UNIVERSITY OF CALIFORNIA
SANTA CRUZ

GENOMIC, PROTEOMIC AND COMPUTATIONAL APPROACHES TO THE STUDY OF HOST-MICROBE SYSTEMS

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in
BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Gabriel A. Penunuri

August 2026

The Dissertation of Gabriel A. Penunuri
is approved:

Professor Rebecca DuBois, Chair

Professor Russell Corbett-Detig

Professor Shelbi Russell

Professor Benedict Paten

Donald Smith
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Tables and Figures ..... **iv**

Abstract ..... **v**

Genomic, Proteomic, and Computational Approaches to the Study of Host-Microbe Systems v

Acknowledgments ..... **vii**

Introduction **1**

Chapter I : EMS Mutation and SNP Detection in Intracellular Wolbachia Genomes **5**

1.1 Abstract 5

1.2 Introduction 6

1.3 Material and Methods 8

1.4 Results 21

1.5 Discussion 31

1.6 Data Availability 33

1.7 Supporting Information 34

Chapter II: Structural Mimicry Screen In Host-Microbe Systems or Identifying Effector Candidates in Host-Microbe Systems through Protein Structure **39**

2.1 Abstract 39

2.2 Introduction 40

2.3 Methods 46

2.4 Results 55

2.5 Discussion 79

2.6 Data Availability 81

2.7 Supporting Information 82

Chapter III: Automating Systematic Literature Searches **89**

3.1 Introduction 89

3.2 Methods 91

3.3 Results 104

3.4 Discussion 108

Conclusion **112**

List of Supplementary Files **115**

References **115**

List of Tables and Figures

Figure 1.1. Treating wMel-infected JW18 cells with EMS induces C>T and G>A mutations in the wMel genome.....23

Figure 1.2. EMS treatment duration was positively associated with mutation rate, whereas functional aspects of the wMel genome exhibited no association. 25

Figure 1.3. EMS mutations in the wMel genome exhibit biased sequence context..... 30

Figure S1.1. Single Base Mutation Spectra..... 34

Figure S1.2. EMS mutation density across the genome..... 35

Figure S1.3. Circle sequencing error rate decrease..... 36

Figure S1.4. Analysis of EMS Mutation Rates Across the wMel Genome and Genes..... 37

Figure S1.5. Sequence Bias Model Evaluation..... 38

Figure 2.1. Alignment Space Under Expansive Control Method..... 56

Figure 2.2. Mimic Candidate Q5F2U4.....63

Figure 2.3. Mimic Candidate O25981.....67

Figure 2.4. Mimic Candidate Q73HU7 Overview.....71

Figure 2.5. Eato Mimic Candidates Overview.....73

Figure 2.6. Mimic Candidate P61189..... 79

Figure S2.1. Expansive Control Alignment Space Helicobacter pylori-Human. 82

Figure S2.2. Full Q5ZW23 Structure Alignment..... 84

Figure S2.3. Full O25525 Structure Alignment.....84

Figure S2.4. Full Q73IF8 Structure Alignment.....85

Figure S2.5. Mimic Candidate Q73GY6 Overview..... 86

Figure S2.6. Mimic Candidate P61189 Phylogenetic Tree..... 87

Figure 3.1. Human versus LLM rubric grades.....106

Figure 3.2 Distribution of literature based mimicry support scores for Legionella-Human structural alignments..... 108

Abstract

Genomic, Proteomic, and Computational Approaches to the Study of Host-Microbe Systems

Gabriel Alejandro Penunuri

Host-microbe systems are core to some of biology's most consequential interactions, from the pathogens that drive infectious disease to intracellular symbionts mitigating vector-borne disease transmission. Yet unlike the model organisms that have driven most of modern molecular biology, the microbes at the center of these interactions are rarely genetically tractable. Many intracellular bacteria cannot be cultured outside a host, resist standard tools for genetic manipulation, and are annotated largely by homology to distantly related free-living relatives. This dissertation develops genomic, proteomic, and computational methods to work around this lack of infrastructure and contribute techniques and tools to the study and further understanding of host-microbe systems. Using *Wolbachia* cultured in Drosophila melanogaster cell lines, I demonstrate that chemical mutagenesis can be used to perturb intracellular genomes leaving a detectable mutational signal. I employ a low error rate sequencing technique to record and model the mutational landscape left by the mutagen ethyl methanesulfonate (EMS) demonstrating its use for mutagenesis screens of intracellular bacteria. I next utilize structural proteome datasets to screen host-microbe proteomes for strong candidates of molecular mimicry, microbe proteins that have coevolved a eukaryotic-like domain or structure and suggest use for host manipulation or microbe survival in the host environment. Building off of this

screen for novel effectors through structural alignments I develop and test a distributed computing system for performing large scale systematic literature reviews. Altogether these projects represent generalizable approaches to the study of host-microbe systems reaching from classically studied and thoroughly understood to novel and non-model systems.

Acknowledgments

There are many important people in my professional and personal circles that have made this work possible and meaningful to me. Although I might only name some of them here, the greater research community that I am very fortunate to have been a part of has been an invaluable source of inspiration to my work as a scientist and growth as a person. I am grateful to my advisors Russ Corbett-Detig and Shelbi Russell for the guidance throughout this process. I am also grateful to my other dissertation committee members Benedict Paten and the chair of my committee Rebecca DuBois, for their time, insight and support.

I cannot express enough how thankful I am for the wonderful labmates I have had the honor of working alongside, Chris, Alex, Lily and Cade, who have not only fostered my love of science more than anyone, but have become amazing friends over the course of our graduate studies. I owe so much to friends old and new that have been with me throughout my PhD studies. I cannot say for sure this work would have been possible without the wonderful community that I have come to know here in Santa Cruz. With this in mind I would like to thank and acknowledge my family as I can say without a doubt this work would have not been possible without them. I especially want to acknowledge my grandparents, David and Joan, for truly inspiring and supporting my pursuit of education, my curiosity and all that has come from that.

Introduction

Host-microbe interactions underlie some of the most consequential processes in biology, medicine, and agriculture. Pathogenic bacteria manipulate host cell biology to establish infection and cause disease, while symbiotic bacteria can be harnessed to control populations that transmit disease. *Wolbachia pipientis*, an intracellular alphaproteobacterium infecting a majority of arthropod species, exemplifies this latter case: it is now deployed at scale to suppress mosquito-borne diseases and agricultural pests (Hoffmann et al., 2011; Salama et al., 2024; Vooght et al., 2022). Other host-associated bacteria, including *Legionella pneumophila* and *Helicobacter pylori*, cause substantial human disease burden through molecular mechanisms of host manipulation that remain incompletely characterized. Advancing our ability to identify and functionally validate the genes responsible for these interactions is therefore a problem with direct relevance to human and agricultural health.

Despite this relevance, the organisms that mediate host-microbe interactions are frequently the least experimentally accessible. Many host-associated bacteria are obligate intracellular symbionts or pathogens, dependent on host cells for survival and unable to be cultured axenically. Their obligate lifestyle imposes complex growth requirements that make genome-wide functional approaches difficult to scale (Feng and Wilson, 2024; Masson and Lemaitre, 2020), and even within stable culture conditions, these microbes have proven extremely difficult to genetically manipulate (McCutcheon et al., 2024). As a consequence, gene function in these organisms is

frequently inferred by homology to distantly related free-living bacteria (Hönigschmid et al., 2018). This approach provides little insight into genes whose functions have specifically diverged to mediate interaction with a host, and that leaves the identification of essential and host-interacting genes severely constrained (McClure et al., 2017). Meaningful progress has been made in a small number of obligate intracellular pathogens, including *Rickettsia* (Lamason et al., 2018) and *Chlamydia trachomatis* (Hawk et al., 2025; Kari et al., 2011; Kokes et al., 2015), but this progress has been slow, expensive, and largely organism-specific rather than generalizable.

Wolbachia is a particularly striking illustration of this gap. Its genome is enriched in eukaryotic-like elements (Bordenstein and Bordenstein, 2022), it induces a range of unique host-associated phenotypes, and the wMel strain from *Drosophila melanogaster* is important in applied mosquito biological control strategies (Dufault et al., 2022). Despite this, direct experimental characterization of *Wolbachia* gene function remains rare, and most functional assignments are still inferred through homology rather than direct evidence. This disparity between biological significance and functional annotation is not unique to *Wolbachia*, it recurs across host-associated bacteria generally, and motivates the central aim of this dissertation: to develop methods that make functional and mechanistic insight into host-microbe systems achievable even where conventional genetic tools are unavailable.

This dissertation approaches that problem from three complementary directions, corresponding to the stages of a typical functional genetics workflow; generating genetic variation, identifying candidate genes of interest, and interpreting those

candidates against the existing literature, each addressed for a setting in which standard tools do not apply.

The first chapter addresses the earliest of these stages: generating genome-wide genetic variation in an organism with no established mutagenesis protocol. Chemical mutagenesis with ethyl methanesulfonate (EMS) has been a foundational tool for genetic discovery for decades, but its application to obligate intracellular bacteria has been limited because EMS-induced mutation frequencies fall three to four orders of magnitude below the error rates of standard sequencing platforms. Using *Wolbachia* wMel stably infecting a cultured *Drosophila* cell line, I combine EMS mutagenesis with circle sequencing, an ultra-low error rate library preparation method, to detect and model EMS-induced mutation for the first time in an intracellular bacterial genome. This establishes a quantitative foundation for future mutagenesis screens in *Wolbachia* and, by extension, other intracellular symbionts.

The second chapter addresses candidate gene identification in the absence of tractable wet-lab screening. Structural mimicry, in which a microbial protein evolves a structure convergent with a host protein, is a recurring mechanism by which pathogens and symbionts manipulate host biology (Mondino et al., 2020), but its identification has historically relied on painstaking case-by-case biochemical characterization. The recent availability of proteome-scale structure predictions (Jumper et al., 2021; Varadi et al., 2022) and ultrafast structural alignment tools (van Kempen et al., 2022) makes it possible, for the first time, to screen entire host and microbe proteomes for structural mimicry in silico. I develop and validate such a screen using the well-characterized *Legionella pneumophila*-human system, then

apply it to *Helicobacter pylori* and *Wolbachia* wMel, identifying candidate mimic proteins with plausible roles in host manipulation and providing functional evidence for one candidate's role in *Drosophila* oogenesis.

The third chapter addresses a bottleneck common to both preceding approaches and to candidate-gene screens generally: the difficulty of interpreting large candidate lists against the published literature. A full systematic literature review at the scale produced by modern omics and structural screens is increasingly infeasible to perform manually (Waffenschmidt et al., 2023). Recent work has shown that large language models can be deployed for structured literature retrieval and evidence grading at scale, in contexts ranging from candidate gene prioritization in blood transcriptomics to early-stage drug target discovery and broader biomedical systematic review (Khan et al., 2025; Guo et al., 2025; Cassell et al., 2025), an approach made increasingly tractable by advances in efficient, reasoning-capable language models (Jiang et al., 2024; Sun et al., 2024). I develop a distributed, LLM-based literature review pipeline that maps candidate gene identifiers to literature-searchable identities, retrieves and grades relevant primary research against structured rubrics, and synthesizes evidence for candidate relevance.

Together, these three approaches span genetic perturbation, candidate gene discovery, and literature-scale evidence synthesis. They are demonstrated primarily in *Wolbachia*, and in the case of the structural mimicry screen, also applied to *Legionella pneumophila* and *Helicobacter pylori*. Each method is intended to be applicable to other host-associated microbes that remain difficult to study with conventional genetic tools.

[The following is an adaptation of a publication which has been published in mSystems 10.1128/msystems.00660-26]

Chapter I: EMS Mutation and SNP Detection in Intracellular Wolbachia Genomes

1.1 Abstract

Endosymbiotic bacteria such as Wolbachia pose significant challenges to genetic and molecular investigation due to their obligate intracellular lifestyle and complex growth requirements. Current understanding of their protein biology relies heavily on functional assignments inferred by homology, which may not reflect the specific roles endosymbiont proteins play within the host. This work addresses the need for robust genetic perturbation by demonstrating the successful application and detection of chemical mutagenesis in the genome of the wMel strain of Wolbachia grown within a stably infected Drosophila melanogaster JW18 cell line. To accurately detect EMS-induced mutations in a large, unsorted cell culture population, in which mutations remain at very low allele frequency, we implemented an ultra-low error rate sequencing strategy, circle sequencing. This technique enables confident detection of EMS-induced single nucleotide polymorphisms (SNPs) that would be swamped by the inherent error rates of standard next-generation sequencing. Circle sequencing library preparations successfully revealed a clear EMS mutation signal in treated cells, characterized by a significant enrichment of canonical C/G>T/A transitions. Furthermore, we present a model explaining observed EMS mutation rates across the genome for different sequence contexts. These findings show that EMS-treatment can successfully leave detectable mutation signals in intracellular

genomes, and offer promise for the future development of protocols to make targeted edits in Wolbachia genomes.

1.2 Introduction

Interest in harnessing intracellular symbionts for large scale biological control continues to grow, particularly as *Wolbachia* demonstrates effectiveness in controlling mosquito-borne diseases and arthropod pest populations (Hoffmann et al., 2011; Salama et al., 2024; Vooght et al., 2022). Because of active bioengineering development around obligate intracellular symbionts, particularly *Wolbachia*, accurate functional annotation of endosymbiont genes is critical for progress. Current functional assignments are frequently inferred through homology to genes in free-living organisms, which may not accurately reflect their specific roles within the symbiotic context (Hönigschmid et al., 2018).

The absence of robust methodologies for *Wolbachia* genetic perturbation severely constrains the identification of essential genes and the elucidation of specific gene functions (McClure et al., 2017). Intracellular bacteria present significant obstacles to comprehensive investigation; their obligate lifestyle requires complex growth conditions that make genome-wide approaches difficult to scale (Feng and Wilson, 2024; Masson and Lemaitre, 2020). Even within stable growth conditions, researchers have found intracellular microbes extremely difficult to genetically manipulate, greatly limiting the scope of experimental interventions (McCutcheon et al., 2024). While impressive advances have been made in obligate intracellular pathogens, such as *Rickettsia* (Lamason et al., 2018) and *Chlamydia trachomatis*

(Hawk et al., 2025; Kari et al., 2011; Kokes et al., 2015), progress is slow and expensive. With interest in obligate intracellular symbionts such as *Wolbachia* continuing to increase, the development of a high throughput, genome-wide random mutagenesis methodology would be useful to establishing foundational functional genetic research.

Chemical mutagenesis with ethyl methanesulfonate (EMS) has been a foundational tool in genetics for over a century, enabling systematic functional genetic discoveries across a wide range of plant and animal systems(Chen et al., 2023). In contrast, the application of EMS to obligate intracellular bacteria has been limited, with only a small number of studies reporting genome-wide detection of EMS-induced single nucleotide polymorphisms (SNPs)(Brothwell et al., 2016; Kari et al., 2011; Kokes et al., 2015; McClure et al., 2017). A primary barrier has been that EMS induces mutations at frequencies three to four orders of magnitude lower than the intrinsic error rates of standard next-generation sequencing platforms(Snyman et al., 2021; Wang et al., 2022; Yao et al., 2025). This disparity between mutation frequency and technical detectability has prevented the use of EMS in intracellular microbes, despite its proven value for functional genetic analysis in nearly all other biological contexts. Overcoming this limitation requires sequencing approaches capable of substantially suppressing technical error in order to reveal rare, chemically induced mutations.

Here, we overcame these technical barriers by combining chemical mutagenesis with high-fidelity sequencing in a tractable intracellular bacterial system. We leverage a *Drosophila melanogaster* cell line stably infected with the *Wolbachia w*Mel strain,

which permits controlled chemical perturbation of the endosymbiont genome. To enable reliable detection of rare EMS-induced variants, we implement circle sequencing(Dahl et al., 2004; Lou et al., 2013), a low-error rate library preparation strategy that generates consensus sequences from repeated reads of individual DNA molecules, thereby reducing sequencing artifacts well below standard Illumina error rates. Using this approach, we demonstrate clear detection of EMS mutagenesis signals in treated cell cultures, including enrichment of canonical EMS-associated substitutions and non-random sequence context biases in mutation positions. By demonstrating the first intentional mutagenesis and high-confidence variant detection in *Wolbachia*, this approach establishes a critical prerequisite for future targeted genome editing in these endosymbionts.

1.3 Material and Methods

1.3.1 Cell Lines

We maintained *Wolbachia w*Mel-infected immortalized JW18 (Serbus et al., 2012) *Drosophila melanogaster* cell lines in 4 mL of media composed of Sheilds and Sang M3 Medium (Sigma S3652) and 10% v/v Fetal Bovine Serum (Gibco A3160502) in plug-seal T25 flasks (Corning 430639), incubated at 25-26°C.

Cell cultures were maintained on a seven-day schedule, splitting at a 1:2 ratio for *w*Mel-infected lines and 1:6 to 1:4 ratio for uninfected lines. For cell passaging, we removed old media, added fresh media, and then dislodged adherent cells by scraping the flask surface with a sterile bent Pasteur pipette (Fisher, 1367820D). Resuspended cells were transferred at the appropriate ratio to new T25 flasks, and

the volume adjusted to 4 mL with fresh media. Infection states were monitored with fluorescent *in situ* hybridization against *w*Mel 16S ribosomal RNA and whole genome sequencing (as in (Mirchandani et al., 2024b)).

1.3.2 EMS Treatment

We first quantified the *w*Mel-infected JW18 cell concentration in culture with a hemocytometer and mixed media with cells with fresh media in T25 flasks for a final concentration of 1.69e6 cells/mL. These cells were incubated at 26°C for 1 day prior to treatment.

To create EMS-treated and control samples, we added either 20 µl of fresh 0.5% v/v EMS (Sigma M0880) or mock-treatment media, respectively, to 3980 µl of Shields and Sang M3 media supplemented with 10% FBS and sterilized through a 0.2 µm syringe filter. We then removed and replaced the old media from previously quantified cell cultures with either EMS or mock media. Flasks were then incubated at 25-26°C for 3, 5, or 7 days.

1.3.3 Library Prep and Sequencing

We obtained cell pellets from the EMS-treated and controlled cell cultures by scraping the flasks, pipetting 1 mL into 1.5 mL epi tubes, and pelleting the cells by centrifugation at 10,000 × g for 3 minutes at 4℃.

We amplified genomic DNA templates by rolling circle amplification (RCA) using the procedure described in the initial circle sequencing publication by Lou et al.

2013(Lou et al., 2013), with several adjustments. DNA purification steps were done with a SPRI magnetic bead protocol instead of the QIAGEN MinElute clean up kit. We sheared gDNA resuspended in TE buffer over two rounds of five minutes with a spin down between rounds. We set the duty cycle to 30%, peak power to 60W, and 1000 cycles per burst. Phosphorylation was done with the T4 PNK as described in the supplemental methods of Lou et al. (Lou et al., 2013). The DNA was then denatured and flash frozen according to the Lou et al. protocol. For the circularization reactions we used the same master mix except we added no more than 3.5 pmol of ssDNA instead of 5 pmol. Linear DNA remaining in the reaction mix was digested according to the Lou et al. protocol. For the rolling circle amplification (RCA) reactions we skipped the random primer annealing reaction, and instead used 2.5 µl of 10 µM random hexamer primers in the RCA reaction mix. We used 2.5 µl (instead of 1 µl) of 10 mM dNTP and did not include the inorganic pyrophosphatase. After performing RCA, we treated about 20-30 ng of RCA DNA products with 1 µl of DNA Exonuclease I (NEB M0293S) and 0.5 µl of DNA Exonuclease III (NEB M0206S) for 1 hour at 37ºC to digest linear fragments. We then purified samples using SPRI magnetic beads and Qubit quantified with the dsDNA HS Assay Kit (Thermo Fisher Q32851).

We prepared libraries from the RCA products by mixing 10 ng of RCA products with 2 µl charged Tn5 transposase (Tn5 enzyme was expressed and purified in-house) in 12 µl of molecular grade water and 4 µl of 5x TAPS-PEG 8000 for 8 minutes at 55ºC (Mirchandani et al., 2024a). The reaction was halted by adding 0.2% SDS and incubated at room temperature for 7 minutes. We set up PCR reactions using the KAPA Biosystems HiFi Polymerase Kit (Roche) and amplified for 12 cycles using

uniquely indexed i5 and i7 primers. We ran these PCR products on a Thermo Fisher 2% E-Gel EX Agarose Gel (G401002) and cut between ~400-700 bp. The gel was purified with the NEB Monarch DNA Gel Extraction Kit (NEB #T1020) and quantified using the dsDNA HS Assay Qubit Kit and the Agilent TapeStation D1000 tapes (5067-5582). We prepared two pools of circle sequencing Tn5 libraries: one pool containing 14 purified and indexed libraries from 5-day samples, and a second pool containing 14 purified and indexed libraries from 3 and 7-day samples. We sequenced both pools using 150-bp paired-end reads across two Illumina HiSeq 4000 lanes at Fulgent Genetics.

1.3.4 Circle Sequencing Processing

We amplified genomic DNA templates by rolling circle amplification (RCA) using the procedure described in the initial circle sequencing publication by Lou et al. 2013(Lou et al., 2013), with several adjustments. DNA purification steps were done with a SPRI magnetic bead protocol instead of the QIAGEN MinElute clean up kit. We sheared gDNA resuspended in TE buffer over two rounds of five minutes with a spin down between rounds. We set the duty cycle to 30%, peak power to 60W, and 1000 cycles per burst. Phosphorylation was done with the T4 PNK as described in the supplemental methods of Lou et al. (Lou et al., 2013). The DNA was then denatured and flash frozen according to the Lou et al. protocol. For the circularization reactions we used the same master mix except we added no more than 3.5 pmol of ssDNA instead of 5 pmol. Linear DNA remaining in the reaction mix was digested according to the Lou et al. protocol. For the rolling circle amplification (RCA) reactions we skipped the random primer annealing reaction, and instead used 2.5 µl

of 10 µM random hexamer primers in the RCA reaction mix. We used 2.5 µl (instead of 1 µl) of 10 mM dNTP and did not include the inorganic pyrophosphatase. After performing RCA, we treated about 20-30 ng of RCA DNA products with 1 µl of DNA Exonuclease I (NEB M0293S) and 0.5 µl of DNA Exonuclease III (NEB M0206S) for 1 hour at 37ºC to digest linear fragments. We then purified samples using SPRI magnetic beads and Qubit quantified with the dsDNA HS Assay Kit (Thermo Fisher Q32851).

We prepared libraries from the RCA products by mixing 10 ng of RCA products with 2 µl charged Tn5 transposase (Tn5 enzyme was expressed and purified in-house) in 12 µl of molecular grade water and 4 µl of 5x TAPS-PEG 8000 for 8 minutes at 55ºC (Mirchandani et al., 2024a). The reaction was halted by adding 0.2% SDS and incubated at room temperature for 7 minutes. We set up PCR reactions using the KAPA Biosystems HiFi Polymerase Kit (Roche) and amplified for 12 cycles using uniquely indexed i5 and i7 primers. We ran these PCR products on a Thermo Fisher 2% E-Gel EX Agarose Gel (G401002) and cut between ~400-700 bp. The gel was purified with the NEB Monarch DNA Gel Extraction Kit (NEB #T1020) and quantified using the dsDNA HS Assay Qubit Kit and the Agilent TapeStation D1000 tapes (5067-5582). We prepared two pools of circle sequencing Tn5 libraries: one pool containing 14 purified and indexed libraries from 5-day samples, and a second pool containing 14 purified and indexed libraries from 3 and 7-day samples. We sequenced both pools using 150-bp paired-end reads across two Illumina HiSeq 4000 lanes at Fulgent Genetics.

1.3.5 Mutation Rate Estimation

1.3.5.1 Per-Sample Rates

We collected mutation counts from mpileup files with a custom python script (https://github.com/gabepen/ems_effect). We loaded site-level mutation counts from these mpileup files. Sites were restricted to G/C positions. For C sites, only C→T mutations were counted; for G sites, only G→A mutations. We constructed an exclusion mask representing genomic positions where mutated sites appeared in more than one control sample. We used this mask to exclude these positions from all downstream analysis. We extracted sequence contexts (3mers, 5mers and 7mers) from the reference genome, and G-centered contexts were reverse-complemented to C-centered for canonicalization. We used a finite-sites model to estimate an initial per-sample mutation rate for mutation spectra visualization:

$$\hat{\mu}_{\text{high}} = \frac{A_{\text{total}}}{D_{\text{total}}} \tag{1}$$

Where A_{total} is the count of alternative alleles across all positions and D_{total} is the total sequencing depth across all sites.

We modeled per-sample mutation rates using generalized linear models (GLMs) with log(depth) as an offset starting with a background subtraction:

$$Y_{i,\text{adj}} = \max(0, Y_{i,\text{raw}} - \alpha \cdot d_i) \quad (2)$$

These adjusted counts are then used in a GLM:

$$\log(\mathbb{E}[\max(0, Y_{i,\text{raw}} - \alpha \cdot d_i)]) = \log(d_i) + \beta_0 \quad (3)$$

Where:

$Y_{i,raw}$ = raw mutation count at a site *i*

$Y_{i,adj}$ = alpha-adjusted mutation count at a site *i*

d_i = sequencing depth at site *i*

α = background false positive rate (estimated from controls)

β_0 = intercept parameter, estimated per sample

We then estimated rates as:

$$\hat{\mu}_s = \exp(\hat{\beta}_{0,s}) \quad (4)$$

1.3.5.2 Mutation Rate Estimation Across Genomic Categories

We estimated mutation rates for intergenic, synonymous, and non-synonymous rates with a negative binomial GLM. We opted to use a negative binomial distribution to account for the overdispersion and low mutation counts in the finite-sites dataset. We fit this model with treatment and category covariates, including interactions:

$$\log(\mathbb{E}[Y_i]) = \beta_0 + \beta_{\text{treatment}} \cdot T_i + \beta_{\text{syn}} \cdot S_i + \beta_{\text{nonsyn}} \cdot N_i + \beta_{\text{treatment}\times\text{syn}} \cdot T_i \cdot S_i + \beta_{\text{treatment}\times\text{nonsyn}} \cdot T_i \cdot N_i + \log(d_i) \quad (5)$$

We modeled these same genomic categories with an alternative approach to capture the absolute instead of the proportional effect of treatment on category rates starting with a control only model to establish a baseline:

$$\log(\mathbb{E}[Y_i^{\text{control}}]) = \beta_0^{\text{control}} + \beta_{\text{syn}}^{\text{control}} \cdot S_i + \beta_{\text{nonsyn}}^{\text{control}} \cdot N_i + \log(d_i) \quad (6)$$

Next we calculate the expected mutation excess with treatment:

$$Y_{\text{excess}}[i] = Y_{\text{observed}}[i] - \mathbb{E}[Y_i^{\text{control}}] \quad (7)$$

We then fit two GLMs using Gaussian family with an identity link, the first a model for constant absolute effect:

$$\mathbb{E}[Y_{\text{excess}}[i]] = \beta_{\text{treatment}} \times d_i \times T_i \quad (8)$$

And second an interaction model:

$$\mathbb{E}[Y_{\text{excess}}[i]] = \beta_{\text{treatment}} \cdot d_i \cdot T_i + \beta_{\text{treatment}\times\text{syn}} \cdot d_i \cdot T_i \cdot S_i + \beta_{\text{treatment}\times\text{nonsyn}} \cdot d_i \cdot T_i \cdot N_i \quad (9)$$

We also estimated category rates for treatment time with a negative binomial GLM, where no-label is the 5 day reference:

$$\log(\mathbb{E}[Y_i]) = \beta_0 + \beta_{3d} \cdot D_{3d,i} + \beta_{7d} \cdot D_{7d,i} + \log(d_i) \quad (10)$$

1.3.5.3 Mutation Sequence Context Analysis

We modeled per-site EMS mutation counts using generalized linear models (GLMs) with sequence-context features.

Site-level mutation counts were loaded from mpileup files. Sites were restricted to G/C positions. For C sites, only C→T mutations were counted; for G sites, only G→A mutations. Sites in an exclusion mask (those appearing in multiple control samples) were excluded. Sequence contexts (3mers, 5mers and 7mers) were extracted from the reference genome, and G-centered contexts were reverse-complemented to C-centered for canonicalization.

