The evolutionary forces that drive patterns of genetic diversity within and between species

Connor Sean Murray

Jupiter, Florida USA

Bachelor of Science (with Honors, *magna cum laude*)

Florida State University, 2019

A Dissertation presented to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Biology

University of Virginia

June 2024

Abstract

Natural genetic diversity describes the historical patterns of selection, demography, and drift that act upon populations. With the widespread proliferation of next generation sequencing (NGS) technologies and abundance of genomic tools, we can investigate the evolutionary mechanisms that leave distinct signals across the genome. By utilizing genome-wide diversity information and population-based datasets, we can begin to elucidate the ways in which organisms evolve and predict how they might respond to novel environmental changes. My dissertation studies the population genetics and natural selection pressures on several taxa within the *Daphnia pulex* species complex, a group of freshwater microcrustaceans that rapidly adapt to changing environments. I also studied the patterns of genetic diversity and demography influencing a metapopulation of *Drosophila melanogaster* (i.e., the common fruit fly) within Charlottesville, Virginia, USA. In my first chapter, I analyze general patterns of gene family evolution and selection across *Daphnia* genomes. My second chapter delves into the evolutionary mechanisms that maintain variation between cryptic species of *D. pulex*. My third chapter examines the demographic consequences of overwintering bottlenecks in *Drosophila melanogaster*. Ultimately, my dissertation contributes to the evolutionary mechanisms that influence genetic diversity within and between species. Through detailed genomic analyses, I laid the groundwork for understanding genomic change and aided the scientific community by contributing to the understanding of both the long- and short-acting evolutionary processes shaping patterns of variation.

Acknowledgements

This dissertation would not have been possible without the immense support I received from so many people throughout the years. My journey was enriched by my dear friends in the Department of Biology, my family, and the many other wonderful friends I made along the way. I would like to highlight some of the individuals who were particularly instrumental during my time at UVA. First, I extend my gratitude to Alan Bergland, my advisor, who guided me from my rotation semester to the completion of my program. His continuous support and mentorship were invaluable. Aakrosh Ratan, a scientific mentor and committee member, provided crucial assistance with genomic analyses and professional development. Joaquin Nunez introduced me to the intricacies of population genetics and the coding techniques that became essential in my research. Abbey Hayes, a friend and scientific mentor, taught me to confidently express my passion for science and public speaking. Amanda Gibson, the chair of my committee, greatly improved my scientific writing and the development of my ideas. Douglas Taylor, a committee member, shared his profound knowledge of population genetics and his unwavering passion for science, which motivated me. I am also grateful to my fellow graduate students and friends: Kendall Branham, Daniel Nondorf, Adam Lenhart, Christopher Robinson, Keric Lamb, Antoine Perrier, Madison Karram, Madison Doceti, and David Bass who kept me motivated and always offered a helping hand. I would also like to thank my undergraduate advisor, David Houle and friends: Luke Jones, Kevin Doheny, and Ryan Fortune for helping me find community and a love for evolutionary genetics during my time at Florida State University. Additionally, thank you to the many friends I made in Boston during my co-op at Sanofi. A heartfelt thank you goes to my parents, Allison and Sean Murray, and my sister, Taylor Murray, for their unwavering support of my passions and learning. Lastly, I express my appreciation to my grandparents, who encouraged my learning from my earliest memories.

Table of Contents

Introduction

Biological novelty is created when evolution acts upon the genome, yet it remains unclear how different species achieve this diversity and the exact genetic mechanisms that are employed. The genome contains the raw material that evolution acts upon to create biological novelty (Fisher, 1930; Lewontin, 1974; Reed & Frankham, 2003). Within my dissertation, I conduct an in-depth analysis of genome evolution, population genetics, and genetic diversity within and between related species. I assess these components by investigating evolutionary and ecological questions using the basis that organisms adapt to environmental change (Lewontin, 1974). Adaptation is fundamental to the field of biology because all organisms contain genetic material that has variability in the nucleotide sequences that code for amino acids and proteins (e.g., central dogma of biology). The nearly neutral model of evolution contends that most genetic variation gets lost (Ohta, 1992; Ohtsuki et al., 2022), yet there exists extensive diversity across species (Romiguier et al., 2014). Despite being typically regarded as a force that reduces variation through purifying and positive selection, natural selection can also act to maintain diversity through balancing selection (Charlesworth, 2006; Fisher, 1922). In addition, many overlapping evolutionary and demographic forces will influence the standing levels of diversity. Additionally, events that happened before the complete split of a species, like population bottlenecks or ancient hybridization will lead to vastly different situations on the levels of diversity across taxa (Kovach & McCouch, 2008; Vilaça et al., 2021). Overall, even simple models of genome evolution have inherent complexities that make it difficult to disentangle requiring further research.

Genetic diversity can be sustained through various mechanisms, among which balancing selection stands out for its ability to maintain polymorphisms over time and improving adaptive potential in varied environmental contexts (Fijarczyk & Babik, 2015). This phenomenon is exemplified by classic cases such as malaria resistance (Malaria Genomic Epidemiology Network, 2015), genetic incompatibility within and between species (Schierup et al., 2001), and

innate immune response (Ferrer-Admetlla et al., 2008). Other forms of balancing selection including, adaptive tracking, frequency-dependent selection, and fluctuating selection are also implicated in maintaining genetic diversity but parsing these mechanisms apart is difficult (Fitzpatrick et al., 2007; Bangerter, 2021; Rudman et al., 2022). With the advent of low-cost next generation sequencing (NGS) technology, we can now delve into the birth and subsequent maintenance of single nucleotide polymorphisms (SNPs; Head et al., 2018), unraveling 'molecular breadcrumbs' to elucidate the distribution of diversity within and between species. However, it can be difficult to definitively attribute patterns of genomic variation to balancing selection. For instance, hybridization and introgression can reintroduce alleles over time between populations (Mavárez et al., 2006), leading to discordant gene and species trees (Suh et al., 2015; Lenz et al., 2013). Evolutionary forces like hybridization, introgression, convergence, and balancing selection all result in the maintenance of SNPs within the genome. Consequently, comprehensive analyses that consider overlapping influences are crucial for understanding how variation persists within and between species outside of the scope of individual effect sizes (Consuegra et al., 2005).

Biological novelty exists in the genome besides just individual mutations (e.g., SNPs) in the form of gene family evolution. For instance, the expansion and contraction in the number of genes within a specific protein family and variability in the number of gene families can be driven by natural selection (Hahn et al., 2007; Hancock, 2005). Gene family expansion and contraction have been linked to adaptation in myriad examples across the tree of life (De Bie et al., 2006; Richter et al., 2018). Neofunctionalization and subfunctionalization are processes in which newly born genes diversify from their progenitor by the accumulation of slightly different functions (Sandve et al., 2018). We can study gene family evolution by quantifying the number of genes across species and measuring the extent of natural selection through statistics like *dN/dS* (i.e., the rate of non-synonymous to synonymous nucleotide substitutions). Higher *dN/dS* indicates positive selection (*dN/dS* > 1), while lower rates indicate purifying selection (*dN/dS* <

1; Álvarez-Carretero et al., 2023). The diversification hypothesis contends that newly expanding genes will be under higher rates of positive selection. Under this hypothesis, gene family evolution is key to driving new biological functions across species. Yet, these analyses have only begun to gain traction in the last decade due to the increased number of sequenced genomes becoming available.

An ideal model system to tackle the questions of balancing selection and gene family divergence across species is through studying the *Daphnia* genera. *Daphnia* are keystone microcrustaceans that graze on algae and inhabit freshwater systems ranging in size from ephemeral rain puddles to lakes and reservoirs (Colbourne et al., 2011; Ebert, 2022). *Daphnia* have been studied for decades due to their propensity to rapidly adapt to ecological change (Chin & Cristescu, 2021). Species of *Daphnia* reproduce through cyclical parthenogenesis, the switching of asexual and sexual reproduction following environmental stressor cues (Lynch, 1983, 1987). *Daphnia* are also phenotypically plastic in the face of environmental degradation, which allows them to handle extreme ecological stressors like hypoxia, desiccation, crowding, and intraspecific/interspecific competition (Stoks et al., 2016). *Daphnia* can also produce resistant egg cases called ephippia that will hatch to seed the next cycle of clonal lineages within a population (Ban et al., 2009; Heier & Dudycha, 2009). More relevant for my dissertation topic, the *Daphnia pulex* species complex encompasses several cryptic species of *D. pulex* that live in European and North American ponds (Crease et al., 2012). Despite being phenotypically indistinguishable, these *D. pulex* taxa could have vastly different patterns of adaptive evolution.

In my first chapter, I utilize whole-genome assemblies and proteomes from several *Daphnia* taxa to explore comparative genomics across related species and test hypotheses related to the selective pressures acting upon evolving gene families. Through this analysis, I found that across *Daphnia*, gene families related to spermatogenesis, the creation of motile sperm, and general stress responses are expanding and contracting. I show that positive selection occurs within these expanding genes. This analysis is supplemented by an

investigation of ecologically relevant gene families within *Daphnia* with a focus on genes that are expected to contribute to complex phenotypes, like reproduction and sperm production. My first chapter spans tens of millions of years and largely analyses the forces that influence gene family shifts across species ranging from 10 to 100 million years ago (mya).

My second chapter uses large-scale NGS population genomic datasets of wild-sequenced *D. pulex* from a cryptic species-pair that live in European and North American ponds (Barnard-Kubow et al., 2022). Through this investigation, I found evidence for an abundance of shared polymorphisms that exist in the same genomic positions across the cryptic species' genomes. I investigate several evolutionary hypotheses that can create an excess of shared polymorphisms in this chapter. I explicitly test hybridization and introgression, incomplete lineage sorting, convergent evolution, and balancing selection using formal statistics on the incidence of shared mutations. Through this, I identified evolutionary processes influencing the cryptic *D. pulex* and preformed a functional genomic investigation of a blue wavelength opsin gene that has several dozen putatively ancient trans-specific polymorphisms. This opsin gene has two distinct haplotypes with dozens of trans-specific SNPs that confer different rates of motility across clonal lineages depending on light condition. My second chapter explores the in-depth patterns of balancing selection and demographic processes that maintain genetic diversity and mostly span recently diverged species-pairs from one to ten mya.

My third chapter examines a local metapopulation of *Drosophila melanogaster* that experiences seasonal population bottlenecks during the winter months. I tested various demographic models that account for these recurring bottlenecks using forward genetics simulations. The severity of these bottlenecks ranged from near local extinction to maintaining a stable, constant population size. Through these simulations, I estimated genetic differentiation statistics and inferred significant population bottlenecks that likely explain the allele frequency shifts observed in sequenced seasonal fly populations (Bangerter, 2021; Nunez et al., 2024). This chapter allowed me to explore tens of thousands of demographic scenarios, examining

their impact on genetic differentiation within a metapopulation over time. Ultimately, it investigates how demographic processes occurring in wild populations over seasonal timescales influence genetic variation patterns within one to three years.

In my dissertation's appendix, I present a study in which I perform comparative genomics across microbial species, scrutinizing a widely used species concept based on an average nucleotide identity cutoff (Murray et al., 2021). This comparative genomics work was pivotal in shaping my early dissertation, revealing the extensive variability in the genomes of many rapidly evolving species. This project also helped me develop crucial computational skills for my comparative genomic and evolutionary biology chapters. Furthermore, it has significant implications for the field of microbial genomics by highlighting the necessity of avoiding biases that can influence patterns identified in data-mining omics projects.

Ultimately, my dissertation illuminates the evolutionary forces at play in nature. By integrating gene family analyses, comparative genomics, functional genetics, and forward genetic simulations, my research delves into fundamental questions of how and why organisms change, and how variation is preserved. The interpretations presented here elucidate how genetic diversity differentiates and is maintained in the genomes of wild organisms. This work not only advances our understanding of evolutionary biology but also underscores the intricate dynamics that affect diversity.

References

Álvarez-Carretero, S., Kapli, P., & Yang, Z. (2023). Beginner's Guide on the Use of *PAML* to Detect Positive Selection. *Molecular Biology and Evolution*, *40*(4), msad041. https://doi.org/10.1093/molbev/msad041

Ban, S., Tenma, H., Mori, T., & Nishimura, K. (2009). Effects of physical interference on life history shifts in *Daphnia pulex*. *Journal of Experimental Biology*, *212*(19), 3174–3183. https://doi.org/10.1242/jeb.031518

Bangerter, A. (2021). *Dense seasonal sampling of an orchard population uncovers population turnover, adaptive tracking, and structure in multiple Drosophila species* [University of Virginia]. https://doi.org/10.18130/WYR9-FZ68

Barnard-Kubow, K. B., Becker, D., Murray, C. S., Porter, R., Gutierrez, G., Erickson, P., Nunez, J. C. B., Voss, E., Suryamohan, K., Ratan, A., Beckerman, A., & Bergland, A. O. (2022). Genetic Variation in Reproductive Investment Across an Ephemerality Gradient in *Daphnia pulex*. *Molecular Biology and Evolution*, *39*(6), msac121. https://doi.org/10.1093/molbev/msac121

Charlesworth, D. (2006). Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genetics*, *2*(4), e64. https://doi.org/10.1371/journal.pgen.0020064

Chin, T. A., & Cristescu, M. E. (2021). Speciation in *Daphnia*. *Molecular Ecology*, mec.15824. https://doi.org/10.1111/mec.15824

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., Tokishita, S., Aerts, A., Arnold, G. J., Basu, M. K., Bauer, D. J., Caceres, C. E., Carmel, L., Casola, C., Choi, J.-H., Detter, J. C., Dong, Q., Dusheyko, S., Eads, B. D., … Boore, J. L. (2011). The Ecoresponsive Genome of *Daphnia pulex*. *Science*, *331*(6017), 555–561. https://doi.org/10.1126/science.1197761

Consuegra, S., Verspoor, E., Knox, D., & García De Leániz, C. (2005). Asymmetric gene flow and the evolutionary maintenance of genetic diversity in small, peripheral Atlantic salmon populations. *Conservation Genetics*, *6*(5), 823–842. https://doi.org/10.1007/s10592-005-9042-4

Crease, T. J., Omilian, A. R., Costanzo, K. S., & Taylor, D. J. (2012). Transcontinental Phylogeography of the *Daphnia pulex* Species Complex. *PLoS ONE*, *7*(10), e46620. https://doi.org/10.1371/journal.pone.0046620

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). *CAFE*: A computational tool for the study of gene family evolution. *Bioinformatics*, *22*(10), 1269–1271. https://doi.org/10.1093/bioinformatics/btl097

Ebert, D. (2022). *Daphnia* as a versatile model system in ecology and evolution. *EvoDevo*, *13*(1), 16. https://doi.org/10.1186/s13227-022-00199-0

Ferrer-Admetlla, A., Bosch, E., Sikora, M., Marquès-Bonet, T., Ramírez-Soriano, A., Muntasell, A., Navarro, A., Lazarus, R., Calafell, F., Bertranpetit, J., & Casals, F. (2008). Balancing Selection Is the Main Force Shaping the Evolution of Innate Immunity Genes. *The Journal of Immunology*, *181*(2), 1315–1322. https://doi.org/10.4049/jimmunol.181.2.1315

Fijarczyk, A., & Babik, W. (2015). Detecting balancing selection in genomes: Limits and prospects. *Molecular Ecology*, *24*(14), 3529–3545. https://doi.org/10.1111/mec.13226

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594–604), 309–368. https://doi.org/10.1098/rsta.1922.0009

Fisher, R. A. (1930). *The genetical theory of natural selection*. Clarendon Press. https://doi.org/10.5962/bhl.title.27468

Fitzpatrick, M. J., Feder, E., Rowe, L., & Sokolowski, M. B. (2007). Maintaining a behaviour polymorphism by frequency-dependent selection on a single gene. *Nature*, *447*(7141), 210–212. https://doi.org/10.1038/nature05764

Hahn, M. W., Han, M. V., & Han, S.-G. (2007). Gene Family Evolution across 12 *Drosophila* Genomes. *PLoS Genetics*, *3*(11), e197. https://doi.org/10.1371/journal.pgen.0030197

Hancock, J. M. (2005). Gene factories, microfunctionalization and the evolution of gene families. *Trends in Genetics*, *21*(11), 591–595. https://doi.org/10.1016/j.tig.2005.08.008

Head, S. R., Ordoukhanian, P., & Salomon, D. R. (Eds.). (2018). *Next Generation Sequencing: Methods and Protocols* (Vol. 1712). Springer New York. https://doi.org/10.1007/978-1-4939-7514-3

Heier, C. R., & Dudycha, J. L. (2009). Ecological speciation in a cyclic parthenogen: Sexual capability of experimental hybrids between *Daphnia pulex* and *Daphnia pulicaria*. *Limnology and Oceanography*, *54*(2), 492–502. https://doi.org/10.4319/lo.2009.54.2.0492

Kovach, M., & McCouch, S. (2008). Leveraging natural diversity: Back through the bottleneck. *Current Opinion in Plant Biology*, *11*(2), 193–200. https://doi.org/10.1016/j.pbi.2007.12.006

Lenz, T. L., Eizaguirre, C., Kalbe, M., & Milinski, M. (2013). Evaluating patterns of convergent evolution and trans-species polymorphism at MHC immunogenes in two sympatric stickleback species. *Evolution*, *67*(8), 2400–2412. https://doi.org/10.1111/evo.12124

Lewontin, R. C. (1974). *The genetic basis of evolutionary change (Vol. 560).* New York: Columbia University Press.

Lynch, M. (1983). Ecological Genetics of *Daphnia pulex*. *Evolution*, *37*(2), 358. https://doi.org/10.2307/2408344

Lynch, M. (1987). The Consequences of Fluctuating Selection for Isozyme Polymorphisms in *Daphnia*. *Genetics*, *115*(4), 657–669. https://doi.org/10.1093/genetics/115.4.657

Malaria Genomic Epidemiology Network. (2015). A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*, *526*(7572), 253–257. https://doi.org/10.1038/nature15390

Mavárez, J., Salazar, C. A., Bermingham, E., Salcedo, C., Jiggins, C. D., & Linares, M. (2006). Speciation by hybridization in *Heliconius* butterflies. *Nature*, *441*(7095), 868–871. https://doi.org/10.1038/nature04738

Murray, C. S., Gao, Y., & Wu, M. (2021). Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nature Communications*, *12*(1), 4059. https://doi.org/10.1038/s41467-021-24128-2

Nunez, J. C. B., Lenhart, B. A., Bangerter, A., Murray, C. S., Mazzeo, G. R., Yu, Y., Nystrom, T. L., Tern, C., Erickson, P. A., & Bergland, A. O. (2024). A cosmopolitan inversion facilitates seasonal adaptation in overwintering *Drosophila*. *Genetics*, *226*(2), iyad207. https://doi.org/10.1093/genetics/iyad207

Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, *23*(1), 263–286. https://doi.org/10.1146/annurev.es.23.110192.001403

Ohtsuki, H., Norimatsu, H., Makino, T., & Urabe, J. (2022). Invasions of an obligate asexual Daphnid species support the nearly neutral theory. *Scientific Reports*, *12*(1), 7305. https://doi.org/10.1038/s41598-022-11218-4

Reed, D. H., & Frankham, R. (2003). Correlation between Fitness and Genetic Diversity. *Conservation Biology*, *17*(1), 230–237. https://doi.org/10.1046/j.1523-1739.2003.01236.x

Richter, D. J., Fozouni, P., Eisen, M. B., & King, N. (2018). Gene family innovation, conservation and loss on the animal stem lineage. *eLife*, *7*, e34226. https://doi.org/10.7554/eLife.34226

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A. A.-T., Weinert, L. A., Belkhir, K., Bierne, N., … Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, *515*(7526), 261–263. https://doi.org/10.1038/nature13685

Rudman, S. M., Greenblum, S. I., Rajpurohit, S., Betancourt, N. J., Hanna, J., Tilk, S., Yokoyama, T., Petrov, D. A., & Schmidt, P. (2022). Direct observation of adaptive tracking on ecological time scales in *Drosophila*. *Science*, *375*(6586). https://doi.org/10.1126/science.abj7484

Sandve, S. R., Rohlfs, R. V., & Hvidsten, T. R. (2018). Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics*, *50*(7), 908–909. https://doi.org/10.1038/s41588-018-0162-4

Schierup, M. H., Mikkelsen, A. M., & Hein, J. (2001). Recombination, Balancing Selection and Phylogenies in MHC and Self-Incompatibility Genes. *Genetics*, *159*(4), 1833–1844. https://doi.org/10.1093/genetics/159.4.1833

Stoks, R., Govaert, L., Pauwels, K., Jansen, B., & De Meester, L. (2016). Resurrecting complexity: The interplay of plasticity and rapid evolution in the multiple trait response to strong changes in predation pressure in the water flea *Daphnia magna*. *Ecology Letters*, *19*(2), 180–190. https://doi.org/10.1111/ele.12551

Suh, A., Smeds, L., & Ellegren, H. (2015). The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. *PLoS Biology*, *13*(8), e1002224. https://doi.org/10.1371/journal.pbio.1002224

Vilaça, S. T., Piccinno, R., Rota-Stabelli, O., Gabrielli, M., Benazzo, A., Matschiner, M., Soares, L. S., Bolten, A. B., Bjorndal, K. A., & Bertorelle, G. (2021). Divergence and hybridization in sea turtles: Inferences from genome data show evidence of ancient gene flow between species. *Molecular Ecology*, *30*(23), 6178–6192. https://doi.org/10.1111/mec.16113

Chapter 1

Gene family evolution and natural selection pressures across *Daphnia* genomes

Connor S. Murray*, Madison Doceti, Alan O. Bergland

Department of Biology, University of Virginia, Charlottesville, VA, USA

* Corresponding authors emails: C.S.M: csm6hg@virginia.edu

Running title: Gene family evolution in *Daphnia*.

ORCIDs:

Connor S. Murray: 0000-0002-8302-6585

Madison Doceti: 0009-0007-9801-6121

Alan O. Bergland: 0000-0001-7145-7575

Conflict of interest: The authors declare no conflicts of interest.

Keywords: *Daphnia*, gene family evolution, positive selection, spermatogenesis, stress response, neofunctionalization

Abstract

Gene family expansion and contraction underlies many molecular innovations across taxa. Understanding why specific gene families expand or contract requires comparative genomic investigations. This study investigates the gene family change dynamics within several species of *Daphnia*, a group of freshwater crustaceans that are useful model systems for evolutionary genetics. We employ comparative genomics to understand the forces driving gene evolution and draw upon candidate gene families that dynamically change. Our results suggest that genes related to spermatogenesis generally expand across taxa, and we investigate evolutionary hypotheses of adaptation by neofunctionalization that underpin these expansions. Through these analyses, we shed light on the interplay between gene expansions and selection within other ecologically relevant stress response gene families. Additionally, while we show general trends towards positive selection within expanding gene families, individual analysis of spermatogenesis genes undergoing expansion also reveals purifying selection, highlighting the complex nature of diversification and evolution within *Daphnia*. In all, our research enhances our understanding of individual gene family evolution within *Daphnia* species and provides a case study of ecologically relevant genes prone to change.

