Identifying and Investigating *Wolbachia*-to-Arthropod HGT Events

Author: Serafina Nieves
Advisor: Dr. Russell Corbett-Detig

Winter 2021

Acknowledgements

I would like to thank Russell Corbett-Detig for his mentorship and for fostering independence and critical thinking in my research. I would also like to thank him for the opportunity to work in his lab and always advocating for me as a young scientist. I am also deeply grateful for Dr. Zia Isola and the Research Mentoring Internship (RMI) program, which have provided me with the financial ability and academic resources to pursue independent projects and contribute to the greater scientific community. I would also like to thank my peers and the graduate students in the Corbett-Detig Lab, who teach me new things every day. Additionally, I must thank my partner Darby, who supports me each day in every endeavor, and probably now knows my project as well as I do. I would also like to thank my parents, who would do anything to see me succeed and be happy.

Abstract

This study analyzes publicly available whole genome sequencing projects for evidence of horizontal gene transfer from *Wolbachia* to arthropods. BLAST was used to query all arthropod proteins against a database of *Wolbachia* proteins, resulting in a set of 1673 putatively transferred genes. The depth of coverage, gene expression, number of introns, and number of eukaryotic genes on the same contig was determined for each putative transfer. After filtering the putative transfers with these criteria, I found 715 real transfers. Additionally, I found that these genes experience relaxed selection after transferring to arthropod lineages. I found that *Wolbachia*-to-arthropod HGT events may be more common than current literature seems to suggest, but are being lost from the genome rapidly after insertion. These findings may be important as *Wolbachia* infections spread and as it becomes an increasingly popular option for biological control.

List of Figures

1 Tree of Life 1
2 Endosymbiont Ratchet 2
3 Selection on Lineages 4
4 STAR Algorithm 6
5 Workflow Overview 8
6 Genomic Coverage Distribution 13
7 Difference in Gene Expression 14
8 Confidence Categories 15

List of Tables

1 Five Key Features for Testing 7
2 HGT by Species 12
3 Confidence Categories by Species 15

Table of Contents

1 Background **1**
- 1.1 Defining Horizontal Gene Transfer . . . 1
- 1.2 *Wolbachia* . . . 3
- 1.3 Evolution of Proteins . . . 3
- 1.4 BLAST . . . 5
- 1.5 BWA . . . 5
- 1.6 STAR . . . 5
- 1.7 TPMCalculator . . . 6
- 1.8 PAML . . . 6

2 Methods and Results **7**
- 2.1 Workflow Overview . . . 7
- 2.2 Identification of Putative Transfer Events . . . 8
  - 2.2.1 Retrieving Arthropod Assemblies from NCBI . . . 8
  - 2.2.2 Creating the *Wolbachia* BLAST Database and Executing BLAST Search . . . 8
- 2.3 Validation of Putative Transfer Events . . . 9
  - 2.3.1 Checking Read Coverage of Transferred Genes Against Host Genes . . . 9
  - 2.3.2 Checking Gene Expression and Poly-A Selection . . . 9
  - 2.3.3 Parsing Annotation Files for Eukaryotic Features . . . 9
- 2.4 Quantifying Results . . . 10
- 2.5 Evaluating Selection on Protein-Coding Genes . . . 10

3 Results **11**
- 3.1 Characteristics of Putative Transfers . . . 11
  - 3.1.1 Identifying a Set of Putatively Transferred Genes . . . 11
- 3.2 Confirming Real Transfers . . . 12
  - 3.2.1 Genome Coverage of Putative Transfers . . . 12
  - 3.2.2 Gene Expression of Putative Transfers . . . 14
  - 3.2.3 Filtering for Real Transfers . . . 15
  - 3.2.4 Evolution of Protein-Coding Genes . . . 16

4 Conclusion **19**

5 Future Work **19**

A GitHub **22**

B Acquiring Data **22**

1 Background

1.1 Defining Horizontal Gene Transfer

When considering the transfer of genetic information from one organism to another, we are most familiar with instances of vertical gene transfer: the transfer of genes from parents to offspring. Vertical gene transfer occurs in both sexual and asexual reproduction, and has overwhelmingly been the focus of evolutionary genetics until recently. In contrast, horizontal gene transfer (often referred to as HGT, lateral gene transfer, or LGT), is the movement of genetic information from one organism to another without reproduction. This mode of transfer can almost instantaneously introduce foreign genes into a new genome, potentially increasing the genetic variation in the host while providing the means to rapidly adapt to a new environment (Danchin, 2016; Woods et al., 2020).

HGT is extensively documented in prokaryotes, particularly with respect to the acquisition of antibiotic resistance genes (Danchin, 2016; Woods et al., 2020). Antibiotic use has even been thought to trigger the horizontal transfer of antibiotic resistance and other mobile genes between different bacterial taxa (Kent et al., 2020). Although HGT between prokaryotes is commonplace and even fundamental to their evolution, it is less common between prokaryotes and other kingdoms.

Despite occurring at much lower frequencies, HGT does happen between prokaryotes and eukaryotes and can have profound evolutionary implications for the host (Figure 1). For example, eukaryotic organisms have the ability to engulf other organisms and materials and bring them into their cells. When prokaryotes are brought into the cell via this process (endocytosis), they can remain in constant close proximity with the host organisms genome as an endosymbiont (Keeling and Palmer, 2008). Host-symbiont relationships may be mutualistic, where one or both may provide some advantage to the other, such as gene functions that may be missing from either organism. In some cases, the relationship between endosymbiont and host is not just mutualistic but is required for survival (obligate), such as that of the filarial nematode *Brugia malayi* and bacteria in the genus *Wolbachia* (Hotopp et al., 2007).

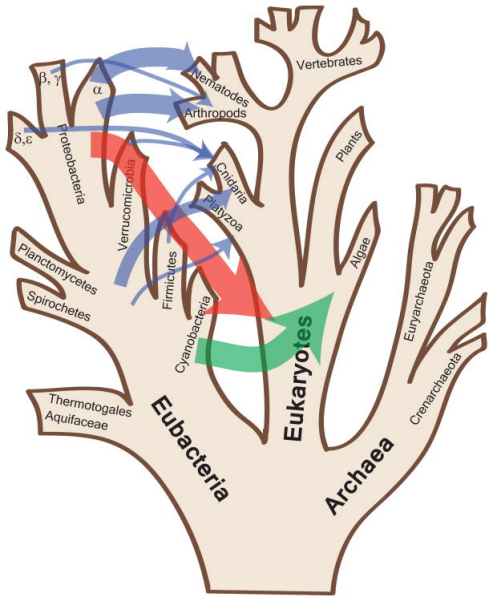

Figure 1: A general tree of life showing multiple extensive genetic contributions of prokaryotes to eukaryotic lineages throughout time (adapted from (Hotopp et al., 2007)).

Mitochondria are thought to be the remnants of alpha-proteobacteria that transferred genes extensively with their eukaryotic hosts. In contrast to the endosymbionts they once were, mitochondria only have a small number of the protein-encoding genes they need to survive (Keeling and Palmer, 2008). The remaining functional proteins that mitochondria require are imported into the double-membrane bound organelle through protein transport systems (Theissen and Martin, 2006). These proteins are encoded by the nuclear genome, indicating a horizontal gene transfer event from alpha-proteobacterial endosymbiont to host occurred at some point in time. These transfers have been shown to occur in large bursts or batches, followed by long periods of dormancy (Keeling and Palmer, 2008).

When HGT occurs from the endosymbiont to its host, the host may acquire novel gene functions. The functions may reduce the host's reliance on its endosymbiont, and may result in the loss of the original gene in the endosymbiont as the host provides the same gene function, or even the loss of the endosymbiont from the host (Figure 2). However, genes acquired by the host may also be purged out of the genome, returning the host genome to its original state and preserving the role of the endosymbiont (Dunning Hotopp, 2011). Depicted in the figure below is a proposed HGT ratchet of a host nuclear genome acquiring new functional genes from its endosymbiont.