We used Negative Binomial GLMs to model the sequence context of EMS mutations. The per-site EMS mutation count served as the response variable, and log(depth) was included as an offset term. The model evaluated the effect of the experimental conditions by including both the treated and control groups as primary covariates.

$$\log(\mathbb{E}[Y_i]) = \beta_0 + \beta_{\text{treatment}} \cdot T_i + \sum_j \beta_j \cdot X_{i,j} + \log(d_i) \quad (11)$$

Where:

Y_i = raw mutation count at a site i

T_i = treatment indicator

$X_{i,j}$ = sequence context features (vary by model type)

d_i = sequencing depth at site i

$\beta_{\text{treatment}}$ = treatment effect on log scale

β_j = sequence context coefficients

β_0 = intercept parameter, baseline log rate for control sites

To account for sequence context dependencies and to determine what context was the best predictor of mutation bias, we implemented five distinct encoding schemes:

1. Positional: We encoded sequence context using 12 binary features representing the base identity (T, G, or C) at each of the 4 varying positions in the canonicalized 5-mer, with A as the reference category for each position. The center position is always C, so 3 features are constant; 12 features vary across sites. Coefficients represent log-enrichment relative to A at each position.
2. 3-mer: We encoded context using 16 binary features representing all possible C centered 3-mers spanning positions -1, 0, 1.
3. 5-mer: We encoded context using a one-hot encoding approach, resulting in 256 binary features. 5-mer refers to the -2,-1,0,1,2 positions with respect to an EMS mutation at position 0.
4. Positional-3-mer: We encoded context using 32 binary features, representing the left and right 3-mers covering the -2, -1, 0 and 0, 1, 2 positions

respectively. This model also evaluates C centered mutations only resulting in fixed ‘0’ positions in both left and right 3-mers.

5. 7-mer: We encoded context using a one-hot encoding approach, with a theoretical maximum of 4,096. 7-mer refers to the -3,-2,-1,0,1,2,3 positions with respect to an EMS mutation at position 0. To manage memory usage, the model required the input contexts to be split into separate groups deterministically by first sorting unique 7-mers. Model fitting was performed independently for each split, and the resultant metrics were subsequently aggregated to obtain the final summary statistics.

After fitting, we predicted mutation counts for each site under control and treated conditions. We converted predicted counts to per-base mutation rates by dividing by depth, then aggregated observed and predicted rates by the relevant sequence-context category for the model using depth-weighted means. We evaluated fit by comparing observed versus predicted aggregated rates using scatter plots and correlation metrics.

1.3.5.4 Cross-Validation Evaluation (Out-of-sample)

We used 5-fold cross-validation to measure out-of-sample predictive performance. We split sites into five folds with sklearn.model_selection.KFold using shuffling (random_state=42). For each fold, we trained each model on four folds and evaluated it on the held-out fold, so every site served as test data exactly once. For each held-out fold, we computed mean squared error (MSE), mean absolute error (MAE), Pearson correlation (r), and a negative-binomial deviance statistic. We

summarized performance across folds by reporting the mean (and, when available, the standard deviation) of each metric.

For the 7-mer split model, we controlled memory usage by fitting the model in groups of 7-mer categories. We deterministically partitioned unique 7-mers into groups using a round-robin assignment over sorted 7-mer categories. Within each cross-validation fold, we fit one GLM per group using only training sites in that group. We generated held-out predictions by applying the corresponding group model to held-out sites in that group (each site received a prediction from exactly one split model), and we computed cross-validation metrics from the concatenated held-out predictions across groups.

1.3.5.5 Prediction Accuracy Metrics

We also evaluated prediction accuracy on the evaluation dataset by comparing observed mutation counts (y) to model-predicted expected counts (mu) while incorporating the log(depth) offset. For the 7-mer split model, we produced predictions by routing each site to the split model corresponding to its 7-mer category, and we evaluated performance only on sites with a valid 7-mer context.

We computed MSE, MAE, RMSE, and MAPE; Pearson correlation (r) and Spearman correlation (rho) with p-values; R-squared and adjusted R-squared; a deviance statistic; MSLE and RMSLE; and median-threshold classification metrics (accuracy, precision, recall, and F1). We also reported the number of evaluated sites and the mean observed and predicted counts.

1.3.5.6 Genome-wide EMS mutation enrichment

We analyzed local genomic variation independent of genomic categories by dividing the genome into fixed-size, non-overallaping 5-mer windows and testing if the observed rates deviated from the 5-mer model for EMS mutation context and rate.

For each mutated site *i* in window W_k we extract the 5-mer context and predict the expected mutation rate using the pre-fitted 5-mer model. We then fit a site-level GLM with a 5-mer offset and a treatment covariate:

$$\log(\mathbb{E}[Y_i]) = \log(\mu_{5\text{mer},i,s}) + \log(d_i) + \beta_{0,k} + \beta_{\text{treatment},k} \cdot T_i \tag{12}$$

Where:

Y_i = observed mutation count at a site *i*

T_i = treatment indicator

$\mu_{5\text{mer,i,s}}$ = pre-computed expected rate from the 5-mer model

d_i = sequencing depth at site *i*

$\beta_{\text{treatment,k}}$ = window-specific treatment effect

$\beta_{0,\text{k}}$ = window-specific intercept

1.3.5.7 Genome-wide EMS mutation enrichment in the context of transcription

We tested the correlation between EMS mutation enrichment and rates of transcription by adapting the fixed-size window models above to encompass genic regions and performed a regression analysis between the predicted window mutation

rates and TPM values for each corresponding gene. To obtain TPM values for the wMel genome, we downloaded the wMel-infected JW18 cell RNAseq dataset produced by Jacobs *et al.* (2025) (PRJNA1240446) and regenerated pseudoalignments according to their methods(Jacobs et al., n.d.). We used ordinary least squares (OLS) linear regression fitting controls and treated samples separately.

1.4 Results

1.4.1 EMS Mutation Signal

Using high-fidelity circle sequencing(Lou et al., 2013), we identified an elevated mutation rate in the EMS-treated *w*Mel genome cultured in *Drosophila* JW18 cells, relative to control treatments. As Wolbachia exist at high copy numbers within host cells, each sample represents a population of bacteria; accordingly, all mutation analyses are conducted at the population level across the treatment group rather than in individual bacterial genomes. Following Illumina sequencing and demultiplexing of the 18 EMS-treated and 10 control samples, we obtained 4,691,282,072 total reads estimated to be derived from 619,355,451 rolling circle amplification (RCA) products amounting to 2,755,257,101 total read copies (Table S1.1, S1.2). When we compared EMS-treated and control consensus circle sequencing reads to the *Wolbachia w*Mel reference genome, GCF_000008025.1, we identified 4 SNPs fixed across all samples and excluded these from the analysis. Treated samples showed a significant enrichment of the canonical EMS-induced transitions C>T and G>A ($p<0.001$, Mann-Whitney U test, Figure 1.1A, S1.1). These

mutations affected the entire genome (Figure S1.2, Table S1.3). We also noted changes to the physical characteristics of the EMS-treated cell lines when compared to the control lines (Figure 1B, 1C). Treated cells exhibited morphological signs of stress including slowed growth, small cell size, condensed cytoplasm, and disrupted membranes.

We represented the mutation spectra as a depth-normalized proportion of sites mutated for each transition type using the finite-sites model, indicating that the EMS mutation signal was only clearly detectable through low error rate sequencing methods (Figure 1.1A). In previously published mutation rates, EMS has been shown to induce per-site mutation rates of ~1×10^{-6} to ~1×10^{-2} across different organisms (Snyman et al., 2021; Wang et al., 2022) while the minimum medium error rate on an Illumina platform is around 5×10^{-3} per-site (Schirmer et al., 2016). We calculated the error rate of our circle sequencing libraries for different counts of RCA tandem repeats used to correct errors observed on the first repeat. The initial repeat consensus versus reference mismatch rate was 6.297×10^{-4} which was reduced to 6.451×10^{-5} with the incorporation of a second tandem repeat in the consensus and further down to 3.049×10^{-5} with the incorporation of four repeats (Figure S1.3).

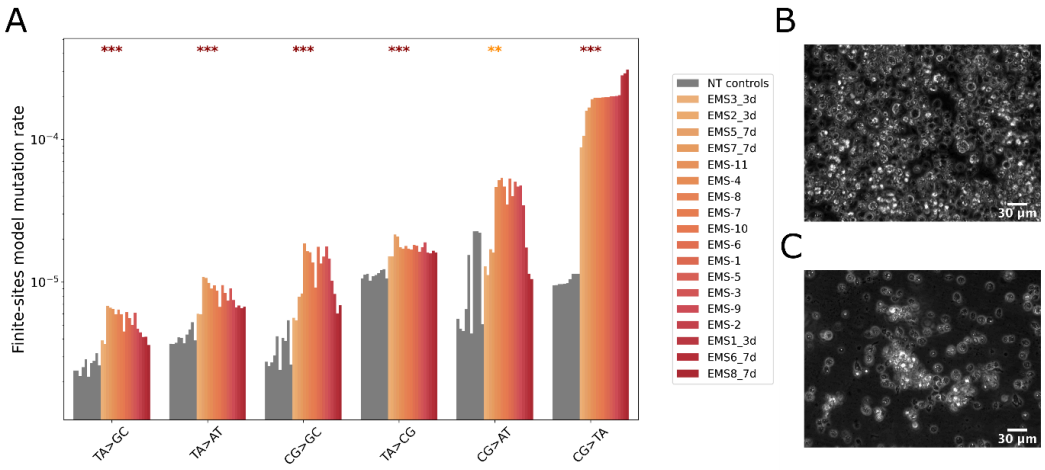

Figure 1.1. Treating *w*Mel-infected JW18 cells with EMS induces C>T and G>A mutations in the *w*Mel genome.

A) EMS mutation spectra demonstrating a large difference between control (grey) and treatment (orange-red gradient by experiment) at the canonical EMS transitions types of C>T and G>A. Asterisks denote one-sided Mann–Whitney tests (EMS > NT) on per-sample strand-collapsed finite-sites mutation rates, with Benjamini–Hochberg FDR across the six mutation classes: **, q < 0.01; ***, q < 1×10^{-4}. Representative tissue culture micrographs of *w*Mel-infected JW18 cells **B)** after the control treatment and **C)** after EMS treatment.

1.4.2 EMS Mutation Rates

We estimated mutation rates across all samples using site-level GLMs, obtaining estimates of per-sample rates ranged from 2.23×10^{-5} to 3.17×10^{-5} mutations per base pair for controls (mean: 2.71×10^{-5} mutations/bp) and 9.44×10^{-5} to 3.24×10^{-4} for treated samples (mean: 2.05×10^{-4} mutations/bp) (Figure 1.2A). Treated samples exhibited significantly elevated mutation rates with average increase of 4.73-fold (95% CI: 4.65 - 4.80) compared to controls ($p < 0.001$, negative binomial GLM with treatment covariate).

The treatment durations used in the EMS dosage protocol had a proportionate effect on mutation rate when we modelled the interaction between treatment duration and mutation rate using a negative binomial GLM with days of treatment as a covariate. We grouped control and treated samples as 5-day (treated=11, control=3), 3-day (treated=3, control=3), and 7-day (treated=4, control=4). All pairwise comparisons between groups were significant with the group-level estimates showing an increase in mutation rates as treatment time increased (Figure 1.2B). However, we note substantial inter-sample variance over each treatment duration category.

We determined that the functional sequence category has a negligible differential effect on EMS mutation incidence by comparing mutation rates across three different functional coding categories: intergenic, synonymous and non-synonomous. We calculated mutation rates by incorporating a baseline mutation rate for each genomic category using the 5-mer model described in the mutation sequence context analysis portion of the methods. This accounted for any significant divergence in sequence content between coding and intergenic regions. We estimated control mutation rates for these three categories to be 4.63×10^{-5}, 2.24×10^{-5} and 2.26×10^{-5} mutations per base, respectively. We estimated treated rates to be 2.16×10^{-4} mutations per base pair for intergenic sites, 2.03×10^{-4} mutations per base pair for synonymous sites and 1.95×10^{-4} mutations per base for non-synonomous sites (Figure 1.2C). Although these rates represent large differences between categories in the proportional changes between the control and treated samples they amount to similar absolute effect, as quantified by the number of mutated bases added by EMS to the background control rates. For the same three categories, the absolute EMS effect, or mutations per base added by treatment, were 1.16×10^{-4}, 1.73×10^{-4}, and

1.62×10^{-4}. While the constant effect and interaction models produce statistically significant effect differences ($p < 10^{-154}$, likelihood ratio test), they only explain about ~5% of the variation. This indicates that the rate differences between intergenic and genic regions are biologically insignificant.

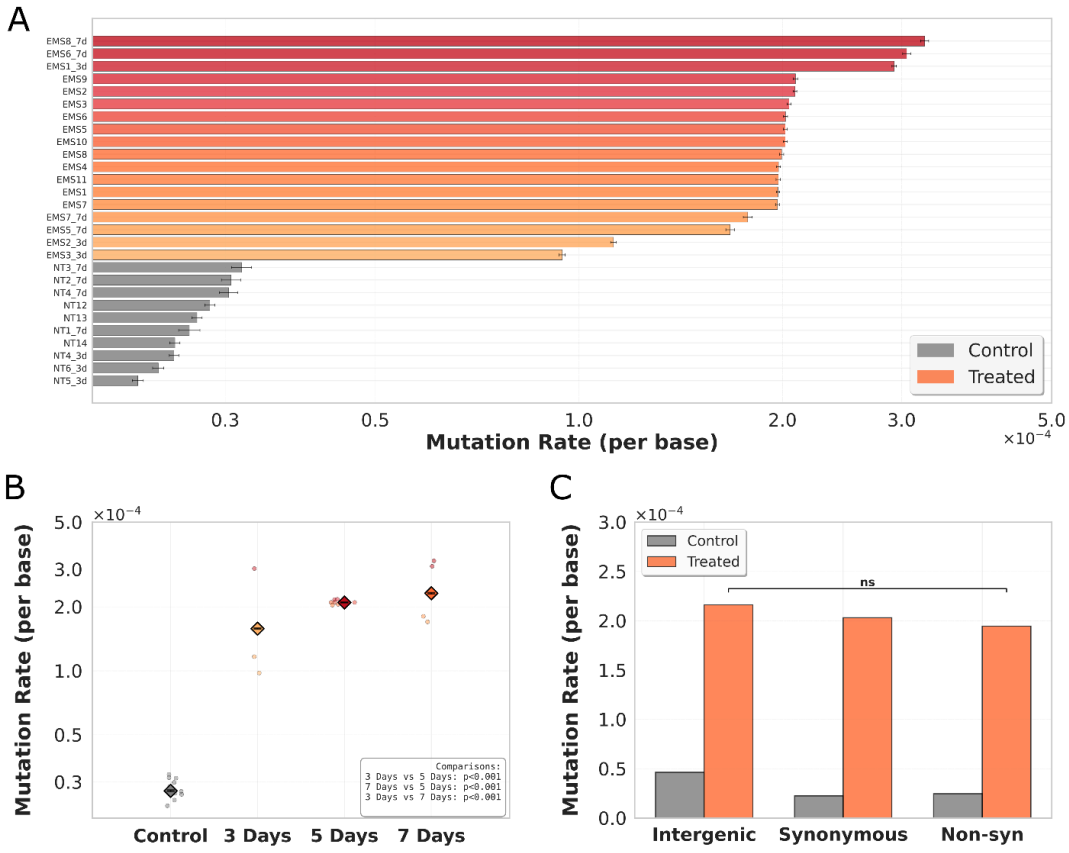

Figure 1.2. EMS treatment duration was positively associated with mutation rate, whereas functional aspects of the wMel genome exhibited no association.

A) Per-sample EMS mutation rates for treated and control samples as estimated by fitting a negative binomial GLM. Treated samples exhibit ~5-fold higher mutation rates than control samples. B) Treatment time had a significant impact on mutation rate when treatment time was modeled as a covariate of mutation rate, with longer treatment times correlating with higher mutation rates. C) Intergenic versus coding mutation rate analysis with treated rates normalized by 5-mer content in each region category. Similar absolute rate differences among coding categories suggests that

EMS treatment added the same number of mutations per base pair across genomic categories.

We found limited evidence to suggest that EMS preferentially mutagenizes particular windows of the *w*Mel genome, or is more likely to impact regions that are open for transcription. We calculated EMS mutation rates within 2500 base pair non-overlapping windows across the genome using our k-mer aware genomic window rate model (Eq. 12) to identify biologically meaningful hot or cold spots. We used a summed EMS mutation count normalized by sequencing depth across all samples as input into the model. We established the genome-wide mean as the average of all window mutation rates calculated by the genomic window rate model (Eq. 12). A binomial test of window clusters, comparing mutation rates to the genome-wide mean, identified no stretches of EMS hot or cold spots beyond a single window. One such example includes a weak difference between mutation rate of the prophage regions, WO-A and WO-B, and the rest of the genome. The rate ratio between the prophage regions and non-prophage was 0.966 (95% CI 0.961-0.971). Paired per-sample comparisons indicated a consistent prophage-associated reduction in treated samples (median per-sample rate ratio 0.962; Wilcoxon $p = 5.3 \times 10^{-5}$) but not in the controls ($p = 0.16$). We also calculated EMS mutation rates for each gene and performed a regression analysis with *w*Mel *Wolbachia* gene expression data collected by Jacobs et al. (2025)(Jacobs et al., n.d.). The expression data were very weakly correlated with EMS mutation rate ($R^2 = 0.025$, Figure S1.4). This combined with the random spatial distribution of significantly elevated windows that have apparent clustering suggests that mutation rate variation in this system is influenced by unmodeled effects. While the treatment induces a global increase in mutation rates, the local variation appears to be largely random, with no evidence for transcription-associated mutagenesis.

1.4.3 EMS Mutation Sequence Context Bias

We selected the 5mer model for determining sequence bias in EMS mutations because it provided the best balance between model fit and generalizability across multiple performance metrics (Table S1.4). The 7mer model achieved the lowest AIC among all evaluated models with a ΔAIC to the next lowest, the 5mer model, of ~18,000 (Table S1.4). However, the 7mer model also reported the highest overall BIC indicating the capturing of a significant amount of noise. The 5mer model reported the lowest BIC with a ΔBIC of ~5000 to the second lowest, the positional-3mer. Thus our prediction accuracy and cross-validation tests supported the 5mer model over the 7mer, 3mer, and positional models (Figure S1.5). For prediction accuracy the 5mer model had the lowest RMSE (0.546), and highest pearson R (0.348) with the positional-3mer model performing second best (RMSE = 0.299, pearson R = 0.343) and the 7mer third (RMSE = 0.300, pearson R = 0.342); while simpler models (positional, 3mer) showed reduced predictive accuracy. Cross-validation (CV) metrics further suggested although the 7mer model was adept at producing rankings, overfitting likely hindered its calibration. The 7mer model had the highest CV-MSE (R=0.319) and lowest CV-pearson (R=0.339). The 5mer model performed the best in these metrics with R=0.298 and 0.367, respectively.

All models performed poorly at predicting EMS mutations from sequence context alone, indicating that additional factors contribute to EMS mutation bias. Nevertheless, our fitted 5mer model identified significant enrichment of certain sequence contexts at EMS-mutated sites (Figure 1.3A, Table S1.5). AT dinucleotides were enriched at the -2 and -1 positions directly upstream of EMS-mutated bases

(Figure 1.3C), while dinucleotides containing A or T were depleted at the downstream +1 and +2 positions where C or G dinucleotides were enriched. Our results suggest that although C and G bases are enriched near EMS-induced SNPs, this enrichment occurs predominantly downstream rather than uniformly surrounding the mutated base. This directional asymmetry distinguishes EMS-induced mutations from spontaneous replication-associated mutations, where G/C enrichment has been reported symmetrically around mutated sites(Schroeder et al., 2016; Sung et al., 2015), suggesting a fundamentally different mechanism of mutation accrual.

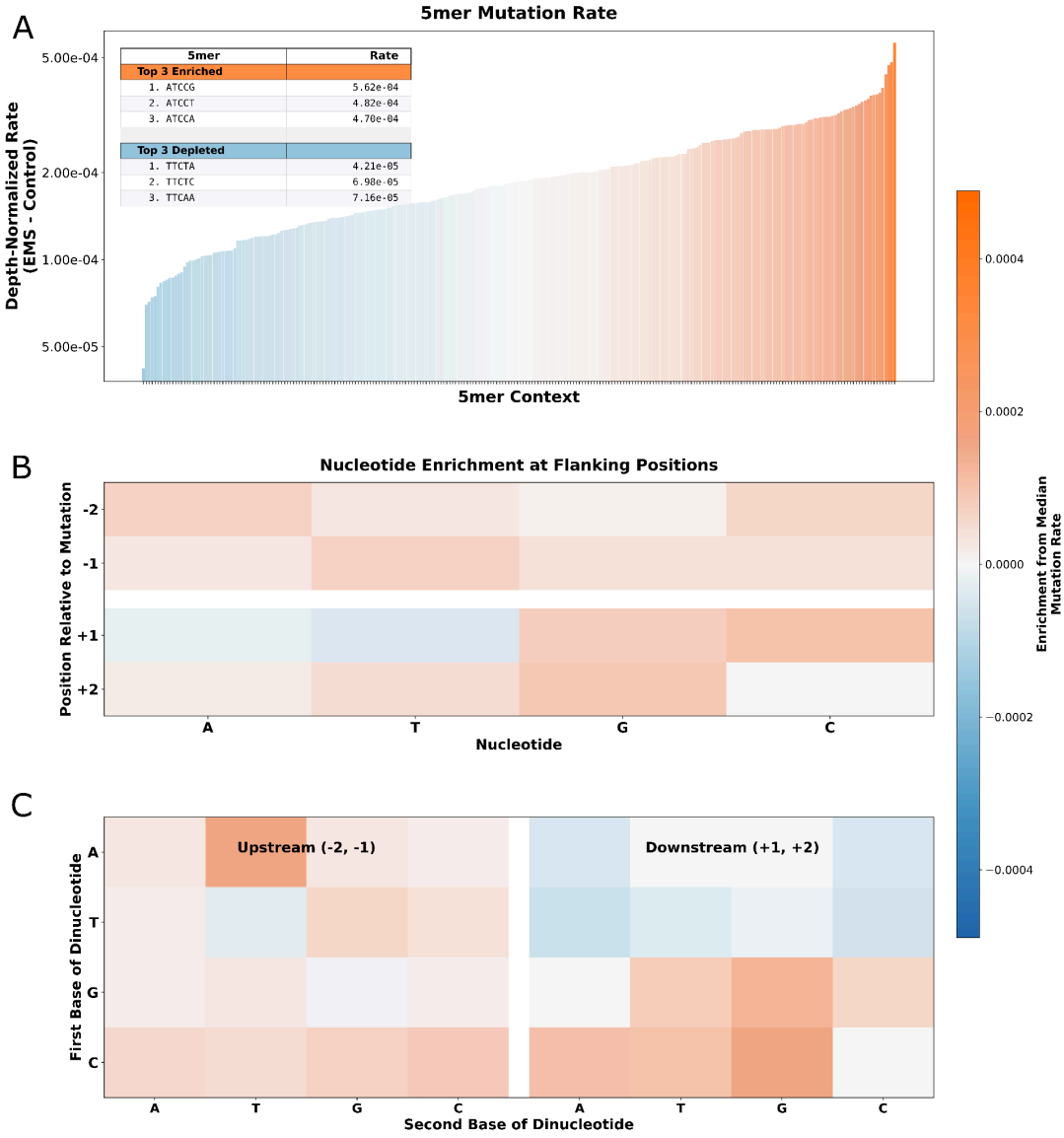

Figure 1.3. EMS mutations in the wMel genome exhibit biased sequence context.

A) Mutation rates of all possible 5mer sequences. Strands are collapsed, such that all canonical EMS mutations are treated as C-centered. The top three enriched and depleted sequences are shown in a table with the associated mutation rate. **B)** Single nucleotide enrichment at each of the four positions flanking a canonical EMS mutation site. **C)** Dinucleotide pair enrichment at the upstream and downstream positions flanking an EMS mutation site.

1.5 Discussion

Researchers have made significant advances in the tractable genetic perturbation of obligate intracellular bacteria (McClure et al., 2017). However genetic perturbation of diverse intracellular bacterial symbionts remains an extended goal of researchers looking to better understand and harness their unique biological niches. Obligate intracellular symbionts like *Wolbachia* are already being used in large scale biological control projects seeking to improve both agricultural and human health (Hoffmann et al., 2011; Salama et al., 2024) yet approaches to modify their genomes are lacking. Here we validated the use and detection of EMS for random chemical mutagenesis of intracellular *w*Mel *Wolbachia* infecting *Drosophila* cell culture. Using an extremely low error rate sequencing method (Lou et al., 2013), we produced mutation profiles with a nearly 5-fold increase above the background mutation and sequencing error rate. Together, these results establish EMS mutagenesis as a tractable and quantifiable approach for generating genome-wide genetic variation in intracellular *Wolbachia*, providing a practical foundation for future functional genetic studies in systems where genetic manipulation has historically been inaccessible.

Beyond establishing the feasibility of detecting EMS-induced mutations in an intracellular genome, our results suggest that EMS imposes a largely uniform mutational burden across functional genomic categories. The minimal difference observed between coding and non-coding mutation rates indicates that EMS does not preferentially target or spare protein-coding regions, consistent with broadly stochastic mutation rather than context-specific targeting or repair. This uniformity is further supported by the absence of extended mutational hot or cold spots across the

genome, arguing against large-scale regional variation in EMS accessibility. This implies that, with sufficient mutational saturation, EMS mutagenesis is expected to impact essentially all genomic regions in *Wolbachia*, making it well suited for genome-wide functional interrogation.

At a finer scale, the lack of a strong relationship between gene expression and mutation rate suggests that transcription-coupled processes play a limited role in shaping EMS mutational outcomes across the *w*Mel genome. However, clear sequence-context preferences are evident, with a 5-mer model providing the most informative description of EMS mutation bias. The fact that even this model performs modestly underscores that EMS mutagenesis is influenced by factors beyond local sequence alone, likely reflecting interactions with genome organization or DNA repair pathways. Together, these findings highlight both the predictability and the limits of sequence-based models of EMS mutagenesis in intracellular genomes.

Our work proposes a future for genomic screens in *Wolbachia* as well as the potential within other host-cell cultured endosymbionts. This demonstration of high-confidence detection of EMS-induced mutations in *Wolbachia* using ultra-low error rate sequencing builds on previous models of EMS mutation rate and bias, and lays the groundwork for larger screens. A key consideration for future mutagenesis screens will be understanding how EMS-induced mutations are maintained or lost over extended culture periods, a necessary step in translating chemical mutagenesis from a tool for generating variation into one capable of supporting phenotypic screens. Doing so will require the development of dose and recovery experimental frameworks that track mutant frequency over time, building on the quantitative

mutational foundation established. Future directions into this phenotypic space will also likely seek to explore correlations between mutation effect, titer and pathogenicity. Another important consideration for future work is the direct comparison of mutation rates between in vitro and in vivo conditions, such as Wolbachia infecting D. melanogaster. With EMS mutagenesis and high-confidence mutation detection now established in cell culture, extending this approach to in vivo systems becomes a more tractable goal.

