 2017.

[56] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, April 2012.

[57] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[58] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.