Moving Targets: Leveraging Geography to Understand the Evolutionary Dynamics of Malaria Vectors

by

Clara Rehmann

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Biology

Dissertation Committee:

Peter Ralph, Chair

Andrew Kern, Advisor

Mara Lawniczak, Core Member

Rori Rohlfs, Core Member

Mike Harms, Institutional Representative

University of Oregon

Spring 2026

© 2026 Clara Rehmann
This work is licensed under a Creative Commons
Attribution 4.0 International License (CC BY-NC-SA).



DISSERTATION ABSTRACT

Clara Rehmann

Doctor of Philosophy in Biology

Title: Moving Targets: Leveraging Geography to Understand the Evolutionary Dynamics of Malaria Vectors

Spatial genetic structure – the observation that individuals that live closer together are more genetically similar, termed "isolation by distance" – is ubiquitious in natural populations. Despite its prevalence and importance for understanding how genetic variation spreads and is shared through populations, much of our contemporary understanding of evolutionary processes is still informed by models that do not account for this spatial population structure. While this structure can obscure signatures of evolutionary processes as we understand them, it also provides another dimension of information that can inform how we analyze genetic variation and uncover the evolutionary histories that underlie it.

In this dissertation, I use a combination of continuous-space simulation and spatially-referenced genomic datasets in order to investigate and uncover the evolutionary dynamics of *Anopheles* malaria vectors and their *Plasmodium* parasites. With this work, I develop and introduce novel spatial methods for detecting evolutionary processes, including migration and positive selection. In application, these approaches uncover previously unobserved co-geographic structure shared by the malaria parasite and its vector, identify novel candidate alleles undergoing positive selection, and reveal convergent selection pressures acting on deeply-diverged vector species.

In Chapter II, I use machine learning to evaluate evidence for co-geography in the *Anopheles-Plasmodium* host-parasite system. Interrogating the spatial patterns of prediction errors from a machine learning method for geographic prediction from genetic data, I develop a method to reduce prediction bias when training on spatially-imbalanced datasets and then use this approach to confirm that errors from both *Anopheles gambiae* and *Plasmodium falciparum* are a biological signal of shared dispersal patterns. In Chapter III, I use continuous-space simulation to uncover a novel spatial signal of selective sweeps, and use this observation to develop an approach for identifying candidate sweeping alleles in spatially-referenced genetic data. Applying this method to *Anopheles gambiae* populations in West Africa, I identify several novel alleles that appear to be under positive selection, including variants with functional relevance and at loci associated with insecticide resistance. Finally, in Chapter IV, I use a comparative population genetic approach to investigate the evolutionary forces shaping and mainintaining patterns of high polymorphism in *Anopheles* vectors. This work reveals common selection pressures – primarily acting on genes relevant to immunity and insecticide resistance – that both maintain functional variation within and drive adaptive divergence between species; in uncovering these patterns, I identify spatially convergent selection on proteins involved in insecticide resistance across two major African *Anopheles* vectors. Altogether, this dissertation demonstrates the potential of leveraging spatial population genetic structure in order to inform our understanding of evolutionary dynamics, providing insight into the evolutionary history of *Anopheles* malaria vectors and their *Plasmodium* parasites while introducing novel methods to identify these processes.

This disseration includes previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Clara Rehmann

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
UCLA, Los Angeles, CA, USA

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2026, University of Oregon
Bachelor of Science, Biology, 2020, University of Oregon

AREAS OF SPECIAL INTEREST:

Evolutionary biology
Population genetics
Vector-borne disease

PUBLICATIONS:

Rivera-Colón, A. G., Rehmann, C. T., & Kern, A. D. (2026). MKado: a toolkit for McDonald-Kreitman tests of natural selection. Preprint: https://www.biorxiv.org/content/10.64898/2026.03.02.709122v1.

Gower, G., Pope, N. S., ... , Rehmann, C. T., ... , & Kern, A. D. (2025). Accessible, Realistic Genome Simulation with Selection Using `stdpopsim`. *Molecular Biology and Evolution*, 42(11), msaf236.

Rehmann, C. T., Small, S. T., Ralph, P. L., & Kern, A. D. (2025). Sweeps in Space: Leveraging Geographic Data to Identify Beneficial Alleles in *Anopheles gambiae*. *Molecular Biology and Evolution*, 42(6), msaf141.

Chevy, E. T., Min, J., ... , Rehmann, C. T., ... , & Ralph, P. L. (2025). Population Genetics Meets Ecology: A Guide to Individual-Based Simulations in Continuous Landscapes. *Ecology and Evolution*, 15(4), e71098.

Rehmann, C. T., Ralph, P. L., & Kern, A. D. (2024). Evaluating evidence for co-geography in the *Anopheles–Plasmodium* host–parasite system. *G3: Genes, Genomes, Genetics*, 14(3), jkae008.

ACKNOWLEDGEMENTS

Graduate school is the most fun I've had since middle school, and I'm incredibly lucky that I've spent the last five years working, learning, and dancing alongside so many brilliant, kind, and supportive people. You all have been instrumental in my development as a scientist and as a person, and I am so grateful to have had your company on this journey.

I have had incredible mentors. Andy - thank you for convincing me to go to graduate school in the first place. I wouldn't be here or really be myself without your mentorship, support, and friendship - you were right that grad school is like science summer camp, and I can't wait to hire you as the Old Man in my lab one day. Peter - I am eternally grateful for your presence and your patience; you have helped me grow so much as a scientist. Thank you for pushing me to think more deeply and openly. I would also like to thank the rest of my committee: Mike Harms, Mara Lawniczak, and Rori Rohlfs. It has been an incredible privelege to have scientists I admire so much take the time to entertain my research, and I am so grateful for your feedback and support throughout this process.

I've also had incredible lab mates. Y'all have been not just coworkers but close friends, and I am so grateful for your support, insights, and for coming with me to '80s night. CJ Battey onboarded me into the world of popgen, Murillo Rodrigues was my favorite roommate, Nate Pope taught me what a z-score is for longer than he should have needed to, and Gilia Patterson and Gabby Coffing always kept me climbing. Silas Tittes, (M')Jordan Anderson, and Amjad Dabi have made me laugh so hard and Thomas Forest and Angel Rivera-Colon have reignited my passion for music. I especially want to thank Jordan Rodriguez for

being my partner in crime through many conferences and for keeping me grounded; I'm so lucky that you joined the lab. I'd also like to thank Scott Small for being the Mosquito Man and never being quite sure. Thank you to the rest of the co lab that has passed through in my time here: Jeff Adrion, Victoria Caudill, Saurabh Belsare, Chris Smith, Jiseon Min, Georgia Tsambos, Bruce Edelman, Anastasia Teterina, Ananya Kapoor, Kevin Korfmann, and everyone else. This group has been the best part of my grad school experience, and a family really can be a lab traveling together for an international conference.

I've been supported by my lab and by a broader community at the University of Oregon. I especially want to thank GrEBES, the IE2 graduate student group, for being a source of camaraderie and support. I'd also like to thank the friends who guided me through this process - Caitlin Smith, Rose Al-Saadi, Savanah Bird, Hope Healey, and Sabrina Mostoufi - as well as the people who have heped me become a guide myself. Rob Porch, Aubrey Miller, Masha Korchagina, and many others have been incredible friends in this journey and helped me grow as a scientist and as a person. I especially want to thank Kendal Tinney for being my grad school mirror and always letting me swing by the Barber Lab.

Before I started my PhD, I had (many) incredible mentors: Matt Barber, Nelson Ting, Ken-ichi Noma, and Eric Selker were kind enough to welcome me into their labs and I am so grateful for their belief in me as I was starting out as a scientist and figuring out what my interests were. I also want to thank my first lab at UCLA, especially my PI Beth Lazazzera and my mentor Makayla Morton, who told me I'd want to go to grad school when I told her I'd be happy pipetting.

I want to thank my family and close friends for their love and support through all of this. Jordan and Kendal - you've been named twice and I can't

begin to express how grateful I am to have you as friends. Thank you for always being there and for making me feel like myself. I was lucky enough to be at UO while my little brother Drew was here too, and grad school would not have been the same without him and our weekly Winco trips. I'm also grateful to my little sister, Julia, who is always there for me and inspires me with her pursuits. I would like to thank my parents, who've supported my curiosity and interest in biology whether it was dinosaurs, *E. coli*, or mosquitoes, and who have patiently listened to many practice talks. Thank you to my mom, Mary, for believing in what I could do and not letting me back away from that; thank you for supporting me through the ups and downs of this process and for always picking up the phone. Thank you to my dad, Dan, for being my first best friend and for showing me how to be myself; thank you for your curiosity and openness and for making sure I still had fun in grad school.

Finally, I want to express my gratitude to my partner, Erin, and our creatures Aeg and Zucchini, who are my home and my heart. Erin - thank you for being crazy enough to date a grad student. Your understanding, support, and love have changed the trajectory of my life and I am so grateful to have met you and to have you by my side through all of this. Thank you for keeping me laughing and for listening to me talk about mosquito genomes and for making every day an adventure. I can't wait to see what the next chapter brings for us.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 23

II. EVALUATING EVIDENCE FOR CO-GEOGRAPHY IN THE *ANOPHELES-PLASMODIUM* HOST-PARASITE SYSTEM 30

2.1. Introduction 30

2.2. Materials and Methods 34

2.2.1. Empirical data 34

2.2.2. Spatial simulations 35

2.2.3. Biased training sets 35

2.2.4. Modifications to training data 36

2.2.4.1. Downsampling 36

2.2.4.2. Sample weighting 37

2.2.5. *A. gambiae* simulations 37

2.3. Results 39

2.3.1. Errors in geographic predictions of *A. gambiae* and *P. falciparum* are correlated 39

2.3.2. Spatial bias in training data leads to biased geographic predictions 42

2.3.3. Spatial distribution of training data does not explain Anopheles residuals 43

2.3.4. Accounting for spatially biased data in geographic prediction 45

2.3.5. Correlation between *A. gambiae* and *P. falciparum* prediction errors is not due to bias 48

2.4. Discussion 49

Chapter Page

2.5. Bridge . . . 60

III. SWEEPS IN SPACE: LEVERAGING GEOGRAPHIC DATA TO IDENTIFY BENEFICIAL ALLELES IN *ANOPHELES GAMBIAE* . . . 61

3.1. Introduction . . . 61

3.2. Methods . . . 65

3.2.1. Spatial simulation . . . 65

3.2.2. Power analysis . . . 66

3.2.3. *Anopheles gambiae* analysis . . . 68

3.2.3.1. Data processing . . . 68

3.2.3.2. Spatial genome scan . . . 70

3.3. Results . . . 72

3.3.1. A novel spatial signal of selection . . . 72

3.3.2. Power to detect selection using space . . . 75

3.3.3. Spatial signal identifies sweeping alleles in *Anopheles gambiae* . . . 77

3.3.3.1. Cytochrome P450 *CYP4H27* . . . 83

3.3.3.2. CUB domain-containing protein . . . 85

3.3.3.3. Gustatory receptors . . . 85

3.4. Discussion . . . 87

3.5. Bridge . . . 95

IV. CONVERGENT SELECTION ON NONSYNONYMOUS VARIATION ACROSS DEEPLY DIVERGED *ANOPHELES* VECTORS . . . 96

4.1. Introduction . . . 96

4.2. Methods . . . 99

4.2.1. Genomic data & processing . . . 99

Chapter Page

4.2.2. Population genetics methods 100

4.2.2.1. Within species 100

4.2.2.2. Between species 101

4.2.2.3. F_{ST} correlation 102

4.3. Results 103

4.3.1. Nonsynonymous polymorphism is abundant and enriched for immune, sensory, and metabolic functions 103

4.3.2. Adaptive divergence targets overlapping functional categories across lineages 108

4.3.2.1. McDonald Kreitman test 108

4.3.2.2. Phylogenetic testing 114

4.3.3. Convergent nonsynonymous spatial differentiation between *An. gambiae* and *funestus* 115

4.4. Discussion 119

V. CONCLUSION 130

APPENDICES

A. SUPPLEMENTAL MATERIAL FOR CHAPTER II: EVALUATING EVIDENCE FOR CO-GEOGRAPHY IN THE *ANOPHELES-PLASMODIUM* HOST-PARASITE SYSTEM 134

B. SUPPLEMENTAL MATERIAL FOR CHAPTER III: SWEEPS IN SPACE: LEVERAGING GEOGRAPHIC DATA TO IDENTIFY BENEFICIAL ALLELES IN *ANOPHELES GAMBIAE* 143

B.1. Approximation to Fisher's wave 143

Chapter Page

C. SUPPLEMENTAL MATERIAL FOR CHAPTER IV: CONVERGENT SELECTION ON NONSYNONYMOUS VARIATION ACROSS DEEPLY DIVERGED *ANOPHELES* VECTORS 148

C.1. GO enrichments for π_N 149

C.1.1. *An. gambiae* 150

C.1.2. An. funestus 150

C.1.3. An. minimus 150

C.2. GO enrichments for Tajima's D_N 151

C.2.1. *An. gambiae* 151

C.2.2. *An. funestus* 152

C.2.3. *An. minimus* 152

C.3. *An. gambiae* and *funestus* $F_{ST_{NS}}$ correlation 153

LIST OF FIGURES

Figure Page

2.1. Predicted locations of *Anopheles gambiae* and *Plasmodium falciparum* samples. (A) Geographic centroids of predictions for each unique sampling site. True sampling locations are connected to their mean predicted location; the size of predicted locations represents error in km. (B, C) Distribution of error magnitudes for *A. gambiae* and *P. falciparum* predictions. (D, E, F) Vector, angle, and magnitude correlations between spatially paired and randomly paired *A. gambiae* and *P. falciparum* residuals. 55

2.2. The effect of training set bias on `Locator` residuals. (A) Training sets sampled with increased bias towards the right side of the landscape. (B) Mean per-run prediction error along the x axis as a function of training set bias, (C) Mean per-run prediction error along the y axis, (D) Mean per-run error magnitude. 56

2.3. Predicted locations of empirical and simulated *Anopheles gambiae* samples. (A) Geographic centroids of predictions for each unique sampling site. True sampling locations are connected to their mean predicted location; the size of predicted locations represents error in km. (B) Distribution of error magnitudes. (C, D, E) correlation coefficients between simulated Anopheles and empirical Plasmodium predictions for the residual vectors, residual angles, and residual magnitude respectively 57

2.4. The effect of training set bias using sampling and training modifications. (A) Mean per-run prediction error along the x axis as a function of training set bias, (B) Mean per-run prediction error along the y axis, (C) Mean per-run error magnitude. (D, E, F) Example predictions at bias = 0.9 for unmodified, downsampled, and weighted training sets, respectively. True sample locations (black points) are connected to their predicted location (colored points). 58

Figure Page

2.5. Predicted locations of *Anopheles gambiae* and *Plasmodium falciparum* samples after training using sample weights. (A) Geographic centroids of predictions for each unique sampling site. True sampling locations are connected to their mean predicted location; the size of predicted locations represents error in km. (B, C) Distribution of error magnitudes for both unweighted and weighted *A. gambiae* and *P. falciparum* predictions. (D, E, F) Vector, angle, and magnitude correlations between spatially paired or randomly sampled *A. gambiae* and *P. falciparum* weighted residuals. . . 59

3.1. A conceptual model of how positive selection impacts an allele's spread across a one-dimensional landscape. Panels A, B, and C show a time series of a beneficial allele (in green) and a neutral allele (in blue) arising at position 50 and spreading through space. The frequency of each allele at a given spatial position is indicated along the y axis. 73

3.2. Selective sweeps in continuous space. (A) Example of simulated allele with selection coefficient 0.1 sweeping through a population with a dispersal distance (σ) of 0.40. Individuals carrying the sweeping allele are colored in green. (B) Joint frequency-area trajectories of sweeping alleles at dispersal distances 0.40, 1.25, and 4.0. Line color represents the strength of selection; allele area is measured as the area of the polygon encompassing allele carriers. 75

3.3. Power to detect selective sweeps using continuous space. (A) Example genome-wide joint distribution between frequency and area of all variants 195 time steps after the introduction of an allele with a selection coefficient of $s = 0.1$ in a population with a mean per-generation dispersal distance of 0.40. The black line represents the tenth percentile cutoff for SF outliers; the sweeping allele is highlighted in green. (B) Example genome-wide windowed analysis of the ratio of SF outliers to non-outliers at the same time point in the same simulation. The dotted line represents the tenth percentile cutoff for WSF outliers; the green triangle highlights the locus of the sweeping allele. (C) Probability of detecting the sweeping allele as a SF outlier as the strictness of the outlier cutoff increases along the x axis. Panels represent σ values 0.40, 1.25, and 4.0; line color represents the strength of selection. The dashed gray line represents the proportion of the genome identified as SF outliers. (D) Probability of detecting the sweeping allele as a WSF outlier in the same simulations; the dashed gray line represents the proportion of the genome identified as WSF outliers. 78

3.4. Joint distribution between frequency and area of all variants within the *Anopheles gambiae* genome. Each point represents a genomic variant; the point's color indicates its annotation. Outlier variants highlighted in vignettes are colored and outlined. Black lines represent SF outlier cutoffs for each of the major inversions and the non-inverted portion of the genome, respectively. INSET: Sampling locations of *An. gambiae* samples. The size of the points represents the number of samples at each location. 81

3.5. Genome-wide windowed analysis of the ratio of SF outliers to non-outliers. Flat horizontal lines represent WSF cutoffs for each of the major inversions and the non-inverted portion of the genome, respectively. Known loci associated with insecticide resistance and outlier variants highlighted in vignettes are labeled. 83

Figure Page

3.6. Impacts of outlier variants highlighted in vignettes. (A) Folding impacts of cytochrome P450 variant CYP4H27:Y274H. The left and right panels show position 274 and local structure for the ancestral tyrosine and the derived histidine, respectively. Relevant amino acids (S324, S326, E327) and hydrogen bonds are shown and the new hydrogen bond is highlighted in yellow. (B) Intronic outlier SNPs at the CUB domain-containing protein locus. Variants are shown at their genomic locations and colored by area; their y-axis position represents their frequency. Coding sequences and intronic segments for genes at the locus are annotated below; arrows represent the direction of transcription. (C) Structural locations of nonsynonymous variants in gustatory receptors. GRs are oriented with their extracellular domain upward and intracellular domain downward; chains are colored from light green at the N terminus to dark green at the C terminus and nonsynonymous mutations are highlighted in yellow. 88

4.1. (A) Geographic distribution and (B) phylogeny of *Anopheles gambiae*, *funestus*, and *minimus*. 99

4.2. Distribution of genic (lighter color) and nonsynonymous (darker color) π, Tajima's D, and F_{ST} values within *An. gambiae*, *funestus*, and *minimus*. 105

4.3. Significant GO enrichments for nonsynonymous π outliers in *An. gambiae*, *funestus*, and *minimus*. 108

4.4. Results of the asymptotic McDonald-Kreitman test for all one-to-one orthologs across *An. gambiae*, *An. funestus*, and *An. minimus*. Light blue points represent genes with significant evidence of adaptive divergence after Boneferroni correction ($p < 0.05$), while dark blue points represent genes with a significant excess of nonsynonymous polymorphism. 109

4.5. (A) Frequency of spatially-constrained nonsynonymous Cyp6p5 mutations in *An. gambiae* (left) and *funestus* (right). (B) Position of spatially-constrained nonsynonymous mutations on the Cyp6p5 protein in *An. gambiae* (left) and *funestus* (right). Arrow indicates channel entrance. 118

Figure Page

4.6. (A) Frequency of spatially-constrained nonsynonymous Cyp6p2 mutations in *An. gambiae* (left) and *funestus* (right). (B) Position of spatially-constrained nonsynonymous mutations on the Cyp6p2 protein in *An. gambiae* (left) and *funestus* (right). Arrow indicates channel entrance. 119

A.1. Predicted locations of all *A. gambiae* and *P. falciparum* samples. A sample's true location is connected to the geographic centroid of its genome-wide windowed predictions. 137

A.2. The magnitude of summed *A. gambiae* and *P. falciparum* residual vectors plotted against the distance between true sample locations. For both unweighted (A) and weighted (B) analyses, distance between samples is negatively correlated with the magnitude of summed residual vectors. 138

A.3. Spatially paired *A. gambiae* and *P. falciparum* principal components. (A) *A. gambiae* sampling locations and (B) principal components, with each sample colored by its spatial location. (C) *P. falciparum* sampling locations and (B) principal components, with each sample colored by its spatial location. (E) Relationship between the first principal component of spatially-paired *A. gambiae* and *P. falciparum* samples, with each pair colored by their spatial location. (F) Relationship between the second principal component of spatially-paired samples. 139

A.4. The effect of training set bias and size on `Locator` prediction error at a dispersal rate of $\sigma = 1.0$. (A) Spatial spread of per-individual windowed predictions as a function of local training set density at different levels of training set bias. (B) Mean per-run prediction error along the x axis as a function of training set bias, (C) mean per-run prediction error along the y axis, (D) mean per-run error. 140

A.5. Results of grid search for λ and bandwidth parameters on $bias = 0.9, \sigma = 1.0$ training sets. (A) Mean x-axis error, (B) mean y-axis error, and (C) mean error for all tested combinations ofλ and bandwidth. 141

Figure Page

A.6. Bandwidth and λ parameters alter the distribution of sample weights. (A) Distribution of sample weights assigned to a given training set holding bandwidth constant and varying λ. (B) Distribution holding λ constant and varying bandwidth. (C, D, E) Biased training sets colored by their sample weights as λ increases. 141

A.7. Predictions from biased training data using `SPASIBA` and `Locator`. Simulations had a per-generation dispersal rate of 1.0 units. (A) Mean x-axis error as a function of training set bias and (B) mean prediction error as a function of training set bias. (C, D) Example `Locator` and `SPASIBA` at $bias = 0.5$ and (E, F) $bias = 0.9$. True sample locations (black points) are connected to their predicted location (colored points). 142

B.1. Approximation of a positively-selected allele spreading according to a Fisher wave across two-dimensional space. (A) Top-down view of the extent of the sweeping allele (r_{edge}) and the extent of the allele's fixation (r_{fixed}). (B) Side view of the method used to approximate the frequency of the sweeping allele as the volume of the conical frustum defined by r_{edge} and r_{fixed} out of the volume of the landscape defined by x, y, and z. (C) Frequency of the positively-selected allele. (D) Joint distribution between frequency and area of the positively selected allele. In both plots, lines are colored by selection coefficient (S). 145

B.2. Probability of detecting a sweeping allele as a spatial outlier over the course of a sweep across dispersal distances (x axis) and cutoff percentiles (y axis). Lines are colored by selection coefficient. 146

B.3. Joint frequency-area distribution for *An. gambiae* SNPs within the 2La (A), 2Rb (B), and 2Rc (C) inversions. SNPs are colored according to their SNPEff annotation; black lines represent the 10th percentile cutoff. 147

C.1. Relationship between π_N and π_G in *An. gambiae* (A), *funestus* (B), and *minimus* (C). 148

C.2. Relationship between D_N and D_G in *An. gambiae* (A), *funestus* (B), and *minimus* (C). 148

Figure Page

C.3. Relationship between $F_{ST_{NS}}$ and $F_{G_{NS}}$ in *An. gambiae* (A), *funestus* (B), and *minimus* (C). 149

C.4. Significantly correlated $F_{ST_{NS}}$ vectors for three *An. gambiae* (dark blue) and *funestus* (light blue) orthologs. Vectors for each species are depicted via a line connecting two points; sampling locations are depicted in points. 153

C.5. Restricted spatial distribution of high-frequency nonsynonymous alleles in *An. funestus* gene AFUN2_013651. 154

LIST OF TABLES

Table Page

2.1. Empirical estimates and scaled simulation parameters for *Anopheles* simulations 38

4.1. Summary of genic and nonsynonymous population genetic statistics for *An. gambiae*, *funestus*, and *minimus*. Tajima's D values are reported as the median across genes; other values are reported as the mean across genes. 104

4.2. Top 10 genes with significant evidence of adaptive divergence in *An. gambiae* 110

4.3. Top 10 genes with significant evidence of adaptive divergence in *An. funestus* 112

4.4. Top 10 genes with significant evidence of adaptive divergence in *An. minimus* 113

C.1. Significant Biological Process GO term enrichments for *An. gambiae* π_N outliers. 149

C.2. Significant Biological Process GO term enrichments for *An. funestus* π_N outliers. 150

C.3. Significant Biological Process GO term enrichments for *An. minimus* π_N outliers. 150

C.4. Significant Biological Process GO term enrichments for *An. gambiae* D_N outliers. 151

C.5. Significant Biological Process GO term enrichments for *An. funestus* D_N outliers. 152

C.6. Significant Biological Process GO term enrichments for *An. minimus* D_N outliers. 152

CHAPTER I

INTRODUCTION

The study of evolutionary histories and processes is essential to our understanding of the natural world and can help inform what will happen in the future. Knowing what happened in the past and how it happened, however, requires inferring these processes from what we observe in the present. To understand past evolutionary events, population genetics relies on the analysis of patterns of genetic diversity in natural populations. Different evolutionary processes imprint different signatures on the genomes of populations: Forces such as mutation, drift, and selection, as well as events such as population size changes or migration, will all drastically impact the patterns of genetic diversity in natural populations. For example, a beneficial mutation that rises to fixation will leave a characteristic pattern of reduced diversity in the genome in a process known as a "selective sweep," (Maynard Smith & Haigh, 1974), or migration between two separate populations will lead to shared regions of the genome that have common patterns of variation between the populations. The goal of population genetics is to identify and leverage these genetic signatures in order to infer the evolutionary history they represent.

In natural populations, these patterns of genetic diversity are also distributed across geographic space, with each individual in a population - and their genome - occupying a location on the landscape. As evolutionary time plays forward, these individuals interact, reproduce, and disperse locally, giving rise to spatial genetic structure within populations: in short, individuals that are closer spatially tend to also be closer genetically. This phenomenon, termed "isolation by distance" by Sewall Wright (Wright, 1946), is ubiquitous in natural populations and has been well-documented across a wide variety of taxa (Schlenke & Begun, 2004; Xue et al., 2021b).

However, despite its prevalence in natural populations, much of our contemporary understanding of evolutionary processes is still informed by models that do not account for this spatial population structure. While the simplifying assumption that all individuals in a population are equally likely to mate regardless of their spatial location makes evolutionary models more tractable, it obscures a key component of the population's evolutionary history and can confound inference of evolutionary processes from genetic data (Bradburd & Ralph, 2019). Further, understanding the relationship between spatial structure and observed patterns of genetic diversity adds another dimension of information that can be leveraged to uncover the evolutionary forces acting on populations. Spatial patterns of genetic variation have been successfully used to infer migration rates, estimate population sizes, and detect selection (Muktupavela et al., 2022; Patterson et al., 2026; C. C. R. Smith et al., 2024); however, the finer-scale patterns of spatial genetic structure now observable in genomic datasets remain unexplored in many contexts.

This dissertation focuses on leveraging spatial population structure to uncover and describe evolutionary processes in the malaria mosquito vector, *Anopheles*. Malaria is an ancient human disease that remains a persistent public health threat despite recent advances in treatments, vaccines, and vector control efforts. Both the parasite and its vector rapidly evolve to counter human interventions, making understanding their evolutionary histories and trajectories crucial to effective disease prevention strategies. Vector control efforts in particular are critical for containing the spread of malaria, and extensive genomic surveillance of *Anopheles* species now allows us to not only understand the population structure of these mosquitoes, but also to identify when and how they are adapting to human interventions such as insecticides (Anopheles gambiae 1000 Genomes Consortium, 2017; The Anopheles gambiae 1000

Genomes Consortium et al., 2020). Using extensive spatial population genomic datasets of *Anopheles* mosquitoes, I investigate patterns of dispersal, selection, polymorphism, and divergence in this system, focusing on how incorporating spatial information into genetic analyses helps uncover processes that would otherwise go undetected.

In Chapter II, I combine a machine learning method for prediction of location from genotype data, called `Locator`, with forward-in-time, continuous-space, individual-based simulations to investigate and identify evidence for common patterns of dispersal between *Anopheles gambiae* and *Plasmodium falciparum*. When using `Locator` to predict locations of *An. gambiae* and *P. falciparum* from genotypic data, we observed similar patterns of prediction errors, which I discovered were significantly correlated between host and parasite across space. Because errors in predictions obtained using `Locator` tend to reflect the locations of a sample's ancestors, we hypothesized that the correlation in errors between *An. gambiae* and *P. falciparum* may be due to common migration patterns - however, the observed correlation could also be due to spatial bias in the distribution of sampling locations. After confirming that spatial bias in sampling locations can indeed lead to biased predictions, I developed a sample weighting method to correct for this bias and reduce the influence of training set distributions on predictions. Significant correlations between prediction errors persist when this method is applied to *An. gambiae* and *P. falciparum*, providing further evidence for these predictions reflecting shared patterns of dispersal across the landscape, particularly in West Africa. This observation of common dispersal patterns between host and parasite across short evolutionary timescales is novel and has important implications for understanding patterns of gene flow and the spread of adaptive alleles in both species. This work was published in

G3 — Genes — Genomes — Genetics in 2024 and is co-authored by Peter L. Ralph and Andrew D. Kern.

Investigating patterns of dispersal in *Anopheles* and *Plasmodium*, I became interested in how beneficial variants were spreading through these spatially-structured populations. Positive selection plays a fundamental role in shaping variation in natural populations, and the rapid spread of beneficial alleles is responsible for many instances of adaptation. This evolutionary process is particularly important to understand in the context of malaria control efforts, as both the parasite and its vector are rapidly evolving to counter human interventions such as insecticides and antimalarial drugs. While the impact of spatial population structure on the spread of beneficial mutations has been understood theoretically (R. A. Fisher, 1937), much of our contemporary understanding of spatial selective sweeps is in regards to how spatial structure obscures the signatures of selection that we look for in genetic datasets (Chotai et al., 2024; Min et al., 2022a). In Chapter III, I use continuous-space simulation to identify a spatial signature of positive selection, then use this observation to develop a method to identify sweeping alleles in spatially-referenced genetic datasets. In simulations of selective sweeps, I observed that positively-selected alleles tend to occupy less of the landscape than neutral alleles at the same frequency. While counterintuitive, this observation can be explained by the fact that neutral alleles, if not lost, will spread through a population - and across the landscape - without increasing in frequency. Positively-selected alleles, on the other hand, will inherently increase in frequency as they sweep through the population, and this increase in frequency outpaces their spatial spread. Using this observation, I develop an approach to detect positively-selected alleles as outliers to the genome-wide spatial-frequency distribution of genetic variants. Applying

this approach to *Anopheles gambiae*, we identify multiple positively selected and functionally impactful variants segregating the population, including variants within known insecticide-resistant loci, potentially novel sources of insecticide resistance, and other potentially selected variants associated with immune response, sensory perception, and development. This work provides a novel method for identifying positively-selected alleles in spatially-structured populations and demonstrates the potential of spatial information to inform, rather than obscure, our understanding of evolutionary processes. This work was published in *Molecular Biology and Evolution* in 2025 and is co-authored by Scott T. Small, Peter L. Ralph, and Andrew D. Kern.

Anopheles gambiae exists as part of a broader genus of malaria vectors, and in Chapter IV, I use a comparative approach to investigate the evolutionary forces driving patterns of polymorphism and divergence in *Anopheles* malaria vectors. Phylogenetic analysis of *Anopheles* vectors has highlighted their rapid rate of genome evolution, while population genomic analyses of *Anopheles* species have revealed that these mosquitoes are incredibly polymorphic and segregating many potentially adaptive variants. Focusing on nonsynonymous polymorphism and divergence in two major African vectors (*Anopheles gambiae* and *funestus*) and one major Asian vector (*Anopheles minimus*), I explore what genes and biological processes are evolving under directional selection in these species and identify how patterns of polymorphism reflect similar evolutionary pressures. I find that genes involved in sensory perception, immune response, and insecticide resistance are both rapidly evolving under directional selection in all three species and segregating high amounts of nonsynonymous polymorphism, suggesting that the variation maintained by *Anopheles* populations may be important for adaptation to rapidly-changing environments and selection pressures. Using patterns of spatial differentiation in

nonsynonymous polymorphism, I also identify candidate genes that may be adapting to local selection pressures in both *Anopheles gambiae* and *Anopheles funestus*, including mutations with similar molecular impacts and spatial distributions in two cytochrome P450 enzymes, which are known to be involved in metabolic insecticide resistance. My work in this chapter provides insight into the evolutionary forces shaping patterns of polymorphism and divergence in *Anopheles* malaria vectors and highlights the importance of considering spatial population structure when investigating these processes. This study is in preparation for submission to *PLOS Genetics* and is co-authored by Andrew D. Kern.

Overall, my dissertation explores how spatial population structure can be leveraged to better understand evolutionary processes in *Anopheles* malaria vectors, with important implications for understanding how these species are adapting in response to human interventions. Chapter II provides some of the first evidence for co-geography between *Anopheles* and *Plasmodium*, suggesting that these species are dispersing in similar trajectories and that their evolutionary histories are intertwined across space. In Chapter III, I identify a novel spatial signature of positive selection and use this observation to develop and apply a method for detecting positively-selected alleles in *Anopheles gambiae*, identifying multiple novel candidate insecticide resistance alleles. Finally, Chapter IV uses a comparative approach to identify and describe the evolutionary forces driving divergence and shaping patterns of polymorphism across three major *Anopheles* malaria vectors, highlighting how variable selection pressures across space may be maintaining important variation in these species. Together, this work demonstrates the potential of spatial information to inform our understanding of evolutionary processes in natural populations and

provides insight into the evolutionary dynamics of malaria parasites and their mosquito vectors.

CHAPTER II

EVALUATING EVIDENCE FOR CO-GEOGRAPHY IN THE *ANOPHELES-PLASMODIUM* HOST-PARASITE SYSTEM

This chapter was published in the journal *G3 — Genes — Genomes — Genetics* in 2024. Andrew D. Kern and Peter L. Ralph are co-authors on this work. Co-authors and I conceptualized the study together and I performed all analyses; I wrote the manuscript with A.D.K and editorial assistance from P.L.R. The citation for this publication is as folllows:

Rehmann, C. T., Ralph, P. L., & Kern, A. D. (2024). Evaluating evidence for co-geography in the *Anopheles–Plasmodium* host–parasite system. *G3: Genes, Genomes, Genetics*, *14*(3), jkae008

2.1 Introduction

A nearly ubiquitous observation in biology is that individuals within a species that live closer together in space are more closely related genetically than individuals that live further apart. Such isolation by distance, as it was referred to by Wright, 1943b, and its fundamental determinant – the spatially limited dispersal of individuals across a landscape – can have profound impacts on patterns of genetic variation within and between populations (Bradburd & Ralph, 2019; Whitlock & McCauley, 1999), genetic association studies (Battey et al., 2020c; Mathieson & McVean, 2012), and the potential for local adaptation (N. Barton, 2001). As befits the central importance of understanding dispersal, over many decades much work has gone in to trying to estimate migration of individuals between local demes or across continuous space (Marcus et al., 2021; Rousset, 1997).

Combining spatial and genetic data can reveal a wealth of information about biological and ecological processes: for example, these data can be used to predict a genetic sample's location (Finch et al., 2020; Guillot et al., 2015; Wasser et al., 2004), estimate dispersal and migration patterns (Marcus et al., 2021; C. C. Smith et al., 2023), or describe major features of spatial genetic structure (Bradburd et al., 2013). Indeed with the increasing availability of geo-referenced genotypic data as well as increasingly sophisticated methods of spatial population genetics simulation (Battey et al., 2020c), there is growing opportunity to explore the potential of identifying biological processes through patterns of isolation by distance. Recently, Battey et al., 2020a introduced `Locator`, a state-of-the-art, deep learning method for predicting the geographic origin of a genetic sample. `Locator` is able to achieve extremely accurate prediction on large, publicly available datasets. Further, Battey et al., 2020a's analysis indicated that errors in geographic prediction of locations (which we refer to as *residuals*) can be informative of those individuals' geographic ancestry: for example, when applied to human data, predicted locations of South African Bantu individuals trended towards west Africa, the historic origin of the Bantu Expansion.

More generally, we would expect various biological processes to affect `Locator`'s residuals, and so may be able to examine residuals to infer the processes that give rise to these patterns. For instance, two processes we expect to affect residuals are heterogeneous dispersal and range expansion. If dispersal is heterogeneous across the landscape, with high dispersal in some regions and low dispersal in others, regions with higher levels of gene flow would have less accurate geographic prediction. In the case of range expansions, one would expect the expanding lineage to genetically resemble individuals from the original range - a prediction we see reflected in the South African Bantu example. In both cases, though we may be unable to identify

individual instances of dispersal, errors in `Locator`'s predictions can be interpreted to provide clues about underlying population processes.

Battey et al., 2020a applied `Locator` to data from the malaria vector *Anopheles gambiae* (Ag1000G; N. Anopheles gambiae 1000 Genomes Consortium, 2017) and its associated parasite *Plasmodium falciparum* (Pf7k, Pearson et al., 2019), and were able to predict geographic origin of individuals to within a median error of 16.9 km and 85 km, respectively. We were interested in exploring if any structure might be present in the residuals from `Locator`'s predictions. In particular, close examination of patterns in prediction error between species on the African continent suggested to us that perhaps residuals between spatially proximate individuals from the *Anopheles gambiae* and *Plasmodium falciparum* datasets might be correlated. Given that we expect `Locator`'s residuals to be affected by heterogeneous gene flow, these errors may reflect shared non-uniform migration patterns between host and parasite, for instance a broad corridor of dispersal throughout West Africa but more restricted gene flow in Central and Eastern parts of the continent.

Seen from the standpoint of host-parasite ecology, such correlations in the residuals of geographic predictions may not be surprising. If parasites depend on hosts for dispersal, then we might indeed expect a close correspondence between the spatial genetic structure of these interacting species. Such patterns have been observed before: for instance, Small et al., 2019 found a tight correspondence between divergence times between geographical locations of the filarial nematode parasite *Wuchereria bancrofti* and known instances of human migration, thus linking parasite divergence to (human) host dispersal. At the limit of low parasite dispersal and/or considered over longer evolutionary time scales, these dynamics can translate into co-speciation, where patterns of splitting in host and parasite closely mirror one another

(e.g., Clayton et al., 2003; Jousselin et al., 2008; Moran & Baumann, 1994), and many phylogenetic-based methods have been developed to identify patterns of host-parasite cospeciation (e.g., Hafner & Nadler, 1988; Huelsenbeck et al., 2000; Huelsenbeck et al., 1997) on a macroevolutionary scale.

While residuals from `Locator`'s predictions may reflect geographic patterns of gene flow - in this case common dispersal patterns between *A. gambiae* and *P. falciparum* populations - the data used to train the network does not represent the complete geographic and genetic diversity of these populations. Indeed in many biological datasets, the geographic locations of samples do not capture an entire landscape and instead are often biased, for instance towards easily sampled locations; this inconsistency in spatial sampling is a well-documented problem in applications such as ecological niche modeling (Kramer-Schadt et al., 2013). Further, it is now well appreciated in the broader machine learning field that models trained from even subtly biased training sets, will yield predictions that reflect those biases with increasingly broad societal implications (e.g., Buolamwini & Gebru, 2018; Mehrabi et al., 2021). There is no reason to believe that such biases would not be present in our own spatial population genetic predictions.