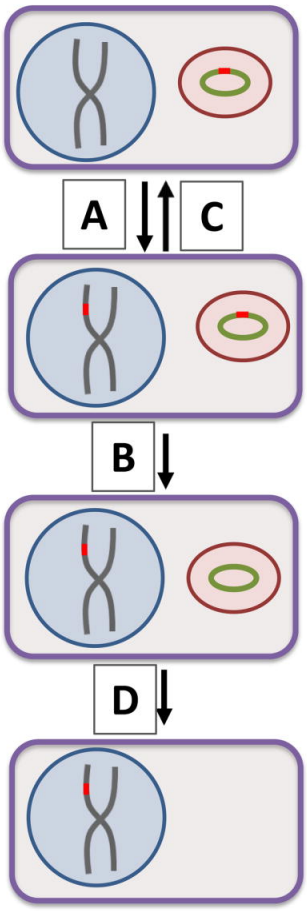

Figure 2: (a) depicts the endosymbiont and host genome in close proximity in the cell, where the gene can transfer from the endosymbiont to the host. (c) negative selection can result in removal of the transferred gene, returning the host genome to its original state. (b) if the transferred gene is functional in the host genome, it can fulfill the needs of the endosymbiont, reducing the need to maintain the endosymbiont gene. (d) the endosymbiont is lost when it no longer fulfills the same role to the host. Figure borrowed from another source Dunning Hotopp, 2011.

1.2 *Wolbachia*

Some of the most prolific endosymbionts are found in the genus *Wolbachia*. These gram-negative bacteria are found in the reproductive systems of many arthropods. *Wolbachia* can also alter the reproductive systems of their host in several ways (Werren, 1997). One of the principal ways *Wolbachia* modifies its hosts is through cytoplasmic incompatibility: sperm from males harboring *Wolbachia* combined with eggs from females without *Wolbachia* do not result in viable embryos. *Wolbachia* can also induce parthenogenesis in species that produce males from unfertilized eggs. The result is that unfertilized eggs actually produce females, which can then vertically transmit *Wolbachia* (Werren, 1997; Werren et al., 2008).

Through manipulating the host reproductive system, *Wolbachia* ensures it can be passed onto offspring and even between species. In fact, *Wolbachia* is so effective at both vertical and horizontal transmission, that studies have estimated that around 40% of arthropod species are infected, with some studies estimating up to 66% (Zug and Hammerstein, 2012). The prevalence of *Wolbachia* infections in arthropod species is of particular interest in research because it could be used as a means of biological control of insect populations. Furthermore, *Wolbachia* might have the potential to influence the evolution of its arthropod hosts (Werren et al., 2008).

There are several examples of HGT between *Wolbachia* and arthropods that have been identified thus far. These span a wide array of taxa and have varying levels of effects on the host. For example, the ancient transfer of salivary gland surface (SGS) proteins from *Wolbachia* to *Aedes aegypti* and *Anopheles gambiae* may aid in the transmission of the malaria parasite from mosquitoes to other organisms (Woolfit et al., 2009). The draft genome of *Formica exsecta* contains approximately 83 protein-coding genes of *Wolbachia* origin (Dhaygude et al., 2019). 30% of wMel strain *Wolbachia* genes were shown to be present in the *Callosobruchus chinensis* nuclear genome (Nikoh et al., 2008). A large transfer of *Wolbachia* genes into Armadillidium vulgare, the common pillbug, may have been responsible for the evolution of new sex-determining chromosomal region, allowing genetic males to develop as phenotypic females (Leclercq et al., 2016). Horizontal transfer and subsequent duplication of *Wolbachia* genes also accounts for up to 2% of the total Drosophila ananassae genome (Klasson et al., 2014).

Wolbachia HGT seems to occur frequently and across diverse taxonomic groups. It is therefore important to understand whether these genes are being utilized by the host, how they affect host fitness and function, and how long they remain in the host genome.

1.3 Evolution of Proteins

If we wish to understand how an organism is evolving and why, we need to be able to quantify changes among the important components of its biological systems. Proteins carry out most of the work in cells and are the basis for diverse interconnected pathways. Proteins are typically evolutionarily conserved, meaning they are sensitive to any amino acid or nucleotide changes that might change their function or expression (Balding et al., 2007). For this reason proteins are often said to be under negative or purifying selection, a process that purges mutations (which are more likely to be deleterious than not), resulting in sequence conservation. On the other hand, positive selection allows beneficial mutations to fix in a population, optimizing a protein or allowing it to adapt to differing functional constraints (Booker et al., 2017).

From this concept a relatively intuitive measurement of selection was created. This method involves calculating the ratio of the nonsynonymous substitution rate vs. the synonymous substitution rate. In essence, the method compares the rate of mutation that causes amino acid changes vs. the rate of mutation that does not cause a change (silent mutation). This ratio, termed the

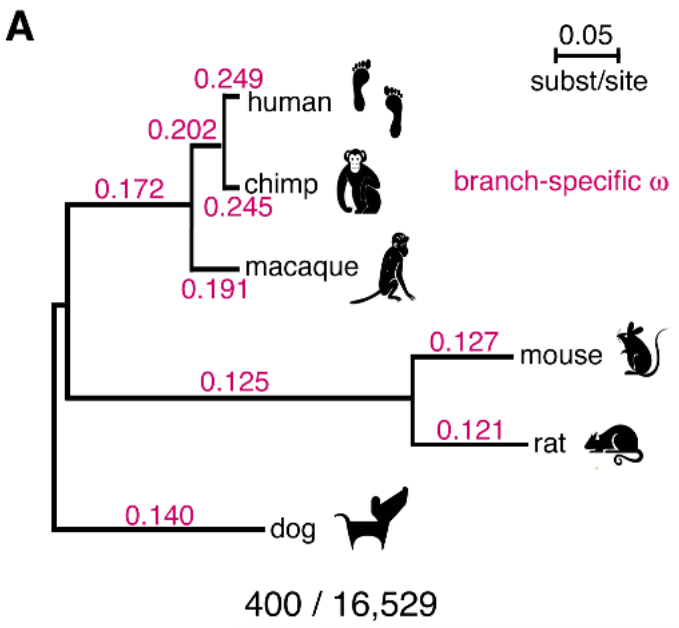

Figure 3: A phylogenetic tree showing the dN/dS ratios along each branch. Overall negative selection is observed in the tree, which is to be expected due to the likelihood of mutations reducing fitness instead of providing new adaptations (figure borrowed from (Kosiol et al., 2008)).

dN/dS ratio (or K_a/K_s), is symbolized by the Greek letter ω. Because synonymous substitutions do not change the amino acid sequence of a protein, they are thought to be evolutionarily neutral (Balding et al., 2007). When the nonsynonymous mutation rate is equal to the synonymous rate, $\omega = 1$, this indicates that selection is not acting in any direction on the gene or protein of interest; when the nonsynonymous mutation is far below the synonymous rate, $\omega < 1$, there is likely some constraint that is acting on the protein or gene of interest, necessitating that changes to the already functional protein be purged. $\omega < 1$ is expected throughout most of the evolutionary history of a gene or protein, as mutations are expected to be more costly than beneficial in terms of fitness (Balding et al., 2007). On the other hand, $\omega > 1$ suggests that positive selection is acting on the gene or protein to fix new potentially adaptive alleles in the population. The fixation of adaptive alleles typically happens in bursts and is flanked by long periods of negative selection removing new mutations from functionally constrained proteins (Zhang et al., 2005).

The concept of dN/dS can be leveraged to explain gene evolution along different lineages. Given a phylogenetic tree constructed by using the same gene across multiple taxa, the dN/dS ratio can be calculated for each branch. In order to understand if the differences in selection along different lineages is truly significant, a null model is used. The null model states that all branches in all taxonomic groups have the same dN/dS, and are therefore under the same selection pressure. The dN/dS ratio is then calculated again in the alternative model, where branches are allowed to have differing dN/dS ratios (Yang, 2017). The ratios can differ along every branch or branches can be grouped together and ratios assigned per group. In Figure 3, every branch has its own dN/dS ratio; however, as mentioned above, one ratio could be assigned to all primates and a different ratio assigned to all rodents. This 2-ratio method allows you to view how selection differs between groups of primates and rodents, rather than between one specific species to another.

1.4 BLAST

The Basic Local Alignment Search Tool (BLAST) is a bioinformatics toolkit primarily used to find similarity between sequences. In contrast to global alignment techniques, which force the alignments of very long sequences end-to-end, local alignment looks for the best subset alignments of potentially dissimilar sequences ("Sequence alignment," n.d.). The local alignment algorithm allows researchers to find sequence homologs between very divergent genes in distantly related or seemingly unrelated taxa.