1.6 Data Availability

All code used for processing circle sequencing data can be found here:

https://github.com/jmcbroome/circleseq

All code used for rate modeling, analysis and data visualization can be found here:

https://github.com/gabepen/ems_effect/tree/manuscript-scope

All raw sequencing data files can be found under NCBI Bioproject PRJNA1427293

1.7 Supporting Information

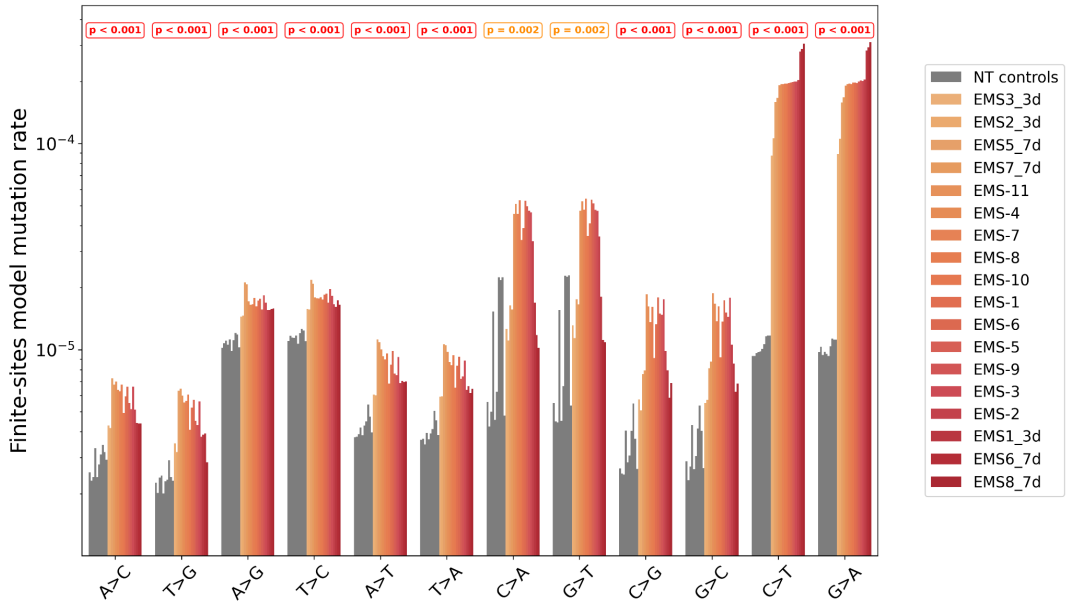

Figure S1.1. Single Base Mutation Spectra.

Rates for each single base mutation type indicate no presence of a strand bias in the mutation data, as expected.

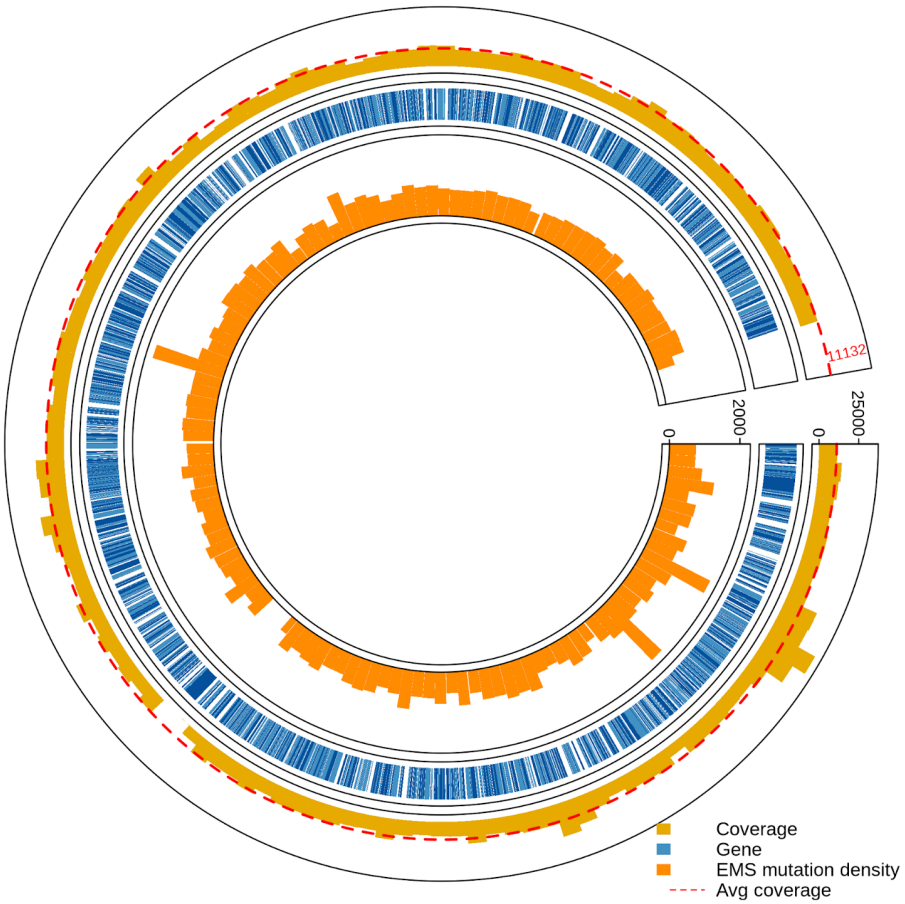

Figure S1.2. EMS mutation density across the genome.

The bars represent 10,000 bp windows. The gap in mutation density coincides with no sequencing coverage. This is due to the drop out of the octomom gene cluster from the JW18 hosted wMel cell line.

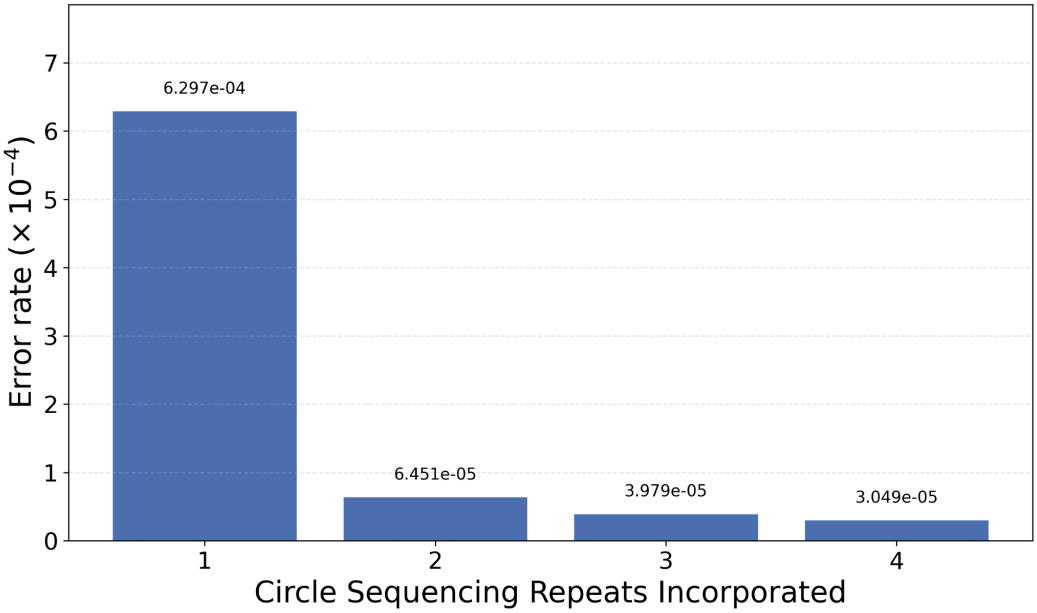

Figure S1.3. Circle sequencing error rate decrease.

The error rate of circle sequencing decreases with the incorporation of tandemly duplicated subreads.

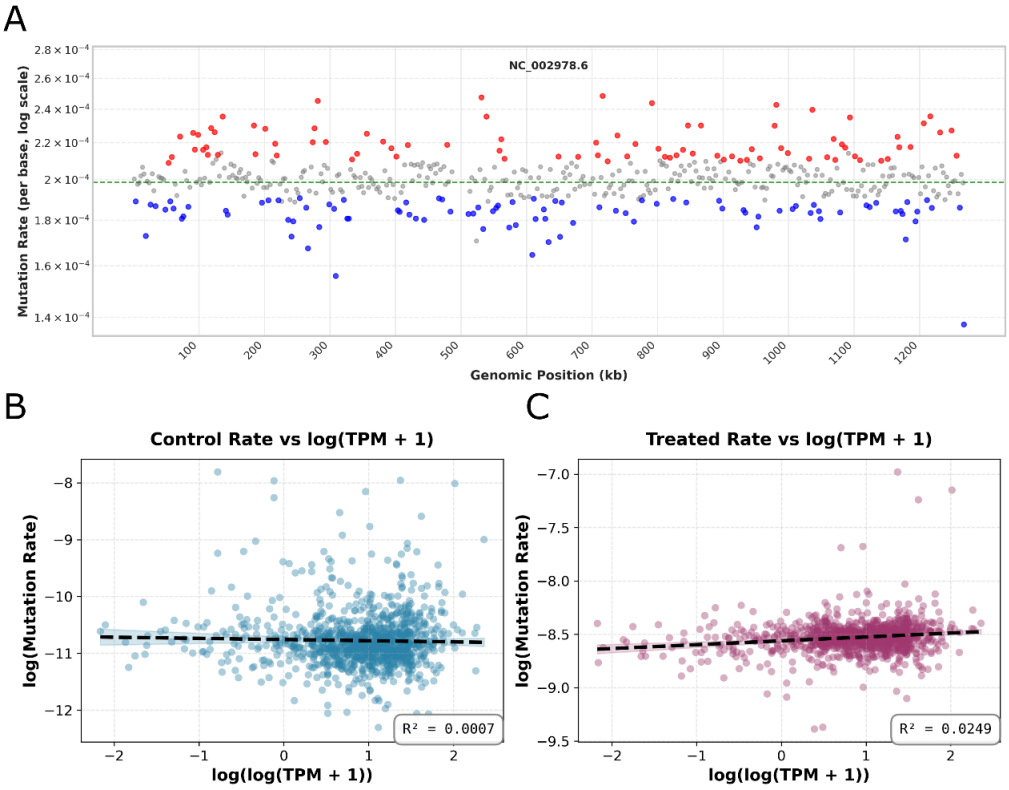

Figure S1.4. Analysis of EMS Mutation Rates Across the wMel Genome and Genes.

A) Plot of 2500 bp windows that have a significantly higher (red) or lower (blue) mutation rate compared to the overall mean. **B, C)** Linear regression analysis of control and treated rates for genes against their expression levels.

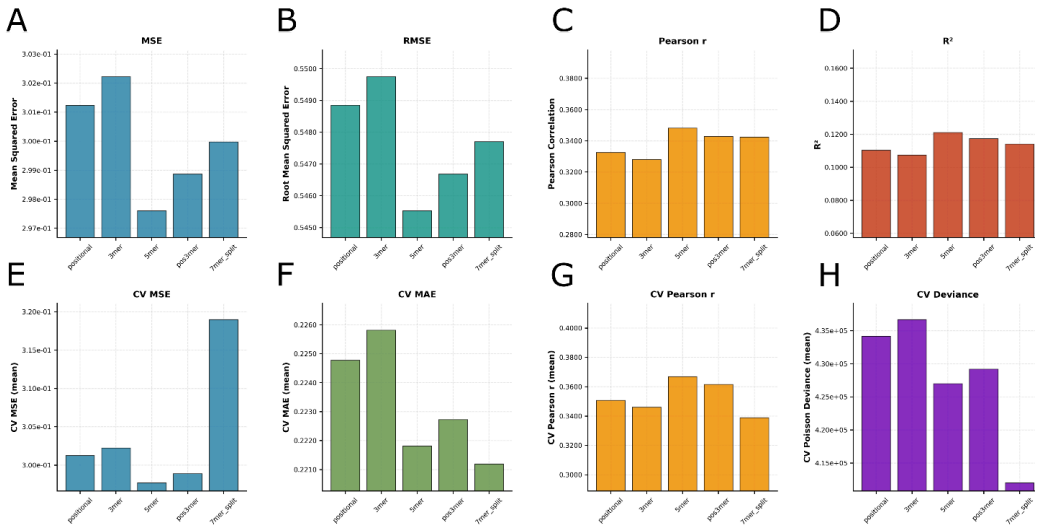

Figure S1.5. Sequence Bias Model Evaluation.

A-D) Within sample prediction accuracy metrics for each model type. **E-H)** Out of sample cross-validation metrics for each model type

Chapter II: Structural Mimicry Screen In Host-Microbe Systems or Identifying Effector Candidates in Host-Microbe Systems through Protein Structure.

2.1 Abstract

Host-microbe systems represent evolutionary niches that produce coevolved biological interactions and are a key component of global health. However, these systems have historically been challenging to research due to their experimental intractability. Impactful advances in global health will be obtained by leveraging *in silico* screens to identify genes involved in mediating interspecific interactions. One signature of host-microbe interaction that can be detected *in silico* is symbiont mimicry of host molecules. Molecular mimicry can occur at any level from DNA to protein structures and functions. Here, we applied protein structure prediction and alignment tools to explore host-associated bacterial structural proteomes for examples of protein structure mimicry. By leveraging the *Legionella pneumophila* proteome and its many known structural mimics, we developed and validated a screen that can be applied to other host-microbe systems to uncover signals of protein mimicry. We have also developed an evolutionary rate analysis workflow for determining sequences undergoing selection to supplement our structural

analysis. These mimics represent candidate proteins that mediate host interactions in microbial proteomes. We successfully applied this screen to other microbes with demonstrated effects on global health, *Helicobacter pylori* and *Wolbachia pipientis*, identifying protein mimic candidates in each proteome. We discuss the roles these candidates may play in important *Wolbachia*-induced phenotypes and show that *Wolbachia* infection can partially rescue the loss of one of these factors. These predictions will progress our understanding of these systems and lay the groundwork for future *in vitro* and *in vivo* experiments and bioengineering projects. This work demonstrates how a genome-wide screen for candidates of host-manipulation and intracellular survival offers an opportunity to identify functionally important genes in host-microbe systems.

2.2 Introduction

The intracellular niche of symbiotic and parasitic bacteria is a complex environment and exhibits pressures that lead to unique coevolutionary products such as molecular mimics(Russell et al., 2024). Intimate cellular and subcellular interactions between hosts and symbionts can create conditions in which it is advantageous for a symbiont species to produce a molecule or protein that mimics the structure of a host component. For example,

molecular mimicry is a major contributor to host immune system evasion and microbe-host and virus-host coevolution (Hebert et al., 2015; Lasso et al., 2020; Mondino et al., 2020; Sansom et al., 2007; Wong et al., 2017). Distant structural homology and convergence with host proteins is likely a key evolutionary step in the generation of successful host-microbe interactions. However, successfully identifying and characterizing protein mimics often requires painstaking molecular approaches. Therefore, novel approaches for rapidly and inexpensively identifying candidates will enable fundamentally new insights and analyses.

Protein mimicry is often assessed in the context of host immune evasion for enacting control of host cellular machinery. Protein mimics have been a research focus of interactions between *Legionella* and its human host(Mondino et al., 2020), where bacterial protein mimicry is thought to contribute to manipulating numerous host systems such as transcription, ubiquitination, and protein phosphorylation. (Black et al., 2019; Sansom et al., 2007; Wan et al., 2019) We thus have a breadth of wet lab validation of known protein mimics that can be used to confirm and control a novel screen in a system with little structural, and biochemical data available. The *Legionella pneumophila*-human system provides a framework that bounds the extent of structural conservation within host-microbe alignments that would be

of interest to identify. *Legionella*-human structural proteomes also provide a type of ‘positive control’ for any approach to identify protein mimicry.

Structural comparison allows for high sensitivity and throughput, which was limited in previous approaches to protein mimicry identification. Sequence based screens for protein mimicry have been successful in identifying candidates, but are limited in their sensitivity because they require high sequence similarity.(Ludin et al., 2011) Hebert *et al.* utilized RNA-seq data of protein coding genes and ORF predictions to construct a pipeline that identified similar protein sequences between host and symbiont genomes(Hebert et al., 2015). Other sequence based screens have also been developed for the purposes of identifying novel *L. pneumophila* effectors. These implement a variety of wet-lab and *in silico* techniques and leverage the large set of known, validated effectors(Burstein et al., 2009; Gomez-Valero et al., 2019). However, while protein sequence analysis is a useful approach for uncovering eukaryotic domain-containing mimics, it disregards newly available structural data that we can now leverage for much more sensitive homology screening. Furthermore, sequence-based approaches miss convergent structural and functional mimics that may have virtually no sequence similarity. While wet lab structural and biochemical data is still the gold standard for confirming and characterizing protein mimicry,

novel structure-aware approaches are needed to identify diverse mimic candidates in genetically intractable organisms.

Modern advances in protein folding prediction from genomic sequences have enabled *ab initio* assays for protein structural mimicry between symbiont proteomes and those of their hosts for the first time. Intracellular *Wolbachia pipientis* alphaproteobacterial symbionts of diverse arthropod and nematode species are ideal candidates for such a mimicry screen. *Wolbachia* genomes are enriched in eukaryotic-like elements (Bordenstein and Bordenstein, 2022), they induce a range of unique host-associated phenotypes, and the wMel strain from *Drosophila melanogaster* is important in applied mosquito biological control strategies.(Dufault et al., 2022) However, the functions of only five of *w*Mel *Wolbachia's* ~1200 genes have been directly experimentally determined because the obligate intracellular lifestyle of these bacteria has made them difficult to characterize genetically.(Adams et al., 2021; Arai et al., 2023; Martin and Newton, 2023; Ote et al., 2016) Thus, a genome-wide structural protein mimicry screen offers an appealing opportunity to identify functionally important *Wolbachia* genes for interacting with and controlling host biology that could be engineered into biological control strategies. The model *D. melanogaster* system naturally infected with the *w*Mel strain offers an experimental system to begin to test and validate these candidates.

Deep learning and generative AI revolutionized our ability to predict protein structure from amino acid sequence data(Chowdhury et al., 2022; Kuhlman and Bradley, 2019; Torrisi et al., 2020), and opened powerful new avenues for identifying structural similarity. Protein structure prediction models such as AlphaFold2(Jumper et al., 2021) coupled with access to large graphical compute power through cloud resources have enabled small budget access to entire structural proteomes in a matter of days. In the first half of 2022, folded proteomes were made even more accessible with the expansion of the AlphaFold Protein Structure Database(Varadi et al., 2022), releasing predictions for nearly every sequenced organism in all domains of life. Ultrafast structural alignment methods provide tools for grappling with these large amounts of structural data. The 3Di+AA method core to Foldseek(van Kempen et al., 2022) has proved to be a lightning fast means of aligning thousands of protein structures and was tested with remote protein homology searches in mind. These new approaches therefore offer an appealing opportunity to investigate molecular mimicry between symbiont and hosts genome-wide.

Here, we utilize these new genomic datasets, protein structure predictions, and rapid structural alignment tools to search for distant protein structural similarity between entire host and symbiont proteomes. We initially developed this screen utilizing the extensive knowledge of structural mimicry in the

Legionella-human system for validation and testing. We then apply the validated framework to *Helicobacter pylori,* which imparts numerous host control and manipulation mechanisms on human gastric epithelial cells that lead to disease. We next apply our framework to *Wolbachia melanogaster* which infects *Drosophila melanogaster* as the host genome is extensively annotated and we have both *in vitro* and *in vivo* systems for potential follow up experiments on identified mimicry candidates(Russell et al., 2018; S. L. Russell et al., 2023). These platforms allow us to establish functional assays that support the findings of the protein mimic screen, demonstrating a basis for a *Wolbachia* wMel mimic of the *Drosophila* protein, eato. We refined the structural homology hits from this process to ensure that the mimicry candidate identified is most likely a result of convergent evolution, opposed to deep phylogenetic conservation for endogenous housekeeping functions. These candidates represent unique novel connections between hosts and their microbes as well as co-evolved products resulting from a symbiotic evolutionary history. This approach advances bioinformatic investigations into convergent evolution within host microbe systems, and produces novel candidate genes for researchers seeking to perturb symbiotic systems.

2.3 Methods

2.3.1 AlphaFold Protein Structure Predictions

We downloaded predicted structural proteomes for a selection of host genomes, host-associated bacterial genomes, and free-living bacterial genomes from the AlphaFold Protein Structure Database(Varadi et al., 2022). We selected host proteomes for *Drosophila melanogaster* and human and proteomes for their associated microbes: the wMel strain of *Wolbachia*, *H. pylori 26695* and the Philadelphia 1 strain of *L. pneumophila.* To control for similar structures constrained by housekeeping functions, we randomly sampled 104 free-living bacteria from across the bacterial tree of life.

2.3.2 Free-Living Proteome Control Database Generation

The free-living control method described above was developed with the goal of ensuring alignments between host and microbes were unique to those systems and not representative of a highly conserved structure found within most bacteria. Highly conserved symbiont proteins required for housekeeping functions could exhibit structural similarity with host proteins due to shared endogenous functions. These proteins would be highly similar to free-living orthologs, given their close phylogenetic proximity and necessity for core cellular processes, allowing us to identify and exclude them from the mimic

candidate gene set. Leveraging this structural conservation logic, we developed a method that would confidently ensure that each candidate mimic protein was not core to free-living bacterial cellular function, and instead had evolved a divergent function through converging on a common structure with its host.

In order to develop a large enough control proteome dataset that would sufficiently describe free-living proteomes and their structural content, we automated the selection process. Specifically, we designed a system of randomly selecting bacterial species and automating the literature search to determine if the selected species was free-living or potentially host-associated. The motivation for this method for developing such a dataset is in part due to the lack of annotation on host-association of bacteria in current databases. We utilized taxonKit(Shen and Ren, 2021) to generate a json representation of the entire NCBI bacterial taxonomy database. We then used a custom python script to randomly search this dataset structure by bacterial order and species. To do this we randomly selected two species from each randomly selected order. The script uses the NCBI Entrezpy python library to validate that a selected species has a genome in the assembly database that is marked as complete. Randomly selected bacteria species were then processed with another custom python script leveraging the unofficial Bard-API (https://github.com/dsdanielpark/Bard-API) accessed

on October 20th, 2023 and the BacDive database API(Reimer et al., 2022). Literature related to the bacteria species queried was processed by a constrained bard session in order to make conservative predictions on whether or not the bacteria was free-living or potentially host-associated. We manually reviewed these selected taxonomy IDs and used them to download associated proteome “shards” from the AlphaFold Protein Structure Databases. Free-living proteomes that contained at least 900 structures were selected for use in the control dataset. The full list of organisms selected for this database along with the literature used in the decision is available in the supplementary data (Supplemental Table S9).

2.3.3 Protein Structure Alignments

We performed structural alignments between proteomes with the tool Foldseek(van Kempen et al., 2022). Foldseek discretizes structures into sequences that describe the spatial relationships between neighboring residues. We selected the 3Di + AA local alignment method for its superior alignment speed while retaining similar sensitivity to the global TM-Align method (van Kempen et al., 2022; Zhang, 2005). We treated microbe proteins as queries and host proteins as targets, generating a TM-score ranging from 0 to 1 for primarily evaluating the structural similarity of an alignment pair. We selected an e-value cutoff of 0.01 for these alignments, an established metric

used in large scale alignment and clustering methods developed by the Steinegger lab(Barrio-Hernandez et al., 2023).

2.3.4 Structure Alignment Controlling and Filtering

We designed and tested our methods with the intent to filter out proteins that are conserved structures in free-living bacteria, and thus likely not examples of host manipulation or symbiosis. We first filtered alignments for mimic candidates by their TM-score, target coverage, query coverage, fraction of identical residues and fraction of free-living proteome alignments. A high fraction of identical residues in alignments is evidence of recent shared ancestry, possibly due to horizontal transfer. This value is expected to be large for recently transferred genes, or highly conserved housekeeping genes(Ma et al., 2008). A high fraction of free-living proteomes containing a significant alignment with a query protein indicated the overall conservation of the structure in free-living proteomes. This may suggest that the microbe protein is less likely to be involved in host interaction or manipulation. In order to further filter the resulting list of candidates to arrive at a final list of high quality protein mimicry candidates, we also considered qualitative metrics such as gene ontology and conservation of critical residues and domains within the structural alignment.

We obtained protein mimic positive control sets for *Legionella* from Mondino *et al* 2020(Mondino et al., 2020). This set of 23 genes represented validated *Legionella* effectors showing structural mimicry. *Legionella* proteins were classified as eukaryotic domain containing or eukaryotic like proteins which we equate with partial and full mimicry, respectively. Alignments between this positive control set and the human proteome were used to determine reasonable filters for alignment statistics. For controlling alignments we removed non-host-associated structural conservation based on alignment to the free-living bacterial proteomes dataset. We aligned the host-associated microbial proteome to each of the proteomes in this dataset, allowing us to report a fraction of proteomes that the microbe protein had a reciprocal best hit alignment within.

2.3.5 Evolutionary Rate Analysis

To perform the evolutionary rate analysis of candidate mimic sequences we selected a taxon where the resulting subtree contained the microbe from which candidate sequences were identified. A subtree for the initial taxon was determined with the tool Taxonkit(Shen and Ren, 2021). The NCBI datasets CLI was then used to download a set of genomes within the subtree. The easy-rbh workflow from MMseqs2(Steinegger and Söding, 2017) was used to determine homologous sequences between the query microbe proteome and each selected genome from the subtree. We then used the NCBI datasets

CLI again to collect the homolog gene nucleic acid sequences for each candidate mimic sequence. We then processed each homolog gene set with the HyPhy(Kosakovsky Pond et al., 2019) codon-msa pre-processing script converting them into codon aware amino acid sequences. The processed codon aware protein sequences were aligned with MAFFT(Katoh and Standley, 2013) v7.525 and then converted to nucleotide sequences with the HyPhy codon-msa post-processing script. We used IQTree(Minh et al., 2020) v2.3.3 to generate the phylogenetic trees of each mimic sequence homolog alignment and then analyzed each tree with aBSREL(Smith et al., 2015) to determine which branches in the phylogenies contained evidence of episodic diversifying selection. We calculated the branch fraction value as the fraction of branches in the phylogenetic tree with a corrected p-value of less than 0.05. All custom scripts for processing intermediate files and parsing the aBSREL output are available individually or within a SnakeMake(Mölder et al., 2021) workflow found in the ‘evorate_workflow’ folder of the github (https://github.com/gabepen/mimic_screen).

2.3.6 Mitochondrial Sequence Prediction

We used targetp(Almagro Armenteros et al., 2019) v2.0 to predict the likelihood of a protein containing a

N-terminal presequences signal peptide (SP), mitochondrial transit peptide (mTP), chloroplast transit peptide (cTP) or thylakoid luminal transit peptide (lTP).

2.3.7 GO Enrichment Analysis

We used host protein UniProt IDs and the PANTHER(Thomas et al., 2022) classification system for the GO enrichment analysis(Ashburner et al., 2000; Gene Ontology Consortium et al., 2023). P-values were obtained with Fisher's exact test. We used FDR for correction.

2.3.8 Phylogenetic Tree Construction

We performed a BLAST+ 2.15.0(Camacho et al., 2009) search with the *Wolbachia* wMel *HtpG* gene sequence (WP_022626167.1) and manually selected samples from the results retaining the amino acid sequences for each hit. We next used ClustalO(Sievers et al., 2011) v1.2.4 to align the amino acid sequences selected from the blast results and manually trimmed the core alignment to 923 amino acid positions. We used IQTree(Minh et al., 2020) v2.2.5 with standard amino acid tree inference parameters and 100 bootstrap replicates. This consensus tree was plotted in FigTree with further annotation in Inkscape.

2.3.9 Ovary fixation and PI staining

We derived our methods from previous work done on the *Drosophila* oocyte(Russell et al., 2018). Newly enclosed flies were transferred to fresh white food with yeast and aged 4–5 days. 10–15 flies from each cross were dissected, fixed, and stained with propidium iodide (PI). Ovaries were dissected, the ovarioles were separated with pins, and fixed in 200 µl devitellinizing solution (2% paraformaldehyde and 0.5% v/v NP40 in 1x PBS) mixed with 600 µl heptane for 20 min at room temperature, with agitation. Oocytes were then washed 5x in PBS-T (0.1% Triton X-100 in 1x PBS), and treated with RNAse A (10 mg/ml) overnight at room temperature. Oocytes were washed six times in PBS-T and were incubated overnight in PI mounting media (20 µg/ml in 70% glycerol and 1x PBS), and mounted on glass slides. The *uninfected OreR* stocks were used as wild-type controls. Uninfected *Eato* mutant and wMel-infected *Eato* mutant samples were processed simultaneously to minimize batch effects. Wild-type control samples were processed after these experimental samples. Slides were imaged immediately or stored at -20°C for no more than three weeks. Oocyte localization variation is the same between runs for wild-type and mutant genotypes.