Introduction

A major goal of biology is to understand how genome changes affect function (Lewontin, 1974; Mayr, 1963). Gene family expansion, driven by the duplication of genes, is a critical process that enables species to achieve biological innovation (Hahn et al., 2007; Jordan et al., 2001). Such innovations often are a result of neofunctionalization, where gene duplicates acquire new functions (Ohno, 2013), or subfunctionalization, where paralogous genes divide the roles of their progenitors (Lynch, 2002; Lynch & Force, 2000). These processes are central to adaptive evolution across taxa, from microbes to mammals (Huang et al., 2023; Zhang et al., 2014; Hahn et al., 2007; Jordan et al., 2001; Lugli et al., 2017; Richter et al., 2018).

Gene family expansions are driven by many selective pressures and are generally associated with neofunctionalization and subfunctionalization across taxa. For example, reproductive and spermatogenesis proteins, which are often found to be expanding across taxa may evolve through sexual conflict and coevolutionary dynamics (Chang et al., 2011; Rivera & Swanson, 2022; Wang et al., 2023). Sperm and seminal fluid proteins also show elevated signals of positive selection, suggesting neofunctionalization or subfunctionalization (Dorus et al., 2008; Gang et al., 2022). Moreover, gene family expansion facilitates adaptation to environmental changes, such as the proliferation of heat-shock proteins in extreme temperature regimes (Chen et al., 2018; Zhang et al., 2012) and the variability of opsins and chemosensory genes due to differing light and chemical environmental backgrounds (Novales-Flamarique, 2013; Peñalva-Arana et al., 2009). Therefore, we hypothesize that expanding gene families are also under forms of positive selection, driving their evolution and functional diversification through processes of neofunctionalization.

In this work, we assess gene family evolution and infer the strength of selection acting upon *Daphnia*, a genus of freshwater Crustaceans that live in a range of habitats from ephemeral rain-puddles to lakes and estuaries (Fryer 1991). *Daphnia* adaptively radiated roughly 200 million years ago (mya) and encompasses at least 121 species to date. Subspecies

and cryptic speciation is common within *Daphnia* and so this number of species is likely an underestimate (Forró et al., 2008). One of the most studied taxa within Daphniidea is *Daphnia pulex*, a cryptic species complex found across both North American and European ponds (Colbourne et al., 1998; Vergilino et al., 2011; Crease et al., 2012; Murray et al., 2024). The first Crustacean genome described was *D. pulex* (Colbourne et al., 2011), and subsequent studies showed that lineages of *Daphnia* show variability in the number of genes within a gene family (Brandon et al., 2017; Lynch et al., 2017). In addition, *Daphnia* species show fluctuations in the spectrum of gene gain and loss in response to environmental change (e.g,. temperature and oxygen content fluctuations), supporting the case that gene family change is an important evolutionary mechanism (Hamza et al., 2023; Schurko et al., 2009; Zhang et al., 2021).

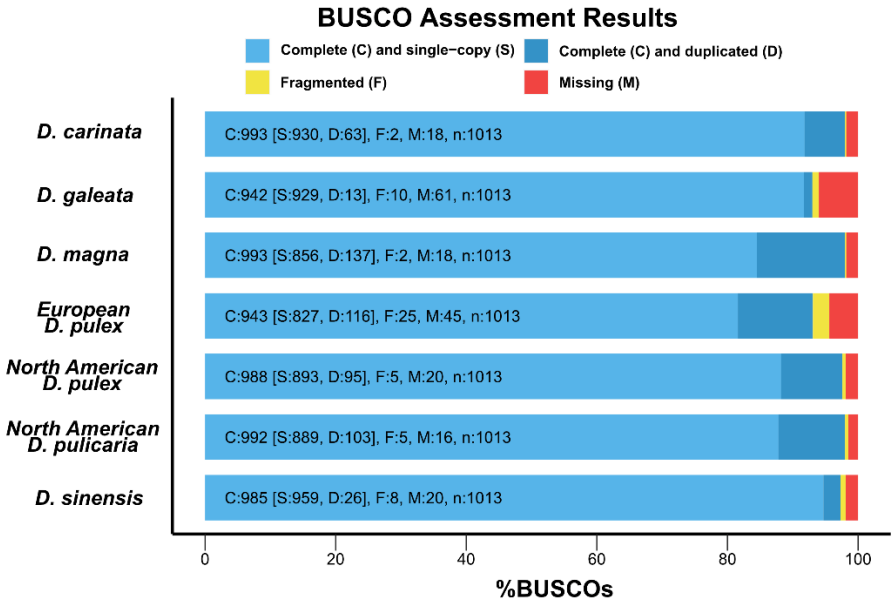
In this work, we analyze gene family evolution across *Daphnia*, highlight the expanding and contracting gene families, and we measure the strength of natural selection acting upon candidate genes. We test an overarching hypothesis that expanding gene families are also under positive selection. Alternatively, gene families that are expanding and or contracting could be under relaxed selection. Our results show substantial gene number shifts across species and that stress-response and reproduction protein gene families are expanding across *Daphnia* genomes. We detect positive selection within these overrepresented gene families, indicating a link between neofunctionalization and gene content expansion in spermatogenesis genes.

Materials and Methods

Daphnia whole-genome dataset. Chromosome-level assemblies of seven species from the genus *Daphnia* were collected from the NCBI Genome search engine (https://www.ncbi.nlm.nih.gov/datasets/genome/) accessed in July 2023 (Kitts et al., 2016). We chose North American *D. pulex* (KAP4; RefSeq: GCF_021134715.1), European *D. pulex* (D84A; GenBank: GCA_023526725; Barnard-Kubow et al., 2022), North American *D. pulicaria* (RefSeq: GCF_021234035.1; Wersebe et al., 2023), *D. sinensis* (GenBank: GCA_013167095.2, Jia et al.,

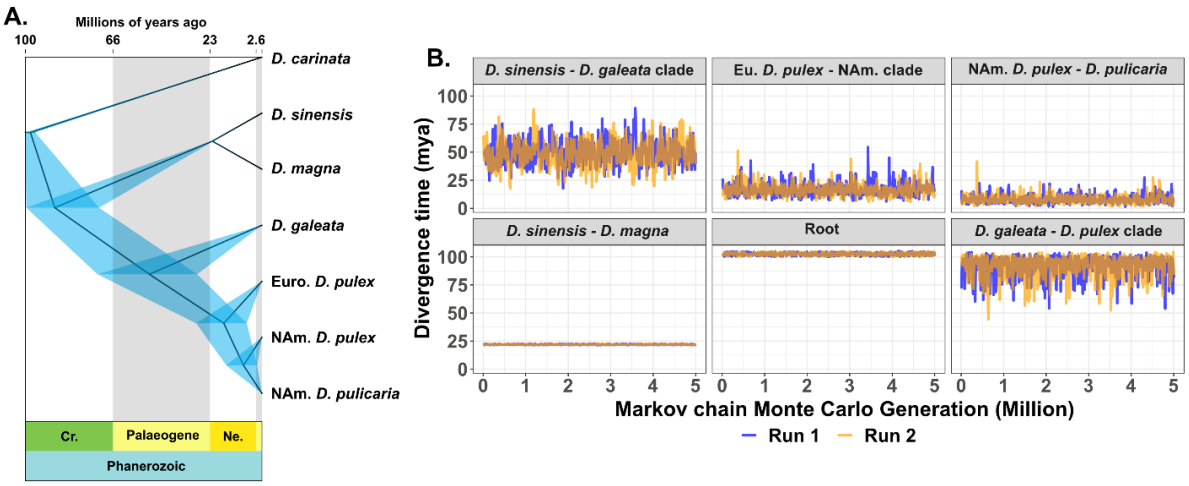
2022), *D. carinata* (RefSeq: GCF_022539665.2), *D. galeata* (GenBank: GCA_030770115.1; Nickel et al., 2021), and *D. magna* (RefSeq: GCF_020631705.1) for analyses because they are the most complete species representatives and were annotated for protein-coding genes. Additionally, the genomes were the highest quality and newest available for each unique species.

Estimating divergence-time. We used *BUSCO v5* on each genome to acquire the BUSCO score (Manni et al., 2021) and used 535 complete single-copy genes for phylogenomic analyses of divergence dating across the genus (Supplemental Figure 1). To do this, we extracted each amino acid alignment for each gene using *seqkit faidx v2.2.0* (Shen et al., 2016) and aligned them using *mafft –auto v7.505* (Katoh & Standley, 2013). We then used *clipkit -m smart-gap v2.1.1* (Steenwyk et al., 2020) to clip out regions with large gaps. After this, we concatenated all of the protein sequences together using *seqkit concat v2.2.0* and used the *mcmctree_tree_prep.py* script (https://github.com/kfuku52/mcmctree_tree_prep) to create the necessary input files for *MCMCtree v4.9e* (Dos Reis & Yang, 2019), which is part of the *PAML* package of software (Yang, 2007). We assembled a gene-tree using *IQtree v2.2.0.3* with the *modelfinder* option (Kalyaanamoorthy et al., 2017).



<u>Supplemental Figure 1:</u> Benchmark universal single copy ortholog (BUSCO) gene tree. BUSCO scores denoting the number of genes that are complete, single-copy, duplicated, fragmented, missing, and the total across each of the seven genomes used in this study.

We used several time-calibration points from previous phylogenetic investigations in *Daphnia.* Specifically, we used *D. carinata* - *D. magna* [100.4 - 104.8 mya] in place of the subgenus *Daphnia* - *Ctenodaphnia* root comparison in Cornetti et al., 2019. North American *D. pulex* – *D. magna* [130 - 150 mya] was taken from *Timetree5* (Kumar et al., 2022; Mathers et al., 2013), and *D. magna* - *D. sinensis* [21.5 - 22.4 mya] was from Cornetti et al., 2019. We used the *MCMCtreeR v1.1* package in *R* to plot the 95% confidence interval of divergence estimates across taxa (Supplemental Figure 2A; Puttick, 2019). *MCMCtree* was run twice to ensure model convergence by showing minor deviations in the estimate of node divergence times and the mean time for each node (Supplemental Figure 2 B).



<u>Supplemental Figure 2:</u> ***Daphnia*** **species tree and model convergence**. **A)** Time-calibrated phylogenetic tree of the seven whole-genomes. The tree was built with the benchmark universal single copy orthologous (BUSCO) genes that were complete and present within each genome. **B)** Trace plots of the distribution of divergence times faceted by internal nodes across two *MCMCtree* runs, purple for the first run and orange for the second. Higher variance in estimates across the MCMC leads to larger confidence intervals in the age estimates for nodes. We ran each tree for five million Markov chain Monte Carlo generations.

<u>Gene family evolution analyses and ontology enrichment.</u> We identified and retained the longest transcript with an open-reading frame in the sequence for each gene using the *primary_transcript.py* script within the *OrthoFinder v2.5.5* tool set (Emms & Kelly, 2019). We then classified orthologous genes between the seven species using *OrthoFinder*. After identifying the orthologous genes between the seven species, we annotated the phylogenetic hierarchical orthologous groups (HOGs) with the most common gene name by a majority vote using the *annotate_orthogroups* function. We used the HMMER dataset and *hmmer2go v0.18.2* functions to assign gene ontology (GO) terms to the HOGs for use in enrichment analyses (Eddy, 2011). We performed quality control for some genes families identified in *OrthoFinder* by

BLASTing each amino acid sequence against the NCBI database using *blastp v2.13.0* (Sayers et al., 2022). We use each HOG as a gene family in all subsequent analyses.

We used *CAFE5 v1.1* to estimate the expansions and contractions of gene families across the seven *Daphnia* proteomes (Mendes et al., 2021). Before running *CAFE5*, we excluded any HOGs that had over 100 genes present within any one species and any genes that were exclusively present in only one species. This was done to avoid false-positives related to expansion and contractions. *CAFE5* uses a birth-and-death process to model gene gain and loss across a phylogeny. We ran three different *CAFE5* models: the base-model (*default*; -log-likelihood = 130,984), a model with varying gamma rate categories (*-k 3*; -log-likelihood = 127,687), and a model with varying gamma rate with a root Poisson distribution (*-k 3 -p*; -log-likelihood = 74,164). These three models were chosen to test the fit of the data and convergence, as recommended by the developers (Mendes et al., 2021). We found that the base model is the best fit with our data by the highest negative log-likelihood. After this, we extracted the HOGs found to be significantly expanding or contracting within each species and used those genes as the foreground and each species' genome as the background to test enrichment of GO terms with *clusterProfiler v3.14.3* in *R* (Wu et al., 2021). We used *REVIGO v1.8.1* as a semantic reduction tool to minimize GO term redundancy for any that had over 15 enriched terms (Supek et al., 2011), and preformed Bonferroni-Holm multiple testing corrections on *p*-values.

Hypothesis testing of positive selection. To test for positive selection across gene families, we used *hyphy v2.5 aBSREL* (Kosakovsky Pond et al., 2020) on aligned codon FASTAs. *aBSREL* tests for positive selection by varying the rate of selection (*dN/dS, ω*) across both sites and branches, thus modeling site-level and branch-level *dN/dS* heterogeneity (Smith et al., 2015). *aBSREL* fits a model of *dN/dS* and performs a likelihood ratio test at each branch, comparing the full model to a null model where branches are not allowed to have rate classes of

dN/dS > 1 (Kosakovsky Pond et al., 2020). We performed one test on each tree, comparing all leaf nodes (tip branches) in a pairwise manner (Spielman et al., 2019), and examined trees to understand the patterns of selection in those potentially undergoing neofunctionalization (Hou et al., 2013; Saad et al., 2018; Mulhair et al., 2023; Wang et al., 2023). We used *pal2nal.pl v14* to align amino acid alignments and corresponding nucleotide sequences while excluding premature stop codons and gaps (Suyama et al., 2006). Sequences were aligned with *MAFFT*, visually inspected for quality, and any alignments with evidence of artificial frameshifts were removed. Each codon FASTA was run independently using a fixed gene-tree with *aBSREL*. We also used the *Datamonkey v2.0* webserver to export trees from the *aBSREL* models (Weaver et al., 2018). Our filtering efforts excluded gene families with too many orthologous genes within a species (>5 orthologs) and those without representation in most species (>5 species). We also removed gene families where *dN/dS* (ω) was ≥ 10. A gene family was considered under significant positive selection if *dN/dS* > 1 and the multiple-testing corrected *p*-value was < 0.05. We tested gene families from expanded, contracted, and non-fluctuating categories, using a two-tailed Fisher's exact test.

Single nucleotide polymorphism calling and population genetics within a wild-sequenced European D. pulex dataset. For some gene families, we wanted to understand if selection is maintaining segregating non-synonymous variation within the focal species of European *D. pulex* because we have done extensive work on the species. We tested the presence of non-synonymous single nucleotide polymorphisms (SNPs) within a metapopulation of European *D. pulex* collected from several small ponds and lakes across the United Kingdom as described in Barnard-Kubow *et al.*, 2022. We used genome-aligned bam files generated from our previous work (see Barnard-Kubow et al., 2022 and Murray et al., 2024). In short, we mapped all samples to the European *D. pulex* genome, called genotypes, merged them, performed genotype calling across the entire dataset, and annotated SNPs (Layer et al., 2014; Pedersen et

al., 2020). We chose to remove SNPs that have a minor allele frequency lower than 0.01 and we derive population genetics statistics like *pN/pS*, the number of non-synonymous (*pN*) to synonymous (*pS*) SNPs, to gauge pressures of selection affecting within-species genetic diversity.

Statistics and visualization in R. Statistical analyses were performed using *R v4.0.3* (R Core Development Team). We used the following *R* packages for general analysis and visualization: *tidyverse v1.3.1* (Wickham et al., 2019), *ggplot2 v3.3.5* (Villanueva & Chen, 2019), *ggtree v2.0.4* (Xu et al., 2022), *patchwork v1.0.1* (Pedersen, 2022), *data.table v1.12.8* (Dowle & Srinivasan, 2023), *foreach v1.4.7*, *doMC v1.3.5* (Daniel et al., 2022).

Data accessibility statement. All scripts and data used in each analysis are deposited on our GitHub repository: https://github.com/connor122721/chapter2.

Results

Daphnia genomes and the gene family dataset. From the seven *Daphnia* genomes, we first identified 117,470 unique genes across the whole dataset after extracting the longest open reading frame per protein coding transcript. From these, *Orthofinder* found 16,431 hierarchical phylogenetic orthologous gene groups (HOGs). We use these HOGs as input into *CAFE5* to estimate the evolutionary rate of gene family gain and loss and to identify any gene groups that are under positive selection. Below we use this gene grouping information to expand our understanding of the phylogenetic relationship between taxa.

Phylogeny of the represented Daphnia genomes. We built a time-calibrated phylogenetic tree to understand the relatedness of each *Daphnia* (Supplemental Figure 2A). We find that North American and European *D. pulex* diverged 15 million years ago (mya) [95% confidence

intervals; 6.4, 28.5]. Recent work highlights a split well in range of our estimates (Murray et al., 2024). We also show that North American *D. pulex* and *D. pulicaria* diverged 9 mya [2, 15.2], an estimate higher than previous at 0.5 - 2 mya (Crease et al., 2012). We are confident that the internal node split-times are converging across the trees because of the lack of deviation from two independent runs (Supplemental Figure 2B).

Trends of gene family expansion and contraction. We used *CAFE5* to identify expanding and contracting gene families across *Daphnia* (Figure 1; Mendes et al., 2021). The base *CAFE5* model maximized the negative log-likelihood value (-log-likelihood = 130,984), so we are reporting its output. The first finding is that *D. magna* has the largest expansion within their genome (N_{Genes} = 634; Figure 1), while European *D. pulex* has the largest contraction (N_{Genes} = 745; Figure 1) compared to their most recent common ancestor. For the remainder of this work, we investigate the genes belonging to the 1,606 phylogenetic hierarchical ortholog groups (HOGs) identified as being significant candidates that are expanding and contracting across the tree (Supplemental Table 1).

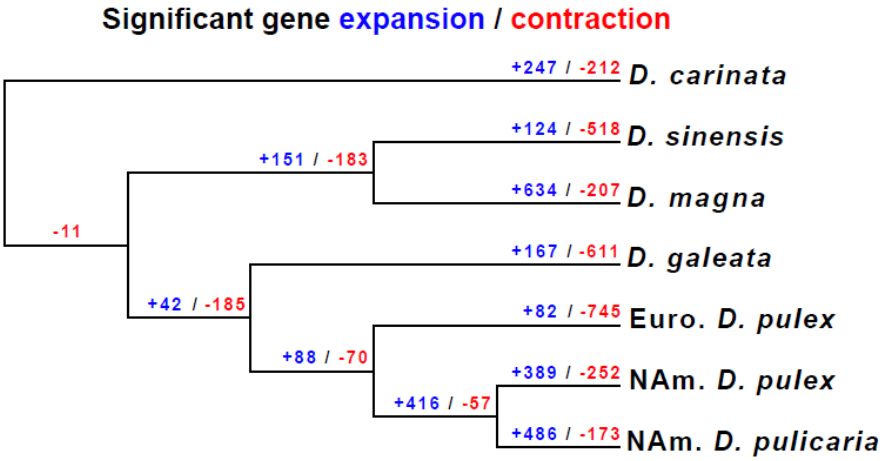
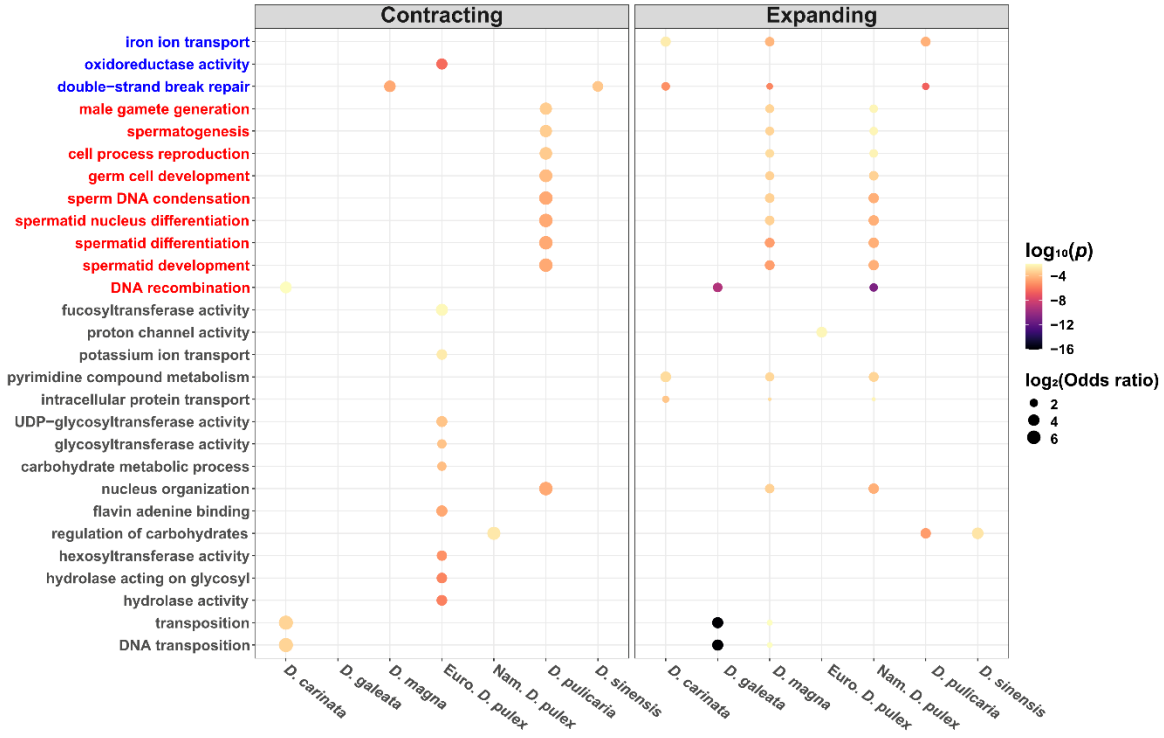


Figure 1: Patterns of gene family evolution across *Daphnia*. Results of gene family evolution analyses across the phylogenetic tree from the base *CAFE5* model run. The blue

colored numbers indicate the number of genes gained and the red indicate the number of genes lost within each node and terminal leaf.