In this study we explore the evidence of correlation between `Locator`'s geographic predictions of *A. gambiae* and *P. falciparum* in sub-Saharan Africa, and investigate whether these patterns emerge from spatial imbalance in training data. Using extensive continuous space forward-in-time population genetic simulations, we demonstrate that non-uniform spatial sampling in training data is a pernicious problem in geographic prediction, but does not alone account for the residuals observed in *A. gambiae* and *P. falciparum* populations. We explore and implement methods to counter the bias created by non-uniform training sets, and demonstrate

conditions under which we might expect geographic prediction to be more or less biased. Applying these methods to our predictions of *A. gambiae* and *P. falciparum* locations, we show that while non-uniform sampling in our training sets may elevate correlations between residuals, our signal is beyond what would be expected from sampling considerations alone. We hypothesize that the correlated patterns between host and parasite residuals demonstrate the evolutionary consequences of linked dispersal between *A. gambiae* and *P. falciparum* on a short timescale.

2.2 Materials and Methods

2.2.1 Empirical data. We applied `Locator` to whole-genome sequencing data from the MalariaGEN Ag1000G Project and Pf7k Project (N. Anopheles gambiae 1000 Genomes Consortium, 2017; Pearson et al., 2019). From the Ag1000G dataset, we used 2,416 field-collected mosquitoes from the *Anopheles gambiae/coluzzii* species complex in sub-Saharan Africa. Samples used in the analysis belonged to *Anopheles gambiae* based on ancestry informative markers (N. Anopheles gambiae 1000 Genomes Consortium, 2017). From the Pf7k dataset, we used 3,315 *Plasmodium falciparum* field-collected blood samples collected in sub-Saharan Africa. For the *A. gambiae* dataset, we ran `Locator` in a windowed analysis across all chromosomes, using 2Mbp windows that captured roughly 100,000 SNPs each. A random 10% of samples were held out during training and used as a test set. This was repeated until all samples had been used in the test set, thus obtaining predicted locations for all 2,416 samples. For the *P. falciparum* dataset, we ran `Locator` using 500 kb windows, again capturing roughly 100,000 SNPs each, and held out a random 5% of samples during training. This same process was repeated when running `Locator` with sample weights.

2.2.2 Spatial simulations. Simulations used for assessing `Locator`'s performance on spatially imbalanced datasets were run under the continuous space model described in Battey et al. (2020c) using `SLiM v3` (Haller & Messer, 2019a). Individuals were simulated across a 50 × 50 unit square landscape with an expected population density of 5 individuals per unit area and a mean parent-offspring dispersal distance that varied between 0.4 and 1.0 units. Simulated individuals were diploid, with two copies of a 10^8 bp chromosome that recombined at a rate of 10^{-8} per bp per generation. Simulations were run for 50,000 generations, at which point nearly all tree sequences had coalesced. After simulations were completed, the resulting tree sequence was recapitated and neutral mutations were added at a rate of 10^{-8} per bp per generation. Finally, location metadata and genotype calls were saved for a given simulation. Genotypes were output as VCFs using the `.write_vcf()` function in `tskit`, then converted to Zarr format using `scikit-allel`.

2.2.3 Biased training sets. To create spatial bias in training data, we sampled training sets from the previously-described 50 × 50 unit continuous-space simulations with increasing probability that a sampled individual would be from the right half of the landscape (ie, $x > 25.0$). We refer to this as *training set bias*, which ranged from 0.5 (uniformly sampled) to 0.9 (a given sample having a 0.9 chance of being from the right half of the landscape). Individuals used for prediction (which we refer to as the *test set*) were sampled uniformly, enabling us to assess the trained model's accuracy across the entire 50 × 50 landscape.

To assess how this bias in training data affects `Locator`'s predictions, we applied this method to simulations with per-generation dispersal rates (referred to as σ) between 0.4 and 1.0 units. We used biased sampling to sample 450 individuals as a training set, then trained `Locator` on 100,000 randomly chosen SNPs from these

samples. The trained network was then used to predict the locations of 50 new uniformly-sampled individuals from the dataset. We defined a residual as the vector between a test set individual's true location and their predicted location. To quantify bias in residuals across a given `Locator` run, we separately assessed the mean x and y component of residual vectors. Because the prediction set is uniformly sampled from the landscape, these values will be distributed around zero across multiple runs if predictions are unbiased.

For a set of simulations where $\sigma = 1.0$, we ran `Locator` in a full-genome windowed analysis, using 2Mbp windows containing roughly 8,000 SNPs each. Alongside this, we varied the number of individuals in the test set between 50 and 500, preserving a 9:1 train-to-test ratio in these analyses. Here, we defined a residual as the vector between a test set individual's true location and the geographic centroid of their windowed predictions. These runs were used to assess the impact of training set bias on individual-level prediction uncertainty.

2.2.4 Modifications to training data.

2.2.4.1 Downsampling. To downsample densely-sampled regions of the landscape, we relied on the Generalized Random Tessellation Stratified (GRTS) method (Stevens. & Olsen, 2004), implemented using the `geosample` Python library. This method splits the landscape recursively into quadrants, then randomly samples from each quadrant to create a spatially balanced dataset. In order to assess how downsampling can affect `Locator`'s performance on datasets that are unevenly sampled across space, we applied GRTS sampling to our skewed training sets, resulting in a spatially even subset of a skewed training set's samples being used during training. The same test sets of individuals were used for prediction.

2.2.4.2 *Sample weighting.*

Sample weighting scales each sample's contribution to the overall loss function when fitting `Locator` to training data. Weights were implemented using Tensorflow's built-in `sample_weight` argument to `model.fit()`.

To calculate sample weights, we used `scikit-learn` (Pedregosa et al., 2011) to produce a Gaussian kernel density estimate of the density of spatial locations of training data across the landscape. Two free parameters are involved in assigning sample weights: bandwidth and λ. Bandwidth is used when fitting the kernel density estimate, and λ is used to exponentially scale the weights: if the estimated density at the location of the i^{th} sample is P_i, its weight W_i is

$$W_i = \frac{1}{P_i^{\lambda}}.$$

Bandwidth and λ values can be directly passed to `Locator` as a command-line argument or inferred using a grid search within `Locator`.

In our analysis, we used bandwidth values uniformly distributed between 0.25 and 5, and λ values ranging from 10^{-5} to 10.0. Methods to calculate sample weights using kernel density estimates or 2D histograms are now included in `Locator` using the `--weight_samples` argument. Sample weights can also be calculated externally and provided as a TSV file. Finally, bandwidth and λ parameters for calculation of kernel density estimates can be provided as command-line arguments or identified using the `GridSearchCV` function from `sklearn.model_selection` (Pedregosa et al., 2011).

2.2.5 *A. gambiae* simulations.

Populations used for assessing `Locator`'s performance on empirical *A. gambiae* data were also simulated using the same continuous space model in `SLiM`. Individuals were simulated across a map with boundaries representing the Sub-Saharan range of *A. gambiae* and that contained

all empirical sampling locations. The landscape represented is 6900 kilometers across and 4750 kilometers high; to make this computationally feasible we reduced the scale of landscape and spatial interactions by a factor of 1/20.

An expected population density of 5 individuals per unit area was used in these simulations, resulting in a simulated population size of roughly 175,000 individuals. The effective population size of *Anopheles gambiae* is estimated to be 1,000,000 individuals, with a mutation rate of 3.5×10^{-9} and a recombination rate of 1.3×10^{-8} per base pair per generation (considering chromosome 2L) (N. Anopheles gambiae 1000 Genomes Consortium, 2017). These rates were scaled with the decreased population size in order to keep θ and ρ constant ($\theta = 4N\mu, \rho = 4Nr$).

	Empirical estimate	Scaled simulation
Population size	1000000	175000
σ	13 km	0.65 units
μ	3.5×10^{-9}	2.5×10^{-8}
r	1.3×10^{-8}	7.4×10^{-8}

Table 2.1. Empirical estimates and scaled simulation parameters for *Anopheles* simulations

We ran ten simulations and performed the same analysis on each. After running these simulations for 50,000 generations, we used coalescent simulation in `msprime` to recapitate the resulting tree sequences and added neutral mutations at a rate of 2.5×10^{-8} per bp per generation. To replicate the spatial distribution of empirical *A. gambiae* samples, we sampled an equivalent number of simulated individuals from each location where *A. gambiae* were sampled. Because multiple individuals in `SLiM` cannot occupy the exact same spatial coordinates, we used a k-nearest neighbors algorithm to sample simulated individuals around each empirical sampling point, then assigned those individuals' locations to the empirical point they were sampled by. This method is similar to real-world sampling strategies, where individuals existing

in a small geographic region are often assigned the same spatial coordinates during sample collection.

We applied `Locator` to these simulated datasets in the same fashion as our empirical analysis, holding out a random 10% of samples during training. We again ran a windowed analysis using 2 Mbp windows, which each captured roughly 80,000 SNPs across a 10^8 bp genome, representing slightly less genomic information than the empirical *A. gambiae* data.

2.3 Results

2.3.1 Errors in geographic predictions of *A. gambiae* and *P. falciparum* are correlated.

A. gambiae mosquitoes are the vector of the malaria-causing parasite *P. falciparum*. While the parasite spends most of its life cycle in humans, female mosquitoes acquire *P. falciparum* gametocytes when feeding on an infected human, which then sexually reproduce before being transmitted as fertilized sporozooites to a human host during the next blood meal (Hall et al., 2005). Though per generation dispersal is limited to perhaps as little as roughly six kilometers per generation (C. C. Smith et al., 2023), the dispersal of infected *A. gambiae* mosquitoes will consequently bring along the infecting *P. falciparum*. It is less clear whether this host's dispersal meaningfully affects *P. falciparum* movement and genetic structure.

When applying `Locator` to empirical data from *A. gambiae* and *P. falciparum*, we observed similar patterns in prediction errors (i.e. residuals) between spatially proximate samples of the host and its parasite (Figure 2.1 A, B, C). Battey et al., 2020a showed that `Locator` predictions reliably reflect individuals' geographic ancestries in some well-documented cases from humans. Thus we hypothesize that our observation of similar prediction errors between between host (*A. gambiae*) and parasite (*P. falciparum*) may reflect common, recent co-geographic patterns.

These patterns are driven primarily by large prediction errors in West Africa, potentially representing a region of relatively high gene flow for both host and parasite. Essentially because *P. falciparum* depends on *A. gambiae* mosquitoes for a portion of its life cycle, some degree of dispersal will be due to dispersal of infected mosquitoes, which should induce co-variation in their evolutionary histories across the landscape.

We first aimed to quantify the relationship between the two datasets' residuals by applying Crosby et al., 1993's technique for measuring correlation between pairs of vector-valued time series. Originally developed for geophysical applications, this vector correlation metric accounts for both magnitude and direction of a set of vectors, and ranges from zero (no correlation) to two (complete correlation). To apply this method to our datasets, we tested for correlation between paired *A. gambiae* and *P. falciparum* individuals from the same approximate geographic area, defining vectors from an individual's true location to their predicted location (i.e. the residuals of our predictions). Because multiple samples can occur in one location, we ran this analysis multiple times, with each iteration resulting in different pairings of spatially proximate individuals. For comparison, we also permuted spatial pairings between species to create random pairs with respect to space, and then applied the same correlation procedure to those randomized datasets. This set of randomly paired data represents a naive null distribution to which we compare our observed correlations.

The mean vector correlation statistic between spatially paired residuals was 0.284 ($SD = 0.038$), which was significantly different (Wilcoxon test, $p < 2.2 \times 10^{-16}$) from the correlation observed in our permuted pairs (Figure 2.1 D). This suggests that there are common patterns of error in `Locator`'s predictions of *A. gambiae* and *P. falciparum* locations. However, Crosby et al., 1993's vector correlation statistic can be influenced by both the angles and magnitudes of the vectors in question, making

interpretation hazy. Because of this, we also performed separate Spearman rank correlation tests on the angle and magnitude of residual vectors individually from the same spatially- and randomly-paired datasets (Figure 2.1 E, F). Both the angles and magnitudes of error vectors were significantly correlated between spatially-proximate *A. gambiae* and *P. falciparum* samples ($R = 0.163$ ($SD = 0.041$), $p < 2.2 \times 10^{-16}$ and $R = 0.924$ ($SD = 0.004$), $p < 2.2 \times 10^{-16}$), further supporting that geographic errors in predictions are similar between the two species.

We further examined the correlation in genetic structure between the species by comparing principle components of species' genotype matrices (Figure A.3). This reveals a similar pattern whereby spatially proximate samples are embedded in PCA space in a correlated fashion. Indeed, longitudinal differention seems to be shared among *A. gambiae* and *P. falciparum*.

Is this a biological signal? The observed correlation between *A. gambiae* and *P. falciparum* residuals is beyond what we would expect under a naive null, particularly in West Africa, but does this implicate the movement of mosquito vectors in the spatial spread of malaria parasites? Or a genetic signal of shared geographic history between these species? Further, while our observed residuals align with known corridors of gene flow, particularly in West Africa (N. Anopheles gambiae 1000 Genomes Consortium, 2017), they also trend towards locations with increased sampling density. This raises a possible statistical issue – are our predictions simply pointing towards geographic regions that are over-represented in our training set? Unbalanced training data is well-known to create issues for machine learning based inference, so we were interested to assess how geographic prediction is influenced by training data that is non-uniformly sampled throughout space and if (or to what

degree) such an effect might be responsible for our observed correlation between *A. gambiae* and*P. falciparum* residuals.

2.3.2 Spatial bias in training data leads to biased geographic predictions. To understand the impact of spatially imbalanced training data on predictions, we applied `Locator` to samples from populations simulated using a continuous space, forward-in-time model implemented in `SLiM` (Haller & Messer, 2019a). In these simulations, the spatial aspect of dispersal, competition, and mating of individuals are each controlled by a Gaussian kernel with mean zero and standard deviation σ in each dimension, which we vary across simulations.

To generate spatially imbalanced data from these simulations, we created training sets where samples were increasingly likely to be sampled from the right half of the landscape (Figure 2.2 A). In these simulations, the relatedness between individuals varies smoothly across the landscape (i.e. the degree of spatial population structure is relatively continuous (Battey et al., 2020c)), so we expect any systematic spatial patterns in prediction error to be attributed to sampling imbalance in training data. After training `Locator` on 450 individuals sampled in this fashion, we then used the trained network to predict the locations of held-out individuals that were sampled uniformly across the landscape. By separating the vector formed between an individual's true and predicted location into its x and y components, we were able to separately quantify error along the axes of imbalanced sampling (x) and random sampling (y). These values were then averaged across each `Locator` run, summarizing the trends in prediction errors for a given training set.

We found that when `Locator` is trained with spatially biased data, predictions are biased towards more densely sampled regions of the landscape. In other words, training sets that were biased towards the right half of the landscape resulted in an

excess of predictions to the right of samples' true locations, with this effect becoming more pronounced as dispersal rate (σ) increases (Figure 2.2 B). We quantified the contribution of training set bias and per-generation dispersal rate (both normalized between 0 and 1) to x-axis error using multiple linear regression; both variables have a significant positive effect on x-axis error ($\beta_1 = 0.128, p = 1.61 \times 10^{-15}$ and $\beta_2 = 0.090, p = 1.17 \times 10^{-8}$ for bias and σ, respectively).

While bias in prediction errors is observed along the axis of imbalanced sampling, there is no relationship between y-axis error and training set bias (Figure 2.2 C), suggesting that bias in residuals reflects the spatial imbalance in training data. A linear model confirms that training set bias has no effect on y-axis error, ($p = 0.244$) but dispersal rate has a small but significant positive effect ($\beta_2 = 0.039$, $p = 0.013$). This phenomenon is likely due to overfitting: when learning the relationship between genotype and geography, `Locator` aims to minimize the distance between true and predicted locations of training data, which, with biased training sets, may be most efficiently done by learning to predict towards densely-sampled regions of the landscape, regardless of true sample location. This prediction bias is amplified at higher dispersal rates due to the reduction in spatial genetic structure that `Locator` is picking up on, and predictions are increasingly based on locations the network has already seen.

2.3.3 Spatial distribution of training data does not explain Anopheles residuals. To assess the contribution of spatial sample imbalance to observed patterns in empirical residuals, we again applied `Locator` to simulated randomly dispersing populations, this time using training samples that mimic the spatial distribution of the Ag1000G dataset. To accomplish this we performed simulations which reflect Africa's *Anopheles gambiae* population, which has an

estimated per-generation dispersal rate of 13 kilometers per generation and an effective population size of approximately 1 million individuals (N. Anopheles gambiae 1000 Genomes Consortium, 2017; C. C. Smith et al., 2023). Populations were simulated across a landscape representing Sub-Saharan Africa, which contains all sampling sites and reflects the general range of *Anopheles* mosquitoes. After running these simulations, we sampled individuals in the same spatial pattern as the Ag1000G dataset and applied `Locator` in a windowed analysis across a 10^8 bp genome, using 2Mbp windows that capture roughly $80,000$ SNPs each - less genetic information than the empirical data. A rotating 10 percent of samples were held out during training runs and used as a prediction set in the same fashion as our empirical analysis. The outcome of these simulations is a rough approximation of African *A. gambiae* populations under neutral spatial dynamics, not accounting for variation in population density, seasonal changes, or mosquito-host dynamics (among other factors). However, we can use these simulations to represent a baseline estimate of population structure across space under random dispersal. As in our right-biased training sets from randomly dispersing simulations, we can then assess the impact of spatially imbalanced sampling on `Locator`'s predictions, this time comparing them to predictions from the true Ag1000G dataset and the data from *P. falciparum*.

Despite training data being sampled unevenly from the landscape, `Locator`'s performance on these simulated populations was relatively accurate (Figure 2.3). While predictions for individuals near landscape edges were slightly biased towards the center, predictions near the center of the landscape - particularly in West Africa - were surprisingly close to individuals' true locations. This is in contrast to `Locator`'s performance on the empirical *A. gambiae* dataset, where samples near the edges of the continent are predicted accurately and West African samples have large, inward-

pointing residuals (Figure 2.3A). This increase in prediction accuracy is also visible in the distribution of error magnitudes, which is bimodal in the case of empirical data but more evenly distributed in the case of the simulated data (Figure 2.3B). We then tested for correlation between empirical *P. falciparum* and simulated *A. gambiae* residuals. The goal of this comparison in essence is to remove any possibility of shared or biased dispersal histories between species while still allowing for continuous population structure. We find that vector and angle correlations when testing against simulated data are significantly less that what we observe in the real data, though a correlation remains in the magnitude of residuals (Figure 2.3C, D, E).

While the correlation between empirical *P. falciparum* and simulated *A. gambiae* error magnitude indicates that the spatial distribution of samples may influence similarities in predictions, many major features of the empirical predictions of *A. gambiae* locations are not reproduced by simulated neutral spatial processes. In particular, residuals of empirical samples in West Africa are large and trend towards the center of the region in a similar fashion to *P. falciparum*, while residuals of simulated samples are relatively small. *A. gambiae* populations in this region have high levels of gene flow, despite large geographic distances, while populations in Central Africa show more genetic differentiation (N. Anopheles gambiae 1000 Genomes Consortium, 2017; Clarkson et al., 2020). These results suggest that the correlation between prediction errors observed when `Locator` is applied to the Ag1000G and Pf7k datasets are not only due to the spatial distribution of sampling locations, and may be indicative of common gene flow and non-uniform migration patterns.

2.3.4 Accounting for spatially biased data in geographic prediction.

While our simulations approximating *A. gambiae* populations suggested that patterns

in `Locator`'s prediction errors were not wholly due to spatial bias in empirical datasets, we were interested in reducing `Locator`'s prediction bias when applied to spatially biased datasets. Because most natural populations are not sampled in a way that accurately reflects their range and density, it is possible that spatial sampling bias may impact `Locator`'s predictions in other empirical applications. To see if prediction bias could be reduced in these situations, we again applied `Locator` to our simulation data. Because the relationship between training set and prediction bias is stronger at higher dispersal rates, we focused on simulations with a dispersal rate of $\sigma = 1.0$ per generation. We first attempted to reduce prediction bias by downsampling individuals in densely-sampled regions of the landscape - this successfully reduced mean prediction bias (though not significantly), but in turn increased prediction error (Figure 2.4 A, C, D, E). This makes sense, as downsampling reduces the overall amount of training data available. In an attempt to make use of all the samples in our dataset while still accounting for spatial bias, we then implemented a sample weighting scheme in `Locator`'s training process.

During each epoch of training, `Locator` aims to minimize the sum of distances between true and predicted locations (the *loss*). In the case of spatially biased datasets, this is done most efficiently by learning to predict in areas with dense sampling, leading to prediction bias. To reduce bias towards densely sampled portions of the landscape during training, we weighted each sample's contribution to the overall loss according to its local spatial density: samples in sparse regions were upweighted, and samples in dense regions were downweighted.

To calculate sample weights we fit a kernel density estimate over the landscape, then assigned each sample a weight value inversely proportional to the sampling density in that region raised to a power λ. We varied two parameters when

assigning sample weights: bandwidth, which controls the "smoothness" of the kernel density estimate (Figure A.6 B), and λ, which effectively increases the difference between the lowest and highest weights (Figure A.6 A). Further discussion of sample weight parameterization is given in the Methods (section "Sample Weighting"). As kernel bandwidth and λ are effectively hyper-parameters of our model used to tune weighting, we performed a grid search across a discretized space of λ and bandwidth to identify effective parameter combinations that reduced loss for a given training set.

In application, it is difficult to disentangle biological causes of prediction error from potential bias caused by a sampling scheme. Therefore in our grid search we aimed to identify the parameter combination that minimized *overall* prediction error as opposed to prediction bias (Figure A.5). The parameter combination that most significantly reduced error on training sets with a sampling bias of 0.9 was a bandwidth value of 0.25 and a λ value of 10^{-5} which reduced median prediction error by 14% (from 4.374 to 3.753 units). Examining prediction bias along the x axis, this reduces the linear relationship between normalized bias and x-axis error from 0.3838 ($p < 2\times10^{-16}$) to 0.1473 ($p = 0.1449$) (Figure 2.4 A). Like downsampling, sample weighting also results in a bias-variance tradeoff, though less dramatically: for a training set bias of 0.9, sample weighting reduced a run's mean x-axis error from 0.5398 to 0.2947 units, but increased the standard deviation in x-axis error from 0.1529 to 0.8224 units (Figure 2.4 A, C).

Overall, weighting sample importance during training provides a way to reduce, though not completely eliminate, the issues caused by spatially biased training datasets. Methods to calculate and implement sample weights during training are now included in `Locator`, and future applications of the neural network to empirical

datasets may benefit from considering the spatial distribution of available training samples during analysis.

2.3.5 Correlation between *A. gambiae* and *P. falciparum* prediction errors is not due to bias. Having identified a method of reducing prediction bias when training with spatially uneven data, we then used sample weights when applying `Locator` to the Ag1000G and Pf7k datasets. To identify parameter values for assigning sample weights, we performed a grid search over bandwidth values between 1 and 10,000 and λ values between 1 and 10. To reduce compute time, we ran `Locator` on 100,000 randomly chosen SNPs from chromosome 2R for the *A. gambiae* data and on the first 500kb of chromosome 0 for the *P. falciparum* data, then used the final validation loss after training to assess performance. For both datasets, a combination of bandwidth = 100 and $\lambda = 1.0$ resulted in the largest decrease in validation loss. The mean validation loss for the unweighted *A. gambiae* dataset was 0.115 ± 0.018; with sample weighting, the mean validation loss was reduced to 0.098 ± 0.012. For the unweighted *P. falciparum* dataset, the mean validation loss was 0.351 ± 0.039; with sample weighting, this was reduced to 0.297 ± 0.026.

We then ran `Locator` on the full *A. gambiae* and *P. falciparum* datasets using sample weights, setting bandwidth = 100 and $\lambda = 1.0$ (Figure 2.5 A). This only reduced mean prediction error in the *A. gambiae* dataset from 130.3 to 119.8 km (not significant by Wilcoxon test), and slightly elevated median prediction error from 12.8 to 13.4 km. In the *P. falciparum* dataset, mean prediction error was significantly reduced from 203.1 to 151.2 km (Wilcoxon test, $p = 2.24 \times 10^{-14}$), and median prediction error was reduced from 8.2 to 6.9 km (Figure 2.5 B, C). With sample weighting, the correlation between spatially proximate *A. gambiae* and *P. falciparum* residuals is still significant , though reduced in magnitude (Figure 2.5 D, E, F):

correlation between vectors is $R = 0.079 \pm 0.033$, correlation between residual angles is $R = 0.076 \pm 0.037$, and correlation between error magnitudes is $R = 0.907 \pm 0.006$ (all metrics $p < 2.2 \times 10^{-16}$).

Using sample weighting during training, we were able to reduce prediction error in our empirical applications, and likely in turn reduce the degree of spatial bias in predictions. Even so, the resulting prediction errors when applying weighted `Locator` to *A. gambiae* and *P. falciparum* populations still display similar patterns and are significantly correlated between the two species. This indicates that while the spatial distribution of sampling locations may contribute to prediction errors, `Locator` is nonetheless influenced by genetic signals of dispersal.

2.4 Discussion

Motivated by the observation that the residuals in spatial predictions from population-scale genomic samples of *Anopheles gambiae* and *Plasmodium falciparum* are correlated, here we ask to what degree is this correlation the biological signal of the linked evolutionary history between host and parasite versus an artifact of spatially biased sampling? Battey et al., 2020a showed that in some well-documented cases, large `Locator` residuals were associated with known admixture events or population movements. Because *Plasmodium* lives a portion of its life cycle in an *Anopheles* host, the geographic histories of these two species are correlated by necessity, thus spatial assignment predictions from genetic variation would be expected co-vary between host and parasite. While that is so, the spatial sampling of genotypes across the landscape is highly non-uniform, even with the massive publicly available datasets of the Ag1000G and Pf7k consortium projects. This of course makes sense – for instance some sampling locations will be more easily accessible than others – but to what degree will this impact our predictions? Here we use simulation to explore the impact of

spatial sampling bias on the correlation of spatial assignment predictions between host and parasite. We find that the correlation of spatial assignment predictions between *A. gambiae* and *P. falciparum* is sensitive to the spatial sampling bias of the host, that biased spatial sampling leads to biased predictions, and demonstrate conditions under which we may or may not expect geographic bias in prediction error. We then examine how the effect of spatially biased training data might be mitigated and evaluate alternative strategies towards this end. Finally, by implementing algorithmic improvements to `Locator` and comparing spatially explicit simulations to empirical data, we show that the magnitude and degree of residual correlation between *A. gambiae* and *P. falciparum* significantly exceeds that expected from algorithmic bias, identifying an instance of co-geographic patterns of genetic variation between parasite and host.

Sample weights reduce biased inference A fundamental problem in statistical modeling is that biases in training data can easily lead to biased predictions. This is true of all methods to various degrees, but it could be argued that supervised learning approaches might be expected *a priori* to be more susceptible. Given the increasing ubiquity of supervised machine learning in society, as well as its growing popularity in genetics applications, it is important to understand the specific ways in which biased training data can affect predictions and how best to deal with them (e.g., Buolamwini & Gebru, 2018). Here, we examined `Locator`'s predictions from spatially biased training data and demonstrated methods that can reduce prediction bias caused by these training sets. Unsurprisingly, sample weighting during training was more effective at reducing bias than simply downsampling data; this makes sense as sample weighting still exposes the network to all available training samples. However our methods of sample weighting did not fully eliminate prediction bias, and in many

regions of parameter space had either negligible effects on performance or made performance worse (Figure A.5). While we tested a number of methods before settling on kernel density estimates, exploring more sophisticated applications such as using an adaptive bandwidth for the kernel density estimate or using a different function to scale sample weights may further improve accuracy and generalizability. Additionally, investigating the effects of sample weighting on different spatial sampling sampling schemes, such as point or midpoint sampling, would be informative as to how the method performs on empirical datasets.

We have described a method that can reduce prediction bias when training with unbalanced datasets, but this solution is ultimately limited by the bias-variance tradeoff. In our case, sample weighting successfully reduced both error magnitude and bias in our simulated datasets, however the overall variance in these values is greater than when training without sample weights. At its core, we are running up against a inferential limitation of our model. As we try to make our model more generalized - training `Locator` to predict locations across the entirety of the landscape - we sacrifice the precision of our inference. Ultimately, and like all inference methods, `Locator` will be limited by the distribution of data used during training. Spatial variation in training set density may represent the ecology of the organism: undersampled regions may reflect local changes in population density and regions with an absence of samples may represent locations from which the species is excluded. On the other hand sparsely-sampled portions of the landscape may occur due to simple sampling biases, such as unequal sampling effort, and thus pose a statistical challenge to correct for. Our implementation of sample weights has the power to increase prediction accuracy and reduce the bias that results from spatially imbalanced training data, but like in

other biased inference problems, performance, parameterization, and interpretation of predictions are dependent on the given dataset.

Empirical residuals reflect signal beyond sampling bias In this study, we carefully assessed the potential for prediction bias in spatial genotype assessments of our empirical datasets. Using a novel modification to our methods and spatially explicit simulation, we demonstrated that the correlation between *A. gambiae* and *P. falciparum* prediction errors in Sub-Saharan Africa is significantly greater than we would expect in comparison to a simple null distribution This correlation in patterns of error persisted even after applying sample weights and improving overall prediction accuracy. Further, comparing empirical *A. gambiae* predictions to those from forward-in-time, continuous space population genetic simulations meant to mimic mosquito population dynamics, we demonstrated that the correlated residuals we observe in *A. gambiae* and *P. falciparum* predictions can not be explained by sampling alone. Although our model of *A. gambiae* is relatively unsophisticated, and does not incorporate major features of population ecology and evolution such as habitat variation, local adaptation or segregating chromosomal inversions, we have no reason to suspect that including these features would affect our results. In comparison to empirical *A. gambiae* predictions, residuals from our our simulated data were unimodal and generally smaller than the empirical residuals. This is particularly noticeable for individuals from West Africa, where the residuals are systematically larger than what we observe in the rest of the data set and suggest high migration rates and source-sink dynamics over that portion of the range in agreement with previous observation e.g. Clarkson et al., 2020. While our simulated *A. gambiae* model provides a decent baseline for comparison, we are not directly modeling the linked life-histories of host and parasite. A more complex, two-species simulation,

which models the interaction between *A. gambiae* and *P. falciparum* could help to reveal the degree to which host-parasite co-dispersal influences shared patterns of gene flow and how these patterns can be detected using inference methods like `Locator`.

We hypothesize that the correlation we observe here can be attributed to common patterns of dispersal between the two species - as *P. falciparum* depends on the *A. gambiae* mosquito for part of its life cycle, its movement will be partially limited by the movement of its mosquito host. The majority of this correlation structure seems to be driven by prediction errors in West Africa, where *A. gambiae* populations have been shown to have increased levels of gene flow and reduced differentiation Clarkson et al., 2020. Indeed when applied to the Ag1000g dataset, `Locator`'s high prediction errors in West Africa and accurate predictions in Central and East Africa align with known patterns of spatial population structure in the *A. gambiae* species complex Clarkson et al., 2020. Under the assumption of paired host-parasite dispersal, we would expect *P. falciparum* gene flow to also be elevated in West Africa and that prediction errors would have similar structure to *A. gambiae*, just as we observe. Further, humans are a key third player in the *A. gambiae*-*P. falciparum* interaction, as humans are hosts to both species, and accordingly *P. falciparum* dispersal has been shown to correlate with human mobility (Wesolowski et al., 2012). West Africa is a region of high human mobility (Wesolowski et al., 2014) - given that *A. gambiae* and *P. falciparum* are both human parasites (though with different degrees of linked dispersal), these patterns of gene flow may in part reflect anthropogenic factors.

Hosts and parasites share evolutionary history The degree to which a parasite depends on its host for dispersal shapes the strength of shared evolutionary history among taxa. Over longer timescales host-dependant dispersal can lead to co-diversification or even co-speciation between host and parasite (reviewed in Althoff

et al., 2014). At shorter timescales, hosts and parasites might share features of genetic variation that are correlated, for instance reductions in variation due to shared population bottlenecks. Accordingly empirical population genomic surveys of parasite species have repeatedly recovered the signature of shared evolutionary history among host and parasite. For instance Doyle et al. (2022) showed that estimates of historical demography for the human whip worm, *Trichuris trichiura*, shared signatures of the human out-of-Africa bottleneck event. Similarly Small et al., 2019 convincingly demonstrated that the biogeographic history of the filarial parasite *Wuchereria bancrofti* was shaped by a series of human dispersal events throughout the Indo-pacific. Counter examples of course exist. When host and parasite dispersal are less tightly linked, such as for the widely distributed intracellular parasite Hamiltosporidium and its host Daphnia species, co-biogeography and co-demography are not apparent (Angst et al., 2022). We speculate that the correlated residuals on geographic prediction between *A. gambiae* and *P. falciparum* is the result of their shared evolutionary interaction as host and parasite. Indeed Plasmodium is host-dependant with respect to dispersal, so our *a priori* expectation should be that of co-geography. The timescale of the processes which we are observing are very roughly on the order of the population size number of generations or fewer - using supervised machine learning methods, we are able to identify an instance of co-geography at the population scale.

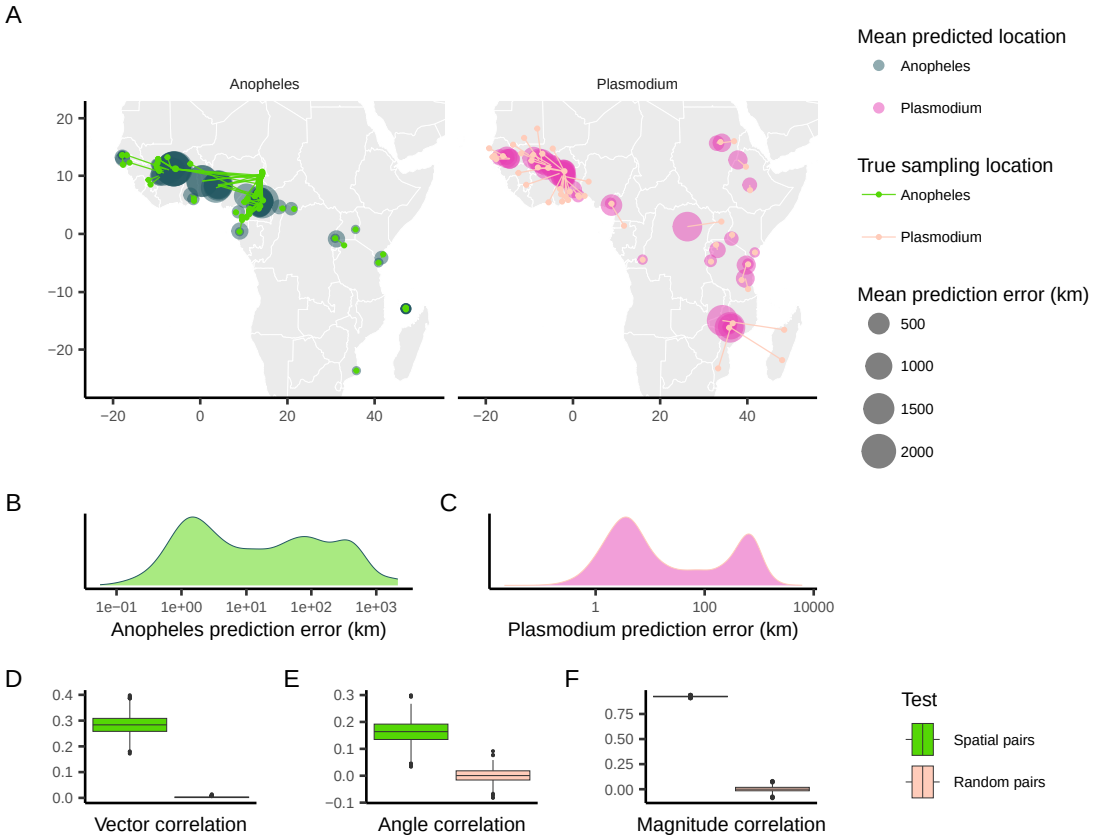


Figure 2.1. Predicted locations of *Anopheles gambiae* and *Plasmodium falciparum* samples. (A) Geographic centroids of predictions for each unique sampling site. True sampling locations are connected to their mean predicted location; the size of predicted locations represents error in km. (B, C) Distribution of error magnitudes for *A. gambiae* and *P. falciparum* predictions. (D, E, F) Vector, angle, and magnitude correlations between spatially paired and randomly paired *A. gambiae* and *P. falciparum* residuals.

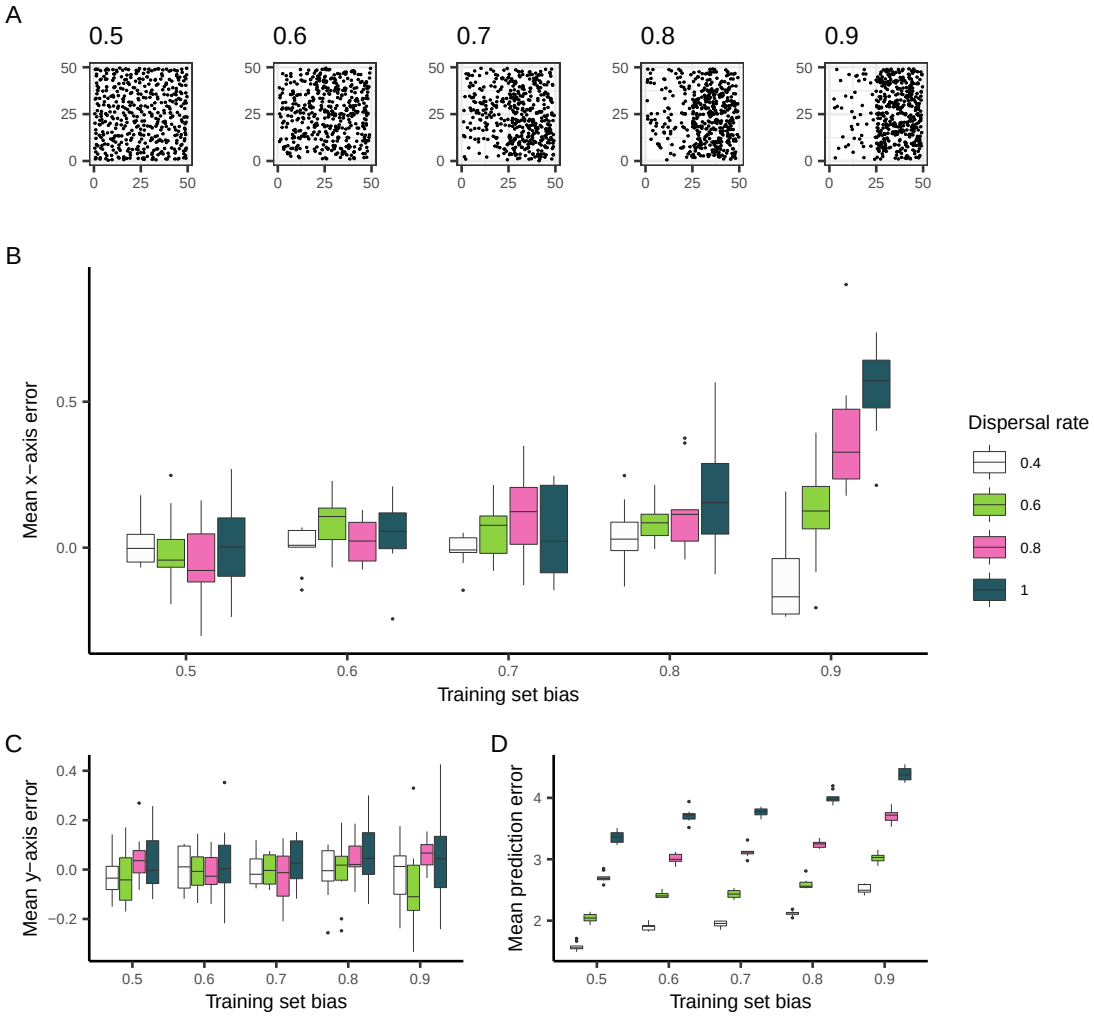


Figure 2.2. The effect of training set bias on `Locator` residuals. (A) Training sets sampled with increased bias towards the right side of the landscape. (B) Mean per-run prediction error along the x axis as a function of training set bias, (C) Mean per-run prediction error along the y axis, (D) Mean per-run error magnitude.