BLAST accepts sequences of interest, called query sequences, and looks them up against a given database. A lookup table is constructed by reading and filtering the query sequence for important subsets of the query, called words. The subject sequences (the sequences in the database of interest) are then scanned for the words contained in the lookup table. Alignments are scored and penalized for features such as gaps resulting from insertions or deletions. All hits that match user parameters are reported (Camacho et al., 2009).

The BLAST+ suite for command line usage contains several programs for different use cases. Two of the most popular programs are the BLASTp and BLASTn tools, used for local alignment of proteins and nucleotides respectively. "makeblastdb" is also a popular BLAST program used to compile custom databases of an individual's choosing, for use as the subject sequences of a BLAST search. The National Center for Biotechnology Information (NCBI) also provides pre-formatted BLAST databases containing vast amounts of non-redundant sequence data, such as the "nr" database for proteins (*BLAST® Command Line Applications User Manual*, 2008).

1.5 BWA

Burrows-Wheeler Alignment is a local alignment-based algorithm that can align many query sequences to a reference genome. It is usually used to align short-read whole genome sequencing data to a whole genome reference. The reference genome is first indexed to enable faster lookup of matching sequences. The algorithm uses dynamic programming to align seed sequences to the reference genome, allowing for gaps and mismatching letters (Li and Durbin, 2010). BWA then extends these aligned seeds by the well-documented Smith-Waterman alignment algorithm. In short, the Smith-Waterman algorithm considers all possible subsets of a sequence when aligning it, and then selects the match that optimizes the similarity score (Ansari, 2018). The BWA-mem algorithm is specifically designed for use with Illumina reads from 70bp up to 1Mbp in length, and works with paired-end or single-end sequencing data ("bwa.1," 2013).

1.6 STAR

Spliced Transcripts Alignment to a Reference (or STAR), is an alignment program used to align RNA-seq data to a reference genome. Aligning transcriptomic data has its unique challenges compared to aligning genomic data, because transcripts are typically non-contiguous and need to be spliced together to form mature transcripts in the cell (Dobin et al., 2013). This makes aligning these transcript fragments a difficult task. STAR uses a similar seed-extend approach as BWA. Seeds are formed from finding the longest subsequence in a read that matches a region of the reference genome. Each read can have more than one seed. After extending seeds, allowing for mismatches and clipping unalignable regions, several seeds are clustered together based on proximity and then stitched together based on the best alignment to the reference genome of the full read (**piperAlignmentSTAR2017**).

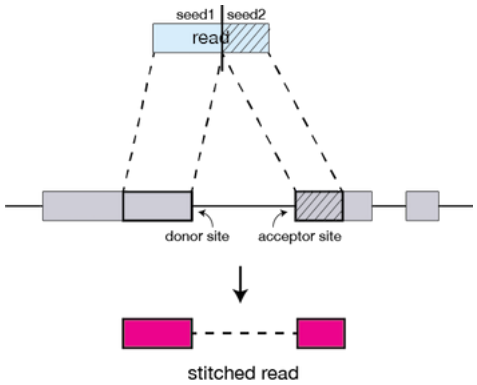

Figure 4: Reads are split into seeds based on the longest best match to the reference genome. After accounting for mismatches, gaps, and unalignable regions, seeds are clustered together based on distance and then joined based on the best alignment to the genome (schematic borrowed from (**piperAlignmentSTAR2017**)).

1.7 TPMCalculator

Determining if a gene is functional and being actively transcribed is key to the understanding of horizontal gene transfer events. One way to do this is to determine the abundance of mature transcripts for each gene. This is usually done by quantifying the number of transcripts from RNA-seq data that map to a particular gene, normalized by gene length. This number is thought to be proportional to the expression levels of the gene: the more transcripts we find for a gene of a given length, the higher we expect its expression to be.

TPMCalculator is a software that can be used to calculate mRNA transcript abundance from a genome annotation file in Gene Transfer Format (GTF) and a file containing information about aligned RNA-seq reads in Binary Alignment Map (BAM) format (Vera Alvarez et al., 2019). Transcripts per million (TPM) is a measure of transcript abundance, and a growing alternative to the popular Fragments per Kilobase Million (FPKM) and Reads per Kilobase Million (RPKM) measures. TPM involves counting the number of reads mapping to a gene, normalizing for gene length, and then adjusting this number to the relative number of reads you would expect to map to the gene of interest if 1 million full-length transcripts were sequenced (Zhao et al., 2020). TPM allows for a simple comparison of gene expression levels between genes, without having to worry about gene-length or sample bias.

1.8 PAML

Phylogenetic Analysis by Maximum Likelihood (PAML) is a collection of programs designed to test evolutionary hypotheses of DNA and protein sequences using statistical maximum likelihood methods. PAML takes information from the user, such as multiple sequence alignments, phylogenetic trees, and a control file, to perform an array of different analyses (Yang, 2017). The control file contains information about the user's model of interest, file locations, and parameters to estimate.

PAML is capable of estimating species divergence time, substitution rates, base composition of sequences, codon frequency, and more. CODEML is a widely used program in the PAML package that can determine the rate of nonsynonymous to synonymous substitutions (dN/dS) of protein coding genes and how they vary along lineages. PAML also computes the log-likelihood of each model under consideration, which makes for easy comparison of the fit of two differing models using a likelihood ratio test (LRT) (Yang, 2007).

2 Methods and Results

2.1 Workflow Overview

The following methods section describes the process of discovering HGT events from publicly available data. I started with downloading all publicly available protein coding sequences for *Wolbachia* and arthropods from the National Center for Biotechnology Information (NCBI). I then determined a set of putatively transferred genes by finding similarities between *Wolbachia* and arthropod proteins. I gathered short read whole genome sequencing (WGS) data for this gene set to test each gene for 5 key features, shown in Table 1.

Feature	Function
Genome Coverage	Determining whether number of reads mapped for transferred genes is normal (no *Wolbachia* contamination)
Gene Expression	See if host organism is expressing transferred genes at all or differently than host genes
Poly-A Selection	Demonstrate that transferred genes have eukaryotic mRNA processing characteristics
Introns	Demonstrate that transferred genes have eukaryotic gene features
Flanking Genes	Show that transferred genes were sequenced between two eukaryotic genes (likely real insertions)

Table 1: This table depicts the 5 key features that we tested. All tests were designed to ascertain whether or not the putative transfers were similar to host genes in DNA and RNA sequencing data sets, and that they possessed eukaryotic gene features.

Below is a graphic depicting the technical details of the DNA and RNA-seq data workflow. Both custom Python scripts and publicly available bioinformatics tools were used to analyze and visualize data. While this section provides an overview of the technical steps of the workflow, it does not contain program parameters or best practices for each program.

In addition, the usage of these methods is highly variable and depends on the researcher's needs and interests. The parameters chosen for each program are not hard and fast rules and may be adjusted to allow greater flexibility or more conservative estimates. I have chosen to present a protocol as a methods section— it is difficult to prescribe rigid protocol steps when there are many ways to correctly configure this pipeline. In addition, from my understanding of literature and coursework in the field, it is uncommon to present step-by-step protocols for dry lab workflows.

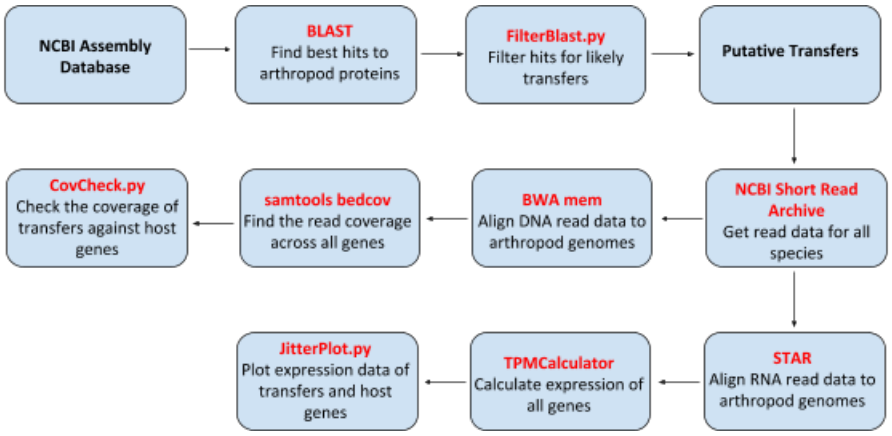

Figure 5: This graphic describes the Genome Coverage and Gene Expression steps described in Table 1. The workflow starts by getting publicly available protein assemblies from the NCBI, and then using BWA to align DNA data and STAR to align RNA data. Custom Python programs were used to analyze the number of reads mapped to each gene. The gene coverage (DNA) or gene expression (RNA) was plotted for the transferred genes and compared to the host distribution of gene coverage or expression.