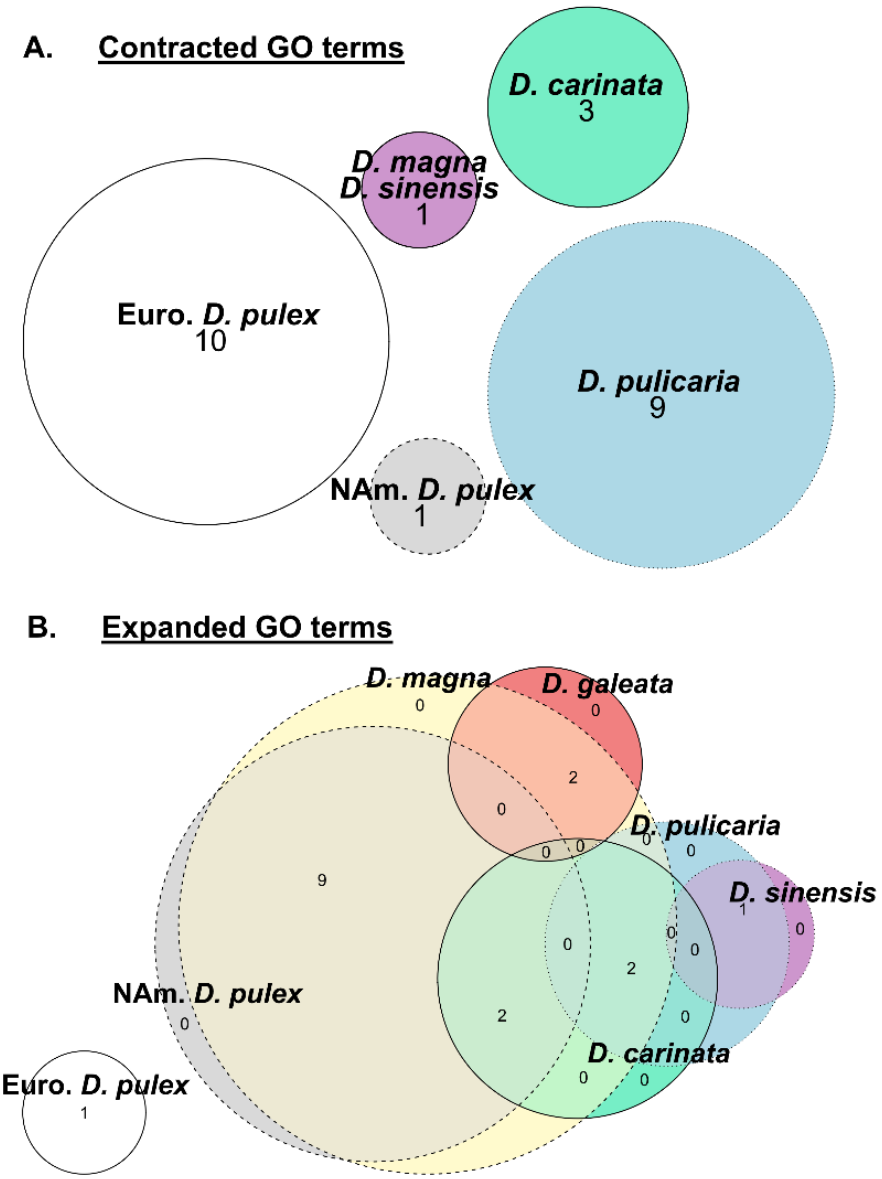
Next, tested how gene expansions and contractions are related to function (Lespinet et al., 2002; Sánchez-Gracia et al., 2009). To investigate this, we measured the enrichment of gene ontology (GO) terms associated with expanding and contracting genes within each species' genome. We identified the most common expanding genes in our dataset and found that five spermatogenesis-related terms were enriched in *D. magna* and North American *D. pulex* (Figure 2). In *D. carinata* and *D. pulicaria*, there was enrichment of terms related to iron ion transport and double-strand DNA repair (Figure 2).

For gene contractions, the most noticeable pattern was observed in *D. pulicaria*, which showed contractions in the same five spermatogenesis terms mentioned above. European *D. pulex* exhibited contractions related to hydrolase and other transferase activities, which may be linked to changes in enzymatic processes related to carbohydrate metabolism (Zeis et al., 2009).



<u>Figure 2:</u> Significantly expanding and contracting gene families and their gene ontology (GO) enrichment across species reveals an excess of spermatogenesis and stress response terms. Presence and absence data of the most enriched terms across species. The y-axis are the enriched terms. The blue labeled GO terms are related to general stress responses and the red colored terms indicate any related to reproduction pathways. All terms have been semantically reduced.

Beyond the clear examples from the enrichment data, many expanding and contracting GO terms are enriched in only one or two species (Supplemental Figure 3 A&B; Supplemental Table 2). Most contracted terms are unique to each species, except for double-strand break repair, which is found in both *D. magna* and *D. sinensis* (Figure 2). In contrast, the expansion terms are often shared across species (Supplemental Figure 3B), particularly those related to spermatogenesis (Figure 2).



Supplemental Figure 3: Enriched gene ontology term sharing related to expansions and contractions of gene families. The size of each circle represents the relative number of enriched terms that are shared or unique to each species and the color represents each species. Faceted by **A)** the contracted terms or **B)** the expanded GO terms. All GO terms have been semantically reduced.

General patterns of positive selection on expanding and contracting gene families: We tested the hypothesis that expanding gene families undergo higher rates of selection by using *hyphy v2.5 aBSREL* (Kosakovsky Pond et al., 2020) on codon sequences from gene families identified by *OrthoFinder*. Gene families with extreme *dN/dS* (*ω*) values greater than 10 were excluded. We classified a gene as being under positive selection if *dN/dS* > 1 and the multiple testing adjusted *p*-value < 0.05. Our findings show that among the genes in expanded gene families, 14 out of 649 trees (2.2%) exhibited signals of positive selection. This is compared to the non-fluctuating gene families, where 112 out of 9,401 trees (1.2%) showed positive selection. A Fisher's exact test indicated a significant positive odds ratio of 1.81 [95% confidence interval: 0.95, 3.19] (two-tailed Fisher's exact test; $p = 0.044$). In contrast, contracted gene families had 2 out of 62 trees (3.2%) showing positive selection, with a positive odds ratio of 2.71 [95% confidence interval: 0.31, 10.4] ($p = 0.18$).

Heme production gene families and natural selection in Daphnia: We investigated whether gene families involved in iron-ion binding and heme production are subject to positive selection using *aBSREL*. Our analysis concentrated on *soma ferritin-like* genes, which demonstrated expansion within the iron-ion binding GO term (Figure 2) and are associated with *Daphnia*'s stress responses (Zeis et al., 2009). We assessed significance with the likelihood ratio test, correcting for multiple testing. *aBSREL* identified positive selection in one out of 19 branches of the *soma ferritin-like* gene tree, specifically in *D. galeata*, which exhibited elevated *dN/dS* ratios among three expanded genes (Figure 3A; $p = 0.0034$). This gene family appears to be a hotspot for expansion, with two expansion events in both *D. pulicaria* and *D. magna*, and an additional copy in North American *D. pulex*. However, positive selection was only detected in the *D. galeata* branch (Figure 3A).

Beyond the *ferritin* gene family, the *iron-sulfur cluster co-chaperone protein HscB-like* family showed significant expansion in North American *D. pulex*, though no positive selection

was observed. For the *heme-binding protein 2-like* family, positive selection was detected in two out of 13 branches, in *D. galeata* (Figure 3B; $p = 2.2x10^{-16}$) and North American *D. pulex* (Figure 3B; $p = 0.0078$). The *cytosolic iron-sulfur assembly* protein family showed positive selection in two out of 13 branches within European D. pulex (Figure 3C; $p = 8.2x10^{-7}$) and *D. galeata* (Figure 3C; $p = 0.0031$). Additionally, the *heme oxygenase-like* family exhibited positive selection in *D. magna* (Figure 3D; $p = 0.017$).

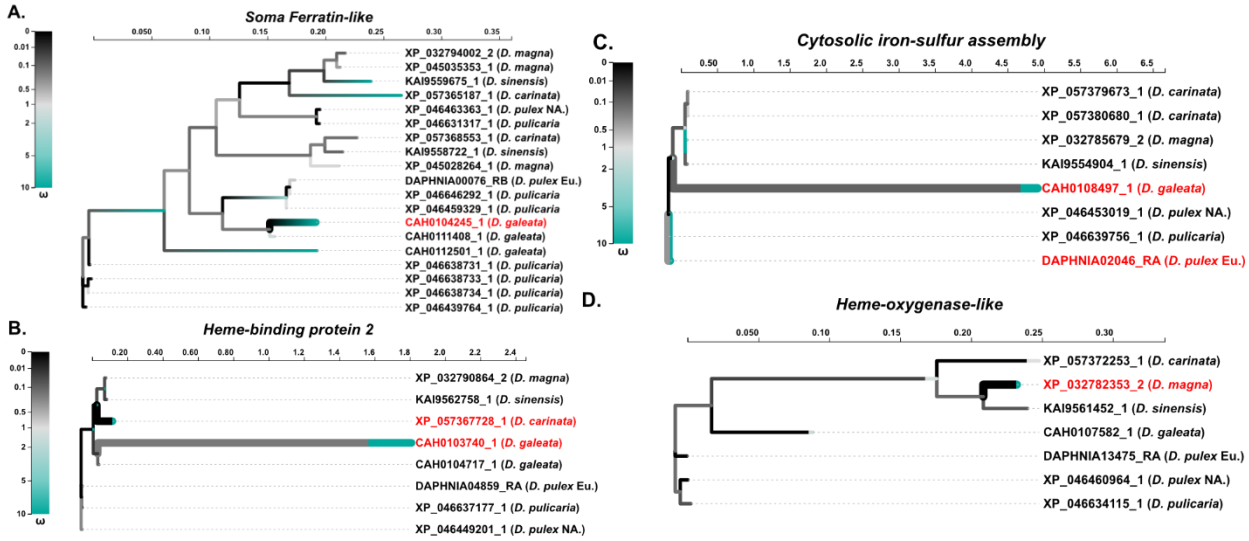


Figure 3: Case study of positive selection on expanded heme production and iron-ion binding genes. *aBSREL* model of codon evolution from gene trees identified as expanded. Red labeled gene names indicate significant gene branches undergoing positive selection. The thickness of the terminal branches indicates the significance after multiple testing correction, the color indicates the *dN/dS* (ω), and the x-axis branch lengths are in substitutions. **A)** *Soma-ferratin-like* gene family shows many expansions in *D. pulicaria*, *D. magna*, and *D. galeata* with a *D. galeata* branch under strong evidence of positive selection. **B)** *Heme-binding protein 2* family with selection in *D. carinata* and *D. galeata*. **C)** *Cytosolic iron-sulfur assembly* family with selection in *D. galeata* and European *D. pulex*. **D)** A *heme oxygenase-like* protein family with

selection in *D. magna*.

Outside of the expansion events, we identified 24 gene families that were related to iron ion binding and heme production and a total of 248 individual genes are classified. We highlight that 7 genes are detected to have positive selection, roughly affecting 3% of iron ion binding and heme genes across *Daphnia*. So in all we show evidence for expanded genes being under positive selection and thus could be candidates for neofunctionalization.

Spermatogenesis gene families and natural selection in Daphnia. We investigate whether expanding gene families associated with spermatogenesis proteins undergo positive selection, utilizing *aBSREL*. In *D. magna* and North American *D. pulex*, we observe an expansion linked to sperm production (Figure 2). Consequently, we tested for positive selection within a *spermine oxidase-like* gene family, a gene family with two duplications within both North American *D. pulex* and *D. magna* (Figure 4A). *aBSREL* found no discernible signals of positive selection within the *spermine oxidase-like* gene family. Additionally, a *spermatogenesis-associated protein* gene family, exhibiting duplications of two genes in *D. magna* (Figure 4B; XP_045029128_1), also shows no signal for positive selection. Despite significant expansion patterns identified through *CAFE5* results, both spermatogenesis gene families show evidence for purifying selection based on the absence of positive selection signals.

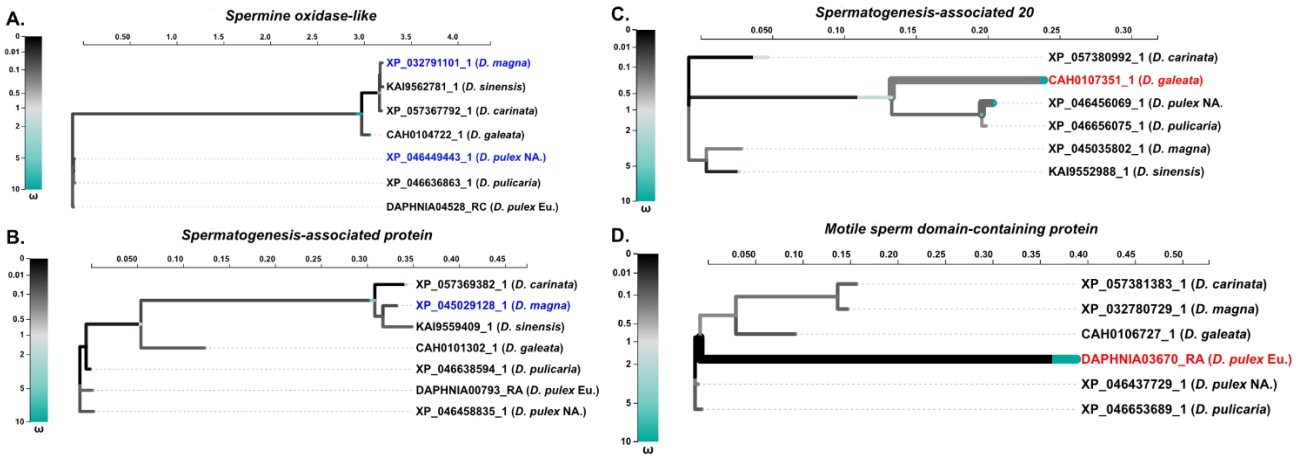


Figure 4: Case study of natural selection on spermatogenesis gene families. *aBSREL* model of codon evolution from gene trees identified as expanding and contracting or stable related to spermatogenesis function. Blue labeled gene labels indicate a branch that underwent a significant expansion, the software reduced sequence redundancy in the two expansions for both North American *D. pulex* (XP_046449443_1) and *D. magna* (XP_032791101_1) due to short branch lengths. (**A, B**). Red labeled gene names indicate significant gene branches undergoing positive selection (**C**, **D**). The thickness of the terminal branches indicates the significance after multiple testing correction, the color of the branch indicates the varying *dN/dS* (ω) rates, and the branch lengths are in substitutions.

Beyond the genes identified as expanded and contracted by *CAFE5*, we detected positive selection in several gene families related to spermatogenesis. In European *D. pulex*, a *motile sperm domain-containing protein I* family shows strong positive selection (p = 1.7x10^{-13}) with a *pN/pS* ratio of 1.07 (16 polymorphic non-synonymous SNPs (*pN*) and 15 polymorphic synonymous SNPs (*pS*)) based on a population genomics dataset from wild-sequenced North American and European *D. pulex* (Barnard-Kubow et al., 2022; Murray et al., 2024). In North American *D. pulex*, the *pN/pS* ratio is 1 (*pN* = 4, *pS* = 4). In *D. sinensis*, we found strong signals of positive selection in a *motile sperm domain-containing protein II* family (p = 2.2x10^{-16}). A

spermatogenesis-associated protein 20-like gene family shows positive selection in two of the six branches in *D. galeata* (p = 0.0007) and *D. magna* (p = 0.01; Figure 4C). A *spermine synthase-like* family shows positive selection within *D. carinata* (p = 0.017). Additionally, the *histone H2A* gene family, which is related to spermatogenesis and DNA packaging within sperm, shows contractions of three genes in European *D. pulex* and *D. galeata*, and the loss of four genes in *D. pulicaria*. Selection could not be determined within this family due to the lack of other unique species' genes. Investigating the mechanisms driving these contractions in *histone H2A* could provide insights into adaptive reproductive strategies, potentially explaining the enrichment of contraction signals in sperm genes observed in Figure 2 for *D. pulicaria*.

In total, out of the 14 hierarchical gene families identified by *OrthoFinder* and 82 individual genes related to spermatogenesis and sperm packaging in our dataset, four genes show significant positive selection across species, constituting approximately 5% of spermatogenesis genes across *Daphnia*.

Discussion

In this study, we test hypotheses concerning the neofunctionalization of gene families that have expanded across *Daphnia*. Our findings reveal an overrepresentation of GO terms related to spermatogenesis and stress response within the expanded gene families (Figure 2). We provide evidence of elevated positive selection affecting approximately 2% of the expanding gene families (Figure 2). This indicates neofunctionalization within the expanding genes, and we illustrate notable examples in heme (Figure 3) and spermatogenesis-related genes (Figure 4). Overall, the patterns of gene family expansion and contraction are consistent with the general hypothesis that gene family evolution should be elevated for traits associated with sexual conflict and environmental adaptation.

Evolutionary genomics of gene family evolution in Daphnia species. Our research explores the evolutionary dynamics of gene families in *Daphnia*, investigating how selective pressures influence gene content and diversity across species. *Daphnia* face similar selection pressures due to predation (Schwartz, 1984) and seasonal adaptation (Winder et al., 2004). We hypothesized that these shared ecological pressures would result in gene expansions and contractions across *Daphnia* (Hebert & Wilson, 1994; Chin & Cristescu, 2021). Contrary to our hypothesis, our findings reveal that *Daphnia* experience selection pressures in a species-specific manner (Figure 2). Many gene families we expected to be common targets for expansion, such as heat-shock proteins, opsins, and chemosensory genes, were not frequently expanded. Instead, a general trend emerged where some *Daphnia* showed expansions in genes related to spermatogenesis, metabolic processes, and stress responses involving heme production (Figure 2). This indicates that while *Daphnia* face similar ecological challenges, their genetic responses to these pressures are highly individualized.

Comparative genomic analyses have revealed significant gene family expansions in *Daphnia*. A study by Zhang *et al.*, (2021) using three *Daphnia* genomes identified expansions related to methylation in *D. pulicaria* and North American *D. pulex*, and structural morphogenesis in *D. carinata* and North American *D. pulex* (Supplemental Table 2). Similarly, Ye *et al.*, (2017) found terms associated with chitin binding and oxidative stress in their analysis of two *Daphnia* species. They also noted high variability in heme production genes in North American *D. pulex* and *D. magna*, related to the adaptation to hypoxic environments via hemoglobin proteins (Fox et al., 1951; Kobayashi et al., 1994). While there is some overlap with our results, which primarily show expansions in spermatogenesis terms, our findings differ due to the inclusion of a broader set of ortholog groups and additional *Daphnia* species. This underscores the importance of comparative genomics in understanding *Daphnia* adaptation across multiple genomes.

The analysis of *Daphnia* genomes has provided valuable insights into their genetic complexity and evolutionary dynamics. The first *Daphnia* genome, *D. pulex arenata*, revealed over 30,000 genes, more than twice the number found in humans (Colbourne et al., 2011). Although many of these genes are now considered erroneous due to fragmented models (Denton et al., 2014), some might still represent significant evolutionary events (Ye et al., 2017). By using the error prediction feature (Han et al., 2013), which estimates the influence of assembly error on gene expansion and contraction estimates, we found an error rate of 5.6%. This rate is comparable to other genomic projects (Neale et al., 2017) and like that observed in *Drosophila* genomes (Da Lage et al., 2019). The gene family gain and loss rate (λ) across the phylogeny is λ = 0.00084, about an order of magnitude lower than estimates in similar *Drosophila* studies (Hahn et al., 2007; Da Lage et al., 2019). This suggests that the *Daphnia* genomes examined have relatively low evolutionary rates of gene gain and loss, indicating a conservative estimate. To minimize assembly bias, we included the highest quality genomes available and will incorporate more in future investigations.

Evolution of spermatogenesis gene families in Daphnia. Our study provides valuable insights into species-specific adaptations and fundamental evolutionary processes, particularly in spermatogenesis. *Daphnia* sperm exhibit significant size variability and extensive phenotypic diversification (Duneau et al., 2022). At the sequence level, we observed positive selection and diversification in genes associated with sperm morphology. For instance, in European *D. pulex*, a motile sperm protein (MSP) and other spermatogenesis proteins show high rates of positive selection (Figure 4C&D). The MSP gene is particularly intriguing, with a *pN/pS* ratio ≥ 1 in both North American and European *D. pulex*, indicating the accumulation of functional diversity through non-synonymous substitutions and polymorphisms, suggesting neofunctionalization. Additionally, *D. pulex* species exhibit highly variable male production rates (Ye et al., 2019;

Barnard-Kubow et al., 2022), implying similar selective pressures on the MSP gene family across distinct populations.

Despite these findings, our understanding of the evolutionary genetics of spermatogenesis in *Daphnia* remains limited (Wuerz et al., 2017). This gap in knowledge is likely due to a predominant research focus on parthenogenesis and meiosis pathways (Gómez et al., 2016; Schurko et al., 2009). Although some *Daphnia* rarely produce males during much of the growing season (Wuerz et al., 2017), the genes controlling spermatogenesis could still be under selective pressures that are not yet fully explored. For example, obligately asexual North American *D. pulex* males exhibit varying sperm ploidy and can spread asexuality to other clones (Xu et al., 2013, 2015), indicating that sperm evolution might drive diversification pressures across the species range. In a broader taxonomic context, major sperm proteins in nematodes show highly conserved sequences despite extensive gene family expansion (Kasimatis & Phillips, 2018). Similarly, sperm motility genes are often associated with positive selection in mammals (Torgerson et al., 2002; Vicens et al., 2014), and *Drosophila* sperm demonstrate extreme variability (Civetta et al., 2006; Wong et al., 2008). Thus, the genes involved in spermatogenesis in *Daphnia* are likely subject to significant evolutionary pressures, underscoring the need for more detailed studies to uncover their roles and adaptive significance related to neofunctionalization and subfunctionalization.

In summary, spermatogenesis emerges as a compelling set of genes under diversification in *Daphnia*, showing enrichment for expansions and contractions. Genes related to spermatogenesis and sperm morphology exhibit high levels of positive selection across branches and sites (Figure 4C&D), with approximately 5% of spermatogenesis genes identified as under positive selection. Additionally, some expanding spermatogenesis genes appear to be under purifying selection (Figure 4A&B). These candidate genes warrant further investigation to understand their relevance to adaptation.

Conclusion

Our study elucidates the gene family evolution of several members of *Daphnia,* and we provide evidence that spermatogenesis and stress response genes are under gene number evolution. We also show that these genes prone to turnover are also under some incidents of positive selection, leading us to understand the early phases of gene diversification and neofunctionalization within *Daphnia*. Our study has important implications for continuing the work to elucidate the mechanisms that drive divergence across species, and we highlight the need to further validate how spermatogenesis genes are functional within species (Genereux et al., 2020). Ultimately though, we linked gene evolution with diversification across an interesting group of taxa.

Author contributions. CSM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing - original draft, Writing - review & editing. MD: Formal analysis, Investigation, Visualization, Writing - review & editing. AOB: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing

Acknowledgments. The authors wish to acknowledge members of the Bergland lab for their discussion and feedback related to the manuscript's development. The authors acknowledge Research Computing at the University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication. URL: https://rc.virginia.edu.

Funding information. A.O.B. was supported by grants from the National Institutes of Health (R35 GM119686), and by start-up funds provided by the University of Virginia. C.S.M. was supported by an Expand National Science Foundation Research Traineeship program at UVA.