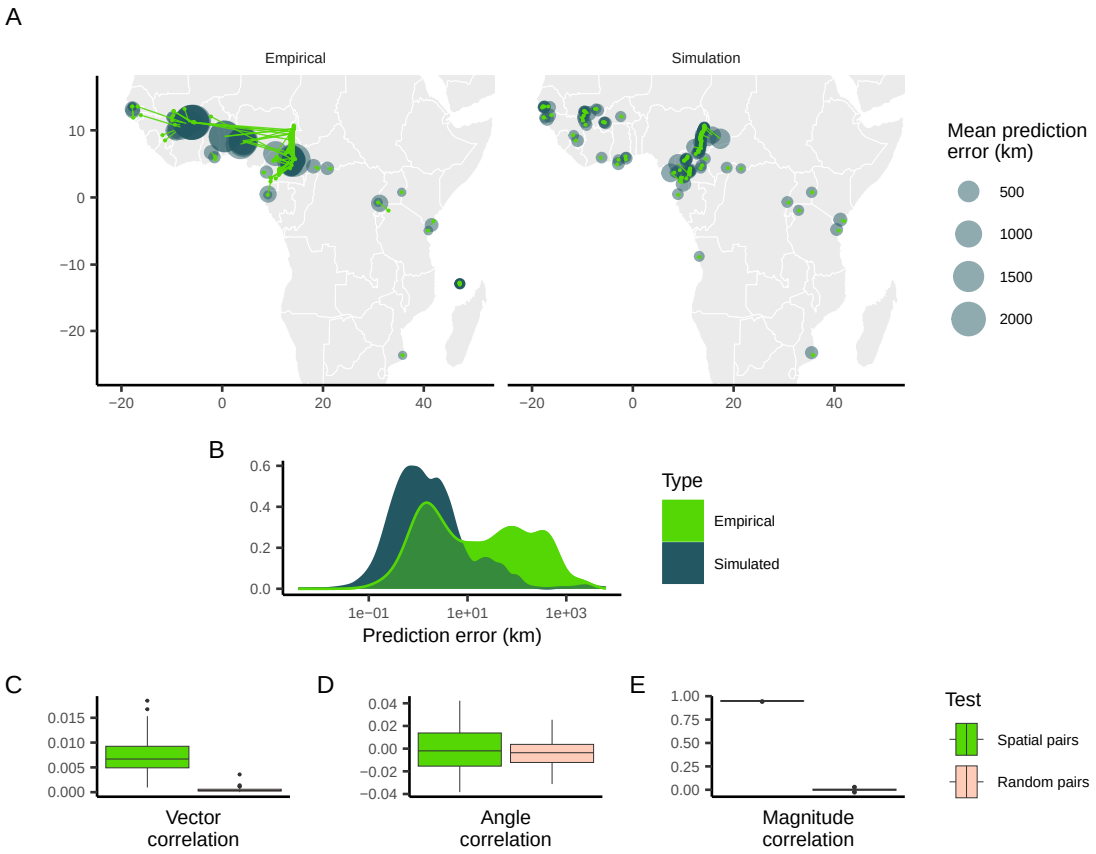


Figure 2.3. Predicted locations of empirical and simulated *Anopheles gambiae* samples. (A) Geographic centroids of predictions for each unique sampling site. True sampling locations are connected to their mean predicted location; the size of predicted locations represents error in km. (B) Distribution of error magnitudes. (C, D, E) correlation coefficients between simulated Anopheles and empirical Plasmodium predictions for the residual vectors, residual angles, and residual magnitude respectively

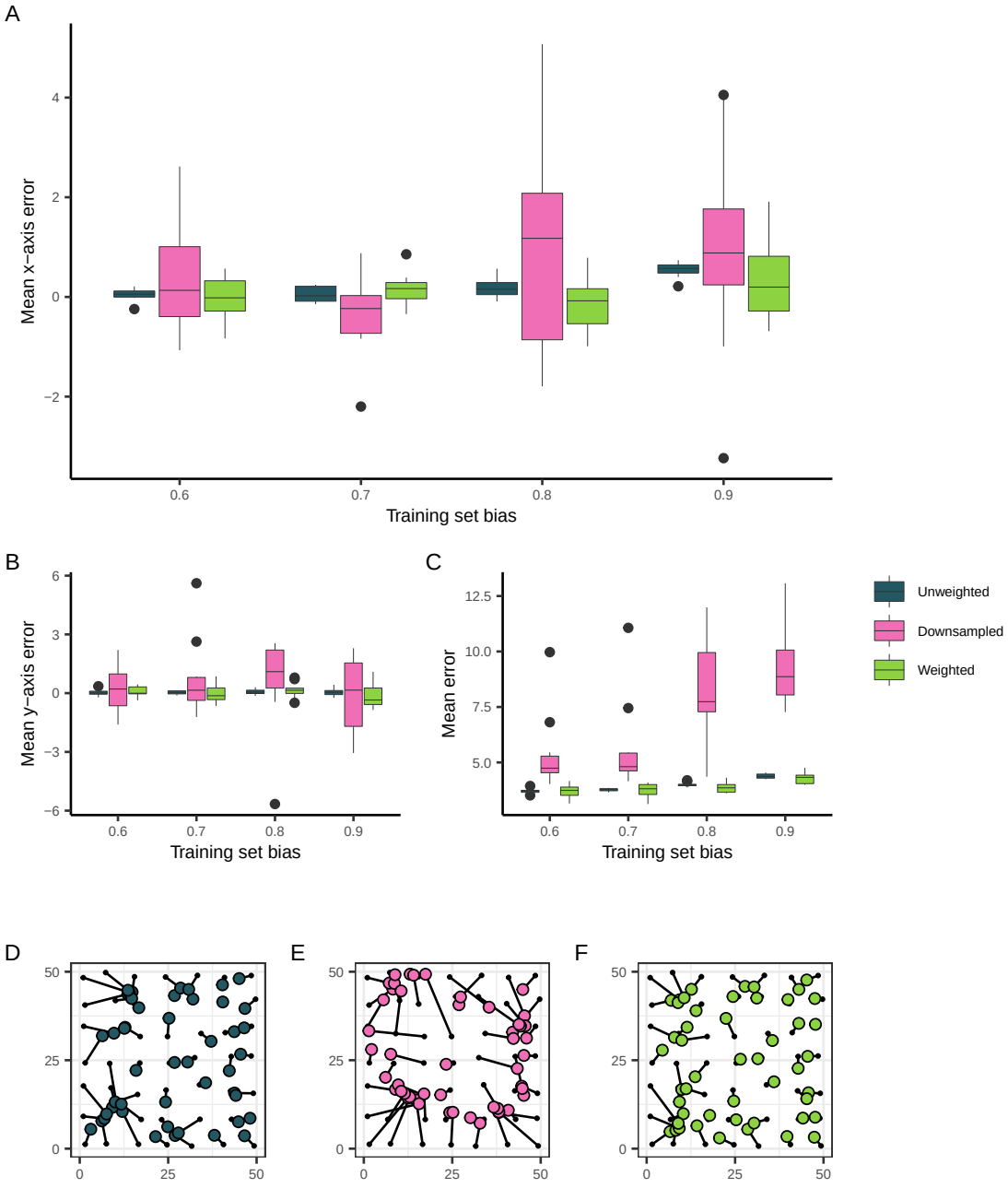


Figure 2.4. The effect of training set bias using sampling and training modifications. (A) Mean per-run prediction error along the x axis as a function of training set bias, (B) Mean per-run prediction error along the y axis, (C) Mean per-run error magnitude. (D, E, F) Example predictions at bias = 0.9 for unmodified, downsampled, and weighted training sets, respectively. True sample locations (black points) are connected to their predicted location (colored points).

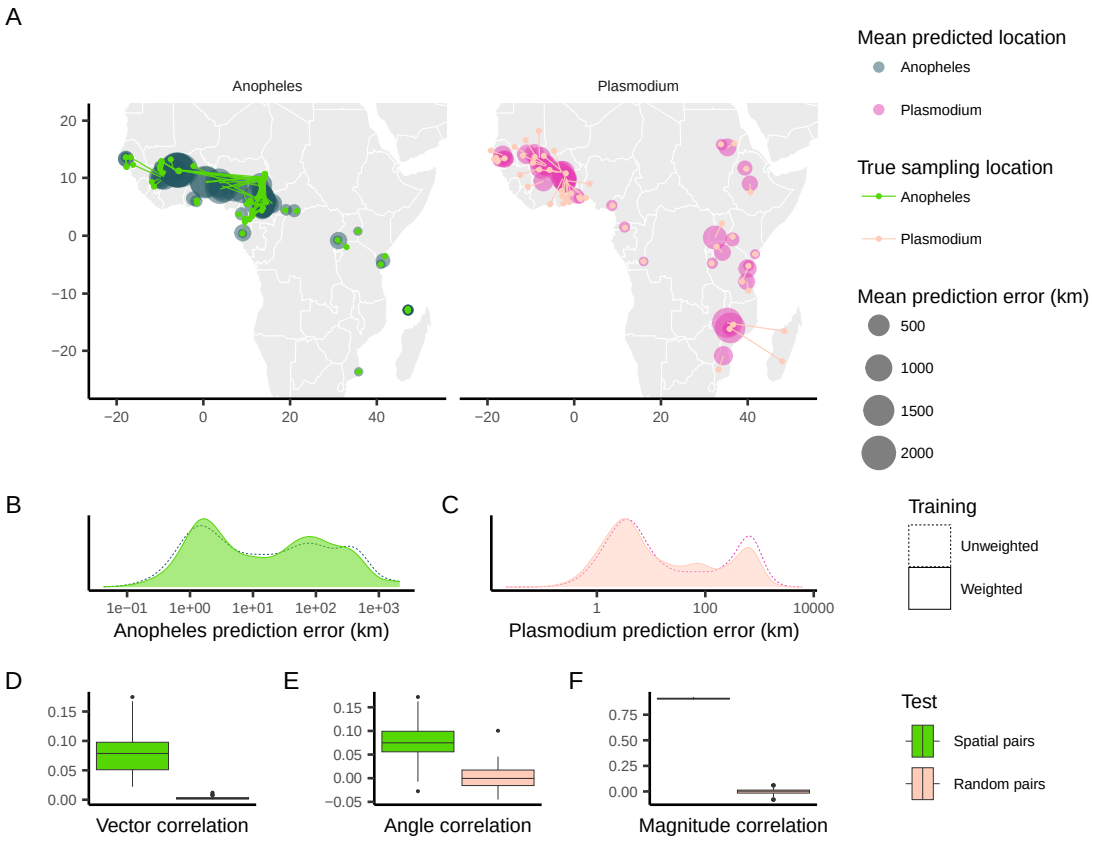


Figure 2.5. Predicted locations of *Anopheles gambiae* and *Plasmodium falciparum* samples after training using sample weights. (A) Geographic centroids of predictions for each unique sampling site. True sampling locations are connected to their mean predicted location; the size of predicted locations represents error in km. (B, C) Distribution of error magnitudes for both unweighted and weighted *A. gambiae* and *P. falciparum* predictions. (D, E, F) Vector, angle, and magnitude correlations between spatially paired or randomly sampled *A. gambiae* and *P. falciparum* weighted residuals.

2.5 Bridge

In Chapter II, I use continuous space simulation and machine learning to evaluate evidence for co-geography in the *Anopheles-Plasmodium* host-parasite system, ultimately finding evidence for common dispersal patterns between the parasite and its mosquito vector. Identifying patterns of spatial population structure and dispersal in this system, however, led me to ask further questions about how spatial structure may be shaping the evolutionary dynamics of these species. How do beneficial alleles spread through these spatially-structured populations, and can we use spatial structure to detect them? In Chapter III, I present my work investigating these questions, using a similar continuous-space simulation framework to uncover a novel spatial signature of selective sweeps and then applying this observation to identify positively-selected alleles in a spatially-referenced genomic dataset of *Anopheles gambiae* mosquitoes.

CHAPTER III

SWEEPS IN SPACE: LEVERAGING GEOGRAPHIC DATA TO IDENTIFY BENEFICIAL ALLELES IN *ANOPHELES GAMBIAE*

This chapter was published in the journal *Molecular Biology and Evolution* in 2025. Scott T. Small, Peter L. Ralph, and Andrew D. Kern are co-authors on this work. I conceptualized this study together with A.D.K. and P.L.R.; S.T.S. assisted in processing the *Anopheles gambiae* genomic dataset and P.L.R. assisted with our conceptual model. I performed all analyses with input from and wrote the manuscript with editorial assistance from all co-authors.

The citation for this publication is as follows:

Rehmann, C. T., Small, S. T., Ralph, P. L., & Kern, A. D. (2025a). Sweeps in space: Leveraging geographic data to identify beneficial alleles in anopheles gambiae. *Molecular Biology and Evolution*, *42*(6), msaf141. https://doi.org/10.1093/molbev/msaf141

3.1 Introduction

Positive selection is a fundemental force in population genetics, and selective sweeps are responsible for many well-known examples of adaptation such as changes in melanin production in humans (Williamson et al., 2007), insecticide resistance in *Drosophila* (Schlenke & Begun, 2004), and a major architectural differences between teosinte and maize (Clark et al., 2004). The foundational model of a selective sweep describes how a novel, positively selected variant increases in frequency and eventually fixes (S. Fisher, 1930; Kojima & Kelleher, 1962). During this process, neutral variants linked to the sweeping variant also increase in frequency, ultimately reducing variation around the site of the sweeping allele (i.e. genetic hitchhiking; Kojima and Schaffer,

1967; Maynard Smith and Haigh, 1974). This observation has been the basis of numerous tests for identifying positive selection in genomic data, including those based on summary statistics (Fay & Wu, 2000; Tajima, 1989), explicit likelihood models (Chen et al., 2010; DeGiorgio & Szpiech, 2022; DeGiorgio et al., 2016; Nielsen et al., 2005), and more recently through the use of supervised machine learning (Arnab et al., 2023; Kern & Schrider, 2018; Lauterbur et al., 2023; Schrider & Kern, 2016; Sugden et al., 2018; Xue et al., 2021b). However, this foundational model of a selective sweep - and the methods based on it - assume populations are panmictic and without spatial heterogeniety. While these assumptions simplify models of sweeps, they also overlook potentially confounding variables, as well as those that could better inform analysis.

One such factor is that most natural populations live in landscapes where their dispersal and interactions are geographically limited, leading to a near ubiquitous observation in population genetics - isolation by distance (IBD) - that individuals tend to be more closely related the closer they are in geographic space (Wright, 1943a). This effect of spatially limited dispersal can have profound impacts on how genetic variation is shared among individuals in a population, both with respect to common summaries of variation such as heterozygosity and the site frequency spectrum, as well as downstream inferences based on that variation such as inferred effective population sizes and genomic association tests (e.g., Battey et al., 2020). These effects have implications for the ways that evolutionary processes, such as selective sweeps, are identified in natural populations.

Prior investigation into selective sweeps in spatially structured populations has revealed that a beneficial allele will move across space and through the population as a travelling wavefront, historically described in a one-dimensional population by

R. A. Fisher, 1937 and in two dimensions by Kolmogorov et al., 1937. In this model, dispersal distance determines the speed at which a sweep progresses, with smaller dispersal distances slowing the allele's frequency trajectory relative to the panmictic expectation (N. H. Barton et al., 2013). Min et al., 2022b demonstrated how structure across one-dimensional space impacts the population genetic signatures expected after a selective sweep, most notably causing an excess of variants at moderate frequency that were able to recombine onto the sweeping haplotype due to the slower nature of the sweep. Further, Chotai et al., 2024 recently explored the signatures of sweeps in populations structured across continuous two-dimensional space. This study corroborated results from one-dimensional explorations of sweeps in showing how sweeping alleles move more slowly in spatially structured populations and demonstrated that spatial structure "softens" the genetic signals expected from selective sweeps, causing them to appear more like soft sweeps (*c.f.*, Hermisson & Pennings, 2005). While our understanding of how spatial population structure changes expectations for the genetic signatures left by selective sweeps has increased, particularly in how it may reduce our capacity to detect them in genomic data, it remains unexplored whether this structure can be leveraged to inform genetic analysis of positive selection.

Spatial patterns of genetic variation can reveal information about evolutionary process, rather than simply confound. For example, Battey et al., 2020b showed how machine learning approaches to geo-referenced data allow for geolocation of genotypes across continuous landscapes, and that the residuals from such predictions can be used to infer historical processes (Battey et al., 2020b; Rehmann et al., 2024). Further, spatial genetic data can be used to infer how organisms move across landscapes (Al-Asadi et al., 2019; Marcus et al., 2021; Ringbauer et al., 2017; Rousset, 1997;

C. C. Smith & Kern, 2023; C. C. Smith et al., 2023) as well as give information about local density of the population (Al-Asadi et al., 2019; C. C. Smith et al., 2024). While spatial data has been being utilized to inform genetic analysis of explicitly spatial processes, other fundamental phenomena, such as selective sweeps, are also impacted by spatial population structure. Given that selective sweeps in spatially structured populations behave differently from those in panmictic populations, most notably in their speed and the impact of that on linked neutral variation, this spatial dimension of how genetic variation is shared may contain information that reveals sweeping alleles.

Identifying selective sweeps is of particular importance in populations relevant to human health, such as the malaria-causing parasite *Plasmodium* and its vectors. The malaria vector *Anopheles* is under strong selective pressures from public health interventions (e.g., Bhatt et al., 2015), and thus is frequently reported as having experienced or or is currently undergoing selective sweeps in part due to these efforts (Anopheles gambiae 1000 Genomes Consortium, 2017; Clarkson, Miles, Harding, O'Reilly et al., 2021; Xue et al., 2021b). Genomic surveillance in these populations is crucial, as variants under positive selection are notably often in genes related to insecticide resistance, (Clarkson, Miles, Harding, O'Reilly et al., 2021; Hemingway et al., 2016), and thus relevant to advances in vector control and prevention of disease transmission. While that is so, the studies identifying these variants often rely on methods that may be confounded by spatial population dynamics, reducing their power to identify sweeping alleles. By using spatial information to inform searches for selective sweeps, we may be able to identify more (or different) variants under positive selection.

In this study, we use forward-in-time simulation to model selective sweeps across a continuous landscape and leverage the joint distribution between an allele's spatial spread and its frequency to reveal a novel signal of positive selection. We demonstrate this spatial signal's potential to reveal alleles under positive selection both through SNP-centric and windowed approaches, then apply these methods to a spatially referenced dataset of *Anopheles gambiae* genomes (Ag1000G Selection Atlas, 2017). Our analysis recovers known instances of positive selection in *An. gambiae*, identifying variants within known insecticide-resistant loci, and identifies potentially novel sources of insecticide resistance as well as other potentially selected variants associated with immune response, sensory perception, and development. Through simulation and empirical analysis, we demonstrate the potential of spatial population structure to identify, rather than obscure, segregating variants under positive selection.

3.2 Methods

3.2.1 Spatial simulation. To simulate selective sweeps in continuous space, we used a `SLiM v3` (Haller & Messer, 2019b) model as described in Battey et al., 2020 and elaborated on in Chevy et al., 2024. We model selection by having a selected allele modify mortality of its carriers rather than fecundity. Following Chevy et al., 2024, we define $f(u)$ as the per-capita mean number of offspring and $\mu(u)$ as the probability of mortality per unit time at local population density u, meaning the population reaches an equilibrium density when $f(u)$ and $\mu(u)$ are equal. A positively-selected allele has additive effects ($h = \frac{1}{2}$) such that an individual with k copies of the allele has a reduced mortality $\mu_k(u) = 1 - (1 - \mu(u))(1 + \frac{ks}{2})$. The positively-selected allele initially grows in copy number at a rate $G(u) = (1 - \mu(u))(1 + f(u))(1 + \frac{s}{2}) - 1$, and because individuals carrying the allele have reduced mortality $\mu_k(u)$, its spread

ultimately causes population density to increase (similar to what was observed in Chevy et al., 2024).

Individuals were simulated across a 25 $\times$ 25 unit square continuous landscape with an initial average population density of 5 individuals per unit square area. Over the course of a simulation time step, individuals mate with proximate individuals, newly born individuals disperse from their focal parent, then living individuals on the landscape compete locally for resources and space. Mating, dispersal, and competition distances were drawn from a Gaussian kernel with mean zero and SD σ, which was varied between 0.40, 1.25, and 4.0 to achieve initial Wright neighborhood sizes between 10, 100, and 1000 (though these values changed as the sweeping allele spread and population density increased).

Each individual had a diploid 10^8 bp genome that recombined at a rate of 10^{-8} per bp per generation. After an initial burn-in of 1000 time steps, a positively selected mutation with a selection coefficient chosen from $[1 \times 10^{-4}, 0.001, 0.01, 0.1]$ was added to the center of a chromosome from a newly born individual at the center of the landscape. As the simulation progressed, the frequency of the sweeping allele was tracked, alongside the landscape area occupied by individuals carrying the sweeping allele, calculated as the area of the convex hull containing all allele carriers (Barber et al., 1996). We ran 100 total simulations to fixation across each parameter combination; simulations where the allele did not fix were restarted from the time point just before the addition of the sweeping allele.

3.2.2 Power analysis. To assess the power of this spatial signal to identify variants under positive selection, we analyzed the spatial distribution of neutral variation across the genome over the course of our simulations. From each previously described simulation, we saved tree sequences over the course of the sweep (at a rate

of 1 tree per 1 time step for a selection coefficient of 0.1, 2 time steps for 0.01, 5 time steps for 0.001, and 10 time steps for 1×10^{-4}, respectively). Tree sequences were saved at different rates for different selection coefficients in order to capture the full dynamics of allele frequency while reducing computational time and memory load (Kelleher et al., 2018).

After simulating, we used `msprime` version 1.3.1 (Baumdicker et al., 2022) to recapitate the tree sequence, adding neutral mutations to the genome based on a population size equal to the population size at the tenth time step of the simulation (after which the population was spatially established), using a mutation rate of 10^{-8} per bp per generation. We then measured the frequency and area (through the same methods described in Methods Spatial simulation) of each variant across the genome, taking note of the sweeping allele. We used the joint distribution between allele frequency and occupied area of all variants across the genome to identify spatial-frequency (SF) outliers by binning every 1,000 variants (grouped by frequency), then classifying the lower Nth percentile in regards to area of these groups as spatial-frequency (SF) outliers (Figure 3.3 A). To assess our power to identify the sweeping allele based on the joint distribution between frequency and area, we tested for the presence of the sweeping allele among these SF outliers while varying the percentile cutoff between 0.5 and 0.01.

Additionally, we analyzed the ratio of SF outliers to non-outliers in 100kb windows along the genome (Figure 3.3 B). To prevent our analysis from being thrown off by windows without many variants, we used as our statistic a Z-score defined as

$$Z = \frac{\frac{O}{S} - Q}{\sqrt{Q \times \frac{(1-Q)}{S}}}$$

where O is the number of SF variants in a window, S is the total number of variants, and Q is the genome-wide proportion of SF outliers to all variants. Genomic windows

that had SF outlier:non-outlier ratios above the Nth percentile of genome-wide ratios were identified as windowed spatial-frequency (WSF) outliers while again varying the percentile cutoff between 0.5 and 0.01, and we again tested for whether the sweeping allele was identified as being within a WSF outlier window.

3.2.3 *Anopheles gambiae* analysis.

3.2.3.1 Data processing.

We applied our method to whole genome sequencing of *Anopheles gambiae* available through the Ag1000G project (The Anopheles gambiae 1000 Genomes Consortium et al., 2020). Genotype data and metadata were downloaded from the Ag1000G phase 3 (Ag3.0) data release (accessed April 2024; https://malariagen.github.io/vector-data/ag3/ag3.0.html). VCFs relating to 28 sampling sets were merged using `bcftools` v1.20 (Danecek et al., 2021) and the command 'merge' with default options. Metadata were filtered to identify *An. gambiae* individuals as defined in the metadata header: 'aim_species' and 'taxon'. The merged VCF was filtered to retain only those *An. gambiae* individuals, resulting in 1470 individuals for downstream analyses. Monomorphic sites were removed using `bcftools` v1.20 and command 'filter' with options -e 'AC==0 ||AC==AN'. Finally, the VCF was filtered to retain high quality sites and accessible regions of the genome using the site filters available through Ag3.0. The final VCF contained 83,924,645 polymorphic sites, where 28% were classified as multiallelic (>1 alt allele).

The allelic state was polarized using the program `est-sfs` v2.03 (Keightley & Jackson, 2018) with *An. coluzzii*, *An. arabiensis*, *An. melas*, and *An. merus* as outgroup species. *An. coluzzii* and *An. arabiensis* are part of the Ag3.0 data release, and their VCFs were downloaded and processed in the same methods as outlined above. These VCFs were also filtered to retain only individuals with an unambiguous

species identification. This left a total of 507 individuals of An. coluzzii and 368 individuals of *An. arabiensis*. The two other outgroup species, *An. melas* and *An. merus*, were downloaded from the available data sets of Fontaine et al., 2015. An input file for `est-sfs` was constructed using the custom script `estsfs_format.py` to partition the VCF into files with 100,000 sites for fast parallelization. The program est-sfs was then run with default parameters and the results integrated back into the *An. gambiae* VCF by adding an Ancestral Allele (AA) state to the INFO column of the VCF. The ancestral allele was determined to be the reference allele if the probability given by est-sfs was >0.90 and then the alternate allele if 1-probability was >0.90. Parsimony was used to polarize all other sites that could not be confidently polarized from the `est-sfs` results. Impacts of derived alleles were then annotated onto the VCF using `SNPEff` v5 and the AgamP4 reference genome (Assembly GCA_000005575.1) (Cingolani et al., 2012).

To reduce impacts of larger-scale population structure, we focused our analysis on 687 *An. gambiae* samples from West Africa grouped using `ADMIXTURE`. Ancestral clusters were estimated using `ADMIXTURE` (Alexander & Lange, 2011). A VCF of the genomic regions 3R:1-37,000,000 and 3L:15,000,000-41,000,000 (as noted in Anopheles gambiae 1000 Genomes Consortium, 2017) were filtered to retain SNPs with a minor allele frequency >5% and then thinned to retain only 1 site every 5,000bp. Thinning was done to reduce the influence of site dependency via linkage. The VCF was converted to bed format using `PLINK` v1.9 (Purcell et al., 2007). `ADMIXTURE` was run for a K from 1 to 5 (number of ancestral populations) with 5-fold cross-validation. Each `ADMIXTURE` analysis was repeated 30 times with different seeds. Final groups were visualized using `CLUMPAK` (Kopelman et al., 2015) and individuals were then

assigned to one of the three clusters, the optimal K being 3 using an admixture proportion cutoff of 0.88.

A recombination map was estimated using `ReLERNN` v1.0.0 (Adrion et al., 2020) and mosquito sequences collected from the Central African Republic (CAR) (see Ag3.0 data release). The CAR population was used because sampling among the other *An. gambiae* populations was variable in both time and space. In contrast, CAR had 55 female mosquitoes collected during the same year (1994) and location (Bangui). `ReLERNN` was run on the phased and filtered VCF containing mosquitoes from CAR and a mask file (MASK) denoting the non-accessible regions of the genome. A custom demographic history (DEMO) trajectory for CAR mosquito population was inferred using `StairwayPlot2` (Liu & Fu, 2020). The classifier was trained using a window size of 26,000bp with default options and setting required parameters: –assumedMu $3.0e-9$ –upperRhoThetaRatio 30 –phased –mask $MASK-n$DEMO. The resulting recombination maps were translated to cM/Mb and HapMap3 format using a custom script `rho2cMMb.py`.

3.2.3.2 Spatial genome scan. We applied our analysis to 687 *An. gambiae* samples from West Africa grouped using `ADMIXTURE`. We recorded the frequency of each variant globally within the sample population alongside the area of the landscape occupied by its carriers. The area of the landscape occupied by allele carriers was determined by fitting a convex hull over the sampling locations of individuals carrying the allele through the same methods described in Methods section Spatial simulation, then calculating the area (in square kilometers) of the polygon determined by this hull using `pyproj.Geod`. Alleles that were not present in at least three unique locations were discarded from our analysis as we were unable to determine the area occupied by these variants.

From the joint distribution of allele frequency and area, we identified SF outliers as those below a spline fit (using the R package `npreg`) through the lower tenth area percentile of every 1000 SNPs binned by frequency after removing SNPs within the three major inversions (2La, 2Rb, 2Rc). To characterize the "distance" of an individual SNP from this cutoff, we used a modified z transformation of each SNP's area compared to the distribution of areas within the defined 1000-SNP frequency bin, centering the areas according to the median of the distribution and then scaling by the median absolute deviation of the distribution. Such z scores give us a sense, although imperfect, of the strength of the outlier relationship. We then used a 100kb windowed analysis to identify WSF outliers beyond the 90th percentile of genomic z scores of SF outlier:non-outlier SNPs. This same approach was applied separately to SNPs within each the three major inversions (Figure B.3). To prevent regions with low variation from inflating genome-wide z score values, we masked windows with a recombination rate below 1.5 cM/Mb before placing the 90th percentile cutoff to identify WSF outlier windows.

To assess potential phenotypic impacts of SF and WSF outliers, we searched for enrichment of SNP annotations obtained using SNPEff (Cingolani et al., 2012) using a χ^2 test. Next, we identified genes containing SF and WSF outliers using VectorBase (Giraldo-Calderón et al., 2015), then used this list of genes to search for enrichment of functional annotations using either `PANTHER` (for SF outliers) or `DAVID` (for WSF and NS outliers) (Dennis et al., 2003; Mi et al., 2019; Thomas et al., 2022). Finally, for nonsynonymous WSF outliers, we obtained `.pdb` files of predicted protein structure for their associated genes using `AlphaFold3`, then used `ThermoMPNN` to assess how the associated amino acid substitution impacted protein stability (Dieckhaus et al., 2024a; Jumper et al., 2021). For *CYP4H27*, we used the pairwise structure alignment

tool available via www.rcsb.org/alignment with method `TM-align` to compare the ancestral and derived versions of the protein (Bittrich et al., 2024).

3.3 Results

3.3.1 A novel spatial signal of selection.

How does a beneficial allele move across space? An allele under positive selection that arises at a point along the landscape (position 50 in Figure 3.1) will increase in frequency where it arises while simultaneously spreading in all directions in the form of a traveling wave (R. A. Fisher, 1937; Kolmogorov et al., 1937). While the beneficial allele is globally increasing in frequency, the portion of the landscape occupied by carriers of said allele is constrained by the wavelike dynamics of its spread. In contrast, a neutral allele that arises and is not lost will diffuse in both directions, but generally remain at low or intermediate frequencies across the landscape, as its frequency dynamics are determined by drift (Figure 3.1). Because positive selection causes a beneficial allele to increase in frequency faster than it can spread through space, these alleles should occupy less space on the landscape than neutral variants at similar global frequencies. We set out to observe this phenomenon in continuous space using forward-in-time, individual-based simulation.

To simulate selective sweeps in continuous space, we used a non-Wright-Fisher framework similar to that described in Chevy et al., 2024, where individuals disperse from their parent across the landscape, compete locally for resources, and reproduce with a spatially proximate mate. Dispersal and interactions are governed by a Gaussian kernel with size σ, which represents the average per-generation movement rate, competition distance, and mate search radius (Methods Spatial simulation). On to this continuous space simulation we introduce a positively selected allele – one that reduces individual mortality – to the middle of the simulated genome of

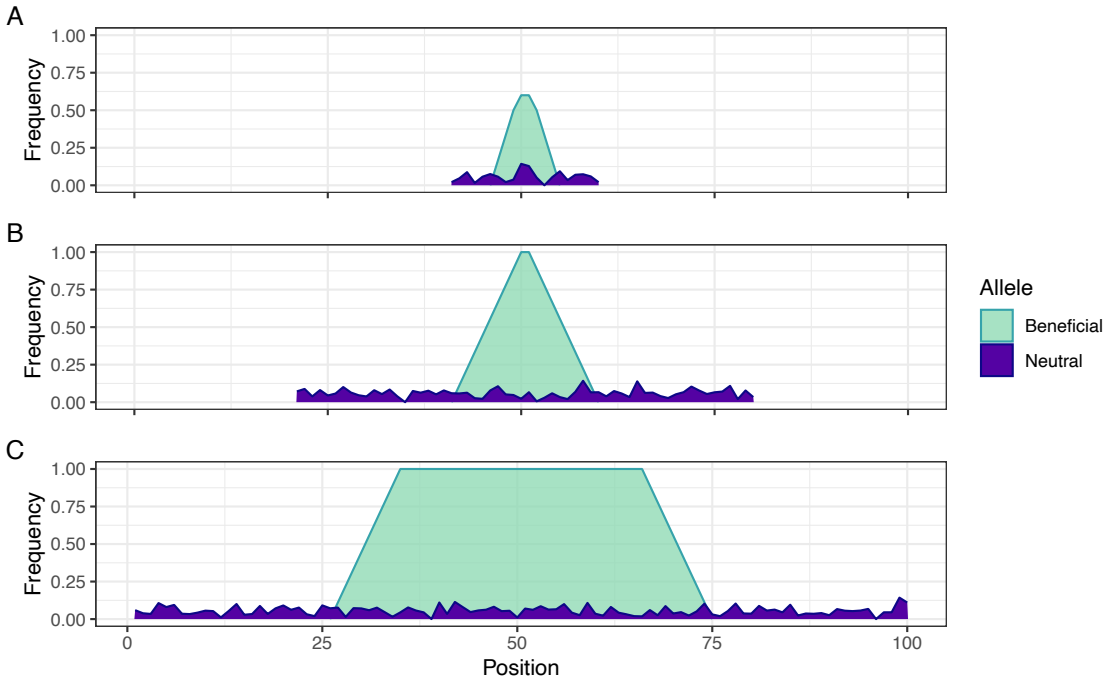


Figure 3.1. A conceptual model of how positive selection impacts an allele's spread across a one-dimensional landscape. Panels A, B, and C show a time series of a beneficial allele (in green) and a neutral allele (in blue) arising at position 50 and spreading through space. The frequency of each allele at a given spatial position is indicated along the y axis.

an individual at the center of the landscape, then track the progress of the focal mutation until eventual fixation or loss (Figure 3.2A). Additionally, we compare results of our simulations to simple theoretical deterministic expectations of the spread of a beneficial allele according to traveling wave dynamics (Supplemental section Approximation to Fisher's wave; R. A. Fisher, 1937; Kolmogorov et al., 1937). In our simulations, we tracked the frequency of the allele as well as the area of the landscape occupied by the allele's carriers, where area was defined by a convex hull formed at the vertices of allele carriers (see Methods).

In Figure 3.2B we show the allele frequency and area trajectories resulting from our spatial simulations. In accordance with the conceptual model shown in Figure 3.1, alleles under strong positive selection occupy less space on the landscape in comparison to neutral or weakly selected alleles at similar frequencies. There is a strong relationship between the strength of selection, the frequency of a beneficial allele, and its area: as the strength of selection increases, beneficial alleles cover *less* of the landscape conditional on frequency. This effect is most pronounced at smaller dispersal rates, where the allele's spread is inherently more limited by spatial population structure, but persists even when per-generation dispersal is high. Selection's effect on this joint distribution - reducing relative area - is initially counterintuitive as one might expect a sweeping allele to spread more quickly across a landscape than a neutral variant. This is true with respect to time, however conditional on its frequency, a neutral variant will diffuse across the landscape without a substantial increase in frequency. Under positive selection, an allele will both spread across the landscape and increase in frequency, reducing the expected area associated conditional on frequency and creating a visible signal in the joint distribution (Figure 3.2B). We note that deterministic models of Fisherian wave dynamics recapitulate

these dynamics as well (Figure B.1), thus drift is not responsible for the general effect.

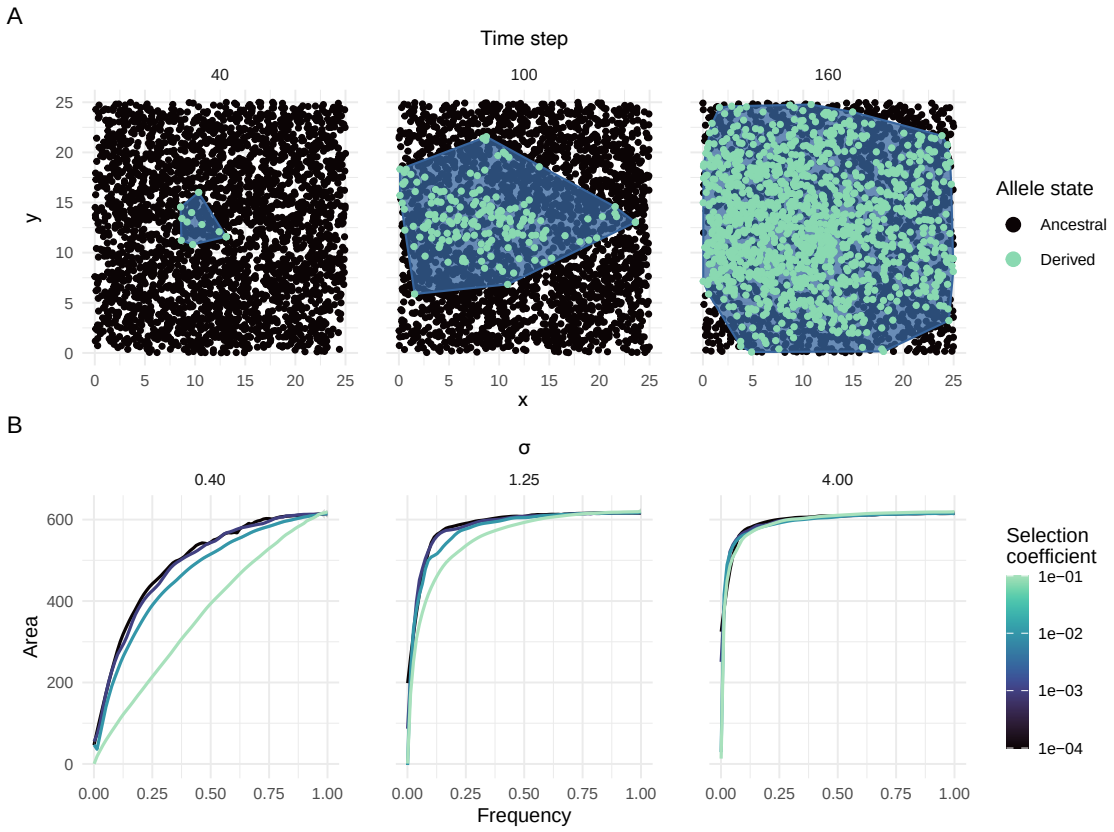


Figure 3.2. Selective sweeps in continuous space. (A) Example of simulated allele with selection coefficient 0.1 sweeping through a population with a dispersal distance (σ) of 0.40. Individuals carrying the sweeping allele are colored in green. (B) Joint frequency-area trajectories of sweeping alleles at dispersal distances 0.40, 1.25, and 4.0. Line color represents the strength of selection; allele area is measured as the area of the polygon encompassing allele carriers.

3.3.2 Power to detect selection using space. We next asked whether we could use this spatial signal to identify alleles under positive selection. Given that positive selection has a striking impact on allele area conditioned on frequency, we decided to use an empirical outlier approach on this joint distribution to localize candidate selected alleles. Simply put, the idea is to identify alleles that occupy

less area than expected given their frequency (spatial-frequency or SF outliers). An example of this joint distribution along with a heuristic cutoff to identify SF outliers is shown in Figure 3.3A. While this distribution is defined based on individual SNPs, sweeping alleles also bring along their linked genetic background, and there may be additional power in aggregating SF outlier signal across windows. To do so, we first identify genome-level SF outliers (as in Fig. 3.3A) in a genomic window, then define a z-score from the proportion of variants found within a given window that are SF outliers (see Power analysis section). The goal is to identify particular windows which harbor an excess of spatial-frequency outliers (WSF outliers, see Power analysis; Figure 3.3B).