2.2 Identification of Putative Transfer Events

2.2.1 Retrieving Arthropod Assemblies from NCBI

I obtained Arthropod assemblies directly from the NCBI's Assembly database. I used "Arthropoda" as the search term, and used the "Download Assemblies" feature to create a zipped tar archive of protein fastas from the GenBank database. I also downloaded sequences from all available *Wolbachia* strains in the same manner. Every download contained a multifasta file including all proteins from that assembly. One species could have more than one assembly.

2.2.2 Creating the *Wolbachia* BLAST Database and Executing BLAST Search

I extracted *Wolbachia* protein fastas from the tar archive. I then concatenated Wolbachia multifasta files into one file. I used the "makeblastdb" command, included in standard installations of NCBI's command-line BLAST program, to make a protein database using the concatenated fasta. I then BLASTed arthropod protein assembly fastas at the *Wolbachia* protein database using "blastp" with "-outfmt 6". The "-outfmt 6" flag specifies the fields in the output of the BLAST program, and in this case specifies these columns, separated by tabs: "qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore". I then filtered the blastp output files for all assemblies to keep only hits that had $\geq 90\%$ sequence identity and an alignment length ≥ 150 amino acids. This filtering step was intended to exclude proteins that may not be homologous or may have high sequence similarity in only a very small conserved region; however, it is possible that my threshold is too high and excludes additional homologous proteins.

2.3 Validation of Putative Transfer Events

2.3.1 Checking Read Coverage of Transferred Genes Against Host Genes

I conducted a test that compares the coverage of putative *Wolbachia* gene insertions and host genes. If the *Wolbachia* genes are real insertions into the host genome, as opposed to contamination, I would expect the coverage of putative transfers and host genes to be similar. I downloaded DNA sequencing data from the NCBI Short Read Archive for every species represented in the putative gene transfer set. I aligned short read sequences to the reference genome for each species using the BWA-mem algorithm with default parameters. I then used the RefSeq annotations in General Feature Format (GFF) and `samtools bedcov` to count the number of reads that aligned to each gene. The coverage of putatively transferred *Wolbachia* genes was compared to the coverage of all other genes in the host genome within 50 bp in length. Putatively transferred genes that fell in the outlier of the distribution of host gene coverage (below the 2.5th percentile and above the 97.5th percentile) failed this test.

2.3.2 Checking Gene Expression and Poly-A Selection

I also conducted a test to determine if any of the putative transfers were being expressed in the host and if putative *Wolbachia* transfers were expressed differently than host genes. The primary goal of this experiment was to determine if each transfer was expressed or not expressed; the level of expression was only a secondary consideration. I aligned RNA-seq data to the reference genome of each species using STAR. TPMCalculator was used to calculate the gene expression levels from the number of reads aligned to each gene. I then used a custom Python script, `JitterPlot.py`, to plot the expression levels of both putatively transferred *Wolbachia* genes and host genes. I performed a Mann-Whitney U test to determine whether the set of *Wolbachia* gene expression values were distributed differently from host genes. I also quantified the number of putative *Wolbachia* genes that originated from poly-A selected libraries. These genes have poly-A mRNA tails, a characteristic feature of eukaryotic gene expression, which might indicate that these are real insertions into the host genome.

2.3.3 Parsing Annotation Files for Eukaryotic Features

I was then interested to know if the putative transfers were embedded in a contig between two host genes. The annotation file for each species can be downloaded through links from the Assembly database. I wrote a custom Python script, contained in the `validate_hgt.ipynb` Jupyter Notebook, to parse the GFF annotation file. For each putative *Wolbachia* transfer, I compiled a list of all other genes on the same contig in the assembly. I then BLASTed the CDS for these genes at the non-redundant protein database. I parsed the output of this BLAST to check how many genes on each contig had best hits to the Eukaryota Kingdom. This would indicate that the genes sequenced on either side of the putative transfer were host genes as opposed to *Wolbachia* derived. Using the same script, for each putatively transferred gene I identified any spaces between exons in the annotation file. I inferred these to be introns, which is a primarily eukaryotic feature that suggests the putative insertion is real.

2.4 Quantifying Results

`MasterList.py` is available in the `validate_hgt.ipynb` Jupyter Notebook. This script takes in the output files from each step in the workflow and places putative transfers into confidence categories based on what and how many tests it passed. The confidence categories are designated as follows:

- **High Confidence**: Passes depth of coverage check AND (contains at least one intron AND is flanked by genes of eukaryotic origin, OR is expressed in poly-A selected RNA-seq libraries)
- **Medium Confidence**: Passes depth of coverage check AND (contains at least one intron OR is flanked by genes of eukaryotic origin)
- **Low Confidence**: Passes depth of coverage check
- **No Confidence**: Does not pass depth of coverage check

This is the definitive step in the workflow that ties together the results from each individual experiment. Others may choose to weigh the results of each experiment differently, which could result in other conclusions about the validity of each transfer. My approach attempts to minimize the effects of insufficient data for some species; for instance, poly-A selected RNA-seq data is only available for a few species in the putative transfer set, so I did not make this an absolute criterion for high confidence transfers. I believe the reasoning for my choices is sound and provides an adequate basis for the rest of my analyses.

2.5 Evaluating Selection on Protein-Coding Genes

I used NCBI Entrez Direct command-line tools to download the CDS protein fastas for each transfer and its closest *Wolbachia* homologs. I then used MACSE to align each transfer and its homologs in protein-space, accounting for frameshifts and stop codons that arose during alignment. TrimAl was used to trim out badly aligned sequences and gaps. I then realigned and refined the alignment with MACSE, removing any early stop codons that would truncate the protein. I wrote custom scripts to filter out badly aligned sequences or sequences that were too similar or too divergent, and then convert the resulting alignments into PHYLIP format. I used the PHYLIP multiple sequence alignment to generate phylogenetic trees for each gene set with IQTREE. The PHYLIP and tree file were used as input to PAML, which calculates the lineage-specific dN/dS ratios and the log-likelihoods for two different models. The null model (model=0) dictated that the dN/dS ratio along every branch was the same, while the alternative model (model=2) dictated that the dN/dS ratio along arthropod branches could be different from the background (*Wolbachia*) branches. I compared the log-likelihoods of each model using the likelihood ratio test (LRT), to determine which model fit the data better. Since PAML outputs the log-likelihoods of each model, the LRT simply requires the difference between the log-likelihood of the null and alternative model, multiplied by a factor of -2. A chi-squared test can then be used to determine whether there is a significant difference between the fit of the models.

3 Results

3.1 Characteristics of Putative Transfers

3.1.1 Identifying a Set of Putatively Transferred Genes

I found 1673 arthropod proteins that could be attributed to recent *Wolbachia* insertions into the host genome. After compiling a list of protein accessions and finding the records in the NCBI Protein Database, I discovered a majority of the proteins (1206 out of 1673) were tagged as bacterial proteins instead of arthropod proteins. Upon closer inspection of the protein records, I learned that all 1206 bacterial proteins were already annotated as *Wolbachia* insertions into the *Armadillidium vulgare* (pill-bug) genome. It was found that 83% of the *Wolbachia* genome had transferred to the *Armadillidium vulgare* genome relatively recently (Leclercq et al., 2016).