References

Barnard-Kubow, K. B., Becker, D., Murray, C. S., Porter, R., Gutierrez, G., Erickson, P., Nunez, J. C. B., Voss, E., Suryamohan, K., Ratan, A., Beckerman, A., & Bergland, A. O. (2022). Genetic Variation in Reproductive Investment Across an Ephemerality Gradient in *Daphnia pulex*. *Molecular Biology and Evolution*, *39*(6), msac121. https://doi.org/10.1093/molbev/msac121

Brandon, C. S., Greenwold, M. J., & Dudycha, J. L. (2017). Ancient and Recent Duplications Support Functional Diversity of *Daphnia* Opsins. *Journal of Molecular Evolution*, *84*(1), 12–28. https://doi.org/10.1007/s00239-016-9777-1

Chang, T.-C., Yang, Y., Yasue, H., Bharti, A. K., Retzel, E. F., & Liu, W.-S. (2011). The Expansion of the PRAME Gene Family in Eutheria. *PLoS ONE*, *6*(2), e16867. https://doi.org/10.1371/journal.pone.0016867

Chen, B., Feder, M. E., & Kang, L. (2018). Evolution of heat-shock protein expression underlying adaptive responses to environmental stress. *Molecular Ecology*, *27*(15), 3040–3054. https://doi.org/10.1111/mec.14769

Chin, T. A., & Cristescu, M. E. (2021). Speciation in *Daphnia*. *Molecular Ecology*, mec.15824. https://doi.org/10.1111/mec.15824

Civetta, A., Rajakumar, S. A., Brouwers, B., & Bacik, J. P. (2006). Rapid Evolution and Gene-Specific Patterns of Selection for Three Genes of Spermatogenesis in *Drosophila*. *Molecular Biology and Evolution*, *23*(3), 655–662. https://doi.org/10.1093/molbev/msj074

Colbourne, J. K., Crease, T. J., Weider, L. J., Hebert, P. D. N., Duferesne, F., & Hobaek, A. (1998). Phylogenetics and evolution of a circumarctic species complex (Cladocera: *Daphnia pulex*). *Biological Journal of the Linnean Society*, *65*(3), 347–365. https://doi.org/10.1111/j.1095-8312.1998.tb01146.x

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., Tokishita, S., Aerts, A., Arnold, G. J., Basu, M. K., Bauer, D. J., Caceres, C. E., Carmel, L., Casola, C., Choi, J.-H., Detter, J. C., Dong, Q., Dusheyko, S., Eads, B. D., … Boore, J. L. (2011). The Ecoresponsive Genome of *Daphnia pulex*. *Science*, *331*(6017), 555–561. https://doi.org/10.1126/science.1197761

Cornetti, L., Fields, P. D., Van Damme, K., & Ebert, D. (2019). A fossil-calibrated phylogenomic analysis of *Daphnia* and the Daphniidae. *Molecular Phylogenetics and Evolution*, *137*, 250–262. https://doi.org/10.1016/j.ympev.2019.05.018

Crease, T. J., Omilian, A. R., Costanzo, K. S., & Taylor, D. J. (2012). Transcontinental Phylogeography of the *Daphnia pulex* Species Complex. *PLoS ONE*, *7*(10), e46620. https://doi.org/10.1371/journal.pone.0046620

Da Lage, J.-L., Thomas, G. W. C., Bonneau, M., & Courtier-Orgogozo, V. (2019). Evolution of salivary glue genes in *Drosophila* species. *BMC Evolutionary Biology*, *19*(1), 36. https://doi.org/10.1186/s12862-019-1364-9

Daniel, F., Revolution Analytics, & Weston, S. (2022). *doMC: Foreach Parallel Adaptor for "parallel."*

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Computational Biology*, *10*(12), e1003998. https://doi.org/10.1371/journal.pcbi.1003998

Dorus, S., Freeman, Z. N., Parker, E. R., Heath, B. D., & Karr, T. L. (2008). Recent Origins of Sperm Genes in *Drosophila*. *Molecular Biology and Evolution*, *25*(10), 2157–2166. https://doi.org/10.1093/molbev/msn162

Dos Reis, M., & Yang, Z. (2019). Bayesian Molecular Clock Dating Using Genome-Scale Datasets. In M. Anisimova (Ed.), *Evolutionary Genomics* (Vol. 1910, pp. 309–330). Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_10

Dowle, M., & Srinivasan, A. (2023). data.table: Extension of “data.frame.” *Https://R-Datatable.Com, Https://Rdatatable.Gitlab.Io/Data.Table, Https://Github.Com/Rdatatable/Data.Table*.

Duneau, D., Möst, M., & Ebert, D. (2022). Evolution of sperm morphology in a crustacean genus with fertilization inside an open brood pouch. *Peer Community Journal*, *2*, e63. https://doi.org/10.24072/pcjournal.182

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Emms, D. M., & Kelly, S. (2019). *OrthoFinder*: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Forró, L., Korovchinsky, N. M., Kotov, A. A., & Petrusek, A. (2008). Global diversity of cladocerans (Cladocera; Crustacea) in freshwater. *Hydrobiologia*, *595*(1), 177–184. https://doi.org/10.1007/s10750-007-9013-5

Fox, H. M., Gilchrist, B. M., and Phear, E. A. (1951). Functions of haemoglobin in *Daphnia*. *Proceedings of the Royal Society of London. Series B - Biological Sciences*, *138*(893), 514–528. https://doi.org/10.1098/rspb.1951.0038

Gang, H., Bin, W., Yali, Y., Weiping, W., Zhihong, Z., & Li, H. (2022). Comparison of Fecundity Traits between *Mauremys reevesii* and Trachemys scripta based on Gene Family Analysis. *Indian Journal of Animal Research*, *Of*. https://doi.org/10.18805/IJAR.BF-1479

Genereux, D. P., Serres, A., Armstrong, J., Johnson, J., Marinescu, V. D., Murén, E., Juan, D., Bejerano, G., Casewell, N. R., Chemnick, L. G., Damas, J., Di Palma, F., Diekhans, M., Fiddes, I. T., Garber, M., Gladyshev, V. N., Goodman, L., Haerty, W., Houck, M. L., … Karlsson, E. K. (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, *587*(7833), 240–245. https://doi.org/10.1038/s41586-020-2876-6

Geoffrey Fryer. (1991). Functional morphology and the adaptive radiation of the Daphniidae (Branchiopoda: Anomopoda). *Philosophical Transactions of the Royal Society of*

London. Series B: Biological Sciences, *331*(1259), 1–99. https://doi.org/10.1098/rstb.1991.0001

Gómez, R., Van Damme, K., Gosálvez, J., Morán, E. S., & Colbourne, J. K. (2016). Male meiosis in Crustacea: Synapsis, recombination, epigenetics and fertility in *Daphnia magna*. *Chromosoma*, *125*(4), 769–787. https://doi.org/10.1007/s00412-015-0558-1

Hahn, M. W., Han, M. V., & Han, S.-G. (2007). Gene Family Evolution across 12 *Drosophila* Genomes. *PLoS Genetics*, *3*(11), e197. https://doi.org/10.1371/journal.pgen.0030197

Hamza, W., Hazzouri, K. M., Sudalaimuthuasari, N., Amiri, K. M. A., Neretina, A. N., Al Neyadi, S. E. S., & Kotov, A. A. (2023). Genome Assembly of a Relict Arabian Species of *Daphnia* O. F. Müller (Crustacea: Cladocera) Adapted to the Desert Life. *International Journal of Molecular Sciences*, *24*(1), 889. https://doi.org/10.3390/ijms24010889

Han, M. V., Thomas, G. W. C., Lugo-Martinez, J., & Hahn, M. W. (2013). Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Molecular Biology and Evolution*, *30*(8), 1987–1997. https://doi.org/10.1093/molbev/mst100

Hebert, P. D. N., & Wilson, C. C. (1994). Provincialism in plankton: endemism and allopatric speciation in australian *Daphnia*. *Evolution*, *48*(4), 1333–1349. https://doi.org/10.1111/j.1558-5646.1994.tb05317.x

Hou, Y., Sierra, R., Bassen, D., Banavali, N. K., Habura, A., Pawlowski, J., & Bowser, S. S. (2013). Molecular Evidence for β-tubulin Neofunctionalization in *Retaria* (Foraminifera and Radiolarians). *Molecular Biology and Evolution*, *30*(11), 2487–2493. https://doi.org/10.1093/molbev/mst150

Huang, Z., Jiang, C., Gu, J., Uvizl, M., Power, S., Douglas, D., & Kacprzyk, J. (2023). Duplications of Human Longevity-Associated Genes Across Placental Mammals. *Genome Biology and Evolution*, *15*(10), evad186. https://doi.org/10.1093/gbe/evad186

Jia, J., Dong, C., Han, M., Ma, S., Chen, W., Dou, J., Feng, C., & Liu, X. (2022). Multi-omics perspective on studying reproductive biology in *Daphnia sinensis*. *Genomics*, *114*(2), 110309. https://doi.org/10.1016/j.ygeno.2022.110309

Jordan, I. K., Makarova, K. S., Spouge, J. L., Wolf, Y. I., & Koonin, E. V. (2001). Lineage-Specific Gene Expansions in Bacterial and Archaeal Genomes. *Genome Research*, *11*(4), 555–565. https://doi.org/10.1101/gr.166001

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). *ModelFinder*: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kasimatis, K. R., & Phillips, P. C. (2018). Rapid Gene Family Evolution of a Nematode Sperm Protein Despite Sequence Hyper-conservation. *G3 Genes|Genomes|Genetics*, *8*(1), 353–362. https://doi.org/10.1534/g3.117.300281

Katoh, K., & Standley, D. M. (2013). *MAFFT* Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kitts, P. A., Church, D. M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., Smith, R. G., Tatusova, T., Xiang, C., Zherikov, A., DiCuccio, M., Murphy, T. D., Pruitt, K. D., & Kimchi, A. (2016). *Assembly*: A resource for assembled genomes at NCBI. *Nucleic Acids Research*, *44*(D1), D73–D80. https://doi.org/10.1093/nar/gkv1226

Kobayashi, M., Ishigaki, K., Kobayashi, M., Igarashi, Y., & Imai, K. (1994). Oxygen transport efficiency of multiple-component hemoglobin in *Daphnia magna*. *Canadian Journal of Zoology*, *72*(12), 2169–2171. https://doi.org/10.1139/z94-289

Kosakovsky Pond, S. L., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., Shank, S. D., Magalis, B. R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S. J., Frost, S. D. W., & Muse, S. V. (2020). *HyPhy 2.5*—A Customizable Platform for

Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution*, *37*(1), 295–299. https://doi.org/10.1093/molbev/msz197

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, *39*(8), msac174. https://doi.org/10.1093/molbev/msac174

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). *LUMPY*: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), R84. https://doi.org/10.1186/gb-2014-15-6-r84

Lespinet, O., Wolf, Y. I., Koonin, E. V., & Aravind, L. (2002). The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. *Genome Research*, *12*(7), 1048–1059. https://doi.org/10.1101/gr.174302

Lewontin, R. C. (1974). *The genetic basis of evolutionary change (Vol. 560).* New York: Columbia University Press.

Lugli, G. A., Milani, C., Turroni, F., Duranti, S., Mancabelli, L., Mangifesta, M., Ferrario, C., Modesto, M., Mattarelli, P., Jiří, K., Van Sinderen, D., & Ventura, M. (2017). Comparative genomic and phylogenomic analyses of the Bifidobacteriaceae family. *BMC Genomics*, *18*(1), 568. https://doi.org/10.1186/s12864-017-3955-4

Lüpold, S., De Boer, R. A., Evans, J. P., Tomkins, J. L., & Fitzpatrick, J. L. (2020). How sperm competition shapes the evolution of testes and sperm: A meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1813), 20200064. https://doi.org/10.1098/rstb.2020.0064

Lynch, M. (2002). Gene Duplication and Evolution. *Science*, *297*(5583), 945–947. https://doi.org/10.1126/science.1075472

Lynch, M., & Force, A. (2000). The Probability of Duplicate Gene Preservation by Subfunctionalization. *Genetics*, *154*(1), 459–473. https://doi.org/10.1093/genetics/154.1.459

Lynch, M., Gutenkunst, R., Ackerman, M., Spitze, K., Ye, Z., Maruki, T., & Jia, Z. (2017). Population Genomics of *Daphnia pulex*. *Genetics*, *206*(1), 315–332. https://doi.org/10.1534/genetics.116.190611

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Mathers, T. C., Hammond, R. L., Jenner, R. A., Hänfling, B., & Gómez, A. (2013). Multiple global radiations in tadpole shrimps challenge the concept of 'living fossils.' *PeerJ*, *1*, e62. https://doi.org/10.7717/peerj.62

Mayr, E. (1963). *Animal Species and Evolution:* Harvard University Press. https://doi.org/10.4159/harvard.9780674865327

Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2021). *CAFE 5* models variation in evolutionary rates among gene families. *Bioinformatics*, *36*(22–23), 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022

Mulhair, P. O., Crowley, L., Boyes, D. H., Lewis, O. T., & Holland, P. W. H. (2023). Opsin Gene Duplication in *Lepidoptera*: Retrotransposition, Sex Linkage, and Gene Expression. *Molecular Biology and Evolution*, *40*(11), msad241. https://doi.org/10.1093/molbev/msad241

Murray, C. S., Karram, M., Bass, D. J., Doceti, M., Becker, D., Nunez, J. C. B., Ratan, A., & Bergland, A. O. (2024). Balancing selection and the functional effects of shared polymorphism in cryptic *Daphnia species*. *bioRxiv*. https://doi.org/10.1101/2024.04.16.589693

Neale, D. B., McGuire, P. E., Wheeler, N. C., Stevens, K. A., Crepeau, M. W., Cardeno, C., Zimin, A. V., Puiu, D., Pertea, G. M., Sezen, U. U., Casola, C., Koralewski, T. E., Paul, R., Gonzalez-Ibeas, D., Zaman, S., Cronn, R., Yandell, M., Holt, C., Langley, C. H., … Wegrzyn, J. L. (2017). The Douglas-Fir Genome Sequence Reveals Specialization of the Photosynthetic Apparatus in Pinaceae. *G3 Genes|Genomes|Genetics*, *7*(9), 3157–3167. https://doi.org/10.1534/g3.117.300078

Nickel, J., Schell, T., Holtzem, T., Thielsch, A., Dennis, S. R., Schlick-Steiner, B. C., Steiner, F. M., Möst, M., Pfenninger, M., Schwenk, K., & Cordellier, M. (2021). Hybridization Dynamics and Extensive Introgression in the *Daphnia longispina* Species Complex: New Insights from a High-Quality *Daphnia galeata* Reference Genome. *Genome Biology and Evolution*, *13*(12), evab267. https://doi.org/10.1093/gbe/evab267

Novales-Flamarique, I. (2013). Opsin switch reveals function of the ultraviolet cone in fish foraging. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1752), 20122490. https://doi.org/10.1098/rspb.2012.2490

Ohno, S. (2013). *Evolution by gene duplication*. Springer Science & Business Media.

Pedersen, Brent S., Layer, Ryan, & Quinlan, Aaron R. (2020). *smoove: Structural-variant calling and genotyping with existing tools* [Computer software]. Apache-2.0.

Peñalva-Arana, D. C., Lynch, M., & Robertson, H. M. (2009). The chemoreceptor genes of the waterflea *Daphnia pulex*: Many Grs but no Ors. *BMC Evolutionary Biology*, *9*(1), 79. https://doi.org/10.1186/1471-2148-9-79

Puttick, M. N. (2019). *MCMCtreeR*: Functions to prepare *MCMCtree* analyses and visualize posterior ages on trees. *Bioinformatics*, *35*(24), 5321–5322. https://doi.org/10.1093/bioinformatics/btz554

R Core Development Team. (2024). *R: A language and environment for statistical computing*.

Richter, D. J., Fozouni, P., Eisen, M. B., & King, N. (2018). Gene family innovation, conservation and loss on the animal stem lineage. *eLife*, *7*, e34226. https://doi.org/10.7554/eLife.34226

Rivera, A. M., & Swanson, W. J. (2022). The Importance of Gene Duplication and Domain Repeat Expansion for the Function and Evolution of Fertilization Proteins. *Frontiers in Cell and Developmental Biology*, *10*, 827454. https://doi.org/10.3389/fcell.2022.827454

Saad, R., Cohanim, A. B., Kosloff, M., & Privman, E. (2018). Neofunctionalization in Ligand Binding Sites of Ant Olfactory Receptors. *Genome Biology and Evolution*, *10*(9), 2490–2500. https://doi.org/10.1093/gbe/evy131

Sánchez-Gracia, A., Vieira, F. G., & Rozas, J. (2009). Molecular evolution of the major chemosensory gene families in insects. *Heredity*, *103*(3), 208–216. https://doi.org/10.1038/hdy.2009.55

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Schurko, A. M., Logsdon, J. M., & Eads, B. D. (2009). Meiosis genes in *Daphnia pulex* and the role of parthenogenesis in genome evolution. *BMC Evolutionary Biology*, *9*(1), 78. https://doi.org/10.1186/1471-2148-9-78

Schwartz, S. S. (1984). Life History Strategies in *Daphnia*: A Review and Predictions. *Oikos*, *42*(1), 114. https://doi.org/10.2307/3544616

Shen, W., Le, S., Li, Y., & Hu, F. (2016). *SeqKit*: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Molecular Biology and Evolution*, *32*(5), 1342–1353. https://doi.org/10.1093/molbev/msv022

Smith, V. H., & Schindler, D. W. (2009). Eutrophication science: Where do we go from here? *Trends in Ecology & Evolution*, *24*(4), 201–207. https://doi.org/10.1016/j.tree.2008.11.009

Spielman, S. J., Weaver, S., Shank, S. D., Magalis, B. R., Li, M., & Kosakovsky Pond, S. L. (2019). Evolution of Viral Genomes: Interplay Between Selection, Recombination, and Other Forces. In M. Anisimova (Ed.), *Evolutionary Genomics* (Vol. 1910, pp. 427–468). Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_14

Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X.-X., & Rokas, A. (2020). *ClipKIT*: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLOS Biology*, *18*(12), e3001007. https://doi.org/10.1371/journal.pbio.3001007

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). *REVIGO* Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE*, *6*(7), e21800. https://doi.org/10.1371/journal.pone.0021800

Suyama, M., Torrents, D., & Bork, P. (2006). *PAL2NAL*: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server), W609–W612. https://doi.org/10.1093/nar/gkl315

Thomas Lin Pedersen. (2022). *patchwork*: The Composer of Plots. *Https://Patchwork.Data-Imaginist.Com, Https://Github.Com/Thomasp85/Patchwork*.

Torgerson, D. G., Kulathinal, R. J., & Singh, R. S. (2002). Mammalian Sperm Proteins Are Rapidly Evolving: Evidence of Positive Selection in Functionally Diverse Genes. *Molecular Biology and Evolution*, *19*(11), 1973–1980. https://doi.org/10.1093/oxfordjournals.molbev.a004021

Vergilino, R., Markova, S., Ventura, M., Manca, M., & Dufresne, F. (2011). Reticulate evolution of the *Daphnia pulex* complex as revealed by nuclear markers. *Molecular Ecology*, *20*(6), 1191–1207. https://doi.org/10.1111/j.1365-294X.2011.05004.x

Vicens, A., Lüke, L., & Roldan, E. R. S. (2014). Proteins Involved in Motility and Sperm-Egg Interaction Evolve More Rapidly in Mouse Spermatozoa. *PLoS ONE*, *9*(3), e91302. https://doi.org/10.1371/journal.pone.0091302

Villanueva, R. A. M., & Chen, Z. J. (2019). *ggplot2*: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 160–167. https://doi.org/10.1080/15366367.2019.1565254

Wang, S., Zhang, Y., Yang, W., Shen, Y., Lin, Z., Zhang, S., & Song, G. (2023). Duplicate genes as sources for rapid adaptive evolution of sperm under environmental pollution in tree sparrow. *Molecular Ecology*, *32*(7), 1673–1684. https://doi.org/10.1111/mec.16833

Wang, W., Zhang, X.-S., Wang, Z.-N., & Zhang, D.-X. (2023). Evolution and phylogenetic diversity of the aquaporin gene family in arachnids. *International Journal of Biological Macromolecules*, *240*, 124480. https://doi.org/10.1016/j.ijbiomac.2023.124480

Weaver, S., Shank, S. D., Spielman, S. J., Li, M., Muse, S. V., & Kosakovsky Pond, S. L. (2018). *Datamonkey 2.0*: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Molecular Biology and Evolution*, *35*(3), 773–777. https://doi.org/10.1093/molbev/msx335

Wersebe, M. J., Sherman, R. E., Jeyasingh, P. D., & Weider, L. J. (2023). The roles of recombination and selection in shaping genomic divergence in an incipient ecological species complex. *Molecular Ecology*, *32*(6), 1478–1496. https://doi.org/10.1111/mec.16383

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the

Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Winder, M., Spaak, P., & Mooij, W. M. (2004). Trade-offs in *Daphnia* habitat selection. *Ecology*, *85*(7), 2027–2036. https://doi.org/10.1890/03-3108

Wong, A., Turchin, M. C., Wolfner, M. F., & Aquadro, C. F. (2008). Evidence for Positive Selection on *Drosophila melanogaster* Seminal Fluid Protease Homologs. *Molecular Biology and Evolution*, *25*(3), 497–506. https://doi.org/10.1093/molbev/msm270

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). *clusterProfiler 4.0*: A universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Wuerz, M., Huebner, E., & Huebner, J. (2017). The morphology of the male reproductive system, spermatogenesis and the spermatozoon of *Daphnia magna* (Crustacea: Branchiopoda). *Journal of Morphology*, *278*(11), 1536–1550. https://doi.org/10.1002/jmor.20729

Xu, S., Innes, D. J., Lynch, M., & Cristescu, M. E. (2013). The role of hybridization in the origin and spread of asexuality in *Daphnia*. *Molecular Ecology*, *22*(17), 4549–4561. https://doi.org/10.1111/mec.12407

Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., Dai, Z., Lam, T. T., Guan, Y., & Yu, G. (2022). *Ggtree*: A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta*, *1*(4). https://doi.org/10.1002/imt2.56

Xu, S., Spitze, K., Ackerman, M. S., Ye, Z., Bright, L., Keith, N., Jackson, C. E., Shaw, J. R., & Lynch, M. (2015). Hybridization and the Origin of Contagious Asexuality in *Daphnia pulex*. *Molecular Biology and Evolution*, msv190. https://doi.org/10.1093/molbev/msv190

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

Ye, Z., Molinier, C., Zhao, C., Haag, C. R., & Lynch, M. (2019). Genetic control of male production in *Daphnia pulex*. *Proceedings of the National Academy of Sciences*, *116*(31), 15602–15609. https://doi.org/10.1073/pnas.1903553116

Ye, Z., Xu, S., Spitze, K., Asselman, J., Jiang, X., Ackerman, M. S., Lopez, J., Harker, B., Raborn, R. T., Thomas, W. K., Ramsdell, J., Pfrender, M. E., & Lynch, M. (2017). A New Reference Genome Assembly for the Microcrustacean *Daphnia pulex*. *G3:Genes|Genomes|Genetics*, *7*(5), 1405–1416. https://doi.org/10.1534/g3.116.038638

Zeis, B., Lamkemeyer, T., Paul, R. J., Nunes, F., Schwerin, S., Koch, M., Schütz, W., Madlung, J., Fladerer, C., & Pirow, R. (2009). Acclimatory responses of the *Daphnia pulex* proteome to environmental changes. I. Chronic exposure to hypoxia affects the oxygen transport system and carbohydrate metabolism. *BMC Physiology*, *9*(1), 7. https://doi.org/10.1186/1472-6793-9-7

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P. W. H., Paps, J., Zhu, Y., Wu, F., Chen, Y., Wang, J., … Wang, J. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, *490*(7418), 49–54. https://doi.org/10.1038/nature11413

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Greenwold, M. J., Meredith, R. W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., … Froman, D. P. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*(6215), 1311–1320. https://doi.org/10.1126/science.1251385

Zhang, X., Blair, D., Wolinska, J., Ma, X., Yang, W., Hu, W., & Yin, M. (2021). Genomic regions associated with adaptation to predation in *Daphnia* often include members of expanded gene families. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1955), 20210803. https://doi.org/10.1098/rspb.2021.0803

Chapter 2

Balancing selection and the functional effects of shared polymorphism in cryptic *Daphnia* species

Connor S. Murray[1,*], Madison Karram[1], David J. Bass[1], Madison Doceti[1], Dörthe Becker[1,2], Joaquin C. B. Nunez[1,‡], Aakrosh Ratan[3,4], Alan O. Bergland[1,*]

[1] Department of Biology, University of Virginia, Charlottesville, VA, USA

[2] School of Biosciences, Ecology and Evolutionary Biology, University of Sheffield, Sheffield, UK

[3] Center of Public Health Genomics, University of Virginia, Charlottesville, VA, USA

[4] Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA

*Corresponding authors emails: C.S.M: csm6hg@virginia.edu, A.O.B: aob2x@virginia.edu

‡Current address: Department of Biology, University of Vermont, 109 Carrigan Drive, Burlington, VT 05405, USA

Running title: Shared mutations across *Daphnia*.