We characterized power of the spatial-frequency (SF) outlier and windowed (WSF) outlier methods using forward-in-time spatial simulations for differing combinations of selection coefficients and dispersal rates (σ) while varying the stringency of the empirical cutoff used to define outliers. Unsurprisingly, we found that our method is best powered to detect the sweeping allele when the strength of selection is strong and dispersal is limited (Figs. 3.3C and 3.3D). Using an empirical cutoff at the 10^{th} percentile leads to a reasonable balance of power and false positive rate given the parameterization of our simulations (Figs. 3.3C and 3.3D), here observed as an inflection point in our power curves. We thus chose this cutoff for later empirical analysis.

Across all dispersal distances and selection coefficients, the probability of detecting the sweeping allele changes as a function of its frequency and increases sharply as the allele reaches fixation (Figure B.2). Because there are fewer high-frequency SNPs overall and the landscape area they occupy is inherently high, we become more likely to identify the sweeping allele as an outlier in the joint

distribution. Variation that is linked to this sweeping allele is also likely to take up less area relative to its frequency, contributing to power to detect the genomic window containing the sweeping allele.

While our method demonstrates potential to detect in-progress sweeps, its empirical approach also results in a number of false positives that we control based on the defined outlier cutoff. For example, in our simulations with a single SNP under selection, identifying the lower tenth percentile of the joint area-frequency distribution as SF outliers and the upper tenth percentile of windowed SF outlier ratios as WSF outliers resulted in roughly $20,000$ SNPs (10% of genomic variation) being identified as SF outliers and roughly 3400 SNPs (less than 2% of genomic variation) being identified as WSF outliers. In empirical applications, genomic/biological annotations can help deal with sorting through these SNPs.

3.3.3 Spatial signal identifies sweeping alleles in *Anopheles gambiae*. After demonstrating the potential of this spatial signal to identify positively selected variants, we next applied our approach to 687 West African *Anopheles gambiae* samples from the Ag1000G project (spatial sampling shown in Figure 3.4 inset; data described in The Anopheles gambiae 1000 Genomes Consortium et al., 2020). Because this species is under strong selective pressures from human interventions, this is a useful system for testing our method with potential consequences for vector control efforts. For each SNP in the genome, we recorded its global frequency along with the locations (in latitude and longitude) of individuals carrying that SNP. After discarding SNPs that were found in fewer than three unique locations (706,991 alleles or approximately 71.2% of genomic variation, leaving 297,305 SNPs used in our analysis), we estimated the area on the landscape that each SNP occupies as the convex hull encompassing all individuals carring the allele. Using

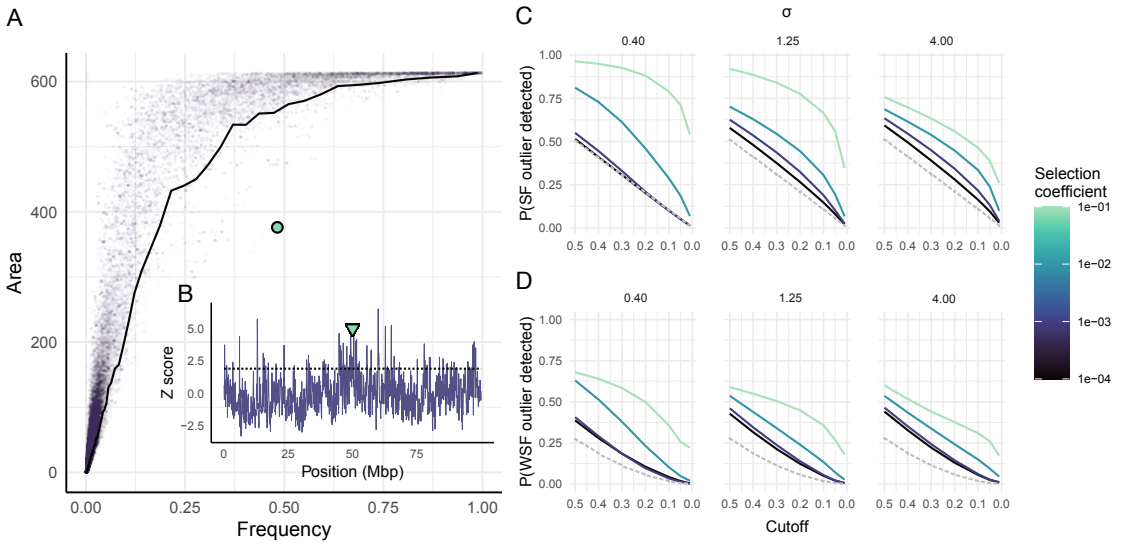


Figure 3.3. Power to detect selective sweeps using continuous space. (A) Example genome-wide joint distribution between frequency and area of all variants 195 time steps after the introduction of an allele with a selection coefficient of $s = 0.1$ in a population with a mean per-generation dispersal distance of 0.40. The black line represents the tenth percentile cutoff for SF outliers; the sweeping allele is highlighted in green. (B) Example genome-wide windowed analysis of the ratio of SF outliers to non-outliers at the same time point in the same simulation. The dotted line represents the tenth percentile cutoff for WSF outliers; the green triangle highlights the locus of the sweeping allele. (C) Probability of detecting the sweeping allele as a SF outlier as the strictness of the outlier cutoff increases along the x axis. Panels represent σ values 0.40, 1.25, and 4.0; line color represents the strength of selection. The dashed gray line represents the proportion of the genome identified as SF outliers. (D) Probability of detecting the sweeping allele as a WSF outlier in the same simulations; the dashed gray line represents the proportion of the genome identified as WSF outliers.

the resulting distribution of SNP frequency and area, we used an empirical 10% area cutoff, conditioned on frequency, to identify SF outlier SNPs (Figure 3.4).

As above, because positive selection reduces the area an allele occupies conditional on its frequency, we focused on SNPs in the joint distribution that match this pattern (Table S1). As linkage betweeen SNPs might lead to genetic hitchiking between neutral and non-neutral SNPs, we further refined our analysis using a genomic scan to identify 100 kb windows of the genome that contain an excess of SF outliers after masking low-recombination regions, labeling SF outliers within these regions as windowed SF outliers (WSF outliers, Figure 3.5). To reduce the impact of other non-neutral processes on our analysis, we identified outliers separately for the major inversions (2La, 2Rb, 2Rc) and the rest of the genome, respectively, then SF and WSF outliers were pooled for downstream analyses. In total, we identified 36,500 SNPs as SF outliers and 3,875 SNPs as WSF outliers. It's worth noting that our WSF outlier statistic agrees to some degree with previous searches for selective sweeps done in this system, and we find weak but significant correlations with windowed iHS statistics ($p = 4.658 \times 10^{-8}$; Anopheles gambiae 1000 Genomes Consortium, 2017) and windows identified using partialSHI/C ($p < 2.2 \times 10^{-16}$; Xue et al., 2021b), providing independent evidence that we are indeed picking up signals of positive selection.

To assess what biological signals were being identified through our spatial approach, we first investigated the SF outliers' annotations and associated genes. SF outliers were significantly enriched for coding SNPs (5.86% of SF outliers were located in coding regions compared to 5.31% of all SNPs retained in our analysis), with significant enrichments for both synonymous ($p = 7.196 \times 10^{-7}$, χ^2 test) and nonsynonymous variants($p = 0.046$, χ^2 test) (Figure 3.4). Furthermore, SF outlier SNPs that are significant outliers in the joint genome-wide distribution of frequency

and area fall within a number of loci associated with insecticide resistance, including an intronic variant within the *VGSC* locus (2L:2395567, $z = -6.66$; Martinez-Torres et al., 1998), 31 intronic variants and 6 synonymous variants within the *Cyp6p* locus (2R:28480576-28505816, z scores range from -0.82 to -7.71; Müller et al., 2008b), and 29 intronic variants within the *Rdl* locus (2L:25363652 - 25434556, z scores range from -0.66 to -13.21×10^{10}; Du et al., 2005). Beyond these loci, the genes containing SF outliers (both coding and noncoding) were significantly enriched after Bonferroni correction for a number of GO biological processes important for sensory processing, including action potential (2.84 times enrichment, $p = 1.64 \times 10^{-02}$), neuron recognition (2.63 times enrichment, $p = 1.92 \times 10^{-2}$), and axon guidance (2.38 times enrichment, $p = 1.21 \times 10^{-13}$) (Table S2). Sensory perception and processing is often implicated in mechanisms of insecticide resistance in *Anopheles* species, and these genes are likely under further selective pressures from other environmental factors and vector control measures.

We next turned to our windowed WSF outliers (Figure 3.5). We again tested for enrichment of SNP annotations within these WSF outliers and found a significant enrichment of upstream ($<$ 5kb from the transcription start site) gene variants ($p = 8.004 \times 10^{-6}$, χ^2 test) and intronic variants ($p = 1.015 \times 10^{-6}$, χ^2 test), which may indicate that a substantial portion of variants under positive selection are regulatory in this system. Further, ten WSF outliers fall within a well-known locus of insecticide resistance, Ace1 (2R:3484107-3495790, z scores range from -1.03 to -2.67). Genes containing WSF outliers were also significantly enriched for drug metabolism function (Enrichment score = 2.41, $p = 0.0062$) and enriched, though not significantly, for other relevant functions and annotations, including CRAL-TRIO domain-containing proteins, which are associated with eyesight (Enrichment score = 1.64, $p = 0.47$;

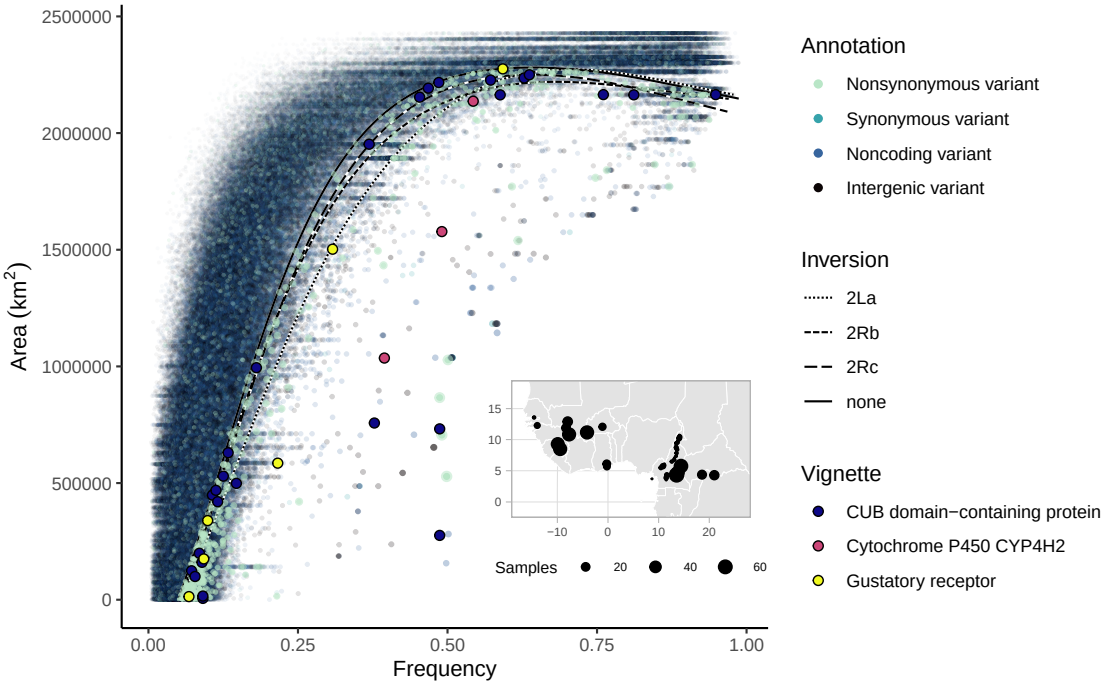


Figure 3.4. Joint distribution between frequency and area of all variants within the *Anopheles gambiae* genome. Each point represents a genomic variant; the point's color indicates its annotation. Outlier variants highlighted in vignettes are colored and outlined. Black lines represent SF outlier cutoffs for each of the major inversions and the non-inverted portion of the genome, respectively. INSET: Sampling locations of *An. gambiae* samples. The size of the points represents the number of samples at each location.

G. Smith and Briscoe, 2015) and G protein-coupled receptors (Enrichment score = 1.78, $p = 0.96$, Bonferroni correction) (Table S3).

Finally, we searched for nonsynonymous outlier SNPs that could potentially be associated with functional changes. Investigating the functional consequences of nonsynonymous outlier SNPs, we found 37 WSF outliers and 237 SF outliers that impact protein stability (274 in total out of the 526 nonsynonymous outliers identified; Table S4). Among these are 44 variants within genes potentially associated with insecticide resistance as defined by the Ag1000g Selection Atlas, including methylenetetrahydrofolate reductase (AGAP007479:Y174D, $z = -0.62$), heme peroxidase 7 (AGAP004036:E564D, $z = -0.86$), and cytochrome P450 enzymes CYP6N2 (AGAP008206:K392N, $z = -0.76$), CYP4D15 (AGAP002418:L80M, $z = -0.88$), and CYP4H27 (AGAP008552, highlighted below) (Ag1000G Selection Atlas, 2017). The largest stability changes were observed within genes coding for metabolic proteins, which may be indicative of selective pressures exerted by insecticides, which target metabolic processes. High-magnitude stability changes are also observed within genes involved in sensory function, which may be due to changes in behavior associated with insecticide resistance (Thomsen et al., 2016), and genes involved in the regulation of gene expression. Changes in gene expression are often the mechanism behind insecticide resistance in *Anopheles* and other mosquitoes, and alterations in regulatory proteins may contribute to resistant phenotypes (Müller et al., 2008b).

While there are a number of individual loci that we could highlight, in what follows we provide three particularly interesting vignettes of loci that represent strong spatial-frequency outliers in our scan and may be of functional consequence (see Figure 3.6).

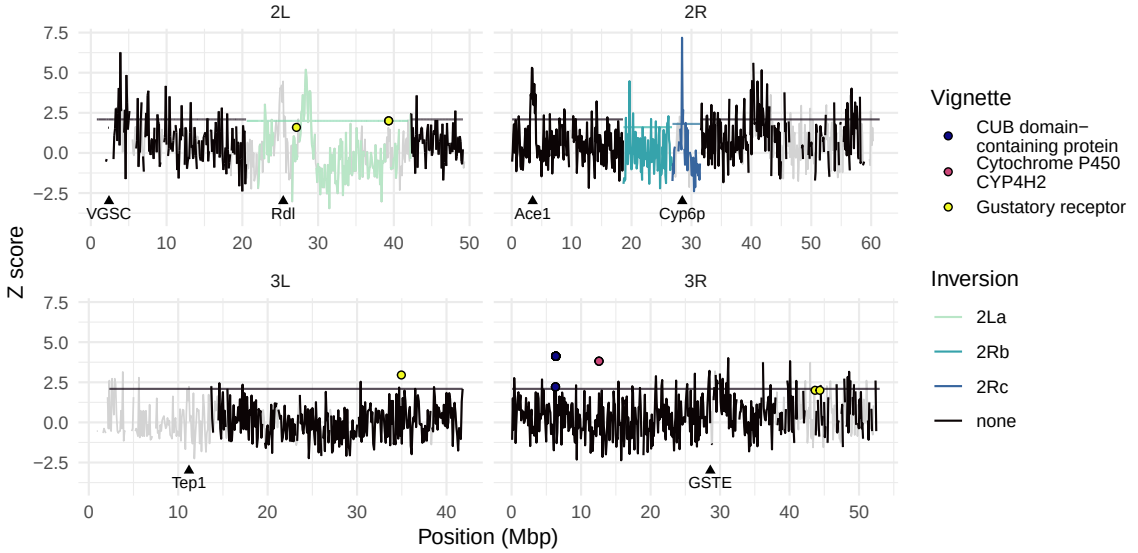


Figure 3.5. Genome-wide windowed analysis of the ratio of SF outliers to non-outliers. Flat horizontal lines represent WSF cutoffs for each of the major inversions and the non-inverted portion of the genome, respectively. Known loci associated with insecticide resistance and outlier variants highlighted in vignettes are labeled.

3.3.3.1 Cytochrome P450 CYP4H27. Cytochromes P450 are frequently implicated in the evolution of insecticide resistance across multiple taxa. In *Anopheline* mosquitoes, positive selection linked to Cyp450-associated resistance is most often associated with increased protein expression level due to copy number or transcriptional changes, rather than changes in protein function (Müller et al., 2008b). However, using our SF outlier approach, we identify three nonsynonymous variants within one cytochrome P450 *CYP4H27* (Gene ID AGAP008552) that impact the folding stability and binding site conformation of this protein (Figure 3.6A). The first variant of interest, 3R:12573245T>C, is found only in five sampling locations, but is at a global frequency of almost 40%, and thus represents a highly significant outlier from the genome-wide joint distribution, the strength of which we can characterize with a z-score ($z = -12.61$). This mutation results in a shift from tyrosine to histidine at amino acid position 274 (Y274H) within the G helix of the CYP4H27 protein. The

substitution shifts the orientation of the S326 residue on the N-terminal end of the I helix within the protein's active site, allowing it to join a helix-stabilizing hydrogen bond between the residues of E327 and S324 and the carboxyl group of E27 and reducing the volume of the ligand binding pocket from 630.271 $Å^3$ to 595.784 $Å^3$ as modeled by AlphaFold2 predictions (Abramson et al., 2024; Ye et al., 2024). Given that positively selected Cyp450 variants tend to be associated with increased protein expression that aids with insecticide detoxification, this modification to the structure of the CYP4H27 active site may have implications for its enzymatic activity, however further investigation is required to confirm this.

The other two variants of interest, 3R:12573798A>C and 3R:12573814T>A, are at 54% and 49% frequency ($z = -2.62$ and $z = -10.36$), respectively, and both result in charged amino acid changes between the L and M helices: 3R:12573798A>C causes a negatively charged serine at position 458 to change to a positively charged arginine (S458R), and 3R:12573814T>A causes a nonpolar valine at position 463 to change to a negatively charged aspartic acid (V463D). Among the individuals analyzed, we identified three mosquitoes with both the Y274H and S458R mutations segregating on the same haplotype and three mosquitoes with both Y274H and V463D segregating together, though we were unable to definitively find one haplotype with all three variants. AlphaFold predictions of the S458R and V463D carrying structures suggest these amino acid substitutions occur at the entrance of the ligand binding pocket, shifting its conformation, and that these changes are amplified when in combination with the Y274H mutation. Though the phenotypic effects of these three Cyp450 variants needs to be further investigated *in vivo*, their strong signals of positive selection may indicate another Cyp450-mediated mechanism of insecticide resistance.

3.3.3.2 *CUB domain-containing protein.*

A second interesting vignette from our analysis of outlier SNPs was AGAP029542, a relatively undescribed CUB domain-containing protein located on chromosome arm 3R (Figure 3.6B). While this gene's function is unknown in *An. gambiae*, its ortholog in *Anopheles stephensi*, XP_035915664.1, has recently been implicated in contributing to insecticide resistance (J. Kumar et al., 2024). Our analysis identified 27 WSF outlier variants within this gene, exclusively in noncoding regions. These variants are some of the strongest outliers in the joint distribution between frequency and area (individual z scores range from -0.60 to -30.25 with 21 SNPs having z scores below -1.0; Figure 3.4), and we positively identify as many as 13 variants segregating within the same individual. While we cannot directly associate these variants with phenotypic changes, we speculate that positively selected variants in this gene might be associated with changes in transcription. Noncoding variants are commonly identified in selection scans in *Anopheles* and are understood to impact transcriptional regulation, a key feature in known mechanisms of insecticide resistance (Müller et al., 2008b). Further, the functional domains of this protein have been shown to be highly conserved across *Anopheles* species (J. Kumar et al., 2024), making noncoding regions a more likely candidate for positive selection. Though more investigation is necessary to uncover the potential role of AGAP029542 in *An. gambiae* insecticide resistance, the density of outlier variants within AGAP029542 uncovered using our analysis highlights the strength of selection scans for generating hypotheses.

3.3.3.3 *Gustatory receptors.*

Chemosensory genes play a critical role in regulating insect behavior, and their evolution is implicated in broad phylogenetic as well as population-level adaptation such as avoidance or preference) of compounds such as those derived from plant secondary metabolites (e.g., Matsuo et al.,

2007). Beyond identifying an enrichment of sensory genes amongst SF and WSF outliers, we identified six nonsynonymous mutations that were inferred to impact the stability of five gustatory receptors (Figure 3.6C). These proteins are cell-surface GPCRs, consisting of an extracellular N-terminus and ligand-interacting domain, seven transmembrane alpha helices, and an intracellular N-terminus and G protein-interacting domain. Two mutations we identified, Gr32:M251L ($z = -1.79$) and Gr9:D33Y ($z = -0.98$), are both located in the extracellular ligand-interacting domains of their respective proteins and result in nonconservative amino acid substitutions (shifting from sulfur-containing methionine to aliphatic leucine and acidic aspartate to aromatic tyrosine, respectively). Another outlier mutation within Gr32, Gr32:V195M ($z = -2.63$), is located proximate an extracellular ligand-interacting domain and results in a substitution from aliphatic valine to sulfur-containing methionine. These substitutions of amino acids with different biochemical properties in their extracellular domains may impact the receptor functions of these chemosensory proteins.

The other three nonsynonymous mutations we identified are more proximate to the intracellular domain, with two (Gr56:G147V, $z = -0.76$ and Gr4:A174S, $z = -0.41$) located on the intracellular ends of transbmembrane helices and one (Gr35:V446I, $z = -1.17$) located on the C terminal end of the protein. The mutations within Gr56 and Gr35 are both conservative substitutions of aliphatic amino acids to slightly larger ones and are predicted to cause small stability increases. Interestingly, the substitution of alanine with serine in Gr4 introduces a sulfur-containing amino acid to the intracellular region that destabilizes the protein, however this variant is observed at relatively high frequency (59.3%) and area (2.28e6 km^2), indicating that this outlier variant is potentially under weaker selection than the others. While these

gustatory receptors may be evolving in *An. gambiae* due to insecticide-mediated pressures, there are also broad biological reasons for chemosensory proteins to be evolving on a macroevolutionary scale within *Anopheline* mosquitoes.

3.4 Discussion

Tests for positive selection are a common tool in population genetic analyses - however, the signals associated with positive selection, such as shifts in the site frequency spectrum and levels of observed heterozygosity, are often also impacted by spatial population structure (Battey et al., 2020; Chotai et al., 2024). In this paper, we leverage such spatial structure to identify variants under positive selection using the joint distribution between allele frequency and area as our signal. Through simulation, we show how the increase in frequency associated with positive selection impacts an allele's spread through space relative to its neutral counterparts, causing a striking shift in the expected frequency-area distribution. We then use this observation to develop a method for identifying sweeping alleles by identifying outliers from the genomic distribution of variants' frequency and area. After demonstrating our method's power to identify sweeping alleles in simulated datasets, we apply our approach to samples from the Ag1000G dataset to discover candidate variants undergoing positive selection. By considering spatial population structure, we uncover previously undescribed, epidemiologically relevant variants that display spatial signals of positive selection, demonstrating the potential for spatial analysis in uncovering evolutionary processes.

Our approach is based on the observation that, conditioning on frequency, variants that are positively selected take up less area on a continuous landscape than neutral variants. This is due to the fact that while positively selected alleles spread through a population more quickly than neutral ones, their spatial extent

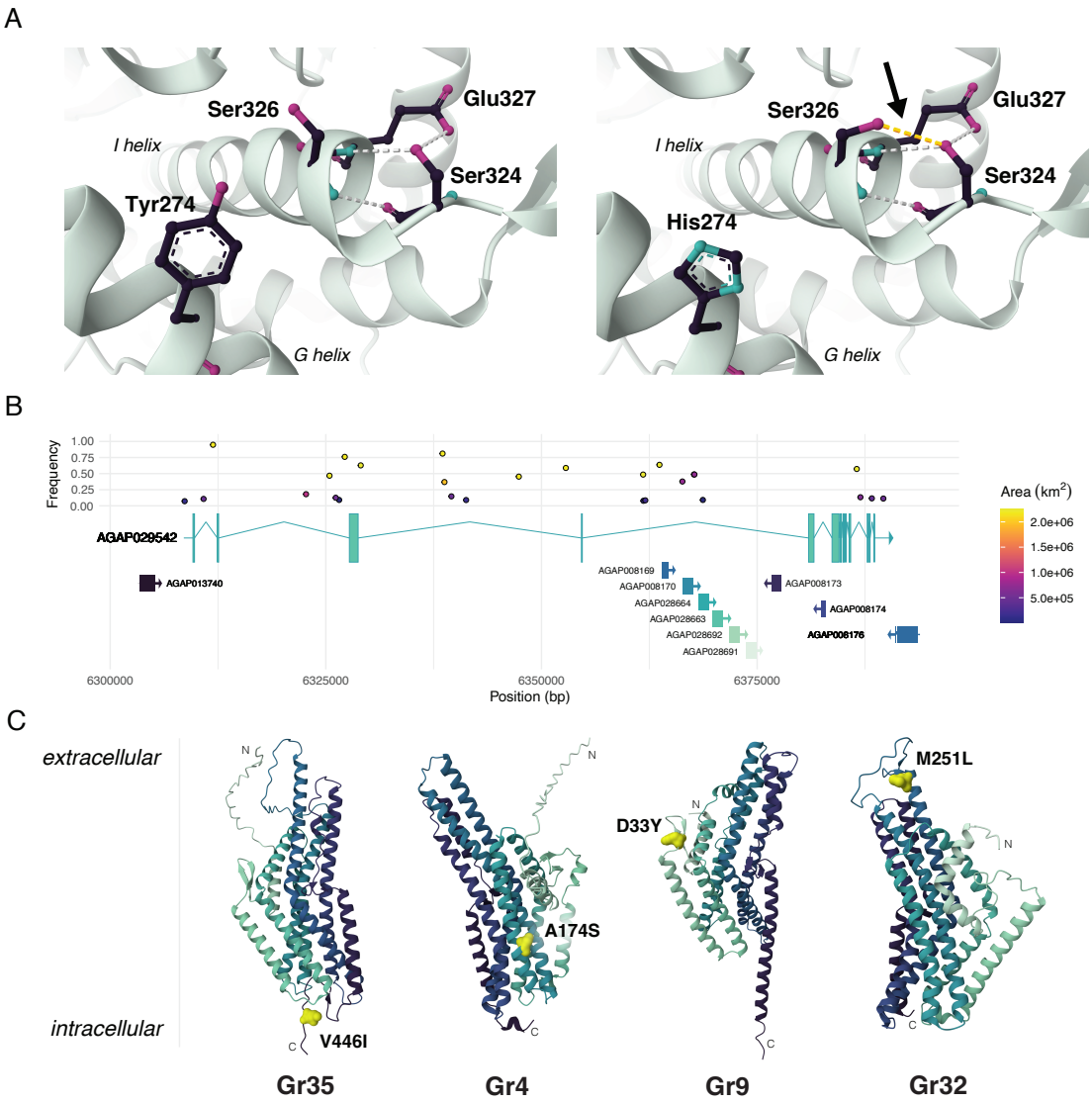


Figure 3.6. Impacts of outlier variants highlighted in vignettes. (A) Folding impacts of cytochrome P450 variant CYP4H27:Y274H. The left and right panels show position 274 and local structure for the ancestral tyrosine and the derived histidine, respectively. Relevant amino acids (S324, S326, E327) and hydrogen bonds are shown and the new hydrogen bond is highlighted in yellow. (B) Intronic outlier SNPs at the CUB domain-containing protein locus. Variants are shown at their genomic locations and colored by area; their y-axis position represents their frequency. Coding sequences and intronic segments for genes at the locus are annotated below; arrows represent the direction of transcription. (C) Structural locations of nonsynonymous variants in gustatory receptors. GRs are oriented with their extracellular domain upward and intracellular domain downward; chains are colored from light green at the N terminus to dark green at the C terminus and nonsynonymous mutations are highlighted in yellow.

is constrained relative to their increasing frequency. In contrast, neutral alleles that are not lost can be expected to diffuse across the landscape while remaining at low frequency, creating a dramatic difference between the joint frequency-area trajectory of selected and neutral variants. This effect dissipates as selection and spatial structure get weaker: for the former, this is due to the sweeping allele being less likely to increase in frequency, and for the latter, the spatial spread of the sweeping allele is less restricted. While we modeled this frequency-area signal in simulations of a globally beneficial allele undergoing a selective sweep, there are other evolutionary phenomena that could lead to variants occupying a small area while being at high frequency. In particular, local adaptation would produce similar signals due to a beneficial allele being geographically restricted. The salient difference here is that an allele, rather than being globally beneficial across the landscape, may only be beneficial over a subset of the range, however from the standpoint of our frequency-area distribution calculated at a static point in time the two may be indistinguishable. This raises the possibility that some fraction of the candidate alleles which we have identified as outliers in the Ag1000G dataset may represent local adaptation, rather than in-progress global sweeps.

Our approach is in some sense analogous to earlier work which sought to contrast the frequency of an allele with its age. For instance, Slatkin and Bertorelle, 2001 compared allele frequency to linked variation to identify the signature of selection, as neutral alleles at higher frequencies should be linked to relatively more polymorphism. Similarly, haplotype scan statistics aimed at detecting selection such as EHH (Sabeti et al., 2002) or iHS (Voight et al., 2006) use as signal haplotypes that are long relative to their frequency because recombination has had little time to break apart sequence surrounding beneficial alleles that are pushed to high frequency by selection.

Here, rather than search for alleles that are young relative to their frequency, we are searching for alleles that are spatially restricted relative to their frequency. Our method is conceptually similar, but focused on spatial rather than temporal information.

At its most basic, the methodology we use here is a form of outlier detection, which comes with it a number of well described issues (e.g. Smiti, 2020 or in the more narrow context of genetics (Teshima et al., 2006). Our simulations show that our method is relatively well-powered; by focusing on the lower tenth percentile of variants in the joint frequency-area distribution (identified as SF outliers), we are able to successfully identify a sweeping allele up to 79% of the time in highly structured populations. However we expect there to be a concomitant high rate of false positives – 10 percent of genomic variation is always identified as outliers. We further refined our spatial analysis by taking advantage of the fact that sweeping variants bring along linked neutral alleles. Using a window-based z-score to identify genomic regions with an excess of SF outliers, we identified the sweeping allele up to 35% of the time at a selection coefficient of $s = 0.1$ while reducing the number of false positives to roughly 1.7% of genomic variants. Unsurprisingly, in simulation both approaches were most successful at identifying a sweeping allele when both spatial structure and selection pressure were strong, as this is when an allele's spread is most constrained relative to its increasing frequency. The windowed approach reduces the number of outliers identified, making downstream analysis simpler, but spatial outliers that are not within outlier windows may still be under positive selection. While we found that a tenth percentile cutoff was the most effective at minimizing both the false negative and false positive rates when we applied our method to simulated data, our approach is constrained by the trade-off between the two. This is however not unique to spatial

scans for selection, and high false negative and false positive rates occur in empirical applications of traditional selection scans, even under relatively simple demographic models (Teshima et al., 2006). Additionally, while our simulations used uniform recombination maps, we found in our *An. gambiae* analysis that regions with low recombination are prone to inflated z-scores due to there being fewer variants overall identified in a window. We removed these segments from our windowed analysis before setting the WSF outlier cutoff, however many of these masked regions were still found to have relevant SF outlier SNPs.

Ultimately the strength of a method for finding selection rests in its application. We applied our approach to 687 samples of the malaria vector *Anopheles gambiae* from West Africa in order to identify variants under positive selection throughout this range (The Anopheles gambiae 1000 Genomes Consortium et al., 2020). Anopheles mosquitoes in West Africa are known to be under strong selective pressure due to human intervention, and consequently genes that are insecticide targets, as well as genes involved in metabolism, behavior, and cuticle formation are often identified as undergoing or having undergone selective sweeps (e.g. Clarkson, Miles, Harding, O'Reilly et al., 2021; Xue et al., 2021b). Further, surveillance of currently sweeping variants and discovery of novel ones is crucial for maintaining successful vector control efforts for this critical public health threat. By considering spatial population structure, our method uncovers a number of epidemiologically relevant variants that display signals of positive selection and have not yet been noted by conventional selection scans.

The suite of *An. gambiae* genes containing SF and WSF outliers is enriched for genes involving sensory processing - a mechanism implicated in behavioral responses to insecticide pressures - and insecticide metabolism. Further, we identify

outliers in a number of IR-associated loci, including *Rdl*, *Ace1*, and *Cyp6p*, and we highlight three outlier signals of particular interest. The first is three structure-altering nonsynonymous variants in the cytochrome P450 *CYP4H27*, a member of a protein family associated with insecticide resistance. Though we were unable to link all three variants to one haplotype, the variant displaying the strongest signal of positive selection, CYP4H27:Y274H, has a considerable impact on the shape and size of the protein's binding pocket when modeled using Alphafold3 (Abramson et al., 2024). The second vignette of interest occurs in a CUB-domain containing protein that has 27 WSF outlier variants within it. While the protein is undescribed in *An. gambiae*, it is orthologus to an IR-associated protein in *An. stephensi* (J. Kumar et al., 2024). All of these variants are noncoding, but are strong outliers in the joint frequency-area distribution, underscoring the potential importance of noncoding variants to adaptation in this system. The last vignette we highlight focuses on six stability-altering nonsynonymous variants in gustatory receptors. These variants tend to be radical amino acid substitutions and are primarily located on either extracellular receptor domains or intracellular protein-binding domains of these receptors, indicating potential changes to both ligand binding and cell signaling. *Anopheles* GRs experience selection pressures beyond those caused by human vector control measures, and this signal demonstrates how our method can identify positive selection across a wide range of potential selective pressures and functional consequences. These findings generate new hypotheses for potential avenues of adaptation it *An. gambiae* and demonstrate the potential of this method in identifying otherwise-overlooked sweeping alleles.

A major consideration in benchmarking the power of our method when applied to real populations is how appropriate the simulation was. This spatial signal assumes

a population where dispersal is continuous across the landscape - i.e., there are no major barriers to dispersal and every individual is equally likely to disperse in every direction. Further we assumed that selection pressure is homogeneous across the landscape as well. Heterogeneity of dispersal and selection may have major impacts on how an allele spreads across a landscape, and should be taken into consideration when applying our method to real populations. As mentioned earlier, other types of positive selection could generate this signal, such as local adaptation, which would result in a high-frequency allele occupying a geographically restricted area. Further, our power analysis assumes equal, dense sampling coverage across the landscape, which is unrealistic for real populations. Though we recover signal of selection in the unevenly sampled *An. gambiae* dataset, consideration of the spatial distribution of samples is also necessary for empirical applications of this method, and further simulation-based analysis should include exploration of different spatial sampling schemes, as uneven sampling has been shown to bias other spatial-genetic analyses (Rehmann et al., 2024). Finally, as previously discussed, our approach strikes a balance between minimizing false negatives and false positives, and is likely to both overlook and falsely identify alleles as being under positive selection (Teshima et al., 2006). For the latter case, genome annotation in empirical applications can help in identifying biologically relevant signals.

Here, we focus on the dynamics of beneficial alleles in continuous space; more generally, spatially informed genetic analyses have great potential to pick up on signals obscured by population structure in traditional analyses. While further inquiry is necessary into how the spatial impacts of other forms of selection, such as local adaptation, impact the frequency-area trajectory of an allele, we anticipate that data such as historical samples or ecologically relevant climatic information could help

differentiate between different forms of adaptation. Generally we see great potential for integrating spatial information into genetic scans for selection; as the genetic history of a population unfolds in time and space, it is increasingly possible to ask questions about the geographic origins of genetic variants and to unravel the spatial history of adaptation.

3.5 Bridge

As presented in Chapter III, spatial structure can be used to uncover adaptive variation at the population level, and indeed we successfully identify several candidate sweeping alleles in *Anopheles gambiae*. However, *Anopheles gambiae* is not the only malaria vector, and many *Anopheles* species are segregating large amounts of potentially-adaptive variation. In Chapter IV, I use a comparative approach to investigate the evolutionary forces shaping patterns of adaptive variation and divergence in three major *Anopheles* malaria vectors. This work reveals common selection pressures acting on these deeply-diverged vectors, with immune and metabolic pathways being particularly affected across a broad range of evolutionary timescales, and highlights how local adaptation and diversifying selection maintain genetic variation in the same pathways.

CHAPTER IV

CONVERGENT SELECTION ON NONSYNONYMOUS VARIATION ACROSS DEEPLY DIVERGED *ANOPHELES* VECTORS

This chapter is in preparation for submission to the journal *PLOS Genetics* and is coauthored with Andrew D. Kern. I conceptualized this study together with A.D.K. and performed all analyses; I wrote the manuscript with input and editorial assistance from A.D.K.

4.1 Introduction

The *Anopheles* mosquitoes that transmit human malaria are among the most medically important insects on earth, yet they remain remarkably difficult to control. The genus is deeply diverged: new world *Anopheles* split from old world lineages approximately 100 million years ago (MYA), and more recent radiations across Africa and Asia have given rise to a number of major vector species and species complexes. Recent population genomic studies have begun to characterize the genetic diversity and evolutionary histories of these species (Boddé et al., 2025; Neafsey et al., 2015; St. Laurent et al., 2022; Tennessen et al., 2025; The *Anopheles gambiae* 1000 Genomes Consortium, 2017), revealing a consistent pattern: *Anopheles* vectors are highly adaptable and readily evolve to counter vector control efforts (Boddé et al., 2025; The *Anopheles gambiae* 1000 Genomes Consortium, 2017). Understanding how these mosquitoes adapt so rapidly requires understanding the evolutionary forces that shape patterns of genetic variation across the genus.

Comparative genomic studies across the genus have revealed a striking capacity for genomic change: *Anopheles* genomes exhibit faster rates of gene gain and loss, elevated gene shuffling on the X chromosome, and rapid diversification of genes involved in chemosensation, immunity, and insecticide detoxification (Neafsey et al.,

2015). These patterns point to a genus whose genomes are unusually dynamic, but they describe evolution between species and leave open the question of what is happening within them.

While phylogenetic study of *Anopheles* evolution has revealed the genus's evolvability and highlighted the rapid rate at which their genomes adapt, population genetics studies suggest that this evolvability may be due in part to exceptionally high levels of segregating polymorphism. In the African major vector *Anopheles gambiae*, individual mosquitoes have 1.7 to 2.7 million variant sites and 21% of variants identified are triallelic; in another major African vector *An. funestus*, individuals have 2 to 3 million variant sites and 15% of variants have more than two alleles (Boddé et al., 2025; The *Anopheles gambiae* 1000 Genomes Consortium, 2017). Similar levels of diversity and proportions of multiallelic sites have been reported in *Anopheles minimus*, a major vector across Southeast Asia, and *An. darlingi*, the primary malaria vector in South America (St. Laurent et al., 2022; Tennessen et al., 2025). Despite high amounts of global polymorphism, these population genetic surveys have identified many instances of selection acting on local populations, particularly regarding adaptation to insecticides (Barnes et al., 2017; Clarkson, Miles, Harding, O'Reilly et al., 2021; Xue et al., 2021a), and indeed while population structure in these species allows for local adaptation, strong enough selection pressures can lead to the rapid spread of adaptive variants such as insecticide resistance at the *Vgsc* locus (The *Anopheles gambiae* 1000 Genomes Consortium, 2017). This raises the question of what forces maintain polymorphism in these populations while also allowing for rapid adaptation to local environmental pressures. In most systems, nonsynonymous polymorphism is assumed to be mildly deleterious, maintained by drift and purged inefficiently in small populations. *Anopheles* mosquitoes, however, have massive

effective population sizes spread across heterogeneous environments with differential exposure to vector control measures across space. We would therefore expect selection to act efficiently in these species, both fixing beneficial alleles and removing deleterious ones. Despite this, *Anopheles* populations continue to segregate large amounts of nonsynonymous variation, suggesting that some of this polymorphism may not be deleterious but instead actively maintained by balancing selection or local adaptation.