2.3.10 Confocal imaging

Stage 14 egg chambers were imaged on a Leica SP5 confocal microscope with a 63x objective. Optical sections were taken at the Nyquist value for the objective, every 0.34 or 0.38 μm, at a magnification of 1x. Propidium iodide was excited with the 514 and 543 nm lasers, and emission from 550 to 680 nm was collected. Images were processed in ImageJ(Schneider et al., 2012).

2.3.11 Image and plotting analysis

To quantify engulfment, we counted the number of persisting NCs in each stage 14 egg chamber through z-stack and 3D reconstruction on ImageJ. The criteria for a stage 14 egg chamber was fully developed dorsal appendages(Jia et al., 2016). For better visualization on ImageJ, the brightness/contrast tool was used to increase the threshold on the image, and all background noise was rendered black. The resulting NC counts for all genotypes were plotted in R through ggplot2(“Create Elegant Data Visualisations Using the Grammar of Graphics,” n.d.). The resulting NC counts between uninfected and wMel infected *Eato* mutants were compared with the Wilcoxon Rank Sum Test.

2.4 Results

2.4.1 Establishing Expectations for Protein Mimic Alignments using Known Legionella Examples

We leveraged well-characterized examples of *Legionella* mimicry of human proteins and a large database of free-living proteomes to establish our baseline expectations for genome-wide candidate structural mimicry detection. First, we examined the expected TM-Scores, target coverage, query coverage, fraction of identical residues and fraction of free-living proteomes of protein structure alignments using 14 biochemically characterized protein mimics from *Legionella*-human examples (Supplemental Table S2.1). We also considered their classification according to Mondino *et al* 2020, which categorizes effectors into eukaryotic domain containing and eukaryotic resembling proteins. We referred to these classifications here as partial and full mimics. This set of 14 represents the minimal set of true mimics in the *Legionella* genome, but there are likely others that have not yet been characterized. To identify unknown *Legionella* mimics, we generated a structural proteome alignment between the human proteome and *Legionella pneumophila* subsp. *pneumophila* str. Philadelphia 1 to use as a genome-wide validation set for investigating further host-microbe pairs *ab initio*.

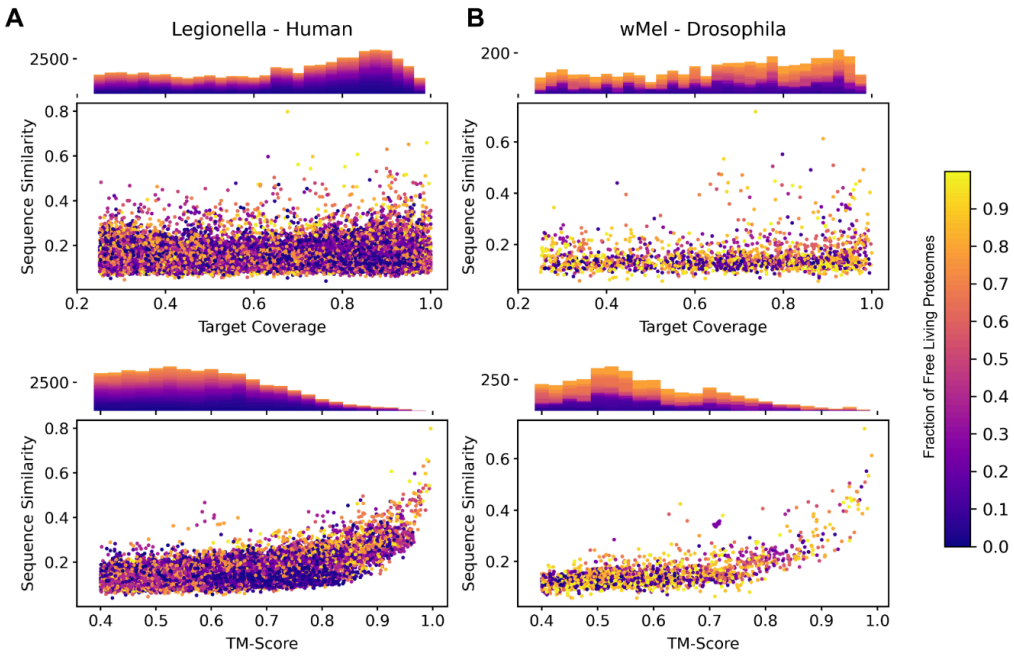

Figure 2.1. Alignment Space Under Expansive Control Method.

A. *Legionella*-Human alignment space sequence similarity as it relates to target coverage and TM-Score. The averages for the fraction of free-living proteomes aligned for both systems were 0.416 for *Legionella*-Human and 0.605 for wMel-*Drosophila*. **B.** wMel-*Drosophila* alignment space sequence similarity as it relates to target coverage and TM-Score. Alignments are colored by the fraction of control proteins the query protein was aligned with.

Analysis of the validation mimic set for *Legionella* confirms that our screen can robustly detect both full and partial structural protein mimics. We found that TM-Scores of the known eukaryotic domain-containing *Legionella* proteins ranged from 0.4-0.8, with an average of 0.64, target (host) coverage ranged from 0.1-0.95, with an average of 0.43 and query (microbe) coverage ranged from 0.23-0.99, with an average of 0.67. This large target coverage range tightened significantly for the full structural mimics to 0.6-0.97 with an

average of 0.82 while the TM-Score range for full mimics encompassed values from 0.4-0.95 with an average of 0.65. We found that the fraction of free-living proteomes aligned differed greatly between *Legionella's* five known full structural mimics and the nine partial ones with averages of 0.6 and 0.2, respectively.

Overall this validation informed our expectations on the search for structural mimicry to require generous thresholding around the different expected ranges for the two general classifications of mimicry extent. This was necessary in order to encompass all possible examples of mimicry in our alignment results. Using an E-value cutoff of 0.01, we decided to select results for further ontological evaluation with structural statistical thresholds of 0.4 for TM-score and 0.25 for target coverage. We also determined there are cases where symbiont proteins mimic a eukaryotic domain within a large host protein, which appear as partial structural target alignments. Considering that bacterial proteins tend to be shorter and contain fewer domains than eukaryotic proteins(Ye and Godzik, 2004), this is intuitive. In order to catch these cases we did not filter alignments with lower than 0.25 target coverage if the query coverage was at least 0.5. In the case of our *Legionella* and Human proteome alignment ~35,000 of the 45,752,796 alignments performed met these initial filters. After filtering based on TM-score, target and query coverage we still retained 1,670 microbe proteins participating in thousands of

alignments (Supplemental Table S2.2, S2.3). Next, we controlled the hits based on the fraction of free-living proteomes they aligned to, which further reduced the total number of alignments considered. 2600 alignments containing 464 unique microbe proteins exhibited free-living frequencies equal to or less than the average for partial mimics in *Legionella* of 0.2.

2.4.2 Comparing Host-Microbe Structural Proteomes Identifies Candidate Mimics

We applied our framework to three host-microbe systems: *Legionella*-human, *H.pylori*-human and *Wolbachia* wMel-*Drosophila melanogaster*, identifying strong candidate mimic-like effectors in each. Comparing the overall trends between microbe-host and microbe-free-living bacteria alignments revealed key differences likely attributable to differences in gene content between *w*Mel, *H. pylori,* and *Legionella* (Figure 2.1, Figure S2.1). Consistent with its larger genome (3.5 Mb vs 1.3 Mb and 1.6 Mb), *Legionella* encodes more proteins with significant structural alignments to its host genome than *Wolbachia* or *H. pylori* encode. We noticed that *Wolbachia* and *H. pylori* appeared to have a higher fraction of free-living conserved proteins than *Legionella*, but found that the total number of proteins highly conserved in free-living proteomes between the three organisms was comparable. At a fraction of free-living proteomes aligned cutoff of 0.8, the number of query structures present in free-living proteomes from *Wolbachia* and *H. pylori* was 297 and 364 respectively. In *Legionella* this value was 384, despite *Legionella*

having twice the number of protein structures. The difference between the conservation statistics between these endosymbiotic bacteria is likely due to the nature of host association in these systems. In the case of *Wolbachia*, strains verge on mutualism with many of their various hosts(Hedges and Johnson, 2008; Hosokawa et al., 2010; S. L. Russell et al., 2023). *H. pylori* infects a narrow range of hosts and is highly adapted for human gastric mucosa(Lee et al., 1993; Thorell et al., 2023). In both these cases the highly specified nature of association with a host has led to an eroded genome with reduced functional gene content. In contrast, *Legionella* is a damaging pathogen driven by a large genome overrun by elements from many previous eukaryotic hosts that it has evolved within(Gomez-Valero and Buchrieser, 2013). However all of these systems have protein structures that are rarely found within free-living bacteria from which we have identified as strong candidates for protein mimicry of host cellular machinery.

2.4.3 Evolutionary Rate Analysis Further Selects Sequences

We next tested the proteomes of the query microbes with an evolutionary rate analysis workflow. We wanted to know if microbial proteins that evolved molecular mechanisms to interact with host cellular machinery would show increased or decreased rates of evolution in localized symbiont clades compared to orthologs from free-living taxa. To explore this, we collected and aligned orthologous sequences from neighboring clades to our queried

microbes, inferred a maximum-likelihood phylogeny for each gene, and determined the average dn/ds across the free-living and host associated ortholog branches within each gene tree. For *Legionella pneumophila* we selected genomes from Legionellales, Enterobacterales, Pseudomonadaceae, Vibrionaceae and Aeromonadaceae finding 100 host associated and 63 free living genomes meeting our criteria for selection. For *Wolbachia* we selected from Rickettsiales, Caulobacterales, Rhodospirillales also finding 100 host associated and 63 free living genomes. For *Helicobacter pylori* the available genomic data was much lower and even selecting taxa from the entirety of Epsilonproteo bacteria we were only able to find 25 host associated and 17 free living genomes matching the criteria. Details on selection methods can be found in the method section.

2.4.4 Structural Mimicry in *Legionella pneumophila*

Structural mimicry of both partial and full host proteins is a common method of host cell manipulation used by *Legionella pneumophila* strains, and is why we selected this pathogen for validation of this screen. *L. pneumophila* is a gram-negative bacterium and the causative agent of Legionnaires' disease, a severe form of pneumonia. *L. pneumophila* strains, like other bacterial pathogens, must avoid phagosome fusion with lysosomes through evolution of numerous host cell manipulation processes(Hubber and Roy, 2010). A vast

arsenal of host effectors and mimics has been painstakingly uncovered through years of biochemical research(Mondino et al., 2020). However, *L. pneumophila*'s bloated pathogenic genome is influenced by countless eukaryotic host associations through which it acquired mobile genetic elements and genes by HGT, and is likely to still have mimic-like effectors yet to be uncovered(Gomez-Valero and Buchrieser, 2013). We also performed a GO analysis of the host target proteins aligned to the 2,600 microbe protein subset that aligned to less than 0.2 of the free-living proteome database (Supplemental Table S2.4). Many of the top enriched biological processes were associated with transmembrane transport. It is important to note that bacterial queries were not limited to alignments with a single host target structure and single query structures contributed multiple targets to the protein IDs used in the GO analysis. Hence, we did not rely on GO analysis results for finding strong examples in the data, and rather noted the results as potentially informative of mimic trends. In our use of *L. pneumophila* as a validation organism for our framework, we were able to identify strong candidates for mimic-like effectors previously uncharacterized. In addition to these effector protein candidates, we were able to suggest new interactions and mechanisms of known effectors, such as Mip, through reconsidering them as mimic-like.

2.4.4.1 Q5F2U4 (Mip)

Mip is a well known *L. pneumophila* effector protein proven to be essential for virulence in various host genetic backgrounds(DebRoy et al., 2006). While Mip has been previously characterized, it has not been conclusively determined what components of human cells Mip interacts with. A recent paper showed that Mip has a general interaction with and regulates the lncRNA GAS5/miR-21/SOCS6 axis, influencing phagocytosis and chemotaxis in the RAW264.7 macrophage-like cell line(Shen et al., 2022). Similarly it has also been shown that *L. pneumophila* infections dampen the cytokine response of infected macrophages(McCoy-Simandle et al., 2011). The *L. pneumophila* structure Q5F2U4 aligns with a human peptidyl-prolyl cis-trans isomerase, FKBP1A which has been shown to have a binary interaction with suppressor of cytokine signaling 6 (SOCS6)(Haenig et al., 2020). These protein structures aligned with a TM-score of 0.89, target coverage of 0.68, query coverage of 0.4, and fraction of identical residues of 0.36 (Figure 2.2). An alignment between Mip and a structure in the free-living proteomes was not found. Thus, this potential example of mimicry supports previous predictions about Mip interactions with host lncRNA pathways. Furthermore, mimicry of FKBP1A, or other structurally similar host peptidyl-prolyl cis-trans isomerases, by Mip suggests a mechanism by which *L. pneumophila* directly suppresses cytokine expression for cellular persistence.

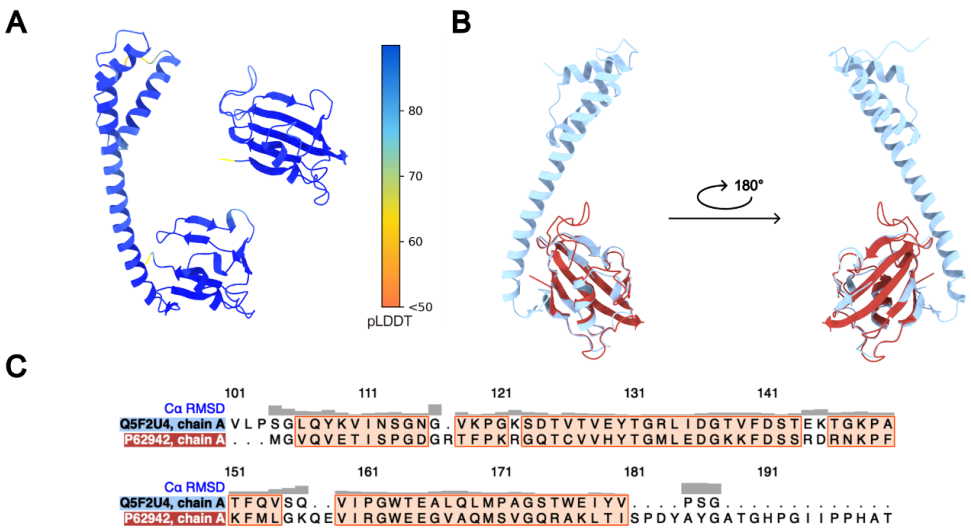

Figure 2.2. Mimic Candidate Q5F2U4.

A. Structures of *L. pneumophila* Q5F2U4 and human FKBP1A colored by pLDDT. **B.** Structural alignment of *L. pneumophila* (blue) and human (red) proteins. **C.** Phylogenetic tree of orthologous sequences used in evolutionary rate analysis **D.** Portion of sequence alignment between microbe and host proteins, showing low sequence similarity. Similarity between the two structures is shown by the RMSD at the top of the plot.

2.4.4.2 Q5ZW23

This *L. pneumophila* protein is predicted to be a membrane associated proton/peptide symporter and it aligns to the full human Solute Carrier Family 15 member 3 (SLC15A3) protein. It shares the strongest alignment with Q8IY34, SLC15A3 (TM-score: 0.73, target coverage: 0.91, query coverage: 0.98 fraction identical residues: 0.15, Supplemental Figure S2.2). The structure is not found in any of the free living proteomes screened. Human SLC15A3 protein is associated with the lysosomal membrane(Schröder et al., 2007) and is thought to be involved in the detection of microbial pathogens by

toll-like receptors and NOD-like receptors through studies in mice(Nakamura et al., 2014). This example of potential mimicry could thus coincide with *L. pneumophila'*s ability to avoid fusion with the lysosome and represent another tool in its extensive kit of subverting host immune responses.

2.4.5 Structural Mimicry in Helicobacter Pylori

Other host-associated bacteria that are limited in their experimental tractability can be investigated with our approach, including human pathogens such as *H. pylori*. The extracellular bacteria infect the stomach epithelium and the upper part of the small intestine in humans(Kao et al., 2016; Malfertheiner et al., 2023), and has also been found in primates, rodents and other various mammals(Mladenova-Hristova et al., 2017). *H. pylori* can cause various gastrointestinal malignancies, including inflammation of the stomach lining and ulcers. *H. pylori* infection is quite common, estimated to infect about half of the world's population(Malfertheiner et al., 2023). Its culture requires a microaerophilic environment and complex blood-containing media, making it challenging to study *in vitro*(Talebi Bezmin Abadi, 2018). In order to establish such persistent and widespread infections, *H. pylori* utilizes highly specialized mechanisms to avoid host defense and adaptive immune mechanisms in gastric epithelial cells. The bacteria interact closely with host cells, leading to manipulation of their cellular process and functions(Alzahrani et al., 2014). These interactions cause a wide range of effects including DNA damage,

apoptosis, proliferation, stimulated cytokine production, and even cell type transformations(Alzahrani et al., 2014; Wizenty et al., 2020). Although dozens of effectors secreted by *H. pylori*'s type IV secretion system (T4SS) have been identified, many more effectors remain to be discovered(Kao et al., 2016; Lina et al., 2014; Skoog et al., 2020).

We produced 24,139,512 pairwise alignments between the *H. pylori* strain 26695 and human proteomes. 3516 alignments to host proteins containing 782 unique microbe protein structures met our significant alignment criteria (Supplemental Table S2.5, S2.6). Of these, 888 were depleted among free-living proteomes (aligned to a fraction less than 0.2). We performed a GO analysis of the host target proteins aligned to the microbe protein subset (Supplemental Table S2.7). Many of the top hits were associated with cellular transport such as asparagine, leukotriene, (R) - carnitine and leucine transport into and between cells. Similar to our screen of the *Legionella* proteome, the application to *H. pylori* identified both novel candidates for structural mimicry as well as suggested mimicry mechanisms for existing effector candidates.

2.4.5.1 O25525 (guaC)

The *H. pylori* O25525 protein is a GMP reductase responsible for catalyzing NADPH-dependent deamination of GMP to IMP in purine salvaging and

processing pathways. It is aligned with the full structure of human GMP reductase 2 (GMPR2). *H. pylori* is reliant on deamination pathways for purine nucleotide biosynthesis, as it lacks the enzymatic machinery necessary for *de novo* production of IMP and must salvage purines from the environment(Mendz et al., 1994). A study into bacterial mutants lacking each of the critical components of the pathway have shown varying capabilities to utilize different purine bases and nucleosides, with guaC mutants having limited conditions for growth(Liechti and Goldberg, 2012). *H. pylori* has a strong, full structural alignment with the human GMPR2 protein, with a TM-score of 0.876 a target coverage of 0.98 and a fraction of identical residues of 0.32 (Supplemental Figure S2.4). O25525 was nearly non-existent in the free-living control set with an alignment fraction of 0.02. It has been shown that GMPR2 overexpression can promote differentiation of leukemia cells(Zhang et al., 2003). This could suggest that *H. pylori* expression of a highly similar structure could be promoting gastric cancer development.

2.4.5.2 O25981

A periplasmic protein of unknown function is encoded by the *H. pylori* gene *HP1440*, which exhibits full structural alignment with a human serine protease. *HP1440* has been recently shown to belong to the CrdRS regulon(Allen et al., n.d.). The CrdRS regulon is a set of genes controlled by

the CrdRS two-component system in response to changes in its environment, specifically in the context of host-induced stress. *H. pylori* O25981 aligned with the human chymase CMA1 structure P23946, with a TM-score of 0.49, target coverage of 0.99, query coverage of 0.96, and fraction of identical residues of 0.11 (Figure 2.3). This serine protease is expressed in mast cells where it activates Matrix metalloproteinase (MMP)-9 and induces inflammation(Takai and Jin, 2022). In *H. pylori* infections, MMP-9 gene expression is significantly increased after *H. pylori* infection, potentially by activation with its own chymase-mimic(Mori et al., 2003).

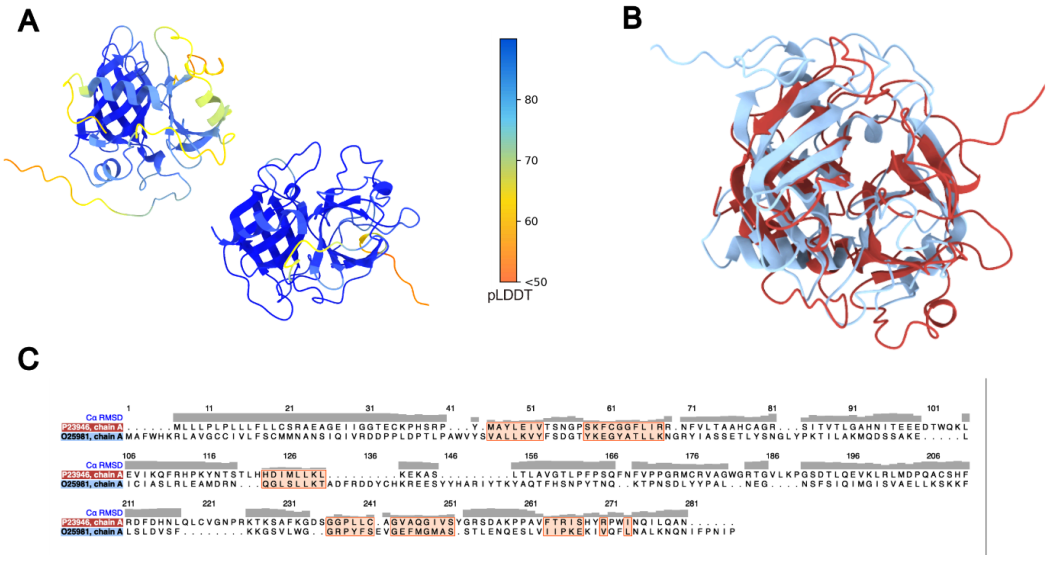

Figure 2.3. Mimic Candidate O25981.

A. Structures of *H. pylori* O25981 and human P23946 colored by pLDDT. Structures exclude extension of terminal disorder tails. **B.** Structural alignment of *H. pylori* (blue) and human (red) proteins. **C.** Sequence alignment between microbes and host proteins showing low sequence similarity. Structural alignment similarity is shown by the RMSD at the top of the plot.

2.4.6 Structural Mimicry in Wolbachia wMel

We performed ~13,700,000 pairwise structural alignments between the *wMel* and *Drosophila* proteomes. From this full proteome alignment, 4,644 alignments from 460 unique *wMel* protein structures fell within the predetermined range of interest in the alignment space for candidate structural mimics. Using the 0.2 presence in free-living alignments threshold for filtering left 33 *wMel* proteins with significant alignments.

The protein mimicry screen and subsequent ontological analysis identified examples of both full and partial structural mimicry that corroborated hypothetical mechanisms underlying *Wolbachia* wMel-*Drosophila* interactions. We identified 1,877 *Drosophila* protein targets with alignment metrics that passed our initial filters. Of these, 279 aligned to fewer than 20% of free-living proteome query structures (Supplemental Table S2.8, S2.9). GO enrichment analysis identified terms for numerous biological processes significantly enriched in this *Drosophila* target protein set, including symbiotic interactions such as modulation by symbiont of entry into host, regulation of biological processes involved in symbiotic interaction, movement within host, and golgi vesicle fusion to target membrane. The full table of results (Supplemental Table S2.10) as well as files mapping query IDs to GO terms can be found in the supplemental data. The candidate proteins we identified through alignment statistics are predicted to encode functions that are consistent with

known host-manipulation phenotypes, such as cytoplasmic incompatibility (CI), and also include multiple candidates associated with intracellular processes known to be modulated through *Wolbachia* infection.

2.4.6.1 Q73HU7

The wMel protein Q73HU7 contains a partial structural mimic to a *Drosophila* deubiquitinase. Q73HU7 contains an ovarian tumor (OTU) domain from the OTU sub-family of deubiquitinases. OTU domains are found in eukaryotes, viruses and some pathogenic bacteria(Makarova et al., 2000)**.** Their biochemical function is predicted to involve serine or cysteine protease activity, although loss-of-function OTU-class pseudo deubiquitinases have been reported(Ceccarelli et al., 2019). This protein was identified through a partial structural alignment with a *Drosophila* cysteine-type deubiquitinase Q9VUN9 at a conserved OTU domain where the target and query coverages were 0.452 and 0.532 respectively. The TM-Score of the alignment was 0.62, but from the superimposed structures it is clear that the OTU domain is conserved between the two proteins (Figure 2.4C). Furthermore the residues thought to be critical for OTU domain function and substrate recognition are also consistent at the sequence level between the wMel and *Drosophila* proteins (Figure 2.4B). The Q73HU7 structure had a free-living proteome alignment fraction of 0.02. It is well known that host manipulation tactics

employed by pathogenic bacteria such as *Legionella* often use deubiquitinase effectors(Liu et al., 2020; Wan et al., 2019). And although their full pathways are not explicitly understood yet, CI factors used by *Wolbachia* are known to have deubiquitinase activity (Beckmann et al., 2017; Chen et al., 2019).

2.4.6.2 Q73IF8

The uncharacterized *Wolbachia* protein Q73IF8 aligned with the full structure of the *Drosophila* Phosphomevalonate Kinase (Pmvk) gene product, Q9VIT2 (Supplemental Figure S2.4) . The alignment had a TM-Score of 0.53 covering 0.96 of the target structure and 0.89 of the query. The fraction of free-living proteomes aligned for Q73IF8 was 0.17. Expression of Pmvk has been shown to increase peroxisome volume in *Drosophila*(Pridie et al., 2020). Modulation of peroxisome volume and abundance may underlie the lipid metabolism modifications induced by *Wolbachia* infection, which have been attributed to influencing RNA virus replication in *Aedes albopictus* mosquito cells(Molloy et al., 2016).

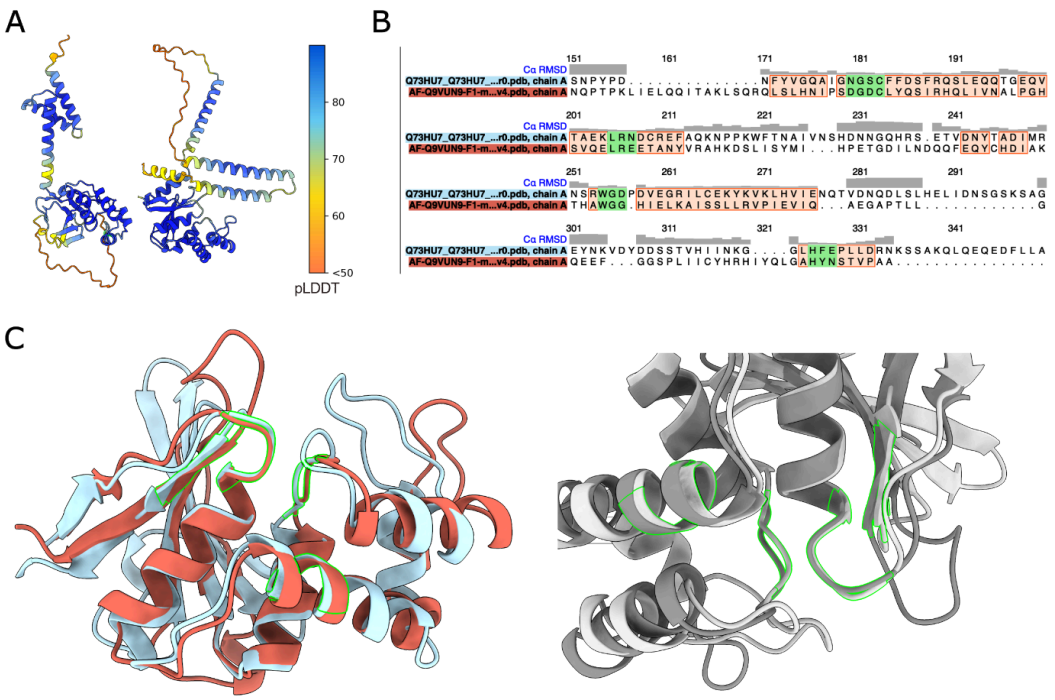

Figure 2.4. Mimic Candidate Q73HU7 Overview.

A. Full structures of query *w*Mel *Wolbachia* Q73HU7 (left) and *D. melanogaster* target Q9VUN9 (right) colored by pLDDT showing strong model confidence in the OTU domain region. **B.** Sequence alignment at structural alignment core showing low fraction of identical residues, but high similarity of critical residues at each of the four motifs characteristic of an OTU domain. Structural alignment similarity is shown by the root mean square deviation (RMSD) at the top of the plot. **C.** Structural alignment between host and microbe proteins at the region of interest (the OTU domain). Conserved residues critical to enzyme function and substrate recognition are highlighted in green.