ORCIDs:

Connor S. Murray: 0000-0002-8302-6585

Madison Karram: 0009-0001-4331-8292

David J. Bass: 0000-0001-7414-2747

Madison Doceti: 0009-0007-9801-6121

Dörthe Becker: 0000-0002-4167-5990

Joaquin C. B. Nunez: 0000-0002-3171-8918

Aakrosh Ratan: 0000-0002-0782-3056

Alan O. Bergland: 0000-0001-7145-7575

Conflict of interest. The authors declare no conflicts of interest.

Keywords: Shared polymorphism, *Daphnia*, balancing selection, incomplete lineage sorting, hybridization, convergent evolution, overdominance, opsins

Abstract

The patterns of genetic variation within and between related taxa represent the genetic history of a species. Shared polymorphisms, loci with identical alleles across species, are of unique interest as they may represent cases of ancient selection maintaining functional variation post-speciation. In this study, we investigate the abundance of shared polymorphism in the *Daphnia pulex* species complex. We test whether shared mutations are consistent with the action of balancing selection or alternative hypotheses such as hybridization, incomplete lineage sorting, or convergent evolution. We analyzed over 2,000 genomes from North American and European *D. pulex* and several outgroup species to examine the prevalence and distribution of shared alleles between the focal species pair, North American and European *D. pulex*. We show that while North American and European *D. pulex* diverged over ten million years ago, they retained tens of thousands of shared alleles. We found that the number of shared polymorphisms between North American and European *D. pulex* cannot be explained by hybridization or incomplete lineage sorting alone. Instead, we show that most shared polymorphisms could be the product of convergent evolution, that a limited number appear to be old trans-specific polymorphisms, and that balancing selection is affecting young and ancient mutations alike. Finally, we provide evidence that a blue wavelength opsin gene with trans-specific polymorphisms has functional effects on behavior and fitness in the wild. Ultimately, our findings provide insights into the genetic basis of adaptation and the maintenance of genetic diversity between species.

Introduction

Genetic diversity reflects a species' history and serves as the foundation for adaptation to ecological change. In nature, mutations arise, and their persistence time is a function of their selective value and the effective population size of the focal species (Crow & Kimura, 1970). One distinct type of genetic variant is a shared polymorphism, in which mutations are identical by state across closely related species (Wang & Mitchell-Olds, 2017). The abundance and frequency of shared polymorphisms between two species can provide insight into some of the most interesting processes in evolution. Shared polymorphism that arose prior to the split of two species are generally referred to as trans-species polymorphism (Hedrick, 2013; Wiuf et al., 2004; Wu et al., 2017). Trans-specific polymorphisms can be used to study the speciation process (Klein et al., 1998), helping refine estimates of the timing and population sizes at divergence (Edwards et al., 2000). Unless divergence happened recently or there is ongoing gene flow, it is unlikely that neutral polymorphism will be retained in both species for long (Leffler et al., 2013). Therefore, the presence of shared polymorphism between two species with limited gene-flow can be a powerful way to identify balanced polymorphisms (Clark, 1997). These polymorphisms are presumed to be maintained by temporal or spatial variation in the direction of natural selection (Bergland et al., 2014; Ségurel et al., 2012, 2013), or by genetic overdominance (Wills, 1975). Shared polymorphisms can also indicate convergent adaptive evolution (Castoe et al., 2009) or adaptive introgression (Hedrick, 2013), and these polymorphisms themselves can also be the target of balancing selection (Wang & Mitchell-Olds, 2017).

Identifying the forces that generate and maintain shared polymorphism is therefore an important problem in evolutionary genetics. However, identifying the contribution of neutral, demographic, and adaptive evolutionary processes to the generation and maintenance of shared polymorphisms is challenging. This is especially so for young species because there has not been enough time for species-specific alleles to fix, and for drift to erode shared

polymorphism present due to incomplete lineage sorting. In contrast, testing alternative hypotheses for the generation of shared polymorphism can sometimes be more tractable for slightly older species pairs. This is because neutral trans-species polymorphisms are expected to be rare thereby eliminating incomplete lineage sorting as a main driver of shared polymorphism. If sufficient time has occurred for fixation of species-specific alleles, then adaptive introgression will be relatively easy to identify (Huerta-Sánchez et al., 2014) especially if it occurred recently. However, if only a single trans-specific polymorphism is functional, recombination will erode the ancestral haplotypes (Gao et al., 2015) and gene trees will align with species trees causing ambiguity when differentiating between convergent evolution and trans-specificity (Unckless et al., 2016). In contrast, long-term balancing selection of a trans-specific polymorphism can be relatively unambiguous if there are multiple sites at a locus that are shared polymorphisms, and these are tightly linked causing allele trees to not align with species trees (Wang et al., 2020). The presence of trans-specific haplotypes suggests that multiple functional sites at the locus are the target of some form of balancing selection (Charlesworth, 2006).

Daphnia species are an excellent model to study the mechanisms that generate and maintain shared polymorphism. *Daphnia* are freshwater microcrustaceans that have been the focus of ecological and evolutionary research for over a century (Ebert, 2022). Among the most widely studied taxa within this genus are *D. magna* (Decaestecker et al., 2007), *D. obtusa* (Spitze, 1993), as well as *D. pulex* (Lynch et al., 2017) and its close relatives (Colbourne et al., 2011). The *D. pulex* species group is currently in the process of an adaptive radiation (Fryer, 1991). Owing to their recent divergence time, some members of the North American *D. pulex* species group are known to hybridize in the wild (Held et al., 2016), resulting in contagious obligate asexuality (Xu et al., 2015). Members of the *D. pulex* species group, including *D. obtusa*, are found across the Palearctic and Nearctic (Crease et al., 2012), and recently established populations can be found in other regions of the world (So et al., 2015). Although *D.*

obtusa, *D. pulicaria*, and *D. pulex* have been identified on multiple continents, each of these three taxa represent polyphyletic groups (Černý & Hebert, 1999). For instance, based on mitochondrial sequence, *D. pulex* found in North America is more closely related to North American *D. pulicaria* than it is to European *D. pulex* (Crease et al., 2012). The confusion of species identification and naming in this genus is due to the similar morphology (Dodson, 1981) and ecological niches of these taxa (Chin & Cristescu, 2021) plus their capacity to interbreed (Pantel et al., 2011), generally reflecting the taxonomic ambiguities within the species group (Hebert & Wilson, 1994), and among zooplankton in general (Brooks, 1957).

The evolutionary and ecological history of the *D. pulex* species group affords us an ideal opportunity to study the evolutionary forces that have shaped patterns of shared polymorphism. Here, we assessed alternative evolutionary mechanisms that can generate and maintain shared polymorphisms between North American and European *D. pulex*. Using population genomic data from samples in North America and Europe, we first confirm that North American and European *D. pulex* are distinct species that have diverged millions of years ago. Next, we show that North American and European *D. pulex* possess tens of thousands of shared polymorphisms, whose abundance cannot be explained by incomplete lineage sorting, hybridization, introgression, or gene-flow. Therefore, we conclude that many of these shared polymorphisms arose either via convergent evolution or have been maintained since the split between these taxa. For many shared polymorphisms, we cannot differentiate which of these two mechanisms is most likely. However, a limited number of genes show a strong excess of shared polymorphisms that are in linkage disequilibrium, consistent with long-term balancing selection operating on a haplotype. One of these genes is a single-copy blue wavelength opsin, part of a gene family that has previously been identified as a target of rapid adaptive evolution in *Daphnia* (Brandon et al., 2017; Ye et al., 2023). We show that European *D. pulex* clones harboring alternate genotypes for this blue wavelength opsin have differences in movement and activity that is dependent on light conditions and provide evidence for overdominance in the

wild. Taken together, our results highlight the abundance, selective history, and function of shared polymorphisms in *Daphnia* and contributes to the understanding of the phylogeography for this classic model system.

Materials and Methods

Sampling and sequencing European Daphnia genomes. *Daphnia* were sampled from 16 ponds throughout England in 2018. Samples were transported to the University of Virginia and clonally derived isofemale lines were established. Samples were identified as either *D. pulex*, *D. pulicaria*, or *D. obtusa* based on morphological characteristics using an online dichotomous key (http://cfb.unh.edu/cfbkey/html/anatomy/daphnia/daphnia.html). DNA extraction and library preparation followed methods outlined in Kubow *et al.* (2022). Briefly, for each isofemale line multiple individuals were exposed to antibiotics (streptomycin, tetracycline, and ampicillin, 50 mg/L of each) and fed Sephadex G-25 beads to clear their gut of algae. Samples were homogenized using metal beads and a bead beater and DNA was extracted using the Agencourt DNAdvance kit (Beckman-Coulter). RNA was removed using RNase followed by an additional bead cleanup. DNA was quantified using the broad-range Quant-iT dsDNA kit (ThermoFisher Scientific) and normalized to 1 or 2 ng/μL before library construction. Full genome libraries were constructed using a scaled down Nextera protocol (Baym et al., 2015). Libraries were size selected for fragments ranging from 450 to 550 bp using a Blue Pippin and quality checked using a BioAnalyzer. Samples were sequenced on a HiSeq X platform, paired-end 150bp.

Publicly available Daphnia genomes. Genome sequences of North American and European *D. pulex*, *D. pulicaria*, and *D. obtusa* were obtained from NCBI's Sequence Read Archive (SRA; Leinonen et al., 2011). We incorporated wild-sequenced or isogenic female lineages (Barnard-Kubow et al., 2022; Lynch et al., 2017; Xu et al., 2015; Ye et al., 2022), and excluded samples

that were from mutation accumulation studies. Species identity for these samples was based on annotations provided in each SRA record.

Short-read mapping. Prior to short-read mapping of all samples, sequencing adaptors were removed using *trimmomatic v0.39* (Bolger et al., 2014), and overlapping reads were merged using *pear v0.9.11* (Zhang et al., 2014). All samples were mapped to the European *D. pulex* genome (Barnard-Kubow et al., 2022) using *bwa mem v0.7.17* (H. Li, 2013), and downstream data manipulation was performed using *samtools merge v1.12* (H. Li et al., 2009). Duplicate reads for every bam file were marked and removed using *picard v2.23.4* (https://broadinstitute.github.io/picard/). Quality control metrics were assembled using *fastqc v0.11.5* (Andrews, 2010) and *MultiQC v1.11* (Ewels et al., 2016). S*amtools flagstat* counted the mapped and properly paired reads.

Additionally, we mapped North American *D. pulex* to the North American *D. pulex* reference genome (KAP4; GenBank assembly: GCF_021134715.1) using the same mapping strategy outlined above. We created a liftOver file to translate features in KAP4 to D84A. The chains exhibited good coverage, allowing us to translate 72.6% of the KAP4 genome to D84A (Lee et al., 2022). We created the liftOver file by running pairwise alignments using *lastz* followed by the use of various UCSC tools to chain the alignments, sort them, filter them, and convert them into UCSC nets and chains (Harris, 2007). We used *LiftOverVCF* from *picard* to convert the KAP4 aligned VCF to the D84A genome coordinates. We then assessed the concordance of the SNP classifications between North American samples mapped to D84A and the liftOver VCF (Supplemental Figure 1C-D).

All analyses using North American and European *D. pulex, D. pulicaria,* and *D. obtusa* were conducted using samples mapped to the European *D. pulex* reference genome. We assessed reference allele bias for these interspecific mappings by calculating the proportion of alternative and reference allele dosage for 1,000 biallelic heterozygous BUSCO gene SNPs

across the genome (with 100 bootstrap resampling). All analyses focusing exclusively on shared polymorphisms between European and North American *D. pulex* were conducted using the intersection of SNPs identified by mapping North American *D. pulex* and European *D. pulex* to their respective reference genomes.

SNP calling and filtering. We used *HaplotypeCaller* and *GenotypeGVCFs* from *gatk v4.1.6.0* to create VCF files (Poplin et al., 2017). *VariantFiltration* in *gatk* removed low-quality SNPs recommended for organisms without reference panels: ("QD<2.0", "QUAL<30.0", "SOR>3.0", "FS>60.0", "MQ<40.0", "MQRankSum<-12.5", "ReadPosRankSum<-8.0"). We removed sites flanking ±10bp any indels using *bcftools filter --SnpGap 10* (Li et al., 2009) and removed indels using *SelectVariants* in *gatk*. We annotated SNPs using *snpEff v4.3t* (Cingolani et al., 2012).

Samples with average genome-wide missingness >10% were removed from analyses as was any genomic region with more than 10% missingness across the remaining samples. We removed regions with high (DP≥35) and low mean site read depth (DP≤8), along with chromosomal endpoints, regions of the reference genome with large stretches of gaps, and regions of Ns as described in Barnard-Kubow et al., 2022. Repetitive elements identified in the European *D. pulex* genomes were classified with *RepeatMasker v4.0.8* and were removed (Tarailo-Graovac & Chen, 2009). We restricted analyses to the genic and non-genic regions associated with the 6,544 single-copy ortholog genes between European and North American *D. pulex* from *OrthoFinder v5* (Emms & Kelly, 2019) for the SNPs that were retained from the liftOver. Most analyses removed SNPs that have a minor allele frequency (MAF) less than 0.01 within-species. After filtering, 347,200 SNPs represent the whole-genome SNP set. We restricted phylogenetic analyses to BUSCO genes identified with *Panther* annotations (Mi et al., 2013; Seppey et al., 2019; Simão et al., 2015). The BUSCO gene SNP set includes 138,024 SNPs. Principal component analysis (PCA) of SNPs was conducted in *SNPRelate v1.24.0* while

excluding sites with MAF<0.01 (Zheng et al., 2012). D_{xy} was calculated using *PopGenome v2.7.5* (Revell, 2019).

Assigning multi-locus genotypes. Every sample was assigned to a multi-locus genotype (MLG) using the *poppr v2.9.3* package (Kamvar et al., 2014) in *R v4.0.3*. Unless otherwise noted, every analysis was subset based on picking a representative sample with the highest coverage for each MLG (Supplemental Table 1).

Mitochondrial tree. We annotated the D84A mitochondrion using *MITOs v1* (Bernt et al., 2013). We aligned and called SNPs using *bcftools mpileup v1.9* and *bcftools call*. We excluded reads that had low-quality scores (Q<20) and high depth (DP>100) using *bcftools filter*. And generated consensus FASTA files using *bcftools consensus*. We mapped North American *D. pulex* and *D. pulicaria* to the North American *D. pulex* mitochondrial genome sequence (GenBank accession: NC_000844.1) and mapped both North American and European *D. obtusa* samples to the North American *D. obtusa* mitochondrial genome sequence (GenBank accession: CM028013.1). The mitochondrial sequence of European *D. magna* was used as an outgroup (GenBank accession: NC_026914.1). We assembled homology blocks using *exonerate v2.4.0* (Slater & Birney, 2005) for the 13 protein-coding genes and found high sequence similarity (>80%), except for *atp8*. Therefore, we assembled trees excluding *atp8*. We then used *mafft v7.475* (Katoh & Standley, 2013) to assemble multiple sequence alignments (MSA). We concatenated these MSAs for each gene using *seqkit concat v2.2.0* (Shen et al., 2016) and ran *iqtree2 v2.1.2* with 1,000 bootstraps (Supplemental Figure 2; Hoang et al., 2018; Kalyaanamoorthy et al., 2017).

Estimating divergence-time. We used *Snapp v1.6.1* within *Beast2 v2.6.6* to estimate the split-time within the species complex (Bouckaert et al., 2014). We used two representative

individuals with the highest coverage for each species. We used 3,000 randomly sampled BUSCO gene SNPs, 1 million iterations, and a 10% burn-in. The output tree was time-constrained for the outgroup species, European *D. obtusa,* to 31 million years ago (MYA with a confidence interval of 1 MYA based upon a genus-wide tree; Chin & Cristescu, 2021; Cornetti et al., 2019). We used *Tracer v1.7.1* to quantify MCMC convergence (Rambaut et al., 2018).

Hybridization statistics. We used *ADMIXTURE v1.3.0* (Alexander & Lange, 2011), excluding any sites with MAF<0.01 and thinned every 500 SNPs. We varied the number of clusters (*k*) from 2-25 and calculated the cross-validation error (CV) at every *k* model. We chose *k*=9 because the minimum CV score was reached. We quantified the magnitude of introgression using *Dsuite v0.5* (Malinsky et al., 2021) with European *D. obtusa* as the outgroup.

Historic N_e and demographic inference of migration. To calculate historical N_e for European and North American *D. pulex,* we ran *MSMC2 v2.1.1* and *SMC++ v1.15.4* (Schiffels & Wang, 2020; Terhorst et al., 2017). We performed demographic inference with *moments v1.1.0* in *python3* (Jouganous et al., 2017). We tested two models: one with-migration and one without-migration. For the former model, we used *moments*' *split_mig* model. For the latter, we used *split_mig* with no migration. We ran inference on 20x20 SFS projections until model convergence and classified a shared polymorphism as an allele whose allele frequency is above 1/20 in both species. We used a mutation rate μ=5.69x10^{-9} (Lynch et al., 2017). We followed the methods of McCoy et al., (2014) to convert coalescent units of into standard units. We estimated the ancestral population size as N_{ANC}=200,000/η_{EU}, where η_{EU}=N_e European *D. pulex* is an approximate from historic demographic inference (Supplemental Figure 3 left panel). Then, with *moments*' estimates η_{NA} for N_{NA}, τ for t_{split}, and *M* for migration, we calculated: N_{NA}=N_{ANC}×η_{NA}, N_{EU}=N_{ANC}×η_{EU}, t_{split}=2N_{ANC}×τ, and m=M÷2N_{ANC}. We chose to associate European

D. pulex with the ancestral species because the reference genome isolate is a European *D. pulex* clone.

Classifying shared polymorphisms between North American and European D. pulex. We classified each mutation as a fixed difference between species, polymorphic within-species, or a shared polymorphism between species. We classified sites as polymorphic (within species or shared) if the minor allele frequency in either or both species was greater than 0.01 (Supplemental Table 2).

We tested whether the extent of shared polymorphism can be explained by incomplete lineage sorting using methods outlined elsewhere (Novikova et al., 2016; Wiuf et al., 2004). The formula in Novikova et al., (2016) estimates the number of expected shared polymorphisms between species, where $d_{between}$ is D_{xy} between North American and European *D. pulex*, and d_{NAm} & d_{Euro} are within-species polymorphism.

$$\exp\left(-2d_{\text{between}} + \frac{(d_{\text{NAm}} + d_{\text{Euro}})\max\{d_{\text{NAm}},\, d_{\text{Euro}}\}}{d_{\text{NAm}}\, d_{\text{Euro}}} \right)$$

Balancing selection statistics. *BetaScan v1* was used to calculate β^{1} statistic within species using the folded site frequency spectrum (Siewert & Voight, 2017). The α_b statistic tested for the proportion of sites under balancing selection between species-pairs from Soni et al., 2022. Where *Poly.* are SNPs within-species and *SP* are shared polymorphisms between-species. *SYN* are synonymous sites and *NS* are non-synonymous sites:

$$\alpha_b = 1 - \frac{Poly.\ Syn\ x\ SP\ NS}{Poly.NS\ x\ SP\ SYN}$$

Phylogenetic tree test using pairwise cophenetic distances. We tested the local sequence genealogy to test for trans-specificity versus convergent evolution (Koenig et al., 2019; Nunez et

al., 2021). This test used trees built from 500bps flanking high-frequency non-synonymous shared polymorphisms (MAF>0.1). We calculated the median pairwise cophenetic distances (CPD) between samples (Cardona et al., 2013). We extracted haplotypes from a *WhatsHap v1.1* phased VCF (Martin et al., 2023) from 30 high-read depth individuals for each species. We chose 30 samples to keep the sample size consistent across species while decreasing model convergence time. We aligned the parental haplotypes (n=60 per species) using *mafft* and built trees using *iqtree2* (1,000 bootstraps). The null hypothesis was that the tree would be concordant with the species-tree topology. The alternative was that the tree would be discordant with the species-tree and that median CPD between North American and European *D. pulex* is higher within-species than between-species (i.e., $CPD_{Within} > CPD_{Between}$). $CPD_{Within-Between} = CPD_{Within} - CPD_{Between}$, where CPD_{Within}=within-species, $CPD_{Between}$=between-species. Positive $CPD_{Within-Between}$ indicates an allele-specific topology and negative $CPD_{Within-Between}$ indicates a concordant topology with the species-tree. A cartoon depicting these hypotheses is in Figure 4A.