What is missing is a comparative, gene-level understanding of nonsynonymous variation across *Anopheles* species — which genes are segregating functional polymorphism, how that polymorphism relates to directional selection along lineages, and whether similar selective pressures - in particular, vector control efforts - are producing convergent patterns across deeply diverged vectors. Here, we address these questions using two African species (*An. gambiae* and *An. funestus*) and one Asian species (*An. minimus*) (Figure 4.1A). The lineage leading to *An. gambiae* diverged 30–50 MYA and has more recently radiated into the *gambiae* species complex, while *An. funestus* and *An. minimus* are more closely related, diverging approximately 22 MYA (Figure 4.1B) (Neafsey et al., 2015). *An. gambiae* and *An. funestus* occupy similar ranges in sub-Saharan Africa and are both highly anthropophilic; they differ in larval site preference but are jointly responsible for the majority of the continent's malaria burden, transmitting primarily *Plasmodium falciparum*. *An. minimus*, by contrast, is a major vector in the Greater Mekong Subregion of Southeast Asia with variable behavior and habitat preference; it transmits primarily *Plasmodium vivax* (Manguin, 2013). All three species occupy heterogeneous environments with variable insecticide exposure across both space and time. Investigating both population genetic dynamics and patterns of divergence, we identify parallel and convergent selection pressures that act on these highly-diverged

vectors and shape their evolutionary trajectories. We find that genes involved in immune response, insecticide resistance, redox metabolism, cuticle formation, and sensory perception both harbor excess nonsynonymous polymorphism within species and are evolving under directional selection along lineages. In *An. gambiae* and *An. funestus*, we further identify convergent spatial patterns of nonsynonymous differentiation at two cytochrome P450 loci involved in insecticide resistance, with parallel molecular changes in both species. Together, these results suggest that local adaptation maintains a reservoir of functional nonsynonymous variation in *Anopheles*, providing substrate for rapid evolutionary responses when selection pressures intensify or shift.

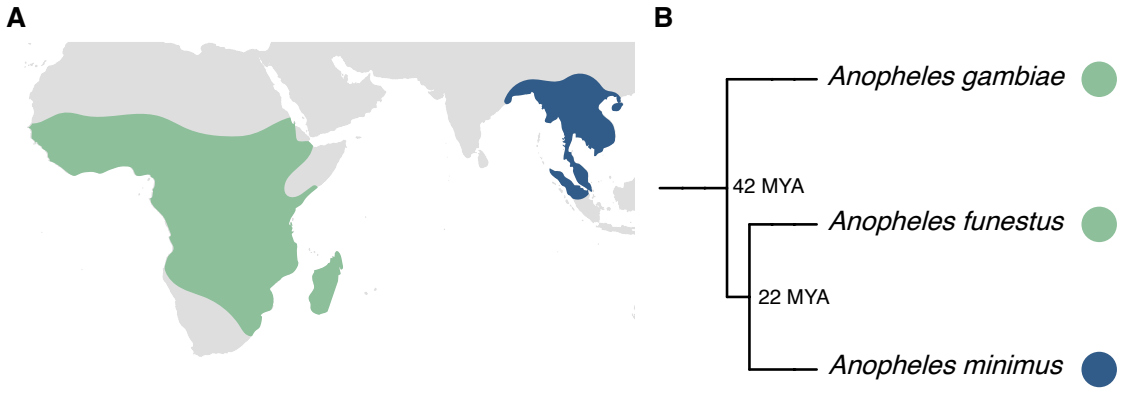


Figure 4.1. (A) Geographic distribution and (B) phylogeny of *Anopheles gambiae*, *funestus*, and *minimus*.

4.2 Methods

4.2.1 Genomic data & processing. *An. gambiae, funestus,* and *minimus* population genetic datasets were obtained from the MalariaGen Vector Observatory. All population VCFs were merged and then filtered (in the case of *An. gambiae* and *funestus*) using the provided site filters. Genome annotations were obtained in GFF format from VectorBase and processed using `gffutils 0.13`; reference genome and

proteome sequences were obtained in Fasta format from VectorBase. For all genes, the longest transcript was used for coding sequence analysis.

Prior to comparing patterns of nonsynonymous polymorphism and fixation between species, we identified one-to-one orthologs between *An. gambiae, funestus*, and *minimus* using `SonicParanoid 2.1.9` and aligned the protein sequence of all samples for a given ortholog using `MUSCLE 5.3`. Aligned amino acid sequences were mapped back to aligned nucleotide coding sequences using `Bio.Align.CodonAligner` from `BioPython 1.85`.

4.2.2 Population genetics methods.

4.2.2.1 Within species. For every protein-coding gene in each species, we calculated within-species population genetics statistics considering the following sequences: the entire length of the gene, the coding sequence, the amino acid sequence, nonsynonymous sites (N), and synonymous sites (S). Tajima's D, nucleotide diversity (π), and Weir and Cockerham's F_{ST} were calculated using a custom Python library, `WGSstats`, based on functions available in `scikit-allel` and `BioPython` (available at github URL). For *An. gambiae* and *An. funestus*, statistics except for F_{ST} were calculated on spatial cohort groups predetermined by the MalariaGEN data release (Boddé et al., 2025; The *Anopheles gambiae* 1000 Genomes Consortium, 2017) and then averaged across cohorts in order to minimize the impact of rare variants; for *An. minimus* all samples were used at once. To calculate F_{ST} within species, we used the `HBDSCAN` algorithm implemented through `scikit-learn` spatially cluster sampling locations (with a minimum cluster size of five samples) and then calculated Weir and Cockerham's F_{ST} between all clusters for a given gene. To assess the relationship between nonsynonymous and genic statistics, we fit linear models using the `stats` R package with the formula $X_N \sim 0 + X_G$ where X_N is the statistic for nonsynonymous

sites sand X_G is the statistic for the entire gene (R Core Team, 2025). We obtained p values for each statistic based on their relationship to the overall within-species distribution and corrected for FDR using the Benjamini-Hochberg method. Outlier genes were then identified as genes with $p < 0.05$ for D_N, π_N, F_{ST_N}, and as genes with $p_N/p_S > 1$ for nonsynonymous to synonymous polymorphism ratios. To test for GO term enrichment amongst outlier genes, performed gene set enrichment analysis using R package `topGO` (Alexa & Rahnenfuhrer, 2025) and GO term annotations obtained from VectorBase (Alvarez-Jarreta et al., 2024). To test for the statistical significance of overlap between lists of outlier genes, we used R package `SuperExactTest` (Wang et al., 2022).

4.2.2.2 Between species. We performed polarized gene-wise McDonald-Kreitman tests on each one-to-one ortholog identified in Genomic data & processing using Python package `mkado` (Rivera-Colón et al., 2026). In order to reduce the influence of rare variants on our analysis, we used a subset of samples from each species selected from the same geographic region and time to conduct the test. For *An. gambiae* we used the AG1000G-GN-A cohort, for *funestus* we used set 1230-VO-GA-CF-AYALA-VMF00045, and for *minimus* we used samples taken from coordinate location $(13.667, 104.982)$ in 2014. For *An. gambiae*, we used *funestus* as an outgroup and polarized using *An. minimus*; for *An. funestus* we used *gambiae* as an outgroup and polarized using *An. minimus*; for *An. minimus* we used *funestus* as an outgroup and polarized using *An. gambiae*. With `mkado`, we also performed asymptotic genome-wide McDonald-Kreitman tests using the same ingroup-outgroup-polarization pairings.

To test for positive selection along the *gambiae*, *funestus*, and *minimus* branches, we performed multi-ratio branch model tests on each one-to-one ortholog with the

focal species as the foreground branch using `crabML` (Kern, 2025; Yang, 2007). In order to account for the high degree of polymorphism within species, we ran ten test replicates for each gene, using randomly-chosen samples from each species as representative inputs.

4.2.2.3 F_{ST} correlation. To investigate spatial correlation between orthologous nonsynonymous F_{ST} outlier genes in *An. gambiae* and *funestus*, we calculated pairwise Hudson's F_{ST_N} (within each species) between clusters of sampling points for each gene. Sample clusters were identified using an HDBSCAN algorithm implemented through `scikit-learn` with a minimum cluster size of five individuals. To identify the average spatial axis of genetic differentiation for a given gene, we then treated each cluster pair as a west→east vector and averaged these vectors, weighted by their F_{ST_N}:

$$(\bar{x}_{F_{ST_N}}, \bar{y}_{F_{ST_N S}}) = \frac{\sum_{i=1}^{n} F_{ST_{NS_i}}(x_{i1}, y_{i1}, x_{i2}, y_{i2})}{\sum_{i=1}^{n} F_{ST_{NS_i}}}$$

We then tested for spatial similarity between *An. gambiae* and *funestus* differentiation using the cosine similarity between average F_{ST} vectors for a given pair of orthologs. To test for significance, we randomly permuted the F_{ST_N} values used to calculate the average F_{ST_N} vector ($N = 1000$) and compared observed and permuted cosine similarities.

For genes with significantly similar nonsynonymous F_{ST}, we probed which mutations were driving west→east differentiation by calculating F_{ST} between western and eastern sampling locations for each nonsynonymous mutation. Mutations with F_{ST} greater than 0.2 were considered driving mutations and retained for further analysis. To identify convergent driving mutations between *An. gambiae* and *funestus*, we analyzed the spatial distribution of each variant and used `ThermoMPNN` to assess the impacts of mutations on protein stability (Dieckhaus et al., 2024b).

4.3 Results

4.3.1 Nonsynonymous polymorphism is abundant and enriched for immune, sensory, and metabolic functions. To characterize nonsynonymous variation within each species, we examined patterns of polymorphism, site frequency spectra, and spatial differentiation at nonsynonymous sites relative to genic and synonymous baselines. Reported genome-wide nucleotide diversity is high in all three species: $\pi = 1.5 \times 10^{-2}$ in *An. gambiae*, between 1.4×10^{-2} and 1.5×10^{-2} in *An. funestus*, and between 1.3×10^{-2} and 1.4×10^{-2} in *An. minimus*, consistent with their large effective population sizes (Boddé et al., 2025; St. Laurent et al., 2022; The *Anopheles gambiae* 1000 Genomes Consortium, 2017). In our analysis, genic π (π_G) is approximately equal to genome-wide π in each species — except for in *An. gambiae*, where it is slightly higher — and nonsynonymous π (π_N) is one order of magnitude lower than π_G ($\pi_N = 2.58 \times 10^{-3}$, 4.17×10^{-3}, and 4.10×10^{-3} for *gambiae*, *funestus*, and *minimus*, respectively) (Figure 4.2, Table 4.1). For comparison, genome-wide π in African *D. melanogaster* is 7.5×10^{-3}, with $\pi_N = 1.5 \times 10^{-3}$ and $\pi_S = 1.9 \times 10^{-2}$ (Langley et al., 2012). Nonsynonymous π in *An. gambiae* is thus almost double that of African *D. melanogaster*; despite *An. gambiae*'s substantially higher genic diversity, in *An. funestus* and *minimus* π_N is higher still. Nonsynonymous π is significantly correlated with gene-wide and synonymous π in all three species (Figure C.1). However, the relationship between nonsynonymous and genic polymorphism differs among species: *An. gambiae* appears more constrained, while *An. funestus* and *An. minimus* show lower genic π but higher nonsynonymous π.

Turning to the site frequency spectrum, almost all genes in each species have negative Tajima's D values (median D_G = -1.48, -1.38, -1.83 and median D_N = -1.97, -1.88, -2.27 for *gambiae*, *funestus*, and *minimus*, respectively), consistent with

recent population expansions. Negative genome-wide Tajima's D is also observed in *D. melanogaster*, where it is driven primarily by linked selection rather than demography (Langley et al., 2012). Although across the genome the distribution of nonsynonymous Tajima's D is shifted below genic Tajima's D in each species, linear models of the per-gene relationship between nonsynonymous and genic D reveals that individual genes tend to have higher nonsynonymous than genic D (coefficient $\beta = 1.227, 1.237, 1.173$ for *gambiae, funestus*, and *minimus*; $p < 2 \times 10^{-16}$ and $\rho > 0.94$ for all three models; Figure C.2). This pattern — nonsynonymous D shifted upward relative to genic D on a per-gene basis — is consistent with balancing forces maintaining nonsynonymous variation at a subset of genes.

Finally, genome-wide distributions of genic and nonsynonymous F_{ST} are nearly identical within each species, with a close to 1:1 relationship between the two measures (Figures 4.2, C.3). In most organisms, purifying selection would be expected to depress nonsynonymous F_{ST} relative to genic F_{ST}. The absence of this pattern here suggests that local adaptation may be elevating nonsynonymous differentiation, consistent with the high levels of nonsynonymous polymorphism we observe across species.

Species	π_G	π_N	**Tajima's** D_G	**Tajima's** D_N	F_{ST_G}	F_{ST_N}
An. gambiae	0.02348	0.00258	-1.48478	-1.96957	0.04450	0.03859
An. funestus	0.01467	0.00417	-1.38169	-1.88379	0.08568	0.07314
An. minimus	0.01206	0.00408	-1.82551	-2.26612	0.01562	0.01224

Table 4.1. Summary of genic and nonsynonymous population genetic statistics for *An. gambiae*, *funestus*, and *minimus*. Tajima's D values are reported as the median across genes; other values are reported as the mean across genes.

Taken together, the π, Tajima's D, and F_{ST} results point to a shared pattern across all three species: nonsynonymous variation is abundant, shows signatures of

balancing selection on a per-gene basis, and is not depleted by purifying selection to the extent seen in other organisms. We next examine which genes and functional categories drive these genome-wide patterns in each species.

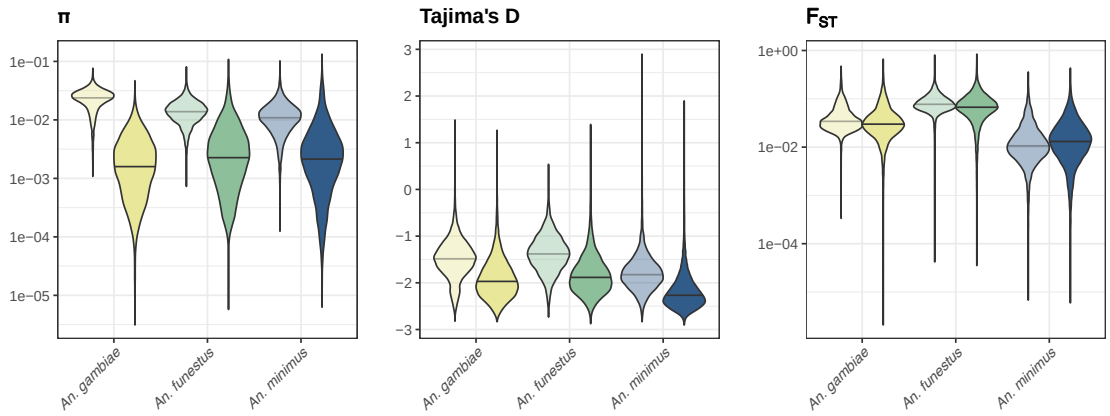


Figure 4.2. Distribution of genic (lighter color) and nonsynonymous (darker color) π, Tajima's D, and F_{ST} values within *An. gambiae*, *funestus*, and *minimus*.

In *An. gambiae*, genes with elevated nonsynonymous polymorphism include cytochrome P450s, which are involved in insecticide resistance, immune genes such as CLIP-domain serine protease (AGAP011790), and cuticular genes such as chitinase (AGAP006191). Among high-π_N genes, we find significant enrichment for sensory perception of smell, stimulus response, stress response [DNA damage response, cellular response to stress], proteolysis, and proteoglycan metabolism (Figure 4.3, Table C.1). The overrepresentation of these functional groups likely reflects varying environmental pressures across *An. gambiae*'s range, which covers the majority of sub-Saharan Africa: sensory perception and stimulus response may involve host-seeking behavior or insecticide avoidance, both of which vary across the landscape, while stress response is also associated with insecticide resistance. The enrichment of proteoglycan metabolism genes may reflect balancing selection on an immune trait, as proteoglycans mediate *Plasmodium* interactions (Armistead et al., 2011).

Genes with elevated nonsynonymous Tajima's D echo several of these categories, with enrichment for sensory perception of smell and proteolysis, a key component of blood meal digestion, as well as glutathione, sulfur compound, and NADP metabolic processes (Table C.4). These metabolic categories are all involved in maintaining redox equilibrium (Champion & Xu, 2018). The production of reactive oxygen species is a key part of *Anopheles* immune defense, and redox states have been shown to impact fitness in *An. gambiae* (Carlos et al., 2025; Champion & Xu, 2018; S. Kumar et al., 2003). The convergence of π_N and Tajima's D_N enrichments on overlapping functional categories suggests that balancing selection and local adaptation help maintain functional polymorphism in *An. gambiae* redox metabolism, with potential downstream impacts on immunity, fecundity, and insecticide susceptibility.

A similar pattern emerges in *An. funestus*, where genes with elevated nonsynonymous polymorphism include a number of immune genes — we observe 13 CD109 antigen-like genes, leucine-rich repeat proteins, and a TsetseEP-like gene among outliers. Among high-π_N genes, we find significant enrichment for reproduction, development, and protein metabolism (Figure 4.3, Table C.2).

Genes with elevated nonsynonymous Tajima's D also include immune genes, as well as multiple phenoloxidase-activating enzymes. GO enrichment for high-D_N genes reveals protein metabolic processes along with a number of catabolic and transport processes (Table C.5). Notably, metabolic enrichments center on nucleoside diphosphate and carbohydrate metabolism, particularly purine ribonucleotide diphosphates, which are involved in the pentose phosphate pathway (PPP), a critical component of *Anopheles* redox metabolism (Kanehisa et al., 2024). The presence of phenoloxidase-activating enzymes among high-D_N genes reinforces this redox connection, as phenoloxidase activity is tightly linked to reactive

oxygen species production. As in *An. gambiae*, the overlap between π_N and Tajima's D_N enrichments on metabolic processes suggests that balancing selection and spatially varying pressures maintain functional polymorphism in *An. funestus*, with protein and carbohydrate metabolism genes likely subject to varying dietary and immunological demands across the species range.

The picture differs in *An. minimus*. Genes with elevated nonsynonymous polymorphism include cuticular proteins and immune genes such as a CLIP-domain serine protease, and elevated nonsynonymous Tajima's D flags additional immune genes. Among high-π_N genes, we find enrichment primarily for broad regulatory functions (Figure 4.3, Table C.3). High-D_N genes show a similar regulatory signature, with enrichment for regulation of catalytic activity, protein metabolic process, and macromolecule metabolic process, as well as sensory perception of chemical stimulus (Table C.6). The sensory enrichment parallels the pattern observed in *An. gambiae*, but the overall enrichment profile in *An. minimus* is notably distinct from the other two species: where *An. gambiae* and *An. funestus* show enrichment for specific metabolic and immune pathways, *An. minimus* enrichments center on higher-level regulatory processes, suggesting that nonsynonymous variation in this species may be maintained through different selective pressures or different evolutionary responses to similar pressures.

Across all three species, immune genes and sensory genes consistently appear among nonsynonymous outliers, while the metabolic signatures of balancing selection — redox metabolism in *An. gambiae* and *An. funestus*, broad regulatory processes in *An. minimus* — suggest both shared and lineage-specific pressures shaping nonsynonymous variation in these vectors.

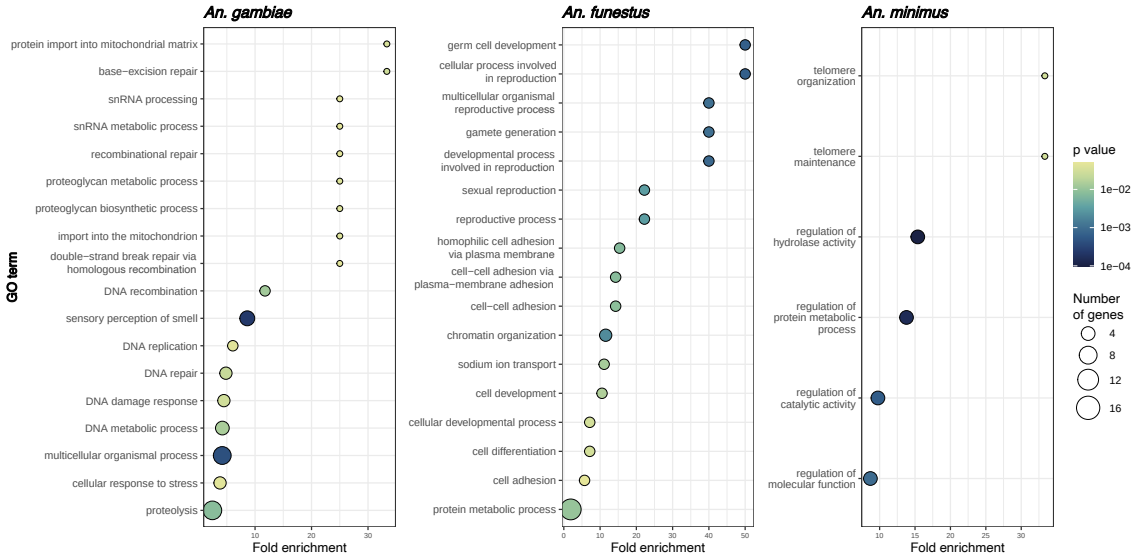


Figure 4.3. Significant GO enrichments for nonsynonymous π outliers in *An. gambiae*, *funestus*, and *minimus*.

4.3.2 Adaptive divergence targets overlapping functional categories across lineages.

4.3.2.1 McDonald Kreitman test. Having established that nonsynonymous variation is abundant and bears signatures of balancing selection, we next asked whether comparisons of polymorphism and divergence show evidence of non-neutral evolution. Using the asymptotic McDonald-Kreitman test (Messer & Petrov, 2013), we estimated the proportion of amino acid fixations driven by directional selection (α) for all one-to-one orthologs across the three species. Despite the high levels of nonsynonymous polymorphism documented above, genome-wide α exceeds 0.60 in all three lineages, with the lowest value in *An. gambiae* ($\alpha = 0.600$) and the highest in *An. funestus* ($\alpha = 0.655$). These values are comparable to or slightly higher than the asymptotic α of 0.571 estimated for *D. melanogaster* (Haller & Messer, 2017; Messer & Petrov, 2013), indicating similarly pervasive adaptive protein evolution despite very different ecology and life history. However, the number of individual

genes with significant evidence of directional selection is small, indicating that while strong positive selection acts on a subset of genes, the majority continue to segregate nonsynonymous variants (Figure 4.4).

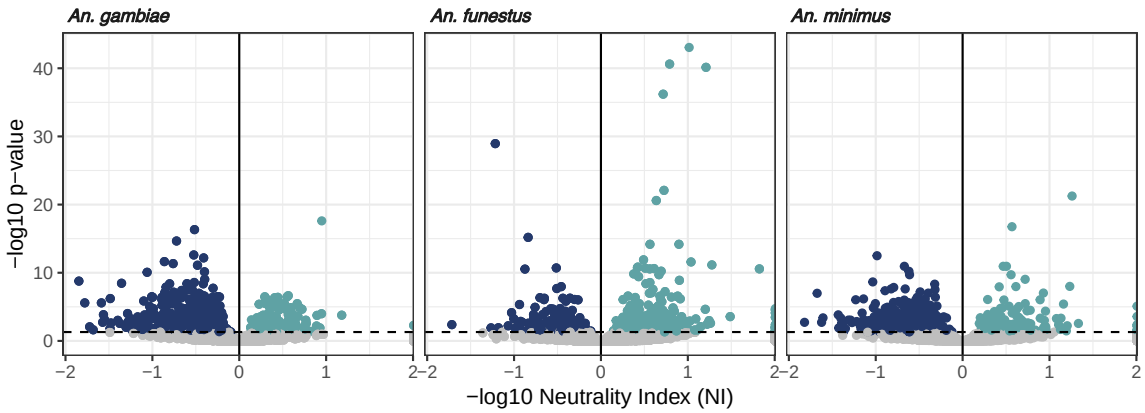


Figure 4.4. Results of the asymptotic McDonald-Kreitman test for all one-to-one orthologs across *An. gambiae*, *An. funestus*, and *An. minimus*. Light blue points represent genes with significant evidence of adaptive divergence after Boneferroni correction ($p < 0.05$), while dark blue points represent genes with a significant excess of nonsynonymous polymorphism.

In *An. gambiae*, 1238 of 5149 one-to-one orthologs (24%) show evidence of non-neutral evolution ($p < 0.05$), but only 171 of these have α values indicating adaptive directional selection ($\alpha > 0.5$). The gene with the highest significant α, fatty acyl-CoA reductase 2 (AGAP004784), is involved in cuticle formation and has an estimated 100% of its nonsynonymous fixations due to positive selection ($p = 3.406 \times 10^{-3}$). Several sensory genes also show strong directional selection, including odorant-binding protein 11 (AGAP002025, $\alpha = 0.900$, $p = 6.926 \times 10^{-3}$), sensory appendage protein 1 (AGAP008051, $\alpha = 0.816$, $p = 2.887 \times 10^{-2}$), and odorant-binding protein 69 (AGAP006368, $\alpha = 0.814$, $p = 1.393 \times 10^{-2}$). Notably, sensory perception of smell is also enriched among high-π_N genes (see above), indicating that this functional category potentially experiences both balancing and directional selection

on different loci. Immune genes with elevated α include a mannose-binding C-type lectin (AGAP029559, $\alpha = 0.818$, $p = 1.449 \times 10^{-12}$), a CLIP-domain serine protease (AGAP011789, $\alpha = 0.681$, $p = 2.162 \times 10^{-2}$), and a leucine-rich immune protein (AGAP006348, $\alpha = 0.616$, $p = 1.044 \times 10^{-3}$). One of the strongest signals of adaptive divergence is in a heparan-sulfate 6-O-sulfotransferase (AGAP001444, $\alpha = 0.937$, $p = 1.195 \times 10^{-2}$). Heparan sulfate mediates *Anopheles–Plasmodium* interactions, and this signal may reflect coevolution between vector and parasite (Armistead et al., 2011; Sinnis et al., 2007).

Gene	p **value**	α	**Product Description**
AGAP004784	3.406×10^{-3}	1.00	fatty acyl-CoA reductase 2
AGAP009154	7.697×10^{-5}	0.96	actin-related protein 2/3 complex subunit 2
AGAP001444	1.195×10^{-2}	0.94	Heparan-sulfate 6-O-sulfotransferase
AGAP005400	4.334×10^{-7}	0.92	four and a half LIM domains
AGAP001903	1.662×10^{-17}	0.91	NAD-dependent malate dehydrogenase
AGAP005696	1.664×10^{-3}	0.91	cuticular protein from two-cysteine family 1
AGAP005991	5.071×10^{-3}	0.91	60S ribosomal protein L14
AGAP007668	3.841×10^{-2}	0.91	Eukaryotic translation initiation factor 3 subunit G
AGAP002025	6.926×10^{-3}	0.90	odorant-binding protein 11
AGAP005958	4.641×10^{-4}	0.86	Yellow-g

Table 4.2. Top 10 genes with significant evidence of adaptive divergence in *An. gambiae*

Testing for functional enrichment among high-α genes, we find enrichment for translation, protein-containing complex assembly, and nuclear transport — categories likely reflecting general species divergence, though changes in protein complex assembly could also involve complement-like complexes central to *Anopheles* immune response (Povelones et al., 2009, 2011). We also find enrichment for stress response, carbohydrate derivative catabolism, and organophosphate catabolism. As with sensory genes, stress response genes appear to experience both balancing selection (excess nonsynonymous polymorphism) and directional selection (elevated α) on

different loci. The enrichment for organophosphate catabolism is particularly striking and may represent a direct adaptation to the phosphate-containing insecticides deployed in vector control.

A similar proportion of genes show non-neutral evolution in *An. funestus* (1249 of 5149, 24.2%), though a notably larger fraction — 710 genes — have α values indicating adaptive divergence, compared to 171 in *An. gambiae*. Cuticle formation, sensory perception, insecticide resistance, and immune response are again common themes among high-α genes. Fourteen genes in *An. funestus* lack any nonsynonymous polymorphism despite nonsynonymous divergence ($\alpha = 1.0$), including Tango10 (AGAP000791, $p = 2.909 \times 10^{-5}$), a circadian rhythm protein, and sensory appendage protein 1 (AGAP008051, $p = 3.472 \times 10^{-4}$), which also shows elevated α in the *gambiae* lineage, indicating that this gene is diverging under directional selection in both species independently. Yellow-g (AGAP005958) provides another case of parallel divergence, with $\alpha = 0.873$ ($p = 1.804 \times 10^{-4}$) in *An. gambiae* and $\alpha = 1.0$ ($p = 7.884 \times 10^{-5}$) in *An. funestus*. Its ortholog inhibits *Plasmodium vivax* in *An. dirus* (Mongkol et al., 2021) and prevents egg desiccation in *Aedes* mosquitoes (Noh et al., 2020), either of which could explain its adaptive divergence along both lineages.

Functional enrichment among high-α genes in *An. funestus* reveals a broader set of categories than in *An. gambiae*, including reproduction, cell cycle, nuclear transport, translational initiation, development, sensory perception, and locomotion. Reproductive and developmental enrichments likely reflect the rapid evolution characteristic of these functional categories, and nuclear transport and translation may be attributed to general species divergence. More striking are the enrichments for locomotion and signal transduction, both of which may be associated with adaptation to insecticide exposure: pyrethroids in particular cause locomotor deficits,

Gene	p value	α	Product Description
AGAP000308	4.426×10^{-2}	1.00	proteasome activator subunit 3 (PA28 gamma)
AGAP000791	2.909×10^{-5}	1.00	Tango10
AGAP001515	3.386×10^{-2}	1.00	unspecified product
AGAP001805	9.179×10^{-3}	1.00	60S ribosomal protein L13
AGAP002053	1.224×10^{-2}	1.00	unspecified product
AGAP002935	1.312×10^{-2}	1.00	Eukaryotic translation initiation factor 3 subunit F
AGAP002936	4.021×10^{-2}	1.00	unspecified product
AGAP003462	2.819×10^{-2}	1.00	protein arginine N-methyltransferase 1
AGAP005991	4.220×10^{-3}	1.00	60S ribosomal protein L14
AGAP005958	7.884×10^{-5}	1.00	Yellow-g

Table 4.3. Top 10 genes with significant evidence of adaptive divergence in *An. funestus*

and resistance is associated with changes in locomotion (Brito et al., 2013; Charreton et al., 2015), while signal transduction is disrupted by several classes of insecticides (Casida & Durkin, 2013) and may also contribute to behavioral resistance via avoidant behavior.

The *An. minimus* lineage shows markedly less non-neutral evolution: only 815 of 5149 orthologs (15.8%) reject neutrality, and just 215 have α values indicating adaptive divergence. As in *An. funestus*, a number of genes lack any nonsynonymous polymorphism despite nonsynonymous divergence ($\alpha = 1.0$), including two post-translational modification enzymes: serine/threonine protein phosphatase 6 (AMIN003725, $p = 1.493 \times 10^{-2}$) and protein arginine N-methyltransferase 1 (AMIN000933, $p = 5.432 \times 10^{-3}$). Other high-α genes include immune proteins such as spaetzle-like cytokine 1 (AMIN005499, $\alpha = 0.808$, $p = 9.834 \times 10^{-5}$) and a mannose-binding C-type lectin (AMIN016218, $\alpha = 0.756$, $p = 3.087 \times 10^{-2}$), as well as cuticular proteins such as fatty acyl-CoA reductase (AGAP001084) and sensory proteins such as odorant-binding protein 40 (AGAP002191). Although we do not find evidence for adaptive divergence in Yellow-g as we do in *An. gambiae* and *An.*

funestus, a paralog, Yellow-g2, is adaptively diverging in *An. minimus* (AMIN010972, $\alpha = 0.796$, $p = 4.141 \times 10^{-3}$). Yellow-g2's ortholog in *Aedes* prevents egg desiccation (Noh et al., 2020), and it may play a role analogous to Yellow-g in *Plasmodium* inhibition (Mongkol et al., 2021), suggesting that selection on this gene family is a recurring theme across *Anopheles* lineages.

Gene	p **value**	α	**Product Description**
AGAP002762	1.975×10^{-2}	1.00	Chaperonin containing TCP1 subunit 7
AGAP003017	1.493×10^{-2}	1.00	serine/threonine-protein phosphatase 6 catalytic subunit
AGAP003462	5.432×10^{-3}	1.00	protein arginine N-methyltransferase 1
AGAP008041	3.147×10^{-3}	1.00	RING finger protein 121
AGAP009863	2.319×10^{-3}	1.00	eukaryotic translation initiation factor 4A, isoform 2
AGAP011706	2.348×10^{-2}	1.00	60S ribosomal protein L27a
AGAP011937	8.950×10^{-4}	1.00	unspecified product
AGAP001805	1.942×10^{-3}	0.96	60S ribosomal protein L13
AGAP004819	4.690×10^{-3}	0.95	receptor expression-enhancing protein 5
AGAP005542	2.030×10^{-9}	0.94	Queuine tRNA-ribosyltransferase subunit QTRTD1 homolog

Table 4.4. Top 10 genes with significant evidence of adaptive divergence in *An. minimus*

Functional enrichment among high-α genes in *An. minimus* highlights programmed cell death, translation, nuclear transport, purine nucleotide metabolism, and phosphorus metabolism. As in the other two species, translation and nuclear transport likely reflect general species divergence. Programmed cell death is a hallmark of both development and immune response; its enrichment among adaptively diverging genes may reflect directional selection on immune function. The enrichment of purine ribonucleotide and ribose phosphate metabolism parallels the polymorphism results in *An. funestus* and may signal functional divergence in redox metabolism, driven by selection on immune defense or insecticide detoxification. Finally, the enrichment for phosphorus metabolism echoes the organophosphate catabolism signal

in *An. gambiae* and likely reflects adaptation to organophosphate insecticides deployed across these species' ranges.

Across all three species, the McDonald-Kreitman results reveal a consistent picture: genome-wide α values comparable to *Drosophila* indicate pervasive adaptive protein evolution, yet the genes individually under directional selection are enriched for the same functional themes — immune response, sensory perception, cuticle formation, and insecticide-related metabolism — that also harbor excess nonsynonymous polymorphism. Several genes, including sensory appendage protein 1 and the Yellow-g family, show parallel adaptive divergence across lineages, reinforcing the convergence of selective pressures acting on these deeply diverged vectors.

4.3.2.2 Phylogenetic testing. To complement the McDonald-Kreitman analysis, we used a branch model (Yang, 2007) to test for positive selection ($\omega > 1$) along the *gambiae*, *funestus*, and *minimus* branches. Because population-level samples contain segregating variation that can influence branch-model inference, we ran the test ten times per ortholog using alignments of randomly drawn haplotypes from each species. This design distinguishes genes where all sampled haplotypes show $\omega > 1$ — indicating fixation of adaptive substitutions — from genes where only some haplotypes do, suggesting that adaptive divergence is still segregating within the species (Between species).

In *An. gambiae*, genes with significant $\omega > 1$ are enriched for signal transduction, and we observe elevated ω on segregating haplotypes in immune genes such as LRRCT domain-containing protein and in cuticular proteins such as fatty acyl-CoA reductase. In *An. funestus*, enrichments shift to translation, mitosis, and polyamine metabolism. LRRCT domain-containing protein again shows elevated ω across all sampled haplotypes — a parallel with *An. gambiae* — while other immune genes

such as mannose-binding C-type lectin show $\omega > 1$ on only a subset of haplotypes, illustrating how adaptive divergence can coexist with segregating nonsynonymous variation at the same locus. In *An. minimus*, enrichments center on ATP metabolism, chitin metabolism, amino sugar metabolism, and purine ribonucleotide metabolism — the last of which echoes the purine metabolic enrichments observed in both the polymorphism and MK analyses. Unlike *An. gambiae* and *An. funestus*, immune and cuticular genes in *An. minimus* do not show haplotype-level variation in ω, consistent with the more generic regulatory enrichments observed among *minimus* nonsynonymous polymorphism outliers.

4.3.3 Convergent nonsynonymous spatial differentiation between *An. gambiae* and *funestus*. The analyses above show that *An. gambiae* and *An. funestus* share many of the same functional targets of selection. Because these two species are broadly sympatric across sub-Saharan Africa, we next asked whether nonsynonymous spatial differentiation within each species follows similar geographic patterns. Among the one-to-one orthologs, seven exhibit elevated nonsynonymous F_{ST} in both species, including three cytochrome P450s, a cholesterol transporter, and a Niemann-Pick Type C-2 protein. To test whether differentiation in these genes follows similar spatial axes, we averaged nonsynonymous F_{ST} vectors across sampling locations for each gene and compared the resulting vectors between species (F_{ST} correlation). Three genes share significantly similar axes of nonsynonymous differentiation: two cytochrome P450s at the Cyp6p locus — Cyp6p5 (AGAP002866) and Cyp6p2 (AGAP002869) — and protein FAM91A1 (AGAP002874) (Figure C.4). The Cyp6p locus is a cluster of cytochrome P450 genes whose mutations are implicated in insecticide resistance, and the shared east-west axis of differentiation across species suggests convergent adaptation to spatially varying

insecticide exposure. FAM91A1 has no well-documented role in *Anopheles* biology, but its convergent spatial differentiation with the Cyp6p genes may indicate a related function in insecticide response.

In Cyp6p5, we identify four nonsynonymous mutations contributing to west-east F_{ST_N} in *An. gambiae* (AGAP002866) and seven in *An. funestus* (AFUN2_013651). Of these, three in each species share the same spatial constraint, occurring almost exclusively in the northeast of the sampling region in the Central African Republic, Uganda, and northwest Tanzania (Figure 4.5A). In both species, the spatially constrained mutations cluster near the N-terminus or channel entrance of the protein, suggesting convergent molecular evolution in this region (Figure 4.5B). We used ThermoMPNN to estimate the impact of individual mutations on protein folding stability (though current methods cannot model the effects of co-occurring mutations). In the *gambiae* Cyp6p5, N343T and V87I are moderately destabilizing ($\Delta\Delta G = 0.243$ and 0.083 kcal/mol, respectively), and H36C, while only slightly destabilizing ($\Delta\Delta G = 0.038$ kcal/mol), substitutes a sulfur-containing cysteine for a histidine near the protein's first α helix, with potentially outsized impacts on folding or catalytic activity. All three mutations segregate on the same haplotype. In the *funestus* Cyp6p5, V14L and R169H are moderately stabilizing ($\Delta\Delta G = -0.136$ and -0.022 kcal/mol), while P171Q is moderately destabilizing ($\Delta\Delta G = 0.248$ kcal/mol), likely due to substitution of a proline in a hairpin turn between two α helices. R169H and P171Q segregate on the same haplotype, while V14L does not co-occur with either. The remaining four *funestus* mutations are spatially constrained to the far southeast or northwest and may represent independent adaptive responses to insecticide exposure (Figure C.5). Although individual mutation effects on protein

stability are moderate, co-occurrence on the same haplotype in both species raises the possibility of non-additive epistatic effects on protein function.