2.4.6.3 Q73GY6

The wMel Q73GY6 RDD domain-containing protein structure partially aligned with the *Drosophila* Peroxisome assembly protein 12 (PEX12), a component of a retrotranslocation channel required for peroxisome organization. PEX12 mediates export of the Peroxisome assembly protein 5 (PEX5) receptor from

peroxisomes to the cytosol(Mast et al., 2011). The proper function of PEX12 is critical for spermatid development and inhibition of the peroxisome complex could lead to developmental defects(Chen et al., 2010). The portion of PEX12 that the *w*Mel protein of undetermined function aligns with is the membrane associated section, and the bacterial protein is missing the second zinc finger domain encoded by PEX12 (Supplemental Figure S2.5). This deletion suggests that this protein could function as a negative regulator of proper peroxisome assembly by blocking functional host components from forming working complexes through a dominant-negative interaction(Gerasimavicius et al., 2022). Similar to Q73HU7 discussed above, this candidate protein had a low fraction of free-living proteomes aligned at 0.04.

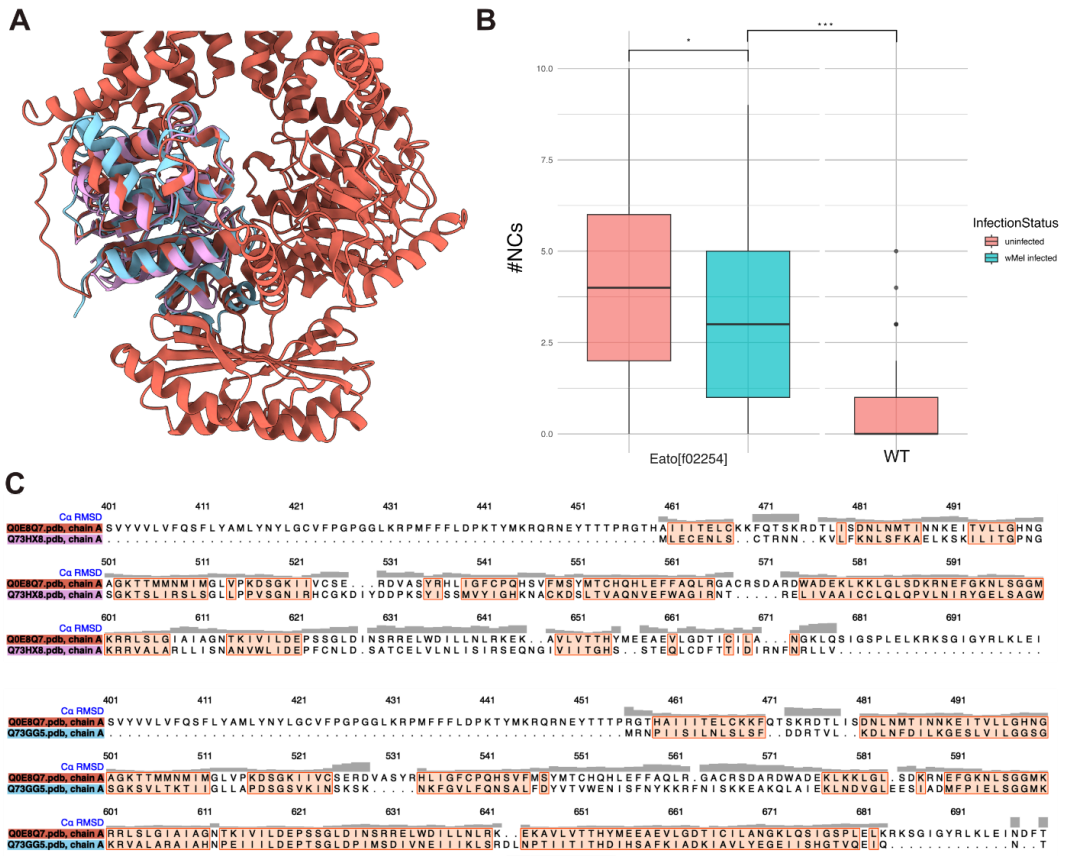

Figure 2.5. *Eato* Mimic Candidates Overview.

A. Full structures of Q73GG5 (light blue) and Q73HX8 (pink) aligned with *Drosophila's Eato* protein. **B.** Nurse cell counts observed in *Eato* knockdown flies both with and without a wMel infection and in the wild type flies expressing *Eato* (Wilcoxon Rank Sum p-value = 0.01167). **C.** Sequence alignment for both candidates with *Eato*.

2.4.6.4 Q73GG5 and Q73HX8

The *w*Mel Q73GG5 and Q73HX8 proteins have inferred ATP-binding and ABC transporter homology and exhibit strong partial alignments with a ABC transporter domain in the *Drosophila eato* gene protein product. The structural alignments have similar TM-scores at 0.86 for Q73GG5 and 0.87 for Q73HX8 (Figure 2.5A). Q73HX8 has a lower fraction of identical residues

at 0.22 compared to the 0.31 of Q73GG5. While host target protein alignment coverage was partial (0.12), a large extent of the *w*Mel query proteins were covered (0.94 and 0.95 for Q73GG5 and Q73HX8, respectively). We found Q73HX8 aligned significantly with 25% of the free living proteomes while Q73GG5 aligned with 90%. The Eato host protein promotes dead cell clearance from the anterior of the oocyte. Knockdowns of the Eato gene prevent the clearance resulting in higher nurse cell counts(Santoso et al., 2018).

2.4.6.5 Functional Investigation of Eato

Using a nurse cell count assay we investigated the impact of wMel infection on Eato function rescue. Our assay demonstrated that Eato function can be partially rescued by *w*Mel infection (Figure 2.5B). *Eato* knockouts have 4.17 nurse cells on average (range 0-10) which is substantially more than *Eato*-knockout infected flies (Wilcoxon Rank Sum p-value = 0.01167 average = 3.25, 0-9) and uninfected wildtype (Wilcoxon Rank Sum p-value = 4.91e-16 average = 0.64, 0-4). The full list of cell counts for each sample are provided in Supplemental Table S2.11. This implies that *w*Mel, which expresses these candidate mimic proteins, is able to partially rescue this phenotype. The structural and functional similarity of the *w*Mel proteins aligned with the *Drosophila* Eato structure suggest that these could be the *w*Mel effectors responsible for the rescue of a component of *D. melanogaster* fertility.

Additional functional evidence is required to demonstrate whether Q73GG5, Q73HX8, or both proteins are specifically responsible for this partial phenotypic rescue effect.

2.4.6.6 P61189

Although filtering protein mimicry candidates by their absence in free-living bacteria was an effective strategy for obtaining unique host-associated candidate genes (Figures 2.1-5), mimics do not need to be unique to host-associated clades to be involved in host-symbiont interactions. Indeed, highly conserved and essential proteins are likely targets for co-evolved interactions due to their stasis over time. Furthermore, targeting these conserved elements may provide a generalistic strategy for associating with different host taxa(Obeng et al., 2023). To search for examples of highly conserved protein structures found in many other free-living bacteria that may be utilized by intracellular organisms for survival or host manipulation, we investigated alignments that had high TM-scores covering nearly all of the target structure. We looked for lower than expected values for fraction of identical residues and fraction of free-living proteomes aligned for alignments within this region.

The *wMel* Chaperone protein P61189 demonstrates full structural similarity with a host chaperone protein, despite high prevalence among free-living

bacteria, suggesting that deep functional conservation may enable a subset of mimicry occurrences. P61189 encodes High temperature protein G (HtpG), which has been inferred from sequence homology to be a heat shock protein generally involved in cellular stress response. HtpG aligns strongly with the *Drosophila* cytoplasmic heat shock protein Hsp83, returning a TM-score of 0.9 for an alignment covering 0.941 of the *Drosophila* protein sequence and 0.984 of the wMel sequence (Figure 2.6). This alignment had a fraction of identical residues of 0.33 and a fraction of free-living proteomes aligned of 0.66. These values were below the averages of the top 100 alignments ranked by TM-score and target coverage which were 0.4 and 0.83 respectively.

Phylogenetic inference of HtpG and Hsp83 evolutionary histories reveal that the structural similarity of these proteins (TM > 0.9) is due to deep structural conservation, and not to recent horizontal gene transfer (HGT) (Supplemental Figure S2.6). As the sequence similarity between *w*Mel and *D. melanogaster* proteins was high compared to other mimic candidates (0.33 vs <0.25 for all but one candidate in Table 1), we wanted to test whether the observed structural similarity was due to horizontal gene transfer (HGT) or structural conservation. We collected sequences related to the *Wolbachia* HtpG protein by BLAST(Altschul et al., 1990) to the non-redundant database(Sayers et al., 2022), resulting in 425 bacterial, 100 archeal, and 382 eukaryotic taxa. Within the eukaryotic taxa, there were 268 insects, 21 other metazoans, and

93 other non-metazoan eukaryotes. The tree indicates HtpG (P61189) and Hsp83 (P02828) are not close relatives. While both proteins group with their congeners and higher order taxonomic affiliations, e.g., HtpG is an alphaproteobacterial gene and Hsp83 is an insect gene. The midpoint root of the tree suggests that eukaryotic sequences diverged from a shared archaeal or mitochondrial ancestor before the bacterial sequences. Deep within the eukaryotes a duplication-pseudogenization event produced two copies of the protein that are retained in some taxa. Extra copies of the gene appear enriched in symbiont-associated insects, although our dataset is enriched for these groups. Archaeal sequences were scattered throughout the tree and often represented the deepest branching clade in new domain radiations (e.g., at the base of the eukaryotic and Pseudomonadota clades). There were several examples of HGT of this protein in other clades (red taxon labels in Figure S2.6), but the direction and certainty of HGT will need to be confirmed by checking for genome mis-assemblies. Overall, the HtpG/Hsp83 chaperone protein is deeply conserved and archeal taxa appear to have mediated a spread of the chaperone across the tree of life.

Heat shock proteins are evolutionarily conserved between bacteria and eukaryotes, suggesting that these elements may be co-opting ancient housekeeping processes. The D. melanogaster Hsp83 protein is associated with the endoplasmic reticulum (ER) chaperone complex(Rosenbaum et al.,

2011) and wMel is known to alter the abundance and distribution of host ER(Fattouh et al., 2019). The wMel HtpG protein could mediate interactions with the ER and aid in the prevention of host stress responses at a cellular location specific to infection. Furthermore the Hsp83 protein has been shown to act as a specific component of the cap-binding complex where it interacts with translational repressor proteins during oogenesis(Pisa et al., 2009). The wMel strain is known to play a role in host germ-line development by rescuing the loss of host proteins essential to regulating transcription and translation(Flores et al., 2015; A. J. C. Russell et al., 2023; Starr and Cline, 2002). Specifically, wMel rescues the loss of proteins essential to regulating transcription and translation. Currently the mechanism behind this rescue phenotype is only partially characterized(Ote et al., 2016), and HtpG could be an important factor. Hsp83 is also a positive regulator of insulin signaling. It has been found that insulin signaling represses bacterial titer through the TORC1 pathway, potentially through increased autophagy(Serbus et al., 2015). This goes to suggest a third potential function of *Wolbachia'*s HtpG protein, mimicking Hsp83 to block signaling and repress TORC1 activation.

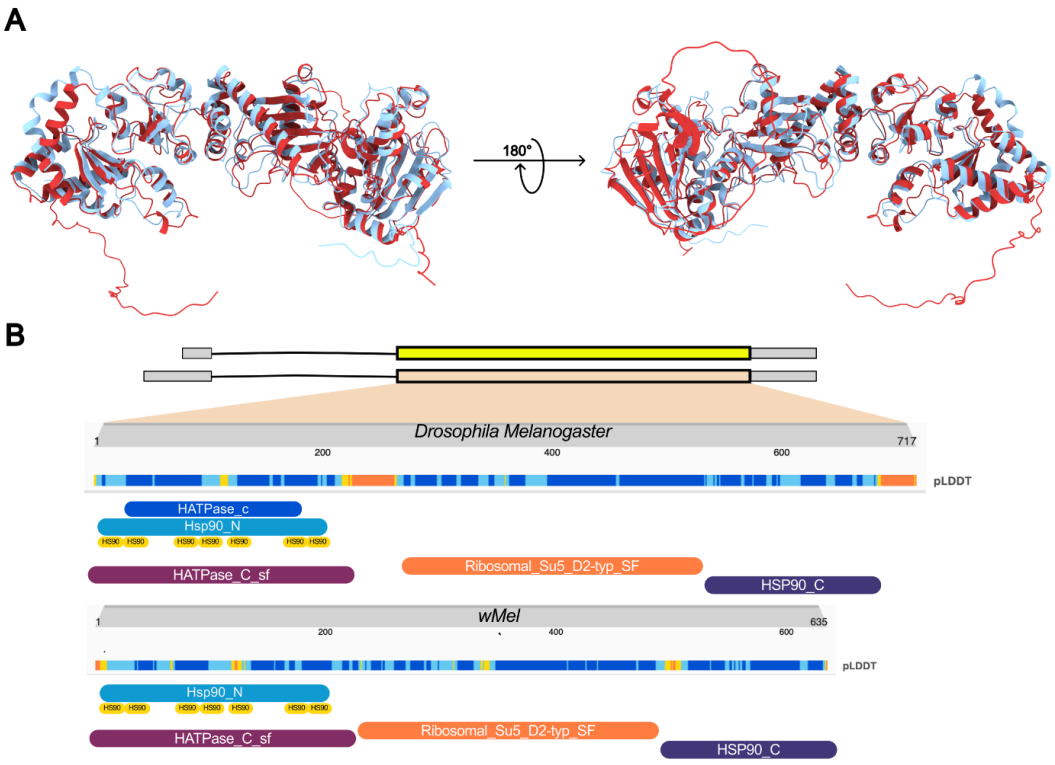

Figure 2.6. Mimic Candidate P61189

A. Structural alignment between the Wolbachia (blue) and Drosophila proteins (red).
B. Conserved domains across the length of each protein

2.5 Discussion

The intricate molecular interaction networks of host-microbe systems are an experimentally challenging set of phenomena to research, but computational protein structure prediction and alignment methods can alleviate some of this burden. This approach offers a means of rapidly identifying critical molecular components of host-microbe systems associated with many human diseases and bioengineering applications. Our use of novel structural proteome datasets and tools represents an effective scan of host-microbe systems for

structural homology that is suggestive of structural mimicry. Using examples of known structural mimicry we established filters and controls to select structural similarity, applied this structural proteome screen to multiple host-microbe systems, produced controlled structural alignments, and identified candidate proteins that may participate in mimicry. Collectively, our results demonstrate that structure-prediction and alignment approaches are a potentially appealing method for identifying molecular mimicry in host-microbe interactions. Being able to quickly identify potential genes and proteins critical to systems of host control through mimicry can accelerate research looking to perturb or harness host-microbe systems.

We were able to identify a large set of potential mimic candidates in multiple host-microbe systems. The ontology of many of these candidate mimics further corroborated their potential to be actively manipulating host biology in their respective systems. The mimic candidates presented here show strong evidence of mimic-like effectors and suggest mechanisms of host interaction for previously known virulence factors. Our initial screen of the *Drosophila*-*wMel* system provided strong examples of candidate structural mimics by using stringent filters and selecting proteins with function that can be correlated with *Wolbachia*-host phenotypes. In this system we demonstrated evidence of the predicted mimicry of Eato, where we showed Eato function in *Drosophila* oogenesis is partially rescued by wMel infection.

We also effectively applied the screen to other medically important host microbe systems. In our *Legionella*-human validation system, we identified new potential effectors and candidate mechanisms of interaction for the well-known effector, Mip. Similarly this mimic screen was able to pose new hypotheses of host manipulation by *H. pylori*, which built on the findings of recent biochemical studies looking to identify novel effectors. These results support the use of this framework as a means of both supplementing and suggesting wet-lab experiments for the interrogation of host-microbe systems.

2.6 Data Availability

All structural proteome data used in this work is publicly available through the AlphaFold Protein Structure Databases. Raw data for nurse cell count assay as well as sequences used in phylogenetic tree construction are available in the supplemental data files. Scripts and workflows used for generating control datasets and performing analysis are available at: https://github.com/gabepen/mimic_screen.

2.7 Supporting Information

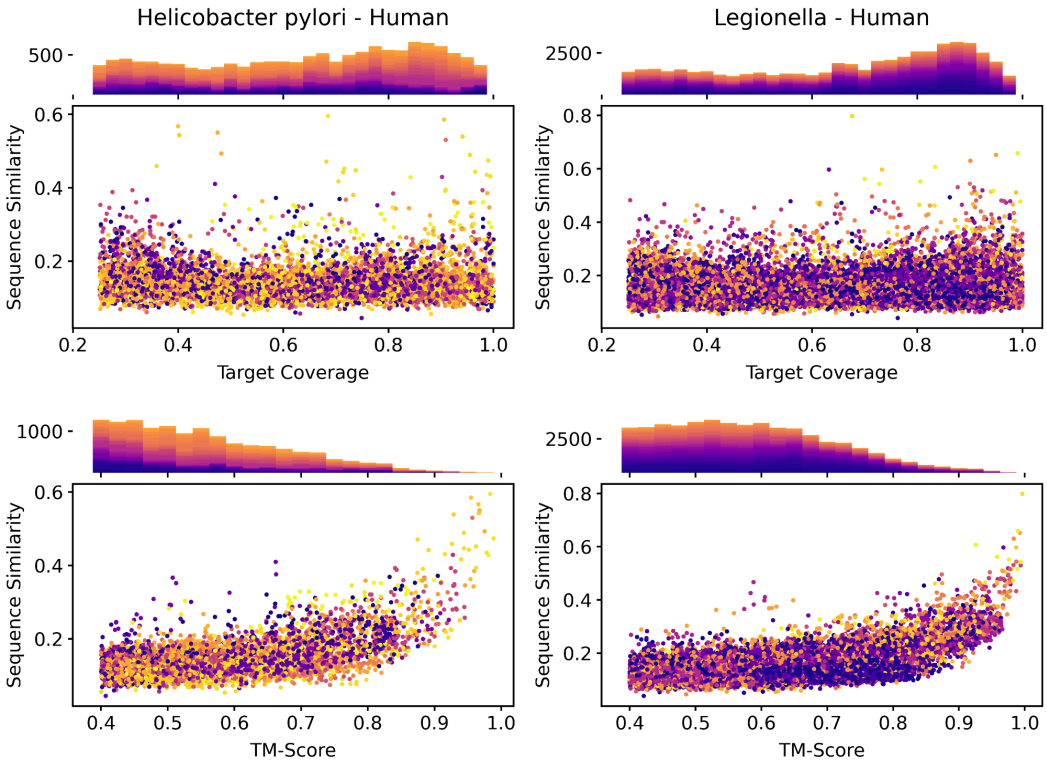

Figure S2.1. Expansive Control Alignment Space *Helicobacter pylori*-Human

Helicobacter pylori alignment space compared to the *Legionella* alignment space as described by the expansive control method. Points are colored by fraction of free-living proteomes aligned from the expansive control dataset.

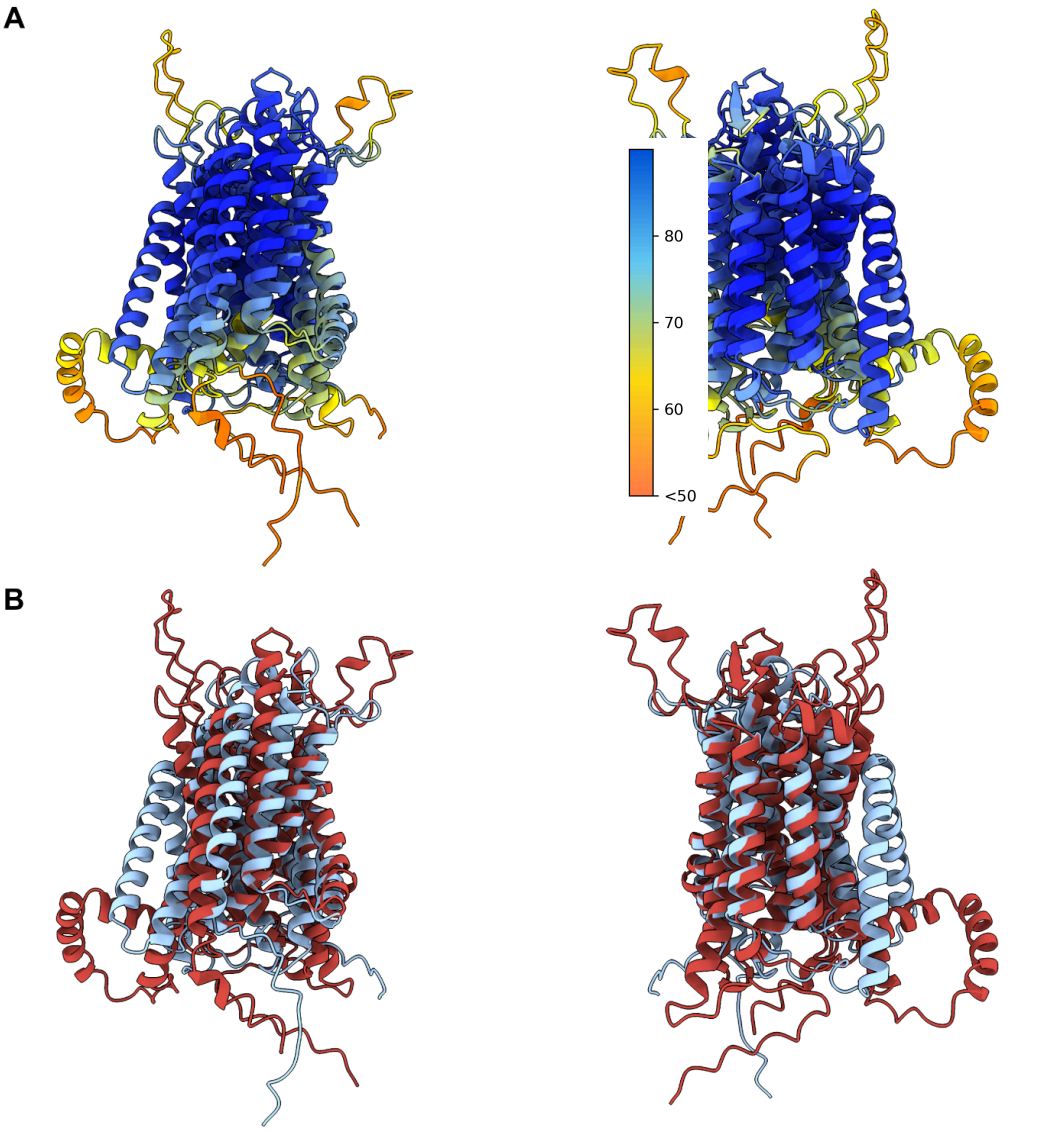
A
80
70
60
<50
B
C

Figure S2.2. Full Q5ZW23 Structure Alignment

Structural alignments between Q5ZW23 (*Legionella pneumophila*) and Q8IY34 (Human). **A.** Structural alignment colored by pLDDT. **B.** Alignment colored by structure. Q5ZW23: light blue, Q8IY34: red. **C.** Sequence alignment and structural RMSD score.

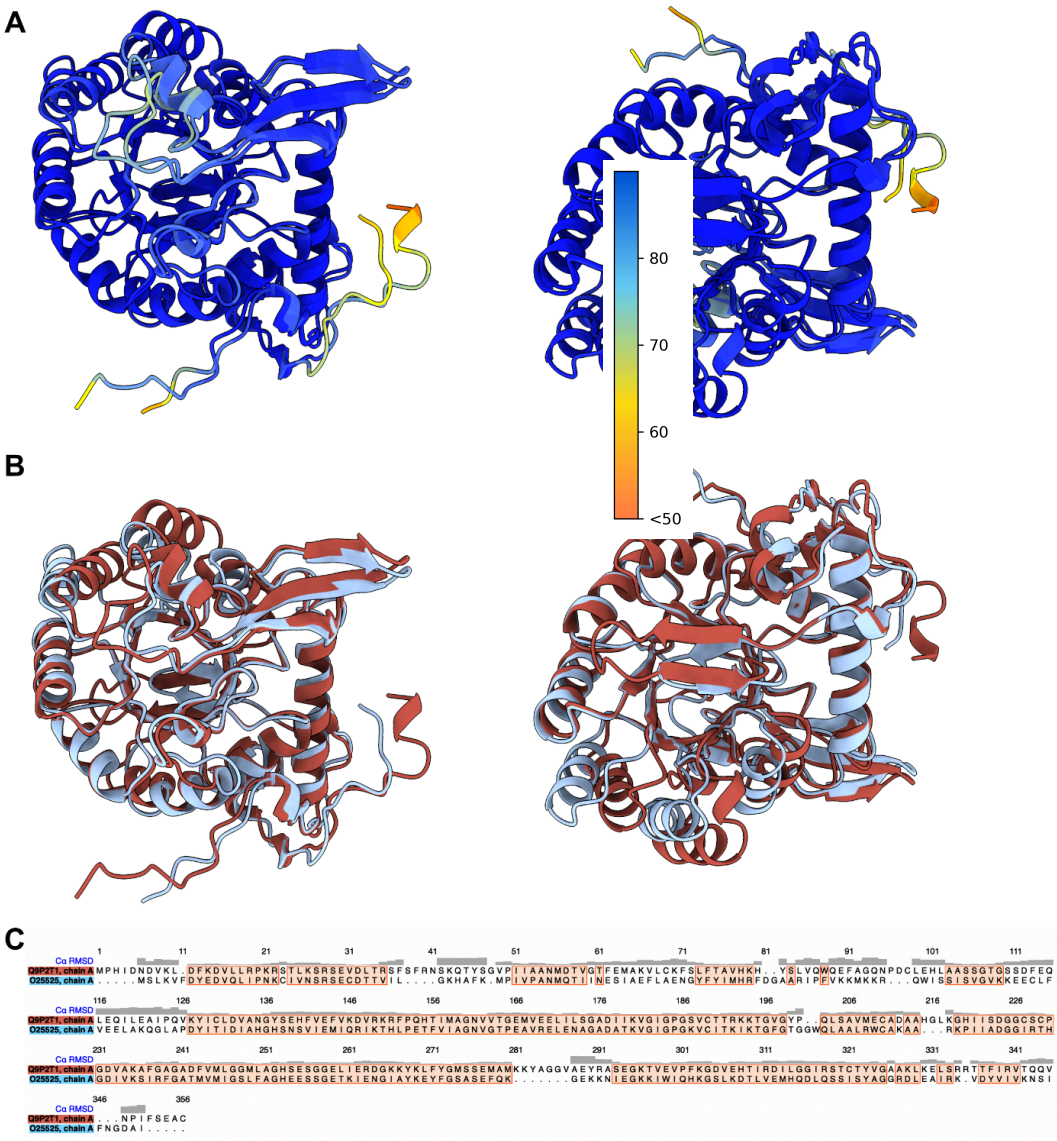

Figure S2.3. Full O25525 Structure Alignment

Structural alignments between O25525 (*Helicobacter pylori*) and Q9P2T1 (Human). **A.** Structural alignment colored by pLDDT. **B.** Alignment colored by structure.

O25525: light blue, Q9P2T1: red. **C.** Sequence alignment and structural alignment RMSD score.