Light exposure experiments on Daphnia activity. We developed a behavioral assay to collect activity data on 12 distinct European *D. pulex* clones using a DAM Trikinetics monitor (Chiu et al., 2010). In total, we measured activity for 216 individuals. The Trikinetics monitor has 32 wells filled with 5mm diameter plastic tubes. Each well has an infrared light beam that when broken by a *Daphnia* individual will count as an activity event. We exposed individuals to white light, blue light, and dark lighting conditions using blackout boxes mounted with LEDs (described in Erickson et al., 2020). Individual *Daphnia* were placed inside a plastic tube with artificial pond water media (ASTM; Standard, 2007) while each Trikinetics monitor collected activity measurements over a twelve-hour experimental period, sampling every 5 seconds. We excluded measurements during the first hour to allow individuals to settle in. For 95% of the 5-second intervals, 0 or 1 beam break was recorded and 99.9% of intervals had 4 or fewer beam

breaks. Therefore, for each 5-second interval, we converted the number of beam breaks recorded into a binary variable (>=1 beam-break vs 0 beam-breaks) and calculated total activity as the fraction of 5-second intervals with more than one beam break per individual over the course of the experiment. We modeled total activity with a generalized linear mixed effect model using *lme4 v1.1-27.1* in *R* (Bates et al., 2015) and performed likelihood ratio tests between the following models:

$$\text{Model 1: } y \sim \text{Light} + \text{Clone} + \text{Block} + \varepsilon$$

$$\text{Model 2: } y \sim \text{Light} + \text{Genotype} + \text{Clone} + \text{Block} + \varepsilon$$

$$\text{Model 3: } y \sim \text{Light} + \text{Genotype} + \text{Light:Genotype} + \text{Clone} + \text{Block} + \varepsilon$$

Where *y* is the fraction of intervals with activity, *Light* is the fixed effect of light treatment (white, blue, dark), *Genotype* is the fixed effect of genotype at the blue wavelength opsin (BLOP) locus, *Light:Genotype* is the fixed interaction effect, (1|*Clone*) is the random effect of clone, (1|*Block*) is the random effect of one of the three experimental blocks run over successive weeks, and ε is the binomially distributed error with weights equal to the number of 5-second intervals (ca. 7800). We conducted likelihood ratio tests between Model 1, Model 2, and Model 3 using the *anova()* function in *R* (Supplemental Table 3). In addition, we performed an additional analysis that explicitly models elapsed time in the experiment as a fixed effect and includes the individual *Daphnia* identifier as a random effect to account for repeated measures. The results of that analysis are in line with the more straightforward model presented here and we show those results in Supplemental Table 4.

Daphnia11806-RA orthologs. We tested the orthology of *Daphnia11806-RA* by BLASTing the amino acid sequence against the NCBI database using *blastp v2.13.0* (Sayers et al., 2022).

Statistics and visualization. Analyses were performed using *R v3.6.2–4.0.3* (R Core Development Team 2013). We used the following packages for analysis and visualization: *tidyverse v1.3.1* (Wickham et al., 2019), *ggplot2 v3.3.5* (Villanueva & Chen, 2019), *ggtree v2.0.4* (Xu et al., 2022), *ape v5.4-1* (Paradis & Schliep, 2019), *patchwork v1.0.1* (Thomas Lin Pedersen, 2022), *viridis v0.5.1* (Garnier et al., 2021), *data.table v1.12.8* (Dowle & Srinivasan, 2023), *foreach v1.4.7*, *doMC v1.3.5* (Daniel et al., 2022), *SeqArray v1.26.2* (Zheng et al., 2017).

Data availability. The D84A mitochondrion was uploaded to NCBI (JAHCQT000000000) and updated to the existing accession: GCA_023526725.1. The novel 93 genomes described here were uploaded to NCBI under the accession: PRJNA982532. The metadata for samples is located in Supplemental Table 1. The VCF and GDS are deposited on dryad: https://doi.org/10.5061/dryad.dncjsxm3p. Scripts and data are deposited on GitHub: https://github.com/connor122721/SharedPolymorphismsDaphnia.

Results

Thousands of Daphnia genomes. We first assembled short-read genomic data for 2,321 samples of *D. pulex*, *D. pulicaria*, and *D. obtusa* collected from North American and European ponds (Figure 1A). This includes whole genomes published elsewhere (Barnard-Kubow et al., 2022; Lynch et al., 2017; Xu et al., 2015; Ye et al., 2022), along with 93 samples reported here for the first time. We aligned samples to the European *D. pulex* assembly (D84A; Barnard-Kubow et al., 2022) and identified 347,200 SNPs after filtering. In brief, our filtering methods removed regions that could prove problematic for population genomic analyses across related species. The SNP that we used represent within-species SNPs, fixed differences, and shared polymorphisms classified between North American and European *D. pulex*. Because lineages could be clonally derived from a recent common ancestor, each sample was assigned to a

multi-locus genotype using the filtered SNP set (MLG; Supplemental Table 1). In all analyses, unless otherwise noted, we restricted to one sample per MLG (n=1,173).

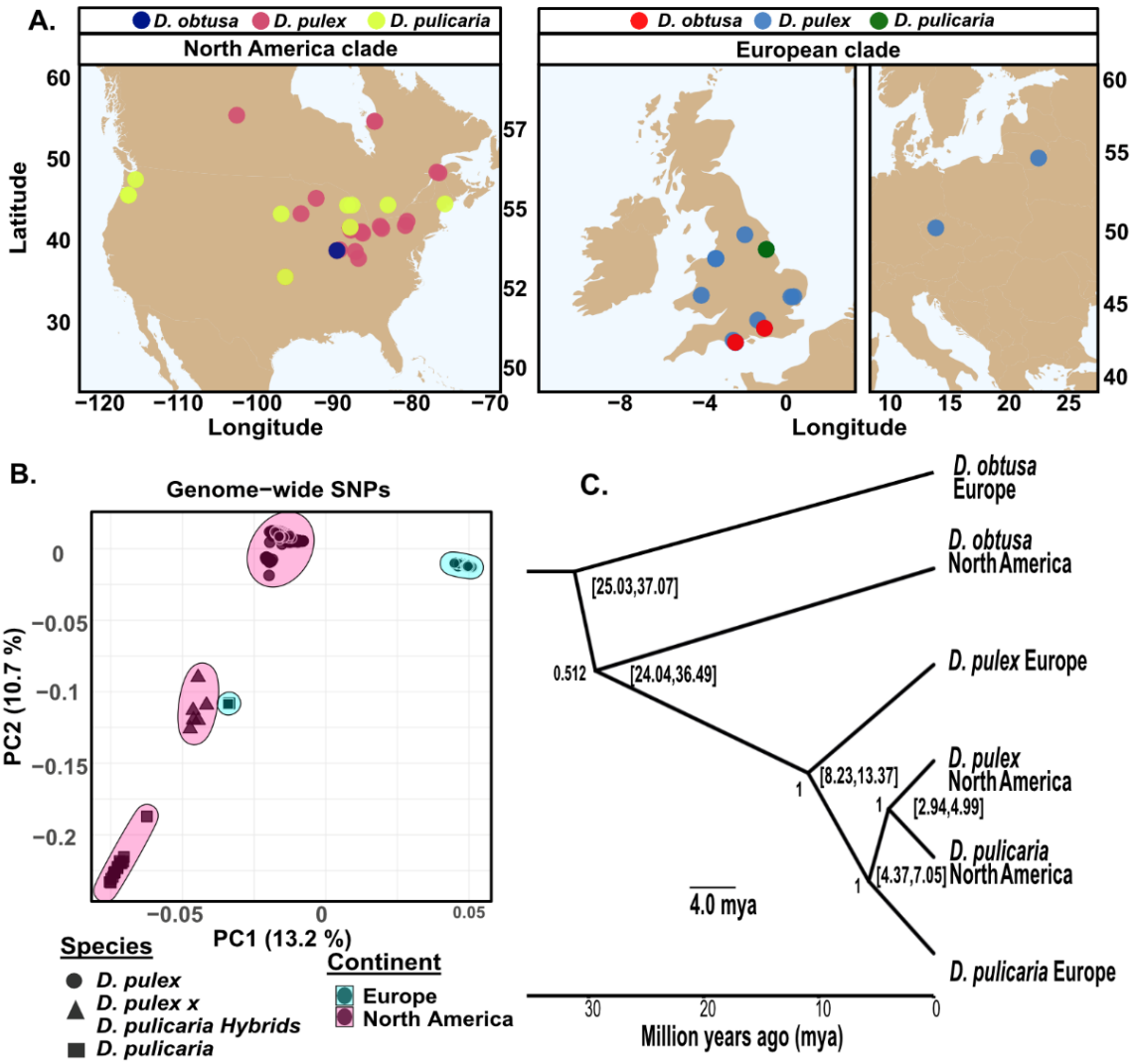


Figure 1. Genetic divergence of the *Daphnia pulex* species complex. A) Sample origin of the North American and European clades, each consisting of *D. pulex*, *D. pulicaria*, and *D. obtusa*. Most of the European clade has samples in the United Kingdom but there is one sample in both the Czech Republic and Lithuania as shown in the rightmost subfigure. **B)** The principal component axes (PC1 and PC2) using filtered genome-wide SNPs (minor allele frequency > 0.01). The proportion of variation explained by each PC is shown in parentheses. We restricted

the principal component analysis to the *D.* pulex and *D. pulicaria* taxa because the *D. obtuse* taxa are so distantly related. **C)** Time-constrained phylogenetic tree restricted to 2 representative individuals within each species based on 3,000 BUSCO gene SNPs. This consensus tree is rooted with European *D. obtusa* to have 31 million years of divergence. Bracketed values are 95% confidence intervals in millions of years ago (mya). Node labels indicate the posterior probabilities estimated from 1 million bootstraps.

Interspecific mapping does not cause systematic biases. A concern for aligning divergent sequences to the same assembly is for reference allele bias to decrease mapping efficiency and cause genotype errors (Günther & Nettelblad, 2019). To assess this, we calculated the proportion of alternative and reference allele dosage for heterozygous BUSCO gene SNPs (N=1,000; 100 bootstraps). On average, SNPs identified in North American or European *D. pulex, D. pulicaria,* or *D. obtusa* had approximately the same alternative and reference allele dosage at heterozygous sites, revealing an absence of systematic reference allele bias (Supplemental Figure 1A). Next, we mapped North American *D. pulex* samples to their species assembly (KAP4) and measured the concordance of SNP classifications between genomes. We show that 88% of SNP classifications are unchanged between assemblies (Supplemental Figure 1B&C). However, this high level of concordance could be an underestimate because of information loss incurred from lifting over assemblies (Chen et al., 2021; Günther & Nettelblad, 2019). Therefore, we conclude that the data is not systematically biased by mapping reads from non-European *D. pulex* to the European *D. pulex* assembly.

Results that highlight genetic divergence, hybridization, and introgression between taxa use the SNP classifications identified by exclusively mapping to the European *D. pulex* reference genome. To be rigorous, all results that focus on shared polymorphisms between North American and European *D. pulex* use sites that were identified as shared polymorphisms

when mapping reads from each species to their respective reference genome, and then lifting over coordinates (N_{SNPs}=28,983; Supplemental Table 2).

Population genetics of the species complex. To understand the extent of divergence between species, we performed principal component analysis (PCA) on the SNP dataset after retaining sites above 0.01 minor-allele frequency (MAF) within-species (Figure 1B). The first and second PC axes are significantly different between the North American and European *D. pulex*, *D. pulicaria,* and hybrids species groups (ANOVA PC1: $F_{4,1154}$ = 70,617, $p < 2 \times 10^{-16}$; ANOVA PC2: $F_{4,1154}$ = 27,940, $p < 2 \times 10^{-16}$). Intriguingly, European *D. pulicaria* clusters near the known hybrids of North American *D. pulicaria* and *D. pulex* (Jackson et al., 2021; Tucker et al., 2013; Supplemental Table 1); below we test whether the samples identified as *D. pulicaria* collected in Europe are related hybrids between North American taxa or are themselves hybrids.

To evaluate the nuclear phylogeny of the *D. pulex* species complex, we built a time-constrained phylogenetic tree using BUSCO gene SNPs. The tree omitted known hybrids of North American *D. pulex* and *D. pulicaria* because they prevented model convergence. Our results show that the nodes that split the *D. pulex* species complex are generally well supported, reflecting a high pairwise sequence divergence (D_{xy}) between taxa. We estimate that the split-time between North American and European *D. pulex* is around 10 million years ago (Figure 1C). The mitochondrial phylogeny also supports a reciprocally monophyletic relationship between North American and European *D. pulex*. However, North American *D. pulex* and *D. pulicaria* are not reciprocally monophyletic (Supplemental Figure 2). The recent split-time between North American *D. pulex* and *D. pulicaria* (Ye et al., 2022), their propensity to hybridize (Pantel et al., 2011), and discordant mitochondrial and nuclear phylogenies support the hypothesis that North American taxa are in the process of incipient speciation (Heier & Dudycha, 2009).

North American and European *D. pulex* possess marked differences in levels of diversity, consistent with long-term divergence. Principal component clusters are more dispersed among North American *D. pulex* than they are among European *D. pulex*, suggesting higher genetic variability within the North American clade (Figure 1B). Second, synonymous site D_{xy} between the two species is large (BUSCO genes D_{xy}=0.054). Third, North American and European *D. pulex* taxa have different historic N_e: the North American *D. pulex* N_e is ~700,000 (95% confidence intervals; 625,090.3, 772,232.5) whereas the European *D. pulex* N_e is ~300,000 (268,445.1, 323,248; Supplemental Figure 3).

***Hybridization in the D. pulex* species group.** Hybridization is common among the group of North American *D. pulicaria* and *D. pulex* species (Pantel et al., 2011), however, signals of hybridization between North American and European *Daphnia* remain less well understood. European *D. pulicaria* and North American *D. pulex-pulicaria* hybrids both exhibit strong signals of hybridization (Figure 2AB; *D*=0.49, *f4-ratio*=0.236, p=2.3x10^{-16} for European *D. pulicaria;* *D*=0.55, *f4-ratio*=0.48, p=2.3x10^{-16} for North American *D. pulex-pulicaria*). However, hybridization between European *D. pulicaria* and North American or closely related circumarctic species is not recent or is with other members of the complex North American *D. pulex-pulicaria* species sub-group. For example, an *ADMIXTURE* analysis reveals that European *D. pulicaria* has distinct ancestry clusters from other species, while the recent hybrids of North American *D. pulex-pulicaria* display split ancestry between North American *D. pulex* and *D. pulicaria* (Figure 2D; Alexander & Lange, 2011). We also examined heterozygosity at fixed differences between North American *D. pulex* and North American *D.* pulicaria in European *D. pulicaria* and North American *D. pulex-pulicaria* hybrids. These fixed differences are heterozygotes 70% of the time in North American hybrids, but only 2% of the time in European *D. pulicaria* suggesting a distinct evolutionary history of the European *D. pulicaria* clade. In summary, our findings imply that European *D. pulicaria* is likely a member of the speciose North American *Daphnia pulex* species

sub-group, consistent with previous reports of a circumarctic *D. pulex* lineage predominant across Northern Eurasia (Colbourne et al., 1998).

However, signals of hybridization are weak between North American and European *D. pulex* (*D*=0.02, *f4-ratio*=0.0077, *p*=1.07x10^{-9}; Figure 2C). *ADMIXTURE* analysis suggests that European *D. pulex* forms several distinct ancestry groups that do not appear within any species of North American *Daphnia* (Figure 2D). Only ~0.5% of fixed differences between North American *D. pulex* and *D. pulicaria* segregate as heterozygous sites in European *D. pulex* (Figure 2E). These results suggest that European *D. pulex* are distinct from the remaining taxa and do not have a recent history of hybridization with the other species studied.

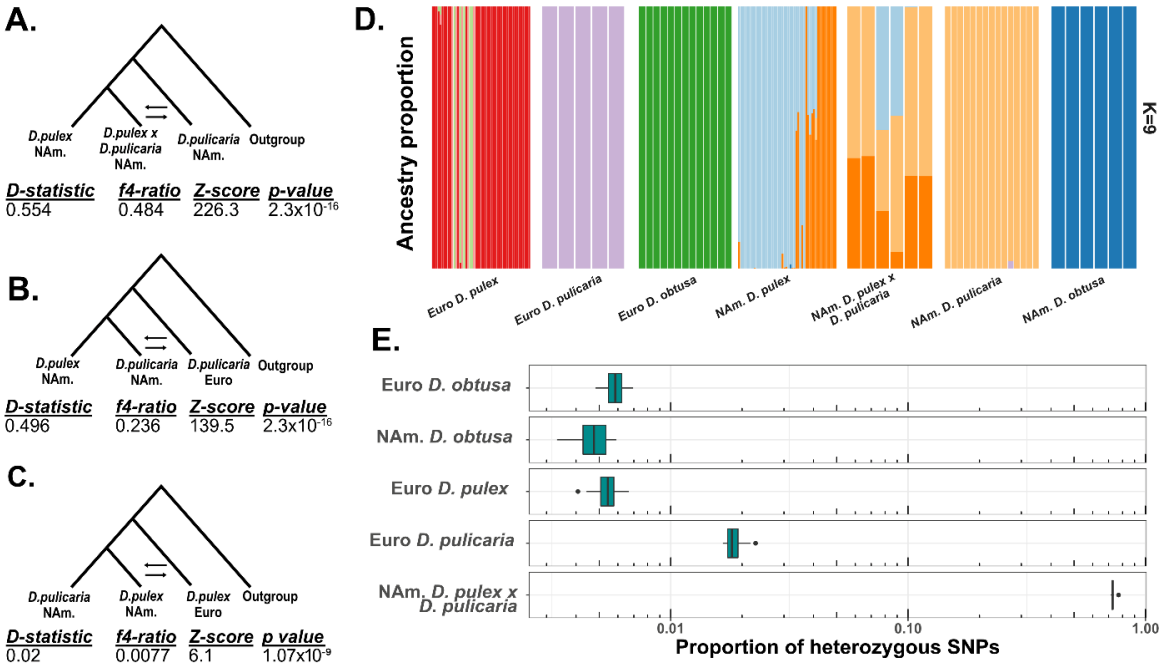


Figure 2. Hybridization across the *D. pulex* species complex. A-C) Introgressions tests using various four-species trees. The outgroup is European *D. obtusa* in all tests. *D-statistic* and *f4-ratio* describe the extent of introgression between the 2nd and 3rd taxa on the tree being tested. **D)** *ADMIXTURE* plot of the *D. pulex* species complex with *k*=9 having the minimal cross-validation error. Each color represents a unique ancestry group for each sample. **E)** We

identified fixed differences between North American *D. pulex* and *D. pulicaria* and calculated the proportion that is heterozygous in a randomly chosen individual from the remaining taxa. The boxplot shows the distribution of these proportions from randomly sampled clones (one per MLG).

Extent of shared polymorphism between North American and European D. pulex is not explained by incomplete lineage sorting or migration. For species with deep split-times and low levels of migration or hybridization, we expect few shared polymorphisms to exist if such polymorphisms are neutral. For instance, based on a simple neutral model with no migration (Novikova et al., 2016; see Material & Methods) we expect to observe 336 shared polymorphisms given the split-time between North American and European *D. pulex* at synonymous sites. Yet, we observe at least 11,000 shared synonymous SNPs between these species (Supplemental Table 2).

This prediction does not account for historic migration, so we preformed demographic inference on the two-dimensional site-frequency spectrum (2D SFS) using *moments* (Figure 3A; Jouganous et al., 2017). First, we contrasted two models, one that allows constant migration (*Split + Migration*) and one where the migration rate was set to zero after population divergence (*Split*). The "*Split + Migration*" model is the best model based on the mean Bayesian information criteria (BIC) across bootstraps ("*Split + Migration*" BIC = 20,637, "*Split*" BIC = 33,216). Notably, the "*Split*" model severely underpredicts the number of shared polymorphisms, reflecting that incomplete lineage sorting alone is insufficient to explain the abundance of shared SNPs. The "*Split + Migration*" model itself underpredicts the number of shared polymorphisms by 25% (Figure 3A), and the model prediction shows a notable deficit of common shared SNPs and an excess of shared SNPs that are at low frequencies (Figure 3B) compared to the empirical SFS.