Additionally, I found potential *Wolbachia* insertions into the genomes of *Vollenhovia emeryi, Trichonephila clavipes, Cinara cedri, Laodelphax striatellus, Formica exsecta, Acromyrmex echinatior, Wasmannia auropunctata, Rhagoletis zephyria, Drosophila ananassae, Ceratina calcarata, Cryptotermes secundus, Solenopsis invicta, Drosophila eugracilis,* and *Sipha flava.* Horizontal gene transfer from *Wolbachia* to arthropods has previously been documented in *Armadillidium vulgare* (Leclercq et al., 2016), *Formica exsecta* (Dhaygude et al., 2019), and *Drosophila ananassae* (Klasson et al., 2014), although *Wolbachia* infections have been documented in almost all of species containing putative transfers. I hope that through my comprehensive screening process I will be able to document new instances of HGT from *Wolbachia* to arthropod hosts.

I also found that the arthropod proteins were most similar to *Wolbachia* proteins from supergroup A and B, and more distantly related to proteins from supergroup C and D. This is consistent with current research that describes supergroups A and B as endosymbionts of arthropods and supergroups C and D as endosymbionts of filarial nematodes (Lo et al., 2002).

Table 2 shows the number of putatively transferred protein-coding genes from each species. It is entirely possible that the transferred *Wolbachia* genes are not functional in the host genomes. In fact, large nuclear *Wolbachia* transfers to the *Drosophila ananassae* genome are predominantly located on the heterochromatic fourth chromosome (Klasson et al., 2014). This chromosome is made of tightly packed chromatin, suggesting an impact on gene expression of transferred genes. This region is also known for the abundance of non-coding DNA regions, repeats, and rearrangements, which could suggest that any transfers to this region will quickly degrade (Klasson et al., 2014).

Scientific Name	Common Name	Number of HGT protein hits
Wolbachia Endosymbiont of Armadillidium vulgare	*Wolbachia (endosymbiont of pill-bug)*	1206
Armadillidium vulgare	Pill-bug, roly-poly	12
Vollenhovia emeryi	Ant species from East Asia	170
Trichonephila clavipes	Golden silk orb-weaver spider	127
Cinara cedri	Cedar bark aphid	70
Laodelphax striatellus	Small brown planthopper	29
Formica exsecta	Narrow-headed ant	20
Acromyrmex echinatior	Panamanian leaf-cutter ant	13
Wasmannia auropunctata	Electric ant, little fire ant	7
Rhagoletis zephyria	Snowberry fruit fly	6
Drosophila ananassae	Fruit fly in the Melanogaster group	3
Ceratina calcarata	Spurred ceratina, carpenter bee species	2
Cryptotermes secundus	West Indian drywood termite	2
Solenopsis invicta	Red imported fire ant	1
Drosophila eugracilis	Fruit fly in the Melanogaster group	1
Sipha flava	Yellow sugarcane aphid	1

Table 2: This table contains the species names, common names, and number of hits identified by BLASTp after filtering out short alignments and sequences with low percent identity. Note that a large number of proteins were identified in *Armadillidium vulgare*, as opposed to the small number of proteins identified in *Drosophila ananassae*. *Wolbachia* HGT events have previously been identified in *Drosophila ananassae*, but the genes were found to be predominantly nonfunctional in their new locations (Klasson et al., 2014).

3.2 Confirming Real Transfers

3.2.1 Genome Coverage of Putative Transfers

As described above, the read depth coverage was computed for all genes, including putative *Wolbachia* transfers and genes believed to be host derived. The coverage of putative *Wolbachia* genes was then compared to the coverage of host genes. Each putative transfer falls into a certain percentile of the host gene coverage distribution. Assuming no biases in the insertion locations of putative transfers, we would assume that the read depth coverage of transfers should be evenly distributed with respect to the coverage of host genes. A majority of genes falling into the extremes of the distribution indicates that these genes had severely lower or higher coverage than host genes. This could occur if the samples were contaminated with a whole *Wolbachia* bacterium, whose genes would be sequenced along with the host genome but not actually inserted.

Figure 6 (page 13) shows the read coverage distributions of all putative transfers for each species. For the most part I observed even distribution of transfer coverage with respect to host gene coverage. Species like *Ceratina calcarata* and *Solenopsis invicta* had very few identified transfers. In these species with little data available, the histogram is not very informative about the uniformity of gene coverage. I kept genes for these species as long as they did not fall below the 5th percentile or above the 95th percentile of coverage. Some species, such as *Laodelphax striatellus* and *Cinara cedri*, showed a large number of genes in bins at the top and bottom of the distribution. The majority of these genes are excluded from future analyses through my filtering process.

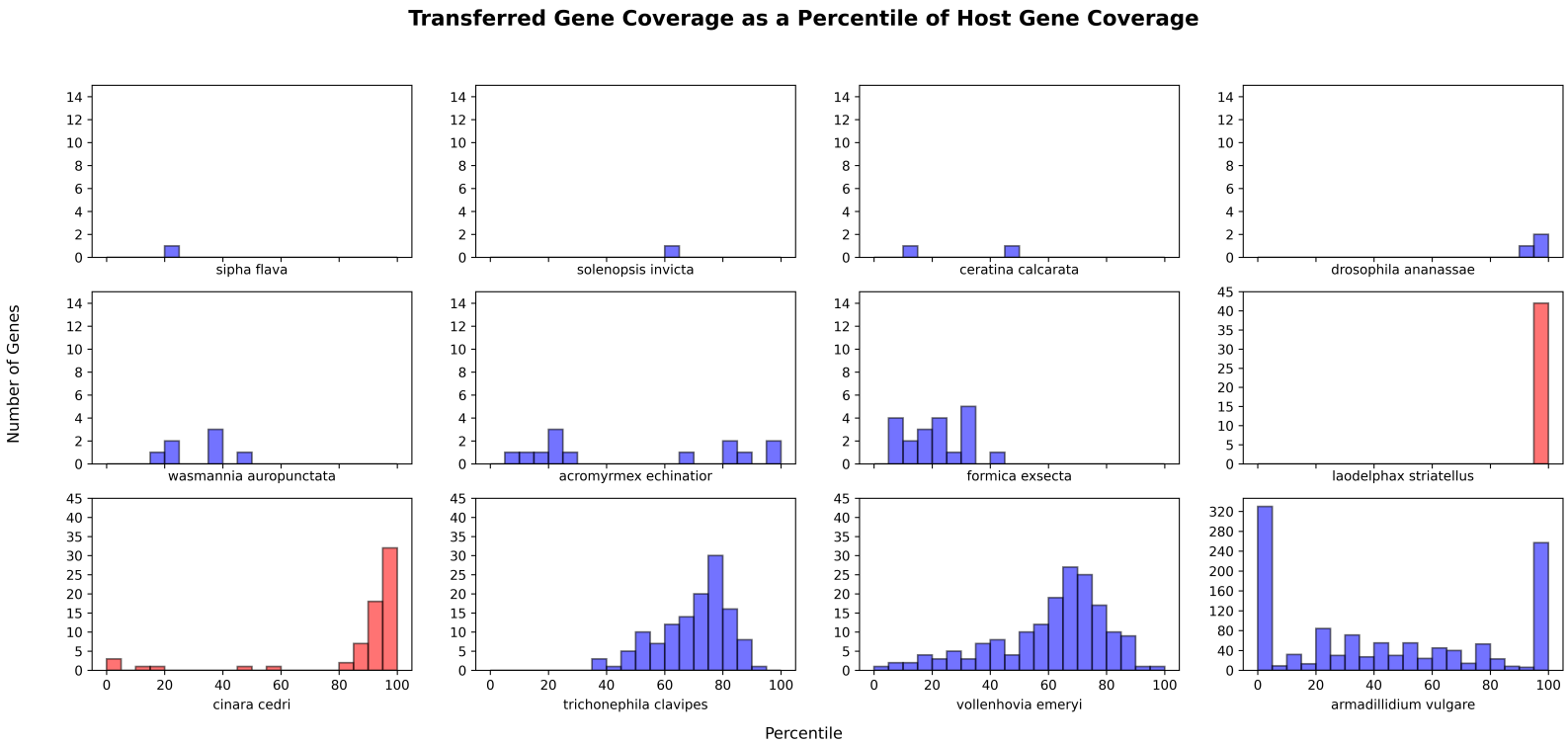

Figure 6: These plots show the coverage of putatively transferred genes compared to the host gene distribution. Histograms that show relatively even distribution across all bins (blue) indicate good evidence for real transfer events. Histograms where a majority of the species' transfers fall in the extremes of the distribution (red) indicate that the transfer likely isn't real.