A similar pattern emerges in Cyp6p2. We identify four nonsynonymous mutations contributing to west-east F_{ST_N} in *An. gambiae* (AGAP002869) and nine in its *An. funestus* ortholog. One mutation in each species — P449S in *gambiae* and Y391D in *funestus* — shares the same spatial constraint and maps to the same region of the protein structure (Figure 4.6). Both are strongly destabilizing: P449S ($\Delta\Delta G$ = 1.067 kcal/mol) disrupts a hairpin turn near the heme-binding pocket, and Y391D ($\Delta\Delta G$ = 2.801 kcal/mol) occurs at a similar position near the channel entrance.

Across both Cyp6p5 and Cyp6p2, the same geographic regions segregate functionally relevant mutations at high frequency in both species, and the mutations themselves have convergent destabilizing effects on protein structure. Cytochrome P450 enzymes are central to insecticide resistance in *Anopheles*, and while most resistance at this locus has been attributed to copy number variation, recent evidence points to a role for nonsynonymous variation as well (Rehmann et al., 2025b). The convergent spatial and molecular patterns we observe at the Cyp6p locus provide strong evidence that *An. gambiae* and *An. funestus* are independently adapting to the same insecticide pressures through parallel nonsynonymous changes in the same genes.

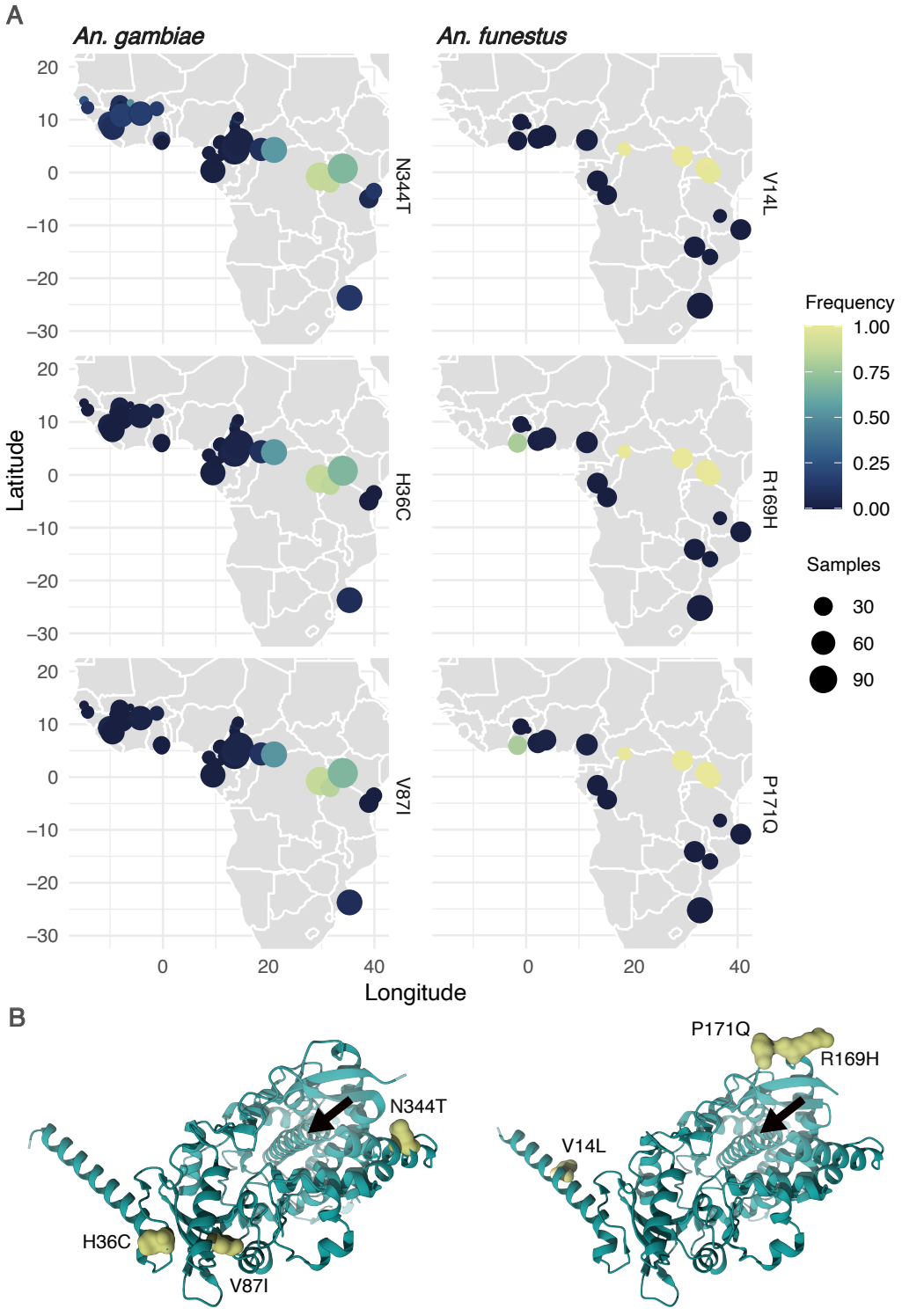


Figure 4.5. (A) Frequency of spatially-constrained nonsynonymous Cyp6p5 mutations in *An. gambiae* (left) and *funestus* (right). (B) Position of spatially-constrained nonsynonymous mutations on the Cyp6p5 protein in *An. gambiae* (left) and *funestus* (right). Arrow indicates channel entrance.

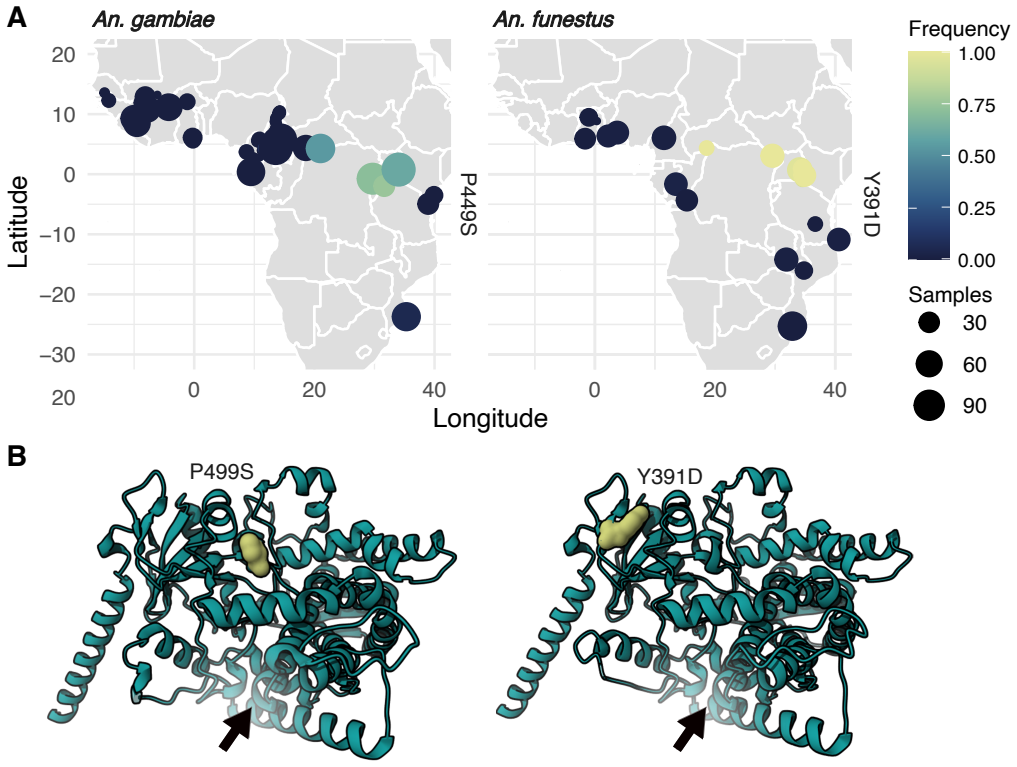


Figure 4.6. (A) Frequency of spatially-constrained nonsynonymous Cyp6p2 mutations in *An. gambiae* (left) and *funestus* (right). (B) Position of spatially-constrained nonsynonymous mutations on the Cyp6p2 protein in *An. gambiae* (left) and *funestus* (right). Arrow indicates channel entrance.

4.4 Discussion

In this study, we use gene-level population and phylogenetic approaches to uncover the evolutionary forces acting on nonsynonymous variation in three deeply diverged *Anopheles* species. Our results point to a scenario where diversifying selection maintains functional nonsynonymous polymorphism within species, and where the same genes and functional categories that harbor this variation are also evolving under directional selection along lineages, providing a substrate for global adaptation when selection pressures are strong enough.

We find that nonsynonymous polymorphism is abundant in *Anopheles* populations, and that this variation is maintained in part by balancing selection, as indicated by the relationship between genic and nonsynonymous Tajima's D, and in part by local adaptation, as indicated by the relationship between genic and nonsynonymous F_{ST}. Genes with elevated nonsynonymous polymorphism are enriched for sensory and metabolic functions, particularly redox metabolism; we also identify a number of specific genes involved in immune response and cuticle formation. Many of these genes and functions are also evolving under directional selection along *Anopheles* lineages, suggesting that the similar selection pressures both maintain polymorphism and drive divergence in these species. We hypothesize that much of the nonsynonymous variation we observe is maintained under hyperlocal adaptation, and can sweep to fixation under strong enough selection pressures. Adaptation to insecticide exposure in particular appears to be under both diversifying and directional selection, with behavioral, metabolic, and cuticular genes both segregating nonsynonymous variants and fixing them adaptively. Sensory functions involved in behavioral resistance are enriched for nonsynonymous variation in *An. gambiae* and *minimus*, but certain genes involved in sensory processing have a significant proportion of adaptive fixations in all three lineages. This is also true of metabolic resistance: while we identify two striking examples of convergent local adaptation in two *An. gambiae* and *funestus* cytochrome P450s in addition to a number of detoxification enzymes segregating excess amounts of nonsynonymous polymorphism, overall metabolism of organophosphates is evolving under directional selection in *gambiae* and *minimus*. Genes involved in cuticle formation are also outliers with respect to nonsynonymous polymorphism, likely driven by local adaptation to desiccation and insecticide exposure, however cuticle formation and maintenance

genes are also fixing adaptive variants on deeper evolutionary timescales. All of this points to the polymorphism maintained in *Anopheles* populations being an abundant reservoir of potentially-adaptive variation relevant to the sort of evolutionary pressures they are facing across their ranges.

The genome-wide asymptotic α values we estimate (0.600–0.655 across the three species) place *Anopheles* among the most adaptively evolving animal lineages studied to date. Across a broad survey of animals, α correlates positively with effective population size, with large-N_e invertebrates such as *Drosophila* ($\alpha \approx 0.57$) at the high end and small-N_e vertebrates such as primates ($\alpha \approx 0.2$) at the low end (Galtier, 2016). The *Anopheles* values exceed those of *D. melanogaster* and are consistent with the large effective population sizes of these species, confirming that selection acts efficiently to fix beneficial amino acid substitutions. This makes the abundance of segregating nonsynonymous polymorphism all the more striking: it persists not because selection is weak, but despite selection being strong. At the same time, relatively few individual genes show significant evidence of directional selection in the per-gene MK tests, even as the asymptotic genome-wide test indicates pervasive adaptation. This apparent discrepancy likely reflects the strong local population structure in *Anopheles*, which generates an excess of rare nonsynonymous variants that inflate the neutrality index at individual loci. The asymptotic MK test accounts for this by fitting the frequency spectrum, but individual gene-level tests lack the power to do so and are consequently biased toward excess amino acid polymorphism. This interpretation is consistent with the deeply negative Tajima's D values we observe across species and with the spatial population structure evident in our F_{ST} analyses.

Our phylogenetic branch model analysis provides a complementary line of evidence connecting within-species polymorphism to between-species divergence. By running branch models on replicate alignments of randomly sampled haplotypes, we can distinguish genes where all sampled haplotypes show $\omega > 1$ — indicating that adaptive substitutions have fixed along a lineage — from genes where only some haplotypes do, indicating that adaptive divergence is still segregating within the species. This distinction is important because it reveals that some of the nonsynonymous variation we observe within species is not merely tolerated but is itself adaptive, with different haplotypes at the same locus under different selective regimes. In *An. gambiae* and *An. funestus*, immune genes such as LRRCT domain-containing protein show elevated ω across all sampled haplotypes, while others such as mannose-binding C-type lectin show $\omega > 1$ on only a subset, illustrating how adaptive divergence can coexist with segregating functional polymorphism at the same locus. These results reinforce the view that nonsynonymous variation in *Anopheles* is actively maintained by selection rather than simply awaiting fixation or loss.

One interesting functional category that is both segregating high amounts of nonsynonymous variation and evolving directionally in *Anopheles* is genes involved in redox metabolism. This includes elevated nonsynonymous polymorphism in genes involved in glutathione metabolism, NADPH metabolism, and potentially the pentose phosphate pathway via enrichment for genes involved in purine ribonucleotide metabolism (Kanehisa et al., 2024). *Anopheles* redox metabolism is critical to their immune response and fecundity, and is also implicated in tolerance of insecticide exposure (Carlos et al., 2025; Champion & Xu, 2018; S. Kumar et al., 2003). Given the importance of reactive oxygen species in crucial *Anopheles* biological functions and its impacts on overall mosquito fitness, it is unsurprising that genes involved in

redox metabolism are both maintaining nonsynonymous variation, possibly due to local adaptation to different pathogen and insecticide pressures, and also evolving directionally along *Anopheles* lineages, with some of these variants becoming globally adaptive and fixing.

Beyond redox metabolism, we also uncover a number of genes and functional categories involved in immune response that are both segregating nonsynonymous variation and evolving directionally in *Anopheles*. This includes genes involved in pathogen recognition, such as C-type lectins and leucine-rich repeat proteins, as well as genes involved in *Anopheles* immune signaling, such as CLIP-domain serine proteases. In *An. funestus*, we observe an interesting outlier for nonsynonymous polymorphism in a gene encoding protein TsetseEP-like (AFUN2_010725), a homolog of TsetseEP, a protein involved in trypanosome defense in tsetse flies. It is not unexpected that we observe immune genes to be both maintaining variation and evolving directionally, as *Anopheles* are under constant pressure from a variety of pathogens and the immune system is a common target of selection in many species. The *Anopheles* immune response is also critical to their ability to transmit the *Plasmodium* parasites, and we observe some evidence that genes specifically involved in *Anopheles–Plasmodium* interactions are either maintaining variation or evolving directionally, such as the heparan-sulfate 6-O-sulfotransferase in *An. gambiae* and potentially Yellow-g in *An. gambiae* and *funestus*. Understanding the genes required for *Anopheles* vector competence is informative for potential malaria transmission-blocking strategies, and some of these genes may be promising candidates for further investigation in this regard, particularly those that are evolving under directional selection as they may be more likely to be involved in critical *Anopheles-Plasmodium* interactions.

Human vector control efforts, particularly insecticide deployment, are a major selective pressure on *Anopheles* populations. Unsurprisingly, we see clear evidence of adaptation to insecticide exposure in all three species, with genes involved in metabolic, behavioral, and cuticular resistance all showing signals of both local and directional selection. Consistent with the relatively recent deployment of these measures (in evolutionary timescales) as well as the variability in this selection pressure across space, we observe many genes relevant to insecticide resistance are segregating an excess of nonsynonymous polymorphism, particularly genes involved in sensory perception (behavioral resistance) and cellular responses to stress. One example of local adaptation to potential insecticide pressure is identified in two cytochrome P450 genes, Cyp6p5 and Cyp6p2, where we observe similar spatial patterns of nonsynonymous differentiation in *An. gambiae* and *funestus* populations. In addition to this local adaptation, we also find evidence for global adaptation to insecticide exposure in *An. gambiae* and *minimus*, with enrichment for genes involved in organophosphate metabolism amongst genes with significantly elevated α values. Organophosphate insecticides are commonly used in vector control efforts and replaced more environmentally persistent organochlorine insecticides such as DDT; however resistance to organophosphates is becoming more widespread and here we observe clear evidence of adaptation to this class of insecticides in *Anopheles* vectors (Lucas et al., 2023; Saad et al., 2025).

One of our most striking results is the convergent spatial differentiation in Cyp6p5 and Cyp6p2 in *An. gambiae* and *funestus*, with the same geographic regions segregating functionally relevant mutations in both species. Both these genes occur at the Cyp6p locus, a cluster of cytochrome P450 genes that are associated with pyretheroid resistance in *Anopheles* (Edi et al., 2014; Müller et al., 2008a). While

much of the resistance at this locus has been attributed to copy number variation leading to overexpression of these detoxification enzymes, more recent evidence points to a role for nonsynonymous variation also being adaptive in these proteins (Rehmann et al., 2025b), and indeed we observe several nonsynonymous mutations in both Cyp6p5 and Cyp6p2 that are spatially constrained, despite being at high frequency locally, and also have destabilizing effects on protein folding. In Cyp6p5, this includes three mutations in *An. gambiae* and three in *funestus* that occur almost exclusively in the northeast of the sampling region and are clustered at the N-terminus and channel entrance of the enzyme. In Cyp6p2, we find even stronger evidence for molecular convergence, with a single mutation in either *An. gambiae* or *funestus* being strongly destabilizing and also being nearly fixed in the same geographic region. The observation of convergent spatial and molecular patterns in these two genes provides an example of local adaptation to insecticide pressures, which is not entirely unsurprising given that *An. gambiae* and *funestus* share a range across which they are subject to many similar environmental pressures. Such common patterns of adaptation have been observed in other species, such as in *Drosophila*, where deeply-diverged species show parallel adaptation to environmental clines across their range (Zhao & Begun, 2017).

Insecticides are of course not the only selective pressure acting on *Anopheles* biology, and on longer evolutionary timescales we observe functional categories of genes evolving under directional selection that, while presently relevant to insecticide resistance, are also acted on by selection pressures that predate insecticide deployment. This includes genes involved in cuticle formation, which are relevant to insecticide resistance but also likely evolving directionally in response to desiccation and other environmental pressures, as well as genes involved in sensory perception,

which are relevant to behavioral resistance but also critical for a variety of other *Anopheles* functions such as host-seeking and mate choice. The fact that these functional categories have been evolving under directional selection along *Anopheles* lineages while still segregating high amounts of nonsynonymous variation suggests that the same genes and pathways relevant to contemporary insecticide pressures have been important targets of directional selection and local adaptation throughout *Anopheles* evolutionary history, and that the variation we observe in these genes today may be a future reservoir of adaptive potential for these species.

One functional category that we observe in one species but not others is genes involved in reproduction, which are both segregating excess nonsynonymous polymorphism and evolving directionally in *An. funestus* but not *gambiae* or *minimus*. One possibility is that this is due to cryptic species divergence within *An. funestus*, as reproductive functions are often rapidly evolving along divergent lineages and may be particularly sensitive to population structure. Indeed sympatric 'ecotypes' have been observed in *An. funestus* populations in Burkina Faso (Small et al., 2023) and deeply-diverged mitochondrial lineages have been observed in sympatric *An. funestus* populations in Zambia and Tanzania (Jones et al., 2018). Further potential evidence for cryptic species divergence may be the presence of large genomic inversions in *An. funestus*, similar to the well-documented inversions in the *An. gambiae* species complex that are associated with speciation (Boddé et al., 2025; The *Anopheles gambiae* 1000 Genomes Consortium, 2017). While we do not find direct evidence for cryptic speciation, the enrichment for reproductive genes among those segregating nonsynonymous variation and evolving directionally in *An. funestus* provides another line of evidence that this may be the case and warrants further investigation.

More broadly, *An. minimus* stands apart from the two African species in several respects. It shows markedly less non-neutral evolution in the McDonald-Kreitman analysis (15.8% of orthologs compared to ∼24% in *An. gambiae* and *funestus*), and its GO enrichments for nonsynonymous polymorphism center on broad regulatory processes rather than the specific metabolic and immune pathways identified in the African species. Several factors may contribute to this difference. *An. minimus* occupies a distinct ecological niche in the forested landscapes of Southeast Asia, with different pathogen exposure and a different history of insecticide deployment than the African vectors. The shorter branch length between *An. minimus* and *An. funestus* may also reduce power to detect lineage-specific adaptive divergence. Additionally, our *An. minimus* samples are drawn from a single country (Cambodia), which may limit our ability to detect the spatially varying selection that is more apparent across the continental-scale sampling of the African species.

Overall, the patterns of nonsynonymous variation and directional selection that we observe in *Anopheles* reflect both their population histories and the convergent selection pressures they are facing across their ranges. In these highly polymorphic, large N_e species, we see evidence for both local adaptation and balancing selection maintaining abundant nonsynonymous polymorphism that provides a reservoir of potentially globally-adaptive variants. The same functions and processes that segregate high amounts of nonsynonymous variation are also evolving directionally on longer timescales, with even some of the same genes showing parallel adaptive divergence across lineages. This parallels the pattern observed in *Drosophila*, where species harbor comparable levels of nonsynonymous polymorphism and also show strong evidence of adaptive protein evolution and parallel adaptation to environmental pressures (Langley et al., 2012; Zhao et al., 2015). Indeed, Zhao

and Begun (2017) found substantial parallelism in *Drosophila* adaptation across two timescales: the same genes that show clinal differentiation within species are enriched among those evolving under directional selection between species, a finding strikingly consistent with our observation that functional categories harboring excess nonsynonymous polymorphism in *Anopheles* are also targets of lineage-level adaptive divergence. However, there are notable differences. The proportion of genes evolving under directional selection is much smaller in *Anopheles* than in *Drosophila* despite much deeper divergence times, possibly reflecting lower rates of amino acid fixation between *Anopheles* lineages and the different ecological histories of these systems. At the same time, the high proportion of genes in *Anopheles* that harbor an excess of nonsynonymous polymorphism is not observed in *Drosophila* despite similar effective population sizes (Langley et al., 2012), providing further evidence that non-neutral processes actively maintain this variation in *Anopheles*. The specific functional targets of adaptation also differ: where *Drosophila* lineages show enrichment for general species divergence functions such as translation and nuclear transport, *Anopheles* lineages are enriched for genes relevant to local adaptation, immune response, and insecticide resistance. These common patterns of selection across deeply diverged lineages highlight the impact of parallel and convergent evolution in shaping the genomic landscape of these vectors.

Through local adaptation and balancing selection, *Anopheles* species maintain abundant nonsynonymous polymorphism in the same functional categories that are evolving under directional selection. Genes involved in insecticide resistance, immune response, cuticle formation, sensory perception, and redox metabolism both segregate excess nonsynonymous polymorphism and are diverging under directional selection along *Anopheles* lineages, pointing to a scenario where standing variation provides

a reservoir of potentially globally-adaptive variants. The forces that maintain this variation and drive divergence are shared across *Anopheles* lineages, and we observe many of the same genes diverging under parallel directional selection. Most notably, we find evidence for convergent local adaptation to insecticide pressures in *An. gambiae* and *funestus*, with two cytochrome P450 genes segregating molecularly similar mutations in the same geographic regions in both species. These results demonstrate the adaptive potential of these mosquito vectors, highlighting the importance of standing variation in their ability to respond to selection pressures across their ranges. In *Anopheles* malaria vectors, where the strongest selection pressures are often driven by human intervention, understanding their evolutionary dynamics will be critical for anticipating and managing resistance as vector control efforts continue to intensify.

CHAPTER V

CONCLUSION

This dissertation sets out to leverage spatial patterns of genetic structure in *Anopheles* malaria vectors in order to uncover the evolutionary dynamics driving adaptation and divergence across their ranges. With continuous space simulation, I develop intuition and approaches for analyzing spatial genetic datasets, then apply these methods to empirical datasets of *Anopheles* to uncover novel insights into the evolutionary histories and trajectories of these important vectors. These findings have relevance for understanding future vector control strategies in this system, and the approaches I present are applicable to a wide range of spatially-structured populations beyond mosquitoes.

In Chapter II, I present our work using correlated patterns of machine learning prediction errors to infer common patterns of dispersal between *Anopheles gambiae* mosquitoes and their *Plasmodium falciparum* parasites. Applying `Locator` — a machine learning method for predicting geographic origin from genetic data — to spatially-referenced genomic datasets of both vector and parasite, I find that prediction errors are significantly correlated across space between the two species, suggesting that they share common dispersal patterns. Using continuous-space simulation, I investigate the circumstances that can lead to biased errors in spatial prediction and develop methods to reduce these effects when applying this approach to empirical data. After correcting for spatial bias in empirical sampling data, correlations between *Anopheles gambiae* and *Plasmodium falciparum* residuals persist, providing evidence that these prediction errors reflect shared dispersal patterns between vector and parasite. This work uncovers previously unobserved co-geographic structure shared by the malaria parasite and its vector on short

evolutionary timescales, with implications for understanding disease transmission dynamics and vector control strategies in this system.

Given these findings, I became interested in how adaptive alleles are spreading through these spatially-structured populations. In Chapter III, I use continuous-space simulation of selective sweeps to discover a spatial signal of positive selection: I find that due to the wave-like nature of their spread, positively-selected alleles take up less geographic area than neutral alleles at the same frequency. Using this observation, I develop a method to detect adaptive alleles as 'space-frequency' outliers in spatially-structured populations. Applying this method to *Anopheles gambiae* populations in West Africa, I identify several novel alleles that appear to be under positive selection, including several functionally-relevant variants at loci associated with insecticide resistance. This work demonstrates the potential of leveraging spatial population genetic structure to identify evolutionary processes, and the method I present here has relevance for detecting adaptive alleles in a wide range of spatially-structured populations. In *Anopheles gambiae* in particular, this work provides insight into novel genomic features that may be contributing to insecticide resistance and introduces a novel approach for identifying adaptive variants in this system.

Anopheles gambiae is one of many *Anopheles* species that transmit malaria, and many of these species are segregating large amounts of potentially-adaptive variation. In Chapter IV, I use a gene-focused comparative population genetic approach to investigate the evolutionary forces that shape patterns of variation and divergence in three major *Anopheles* malaria vectors: African *Anopheles gambiae* and *funestus*, and Asian *minimus*. Focusing on patterns of nonsynonymous polymorphism and divergence reveals common selection pressures acting on these deeply-diverged vectors to both maintain functional variation within and drive adaptive divergence between

species, with immune and metabolic functions being particularly affected across a broad range of evolutionary timescales. The biological functions that show evidence of both diversifying and directional selection are often shared across species and include many mechanisms of insecticide resistance, such as metabolic detoxification and cuticular resistance. Investigating the spatial patterns of diversifying selection acting on *Anopheles gambiae* and *funestus*, I uncover two proteins involved in insecticide resistance that share convergent spatial and molecular patterns of variation, demonstrating how local adaptation maintains functional variation in these relevant pathways. This work provides insight into the parallel evolutionary forces that shape patterns of variation across these deeply-diverged malaria vectors, highlighting how the functional polymorphism maintained by local adaptation and balancing selection provides an abundant source of potentially-adaptive variation for rapid evolutionary responses to novel or intensified selection pressures.

Overall, this dissertation provides insight into how spatial population structure can be leveraged in genetic studies to uncover evolutionary phenomena in the *Anopheles* malaria vector system and beyond. In Chapter II, I use correlated patterns of machine learning prediction errors to identify and confirm co-geographic structure between *Anopheles gambiae* and *Plasmodium falciparum*, identifying a previously undescribed pattern of shared dispersal between vector and parasite. In Chapter III, I use continuous-space simulation to identify a novel spatial signature of selective sweeps and apply this observation to identify positively-selected alleles in *Anopheles gambiae*. Finally, in Chapter IV, I use a comparative, gene-focused approach to investigate the evolutionary forces shaping patterns of variation and divergence in three major *Anopheles* vectors, uncovering common selection pressures across these deeply-diverged species and identifying convergent local adaptation to insecticide

pressures. The novel methods and approaches I present in this dissertation have relevance for a wide range of spatially-structured populations, and the insights into the *Anopheles* system I present here further our understanding of these important vectors and may inform future vector control strategies.

APPENDIX A

SUPPLEMENTAL MATERIAL FOR CHAPTER II: EVALUATING EVIDENCE FOR CO-GEOGRAPHY IN THE *ANOPHELES-PLASMODIUM* HOST-PARASITE SYSTEM

Supplementary material

Windowed analysis and uncertainty. While our analysis focused on trends across individual predictions in a dataset, we were also interested in how estimated uncertainty is affected by training data. By visualizing the spatial spread of predictions using different portions of the genome, windowed analysis can be used to assess the uncertainty in individual-level predictions (Battey et al., 2020a). When examining the windowed predictions for individuals in our simulated datasets, we observed a negative correlation between local density of training data and the spatial spread of an individual's predictions across genomic windows, with this trend becoming more pronounced as training sets became more imbalanced (Figure A.4A). When trained on uniformly-sampled data (bias = 0.5) with a dispersal rate of 1.0, the relationship between spatial density of training data and spatial spread of windowed predictions for a test sample is weakly negatively correlated: $r = -.10, p = 0.02, n = 500$. With a training set bias of 0.9, the relationship is much stronger: $r = -.67, p = 2.8 \times 10^{-65}, n = 500$.

Effect of training set size. We also assessed how increasing the number of training samples affects `Locator`'s performance on spatially imbalanced datasets (Figure A.4, panels B-D). Initial analysis on randomly sampled training sets had indicated that 450 training samples were sufficient for accurate inference of location (Battey et al., 2020a), but this number of training samples did not provide the same results when sampled unevenly from the landscape. Unsurprisingly, increasing

training set size reduced the effects of training set imbalance, even in populations undergoing a high rate of dispersal (1.0 units per generation). While residuals are still biased - correlation between x-axis error and training set skew remains significant with 4,500 training samples ($r = 0.72, p = 3.92 \times 10^{-9}$) - the magnitude of these errors is reduced, likely reflecting the overall increase in training data.

Sample weighting hyperparameters. During our grid search, we observed an inverse relationship between the magnitude of bandwidth and λ parameters required to reduce residual bias when training with spatially imbalanced data. At high values of λ, low bandwidth resulted in reduced bias (as measured by error along the x-axis), while overall prediction accuracy remained unaffected (as measured by y-axis error and residual magnitude). As bandwidth was increased, the value of λ that resulted in reduced prediction bias while maintaining accuracy decreased consistently. These combinations form a diagonal across the parameter space, with `Locator` performing differently on either side of this threshold (Figure A.5).

When both λ and bandwidth were low, sample weights were relatively uniform across the landscape, which had a negligible effect on training outcomes: in the space beneath optimal λ/bandwidth combinations, `Locator`'s predictions were similar to those from unweighted training runs (Figure A.6 C). Above the optimal threshold, where both λ and bandwidth values are high, performance was more inconsistent: residuals were biased along the x-axis and y-axis, and overall error magnitude increased well beyond that observed when training without sample weights (Figure A.6 E). With these parameter combinations, the majority of samples were assigned extremely low weights, causing `Locator` to only focus on a few samples near the landscape boundaries during training and predict samples to be in some landscape midpoint.

Comparison to SPASIBA Prior to our introduction of Locator, the state-of-the-art statistical method for geographic assignment using genotype data was perhaps SPASIBA (Guillot et al., 2015). SPASIBA uses a clever implementation of a geostatistical model which learns smooth surfaces of allele frequencies over a landscape, and then predicts locations of unlabeled samples by maximizing their likelihood on this surface. We were interested to compare the performance of SPASIBA and Locator when using spatially imbalanced training data, in particular because biased training sets may affect traditional statistical and machine learning methods to a greater or lesser extent.

In Figure A.7 we show results from SPASIBA applied to training sets of 10,000 SNPs drawn from samples of 450 individuals. In some cases we were only able to execute seven SPASIBA analyses for a given training set bias, as compared to ten Locator analyses, because jobs were killed after 30 days of execution. As with Locator, SPASIBA's predictions were biased towards the right half of the landscape when run on imbalanced training sets. However, SPASIBA's predictions suffered significant collapse towards the center of the sampling density in some cases, and often covered a small fraction of the landscape, with test individuals across a run all predicted to be in only a few unique locations. These results suggest that SPASIBA may struggle to a greater extent than Locator in the face of imbalanced training data, and more generally that biased training sets can influence prediction in traditional statistical as well as machine learning settings.

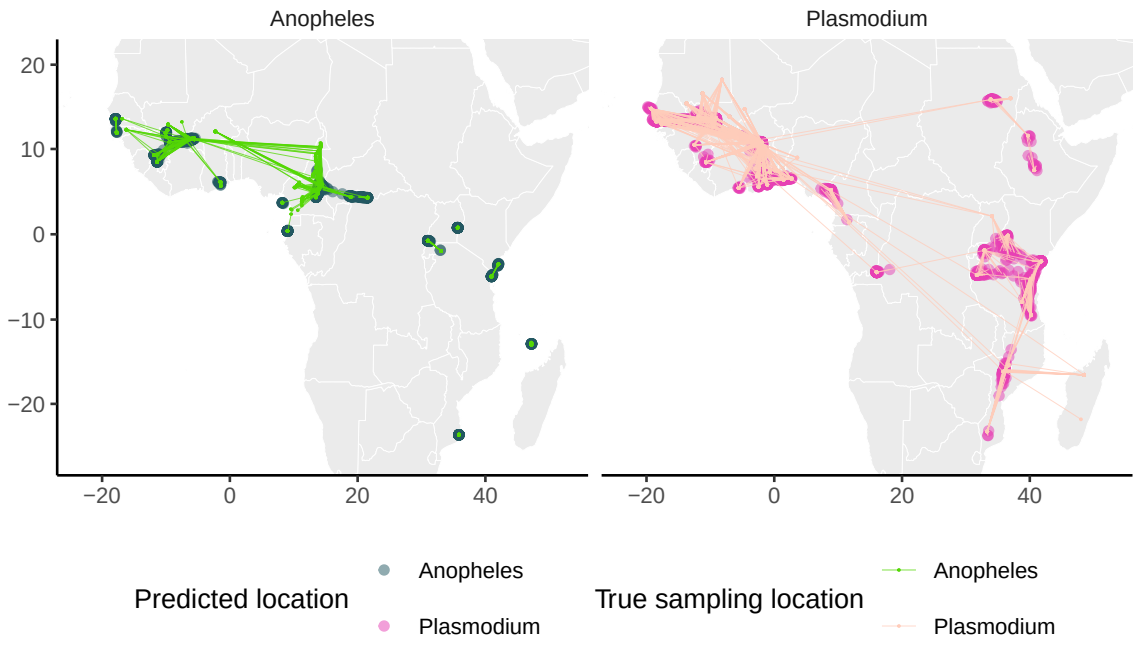


Figure A.1. Predicted locations of all *A. gambiae* and *P. falciparum* samples. A sample's true location is connected to the geographic centroid of its genome-wide windowed predictions.

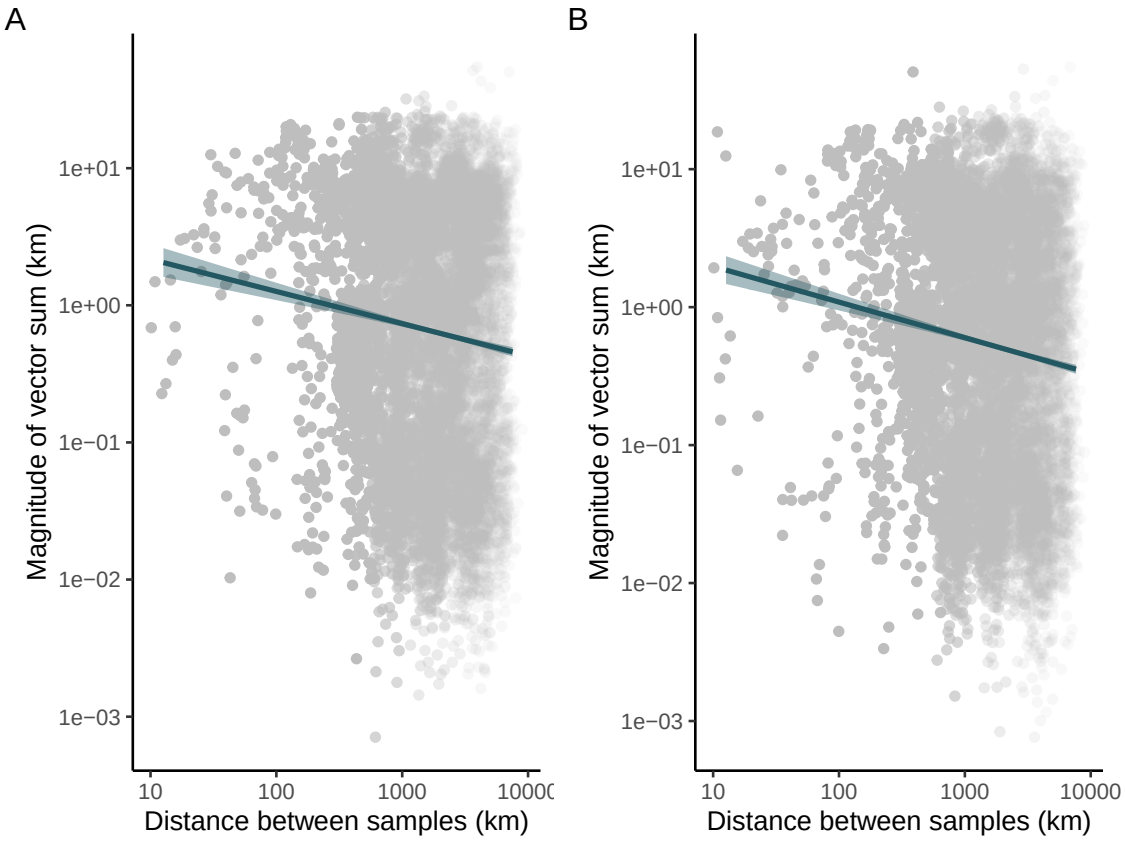


Figure A.2. The magnitude of summed *A. gambiae* and *P. falciparum* residual vectors plotted against the distance between true sample locations. For both unweighted (A) and weighted (B) analyses, distance between samples is negatively correlated with the magnitude of summed residual vectors.