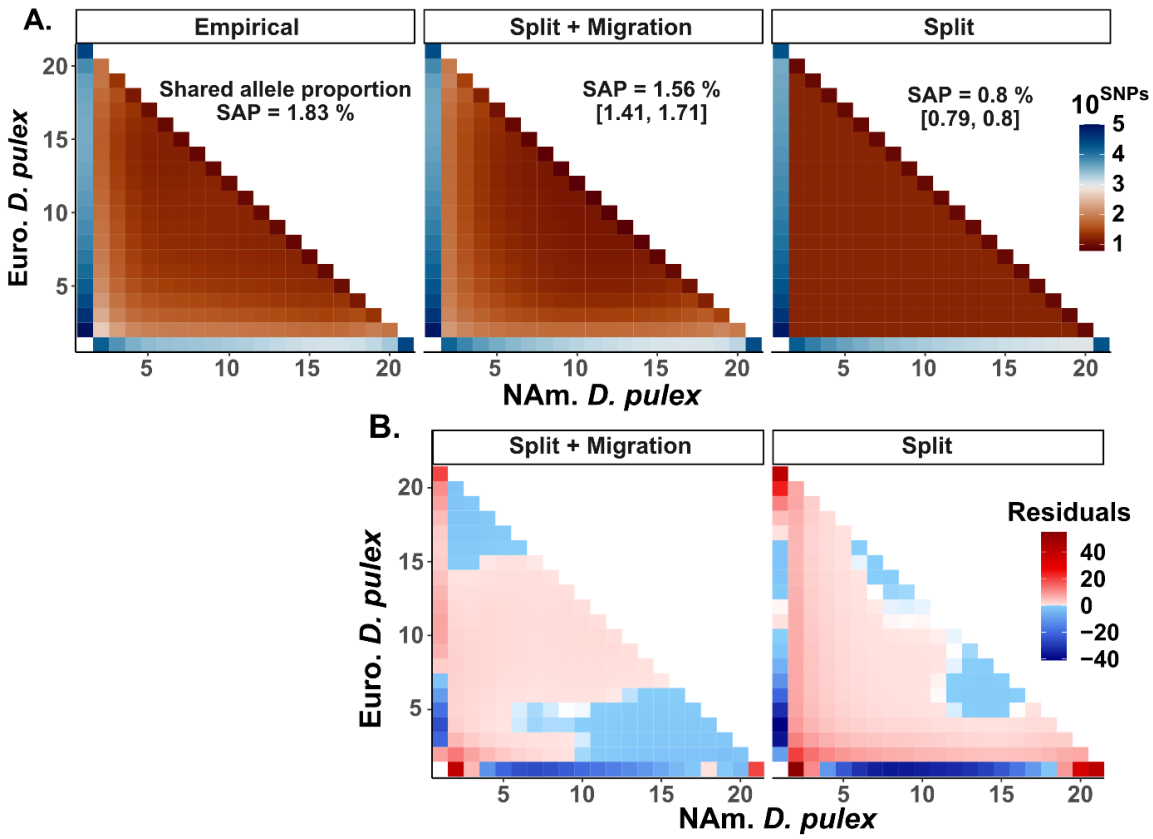


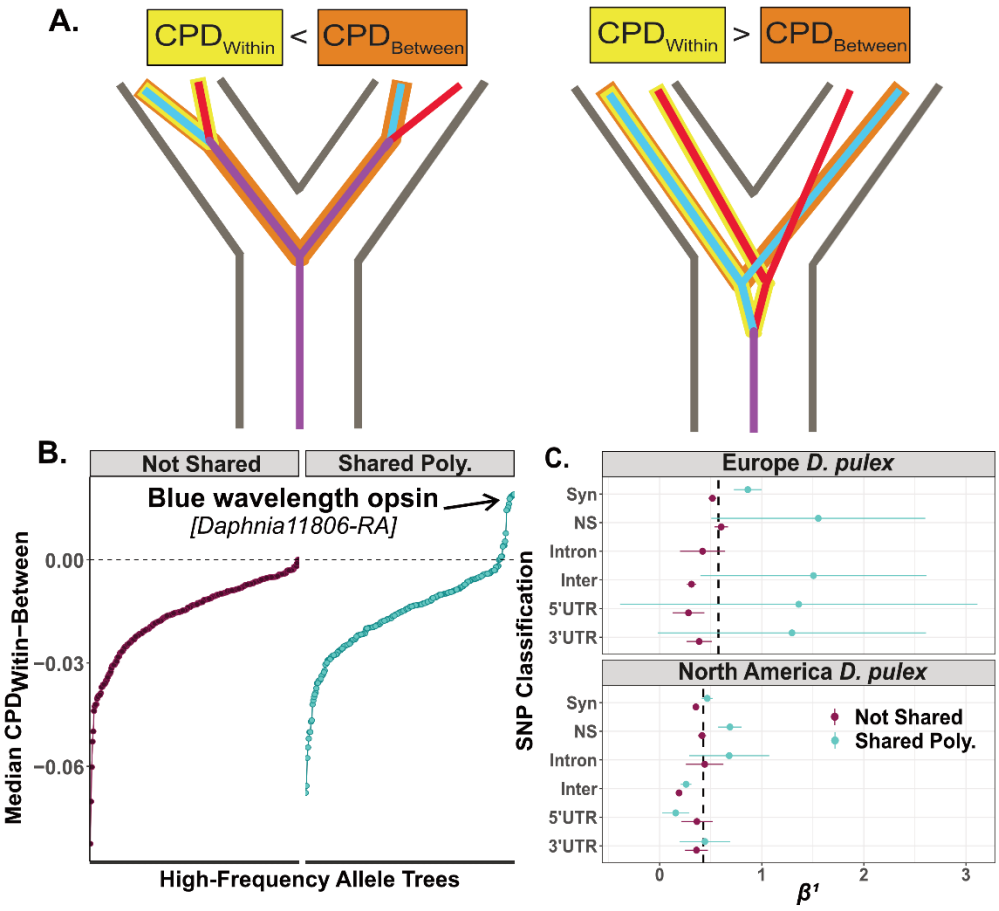
Figure 3. An excess of shared polymorphisms between North American and European *Daphnia pulex*. A) Demographic model inference between North American and European *D. pulex* based on the folded site-frequency spectrum (SFS). The empirical SFS is constructed from the genome-wide SNP dataset. The split with migration ("*Split + Migration"*) and split without migration ("*Split"*) models were generated from *moments* and we are showing the mean projection based on 1,000 bootstraps. The x and y-axis use a 20x20 SFS projection. **B)** Average standardized residuals for both models tested against the empirical SFS. Standardized residuals were calculated from the allele counts for each row and column combination of the SFS with the following formula: $\frac{Empirical - Model_x}{\sqrt{Model_x}}$, where $Model_x$ is "*Split + Migration*" or "*Split*".

Selective forces acting on shared polymorphisms. European and North American *D. pulex* possess an excess of shared polymorphism relative to neutral or demographic models,

suggesting that some form of selection could be maintaining these polymorphisms. We sought to identify old-balanced polymorphisms and convergently evolved polymorphisms by building allele-trees surrounding focal shared polymorphisms. If shared polymorphisms arose via convergence, then allele-trees would be concordant with the species-tree and all parental haplotypes from the same species would be reciprocally monophyletic (Figure 4A). If shared polymorphisms arose prior to the species split, then allele trees will not necessarily be concordant with the species tree. Notably, if there are multiple shared polymorphisms in close linkage, then alleles from the two species will cluster together and be distinct from the species tree. However, it is important to note that if only a single trans-specific polymorphism is the target of balancing selection that arose prior to the species split, then recombination could have eroded the signal of linked ancient polymorphism and the allele trees will be concordant with the species tree (Gao et al., 2015). Thus, our analysis cannot accurately separate convergence from trans-specificity in all cases, but can identify genes that have multiple, linked shared trans-specific polymorphisms that could be the target of long-term balancing selection.

We summarized allele tree discordance by calculating the pairwise cophenetic distances (CPD, Cardona et al., 2013) within and between haplotypes of the same-species from allele-trees that contain high-frequency (MAF>0.25), non-synonymous, shared polymorphisms. When allele trees resemble the species-tree topology, the within-species distances will be lower than the between-species distances (CPD_{Within}-$CPD_{Between} < 0$; Figure 4A). However, if alleles from two species cluster together, and are discordant with the species-tree, the within-species distances will be larger than between-species distances (CPD_{Within}-$CPD_{Between} > 0$). A small number of allele trees surrounding shared polymorphisms have a positive $CPD_{Within-Between}$ value, consistent with balancing selection maintaining trans-specific haplotypes (Figure 4A). However, most shared polymorphisms have negative $CPD_{Within-Between}$ values (Figure 4B), consistent either with convergently evolution or trans-specificity. Although determining the fraction of shared polymorphisms that arose via either selective mechanism is challenging, it seems unlikely that

all shared polymorphisms with negative $CPD_{Within-Between}$ arose via convergent adaptive evolution. This is because the probability of mutation occurring at the same nucleotide in two species is small ($\mu^2 \sim 10^{-17}$; Keith et al., 2016), coupled with the low establishment probability for anything but the most strongly beneficial mutations.



<u>Figure 4:</u> Convergent evolution, trans-specificity, and signatures of balancing selection

A) Visualization of two adaptive hypotheses that produce shared polymorphisms, convergent evolution on the left and trans-specificity on the right. For each tree, we calculated the median pairwise cophenetic distance as the distance within species (CPD_{Within}; yellow highlighted pair) – between species ($CPD_{Between}$; orange highlight) for shared polymorphisms and non-shared polymorphisms. $CPD_{Within-Between} < 0$ describes the consensus species-tree topology (Left), while

$CPD_{Within-Between} > 0$ describes an allele-specific tree topology consistent with an old mutation being maintained within the sequence (Right). The red and blue branches indicate examples of shared polymorphisms between species. B) $CPD_{Within-Between}$ for non-synonymous shared SNPs and non-shared SNPs above 0.25 minor allele frequency (MAF) in both species. Each allele-tree was made from 30 samples from North American and European *D. pulex*. At the focal SNP, we extracted 500bps surrounding the focal SNP. C) β^1 is a statistic that detects balancing selection. We show the mean with 95% standard errors for several SNP classifications (SYN=synonymous, NS=non-synonymous, Intron=intronic, Inter=intergenic, 5' UTR=5' untranslated region, 3' UTR=3' untranslated region). The dotted vertical line is the average β^1 within each species.

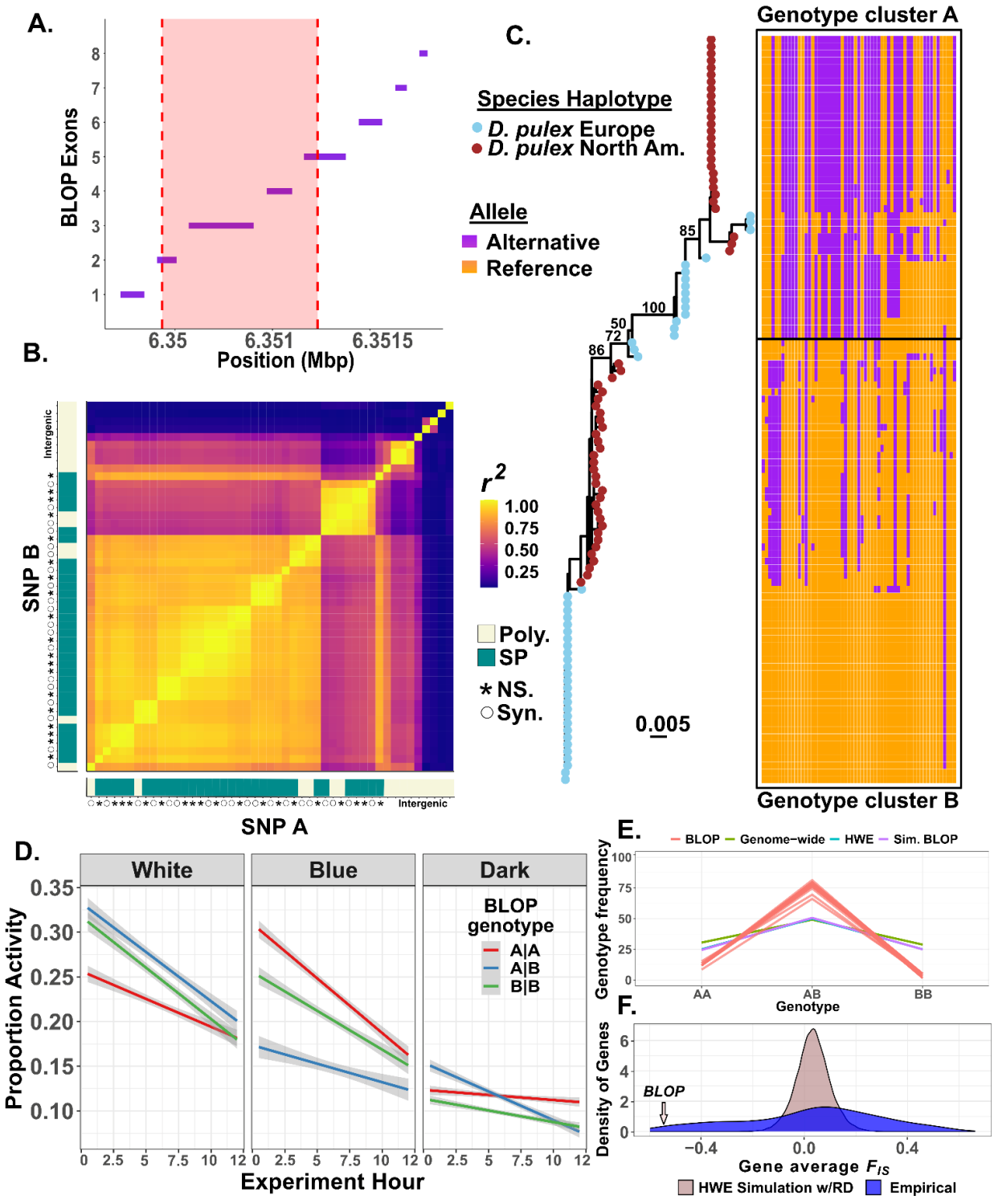
Regardless of whether shared polymorphisms arose via convergent evolution, or prior to the species split, they could have been subject to balancing selection. To test this hypothesis, we first calculated α_b, a statistic to estimate the proportion of non-synonymous sites under balancing selection using a contingency table odds ratio of both private-species' alleles and shared polymorphisms (Soni et al., 2022). We found that α_b is significantly positive across the genome, indicating that balancing selection is influencing non-synonymous shared polymorphisms (α_b=0.082 [0.05, 0.114], p=1.5x10^{-6}). Next, we calculated β^1, a site-frequency spectrum-based statistic for detecting signals of balancing selection (Siewert & Voight, 2017) at both shared and control SNPs. We found that β^1 at shared polymorphisms are significantly higher than zero in both species for non-synonymous SNPs (one sample t-test: Euro. t = 7.8, df = 270, p = 1.75x10^{-13}; NAm. t = 18.8, df = 1563, p = 2.2x10^{-16}; Figure 4C). Shared synonymous sites are also significantly elevated β^1 in both species (NAm. t = 25.12, df = 4367, p = 2.2x10^{-16}; Euro. t = 20.41, df = 1481, p = 2.2x10^{-16}; Figure 4C).

Trans-specific polymorphisms at a blue wavelength opsin affect behavior and show evidence of genetic overdominance. Of the common, non-synonymous shared polymorphisms, 14 (5%) have positive $CPD_{Within-Between}$ values (Figure 4B). Almost all of these shared polymorphisms (13/14) are within a rhabdomeric blue wavelength opsin (BLOP) gene (Brandon et al., 2017). The BLOP that we identify is found as a single copy in European and North American *D. pulex* (Supplemental Figure 4A). The 13 non-synonymous shared SNPs reside across several exons (Figure 5A) and encompass a large linkage block within European *D. pulex* ($r^2 > 0.7$ ~ 1.5kbps; Figure 5B&C), thereby explaining the allele-tree species-tree discordance (Figure 4B; Figure 5C) and suggest that these alleles are trans-specific polymorphisms (TSP) that predate the split between North American and European *D. pulex*.

If these haplotypes at the BLOP have been maintained since prior to the split between North American and European *D. pulex* 10 MYA (Figure 1C), they may have a functional effect. To test this hypothesis, we measured the light-induced activity of European *D. pulex* clones that harbor distinct haplotypes bearing alternate shared alleles. We first assigned clonal haplotypes to one of two genetic clusters (Figure 5C) and tested the activity levels of all three genotypes (AA, AB, BB) in different light conditions. We found that genotype has a significant effect on activity that is dependent on light conditions (χ^2=5,849.71, *df*=4, $p < 2x10^{-16}$; Supplemental Table 3). In general, all genotypes had low activity in dark conditions. Heterozygotes have the highest activity levels when exposed to white light yet have the lowest activity when exposed to blue light consistent with shifts between genetic overdominance and underdominance affecting behavior (Figure 5D).

Overdominance affecting behavior could also translate into overdominance affecting fitness. If trans-specific polymorphisms at the BLOP cause overdominance in fitness, heterozygotes should be more common than expected under Hardy-Weinberg equilibrium. To test the hypothesis, we examined segregation patterns of trans-specific SNPs at the BLOP among F1 offspring derived from a cross between two clones that are both heterozygous for the

trans-specific haplotypes we identified. These clones were previously referred to as “super-clone A” and “super-clone C” by Barnard-Kubow et al., (2022). Both clones had reached high frequency in the southern English (Dorset) pond D8 by the end of the 2017 growing season. In 2018, most individuals in the D8 pond were the F1 offspring between super-clone A and C enabling us to directly test if there is an excess of heterozygotes relative to the expected Mendelian segregation patterns among the F1s. First, we calculated the frequency of AA, AB, and BB genotypes at trans-specific polymorphisms, without downsampling to one clone per MLG, at the BLOP. We find that there is a strong excess of heterozygotes in the wild-caught individuals compared to expectations from Hardy-Weinberg (HWE) and compared to random SNPs in the genome or other TSPs (Figure 5E). Next, we calculated the distribution of F_{IS}, a measure of the departure of HWE, at genes across the genome and found that the BLOP gene is amongst the most strongly negative F_{IS} compared to other genes (F_{IS} = -0.54; Figure 5E). Indeed, the BLOP has amongst the smallest 2.6% of F_{IS} values that we measured. Even if we examine the genotype distribution by only sampling one genotype per clonal lineage we still observe an excess of heterozygotes (Supplemental Figure 5C&F), again suggesting natural selection in the wild. We also examined genotype frequencies in lab-generated AxC and CxC F1s. In contrast to our field-sampled individuals, we do not observe an excess of heterozygotes from a lab-generated cross of the same clones (Supplemental Figure 5A-E).



<u>Figure 5.</u> Behavioral and fitness effects of trans-specific SNPs at a blue-light opsin. A) Gene structure showing the length and position of exons within the BLOP (*Daphnia11806*). The shaded red region indicates the location of a large high-linkage block identified in panel B. **B)**

Pairwise linkage disequilibrium (r^2) for every SNP within the BLOP for European *D. pulex*, filtered for SNPs with a MAF > 0.01. The right and bottom tile objects indicate whether the SNP is polymorphic (Poly; khaki) or shared polymorphism (SP; blue-green). NS refers to non-synonymous polymorphism and Syn refers to synonymous polymorphism represented by asterisks and open circles respectively. **C)** Allele-tree made from the gene for a subset of phased samples of North American and European *D. pulex*. Tip symbols indicate whether the samples are North American or European *D. pulex*. Numbers indicate bootstrap support. The included haplotype plot and multiple-sequence alignment showcase the presence of each SNP within the gene, colored for whether the allele is derived (purple) or reference (gold). **D)** The activity of individual European *D. pulex* was measured for 12 hours for three genotypes in three different light conditions. Lines represent the best fit and 95% standard errors. **E)** Average segregation frequency of F1 genotypes expected based on a double heterozygous cross (i.e., AB x AB) using empirical read depth at each SNP. “BLOP” is the empirical segregation of trans-specific polymorphisms within the blue wavelength opsin gene among F1 genotypes. “Genome-wide” is the segregation for SNPs based on the read depth. “HWE” is the segregation pattern expected for Hardy Weinberg equilibrium. “Sim. BLOP” is the segregation pattern expected for the SNPs within the blue opsin gene based on empirical read depth. **F)** Distribution of average gene F_{IS}. “HWE Simulation w/RD” is the expected F_{IS} for each gene based on the empirical read depth for each SNP within every gene and “Empirical” is the average F_{IS} across genes. The small arrow denotes where the gene average for the blue wavelength opsin falls along the empirical distribution.

Discussion

In this study, we examined the evolutionary forces that generate and maintain shared polymorphisms in the *D. pulex* species complex. This species complex contains several taxa that have played a preeminent role in evolutionary genetics and ecology, yet their phylogenetic

relationship and nomenclature has proven challenging for over 150 years. Here, we used whole-genome sequences coupled with polymorphism data to resolve the nuclear phylogeny of members of this species group, to evaluate mechanisms that can generate shared polymorphisms between species, and to test the functional and fitness effects of ancient mutations. We show that there is an excess of shared polymorphisms between North American and European *D. pulex* that cannot be explained by neutral or demographic processes, thereby implicating some form of natural selection as a force maintaining polymorphism. For one gene, a blue wavelength opsin, we show that shared polymorphism is likely ancient, predating speciation, and has functional consequences on behavior and fitness in the wild.

Phylogenetics of the D. pulex species group. Members of the genus *Daphnia*, and the *D. pulex* species group in particular, have proven challenging from a taxonomic perspective since their early description. For instance, Leydig separated *D. pulex* from *D. magna* and *D. longispina* (Leydig, 1860, p. 117), but did not further describe divisions in the group. Richard (1896) identified *D. obtusa* as a distinct species from *D. pulex* (p. 260), but also described ten subspecies of *D. pulex* found across the Americas and Eurasia (p. 232-255). Scourfield (1942) reinforced the view that *D. obtusa* and *D. pulex* are distinct species and emphasized the view that this species group represents several lineages in various stages of speciation. Johnson (1951), in his description of British members of the *D. pulex* group, noted that American forms resembling species in the *D. pulex* group are not likely monophyletic with Eurasian species of the same name, although these naming conventions have persisted (Brooks, 1957b; Omilian & Lynch, 2009; Ye et al., 2023). The challenge of morphological classification in the *D. pulex* group stems from a limited number of diagnostic characteristics (Brooks, 1957b; Dodson, 1981), coupled with phenotypic plasticity (Colbourne et al., 1997), mating type variation (Heier & Dudycha, 2009; Jose & Dufresne, 2010), and cytological variation (Gómez et al., 2016; Hosseinie, 1966). However, recent phylogenetic analysis of mitochondrial markers has shown

the *D. pulex* group consists of many distinct lineages and that the deepest splits within the *D. pulex* species group occur between Eurasian and North American taxa (Crease et al., 2012; Ye et al., 2022). Consistent with these results, allopatric speciation has been estimated to account for roughly 40% of cladogenetic events within *Daphnia* (Adamowicz et al., 2009), a process possibly enhanced by cycles of glaciation (Chin & Cristescu, 2021). We show that substantial genetic division exists between North American and European taxa and that these taxa are separated by millions of years (Figure 1). Given the relatively deep split time between members of the *D. pulex* species group, it is likely that they have distinct features ranging from their response to environmental stimuli to their impact on the ecosystem. Further study of the behavioral, physiological, and ecological interactions of these taxa is warranted.

The complicated nature of the *D. pulex* species group is compounded by incomplete reproductive isolation between them. North American *D. pulex* and North American *D. pulicaria* are known to hybridize in the wild (Xu et al., 2015; Ye et al., 2019). Hybrids between these lineages are obligately asexual and fail to produce functional males (Tucker et al., 2013; Xu et al., 2015; Ye et al., 2019). These post-zygotic reproductive incompatibilities are a hallmark of taxa undergoing incipient speciation (Coughlan & Matute, 2020). Consistent with this view, we show that the split time between North American *D. pulex* and North American *D. pulicaria* based on the nuclear genome is recent, within 3 million years (Figure 1C). Our estimate is consistent with a study made from mitochondrial genomes (Colbourne et al., 1998), but older than another using a limited number of nuclear markers (Omilian & Lynch, 2009). Nonetheless, genomic data clearly show that hybridization between these North American lineages occurs (Figure 2). Previous analysis of mitochondrial markers placed European *D. pulicaria* as sister to the North American *D. pulex/pulicaria* clade (e.g., Marková et al., 2013), a result consistent with the nuclear phylogeny we constructed (Figure 1C). European *D. pulicaria* also shows evidence of hybridization with members of the North American *D. pulex/pulicaria* clade (Figure 2B&E), although such hybridization is not likely recent or could have occurred with other lineages in this

complex. Although the North American taxa, along with European *D. pulicaria* show signals of hybridization with each other, European *D. pulex* appears to be a well-defined species. We show that European *D. pulex* split from the other *D. pulex/pulicaria* taxa approximately 10 million years ago (Figure 1C) and has little to no evidence of recent hybridization (Figure 2A&E).

The generation of shared polymorphisms. Polymorphisms that are shared between species represent a particularly interesting class of mutation because they can reflect a wide variety of evolutionary processes. On the one hand, shared polymorphisms could reflect neutral processes when they occur between closely related species. For example, species that diverged relatively recently will share many polymorphisms because of incomplete lineage sorting (Hobolth et al., 2011) or ongoing gene-flow (Payseur & Rieseberg, 2016). While the presence of neutral shared polymorphisms due to incomplete lineage sorting or gene-flow is important for understanding features such as historical population size (Suh et al., 2015) or barriers to migration (Kutschera et al., 2014), they can obscure selective forces such as convergent adaptive evolution or balancing selection that can also generate or maintain shared polymorphism. Therefore, to examine these selective forces, it is important to identify species that have diverged long enough ago that incomplete lineage sorting and ongoing gene-flow are limited. Our work identifies European and North American *D. pulex* as two such species because of their deep split time and limited evidence for hybridization.