3.2.2 Gene Expression of Putative Transfers

Putative *Wolbachia* gene transfers were expressed at lower levels than host genes. This is an interesting feature, as it potentially indicates that the function of these genes is not important to the host organism. It could also indicate that the genes have not been present in the host organism long enough to acquire certain eukaryotic gene features that would make successful transcription more likely. I used a Mann-Whitney U test to assess if the transferred genes for each species were expressed at lower levels than host genes as a whole.

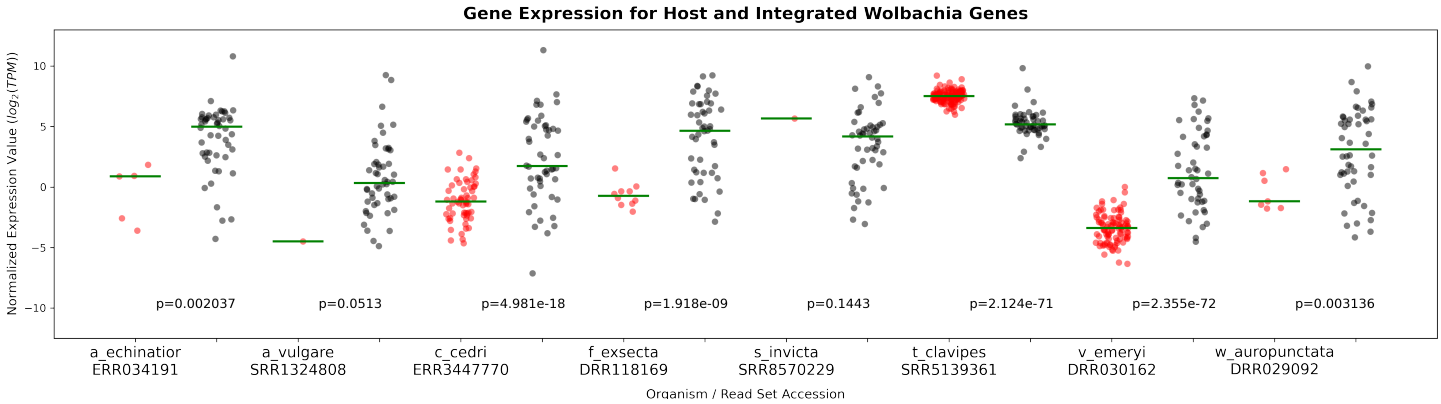

Figure 7: Normalized transcripts per million (TPM) are plotted for each gene (point). Location on the x-axis does not matter and was only performed to maximize visibility of each point. There are two distributions plotted for each species: gene expression data of putative transfers (red) and of a sub-sample of host genes (black). The median of each distribution is represented by a green line. P-values for the Mann-Whitney U test are displayed between the pair of distributions for each species.

As shown in Figure 7, putatively transferred genes as a whole were expressed at lower levels than host genes. For some species, only a small fraction of total putative transfers could be identified in RNA-seq data sets. Putative transfer expression tends to be more tightly packed at a certain level, whereas host gene expression tends to be more spread out. Interestingly, putative transfers are only expressed higher than host genes in one species: *Trichonephila clavipes*. This assembly was marked as anomalous by the NCBI Assembly Database with no reason cited; however, the genomic coverage of these genes seems to be normal (Figure 6, page 13).

This particular test will not be used to filter out potentially spurious transfers. Gene expression can vary considerably between RNA-seq datasets taken from the same organism, and some species only had one or two RNA-seq datasets to analyze. Therefore, the absence of putatively transferred genes from these datasets cannot be used as evidence to disqualify them. However, the presence of putative transfers in these datasets can strengthen the evidence for real insertions. Polyadenylation is a post-transcriptional modification to mRNA that results in poly(A) tails. These poly(A) tails are predominantly a eukaryotic gene feature, although they occur much less frequently in prokaryotes (Régnier and Marujo, 2013). Some of the RNA-seq data we used was constructed from RNA libraries that were poly(A)-selected, meaning only mRNAs with poly(A) tails were retained. The presence of a putative transfer in a poly(A)-selected dataset is strong evidence that it is a real transfer being transcribed with eukaryotic machinery.

3.2.3 Filtering for Real Transfers

After considering the results from the depth of coverage experiment and the gene expression experiment, I sought to categorize the putative transfers into conservative and lenient categories of genes, based on how many total tests they passed: high, medium, low, and no confidence. I used the aforementioned experiments along with counts of flanking genes of eukaryotic origin and introns, to filter all putative transfers. The counts of genes in each category and for each species is reported below in Figure 8 and Table 3. All further evolutionary analysis is restricted to genes from the medium and high confidence categories.

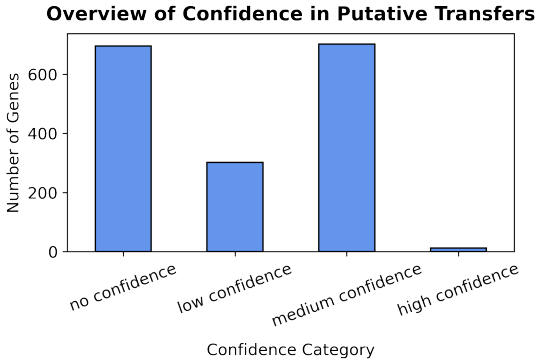

Figure 8: Counts of genes in each confidence category. Criteria for each category is detailed in the Quantifying Results section.

Scientific Name	No Confidence	Low Confidence	Medium Confidence	High Confidence
Wolbachia Endosymbiont of Armadillidium vulgare	586	41	579	0
Armadillidium vulgare	12	0	0	0
Vollenhovia emeryi	0	105	59	5
Trichonephila clavipes	0	68	55	4
Cinara cedri	11	52	3	0
Laodelphax striatellus	41	1	0	0
Formica exsecta	0	15	5	0
Acromyrmex echinatior	0	10	1	2
Wasmannia auropunctata	0	7	0	0
Rhagoletis zephyria	N/A	N/A	N/A	N/A
Drosophila ananassae	2	1	0	0
Ceratina calcarata	0	1	1	0
Cryptotermes secundus	N/A	N/A	N/A	N/A
Solenopsis invicta	0	1	0	0
Drosophila eugracilis	N/A	N/A	N/A	N/A
Sipha flava	0	0	0	1

Table 3: Table showing the breakdown of putative transfers into confidence categories by species. Species with only "N/A" values have putative transfers that were not detected in RNA or Whole Genome Sequences (WGS) datasets. Further investigation and more data is needed to confirm whether these transfers are real.

3.2.4 Evolution of Protein-Coding Genes

The PAML analysis revealed several interesting characteristics of HGT occurring between *Wolbachia* and arthropods. The dN/dS ratio and log-likelihood was calculated independently for each transferred gene and its homologs. I plotted the dN/dS ratios of the alternative model: on the y-axis I plotted the dN/dS ratio along arthropod lineages, and on the x-axis I plotted the dN/dS ratio of *Wolbachia* (background) lineages. If the likelihood ratio test showed that the alternative model was significantly better than the null model (p-value < 0.05), then I plotted the marker for that gene set in red. All other markers plotted in blue did not show significant differences in dN/dS between *Wolbachia* and arthropod lineages.

The data in Figure 9 was very different to what I was expecting. Given most gene sets did not have a significant difference between the fit of the null and alternative hypotheses, we would expect an IID distribution of the p-values, with the exception of a small peak at $p < 0.05$. However, as you can see in Figure 9, we saw a very small elevation of p-values at < 0.05 and a huge peak at > 0.95. This meant there was a large skew of the insignificant gene sets towards high p-values, which was very concerning as it likely indicated a large source of bias in the data.

When investigating this issue, I looked closely at the rate of synonymous substitutions of each gene. Because synonymous substitutions are thought to be neutral, they happen relatively constantly over time. It is common practice to filter out dS ratios below 0.02 and above 2 to avoid analyzing sequences that are too divergent or too similar. However, I had used this filtering protocol and still obtained p-value results skewed in the opposite direction to what I had expected. As you can see in Figure 9, despite filtering for $0.02 < dS < 2$, I still had a large skew towards small dS.