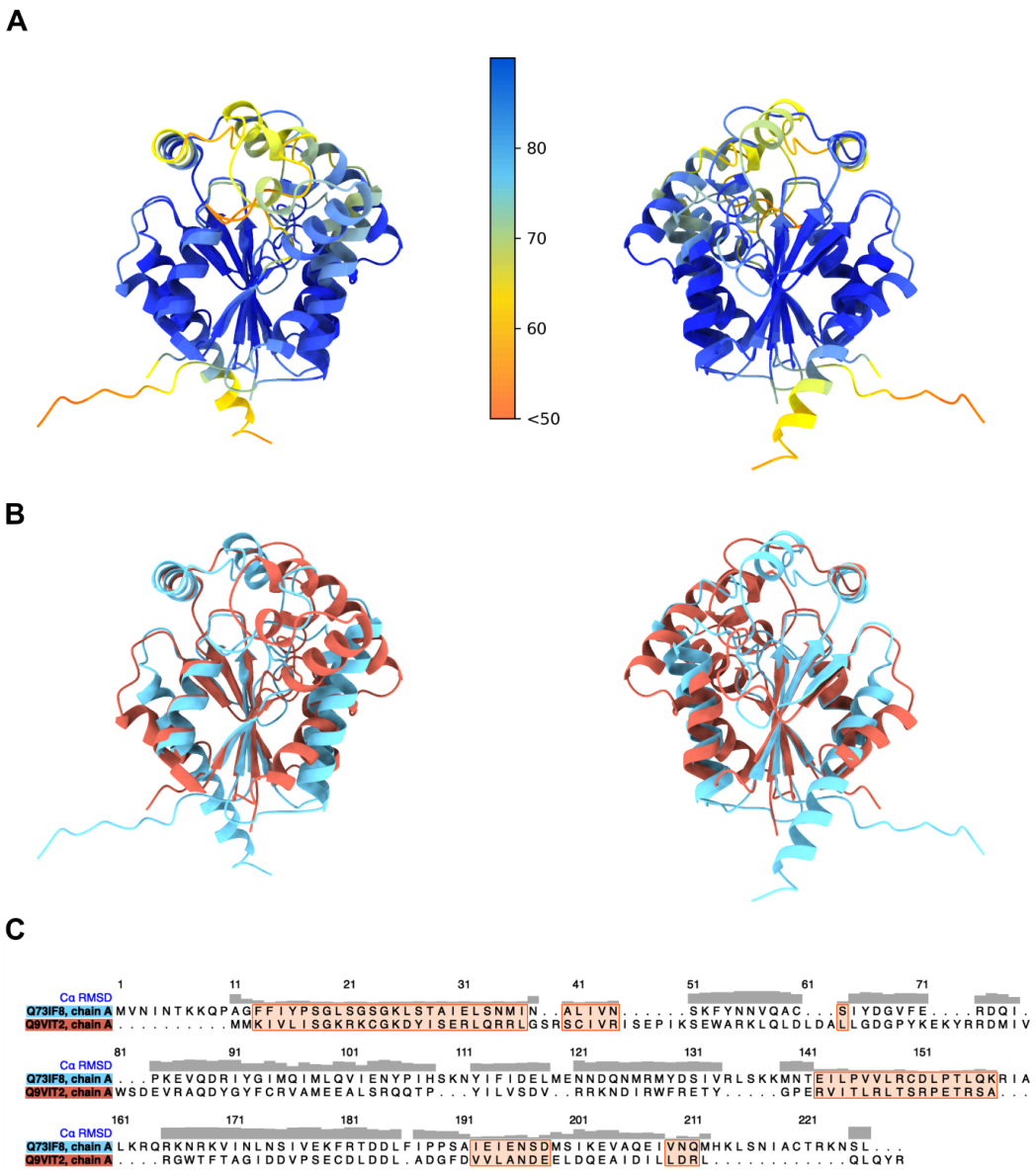

Figure S2.4. Full Q73IF8 Structure Alignment

Structural alignments between Q73IF8 (*Wolbachia wMel*) and Q9VIT2 (*Drosophila melanogaster*). **A.** Structural alignment colored by pLDDT. **B.** Alignment colored by structure. Q73IF8: light blue, Q9VIT2: red. **C.** Sequence alignment and structural alignment RMSD score.

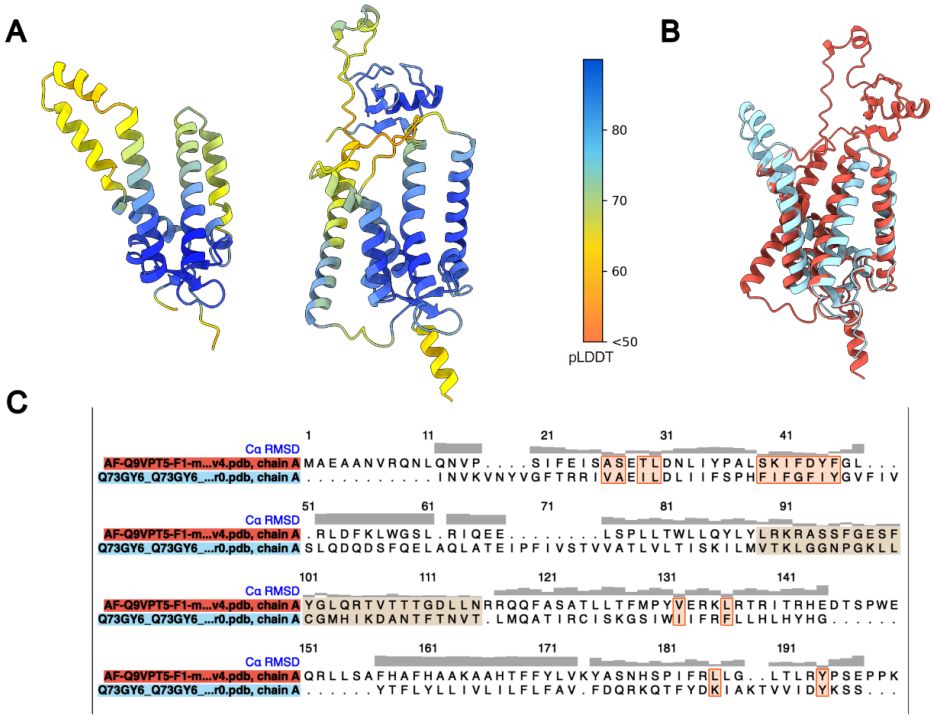

Figure S2.5. Mimic Candidate Q73GY6 Overview.

A. Full structures of *w*Mel *Wolbachia* query Q73GY6 (left) and *D. melanogaster* target Q9VPT5, the PEX12 protein (right) colored by pLDDT, showing strong model confidence in the RDD domain region (arrows). **B.** Structural alignment between the wMel and host proteins. **C**. Sequence alignment of the wMel protein and *D. melanogaster's* PEX12 protein, showing low sequence similarity. Structural alignment similarity is shown by the RMSD at the top of the plot.

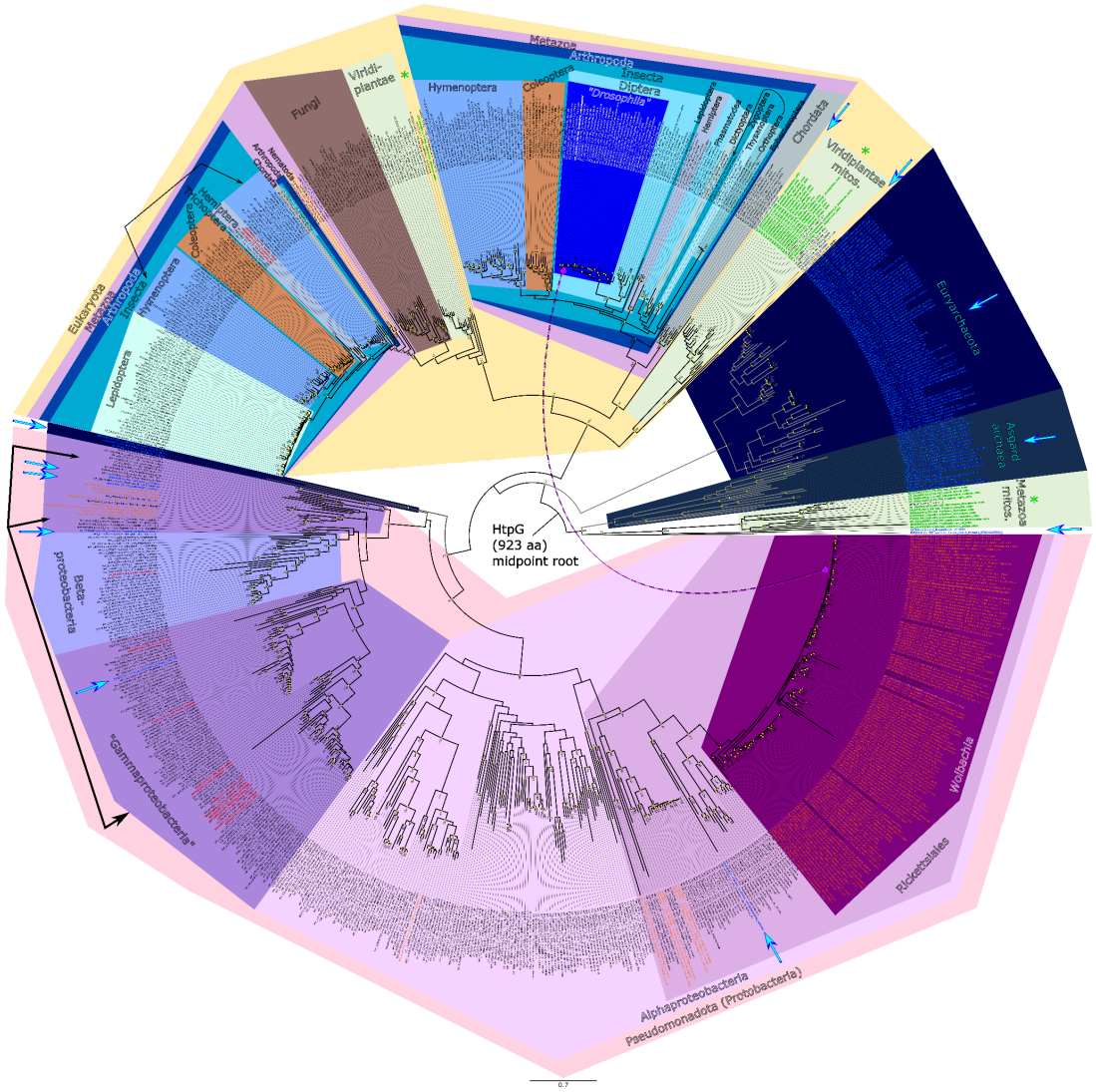

Figure S2.6. Mimic Candidate P61189 Phylogenetic Tree.

The *Wolbachia* HtpG and *Drosophila* Hsp83 proteins exhibit highly similar structural similarity, despite deep phylogenetic divergence. Midpoint rooted maximum likelihood tree of 845 chaperone sequences with detectable sequence similarity to *Wolbachia's* HtpG protein, across 923 homologous positions. Numbers at nodes indicate bootstrap support (% of trees supporting consensus topology out of 100 replicates). Values below 50% are omitted. Nodes with >=50% support are denoted with yellow circles. Annotations were added to highlight taxonomic affiliations across the tree and link *Wolbachia* and *Drosophila* proteins (dashed magenta/black line). Dark arrows on the outer perimeter link split taxonomic groupings within a clade. Archaeal taxa are spread throughout the tree, as indicated by the blue (bright/dark)

arrows and blue text. Symbiotic taxa are common among the pseudomonadota taxa, and are indicated in orange text. Red taxon names are suspected HGT events (enriched in gammaproteobacteria) or genome misassemblies (e.g., eukaryotic contigs in bacterial assemblies). The black taxon names in the Wolbachia clade are either HGT events (from Wolbachia to the host) or misassemblies (e.g., *Wolbachia* contigs in host assemblies).

Chapter III: Automating Systematic Literature Searches

3.1 Introduction

Large-scale omics analyses routinely produce candidate lists of genes, proteins, variants, small molecules, or other biological units that are statistically or computationally flagged for further investigation but whose biological relevance can only be judged against the existing literature. This is true whether the candidates come from a differential expression study, a GWAS, a drug-target screen or, as in our case, a structural alignment screen for host-microbe protein mimicry. In each case, an analysis pipeline can surface anywhere from tens to thousands of hits, and meaningfully triaging them requires reading and synthesizing what is known about each one. Comprehensive, literature informed insight into a candidate may be essential in many research contexts where follow up experiments are costly and must be rigorously vetted before investing resources.

A full systematic literature review at this scale is increasingly time-consuming and, in practice, close to infeasible (Waffenschmidt et al., 2023). This bottleneck has motivated a growing body of tool development aimed at automating literature review as a step embedded directly in the biological data analysis pipeline, rather than treating it as a separate downstream task. Early text-mining approaches focused on structured extraction by pulling gene-disease or gene-phenotype associations out of the literature to support candidate screening. Tools like OmixLitMiner(Steffen et al., 2020) applied this approach directly to omics-scale hit lists. More recently, LLM-based systems have been applied to full retrieval-and-validation literature

screening pipelines, using an LLM to score candidates and retrieve supporting literature before validating that each conclusion is actually grounded in the retrieved text. This general approach has been applied to prioritizing candidate genes from blood transcriptional signatures across a set of over 10,000 genes (Khan et al., 2025). Similar pipelines have been built for early-stage drug target discovery, automatically retrieving and screening literature to surface high-potential targets in diseases lacking established target databases (Guo et al., 2025), and LLM-based systematic review platforms have shown comparable time savings and screening performance across a range of biomedical review tasks more broadly(Cassell et al., 2025). Together these efforts point to a broader shift for many omics-scale analysis, where the limiting step is no longer generating the candidate list but reading about it, and that step is now tractable to automate.

This shift has been enabled by the increasing reliability of LLMs for STEM reasoning tasks, particularly mixture-of-experts (MoE) models, which have brought dense-model-level reasoning capability to a fraction of the inference cost (Jiang et al., 2024; Sun et al., 2024). This makes it feasible to offload thousands of human hours of literature review into a systematic, LLM-based literature analysis pipeline without the scale limitations of earlier NER/extraction-based approaches. We have developed a system and accompanying software for deploying literature analysis at the scale of hundreds of thousands of papers, spanning thousands of individual biological units, and applied it here to conduct a systematic literature review of candidate protein mimics identified in our structural proteome screen of host-microbe systems.

3.2 Methods

3.2.1 ID Mapping

For our application of this tool to the protein mimic screen data set we required an initial step of extending UniProt IDs to more literature aware IDs. In many data sets gene or protein accessions will likely be used as the main identifier throughout previous analysis so collecting standard gene names, locus tags or any other synonyms and aliases that are more likely to be used as reference in scientific text is necessary. This is referred to as ID mapping in the code base and is designed to determine the common gene names, locus tags, accession numbers and other aliases that would be used when referencing the gene or gene product.

3.2.1.1 Trusted Entrez Gene IDs

We treat MyGene.info(Xin et al., 2016) as the only authoritative source of NCBI Gene IDs. We query MyGene in batches (default 1,000 accessions) under UniProt and accession scopes (uniprot, accession, uniprot.Swiss-Prot, uniprot.TrEMBL), request the entrezgene field across all species, and record the first Gene ID when MyGene returns multiple. We deliberately ignore UniProt GeneID cross-references: those links can retain stale or

organism-mismatched NCBI Gene IDs. Accessions that MyGene does not resolve remain Entrez-unmapped and proceed to fallback identifier collection.

3.2.1.2 UniProtKB fallback identifiers

For Entrez-unmapped accessions, we fetch UniProtKB entries via the UniProt REST stream API (batches of 25) and extract: recommended gene symbol; ordered locus name; EMBL/EnsemblBacteria protein accessions; RefSeq accessions (preferring WP_ when present); UniProt gene synonyms and ORF names as aliases; and the protein recommended/alternative/submission name as a display common name. When UniProt lacks an explicit gene name, we prefer a curated ORF name (e.g., Drosophila CG#####), then a symbol-like trailing token parsed from the protein description, then the ordered locus name. We never promote bare enzyme-class words to the primary gene name.

3.2.1.3 UniParc recovery

Accessions absent from UniProtKB are queried one-by-one against the EBI Proteins UniParc API. We retain cross-references whose NCBI taxonomy ID matches the query or target taxid supplied for that side of the alignment, harvesting gene name, locus/ordered-locus tags, and

EMBL/RefSeq/EnsemblBacteria accessions when present. Existing UniProt-derived fields take precedence when both sources contribute values.

3.2.1.4 Enrichment of already Entrez-mapped accessions

After Entrez resolution, we still fetch UniProt (and, for MyGene-mapped IDs, MyGene symbol/alias/locus/accession fields) so both sides of every hit carry gene symbols, aliases, locus tags, and GenBank accessions for text search. When UniProt and MyGene disagree on the primary symbol, we keep the UniProt symbol and retain the MyGene symbol in the alias pool. We never store MyGene's free-text gene "name"/description as a common name, because those strings often encode phenotypes or methods (e.g., "knockdown") that are unsafe search terms.

3.2.1.5 GenBank locus-tag recovery

Where a GenBank/EMBL protein accession is available but no locus tag is yet known, we batch-fetch NCBI protein records (efetch, GenPept XML) and read the first CDS /locus_tag qualifier. We also sanitize locus tags by dropping bare integers and values identical to the row's Entrez Gene ID, which otherwise contaminate quoted text search.

3.2.1.6 Human symbol-as-locus convention

Human UniProt entries typically lack bacterial-style ordered locus names. For accessions on a side whose taxid is 9606, if no locus tag exists we copy the HGNC gene symbol into the locus-tag field so the first Europe PMC pass can still quote-search the symbol in the same way it quotes bacterial locus tags.

3.2.1.7 Search-term hygiene at write-out

Before emitting gene-name and alias columns, we filter tokens through shared usability rules: we reject method words, generic enzyme/functional nouns, short ambiguous symbols (e.g., gap, CS), technical aliases (probe/clone IDs), and multi-word protein phrases. If the primary gene name fails these checks, we substitute a usable locus tag when available. Aliases are deduplicated case-insensitively, stripped of the primary symbol, pipe-joined for CSV storage, and likewise filtered.

3.2.2 Literature database search

We retrieve papers for each mapped alignment row by querying Europe PMC independently for the query (microbe) and target (host) proteins, then merging the hit lists. The search module takes the identifier-mapping table (UniProt accessions plus Entrez IDs, gene symbols, aliases, locus tags, and GenBank accessions) and optional organism taxids / organism-name strings for each side. Before row-wise retrieval, we expand Entrez-linked synonyms: for human genes we pull Symbol and

Synonyms from NCBI's Homo sapiens gene_info, and for all taxa we merge MyGene.info symbol/alias fields. We never expand from free-text gene descriptions.

3.2.2.1 Text search, pass 1 (high-specificity identifiers)

We build a quoted Boolean query from the mapping-row locus tag (skipping numeric Entrez-ID junk) and GenBank/RefSeq accession. When the accession carries a version suffix, we also search the unversioned stem, because Europe PMC sometimes indexes one form but not the other. We OR the terms and require each to appear in TITLE_ABS or BODY. Pass 1 is not constrained by organism taxid.

3.2.2.2 Text search, pass 2 (gene symbols and synonyms)

We split pass 2 into a base query and a synonym-only query so we can audit which term class contributed hits. Base terms are the mapped gene name plus pipe-separated gene aliases from the idmap table. Synonym terms are the Entrez-expanded names collected above. Before querying we apply the same usability filters used at mapping write-out (reject method words, generic enzyme nouns, short ambiguous symbols, technical aliases) and cap each query at 30 unique terms. Primary gene-name clauses may match TITLE_ABS or BODY; alias and synonym clauses require TITLE_ABS only, which preserves synonym recall while limiting body-only coincidences. When

organism taxids are supplied, we AND an ORGANISM_ID clause (single taxid or OR of several). If a taxid-constrained pass-2 query returns no papers and organism-name strings are provided, we retry the same identity terms with a text organism clause (TITLE_ABS/BODY matches to those names) instead of ORGANISM_ID, and we record the fallback for auditing.

3.2.2.3 Result filtering and paper identity

From each Europe PMC response we keep records that look like primary research articles (explicit research/journal-article pubtypes, or abstract-bearing records when pubtype is missing) and drop reviews, case reports, editorials, letters, preprints, protocols, corrections, retractions, and similar non-research types. We assign each kept record a stable paper ID preferring DOI, then PMID, then PMC ID, then an Europe PMC internal ID.

For each side we union accession-citation hits with pass-1 and pass-2 text hits, deduplicating by paper ID while preserving titles. We store per-source DOI lists (europepmc_accession, text_pass1, text_pass2_base, text_pass2_synonym, and their overlap) alongside the merged query_paper_dois / target_paper_dois columns. Optionally, when accession–text overlap filtering is enabled for a side, we drop DOIs that appear only in the UniProt accession search and never in the text-search

union, which removes accession-index noise unsupported by textual mention of the gene identifiers.

3.2.3 General System Design of Paper Text Collection and Analysis

We developed this automated literature analysis tool for processing the alignment hits from the mimic screen. Although it was initially built around answering the specific research question left within our data set, we designed its use to be deployable on most research institution High-Performance Computing (HPC) clusters and to be able to perform literature searches on datasets not representing protein-protein alignments but that with any numerable units of biological concepts and the resulting list of DOIs mentioning the unit IDs. The core config components that must be customized for each research question are the paper grading rubrics for each unit type, and the synthesis instructions for the hypothesis we seek to verify through literature search. In the structural proteome alignment use case the unit types are microbe and host protein.

The paper analysis is deployed through a separate launcher to the ID mapping and literature search stage which runs as initial pipeline orchestration script locally. It initializes numerous GPU nodes to act as servers for receiving packets of papers or downstream analysis components

to process. The number and type of these nodes as well as the models loaded onto each are all modifiable through the config files. The orchestrator initializes server nodes for grader LLM, synthesis LLM, and Docling7 PDF conversion. Once nodes are reporting healthy back to the orchestrator it launches a CPU hub job responsible for the retrieval and posting of paper text to the grader and synthesis nodes. Paper packets for each gene in each alignment are downloaded with text prioritisation, PDF fallback. PDFs are converted to text using GPU powered Docling. Text packets are posted to grader nodes and completed score cards sent back to the CPU hub before being posted to the synthesis LLMs for producing the final conclusion on the research question posed.

3.2.4 Design Principles for Rubrics

We graded literature using paired, system-specific JSON rubrics: a host-protein rubric and a companion microbe-gene rubric for each host–microbe pair (L. pneumophila–human, H. pylori–human, Wolbachia wMel–Drosophila). Papers were scored under the rubric matching paper role (host vs query/microbe). Host and microbe scorecards were kept separate until synthesis; scores were never merged across rubrics before that stage.

3.2.4.1 Evaluation unit

Each scorecard scored one gene (or protein) as described in one paper. The same paper for a different gene produced an independent scorecard. Papers were compared only within a gene's evidence pool, not across genes. Graders saw a paper excerpt plus gene identifiers/aliases and judged only what the excerpt stated—not alignment metrics, generic "relevance," or unstated experiments.

3.2.4.2 Scoring scale and criteria

Every scored criterion used an ordinal 0–2 scale: 0 = absent; 1 = partial/indirect/caveated; 2 = direct and unambiguous. Each criterion carried a prompt and explicit anchors for 0/1/2. Criteria were weighted high (2×), medium or low (1×), or flag (excluded from numeric totals). Flag criteria (e.g. mimicry potential on the host side; mechanism novelty on the microbe side) were passed downstream as tags.

3.2.4.3 Axes

Rubrics were organized into independent axes with synthesis-facing role definitions.

Host side (three axes):

1. **Protein characterisation quality**: depth of functional/biochemical knowledge independent of infection context (e.g. experimental

directness of function, molecular property definition, localisation, isoform clarity, interaction evidence).

2. **Infection-process relevance:** overlap between the protein's characterised biology and processes the microbe is known to exploit, scored from functional description even when the paper never mentions the pathogen; plus infection-context evidence, system-specific evidence, and pathway proximity. A mimicry-potential flag marked biochemical similarity to known effector mechanisms.
3. **Disease/population relevance (host-only)**: genetic association, immunodeficiency links, expression in infection-relevant cell types, and clinical/cohort evidence. Reported separately and not averaged into axes 1–2.

Microbe side (two axes):

1. **Evidence quality**: how directly and rigorously the gene was studied in host context (experimental directness, host-cell evidence, secretion/translocation confirmation, phenotype connection such as intracellular replication/colonisation/symbiont phenotype, claim assertiveness).
2. **System relevance**: applicability to the focal host–microbe interaction (host cell-type match, engagement of known host target systems,

vacuole/colonisation/transmission biology, strain/isolate relevance, conservation), with a novelty flag for previously uncharacterised mechanisms.

3.2.4.4 Grading Versus Aggregation

An LLM grader produced per-criterion scores, short evidence notes, mention type, optional claim summary, and flags. It did not compute axis totals or paper grades. The pipeline aggregated weighted criterion scores into axis totals and an overall paper grade deterministically. Primary and bonus axes (configurable per rubric) guided downstream scorecard use: host primary typically infection-process relevance (with disease/population as bonus); microbe primary typically system relevance.

3.2.5 Design Principles for Synthesis

A synthesis instruction document is a short, system-specific brief for the synthesis model. Conceptually it should be written as a lens through which the LLM should view the raw text excerpts and paper grades that are passed in the data packet it receives after the instructions document. Across systems it has the same shape:

1. **System framing**: Name the microbe (query) and host (target). Tell the model which biology it is working in.

2. **Research questions**: Ask the same four questions every time, filled in with that system's language:
    a. Does the host protein sit in processes the microbe can exploit?
    b. Does the microbe gene look like an effector / secreted / host-targeting factor?
    c. How plausible is mimicry or pathway-level manipulation for this pair?
    d. How high should we prioritize the pair for follow-up?

3. **Pair context:** State why the pair exists (structural similarity), what counts as evidence, and what not to demand. We add any system-specific caveats here.

4. **Per-paper work**: Summarize each paper's bearing on query or target, using rubric scores, grader notes, and tags.

5. **Final conclusion:** Write a discussion that answers the research questions, then close with a fixed numeric score summary (0–100), not free-form labels.

3.2.6 Validation

3.2.6.1 Grader Validation

To check whether automated paper grades match an independent reading of the same evidence, we scored a blinded sample of paper-gene pairs already processed by the pipeline. The sampling frame was all graded papers from the L. pneumophila–human alignment set. Twenty pairs were drawn with a fixed random seed, stratified to 10 query-side (microbe rubric) and 10 target-side (host rubric) papers. Approximately one quarter of the sample was drawn from papers the pipeline had scored as zero relevance; the remainder was allocated across low (0–0.25), mid (0.25–0.5), and high (0.5–1) pipeline relevance bins in a 40 / 40 / 20 split.

Each item was a single paper scored for a single alignment gene (the query gene on the microbe side, the human gene on the host side). The rater received the full text, the gene's identifiers and search aliases, the paper role, and only the rubric criteria that apply to that role. Pipeline scores, rationales, and overall grades were withheld in a sealed answer key.

Scoring followed the same 0 / 1 / 2 criterion scale used by the grader (absent, partial or inferred, or direct evidence for that gene in that paper). If the alignment gene was not discussed, or appeared only incidentally, every criterion was scored 0. Evidence was not transferred from homologs, paralogs, or a different gene that was the paper's main subject. After scoring, human criterion values were aggregated with the same weighted axis math as the pipeline and compared to the withheld grades.

3.3 Results

This first application of the systematic literature analysis tool was tested with the result files of the structural proteome screen developed in chapter 2, collecting literature mentions of both the host and microbe gene in each alignment, evaluating the mentions with the various host-microbe rubrics, and then synthesizing the literature mentions of each gene into a literature backed plausibility score for the alignment representing evidence of protein mimicry. We used the pipeline to assign a 0-100 plausibility for mimicry value based on the instructed synthesis model as well as a list of papers that were graded to have some relevance to the gene mentioned within. Along with each score we connected the associated reasoning the synthesis LLM produced for that alignment as well as the summary of grading for each paper that contributed to the score.

For the Legionella-Human data set we collected, graded and synthesized the contents of 190,228 unique DOIs. This translated to a plausibility score for the research question for each alignment in the structural mimicry screen output, about 1500 alignments in total. For the wMel-Drosophila results literature review this amount was much less, mostly due to the absence of literature on human genes, at 14,684 unique DOIs. Finally we processed and analyzed 117,772 papers related to the Helicobacter pylori-human structural alignment results.

3.3.1 Grader Validation Results

Automated and blinded human grades matched on 218 of 230 criteria (94.8%), and every disagreement was a single step on the 0–2 scale. We scored 20 paper-gene

pairs from the L. pneumophila–human alignment set (10 query / microbe-rubric papers and 10 target / host-rubric papers), withholding pipeline grades. We used the same 0 / 1 / 2 criterion scale and weighted axis aggregation as the grader.