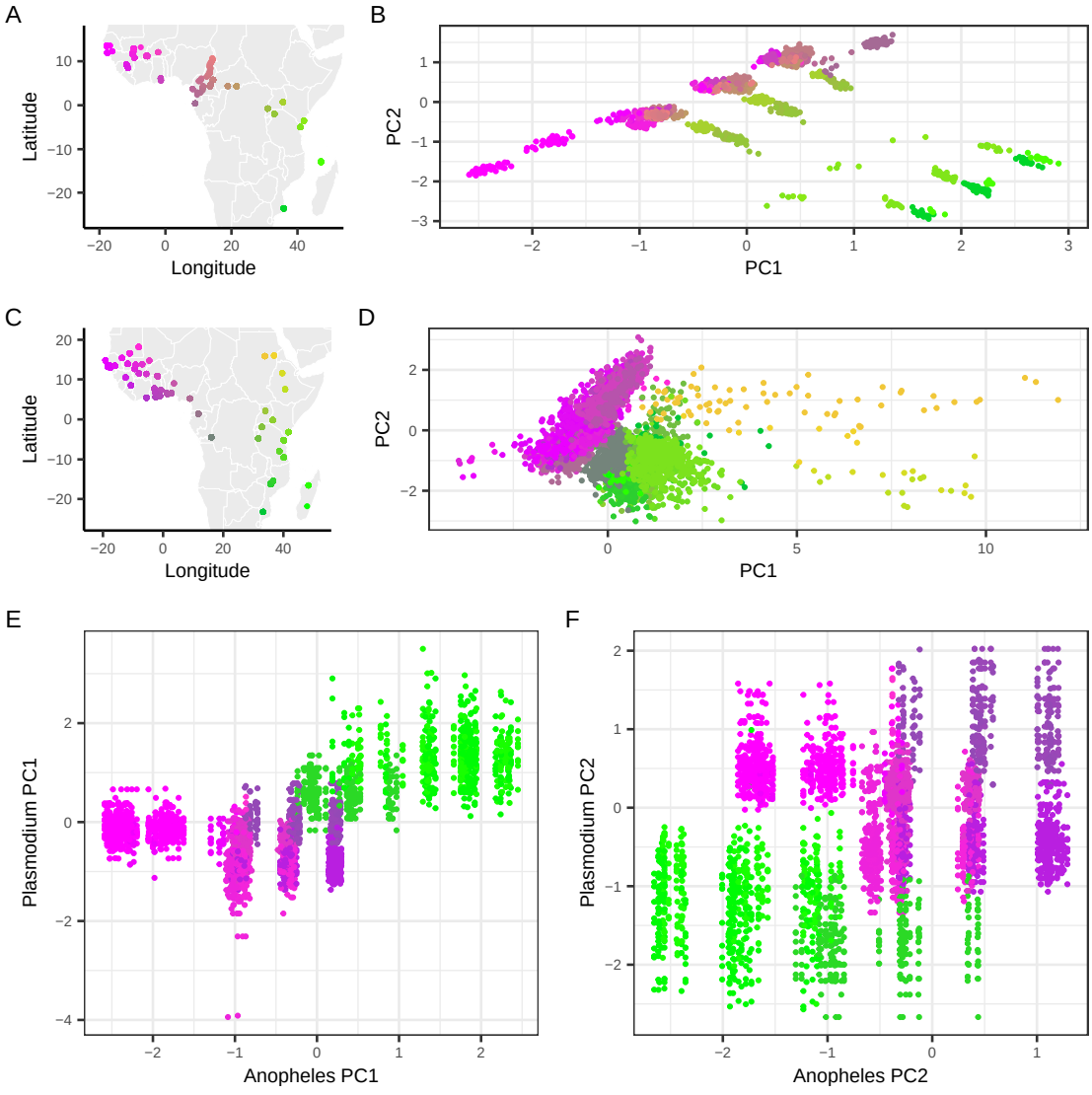


Figure A.3. Spatially paired *A. gambiae* and *P. falciparum* principal components. (A) *A. gambiae* sampling locations and (B) principal components, with each sample colored by its spatial location. (C) *P. falciparum* sampling locations and (B) principal components, with each sample colored by its spatial location. (E) Relationship between the first principal component of spatially-paired *A. gambiae* and *P. falciparum* samples, with each pair colored by their spatial location. (F) Relationship between the second principal component of spatially-paired samples.

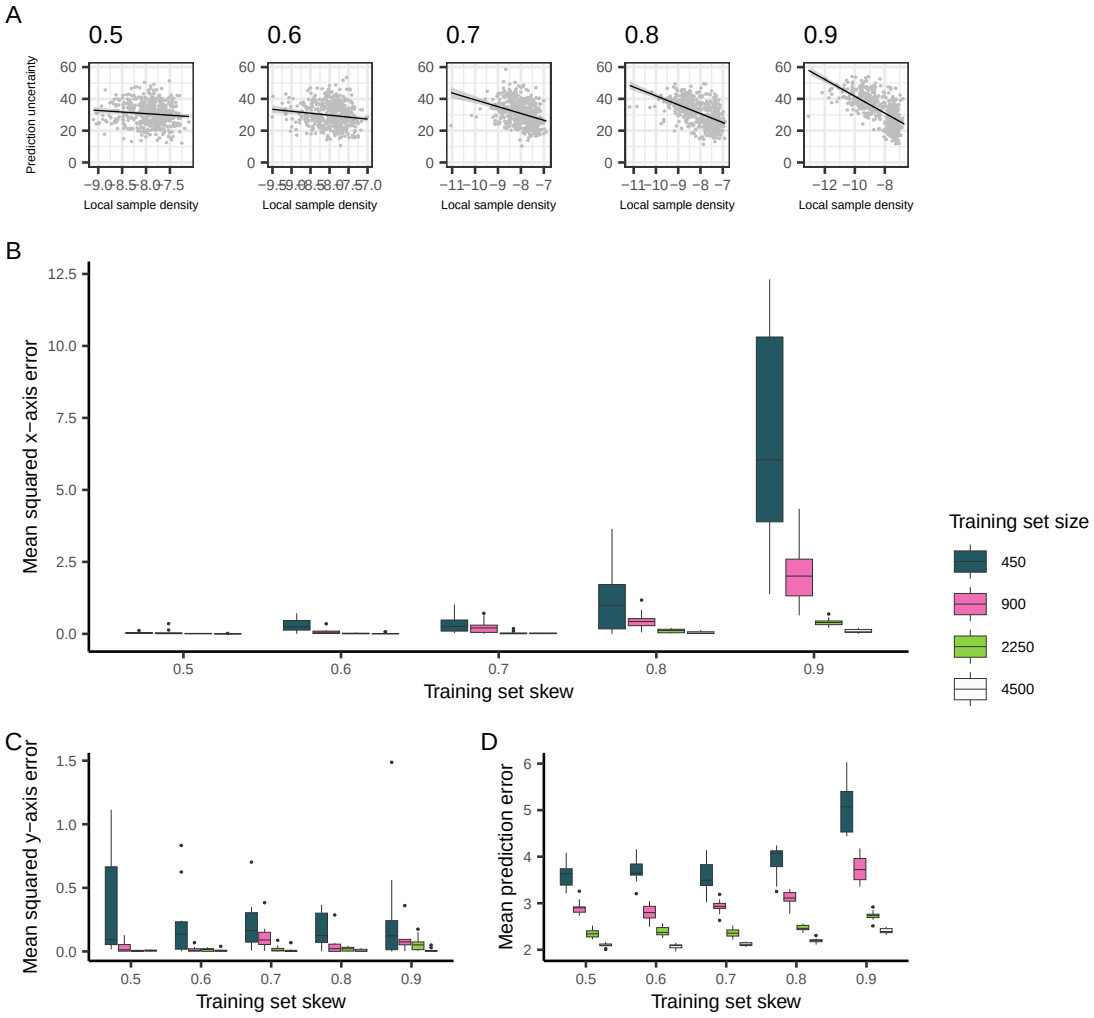


Figure A.4. The effect of training set bias and size on `Locator` prediction error at a dispersal rate of $\sigma = 1.0$. (A) Spatial spread of per-individual windowed predictions as a function of local training set density at different levels of training set bias. (B) Mean per-run prediction error along the x axis as a function of training set bias, (C) mean per-run prediction error along the y axis, (D) mean per-run error.

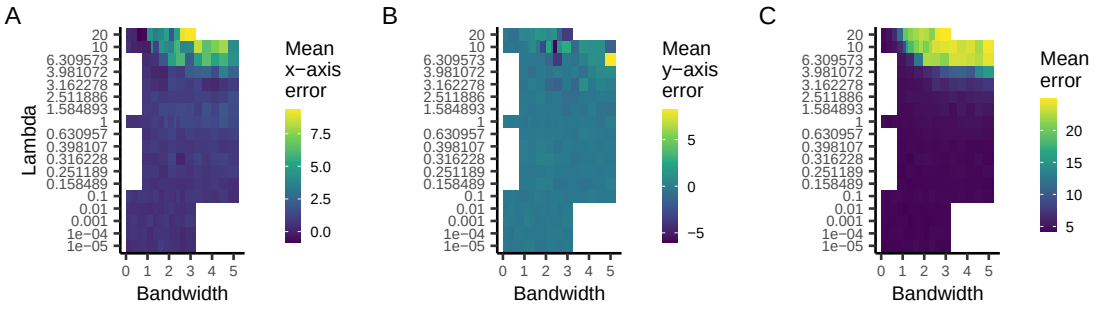


Figure A.5. Results of grid search for λ and bandwidth parameters on $bias = 0.9$, $\sigma = 1.0$ training sets. (A) Mean x-axis error, (B) mean y-axis error, and (C) mean error for all tested combinations ofλ and bandwidth.

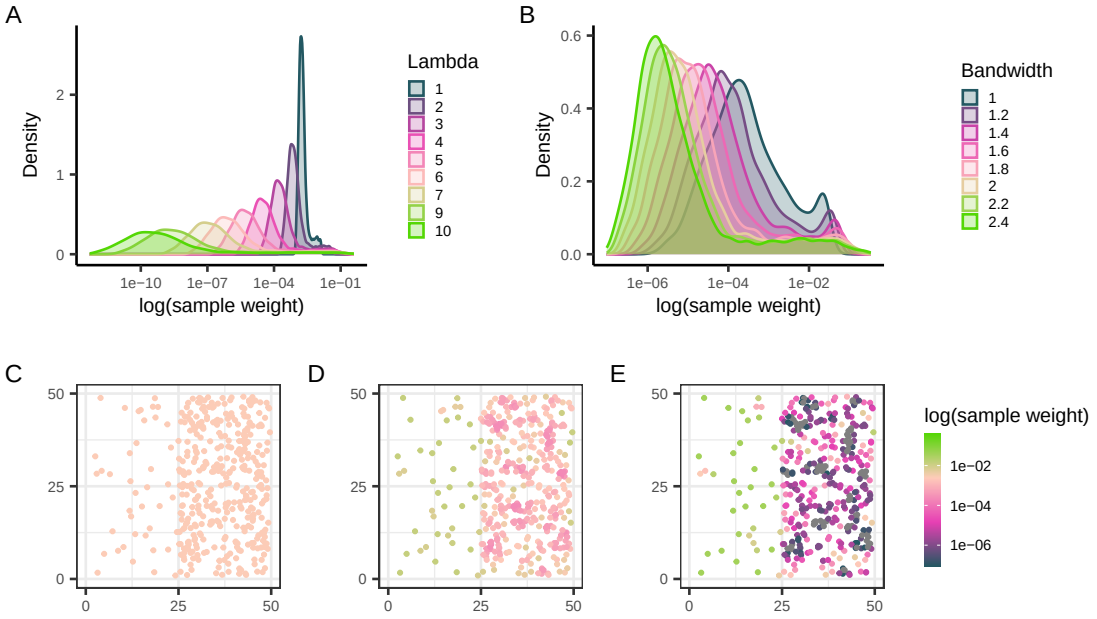


Figure A.6. Bandwidth and λ parameters alter the distribution of sample weights. (A) Distribution of sample weights assigned to a given training set holding bandwidth constant and varying λ. (B) Distribution holding λ constant and varying bandwidth. (C, D, E) Biased training sets colored by their sample weights as λ increases.

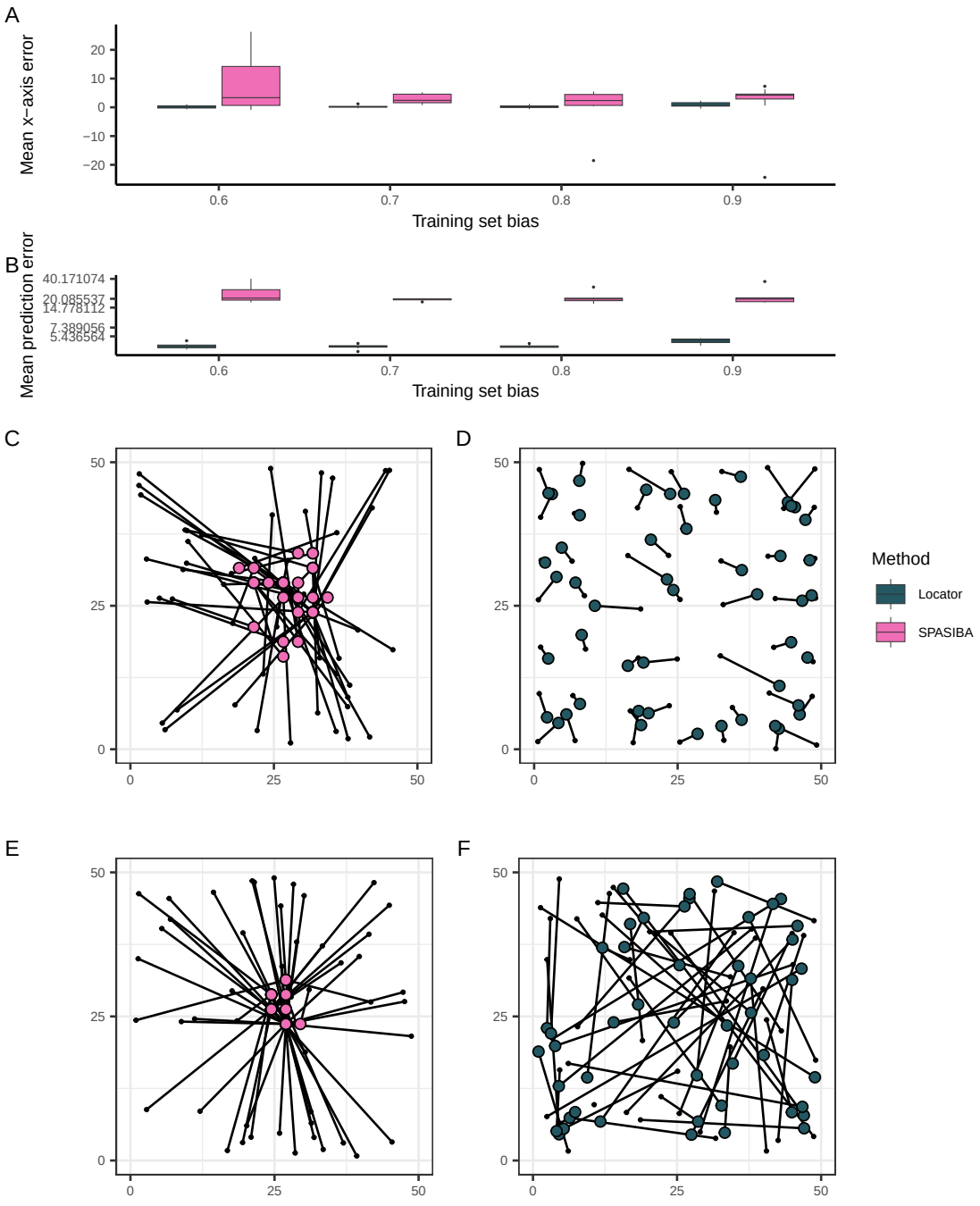


Figure A.7. Predictions from biased training data using `SPASIBA` and `Locator`. Simulations had a per-generation dispersal rate of 1.0 units. (A) Mean x-axis error as a function of training set bias and (B) mean prediction error as a function of training set bias. (C, D) Example `Locator` and `SPASIBA` at $bias = 0.5$ and (E, F) $bias = 0.9$. True sample locations (black points) are connected to their predicted location (colored points).

APPENDIX B

SUPPLEMENTAL MATERIAL FOR CHAPTER III: SWEEPS IN SPACE: LEVERAGING GEOGRAPHIC DATA TO IDENTIFY BENEFICIAL ALLELES IN *ANOPHELES GAMBIAE*

B.1 Approximation to Fisher's wave

In order to compare the allele frequency and area trajectories of our simulated sweeps to a deterministic model, we modeled an approximation of a positively-selected allele spreading across two-dimensional space as a Fisher wave (R. A. Fisher, 1937). The positively-selected allele originates at the center of a 25x25-unit square landscape and spreads in expanding circle whose radius r increases at a velocity $V = M\sqrt{S}$ where M is the rate of per-generation dispersal (similar to the σ parameter in the SLiM simulation) and S is the selection coefficient of the allele. We can therefore define the radius of the circle representing the extent of the sweeping allele at time point t as $r_{edge}(t) = M\sqrt{S}t$ and use this to calculate the area occupied by the allele (Figure B.1A). The width of the wavefront is defined as $W = 1/4\sqrt{S}$, thus the spatial extent the fixed allele at time point t can be defined as $r_{fixed}(t) = M\sqrt{S}t - \frac{\sqrt{S}}{4}$. The circle representing the sweeping wavefront exits the landscape at $R = 12.5$ and completely encompasses the landscape at $R_{max} \approx 17.68$. If $r_{edge}(t) \leq R$, the area of the sweeping allele is simply $\pi r_{edge}(t)^2$. If $r_{edge}(t) > R$, the area of the sweeping allele is defined as $(\pi - 4\arccos(\frac{R}{r_{edge}}))r_{edge}^2 + 4Rr_{edge}(t)\sin(\arccos(\frac{R}{r_{edge}(t)}))$.

To approximate the frequency of the allele, we modeled the sweeping wave in three dimensions as a conical frustum - a cone sliced horizontally - with height of 1 along the z axis, an upper radius r_{fixed}, and base radius of r_{edge}, and calculated its frequency out of the $25 \times 25 \times 1$ volume of the landscape (Figure B.1B). If $r_{edge}(t) < R$, then the volume of the conical frustum, $V(t)$, is equal to $\frac{\pi(M\sqrt{S}t)^2}{3} - \frac{\pi(M\sqrt{S}t - \frac{\sqrt{S}}{4})^2}{3}$. If

$R \leq r_{edge} \leq R_{max}$ and the base begins to exit the landscape border, then

$$V(t) = (\frac{\pi(M\sqrt{S}t)^2}{3} - \frac{\pi(M\sqrt{S}t - \frac{\sqrt{S}}{4})^2}{3}) - (4(\frac{1}{3}(-2R\sqrt{r_{edge}^2 - R^2}) + (\frac{R^3}{r_{edge}} \ln(\sqrt{r_{edge}^2 - R^2}) + r_{edge}) + (r_{edge}^2 \arccos\frac{R}{r_{edge}} - \frac{R^3}{r} \ln R))), \quad \text{(B.1)}$$

removing the volumes of the frustum that are outside the landscape. Finally, if $r_{edge} > R_{max}$ and the base completely encompasses the landscape, the volume is calculated as the sum of the volume of a 25×25 rectangular prism with height $\frac{r_{edge}(t) - R_{max}}{r_{edge}(t) - r_{fixed}(t)}$ and a conical frustum with height $1 - \frac{r_{edge}(t) - R_{max}}{r_{edge}(t) - r_{fixed}(t)}$, base radius R_{max}, and upper radius $r_{fixed}(t)$ (after subtracting off volumes of the frustum outside the landscape as described in Equation S1).

Our model recapitulates both the logistic frequency trajectories of a sweep (Figure B.1C) and the relationship between selection, frequency, and area that we observe in our continuous-space simulations (Figure B.1D). We note that the key to the key to the nonlinear relationship between an allele's frequency and area depends on the wavefront where the allele is still segregating; as the strength of selection increases and the width of the wavefront goes to zero the relationship between frequency and area becomes linear.

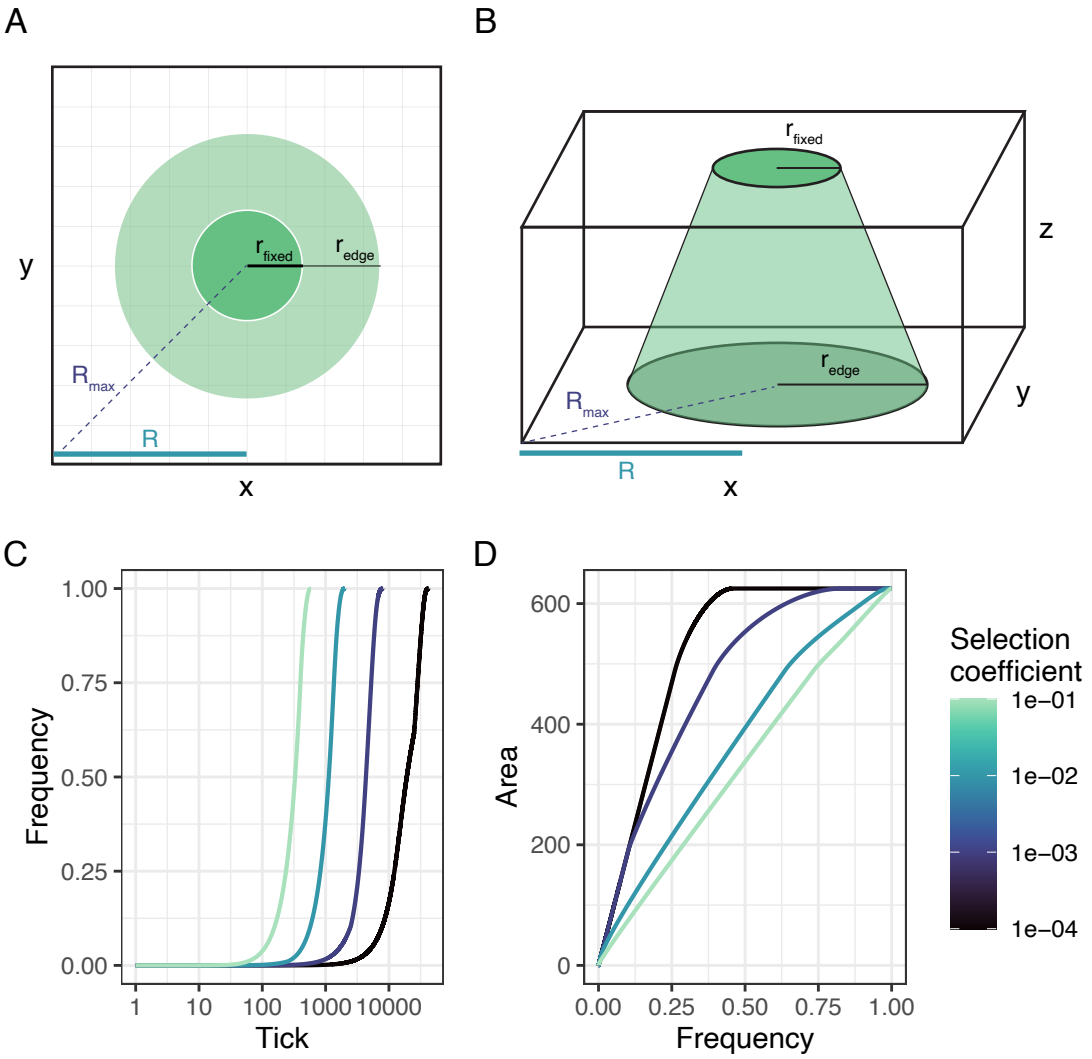


Figure B.1. Approximation of a positively-selected allele spreading according to a Fisher wave across two-dimensional space. (A) Top-down view of the extent of the sweeping allele (r_{edge}) and the extent of the allele's fixation (r_{fixed}). (B) Side view of the method used to approximate the frequency of the sweeping allele as the volume of the conical frustum defined by r_{edge} and r_{fixed} out of the volume of the landscape defined by x, y, and z. (C) Frequency of the positively-selected allele. (D) Joint distribution between frequency and area of the positively selected allele. In both plots, lines are colored by selection coefficient (S).

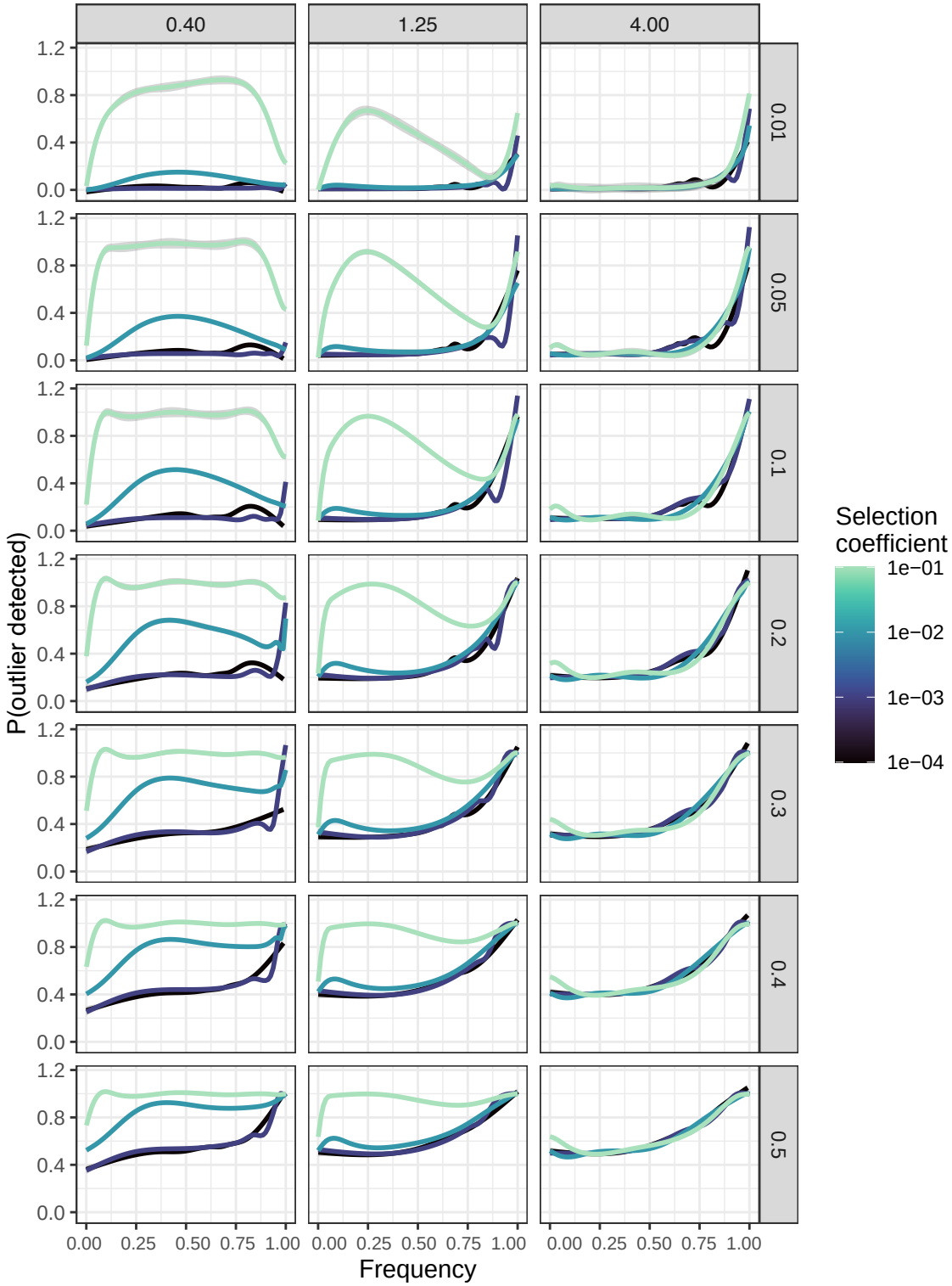


Figure B.2. Probability of detecting a sweeping allele as a spatial outlier over the course of a sweep across dispersal distances (x axis) and cutoff percentiles (y axis). Lines are colored by selection coefficient.

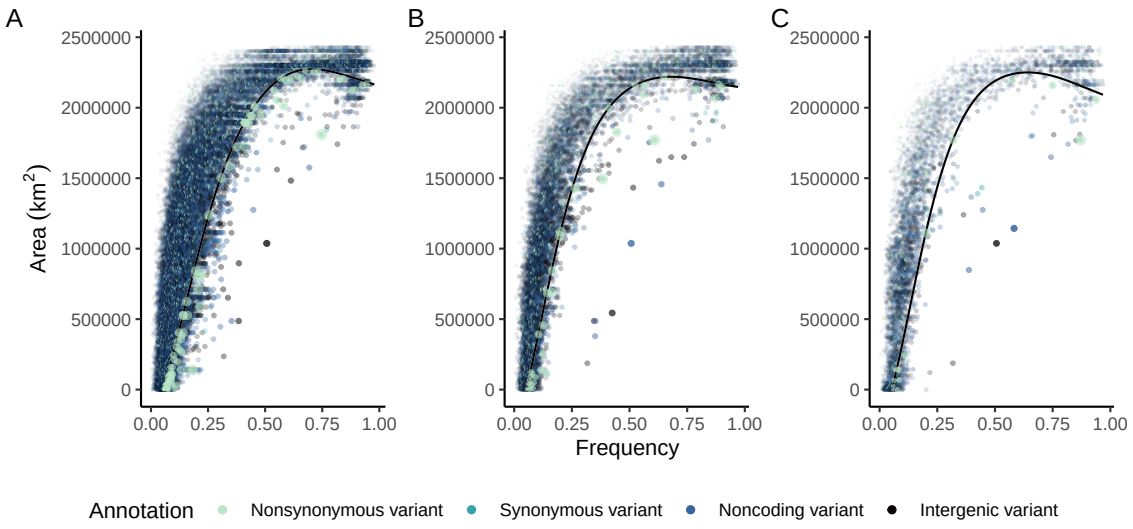


Figure B.3. Joint frequency-area distribution for *An. gambiae* SNPs within the 2La (A), 2Rb (B), and 2Rc (C) inversions. SNPs are colored according to their SNPEff annotation; black lines represent the 10th percentile cutoff.

APPENDIX C

SUPPLEMENTAL MATERIAL FOR CHAPTER IV: CONVERGENT SELECTION ON NONSYNONYMOUS VARIATION ACROSS DEEPLY DIVERGED *ANOPHELES* VECTORS

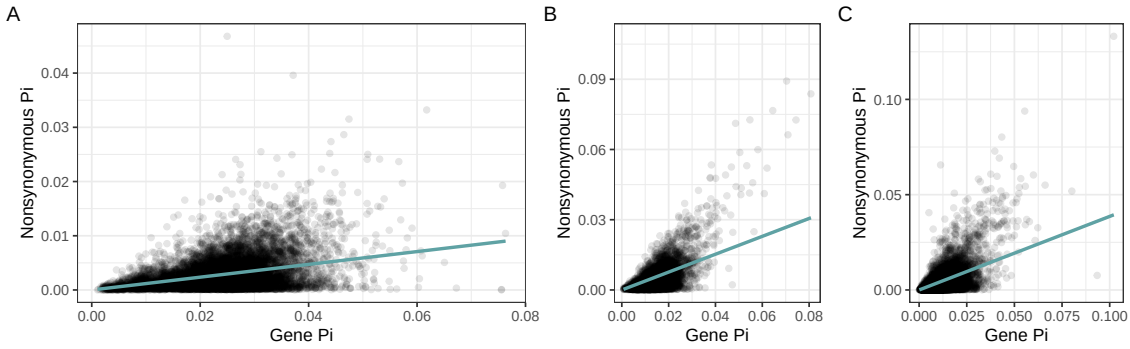


Figure C.1. Relationship between π_N and π_G in *An. gambiae* (A), *funestus* (B), and *minimus* (C).

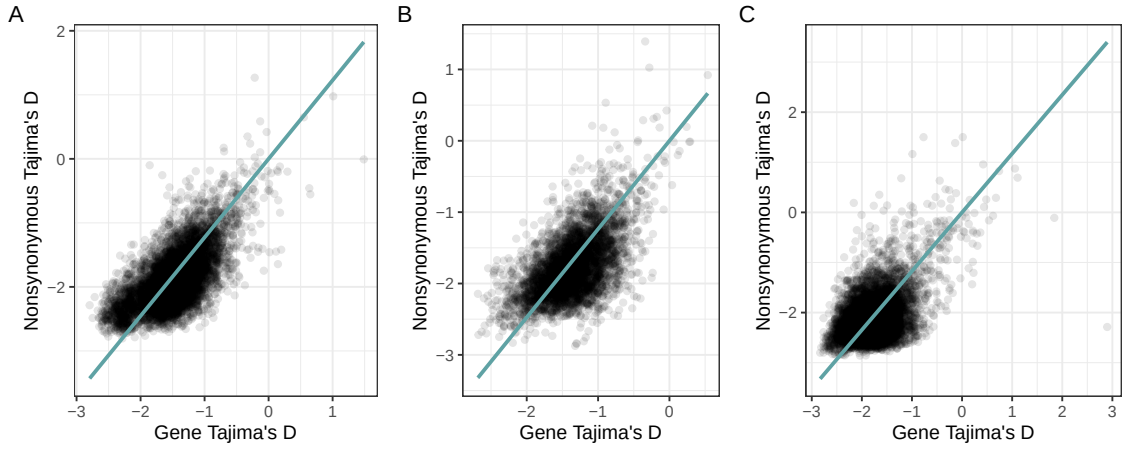


Figure C.2. Relationship between D_N and D_G in *An. gambiae* (A), *funestus* (B), and *minimus* (C).

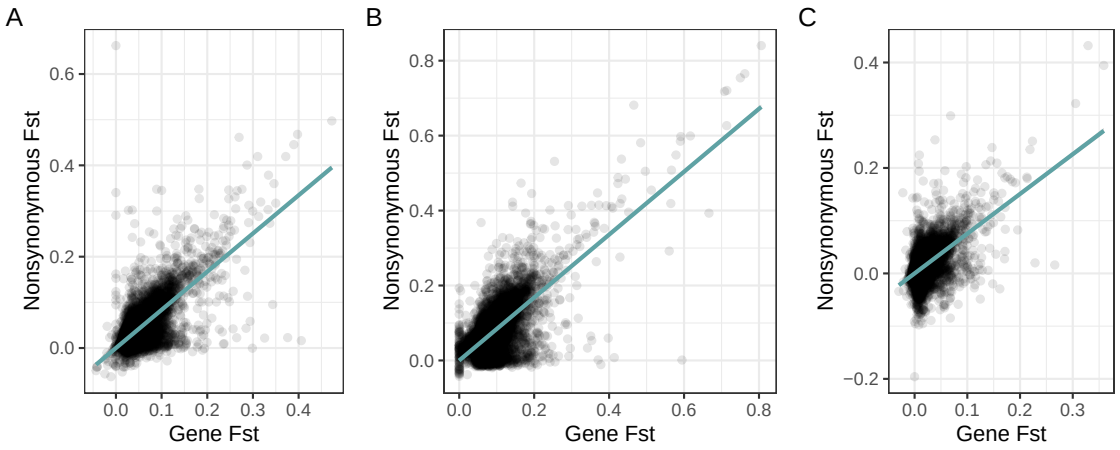


Figure C.3. Relationship between $F_{ST_{NS}}$ and $F_{G_{NS}}$ in *An. gambiae* (A), *funestus* (B), and *minimus* (C).

C.1 GO enrichments for π_N

GO ID	Term	p value
GO:0007608	sensory perception of smell	0.00024
GO:0006508	proteolysis	0.00793
GO:0006310	DNA recombination	0.01196
GO:0006259	DNA metabolic process	0.01383
GO:0006281	DNA repair	0.02287
GO:0006284	base-excision repair	0.02768
GO:0006974	DNA damage response	0.02866
GO:0030150	protein import into mitochondrial matrix	0.03312
GO:0006029	proteoglycan metabolic process	0.03854
GO:0030166	proteoglycan biosynthetic process	0.03854
GO:0170036	import into the mitochondrion	0.03854
GO:0006260	DNA replication	0.04256
GO:0033554	cellular response to stress	0.04316
GO:0000724	double-strand break repair via homologous recombination	0.04393
GO:0000725	recombinational repair	0.04393
GO:0016073	snRNA metabolic process	0.04393
GO:0016180	snRNA processing	0.04393

Table C.1. Significant Biological Process GO term enrichments for *An. gambiae* π_N outliers.

C.1.1 *An. gambiae.*

GO ID	Term	p value
GO:0007281	germ cell development	0.00066
GO:0022412	cellular process involved in reproduction	0.00066
GO:0003006	developmental process involved in reproduction	0.00085
GO:0007276	gamete generation	0.00106
GO:0006325	chromatin organization	0.00207
GO:0007156	homophilic cell adhesion via plasma membrane	0.00727
GO:0098742	cell-cell adhesion via plasma-membrane adhesion	0.00782
GO:0098609	cell-cell adhesion	0.00840
GO:0006814	sodium ion transport	0.01294
GO:0048468	cell development	0.01515
GO:0030154	cell differentiation	0.03040
GO:0048869	cellular developmental process	0.03040
GO:0007155	cell adhesion	0.04726

Table C.2. Significant Biological Process GO term enrichments for *An. funestus* π_N outliers.

C.1.2 An. funestus.

GO ID	Term	p value
GO:0051336	regulation of hydrolase activity	0.00011
GO:0051246	regulation of protein metabolic process	0.00016
GO:0050790	regulation of catalytic activity	0.00060
GO:0065009	regulation of molecular function	0.00093
GO:0000723	telomere maintenance	0.02931
GO:0032200	telomere organization	0.02931

Table C.3. Significant Biological Process GO term enrichments for *An. minimus* π_N outliers.

C.1.3 An. minimus.

C.2 GO enrichments for Tajima's D_N

GO ID	Term	p value
GO:0006508	proteolysis	0.0014
GO:0006749	glutathione metabolic process	0.0054
GO:0006575	modified amino acid metabolic process	0.0092
GO:0043603	amide metabolic process	0.0182
GO:0006790	sulfur compound metabolic process	0.0197
GO:0006098	pentose-phosphate shunt	0.0229
GO:0006740	NADPH regeneration	0.0229
GO:0051156	glucose 6-phosphate metabolic process	0.0257
GO:0006739	NADP metabolic process	0.0285
GO:0007608	sensory perception of smell	0.0353

Table C.4. Significant Biological Process GO term enrichments for *An. gambiae* D_N outliers.

C.2.1 *An. gambiae.*

GO ID	Term	p value
GO:0030301	cholesterol transport	0.027
GO:0032367	intracellular cholesterol transport	0.027
GO:0032365	intracellular lipid transport	0.030
GO:0032366	intracellular sterol transport	0.030
GO:0006096	glycolytic process	0.036
GO:0006195	purine nucleotide catabolic process	0.036
GO:0009134	nucleoside diphosphate catabolic process	0.036
GO:0009137	purine nucleoside diphosphate catabolic process	0.036
GO:0009154	purine ribonucleotide catabolic process	0.036
GO:0009181	purine ribonucleoside diphosphate catabolic process	0.036
GO:0009191	ribonucleoside diphosphate catabolic process	0.036
GO:0009261	ribonucleotide catabolic process	0.036
GO:0015850	organic hydroxy compound transport	0.036
GO:0015918	sterol transport	0.036
GO:0019364	pyridine nucleotide catabolic process	0.036
GO:0046032	ADP catabolic process	0.036
GO:0072526	pyridine-containing compound catabolic process	0.036
GO:0009135	purine nucleoside diphosphate metabolic process	0.039
GO:0009179	purine ribonucleoside diphosphate metabolic process	0.039
GO:0009185	ribonucleoside diphosphate metabolic process	0.039
GO:0046031	ADP metabolic process	0.039
GO:0009132	nucleoside diphosphate metabolic process	0.042
GO:0016052	carbohydrate catabolic process	0.048

Table C.5. Significant Biological Process GO term enrichments for *An. funestus* D_N outliers.

C.2.2 *An. funestus.*

GO ID	Term	p value
GO:0050790	regulation of catalytic activity	0.00044
GO:0051246	regulation of protein metabolic process	0.00112
GO:0043462	regulation of ATP-dependent activity	0.0408
GO:0007606	sensory perception of chemical stimulus	0.0498

Table C.6. Significant Biological Process GO term enrichments for *An. minimus* D_N outliers.

C.2.3 *An. minimus.*

C.3 *An. gambiae* and *funestus* $F_{ST_{NS}}$ correlation

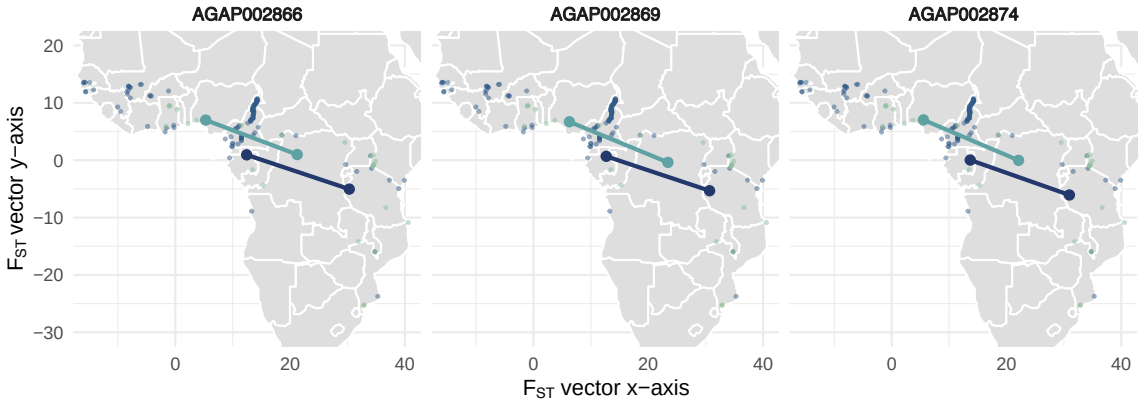


Figure C.4. Significantly correlated $F_{ST_{NS}}$ vectors for three *An. gambiae* (dark blue) and *funestus* (light blue) orthologs. Vectors for each species are depicted via a line connecting two points; sampling locations are depicted in points.