Upon inspection of my data, I realized that filtering the data on the total tree dS was resulting in my large dS and p-value peak. The total tree dS is the sum of the dS on all *Wolbachia* branches and all arthropod branches. However, the gene insertions into arthropods were so recent, that the rate of synonymous substitutions, and therefore dS, was very small. The dS of *Wolbachia* lineages, however, was within the expected range. When the total tree dS was used to filter the data, the *Wolbachia* dS effectively masked the low dS values seen across arthropod branches. When I filtered the data using the dS values for arthropod and *Wolbachia* separately, I found that p-value distribution looked more evenly distributed overall, with a skew toward significant p-values (as expected). The total dS distribution after filtering in this new manner also looked more evenly distributed. This was a heartening result as I initially thought I would have to completely scrap my data and start over. In reality, using aggregate data was masking important characteristics of each data component.

An interesting consequence of obtaining many low dS gene sets is that it points to a trend of gene loss. As mentioned earlier, low dS indicates that the arthropod sequences have not diverged sufficiently from the rest of the tree so that PAML can accurately model differences in dN/dS. For these sequences to be experiencing such little divergence, they had to have inserted extremely recently into arthropod lineages, or they must be lost at such a rate that we never get to observe ancient transfers.

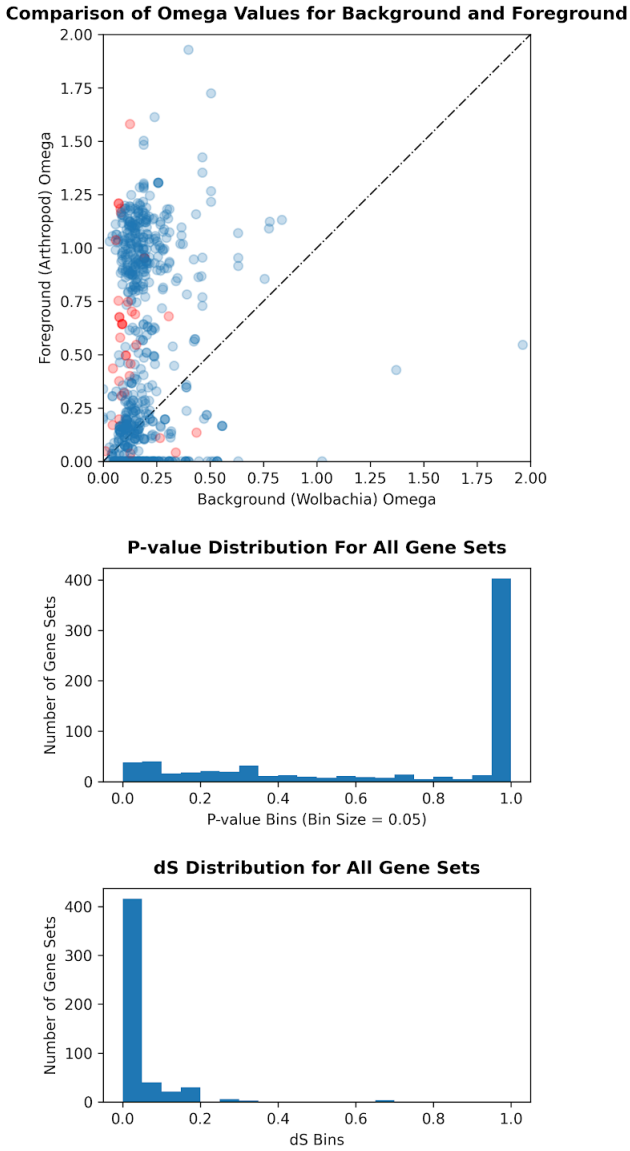

Figure 9: Plot of omega, p-values, and synonymous substitution rate for all genes. One point is the ratio between arthropod and *Wolbachia* omega for one gene in the main figure. This figure is based on filtering by the whole tree dS. A histogram of p-value distribution shows a significant skew towards large p-values. A histogram of dS values shows a large skew towards very small dS. Both histograms indicate a potential problem in the filtering method used, as ideally the p-values and dS values would be fairly evenly distributed.

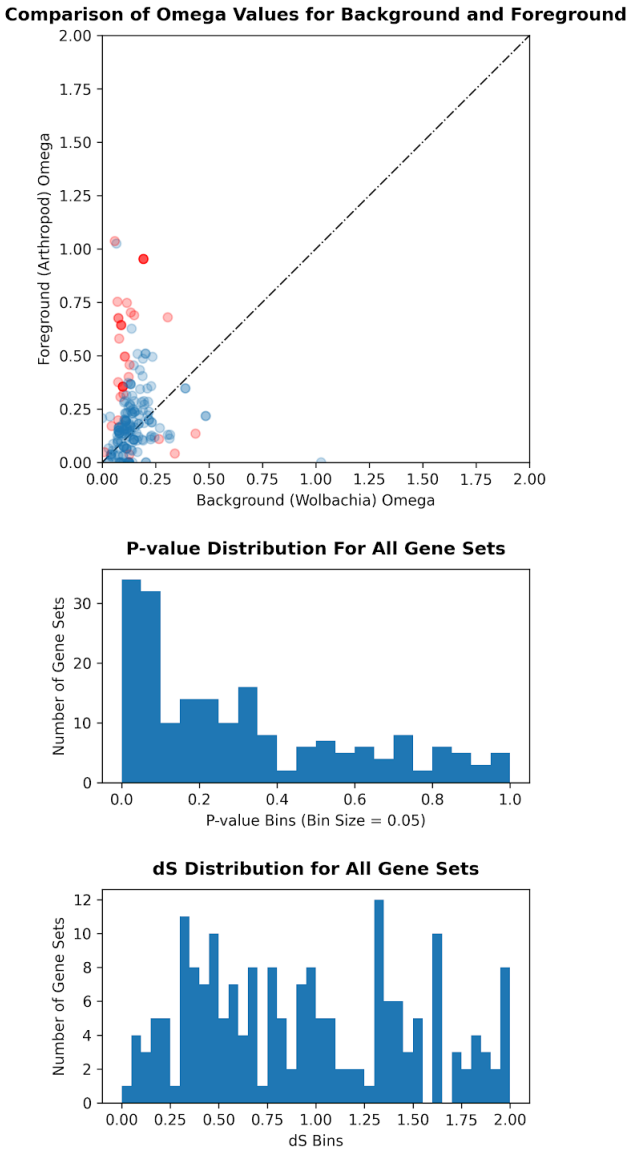

Figure 10: Plot of omega, p-values, and synonymous substitution rate for all genes. One point is the ratio between arthropod and *Wolbachia* omega for one gene in the main figure. This figure is based on filtering by the dS on each individual branch in the tree. Although there are far fewer genes that pass filtration, the dS values immediately appear to be more evenly distributed. The p-value distribution is more evenly distributed with a skew towards significant ($p < 0.05$) p-values. Such a skew is expected if there is a difference in selection for transferred genes overall, depending on the lineage.

4 Conclusion

I have found 1673 putative HGT events by analyzing existing protein sequence records available publicly on NCBI databases. These HGT events are distributed across 16 species, 13 of which do not have existing literature that supports the existence of these events. As expected, the arthropod proteins I tested were most closely related to *Wolbachia* proteins from supergroups A and B which are known to infect arthropods, as opposed to C and D which are known to infect filarial nematodes.

I developed several robust methods for validating putative transfers, which involved computing read-depth coverage, gene expression in transcripts per million, counting introns per gene, and determining the origin of genes flanking the insertion. Altogether, the results of these tests should be able to filter out sequence contamination and elucidate real gene transfers from *Wolbachia* to arthropods. Thus, I found 715 of the 1673 putative transfers were likely to be real insertions. I quantified the direction and strength of selection on each transferred gene by computing the dN/dS ratio, a measure of the nonsynonymous vs. synonymous substitution rates. These data revealed that while many transferred genes are undergoing purifying selection in arthropods, similarly to *Wolbachia*, there were some that appeared to be accumulating amino acid substitutions, although never reaching positive selection. These proteins are likely under reduced functional constraint and actively being lost from the genome, as I only observed recent transfers.