No paper differed by more than 4 points on the 32-point query or 36-point host total. Query-side paper totals tracked closely ($R^2 = 0.97$; mean absolute error 0.9 points). Host-side totals agreed well but with more scatter ($R^2 = 0.90$; MAE 1.0), driven mainly by protein characterisation (Figure 3.1).

Axis totals followed the same pattern. On query papers, evidence quality ($R^2 = 0.97$) and system relevance ($R^2 = 0.95$) both followed the identity line; the largest single-axis miss was a 4-point difference on system relevance. On host papers, infection-process scores agreed closely ($R^2 = 0.90$; MAE 0.2), protein characterisation was the weakest axis ($R^2 = 0.88$; MAE 0.8), and disease / population scores matched exactly ($R^2 = 1.00$).

Disagreements clustered on a few criteria rather than on gene identity. Typical misses were one-point differences on functional characterisation depth, biochemical property definition, conservation, translocation, LCV connection, or host-cell-type match. The grader did not assign high scores to papers we scored as off-gene or irrelevant, and it recovered the sample's zero-relevance papers as zeros.

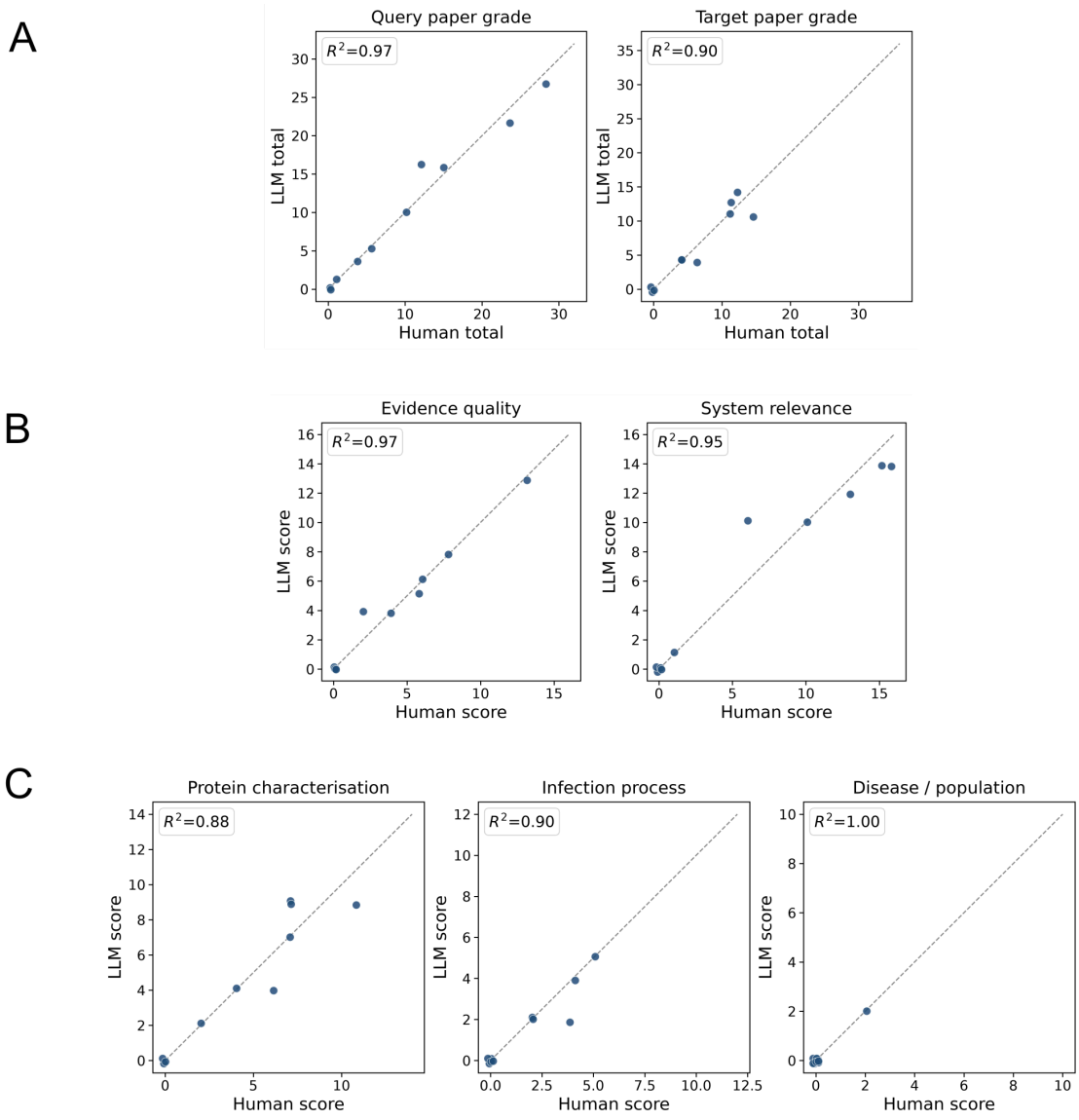

Figure 3.1. Human versus LLM rubric grades.

Each point is one paper-gene pair (n = 10 per panel). The dashed line is y = x. R² is Pearson r squared on the raw axis or paper totals. **(A)** Shows the correlation of totals for grades given on the query and target papers between the human given grades and the LLM given grades. **(B)** Per-axis comparisons for the query paper grades. **(C)** Per-axis comparisons for the target paper grades.

3.3.2 Literature Score Distributions

Known molecular mimics from L. pneumophila scored far above the background of literature reviewed alignments (Figure 3.2). Across 1,550 pairs with a completed synthesis scorecard, mimicry plausibility had mean 27.6 and median 30. The 15 control effectors had mean 71 and median 80. Eleven of the 15 sat at or above 80, the background 95th percentile.

The histogram of alignment literature grades is skewed towards no literature evidence with most pairs scoring in the 0–40 range, and only 26 pairs (1.7%) reached 85 or higher. Canonical effectors occupy that upper tail: SidD (Q5ZSQ2), VipD (Q5ZRP9), RomA (Q5ZWA1), DrrA (Q5ZSB6), AnkX (Q5ZXN6), and the GEF Q5ZU58 all score 80–90.

Four controls fall below that cluster. The ankyrin protein Q5ZT65 scores 60 and the kinase Q5ZTM4 scores 50, still above the background median. The TPR protein Q5ZTE0 scores 30, inside the bulk of the distribution. Lgt1 (Q5ZVS2) scores 0 against host Q9UNA3; that pair is a structural hit to a host protein that is not Lgt1's established eEF1A substrate, so the zero reflects the aligned pair, not a failure to recover Lgt1 biology in general.

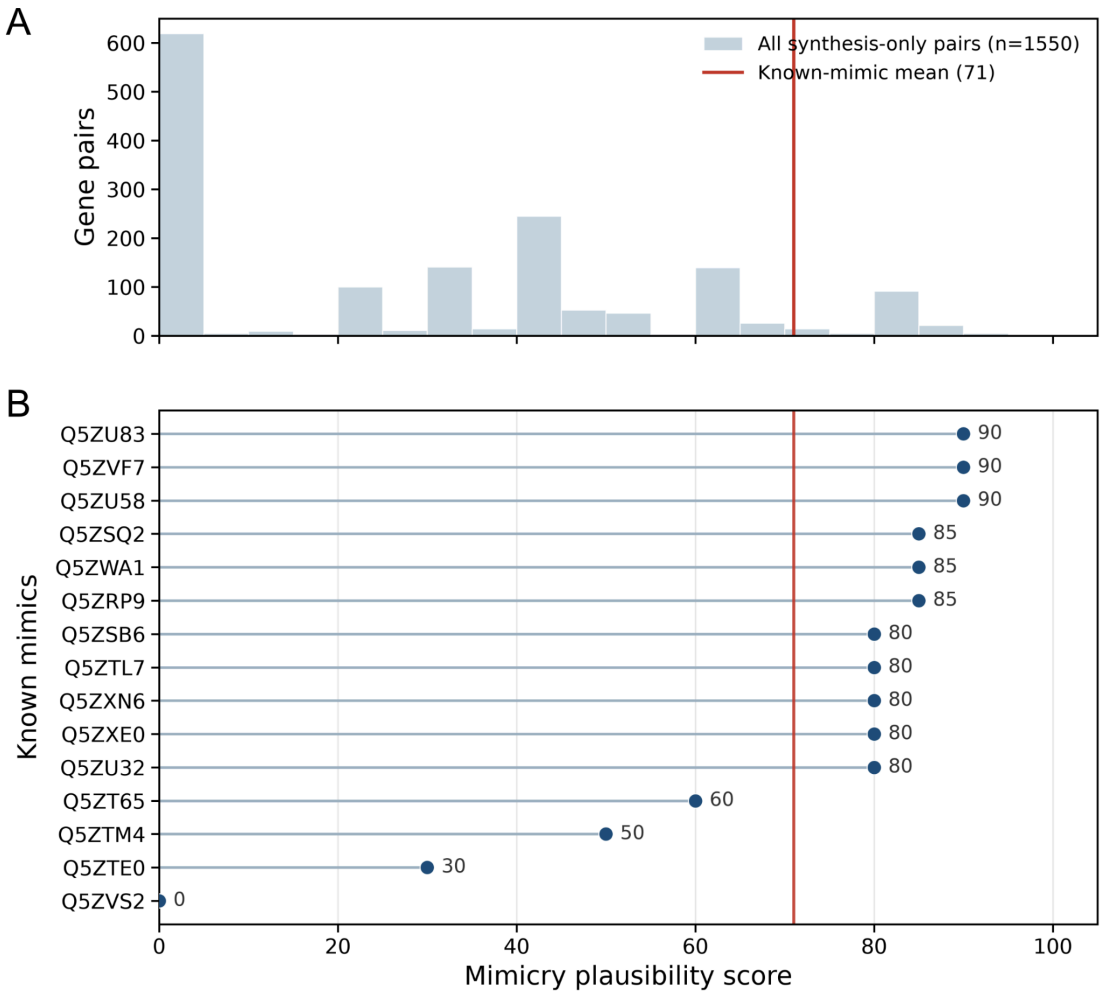

Figure 3.2 Distribution of literature based mimicry support scores for Legionella-Human structural alignments.

(A) Distribution of all scores in the analysis set (1550 host-microbe protein pairs). **(B)** Scores of Legionella effectors known to share domain or full tertiary structural similarity with a human protein.

3.4 Discussion

This literature review pipeline successfully distinguished known molecular mimics from the broader set of structural alignment candidates using literature evidence alone. The 15 previously characterized *Legionella* effectors scored far above the

background distribution of 1,550 alignments (mean 71 vs. 27.6), with the majority falling at or above the 95th percentile of all scored pairs (Table S3.1). Because these effectors were not flagged to the pipeline in any way, the system had no knowledge of which candidates were previously validated, this separation demonstrates that the synthesis scores reflect genuine literature support rather than an artifact of how the candidate list was constructed (Data S3.2). This result directly addresses the bottleneck motivating this chapter: given a large list of structural mimicry candidates, the pipeline can prioritize which pairs already have literature support worth reading in full, without requiring a person to manually review each one.

The reliability of the grading stage is a precondition for interpreting the synthesis scores and plausibility of the research question the literature review aims to answer.. Rather than a uniform decline in accuracy, disagreements between human and automated grading were concentrated in criteria that depend on inference beyond what is explicitly stated in a paper excerpt, such as characterizing a protein's function or confirming subcellular translocation. This pattern suggests the grader performs closest to human judgment on criteria with clear textual evidence and is more prone to error where a criterion requires synthesizing implication rather than reading a stated claim, a distinction that is likely to generalize to other rubric-based grading tasks built on this system. Equally important, the grader did not produce false positives for off-target or irrelevant papers, indicating that its errors are ones of precision on genuinely ambiguous evidence rather than a general tendency toward score inflation.

The distribution of synthesis scores across the full alignment set shows that high scores are specific to genuine literature support rather than a general property of well-characterized effectors. Most of the 1,550 scored pairs fell well below the scores assigned to known *Legionella* effectors, consistent with the majority of structural alignments lacking prior literature characterization for that specific host-microbe pairing. This specificity is illustrated by Lgt1, a control effector that scored much lower despite being a well-characterized effector. Lgt1 scored zero when paired with the host protein identified by the structural alignment, because that protein is not Lgt1's established substrate, eEF1A. The pipeline correctly reported no literature evidence connecting the aligned pair rather than crediting the score with Lgt1's characterization against its true target. This result clarifies what the synthesis score measures, a support for a specific structural pairing, not a general judgment of whether a gene is an established or well-studied effector.

The pipeline's grading and synthesis stages are not specific to host-microbe mimicry, which allows the same system to be applied to other types of candidate lists. Grading is scoped to a single biological unit as mentioned in a single paper, and the rubric criteria define what counts as evidence for that biological entity rather than for a gene or protein or drug’s general reputation in the literature. Nothing in the grading or synthesis stage is specific to Legionella, Wolbachia, or host-microbe mimicry as a topic. The biology-specific components are confined to two inputs, the rubric and the synthesis instruction document, while ID mapping, literature retrieval, and grading operate on any set of biological identifiers and associated DOIs. A researcher could therefore apply this system to a different candidate list, such as GWAS hits or

differential expression results, by writing a new rubric and synthesis brief rather than rebuilding the pipeline itself. This system offers a general path to keep literature interpretation tractable across the many domains where such candidate lists arise.

Within the context of this dissertation this literature review tool has been developed, tested and deployed to successfully complete a systematic literature review representing months of consistent work by multiple experts in their field, however further steps need to be taken to complete its development as a tool applicable to any large scale biological dataset requiring extensive literature review. Further validation of the overall synthesis effectiveness as well as accuracy with cited literature recall should be demonstrated in a similar manner to comparable tools currently available for systematic review such as the methods used by OmixLitMiner (Steffen et al., 2020). Meaningful validation of this tool would likely be a recreation of an already completed and published literature review. An example validation paper could be “Keeping in Touch with Type-III Secretion System Effectors: Mass Spectrometry-Based Proteomics to Study Effector–Host Protein–Protein Interactions” (Meyer et al., 2020). Although this is not a classical literature review the publication contains a structured comparison table of known effector–host interaction (EH-PPI) studies. This provides a set of effectors, their claimed host targets, and the review's own framing/prioritization of which interactions matter. A more classical literature review example that could be used to comprehensively validate this tool is “Cytokines as Biomarkers of Pancreatic Ductal Adenocarcinoma: A Systematic Review” (Yako et al., 2016). This review has a defined biological unit list in the form of cytokine-proteins and follows the 2009 PRISMA guidelines established to insure trustworthy and credible literature analysis (Liberati et al., 2009).

Conclusion

In this dissertation I have explored host-microbe systems that resist study by conventional genetic tools, using *Wolbachia* and other host-associated bacteria as test systems. Each chapter addresses one part of this problem: generating genetic variation, identifying candidate genes, and interpreting those candidates against the published literature.

Chapter I addressed the earliest of these stages by establishing that EMS mutagenesis can generate detectable genetic variation in an intracellular bacterial genome. Combining chemical mutagenesis with low error rate sequencing methods produced a nearly five-fold increase in mutation rate over background, distributed relatively uniformly across the *Wolbachia* genome rather than concentrated in specific regions or functional categories. This result does not itself constitute a phenotypic screen, but it establishes that the genome-wide variation such a screen would require can be generated and confidently detected, which was not previously demonstrated in an obligate intracellular symbiont.

Chapter II addressed candidate gene identification in a system where wet-lab screening remains impractical at omics scale. The structural mimicry screen, validated against known *Legionella pneumophila* effectors before being applied to *Helicobacter pylori* and *Wolbachia* wMel, identified candidate proteins with plausible roles in host manipulation across all three systems. The clearest validation of this approach came from the Eato gene candidates, where a nurse cell count assay

showed that wMel infection partially rescues an Eato knockout phenotype, providing direct functional support for a computationally identified pair.

Chapter III addressed the bottleneck that follows from a successful candidate screen: interpreting hundreds or thousands of hits against the published literature. Applied directly to the alignment output of Chapter II, the systematic literature review pipeline processed literature across all three host-microbe systems and, when validated against the Legionella positive controls, separated known effectors from background candidates using literature evidence alone. The Lgt1 result in particular showed that the pipeline's scores are specific to a given structural pairing rather than a general measure of how well-studied a gene is, a distinction that matters for how these scores should be used to prioritize candidates for follow-up work.

Read together, these three chapters represent tools and techniques that are built with non-model systems in mind. Chapter I shows that EMS mutagenesis can generate genetic variation in an organism with no established transformation or knockout protocol. Chapter II shows that structural mimicry screening can identify candidate genes without a screenable phenotype or existing biochemical data. Chapter III shows that an LLM-based literature pipeline can interpret large candidate lists at a scale no person could read manually. Although initially developed around *Wolbachia*-wMel the methods presented in this dissertation are readily applicable to any host-microbe system or omic study in the case of chapter III.

The organisms this dissertation focuses on, obligate intracellular symbionts and host-restricted pathogens, will likely remain outside the reach of standard genetic tools for the foreseeable future. The methods developed here do not resolve that limitation, but they offer three concrete ways to make progress within it: generating variation where none previously existed, identifying candidates without requiring a screenable phenotype, and making large candidate lists tractable to interpret. Future work extending EMS mutagenesis toward phenotypic screens, applying the mimicry screen to additional host-microbe systems, and using the literature pipeline on other classes of omics-scale candidate lists would test how far these approaches generalize beyond the systems demonstrated here.

List of Supplementary Files

Supplemental File 1: Chapter_1_supplemental_tables.xlsx

Contains all supplemental tables and data for Chapter I.

Supplemental File 2: Chapter_2_supplemental_tables.xlsx

Contains all supplemental tables and data for Chapter II.

Supplemental File 3: Chapter_3_supplemental_table.csv

Supplemental table for Chapter III.

Supplemental File 4: Chapter_3_supplemental_data1.json

Supplemental data for Chapter III.

References

Adams, K.L., Abernathy, D.G., Willett, B.C., Selland, E.K., Itoe, M.A., Catteruccia, F., 2021. Wolbachia cifB induces cytoplasmic incompatibility in the malaria mosquito vector. Nat. Microbiol. 6, 1575–1582. https://doi.org/10.1038/s41564-021-00998-6

Allen, M.G., Bate, M.Y., Tramonte, L.M., Avalos, E.Y., Loh, J., Cover, T.L., Forsyth, M.H., n.d. Regulation of Helicobacter pylori Urease and Acetone Carboxylase Genes by Nitric Oxide and the CrdRS Two-Component System. Microbiol. Spectr. 11, e04633-22. https://doi.org/10.1128/spectrum.04633-22

Almagro Armenteros, J.J., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., Nielsen, H., 2019. Detecting sequence signals in targeting peptides using deep learning. Life Sci. Alliance 2, e201900429. https://doi.org/10.26508/lsa.201900429

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Alzahrani, S., Lina, T.T., Gonzalez, J., Pinchuk, I.V., Beswick, E.J., Reyes, V.E., 2014. Effect of Helicobacter pylori on gastric epithelial cells. World J. Gastroenterol. WJG 20, 12767–12780. https://doi.org/10.3748/wjg.v20.i36.12767

Arai, H., Anbutsu, H., Nishikawa, Y., Kogawa, M., Ishii, K., Hosokawa, M., Lin, S.-R., Ueda, M., Nakai, M., Kunimi, Y., Harumoto, T., Kageyama, D., Takeyama, H., Inoue, M.N., 2023. Combined actions of bacteriophage-encoded genes in Wolbachia-induced male lethality. iScience 26. https://doi.org/10.1016/j.isci.2023.106842

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. https://doi.org/10.1038/75556

Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C.L.M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., Steinegger, M., 2023. Clustering predicted structures at the scale of the known protein universe. Nature 622, 637–645. https://doi.org/10.1038/s41586-023-06510-w

Beckmann, J.F., Ronau, J.A., Hochstrasser, M., 2017. A Wolbachia deubiquitylating enzyme induces cytoplasmic incompatibility. Nat. Microbiol. 2, 17007. https://doi.org/10.1038/nmicrobiol.2017.7

Black, M.H., Osinski, A., Gradowski, M., Servage, K.A., Paw, K., 2019. Bacterial pseudokinase catalyzes protein polyglutamylation to inhibit the SidE-family ubiquitin ligases.

Bordenstein, Sarah R., Bordenstein, Seth R., 2022. Widespread phages of endosymbionts: Phage WO genomics and the proposed taxonomic classification of Symbioviridae. PLOS Genet. 18, e1010227. https://doi.org/10.1371/journal.pgen.1010227

Brothwell, J.A., Muramatsu, M.K., Toh, E., Rockey, D.D., Putman, T.E., Barta, M.L., Hefty, P.S., Suchland, R.J., Nelson, D.E., 2016. Interrogating Genes That Mediate Chlamydia trachomatis Survival in Cell Culture Using Conditional Mutants and Recombination. J. Bacteriol. 198, 2131–2139. https://doi.org/10.1128/JB.00161-16

Burstein, D., Zusman, T., Degtyar, E., Viner, R., Segal, G., Pupko, T., 2009. Genome-Scale Identification of Legionella pneumophila Effectors Using a Machine Learning Approach. PLOS Pathog. 5, e1000508. https://doi.org/10.1371/journal.ppat.1000508

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

Cassell, K., Ologunowa, A., Rastegar-Mojarad, M., Chun, B., Huang, Y.-L., Wang, D., Cossrow, N., 2025. Analysis of article screening and data extraction performance by an AI systematic literature review platform. Front. Artif. Intell. 8. https://doi.org/10.3389/frai.2025.1662202

Ceccarelli, D.F., Ivantsiv, S., Mullin, A.A., Coyaud, E., Manczyk, N., Maisonneuve, P., Kurinov, I., Zhao, L., Go, C., Gingras, A.-C., Raught, B., Cordes, S., Sicheri, F., 2019. FAM105A/OTULINL Is a Pseudodeubiquitinase of the OTU-Class that Localizes to the ER Membrane. Structure 27, 1000-1012.e6. https://doi.org/10.1016/j.str.2019.03.022

Chen, H., Liu, Z., Huang, X., 2010. Drosophila models of peroxisomal biogenesis disorder: peroxins are required for spermatogenesis and very-long-chain fatty acid metabolism. Hum. Mol. Genet. 19, 494–505. https://doi.org/10.1093/hmg/ddp518

Chen, H., Ronau, J.A., Beckmann, J.F., Hochstrasser, M., 2019. A *Wolbachia* nuclease and its binding partner provide a distinct mechanism for cytoplasmic incompatibility. Proc. Natl. Acad. Sci. 116, 22314–22321. https://doi.org/10.1073/pnas.1914571116

Chen, L., Duan, L., Sun, M., Yang, Z., Li, H., Hu, K., Yang, H., Liu, L., 2023. Current trends and insights on EMS mutagenesis application to studies on plant

abiotic stress tolerance and development. Front. Plant Sci. 13, 1052569. https://doi.org/10.3389/fpls.2022.1052569

Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G.M., Sorger, P.K., AlQuraishi, M., 2022. Single-sequence protein structure prediction using a language model and deep learning. Nat. Biotechnol. 40, 1617–1623. https://doi.org/10.1038/s41587-022-01432-w

Create Elegant Data Visualisations Using the Grammar of Graphics [WWW Document], n.d. URL https://ggplot2.tidyverse.org/ (accessed 3.30.24).

Dahl, F., Banér, J., Gullberg, M., Mendel-Hartvig, M., Landegren, U., Nilsson, M., 2004. Circle-to-circle amplification for precise and sensitive DNA analysis. Proc. Natl. Acad. Sci. 101, 4548–4553. https://doi.org/10.1073/pnas.0400834101

DebRoy, S., Aragon, V., Kurtz, S., Cianciotto, N.P., 2006. Legionella pneumophila Mip, a Surface-Exposed Peptidylproline cis-trans-Isomerase, Promotes the Presence of Phospholipase C-Like Activity in Culture Supernatants. Infect. Immun. 74, 5152–5160. https://doi.org/10.1128/IAI.00484-06

Dufault, S.M., Tanamas, S.K., Indriani, C., Utarini, A., Ahmad, R.A., Jewell, N.P., Simmons, C.P., Anders, K.L., 2022. Disruption of spatiotemporal clustering in dengue cases by wMel Wolbachia in Yogyakarta, Indonesia. Sci. Rep. 12, 9890. https://doi.org/10.1038/s41598-022-13749-2

Fattouh, N., Cazevieille, C., Landmann, F., 2019. Wolbachia endosymbionts subvert the endoplasmic reticulum to acquire host membranes without triggering ER stress. PLoS Negl. Trop. Dis. 13, e0007218. https://doi.org/10.1371/journal.pntd.0007218

Feng, H., Wilson, A.C.C., 2024. Experimental uncoupling of hosts and endosymbionts. mBio 15, e0111624. https://doi.org/10.1128/mbio.01116-24

Flores, H.A., Bubnell, J.E., Aquadro, C.F., Barbash, D.A., 2015. The Drosophila bag of marbles Gene Interacts Genetically with Wolbachia and Shows Female-Specific Effects of Divergence. PLoS Genet. 11, e1005453. https://doi.org/10.1371/journal.pgen.1005453

Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., Hill, D.P., Lee, R., Mi, H., Moxon, S., Mungall, C.J., Muruganugan, A., Mushayahama, T., Sternberg, P.W., Thomas, P.D., Van Auken, K., Ramsey, J., Siegele, D.A., Chisholm, R.L., Fey, P., Aspromonte, M.C., Nugnes, M.V., Quaglia, F., Tosatto, S., Giglio, M., Nadendla, S., Antonazzo, G., Attrill, H., Dos Santos, G., Marygold, S., Strelets, V., Tabone, C.J., Thurmond, J., Zhou, P., Ahmed, S.H., Asanitthong, P., Luna Buitrago, D., Erdol, M.N., Gage, M.C., Ali

Kadhum, M., Li, K.Y.C., Long, M., Michalak, A., Pesala, A., Pritazahra, A., Saverimuttu, S.C.C., Su, R., Thurlow, K.E., Lovering, R.C., Logie, C., Oliferenko, S., Blake, J., Christie, K., Corbani, L., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D., Smith, C., Cuzick, A., Seager, J., Cooper, L., Elser, J., Jaiswal, P., Gupta, P., Jaiswal, P., Naithani, S., Lera-Ramirez, M., Rutherford, K., Wood, V., De Pons, J.L., Dwinell, M.R., Hayman, G.T., Kaldunski, M.L., Kwitek, A.E., Laulederkind, S.J.F., Tutaj, M.A., Vedi, M., Wang, S.-J., D'Eustachio, P., Aimo, L., Axelsen, K., Bridge, A., Hyka-Nouspikel, N., Morgat, A., Aleksander, S.A., Cherry, J.M., Engel, S.R., Karra, K., Miyasato, S.R., Nash, R.S., Skrzypek, M.S., Weng, S., Wong, E.D., Bakker, E., Berardini, T.Z., Reiser, L., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Blatter, M.-C., Boutet, E., Breuza, L., Bridge, A., Casals-Casas, C., Coudert, E., Estreicher, A., Livia Famiglietti, M., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Pedruzzi, I., Pourcel, L., Poux, S., Rivoire, C., Sundaram, S., Bateman, A., Bowler-Barnett, E., Bye-A-Jee, H., Denny, P., Ignatchenko, A., Ishtiaq, R., Lock, A., Lussi, Y., Magrane, M., Martin, M.J., Orchard, S., Raposo, P., Speretta, E., Tyagi, N., Warner, K., Zaru, R., Diehl, A.D., Lee, R., Chan, J., Diamantakis, S., Raciti, D., Zarowiecki, M., Fisher, M., James-Zorn, C., Ponferrada, V., Zorn, A., Ramachandran, S., Ruzicka, L., Westerfield, M., 2023. The Gene Ontology knowledgebase in 2023. Genetics 224, iyad031. https://doi.org/10.1093/genetics/iyad031

Gerasimavicius, L., Livesey, B.J., Marsh, J.A., 2022. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. Nat. Commun. 13, 3895. https://doi.org/10.1038/s41467-022-31686-6

Gomez-Valero, L., Buchrieser, C., 2013. Genome Dynamics in Legionella: The Basis of Versatility and Adaptation to Intracellular Replication. Cold Spring Harb. Perspect. Med. 3, a009993. https://doi.org/10.1101/cshperspect.a009993

Gomez-Valero, L., Rusniok, C., Carson, D., Mondino, S., Pérez-Cobas, A.E., Rolando, M., Pasricha, S., Reuter, S., Demirtas, J., Crumbach, J., Descorps-Declere, S., Hartland, E.L., Jarraud, S., Dougan, G., Schroeder, G.N., Frankel, G., Buchrieser, C., 2019. More than 18,000 effectors in the Legionella genus genome provide multiple, independent combinations for replication in human cells. Proc. Natl. Acad. Sci. 116, 2265–2273. https://doi.org/10.1073/pnas.1808016116

Guo, S., Shang, T., Pan, Y., Wu, Y., Li, Y., Zhou, M., 2025. TARGETFLOW:An automated literature mining pipeline for accelerating the discovery of high-potential targets in rare and mature diseases. MethodsX 15, 103735. https://doi.org/10.1016/j.mex.2025.103735

Haenig, C., Atias, N., Taylor, A.K., Mazza, A., Schaefer, M.H., Russ, J., Riechers, S.-P., Jain, S., Coughlin, M., Fontaine, J.-F., Freibaum, B.D., Brusendorf, L.,

Zenkner, M., Porras, P., Stroedicke, M., Schnoegl, S., Arnsburg, K., Boeddrich, A., Pigazzini, L., Heutink, P., Taylor, J.P., Kirstein, J., Andrade-Navarro, M.A., Sharan, R., Wanker, E.E., 2020. Interactome Mapping Provides a Network of Neurodegenerative Disease Proteins and Uncovers Widespread Protein Aggregation in Affected Brains. Cell Rep. 32, 108050. https://doi.org/10.1016/j.celrep.2020.108050

Hawk, C., Hamdzah, N., Dimond, Z., Fields, K.A., 2025. A platform supporting generation and isolation of random transposon mutants in Chlamydia trachomatis. J. Bacteriol. 207, e00500-24. https://doi.org/10.1128/jb.00500-24

Hebert, F.O., Phelps, L., Samonte, I., Panchal, M., Grambauer, S., Barber, I., Kalbe, M., Landry, C.R., Aubin-Horth, N., 2015. Identification of candidate mimicry proteins involved in parasite-driven phenotypic changes. Parasit. Vectors 8, 225. https://doi.org/10.1186/s13071-015-0834-1

Hedges, L.M., Johnson, K.N., 2008. Induction of host defence responses by Drosophila C virus. J. Gen. Virol. 89, 1497–1501. https://doi.org/10.1099/vir.0.83684-0

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., 2011. Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. Nature 476, 454–457. https://doi.org/10.1038/nature10356

Hönigschmid, P., Bykova, N., Schneider, R., Ivankov, D., Frishman, D., 2018. Evolutionary Interplay between Symbiotic Relationships and Patterns of Signal Peptide Gain and Loss. Genome Biol. Evol. 10, 928–938. https://doi.org/10.1093/gbe/evy049

Hosokawa, T., Koga, R., Kikuchi, Y., Meng, X.-Y., Fukatsu, T., 2010. Wolbachia as a bacteriocyte-associated nutritional mutualist. Proc. Natl. Acad. Sci. 107, 769–774. https://doi.org/10.1073/pnas.0911476107

Hubber, A., Roy, C.R., 2010. Modulation of Host Cell Function by *Legionella pneumophila* Type IV Effectors. Annu. Rev. Cell Dev. Biol. 26, 261–283. https://doi.org/10.1146/annurev-cellbio-100109-104034

Jacobs, J., Mirchandani, C., Seligmann, W.E., Sacco, S., Escalona, M., Russell, S.L., n.d. Wolbachia induces host cell identity changes and determines symbiotic fate in Drosophila.