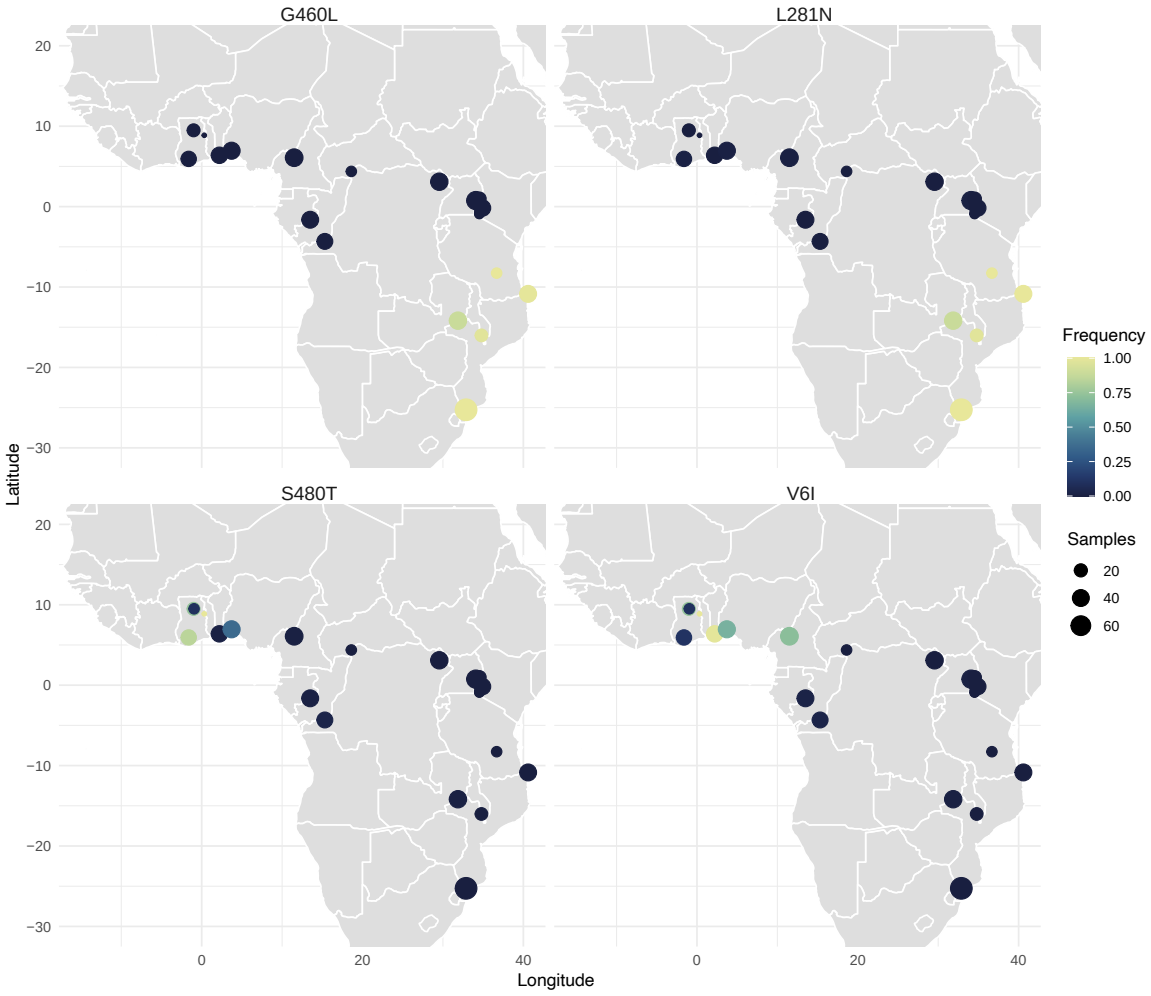


Figure C.5. Restricted spatial distribution of high-frequency nonsynonymous alleles in *An. funestus* gene AFUN2_013651.

Bibliography

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., & et al. (2024). Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the landscape of recombination using deep learning. *Molecular biology and evolution*, *37*(6), 1790–1808.

Ag1000G Selection Atlas. (2017). Ag1000g selection atlas [[Accessed 15-01-2025]].

Al-Asadi, H., Petkova, D., Stephens, M., & Novembre, J. (2019). Estimating recent migration and population-size surfaces. *PLoS genetics*, *15*(1), e1007908.

Alexa, A., & Rahnenfuhrer, J. (2025). *Topgo: Enrichment analysis for gene ontology* [R package version 2.60.1]. https://doi.org/10.18129/B9.bioc.topGO

Alexander, D. H., & Lange, K. (2011). Enhancements to the admixture algorithm for individual ancestry estimation. *BMC bioinformatics*, *12*, 1–6.

Althoff, D. M., Segraves, K. A., & Johnson, M. T. (2014). Testing for coevolutionary diversification: Linking pattern with process. *Trends in Ecology & Evolution*, *29*(2), 82–89.

Alvarez-Jarreta, J., Amos, B., Aurrecoechea, C., Bah, S., Barba, M., Barreto, A., Basenko, E. Y., Belnap, R., Blevins, A., Böhme, U., Brestelli, J., Brown, S., Callan, D., Campbell, L. I., Christophides, G. K., Crouch, K., Davison, H. R., DeBarry, J. D., Demko, R., ... Zheng, J. (2024). Veupathdb: The eukaryotic pathogen, vector and host bioinformatics resource center in 2023. *Nucleic Acids Research*, *52*(D1), D808–D816. https://doi.org/10.1093/nar/gkad1003

Angst, P., Ebert, D., & Fields, P. D. (2022). Demographic history shapes genomic variation in an intracellular parasite with a wide geographical distribution. *Molecular Ecology*, *31*(9), 2528–2544.

Anopheles gambiae 1000 Genomes Consortium. (2017). Genetic diversity of the african malaria vector anopheles gambiae. *Nature*, *552*(7683), 96.

Anopheles gambiae 1000 Genomes Consortium, N. (2017). Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature*, *552*(7683), 96–100. https://doi.org/10.1038/nature24995

Armistead, J. S., Wilson, I. B., van Kuppevelt, T. H., & Dinglasan, R. R. (2011). A role for heparan sulfate proteoglycans in plasmodium falciparum sporozoite invasion of anopheline mosquito salivary glands. *Biochemical Journal*, *438*(3), 475–483. https://doi.org/10.1042/BJ20110694

Arnab, S. P., Amin, M. R., & DeGiorgio, M. (2023). Uncovering footprints of natural selection through spectral analysis of genomic summary statistics. *Molecular Biology and Evolution*, *40*(7), msad157.

Barber, C. B., Dobkin, D. P., & Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, *22*(4), 469–483. https://doi.org/10.1145/235815.235821

Barnes, K. G., Weedall, G. D., Ndula, M., Irving, H., Mzihalowa, T., Hemingway, J., & Wondji, C. S. (2017). Genomic footprints of selective sweeps from metabolic resistance to pyrethroids in African malaria vectors are driven by scale up of insecticide-based vector control. *PLOS Genetics*, *13*(2), e1006539. https://doi.org/10.1371/journal.pgen.1006539

Barton, N. H., Etheridge, A. M., Kelleher, J., & Véber, A. (2013). Genetic hitchhiking in spatially extended populations. *Theoretical Population Biology*, *87*, 75–89. https://doi.org/10.1016/j.tpb.2012.12.001

Barton, N. (2001). The evolutionary consequences of gene flow and local adaptation: Future approaches. In *Dispersal*.

Battey, C. J., Ralph, P. L., & Kern, A. D. (2020). Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data. *Genetics*, *215*(1), 193–214. https://doi.org/10.1534/genetics.120.303143

Battey, C., Ralph, P. L., & Kern, A. D. (2020a). Predicting geographic location from genetic variation with deep neural networks. *eLife*, *9*, e54507.

Battey, C., Ralph, P. L., & Kern, A. D. (2020b). Predicting geographic location from genetic variation with deep neural networks (P. J. Wittkopp, M. Nordborg, C. Huber, A. Harpak & L. Speidel, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *9*, e54507. https://doi.org/10.7554/eLife.54507

Battey, C., Ralph, P. L., & Kern, A. D. (2020c). Space is the place: Effects of continuous spatial structure on analysis of population genetic data. *Genetics*, *215*(1), 193–214.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., … Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229. https://doi.org/10.1093/genetics/iyab229

Bhatt, S., Weiss, D., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C., Henry, A., Eckhoff, P., et al. (2015). The effect of

malaria control on plasmodium falciparum in africa between 2000 and 2015. *Nature*, *526*(7572), 207–211.

Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K., & Rose, Y. (2024). Rcsb protein data bank: Exploring protein 3d similarities via comprehensive structural alignments. *Bioinformatics*, *40*(6), btae370. https://doi.org/10.1093/bioinformatics/btae370

Boddé, M., Nwezeobi, J., Korlević, P., Makunin, A., Akone-Ella, O., Barasa, S., Gadji, M., Hart, L., Kaindoa, E. W., Love, K., Lucas, E. R., Lujumba, I., Máquina, M., Nagi, S. C., Odero, J. O., Polo, B., Sangbakembi, C., Dadzie, S., Koekemoer, L. L., ... Lawniczak, M. K. N. (2025). Genomic diversity of the african malaria vector anopheles funestus. *Science*, *389*(6766), eadu3596. https://doi.org/10.1126/science.adu3596

Bradburd, G. S., & Ralph, P. L. (2019). Spatial population genetics: It's about time. *Annual Review of Ecology, Evolution, and Systematics*, *50*, 427–449.

Bradburd, G. S., Ralph, P. L., & Coop, G. M. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, *67*(11), 3258–3273.

Brito, L. P., Linss, J. G. B., Lima-Camara, T. N., Belinato, T. A., Peixoto, A. A., Lima, J. B. P., Valle, D., & Martins, A. J. (2013). Assessing the effects of aedes aegypti kdr mutations on pyrethroid resistance and its fitness cost. *PLOS ONE*, *8*(4), e60878. https://doi.org/10.1371/journal.pone.0060878

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 77–91.

Carlos, B. C., Voges, K., de Andrade Affonso, P. H., Jaye, A., Tong Rios, C., Tinoco-Nunes, B., Alonso, D. P., MacCallum, R. M., Moreno, M., Vlachou, D., Souza-Neto, J. A., & Christophides, G. K. (2025). Metabolic reprogramming and gut microbiota ecology drive divergent plasmodium vivax infection outcomes in anopheles darlingi. *bioRxiv: The Preprint Server for Biology*, 2025.08.13.670040. https://doi.org/10.1101/2025.08.13.670040

Casida, J. E., & Durkin, K. A. (2013). Neuroactive insecticides: Targets, selectivity, resistance, and secondary effects. *Annual Review of Entomology*, *58*(Volume 58, 2013), 99–117. https://doi.org/10.1146/annurev-ento-120811-153645

Champion, C. J., & Xu, J. (2018). Redox state affects fecundity and insecticide susceptibility in anopheles gambiae. *Scientific Reports*, *8*(1), 13054. https://doi.org/10.1038/s41598-018-31360-2

Charreton, M., Decourtye, A., Henry, M., Rodet, G., Sandoz, J.-C., Charnet, P., & Collet, C. (2015). A locomotor deficit induced by sublethal doses of pyrethroid and neonicotinoid insecticides in the honeybee apis mellifera. *PLOS ONE*, *10*(12), e0144879. https://doi.org/10.1371/journal.pone.0144879

Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome research*, *20*(3), 393–402.

Chevy, E. T., Min, J., Caudill, V., Champer, S. E., Haller, B. C., Rehmann, C. T., Smith, C. C. R., Tittes, S., Messer, P. W., Kern, A. D., Ramachandran, S., & Ralph, P. L. (2024, July). Population genetics meets ecology: A guide to individual-based simulations in continuous landscapes [Pages: 2024.07.24.604988 Section: New Results]. https://doi.org/10.1101/2024.07.24.604988

Chotai, M., Wei, X., & Messer, P. W. (2024, July). Signatures of selective sweeps in continuous-space populations [Pages: 2024.07.26.605365 Section: New Results]. https://doi.org/10.1101/2024.07.26.605365

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *fly*, *6*(2), 80–92.

Clark, R. M., Linton, E., Messing, J., & Doebley, J. F. (2004). Pattern of diversity in the genomic region near the maize domestication gene tb1. *Proceedings of the National Academy of Sciences*, *101*(3), 700–707.

Clarkson, C. S., Miles, A., Harding, N. J., Lucas, E. R., Battey, C. J., Amaya-Romero, J. E., Kern, A. D., Fontaine, M. C., Donnelly, M. J., Lawniczak, M. K., et al. (2020). Genome variation and population structure among 1142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*. *Genome research*, *30*(10), 1533–1546.

Clarkson, C. S., Miles, A., Harding, N. J., O'Reilly, A. O., Weetman, D., Kwiatkowski, D., & Donnelly, M. J. (2021). The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors Anopheles gambiae and Anopheles coluzzii. *Molecular Ecology*, *30*(21), 5303–5317. https://doi.org/10.1111/mec.15845

Clarkson, C. S., Miles, A., Harding, N. J., O'Reilly, A. O., Weetman, D., Kwiatkowski, D., Donnelly, M. J., & gambiae 1000 Genomes Consortium, A. (2021). The genetic architecture of target-site resistance to pyrethroid insecticides in the african malaria vectors anopheles gambiae and anopheles coluzzii. *Molecular ecology*, *30*(21), 5303–5317.

Clayton, D. H., Bush, S. E., Goates, B. M., & Johnson, K. P. (2003). Host defense reinforces host–parasite cospeciation. *Proceedings of the National Academy of Sciences*, *100*(26), 15694–15699.

Crosby, D. S., Breaker, L. C., & Gemmill, W. H. (1993). A proposed definition for vector correlation in geophysics: Theory and application. *Journal of Atmospheric and Oceanic Technology*, *10*(3), 355–367.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools [giab008]. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016). Sweepfinder2: Increased sensitivity, robustness and flexibility. *Bioinformatics*, *32*(12), 1895–1897.

DeGiorgio, M., & Szpiech, Z. A. (2022). A spatially aware likelihood test to detect sweeps from haplotype distributions. *PLoS genetics*, *18*(4), e1010134.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome biology*, *4*, 1–11.

Dieckhaus, H., Brocidiacono, M., Randolph, N. Z., & Kuhlman, B. (2024a). Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, *121*(6), e2314853121.

Dieckhaus, H., Brocidiacono, M., Randolph, N. Z., & Kuhlman, B. (2024b). Transfer learning to leverage larger datasets for improved prediction of protein

stability changes. *Proceedings of the National Academy of Sciences*, *121*(6), e2314853121. https://doi.org/10.1073/pnas.2314853121

Doyle, S. R., Søe, M. J., Nejsum, P., Betson, M., Cooper, P. J., Peng, L., Zhu, X.-Q., Sanchez, A., Matamoros, G., Sandoval, G. A. F., et al. (2022). Population genomics of ancient and modern *Trichuris trichiura*. *Nature Communications*, *13*(1), 3888.

Du, W., Awolola, T. S., Howell, P., Koekemoer, L. L., Brooke, B. D., Benedict, M. Q., Coetzee, M., & Zheng, L. (2005). Independent mutations in the rdl locus confer dieldrin resistance to anopheles gambiae and an. arabiensis. *Insect Mol. Biol.*, *14*(2), 179–183.

Edi, C. V., Djogbénou, L., Jenkins, A. M., Regna, K., Muskavitch, M. A. T., Poupardin, R., Jones, C. M., Essandoh, J., Kétoh, G. K., Paine, M. J. I., Koudou, B. G., Donnelly, M. J., Ranson, H., & Weetman, D. (2014). Cyp6 p450 enzymes and ace-1 duplication produce extreme and multiple insecticide resistance in the malaria mosquito anopheles gambiae. *PLOS Genetics*, *10*(3), e1004236. https://doi.org/10.1371/journal.pgen.1004236

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking under positive darwinian selection. *Genetics*, *155*(3), 1405–1413.

Finch, K. N., Cronn, R. C., Ayala Richter, M. C., Blanc-Jolivet, C., Correa Guerrero, M. C., De Stefano Beltrán, L., García-Dávila, C. R., Honorio Coronado, E. N., Palacios-Ramos, S., Paredes-Villanueva, K., et al. (2020). Predicting the geographic origin of Spanish Cedar (*Cedrela odorata* l.) based on DNA variation. *Conservation Genetics*, *21*(4), 625–639.

Fisher, R. A. (1937). The Wave of Advance of Advantageous Genes [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1937.tb02153.x]. *Annals*

of Eugenics, *7*(4), 355–369. https://doi.org/10.1111/j.1469-1809.1937.tb02153.x

Fisher, S., Ronald Aylmer. (1930). *The genetical theory of natural selection* [https://www.biodiversitylibrary.org/bibliography/27468]. Oxford, Clarendon Press, 1930. https://www.biodiversitylibrary.org/item/69976

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., et al. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, *347*(6217), 1258524.

Galtier, N. (2016). Adaptive protein evolution in animals and the effective population size hypothesis. *PLOS Genetics*, *12*(1), e1005774. https://doi.org/10.1371/journal.pgen.1005774

Giraldo-Calderón, G. I., Emrich, S. J., MacCallum, R. M., Maslen, G., Dialynas, E., Topalis, P., Ho, N., Gesing, S., Consortium, V., Madey, G., et al. (2015). Vectorbase: An updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic acids research*, *43*(D1), D707–D713.

Guillot, G., Jónsson, H., Hinge, A., Manchih, N., & Orlando, L. (2015). Accurate continuous geographic assignment from low- to high-density SNP data. *Bioinformatics*, *32*(7), 1106–1108. https://doi.org/10.1093/bioinformatics/btv703

Hafner, M. S., & Nadler, S. A. (1988). Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, *332*(6161), 258–259. https://doi.org/10.1038/332258a0

Hall, N., Karras, M., Raine, J. D., Carlton, J. M., Kooij, T. W. A., Berriman, M., Florens, L., Janssen, C. S., Pain, A., Christophides, G. K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M. A., Ormond, D., Doggett, J., Trueman, H. E., ... Sinden, R. E. (2005). A comprehensive survey of the ¡i¿plasmodium¡/i¿ life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, *307*(5706), 82–86.

Haller, B. C., & Messer, P. W. (2017). Asymptoticmk: A web-based tool for the asymptotic mcdonald-kreitman test. *G3: Genes—Genomes—Genetics*, *7*(5), 1569–1575. https://doi.org/10.1534/g3.117.039693

Haller, B. C., & Messer, P. W. (2019a). SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Haller, B. C., & Messer, P. W. (2019b). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Hemingway, J., Ranson, H., Magill, A., Kolaczinski, J., Fornadel, C., Gimnig, J., Coetzee, M., Simard, F., Roch, D. K., Hinzoumbe, C. K., et al. (2016). Averting a malaria disaster: Will insecticide resistance derail malaria control? *The Lancet*, *387*(10029), 1785–1788.

Hermisson, J., & Pennings, P. S. (2005). Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics*, *169*(4), 2335–2352.

Huelsenbeck, J. P., Rannala, B., & Larget, B. (2000). A bayesian framework for the analysis of cospeciation. *Evolution*, *54*(2), 352–364.

Huelsenbeck, J. P., Rannala, B., & Yang, Z. (1997). Statistical tests of host-parasite cospeciation. *Evolution*, *51*(2), 410–419.

Jones, C. M., Lee, Y., Kitchen, A., Collier, T., Pringle, J. C., Muleba, M., Irish, S., Stevenson, J. C., Coetzee, M., Cornel, A. J., Norris, D. E., & Carpi, G. (2018). Complete anopheles funestus mitogenomes reveal an ancient history of mitochondrial lineages and their distribution in southern and central africa. *Scientific Reports*, *8*(1), 9054. https://doi.org/10.1038/s41598-018-27092-y

Jousselin, E., Van Noort, S., Berry, V., Rasplus, J.-Y., Rønsted, N., Erasmus, J. C., & Greeff, J. M. (2008). One fig to bind them all: Host conservatism in a fig wasp community unraveled by cospeciation analyses among pollinating and nonpollinating fig wasps. *Evolution: International Journal of Organic Evolution*, *62*(7), 1777–1797.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature*, *596*(7873), 583–589.

Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y., & Ishiguro-Watanabe, M. (2024). Kegg: Biological systems database as a model of the real world. *Nucleic Acids Research*, *53*(D1), D672–D677. https://doi.org/10.1093/nar/gkae909

Keightley, P. D., & Jackson, B. C. (2018). Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics*, *209*(3), 897–906.

Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, *14*(11), 1–21. https://doi.org/10.1371/journal.pcbi.1006581

Kern, A. D. (2025). *Crabml: High-performance phylogenetic analysis*. https://github.com/andrewkern/crabml

Kern, A. D., & Schrider, D. R. (2018). diploS/HIC: An Updated Approach to Classifying Selective Sweeps. *G3 Genes—Genomes—Genetics*, *8*(6), 1959–1970. https://doi.org/10.1534/g3.118.200262

Kojima, K.-I., & Kelleher, T. M. (1962). Survival of mutant genes. *The American Naturalist*, *96*(891), 329–346.

Kojima, K.-I., & Schaffer, H. E. (1967). Survival process of linked mutant genes. *Evolution*, 518–531.

Kolmogorov, A., Petrovskii, I., & Piscunov, N. (1937). A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem. *Byul. Moskovskogo Gos. Univ.*, *1*(6), 1–25. http://books.google.com/books?id=ikN59GkYJKIC&lpg=PP1&dq=A.N.%20Kolmogorov%3A%20Selected%20Works&client=firefox-a&pg=PA242#v=onepage&q=&f=false

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: A program for identifying clustering modes and packaging population structure inferences across k. *Molecular ecology resources*, *15*(5), 1179–1191.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Augeri, D. M., et al. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and distributions*, *19*(11), 1366–1379.

Kumar, J., Kumar, A., Gupta, Y., Vashisht, K., Kumar, S., Sharma, A., Kumar, R., Sharon, A., Tripathi, P. K., Das, R., Singh, O. P., Singh, S., Chakraborti, S., Sunil, S., & Pandey, K. C. (2024). A cub and sushi domain-containing protein with esterase-like activity confers insecticide resistance in the indian

malaria vector ¡em¿anopheles stephensi¡/em¿. *Journal of Biological Chemistry*, *300*(10). https://doi.org/10.1016/j.jbc.2024.107759

Kumar, S., Christophides, G. K., Cantera, R., Charles, B., Han, Y. S., Meister, S., Dimopoulos, G., Kafatos, F. C., & Barillas-Mury, C. (2003). The role of reactive oxygen species on plasmodium melanotic encapsulation in anopheles gambiae. *Proceedings of the National Academy of Sciences*, *100*(24), 14139–14144. https://doi.org/10.1073/pnas.2036262100

Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., Langley, S. A., Suarez, C., Corbett-Detig, R. B., Kolaczkowski, B., Fang, S., Nista, P. M., Holloway, A. K., Kern, A. D., Dewey, C. N., Song, Y. S., Hahn, M. W., & Begun, D. J. (2012). Genomic variation in natural populations of drosophila melanogaster. *Genetics*, *192*(2), 533–598. https://doi.org/10.1534/genetics.112.142018

Lauterbur, M. E., Munch, K., & Enard, D. (2023). Versatile detection of diverse selective sweeps with flex-sweep. *Molecular Biology and Evolution*, *40*(6), msad139.

Liu, X., & Fu, Y.-X. (2020). Stairway plot 2: Demographic history inference with folded snp frequency spectra. *Genome biology*, *21*(1), 280.

Lucas, E. R., Nagi, S. C., Egyir-Yawson, A., Essandoh, J., Dadzie, S., Chabi, J., Djogbénou, L. S., Medjigbodo, A. A., Edi, C. V., Kétoh, G. K., Koudou, B. G., Van't Hof, A. E., Rippon, E. J., Pipini, D., Harding, N. J., Dyer, N. A., Cerdeira, L. T., Clarkson, C. S., Kwiatkowski, D. P., ... Weetman, D. (2023). Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in anopheles gambiae and anopheles coluzzii.

Nature Communications, *14*(1), 4946. https://doi.org/10.1038/s41467-023-40693-0

Manguin, S. (2013, July). *Anopheles mosquitoes - new insights into malaria vectors*. https://doi.org/10.5772/3392

Marcus, J., Ha, W., Barber, R. F., & Novembre, J. (2021). Fast and flexible estimation of effective migration surfaces. *Elife*, *10*, e61927.

Martinez-Torres, D., Chandre, F., Williamson, M. S., Darriet, F., Bergé, J. B., Devonshire, A. L., Guillet, P., Pasteur, N., & Pauron, D. (1998). Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector anopheles gambiae s.s. *Insect Mol. Biol.*, *7*(2), 179–184.

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, *44*(3), 243–246.

Matsuo, T., Sugaya, S., Yasukawa, J., Aigaki, T., & Fuyama, Y. (2007). Odorant-binding proteins obp57d and obp57e affect taste perception and host-plant preference in drosophila sechellia. *PLoS biology*, *5*(5), e118.

Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.*, *23*(1), 23–35. https://doi.org/10.1017/S0016672300014634

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Messer, P. W., & Petrov, D. A. (2013). Frequent adaptation and the mcdonald–kreitman test. *Proceedings of the National Academy of Sciences*, *110*(21), 8615–8620. https://doi.org/10.1073/pnas.1220835110

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., & Thomas, P. D. (2019). Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.*, *14*(3), 703–721.

Min, J., Gupta, M., Desai, M. M., & Weissman, D. B. (2022a). Spatial structure alters the site frequency spectrum produced by hitchhiking. *Genetics*, *222*(3), iyac139. https://doi.org/10.1093/genetics/iyac139

Min, J., Gupta, M., Desai, M. M., & Weissman, D. B. (2022b). Spatial structure alters the site frequency spectrum produced by hitchhiking. *Genetics*, *222*(3), iyac139. https://doi.org/10.1093/genetics/iyac139

Mongkol, W., Pomun, T., Nguitragool, W., Kumpitak, C., Duangmanee, A., Sattabongkot, J., & Kubera, A. (2021). Anopheles dirus yellow-g mediates plasmodium vivax infection. *Tropical Medicine International Health*, *26*(9), 1029–1035. https://doi.org/10.1111/tmi.13635

Moran, N., & Baumann, P. (1994). Phylogenetics of cytoplasmically inherited microorganisms of arthropods. *Trends in Ecology & Evolution*, *9*(1), 15–20.

Muktupavela, R. A., Petr, M., Ségurel, L., Korneliussen, T., Novembre, J., & Racimo, F. (2022). Modeling the spatiotemporal spread of beneficial alleles using ancient genomes (A. M. Andrés, G. H. Perry, I. Alves & A. Eriksson, Eds.). *eLife*, *11*, e73767. https://doi.org/10.7554/eLife.73767

Müller, P., Warr, E., Stevenson, B. J., Pignatelli, P. M., Morgan, J. C., Steven, A., Yawson, A. E., Mitchell, S. N., Ranson, H., Hemingway, J., Paine, M. J. I., & Donnelly, M. J. (2008a). Field-caught permethrin-resistant anopheles gambiae overexpress cyp6p3, a p450 that metabolises pyrethroids. *PLoS Genetics*, *4*(11), e1000286. https://doi.org/10.1371/journal.pgen.1000286

Müller, P., Warr, E., Stevenson, B. J., Pignatelli, P. M., Morgan, J. C., Steven, A., Yawson, A. E., Mitchell, S. N., Ranson, H., Hemingway, J., Paine, M. J. I., & Donnelly, M. J. (2008b). Field-caught permethrin-resistant anopheles gambiae overexpress CYP6P3, a P450 that metabolises pyrethroids. *PLoS Genet.*, *4*(11), e1000286.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., Assour, L. A., Basseri, H., Berlin, A., Birren, B. W., Blandin, S. A., Brockman, A. I., Burkot, T. R., Burt, A., Chan, C. S., ... Besansky, N. J. (2015). Highly evolvable malaria vectors: The genomes of 16 anopheles mosquitoes. *Science*, *347*(6217), 1258522. https://doi.org/10.1126/science.1258522

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data [Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab]. *Genome Res.*, *15*(11), 1566–1575. https://doi.org/10.1101/gr.4252305

Noh, M. Y., Kim, S. H., Gorman, M. J., Kramer, K. J., Muthukrishnan, S., & Arakane, Y. (2020). Yellow-g and yellow-g2 proteins are required for egg desiccation resistance and temporal pigmentation in the asian tiger mosquito, aedes albopictus. *Insect Biochemistry and Molecular Biology*, *122*, 103386. https://doi.org/10.1016/j.ibmb.2020.103386

Patterson, G., Goodfellow, C. K., Ting, N., Kern, A. D., & Ralph, P. L. (2026). Simulation-based spatially explicit close-kin mark–recapture. *Molecular Ecology Resources*, *26*(1), e70074. https://doi.org/10.1111/1755-0998.70074

Pearson, R. D., Amato, R., Kwiatkowski, D. P., falciparum Community Project, M. P., et al. (2019). An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. *bioRxiv*, 824730.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Povelones, M., Upton, L. M., Sala, K. A., & Christophides, G. K. (2011). Structure-function analysis of the anopheles gambiae lrim1/apl1c complex and its interaction with complement c3-like protein tep1. *PLOS Pathogens*, *7*(4), e1002023. https://doi.org/10.1371/journal.ppat.1002023

Povelones, M., Waterhouse, R. M., Kafatos, F. C., & Christophides, G. K. (2009). Leucine-rich repeat protein complex activates mosquito complement in defense against plasmodium parasites. *Science*, *324*(5924), 258–261. https://doi.org/10.1126/science.1171400

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559–575.

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rehmann, C. T., Ralph, P. L., & Kern, A. D. (2024). Evaluating evidence for co-geography in the *Anopheles–Plasmodium* host–parasite system. *G3: Genes, Genomes, Genetics*, *14*(3), jkae008.

Rehmann, C. T., Small, S. T., Ralph, P. L., & Kern, A. D. (2025a). Sweeps in space: Leveraging geographic data to identify beneficial alleles in anopheles gambiae. *Molecular Biology and Evolution*, *42*(6), msaf141. https://doi.org/10.1093/molbev/msaf141

Rehmann, C. T., Small, S. T., Ralph, P. L., & Kern, A. D. (2025b). Sweeps in space: Leveraging geographic data to identify beneficial alleles in *Anopheles gambiae*. *Molecular Biology and Evolution*, *42*(6), msaf141. https://doi.org/10.1093/molbev/msaf141

Ringbauer, H., Coop, G., & Barton, N. H. (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, *205*(3), 1335–1351.

Rivera-Colón, A. G., Rehmann, C. T., & Kern, A. D. (2026). Mkado: A toolkit for mcdonald-kreitman tests of natural selection. *bioRxiv*. https://doi.org/10.64898/2026.03.02.709122

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, *145*(4), 1219–1228.

Saad, H., Elfeky, S. A., El-Gamel, N. E. A., & Abo Dena, A. S. (2025). Organophosphate pesticides: A review on classification, synthesis, toxicity, remediation and analysis. *RSC Advances*, *15*(48), 40802–40822. https://doi.org/10.1039/d5ra05552k

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J.,

et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837.

Schlenke, T. A., & Begun, D. J. (2004). Strong selective sweep associated with a transposon insertion in drosophila simulans. *Proceedings of the National Academy of Sciences*, *101*(6), 1626–1631.

Schrider, D. R., & Kern, A. D. (2016). S/hic: Robust identification of soft and hard sweeps using machine learning. *PLoS genetics*, *12*(3), e1005928.

Sinnis, P., Coppi, A., Toida, T., Toyoda, H., Kinoshita-Toyoda, A., Xie, J., Kemp, M. M., & Linhardt, R. J. (2007). Mosquito heparan sulfate and its potential role in malaria infection and transmission. *The Journal of Biological Chemistry*, *282*(35), 25376–25384. https://doi.org/10.1074/jbc.M704698200

Slatkin, M., & Bertorelle, G. (2001). The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics*, *158*(2), 865–874.

Small, S. T., Costantini, C., Sagnon, N., Guelbeogo, M. W., Emrich, S. J., Kern, A. D., Fontaine, M. C., & Besansky, N. J. (2023). Standing genetic variation and chromosome differences drove rapid ecotype formation in a major malaria mosquito. *Proceedings of the National Academy of Sciences*, *120*(11), e2219835120. https://doi.org/10.1073/pnas.2219835120

Small, S. T., Labbé, F., Coulibaly, Y. I., Nutman, T. B., King, C. L., Serre, D., & Zimmerman, P. A. (2019). Human migration and the spread of the nematode parasite *Wuchereria bancrofti*. *Molecular biology and evolution*, *36*(9), 1931–1941.

Smith, C. C. R., Patterson, G., Ralph, P. L., & Kern, A. D. (2024). Estimation of spatial demographic maps from polymorphism data using a neural network.

Molecular Ecology Resources, *24*(7), e14005. https://doi.org/10.1111/1755-0998.14005

Smith, C. C., & Kern, A. D. (2023). Dispersenn2: A neural network for estimating dispersal distance from georeferenced polymorphism data. *BMC bioinformatics*, *24*(1), 385.

Smith, C. C., Patterson, G., Ralph, P. L., & Kern, A. D. (2024). Estimation of spatial demographic maps from polymorphism data using a neural network. *bioRxiv*.

Smith, C. C., Tittes, S., Ralph, P. L., & Kern, A. D. (2023). Dispersal inference from population genetic variation using a convolutional neural network. *Genetics*, *224*(2), iyad068.

Smith, G., & Briscoe, A. D. (2015). Molecular evolution and expression of the cral trio protein family in insects. *Insect Biochemistry and Molecular Biology*, *62*, 168–173. https://doi.org/10.1016/j.ibmb.2015.02.003

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306.

St. Laurent, B., Harding, N., Deason, N., Oy, K., Sok Loeun, C., Sary, M., Sunly, R., Nhep, S., Drury, E., Rockett, K., Sovannaroth, S., Goncalves, S., Kwiatkowski, D., & Miles, A. (2022). Population genomics reveal distinct and diverging populations of an. minimus in cambodia. *Communications Biology*, *5*(1), 1308. https://doi.org/10.1038/s42003-022-04259-y

Stevens., D. L., & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, *99*(465), 262–278.

Sugden, L. A., Atkinson, E. G., Fischer, A. P., Rong, S., Henn, B. M., & Ramachandran, S. (2018). Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nature communications*, *9*(1), 703.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, *123*(3), 585–595.

Tennessen, J. A., Brosula, R., Chabanol, E., Bickersmith, S., Early, A. M., Laws, M., Kelley, K. A., Grillet, M. E., Gamboa, D., Lucas, E. R., Duchemin, J.-B., Quiñones, M. L., Sallum, M. A. M., Bergo, E. S., Moreno, J. E., Nagi, S., Arisco, N. J., Sooklall, M., Niles-Robin, R., ... Neafsey, D. E. (2025). Population genomics of anopheles darlingi, the principal south american malaria vector mosquito, 2025.03.13.643102. https://doi.org/10.1101/2025.03.13.643102

Teshima, K. M., Coop, G., & Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome research*, *16*(6), 702–712.

The Anopheles gambiae 1000 Genomes Consortium, Clarkson, C. S., Miles, A., Harding, N. J., Lucas, E. R., Battey, C. J., Amaya-Romero, J. E., Cano, J., Diabate, A., Constant, E., Nwakanma, D. C., Jawara, M., Essandoh, J., Dinis, J., Goff, G. L., Robert, V., Troco, A. D., Costantini, C., Rohatgi, K. R., ... Kwiatkowski, D. P. (2020, February). Genome variation and population structure among 1,142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii [Pages: 864314 Section: New Results]. https://doi.org/10.1101/864314

The *Anopheles gambiae* 1000 Genomes Consortium. (2017). Genetic diversity of the african malaria vector anopheles gambiae. *Nature*, *552*(7683), 96–100. https://doi.org/10.1038/nature24995

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P., & Mi, H. (2022). Panther: Making genome-scale phylogenetics accessible to all. *Protein Science*, *31*(1), 8–22. https://doi.org/https://doi.org/10.1002/pro.4218

Thomsen, E. K., Koimbu, G., Pulford, J., Jamea-Maiasa, S., Ura, Y., Keven, J. B., Siba, P. M., Mueller, I., Hetzel, M. W., & Reimer, L. J. (2016). Mosquito behaviour change after distribution of bednets results in decreased protection against malaria exposure. *J. Infect. Dis.*, jiw615.

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS biology*, *4*(3), e72.

Wang, M., Zhao, Y., & Zhang, B. (2022). *Superexacttest: Exact test and visualization of multi-set intersections* [R package version 1.1.0]. https://doi.org/10.32614/CRAN.package.SuperExactTest

Wasser, S. K., Shedlock, A. M., Comstock, K., Ostrander, E. A., Mutayoba, B., & Stephens, M. (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences*, *101*(41), 14847–14852.

Wesolowski, A., Buckee, C. O., Bengtsson, L., Wetter, E., Lu, X., & Tatem, A. J. (2014). Commentary: Containing the Ebola outbreak-the potential and challenge of mobile network data. *PLoS currents*, *6*.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, *338*(6104), 267–270.

Whitlock, M. C., & McCauley, D. E. (1999). Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. *Heredity*, *82*(2), 117–125.

Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., & Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *PLoS genetics*, *3*(6), e90.

Wright, S. (1943a). ISOLATION BY DISTANCE. *Genetics*, *28*(2), 114–138. https://doi.org/10.1093/genetics/28.2.114

Wright, S. (1943b). Isolation by distance. *Genetics*, *28*(2), 114.

Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics*, *31*(1), 39.

Xue, A. T., Schrider, D. R., & Kern, A. D. (2021a). Discovery of ongoing selective sweeps within Anopheles mosquito populations using deep learning. *Molecular Biology and Evolution*, *38*(3), 1168–1183. https://doi.org/10.1093/molbev/msaa259

Xue, A. T., Schrider, D. R., & Kern, A. D. (2021b). Discovery of ongoing selective sweeps within anopheles mosquito populations using deep learning. *Molecular biology and evolution*, *38*(3), 1168–1183.

Yang, Z. (2007). Paml 4: Phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, *24*(8), 1586–1591.

Ye, B., Tian, W., Wang, B., & Liang, J. (2024). Castpfold: Computed atlas of surface topography of the universe of protein folds. *Nucleic Acids Research*, *52*(W1), W194–W199. https://doi.org/10.1093/nar/gkae415

Zhao, L., & Begun, D. J. (2017). Genomics of parallel adaptation at two timescales in drosophila. *PLOS Genetics*, *13*(10), e1007016. https://doi.org/10.1371/journal.pgen.1007016

Zhao, L., Wit, J., Svetec, N., & Begun, D. J. (2015). Parallel gene expression differences between low and high latitude populations of drosophila melanogaster and d. simulans. *PLOS Genetics*, *11*(5), e1005184. https://doi.org/10.1371/journal.pgen.1005184