5 Future Work

In the future, I intend to estimate the rates of transferred gene gain and loss in arthropods. This will help us understand how frequent this event actually is and how long ago these specific events occurred. I would also like to investigate what kinds of genes are being transferred. I will compare the null functional distribution of *Wolbachia* genes to the functional distribution of transferred genes. This analysis could reveal that horizontal gene transfer is enriched for certain biological functions.

I intend on using the workflow developed here to detect and validate more HGT events in the future, especially when data is added or updated for organisms in public repositories. The high throughput nature of this project can help quickly identify HGT events as new data becomes available. A future goal therefore would be to automate the existing workflow to operate on any new data uploaded on the NCBI Assembly database. This would involve creating generalized pipelines that could download new assembly and annotation files into their own directories, and handle a variety of file naming conventions and data availability. This pipeline would have to be deployed on a server to run on a schedule. A feature could be added to notify assembly authors via email if spurious endosymbiont genes or contamination were detected in their assembly.

Understanding the nature of *Wolbachia* HGT in arthropods is key as *Wolbachia* spreads throughout terrestrial ecosystems. In particular, this work could prove to be instrumental in predicting the effects of artificially introducing *Wolbachia* to natural arthropod populations as a biological control method. Only more time and observation can elucidate the fascinating relationship between *Wolbachia* and arthropod genomes.

References

Ansari, A. (2018, January 4). *Smith-waterman algorithm*. Retrieved February 18, 2021, from https://bioinfoguide.com/index.php/algorithms-and-methods/11-smith-waterman-algorithm

Balding, D. J., Bishop, M. J., & Cannings, C. (Eds.). (2007). *Handbook of statistical genetics* (3rd ed) [OCLC: 837064903]. Wiley.

BLAST® command line applications user manual. (2008). National Center for Biotechnology Information (US).

Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biology*, *15*(1), 98. https://doi.org/10.1186/s12915-017-0434-y

Bwa.1. (2013, March 8). Retrieved February 18, 2021, from http://bio-bwa.sourceforge.net/bwa.shtml

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Danchin, E. G. J. (2016). Lateral gene transfer in eukaryotes: Tip of the iceberg or of the ice cube? *BMC Biology*, *14*(1), 101. https://doi.org/10.1186/s12915-016-0330-x

Dhaygude, K., Nair, A., Johansson, H., Wurm, Y., & Sundström, L. (2019). The first draft genomes of the ant formica exsecta, and its wolbachia endosymbiont reveal extensive gene transfer from endosymbiont to host. *BMC Genomics*, *20*(1), 301. https://doi.org/10.1186/s12864-019-5665-6

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dunning Hotopp, J. C. (2011). Horizontal gene transfer between bacteria and animals. *Trends in genetics : TIG*, *27*(4), 157–163. https://doi.org/10.1016/j.tig.2011.01.005

Hotopp, J. C. D., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P., Torres, M. C. M., Giebel, J. D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R. V., Shepard, J., Tomkins, J., Richards, S., Spiro, D. J., Ghedin, E., Slatko, B. E., Tettelin, H., & Werren, J. H. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes [Publisher: American Association for the Advancement of Science Section: Report]. *Science*, *317*(5845), 1753–1756. https://doi.org/10.1126/science.1142490

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution [Number: 8 Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, *9*(8), 605–618. https://doi.org/10.1038/nrg2386

Kent, A. G., Vill, A. C., Shi, Q., Satlin, M. J., & Brito, I. L. (2020). Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial hi-c [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *11*(1), 4379. https://doi.org/10.1038/s41467-020-18164-7

Klasson, L., Kumar, N., Bromley, R., Sieber, K., Flowers, M., Ott, S. H., Tallon, L. J., Andersson, S. G. E., & Dunning Hotopp, J. C. (2014). Extensive duplication of the wolbachia DNA in chromosome four of drosophila ananassae. *BMC genomics*, *15*, 1097. https://doi.org/10.1186/1471-2164-15-1097

Kosiol, C., Vinař, T., Fonseca, R. R. d., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six mammalian genomes [Publisher: Public Library of Science]. *PLOS Genetics*, *4*(8), e1000144. https://doi.org/10.1371/journal.pgen.1000144

Leclercq, S., Thézé, J., Chebbi, M. A., Giraud, I., Moumen, B., Ernenwein, L., Grève, P., Gilbert, C., & Cordaux, R. (2016). Birth of a w sex chromosome by horizontal transfer of wolbachia

bacterial symbiont genome. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(52), 15036–15041. https://doi.org/10.1073/pnas.1608979113

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Lo, N., Casiraghi, M., Salati, E., Bazzocchi, C., & Bandi, C. (2002). How many wolbachia supergroups exist? [Publisher: Oxford Academic]. *Molecular Biology and Evolution*, *19*(3), 341–346. https://doi.org/10.1093/oxfordjournals.molbev.a004087

Nikoh, N., Tanaka, K., Shibata, F., Kondo, N., Hizume, M., Shimada, M., & Fukatsu, T. (2008). Wolbachia genome integrated in an insect chromosome: Evolution and fate of laterally transferred endosymbiont genes [Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab]. *Genome Research*, *18*(2), 272–280. https://doi.org/10.1101/gr.7144908

Régnier, P., & Marujo, P. E. (2013). *Polyadenylation and degradation of RNA in prokaryotes* [Publication Title: Madame Curie Bioscience Database [Internet]]. Landes Bioscience. Retrieved December 11, 2020, from https://www.ncbi.nlm.nih.gov/books/NBK6253/

Sequence alignment. (n.d.). Retrieved February 18, 2021, from https://www.cs.mcgill.ca/~rwest/wikispeedia/wpcd/wp/s/Sequence_alignment.htm

Theissen, U., & Martin, W. (2006). The difference between organelles and endosymbionts. *Current Biology*, *16*(24), R1016–R1017. https://doi.org/10.1016/j.cub.2006.11.020

Vera Alvarez, R., Pongor, L. S., Mariño-Ramírez, L., & Landsman, D. (2019). TPMCalculator: One-step software to quantify mRNA abundance of genomic features. *Bioinformatics*, *35*(11), 1960–1962. https://doi.org/10.1093/bioinformatics/bty896

Werren, J. H. (1997). Biology of wolbachia. *Annual Review of Entomology*, *42*, 587–609. https://doi.org/10.1146/annurev.ento.42.1.587

Werren, J. H., Baldo, L., & Clark, M. E. (2008). Wolbachia: Master manipulators of invertebrate biology [Number: 10 Publisher: Nature Publishing Group]. *Nature Reviews Microbiology*, *6*(10), 741–751. https://doi.org/10.1038/nrmicro1969

Woods, L. C., Gorrell, R. J., Taylor, F., Connallon, T., Kwok, T., & McDonald, M. J. (2020). Horizontal gene transfer potentiates adaptation by reducing selective constraints on the spread of genetic variation. *Proceedings of the National Academy of Sciences*, *117*(43), 26868–26875. https://doi.org/10.1073/pnas.2005331117

Woolfit, M., Iturbe-Ormaetxe, I., McGraw, E. A., & O'Neill, S. L. (2009). An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium wolbachia pipientis. *Molecular Biology and Evolution*, *26*(2), 367–374. https://doi.org/10.1093/molbev/msn253

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang, Z. (2017). PAML: Phylogenetic analysis by maximum likelihood, 67.

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, *22*(12), 2472–2479. https://doi.org/10.1093/molbev/msi237

Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*, *26*(8), 903–909. https://doi.org/10.1261/rna.074922.120

Zug, R., & Hammerstein, P. (2012). Still a host of hosts for wolbachia: Analysis of recent data suggests that 40% of terrestrial arthropod species are infected. *PLoS ONE*, *7*(6). https://doi.org/10.1371/journal.pone.0038544

A GitHub

I have made all code available for viewing and downloading at https://github.com/rafihorse/Wolbachia-Thesis. This includes instructions for how I downloaded the data and made custom databases, selected filtering parameters, and fine-tuned programs / workflows.

B Acquiring Data

All arthropod data can be acquired from https://www.ncbi.nlm.nih.gov/datasets/, using "Arthropoda" as the search term. *Wolbachia* data can be acquired from https://www.ncbi.nlm.nih.gov/assembly/, using "Wolbachia" as the search term.

Protein and gene accessions for arthropod sequences are also available on the GitHub in `prot_gene_ids.txt`.