Jia, Y., Yun, C.-H., Park, E., Ercan, D., Manuia, M., Juarez, J., Xu, C., Rhee, K., Chen, T., Zhang, H., Palakurthi, S., Jang, J., Lelais, G., DiDonato, M., Bursulaya, B., Michellys, P.-Y., Epple, R., Marsilje, T.H., McNeill, M., Lu, W., Harris, J., Bender, S., Wong, K.-K., Jänne, P.A., Eck, M.J., 2016. Overcoming

EGFR T790M and C797S resistance with mutant-selective allosteric inhibitors. Nature 534, 129–132. https://doi.org/10.1038/nature17960

Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D. de las, Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E., 2024. Mixtral of Experts. https://doi.org/10.48550/arXiv.2401.04088

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kao, C.-Y., Sheu, B.-S., Wu, J.-J., 2016. Helicobacter pylori infection: An overview of bacterial virulence factors and pathogenesis. Biomed. J. 39, 14–23. https://doi.org/10.1016/j.bj.2015.06.002

Kari, L., Goheen, M.M., Randall, L.B., Taylor, L.D., Carlson, J.H., Whitmire, W.M., Virok, D., Rajaram, K., Endresz, V., McClarty, G., Nelson, D.E., Caldwell, H.D., 2011. Generation of targeted Chlamydia trachomatis null mutants. Proc. Natl. Acad. Sci. U. S. A. 108, 7189–7193. https://doi.org/10.1073/pnas.1102229108

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010

Khan, T., Toufiq, M., Yurieva, M., Indrawattana, N., Jittmittraphap, A., Kosoltanapiwat, N., Pumirat, P., Sukphopetch, P., Vanaporn, M., Palucka, K., Syed Ahamed Kabeer, B., Rinchai, D., Chaussabel, D., 2025. Automating candidate gene prioritization with large language models: from naive scoring to literature-grounded validation. Bioinformatics 41, btaf541. https://doi.org/10.1093/bioinformatics/btaf541

Kokes, M., Dunn, J.D., Granek, J.A., Nguyen, B.D., Barker, J.R., Valdivia, R.H., Bastidas, R.J., 2015. Integrating chemical mutagenesis and whole genome sequencing as a platform for forward and reverse genetic analysis of Chlamydia. Cell Host Microbe 17, 716–725. https://doi.org/10.1016/j.chom.2015.03.014

Kosakovsky Pond, S.L., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B., Shank, S.D., Magalis, B.R., Bouvier, D., Nekrutenko, A., Wisotsky,

S., Spielman, S.J., Frost, S.D.W., Muse, S.V., 2019. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Mol. Biol. Evol. 37, 295–299. https://doi.org/10.1093/molbev/msz197

Kuhlman, B., Bradley, P., 2019. Advances in protein structure prediction and design. Nat. Rev. Mol. Cell Biol. 20, 681–697. https://doi.org/10.1038/s41580-019-0163-x

Lamason, R.L., Kafai, N.M., Welch, M.D., 2018. A streamlined method for transposon mutagenesis of Rickettsia parkeri yields numerous mutations that impact infection. PLOS ONE 13, e0197012. https://doi.org/10.1371/journal.pone.0197012

Lasso, G., Honig, B., Shapira, S.D., 2020. A sweep of earth's virome reveals host-guided viral protein structural mimicry; with implications for human disease (preprint). Microbiology. https://doi.org/10.1101/2020.06.18.159467

Lee, A., Fox, J., Hazell, S., 1993. Pathogenicity of Helicobacter pylori: a perspective. Infect. Immun. 61, 1601–1610. https://doi.org/10.1128/iai.61.5.1601-1610.1993

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 339, b2700. https://doi.org/10.1136/bmj.b2700

Liechti, G., Goldberg, J.B., 2012. Helicobacter pylori Relies Primarily on the Purine Salvage Pathway for Purine Nucleotide Biosynthesis. J. Bacteriol. 194, 839–854. https://doi.org/10.1128/JB.05757-11

Lina, T.T., Alzahrani, S., Gonzalez, J., Pinchuk, I.V., Beswick, E.J., Reyes, V.E., 2014. Immune evasion strategies used by Helicobacter pylori. World J. Gastroenterol. WJG 20, 12753–12766. https://doi.org/10.3748/wjg.v20.i36.12753

Liu, S., Luo, J., Zhen, X., Qiu, J., Ouyang, S., Luo, Z.-Q., 2020. Interplay between bacterial deubiquitinase and ubiquitin E3 ligase regulates ubiquitin dynamics on Legionella phagosomes. eLife 9, e58114. https://doi.org/10.7554/eLife.58114

Lou, D.I., Hussmann, J.A., McBee, R.M., Acevedo, A., Andino, R., Press, W.H., Sawyer, S.L., 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc. Natl. Acad. Sci. 110, 19872–19877. https://doi.org/10.1073/pnas.1319590110

Ludin, P., Nilsson, D., Mäser, P., 2011. Genome-Wide Identification of Molecular Mimicry Candidates in Parasites. PLoS ONE 6, e17546. https://doi.org/10.1371/journal.pone.0017546

Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Miller, W., Haussler, D., 2008. The infinite sites model of genome evolution. Proc. Natl. Acad. Sci. 105, 14254–14261. https://doi.org/10.1073/pnas.0805217105

Makarova, K.S., Aravind, L., Koonin, E.V., 2000. A novel superfamily of predicted cysteine proteases from eukaryotes, viruses and Chlamydia pneumoniae. Trends Biochem. Sci. 25, 50–52. https://doi.org/10.1016/S0968-0004(99)01530-3

Malfertheiner, P., Camargo, M.C., El-Omar, E., Liou, J.-M., Peek, R., Schulz, C., Smith, S.I., Suerbaum, S., 2023. Helicobacter pylori infection. Nat. Rev. Dis. Primer 9, 1–24. https://doi.org/10.1038/s41572-023-00431-8

Martin, M., Newton, I.L.G., 2023. The Wolbachia WalE1 effector alters Drosophila endocytosis. bioRxiv 2023.02.26.530160. https://doi.org/10.1101/2023.02.26.530160

Masson, F., Lemaitre, B., 2020. Growing Ungrowable Bacteria: Overview and Perspectives on Insect Symbiont Culturability. Microbiol. Mol. Biol. Rev. MMBR 84, e00089-20. https://doi.org/10.1128/MMBR.00089-20

Mast, F.D., Li, J., Virk, M.K., Hughes, S.C., Simmonds, A.J., Rachubinski, R.A., 2011. A *Drosophila* model for the Zellweger spectrum of peroxisome biogenesis disorders. Dis. Model. Mech. 4, 659–672. https://doi.org/10.1242/dmm.007419

McClure, E.E., Oliva Chávez, A.S., Shaw, D.K., Carlyon, J.A., Ganta, R.R., Noh, S.M., Wood, D.O., Bavoil, P.M., Brayton, K.A., Martinez, J.J., McBride, J.W., Valdivia1, R.H., Munderloh, U.G., Pedra, J.H.F., 2017. Engineering of obligate intracellular bacteria: progress, challenges and paradigms. Nat. Rev. Microbiol. 15, 544–558. https://doi.org/10.1038/nrmicro.2017.59

McCoy-Simandle, K., Stewart, C.R., Dao, J., DebRoy, S., Rossier, O., Bryce, P.J., Cianciotto, N.P., 2011. Legionella pneumophila Type II Secretion Dampens the Cytokine Response of Infected Macrophages and Epithelia ▽. Infect. Immun. 79, 1984–1997. https://doi.org/10.1128/IAI.01077-10

McCutcheon, J.P., Garber, A.I., Spencer, N., Warren, J.M., 2024. How do bacterial endosymbionts work with so few genes? PLOS Biol. 22, e3002577. https://doi.org/10.1371/journal.pbio.3002577

Mendz, G.L., Jimenez, B.M., Hazell, S.L., Gero, A.M., O'Sullivan, W.J., 1994. Salvage synthesis of purine nucleotides by Helicobacter pylori. J. Appl. Bacteriol. 77, 674–681. https://doi.org/10.1111/j.1365-2672.1994.tb02818.x

Meyer, M.D., Ryck, J.D., Goormachtig, S., Van Damme, P., 2020. Keeping in Touch with Type-III Secretion System Effectors: Mass Spectrometry-Based Proteomics to Study Effector–Host Protein–Protein Interactions. Int. J. Mol. Sci. 21, 6891. https://doi.org/10.3390/ijms21186891

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirchandani, C., Genetti, M., Wang, P., Pepper-Tunick, E., Russell, S., Corbett-Detig, R., 2024a. Plate Scale Tn5 based tagmentation library prep protocol.

Mirchandani, C., Wang, P., Jacobs, J., Genetti, M., Pepper-Tunick, E., Sullivan, W.T., Corbett-Detig, R., Russell, S.L., 2024b. Mixed Wolbachia infections resolve rapidly during in vitro evolution. PLOS Pathog. 20, e1012149. https://doi.org/10.1371/journal.ppat.1012149

Mladenova-Hristova, I., Grekova, O., Patel, A., 2017. Zoonotic potential of Helicobacter spp. J. Microbiol. Immunol. Infect. 50, 265–269. https://doi.org/10.1016/j.jmii.2016.11.003

Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. https://doi.org/10.12688/f1000research.29032.2

Molloy, J.C., Sommer, U., Viant, M.R., Sinkins, S.P., 2016. Wolbachia Modulates Lipid Metabolism in Aedes albopictus Mosquito Cells. Appl. Environ. Microbiol. 82, 3109–3120. https://doi.org/10.1128/AEM.00275-16

Mondino, S., Schmidt, S., Buchrieser, C., 2020. Molecular Mimicry: a Paradigm of Host-Microbe Coevolution Illustrated by *Legionella*. mBio 11, e01201-20. https://doi.org/10.1128/mBio.01201-20

Mori, N., Sato, H., Hayashibara, T., Senba, M., Geleziunas, R., Wada, A., Hirayama, T., Yamamoto, N., 2003. Helicobacter pylori induces matrix metalloproteinase-9 through activation of nuclear factor κB. Gastroenterology 124, 983–992. https://doi.org/10.1053/gast.2003.50152

Nakamura, N., Lill, J.R., Phung, Q., Jiang, Z., Bakalarski, C., de Mazière, A., Klumperman, J., Schlatter, M., Delamarre, L., Mellman, I., 2014. Endosomes are specialized platforms for bacterial sensing and NOD2 signalling. Nature 509, 240–244. https://doi.org/10.1038/nature13133

Obeng, N., Czerwinski, A., Schütz, D., Michels, J., Leipert, J., Bansept, F., García García, M.J., Schultheiß, T., Kemlein, M., Fuß, J., Tholey, A., Traulsen, A., Sondermann, H., Schulenburg, H., 2023. Bacterial c-di-GMP has a key role in

establishing host–microbe symbiosis. Nat. Microbiol. 8, 1809–1819. https://doi.org/10.1038/s41564-023-01468-x

Ote, M., Ueyama, M., Yamamoto, D., 2016. Wolbachia Protein TomO Targets nanos mRNA and Restores Germ Stem Cells in Drosophila Sex-lethal Mutants. Curr. Biol. 26, 2223–2232. https://doi.org/10.1016/j.cub.2016.06.054

Pisa, V., Cozzolino, M., Gargiulo, S., Ottone, C., Piccioni, F., Monti, M., Gigliotti, S., Talamo, F., Graziani, F., Pucci, P., Verrotti, A.C., 2009. The molecular chaperone Hsp90 is a component of the cap-binding complex and interacts with the translational repressor Cup during Drosophila oogenesis. Gene 432, 67–74. https://doi.org/10.1016/j.gene.2008.11.025

Pridie, C., Ueda, K., Simmonds, A.J., 2020. Rosy Beginnings: Studying Peroxisomes in Drosophila. Front. Cell Dev. Biol. 8, 835. https://doi.org/10.3389/fcell.2020.00835

Reimer, L.C., Sardà Carbasse, J., Koblitz, J., Ebeling, C., Podstawka, A., Overmann, J., 2022. Bac *Dive* in 2022: the knowledge base for standardized bacterial and archaeal data. Nucleic Acids Res. 50, D741–D746. https://doi.org/10.1093/nar/gkab961

Rosenbaum, E.E., Brehm, K.S., Vasiljevic, E., Liu, C.-H., Hardie, R.C., Colley, N.J., 2011. XPORT-Dependent Transport of TRP and Rhodopsin. Neuron 72, 602–615. https://doi.org/10.1016/j.neuron.2011.09.016

Russell, A.J.C., Sanderson, T., Bushell, E., Talman, A.M., Anar, B., Girling, G., Hunziker, M., Kent, R.S., Martin, J.S., Metcalf, T., Montandon, R., Pandey, V., Pardo, M., Roberts, A.B., Sayers, C., Schwach, F., Choudhary, J.S., Rayner, J.C., Voet, T., Modrzynska, K.K., Waters, A.P., Lawniczak, M.K.N., Billker, O., 2023. Regulators of male and female sexual development are critical for the transmission of a malaria parasite. Cell Host Microbe 31, 305-319.e10. https://doi.org/10.1016/j.chom.2022.12.011

Russell, S., Penunuri, G., Condon, C., 2024. Diverse Genetic Conflicts Mediated by Molecular Mimicry and Computational Approaches to Detect Them. https://doi.org/10.20944/preprints202403.1765.v1

Russell, S.L., Castillo, J.R., Sullivan, W.T., 2023. Wolbachia endosymbionts manipulate the self-renewal and differentiation of germline stem cells to reinforce fertility of their fruit fly host. PLOS Biol. 21, e3002335. https://doi.org/10.1371/journal.pbio.3002335

Russell, S.L., Lemseffer, N., Sullivan, W.T., 2018. Wolbachia and host germline components compete for kinesin-mediated transport to the posterior pole of the Drosophila oocyte. PLOS Pathog. 14, e1007216. https://doi.org/10.1371/journal.ppat.1007216

Salama, L.A., Alomar, A.A., Pérez-Ramos, D.W., Caragata, E.P., 2024. Using Wolbachia Bacteria to Control Mosquito-Borne Disease : ENY2115/IN1434, 10/2024. EDIS 2024. https://doi.org/10.32473/edis-in1434-2024

Sansom, F.M., Newton, H.J., Crikis, S., Cianciotto, N.P., Cowan, P.J., d'Apice, A.J.F., Hartland, E.L., 2007. A bacterial ecto-triphosphate diphosphohydrolase similar to human CD39 is essential for intracellular multiplication of Legionella pneumophila. Cell. Microbiol. 9, 1922–1935. https://doi.org/10.1111/j.1462-5822.2007.00924.x

Santoso, C.S., Meehan, T.L., Peterson, J.S., Cedano, T.M., Turlo, C.V., McCall, K., 2018. The ABC Transporter Eato Promotes Cell Clearance in the Drosophila melanogaster Ovary. G3 GenesGenomesGenetics 8, 833–843. https://doi.org/10.1534/g3.117.300427

Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B.W., Pruitt, K.D., Sherry, S.T., 2022. Database resources of the national center for biotechnology information. Nucleic Acids Res. 50, D20–D26. https://doi.org/10.1093/nar/gkab1112

Schirmer, M., D'Amore, R., Ijaz, U.Z., Hall, N., Quince, C., 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics 17, 125. https://doi.org/10.1186/s12859-016-0976-y

Schneider, C.A., Rasband, W.S., Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671–675. https://doi.org/10.1038/nmeth.2089

Schröder, B., Wrocklage, C., Pan, C., Jäger, R., Kösters, B., Schäfer, H., Elsässer, H.-P., Mann, M., Hasilik, A., 2007. Integral and Associated Lysosomal Membrane Proteins. Traffic 8, 1676–1686. https://doi.org/10.1111/j.1600-0854.2007.00643.x

Schroeder, J.W., Hirst, W.G., Szewczyk, G.A., Simmons, L.A., 2016. The Effect of Local Sequence Context on Mutational Bias of Genes Encoded on the Leading and Lagging Strands. Curr. Biol. CB 26, 692–697. https://doi.org/10.1016/j.cub.2016.01.016

Serbus, L.R., Landmann, F., Bray, W.M., White, P.M., Ruybal, J., Lokey, R.S., Debec, A., Sullivan, W., 2012. A Cell-Based Screen Reveals that the Albendazole Metabolite, Albendazole Sulfone, Targets Wolbachia. PLoS Pathog. 8, e1002922. https://doi.org/10.1371/journal.ppat.1002922

Serbus, L.R., White, P.M., Silva, J.P., Rabe, A., Teixeira, L., Albertson, R., Sullivan, W., 2015. The Impact of Host Diet on Wolbachia Titer in Drosophila. PLOS Pathog. 11, e1004777. https://doi.org/10.1371/journal.ppat.1004777

Shen, W., Ren, H., 2021. TaxonKit: A practical and efficient NCBI taxonomy toolkit. J. Genet. Genomics 48, 844–850. https://doi.org/10.1016/j.jgg.2021.03.006

Shen, Y., Xu, J., Zhi, S., Wu, W., Chen, Y., Zhang, Q., Zhou, Y., Deng, Z., Li, W., 2022. MIP From Legionella pneumophila Influences the Phagocytosis and Chemotaxis of RAW264.7 Macrophages by Regulating the lncRNA GAS5/miR-21/SOCS6 Axis. Front. Cell. Infect. Microbiol. 12.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. https://doi.org/10.1038/msb.2011.75

Skoog, E.C., Martin, M.E., Barrozo, R.M., Hansen, L.M., Cai, L.P., Lee, S.-J., Benoun, J.M., McSorley, S.J., Solnick, J.V., 2020. Maintenance of Type IV Secretion Function During Helicobacter pylori Infection in Mice. mBio 11, 10.1128/mbio.03147-20. https://doi.org/10.1128/mbio.03147-20

Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., Kosakovsky Pond, S.L., 2015. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Mol. Biol. Evol. 32, 1342–1353. https://doi.org/10.1093/molbev/msv022

Snyman, M., Huynh, T.V., Smith, M.T., Xu, S., 2021. The genome-wide rate and spectrum of EMS-induced heritable mutations in the microcrustacean Daphnia: on the prospect of forward genetics. Heredity 127, 535–545. https://doi.org/10.1038/s41437-021-00478-x

Starr, D.J., Cline, T.W., 2002. A host parasite interaction rescues Drosophila oogenesis defects. Nature 418, 76–79. https://doi.org/10.1038/nature00843

Steffen, P., Wu, J., Hariharan, S., Voss, H., Raghunath, V., Molloy, M.P., Schlüter, H., 2020. OmixLitMiner: A Bioinformatics Tool for Prioritizing Biological Leads from ‘Omics Data Using Literature Retrieval and Data Mining. Int. J. Mol. Sci. 21, 1374. https://doi.org/10.3390/ijms21041374

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028. https://doi.org/10.1038/nbt.3988

Sun, L., Luo, D., Ma, D., Zhao, Z., Chen, B., Shen, Z., Zhu, S., Chen, L., Chen, X., Yu, K., 2024. SciDFM: A Large Language Model with Mixture-of-Experts for Science. https://doi.org/10.48550/arXiv.2409.18412

Sung, W., Ackerman, M.S., Gout, J.-F., Miller, S.F., Williams, E., Foster, P.L., Lynch, M., 2015. Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation–Accumulation Experiments. Mol. Biol. Evol. 32, 1672–1683. https://doi.org/10.1093/molbev/msv055

Takai, S., Jin, D., 2022. Pathophysiological Role of Chymase-Activated Matrix Metalloproteinase-9. Biomedicines 10, 2499. https://doi.org/10.3390/biomedicines10102499

Talebi Bezmin Abadi, A., 2018. Diagnosis of Helicobacter pylori Using Invasive and Noninvasive Approaches. J. Pathog. 2018, 9064952. https://doi.org/10.1155/2018/9064952

Thomas, P.D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L., Mi, H., 2022. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. Publ. Protein Soc. 31, 8–22. https://doi.org/10.1002/pro.4218

Thorell, K., Muñoz-Ramírez, Z.Y., Wang, D., Sandoval-Motta, S., Boscolo Agostini, R., Ghirotto, S., Torres, R.C., Falush, D., Camargo, M.C., Rabkin, C.S., 2023. The Helicobacter pylori Genome Project: insights into H. pylori population structure from analysis of a worldwide collection of complete genomes. Nat. Commun. 14, 8184. https://doi.org/10.1038/s41467-023-43562-y

Torrisi, M., Pollastri, G., Le, Q., 2020. Deep learning methods in protein structure prediction. Comput. Struct. Biotechnol. J. 18, 1301–1310. https://doi.org/10.1016/j.csbj.2019.12.011

van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Gilchrist, C.L.M., Söding, J., Steinegger, M., 2022. Foldseek: fast and accurate protein structure search (preprint). Bioinformatics. https://doi.org/10.1101/2022.02.07.479398

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50, D439–D444. https://doi.org/10.1093/nar/gkab1061

Vooght, L.D., Ridder, K.D., Hussain, S., Stijlemans, B., Baetselier, P.D., Caljon, G., Abbeele, J.V.D., 2022. Targeting the tsetse-trypanosome interplay using genetically engineered Sodalis glossinidius. PLOS Pathog. 18, e1010376. https://doi.org/10.1371/journal.ppat.1010376

Waffenschmidt, S., Sieben, W., Jakubeit, T., Knelangen, M., Overesch, I., Bühn, S., Pieper, D., Skoetz, N., Hausner, E., 2023. Increasing the efficiency of study selection for systematic reviews using prioritization tools and a

single-screening approach. Syst. Rev. 12, 161. https://doi.org/10.1186/s13643-023-02334-x

Wan, M., Wang, X., Huang, C., Xu, D., Wang, Z., Zhou, Y., Zhu, Y., 2019. A bacterial effector deubiquitinase specifically hydrolyses linear ubiquitin chains to inhibit host inflammatory signalling. Nat. Microbiol. 4, 1282–1293. https://doi.org/10.1038/s41564-019-0454-1

Wang, Caihui, Liu, B., Chen, M., Ning, J., Lu, X., Wang, Chunde, 2022. Mutations in Growth-Related Genes Induced by EMS Treatment in Scallops. Front. Genet. 13. https://doi.org/10.3389/fgene.2022.879844

Wizenty, J., Tacke, F., Sigal, M., 2020. Responses of gastric epithelial stem cells and their niche to Helicobacter pylori infection. Ann. Transl. Med. 8, 568. https://doi.org/10.21037/atm.2020.02.178

Wong, K., Perpich, J.D., Kozlov, G., Cygler, M., Abu Kwaik, Y., Gehring, K., 2017. Structural Mimicry by a Bacterial F Box Effector Hijacks the Host Ubiquitin-Proteasome System. Structure 25, 376–383. https://doi.org/10.1016/j.str.2016.12.015

Xin, J., Mark, A., Afrasiabi, C., Tsueng, G., Juchler, M., Gopal, N., Stupp, G.S., Putman, T.E., Ainscough, B.J., Griffith, O.L., Torkamani, A., Whetzel, P.L., Mungall, C.J., Mooney, S.D., Su, A.I., Wu, C., 2016. High-performance web services for querying gene and variant annotation. Genome Biol. 17, 91. https://doi.org/10.1186/s13059-016-0953-9

Yako, Y.Y., Kruger, D., Smith, M., Brand, M., 2016. Cytokines as Biomarkers of Pancreatic Ductal Adenocarcinoma: A Systematic Review. PLoS ONE 11, e0154016. https://doi.org/10.1371/journal.pone.0154016

Yao, X.-F., Liu, Y., Li, Z., Jiang, G.-Q., Wang, W., Lu, H., Li, H., Lu, Z., Liu, C.-M., 2025. Non-Random Distribution of EMS-Induced Mutations Reveals Preference for Open Chromatin and Expressed Genes in Rice. Adv. Sci. 12, e10034. https://doi.org/10.1002/advs.202510034

Ye, Y., Godzik, A., 2004. Comparative Analysis of Protein Domain Organization. Genome Res. 14, 343–353. https://doi.org/10.1101/gr.1610504

Zhang, J., Zhang, W., Zou, D., Chen, G., Wan, T., Zhang, M., Cao, X., 2003. Cloning and functional characterization of GMPR2, a novel human guanosine monophosphate reductase, which promotes the monocytic differentiation of HL-60 leukemia cells. J. Cancer Res. Clin. Oncol. 129, 76–83. https://doi.org/10.1007/s00432-002-0413-7

Zhang, Y., 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309. https://doi.org/10.1093/nar/gki524