We show that there are tens of thousands of polymorphisms that are shared between European and North American *D. pulex* (Figure 3A, Supplemental Table 2) and suggest that natural selection is responsible for their presence. Natural selection has often been implicated as playing a key role in maintaining shared polymorphism. For instance, polymorphisms at MHC genes in vertebrates are routinely identified to be older than the species split (Aguilar et al., 2004; Azevedo et al., 2015; Klein et al., 1993) and are thought to be maintained as polymorphism via mechanisms such as negative frequency dependence or genetic

overdominance (Key et al., 2014). In other cases, shared polymorphisms in a variety of taxa have possibly arisen via convergent evolution to common selective pressures such as pathogens (Těšický & Vinkler, 2015) and have been maintained in both species via balancing selection (Solberg et al., 2008). North American and European *D. pulex* genes involved in the immune system do not show any systematic evidence of shared polymorphism (results not shown), although the ratio of non-synonymous to synonymous polymorphisms is higher for shared polymorphisms (0.58) than non-shared polymorphisms in either North American *D. pulex* or European *D. pulex* (0.53 and 0.46, respectively; see Supplemental Table 2). Therefore, it is likely that many of these shared polymorphisms are functional and subject to some form of balancing selection.

The shared non-synonymous polymorphisms that we identified have allele trees that largely reflect the species tree (Figure 4A). Taken at face value, this result is consistent with convergent evolution. Others have suggested that widespread convergent evolution is an unlikely mechanism generating shared polymorphisms (Klein et al., 1993). Is this conclusion valid for *Daphnia*? The probability of a beneficial mutation arising in a population is a function of its census size (Pennings & Hermisson, 2006) and its establishment in a population is a function of the selective value of the mutation (Haldane, 1927). While the long-term effective population size of both European and North American *D. pulex* is somewhat limited (N_e < 1 million; Supplemental Figure 3), the census size at any single pond or lake can be quite large, possibly reaching into the millions of individuals (Dudycha, 2004), while the global census size of either species can reach upwards of 10^{12} individuals (Buffalo, 2021). Therefore, across the species range, these taxa are not likely mutation-limited. Indeed, recurrent *de novo* evolution of beneficial mutations have been hypothesized to occur rapidly and contribute to within-population variation in male production rates (Barnard-Kubow et al., 2022) and morphological responses to predators (Becker et al., 2022). Temporally and spatially variable natural selection have also been shown to be a potent force acting on *Daphnia* populations (Chaturvedi et al., 2021; Lynch,

1987; Lynch et al., 2023), suggesting that positive selection on new beneficial mutations could be strong enough to prevent beneficial mutations from being lost (Flynn et al., 2017). Therefore, it is conceivable that such shared polymorphisms between North American and European *D. pulex* arose independently. On the other hand, distinguishing between convergent evolution and old trans-specific polymorphism based on comparisons between allele trees and species trees is not always possible. This is especially so when only a single trans-specific polymorphism is the direct target of selection. In this scenario, the linked neutral trans-specific polymorphisms that generate the footprint of genealogical discordance will be eroded via recombination. Regardless of whether the many shared polymorphisms that we observe between North American and European *D. pulex* arose via convergent evolution or have been maintained since prior to the species split, these mutations tend to be associated with signatures of elevated polymorphism (Figure 4C), suggestive of balancing selection, as seen in other systems (Leffler et al., 2013).

Natural selection maintains functional trans-specific polymorphisms in a blue wavelength opsin gene. We show that one gene, a blue wavelength opsin harbors trans-specific mutations that predates the split between North American and European *D. pulex* (Figure 4B, 5C). At this locus, allele trees differ from species trees, a signal that is consistent with trans-specific polymorphism (Charlesworth, 2006; Fijarczyk & Babik, 2015). This BLOP gene has 15 non-synonymous TSPs and extensive heterozygosity (Supplemental Figure 4B). The extensive heterozygosity and linkage structure of this BLOP makes it a high priority candidate for functional characterization. Research into the North American *D. pulex* genome has shown ancient expansion of opsin genes in general that occurred over 145 mya (Brandon et al., 2017). Recent work showcases that positive selection strength is distinct between North American *D. pulex* and *D. pulicaria* at opsin genes highlighting the complex patterns of selection acting upon opsins across the genome (Ye et al., 2023). It could be that this blue wavelength

opsin mediates behavioral responses like predator avoidance or vertical diel migration seen in most Daphnids (Li et al., 2022). Our laboratory experimental work shows that alternate genotypes at the BLOP have different behavioral activity patterns in response to different light conditions (Figure 5D). Indeed, it even appears that there are changes in dominance as a function of light treatment, a feature that is consistent with the long-term persistence of balanced polymorphisms (Wittmann et al., 2017).

Our genomic analyses show that there is an excess of heterozygotes at this locus. Likewise, our experimental work identified a putative fitness advantage in the wild (Figure 5E). These results are consistent with previous experiments and observations in *Daphnia* (Haag & Ebert, 2007; Hebert et al., 1982). Our result relies on temporal sampling of a single wild population along with the reconstruction of the pedigree using genomic data of wild-caught individuals (Barnard-Kubow et al., 2022). Barnard-Kubow *et al.,* (2022) show that two clones became dominant in a pond and then crossed with each other, producing a population of F1 offspring the following year. The two dominant clones were heterozygous for the trans-specific SNPs at the BLOP and thus we expect their offspring to follow a simple Mendelian 1:2:1 ratio. In contrast, we observe an excess of heterozygous individuals in the population. This pattern is largely explained by heterozygous clones reaching higher frequency in the population by the time they were sampled suggesting that heterozygotes had higher fitness and thus were more likely to survive. By contrasting genotype frequencies from the field to the lab (Supplemental Figure 5), we conclude that the excess of heterozygotes in the field is not likely due to factors such as inbreeding depression or associative overdominance (Ohta, 1971). Instead, these patterns likely emerged due to the action of natural selection. Given the strong link between looming stimulus, movement, and predator avoidance in *Daphnia* (Pijanowska & Kowalczewski, 1997; Ringelberg, 1999; Van Gool & Ringelberg, 2003), we hypothesize that trans-specific polymorphisms at the BLOP locus may play a role in conferring a fitness advantage by reducing encounters with predators or by facilitating migration through the water column.

Conclusion

Our study elucidates the evolutionary history and genetic structure of the *D. pulex* species complex and provides evidence that shared polymorphisms are common between cryptic species. We show that balancing selection broadly influences shared polymorphisms and that a small fraction predates the species-split. We experimentally study the functional significance of shared polymorphisms across specific ecological contexts and show that these polymorphisms are associated with fitness in the wild. While we present four hypotheses related to the origin and maintenance of shared polymorphism (hybridization, incomplete lineage sorting, convergence, and balancing selection), these hypotheses are not mutually exclusive. Additionally, the evolutionary mechanisms presented as hypotheses will all be affected by background levels of recombination, historic shifts in N_e, and patterns of positive and purifying selection acting upon the genome (Charlesworth, 2009; Charlesworth, 2006). Despite this challenge, we laid the groundwork for understanding the mechanisms by which genetic diversity is maintained between cryptic *D. pulex* species.

Author contributions. CSM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing - original draft, Writing - review & editing. MK: Methodology, Investigation, Writing - review & editing. DJB: Formal analysis, Software, Visualization, Writing - review & editing. MD: Investigation, Writing - review & editing. DB: Investigation, Resources, Writing - review & editing. JCBN: Formal analysis, Software, Methodology, Writing - review & editing. AR: Formal analysis, Software, Writing - review & editing. AOB: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing - review & editing

Acknowledgments. The authors acknowledge members of the Bergland lab for their feedback related to the manuscript's development. We thank Benedict Adam Lenhart, Daniel Nondorf, Christopher Robinson, Zoë Ogilvie, and Kendall Branham for their valuable comments on early versions of the manuscript. We also thank Dieter Ebert for his advice about the divergence-time estimates. The authors acknowledge Research Computing at UVA for providing computational resources and technical support that have contributed to the results reported within this manuscript. URL: https://rc.virginia.edu.

Funding information. A.O.B. was supported by grants from the NIH (R35 GM119686), and by start-up funds provided by UVA. C.S.M. was supported by the Expand NSF NRT program at UVA. D.J.B. was supported by a Harrison Undergraduate Research Award from UVA.

References

Adamowicz, S. J., Petrusek, A., Colbourne, J. K., Hebert, P. D. N., & Witt, J. D. S. (2009). The scale of divergence: A phylogenetic appraisal of intercontinental allopatric speciation in a passively dispersed freshwater zooplankton genus. *Molecular Phylogenetics and Evolution*, *50*(3), 423–436. https://doi.org/10.1016/j.ympev.2008.11.026

Aguilar, A., Roemer, G., Debenham, S., Binns, M., Garcelon, D., & Wayne, R. K. (2004). High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences*, *101*(10), 3490–3494. https://doi.org/10.1073/pnas.0306582101

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-246

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*

Azevedo, L., Serrano, C., Amorim, A., & Cooper, D. N. (2015). Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. *Human Genomics*, *9*(1), 21. https://doi.org/10.1186/s40246-015-0043-1

Barnard-Kubow, K. B., Becker, D., Murray, C. S., Porter, R., Gutierrez, G., Erickson, P., Nunez, J. C. B., Voss, E., Suryamohan, K., Ratan, A., Beckerman, A., & Bergland, A. O. (2022). Genetic Variation in Reproductive Investment Across an Ephemerality Gradient in *Daphnia pulex*. *Molecular Biology and Evolution*, *39*(6), msac121. https://doi.org/10.1093/molbev/msac121

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., & Kishony, R. (2015). Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE*, *10*(5), Article 5. https://doi.org/10.1371/journal.pone.0128036

Becker, D., Barnard-Kubow, K., Porter, R., Edwards, A., Voss, E., Beckerman, A. P., & Bergland, A. O. (2022). Adaptive phenotypic plasticity is under stabilizing selection in *Daphnia*. *Nature Ecology & Evolution*, *6*(10), 1449–1457. https://doi.org/10.1038/s41559-022-01837-5

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., & Petrov, D. A. (2014). Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in *Drosophila*. *PLoS Genetics*, *10*(11), e1004775. https://doi.org/10.1371/journal.pgen.1004775

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., & Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, *10*(4), 1–6. https://doi.org/10.1371/journal.pcbi.1003537

Brandon, C. S., Greenwold, M. J., & Dudycha, J. L. (2017). Ancient and Recent Duplications Support Functional Diversity of *Daphnia* Opsins. *Journal of Molecular Evolution*, *84*(1), 12–28. https://doi.org/10.1007/s00239-016-9777-1

Brooks, J. L. (1957a). The species problem in freshwater animals*. Pp. 81-123, in: The Species Problem*. Amer. Assoc. Advan. Sci.

Brooks, J. L. (1957b). The systematics of North American *Daphnia*. Memoirs of the Connecticut Academy of Arts & Science.

Buffalo, V. (2021). Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox. *eLife*, *10*, e67509. https://doi.org/10.7554/eLife.67509

Cardona, G., Mir, A., Rosselló, F., Rotger, L., & Sánchez, D. (2013). Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. *BMC Bioinformatics*, *14*(1), 3. https://doi.org/10.1186/1471-2105-14-3

Castoe, T. A., De Koning, A. P. J., Kim, H.-M., Gu, W., Noonan, B. P., Naylor, G., Jiang, Z. J., Parkinson, C. L., & Pollock, D. D. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences*, *106*(22), 8986–8991. https://doi.org/10.1073/pnas.0900233106

Černý, M., & Hebert, P. D. N. (1999). Intercontinental allozyme differentiation among four holarctic *Daphnia* species. *Limnology and Oceanography*, *44*(6), 1381–1387. https://doi.org/10.4319/lo.1999.44.6.1381

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, *2*(4), 379–384. https://doi.org/10.1371/journal.pgen.0020064

Chaturvedi, A., Zhou, J., Raeymaekers, J. A. M., Czypionka, T., Orsini, L., Jackson, C. E., Spanier, K. I., Shaw, J. R., Colbourne, J. K., & De Meester, L. (2021). Extensive standing genetic variation from a small number of founders enables rapid adaptation in *Daphnia*. *Nature Communications*, *12*(1), 4306. https://doi.org/10.1038/s41467-021-24581-z

Chen, N.-C., Solomon, B., Mun, T., Iyer, S., & Langmead, B. (2021). Reference flow: Reducing reference bias using multiple population genomes. *Genome Biology*, *22*(1), 8. https://doi.org/10.1186/s13059-020-02229-3

Chin, T. A., & Cristescu, M. E. (2021). Speciation in *Daphnia*. *Molecular Ecology*, mec.15824. https://doi.org/10.1111/mec.15824

Chiu, J. C., Low, K. H., Pike, D. H., Yildirim, E., & Edery, I. (2010). Assaying Locomotor Activity to Study Circadian Rhythms and Sleep Parameters in *Drosophila*. *Journal of Visualized Experiments*, *43*, 2157. https://doi.org/10.3791/2157

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 w^{1118} ; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Clark, A. G. (1997). Neutral behavior of shared polymorphism. *Proceedings of the National Academy of Sciences*, *94*(15), 7730–7734. https://doi.org/10.1073/pnas.94.15.7730

Colbourne, J. K., Crease, T. J., Weider, L. J., Hebert, P. D. N., Duferesne, F., & Hobaek, A. (1998). Phylogenetics and evolution of a circumarctic species complex (Cladocera: *Daphnia pulex*). *Biological Journal of the Linnean Society*, *65*(3), 347–365. https://doi.org/10.1111/j.1095-8312.1998.tb01146.x

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., Tokishita, S., Aerts, A., Arnold, G. J., Basu, M. K., Bauer, D. J., Caceres, C. E., Carmel, L., Casola, C., Choi, J.-H., Detter, J. C., Dong, Q., Dusheyko, S., Eads, B. D., … Boore, J. L. (2011). The Ecoresponsive Genome of *Daphnia pulex*. *Science*, *331*(6017), 555–561. https://doi.org/10.1126/science.1197761

Colbourne, J. K.J, Hebert, P. D., Taylor, D. J., & Givnish, T. J. (1997). Evolutionary origins of phenotypic diversity in *Daphnia. Molecular evolution and adaptive radiation*, 163-188.

Cornetti, L., Fields, P. D., Van Damme, K., & Ebert, D. (2019). A fossil-calibrated phylogenomic analysis of *Daphnia* and the Daphniidae. *Molecular Phylogenetics and Evolution*, *137*, 250–262. https://doi.org/10.1016/j.ympev.2019.05.018

Coughlan, J. M., & Matute, D. R. (2020). The importance of intrinsic postzygotic barriers throughout the speciation process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1806), 20190533. https://doi.org/10.1098/rstb.2019.0533

Crease, T. J., Omilian, A. R., Costanzo, K. S., & Taylor, D. J. (2012). Transcontinental Phylogeography of the *Daphnia pulex* Species Complex. *PLoS ONE*, *7*(10), e46620. https://doi.org/10.1371/journal.pone.0046620

Crow, J. F. & Kimura, M. (1970). An introduction to population genetics theory. New York, NY: Harper & Row, Publishers, Inc.

Daniel, F., Revolution Analytics, & Weston, S. (2022). *doMC: Foreach Parallel Adaptor for "parallel."*

Decaestecker, E., Gaba, S., Raeymaekers, J. A. M., Stoks, R., Van Kerckhoven, L., Ebert, D., & De Meester, L. (2007). Host–parasite 'Red Queen' dynamics archived in pond sediment. *Nature*, *450*(7171), 870–873. https://doi.org/10.1038/nature06291

Dodson, S. I. (1981). Morphological variation of *Daphnia pulex* Leydig (Crustacea: Cladocera) and related species from North America. *Hydrobiologia*, *83*(1), 101–114. https://doi.org/10.1007/BF02187155

Dowle, M., & Srinivasan, A. (2023). data.table: Extension of "data.frame." *Https://R-Datatable.Com, Https://Rdatatable.Gitlab.Io/Data.Table, Https://Github.Com/Rdatatable/Data.Table*.

Dudycha, J. L. (2004). Mortality dynamics of *Daphnia* in contrasting habitats and their role in ecological divergence. *Freshwater Biology*, *49*(5), 505–514. https://doi.org/10.1111/j.1365-2427.2004.01201.x

Ebert, D. (2022). *Daphnia* as a versatile model system in ecology and evolution. *EvoDevo*, *13*(1), 16. https://doi.org/10.1186/s13227-022-00199-0

Edwards, ScottV., & Beerli, P. (2000). Perspective: Gene Divergence, Population Divergence, and the Variance in Coalescence Time in Phylogeographic Studies. *Evolution*, *54*(6), 1839–1854. https://doi.org/10.1111/j.0014-3820.2000.tb01231.x

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Erickson, P. A., Weller, C. A., Song, D. Y., Bangerter, A. S., Schmidt, P., & Bergland, A. O. (2020). Unique genetic signatures of local adaptation over space and time for diapause, an ecologically relevant complex trait, in *Drosophila melanogaster*. *PLOS Genetics*, *16*(11), e1009110. https://doi.org/10.1371/journal.pgen.1009110

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Flynn, J. M., Chain, F. J. J., Schoen, D. J., & Cristescu, M. E. (2017). Spontaneous Mutation Accumulation in *Daphnia pulex* in Selection-Free vs. Competitive Environments. *Molecular Biology and Evolution*, *34*(1), 160–173. https://doi.org/10.1093/molbev/msw234

Gao, Z., Przeworski, M., & Sella, G. (2015). Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution*, *69*(2), 431–446. https://doi.org/10.1111/evo.12567

Garnier, S., Ross, N., BoB Rudis, Filipovic-Pierucci, A., Galili, T., Timelyportfolio, Greenwell, B., Sievert, C., Harris, D. J., & JJ Chen. (2021). *sjmgarnier/viridis: Viridis 0.6.0 (pre-CRAN release)* (v0.6.0pre). https://doi.org/10.5281/ZENODO.4679424

Geoffrey Fryer. (1991). Functional morphology and the adaptive radiation of the Daphniidae (Branchiopoda: Anomopoda). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *331*(1259), 1–99. https://doi.org/10.1098/rstb.1991.0001

Gómez, R., Van Damme, K., Gosálvez, J., Morán, E. S., & Colbourne, J. K. (2016). Male meiosis in Crustacea: Synapsis, recombination, epigenetics and fertility in *Daphnia magna*. *Chromosoma*, *125*(4), 769–787. https://doi.org/10.1007/s00412-015-0558-1

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, *15*(7), e1008302. https://doi.org/10.1371/journal.pgen.1008302

Haag, C. R., & Ebert, D. (2007). Genotypic selection in *Daphnia* populations consisting of inbred sibships. *Journal of Evolutionary Biology*, *20*(3), 881–891. https://doi.org/10.1111/j.1420-9101.2007.01313.x

Haldane, J. B. S. (1927). A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *23*(7), 838–844. https://doi.org/10.1017/S0305004100015644

Harris, RS. (2007). Improved Pairwise Alignment of Genomic DNA. *The Pennsylvania State University, University Park, PA.*

Hebert, P. D. N., Ferrari, D. C., & Crease, T. J. (1982). Heterosis in *Daphnia*: A Reassessment. *The American Naturalist*, *119*(3), 427–434. https://doi.org/10.1086/283921

Hebert, P. D. N., & Wilson, C. C. (1994). Provincialism in plankton: endemism and allopatric speciation in australian *Daphnia*. *Evolution*, *48*(4), 1333–1349. https://doi.org/10.1111/j.1558-5646.1994.tb05317.x

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *22*(18), 4606–4618. https://doi.org/10.1111/mec.12415

Heier, C. R., & Dudycha, J. L. (2009). Ecological speciation in a cyclic parthenogen: Sexual capability of experimental hybrids between *Daphnia pulex* and *Daphnia pulicaria*. *Limnology and Oceanography*, *54*(2), 492–502. https://doi.org/10.4319/lo.2009.54.2.0492

Held, C., Koenemann, S., & Schubart, C. D. (Eds.). (2016). Phylogeography and Population Genetics in Crustacea (0 ed.). CRC Press. https://doi.org/10.1201/b11113

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., & Mailund, T. (2011). Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Research*, *21*(3), 349–356. https://doi.org/10.1101/gr.114751.110

Hosseinie, Farammarz. (1966). The ecology and reproductive cytology of *Daphnia middendorffiana,* FischerCladocera) from the arctic*.* Indiana University ProQuest Dissertations Publishing.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., … Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, *512*(7513), 194–197. https://doi.org/10.1038/nature13408

Jackson, C. E., Xu, S., Ye, Z., Pfrender, M. E., Lynch, M., Colbourne, J. K., & Shaw, J. R. (2021). Chromosomal rearrangements preserve adaptive divergence in ecological speciation. *BioRxiv*, 2021-08. https://doi.org/10.1101/2021.08.20.457158

Johnson, D. S. (1951). A study of the physiology and ecology of certain Cladocera. University of London, Bedford College (United Kingdom) ProQuest Dissertations Publishing.

Jose, C., & Dufresne, F. (2010). Differential survival among genotypes of *Daphnia pulex* differing in reproductive mode, ploidy level, and geographic origin. *Evolutionary Ecology*, *24*(2), 413–421. https://doi.org/10.1007/s10682-009-9314-4

Jouganous, J., Long, W., Ragsdale, A. P., & Gravel, S. (2017). Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics*, *206*(3), 1549–1567. https://doi.org/10.1534/genetics.117.200493

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). *Poppr*: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. https://doi.org/10.7717/peerj.281

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Keith, N., Tucker, A. E., Jackson, C. E., Sung, W., Lucas Lledó, J. I., Schrider, D. R., Schaack, S., Dudycha, J. L., Ackerman, M., Younge, A. J., Shaw, J. R., & Lynch, M. (2016). High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. *Genome Research*, *26*(1), 60–69. https://doi.org/10.1101/gr.191338.115

Key, F. M., Teixeira, J. C., de Filippo, C., & Andrés, A. M. (2014). Advantageous diversity maintained by balancing selection in humans. *Current Opinion in Genetics & Development*, *29*, 45–51. https://doi.org/10.1016/j.gde.2014.08.001

Klein, J., Sato, A., Nagl, S., & O'hUigín, C. (1998). Molecular trans-species polymorphism. *Annual Review of Ecology and Systematics*, *29*(1), 1–21. https://doi.org/10.1146/annurev.ecolsys.29.1.1

Klein, J., Satta, Y., Takahata, N., & O'hUigin, C. (1993). Trans-specific *Mhc* polymorphism and the origin of species in primates. *Journal of Medical Primatology*, *22*(1), 57–64. https://doi.org/10.1111/j.1600-0684.1993.tb00637.x

Koenig, D., Hagmann, J., Li, R., Bemm, F., Slotte, T., Neuffer, B., Wright, S. I., & Weigel, D. (2019). Long-term balancing selection drives evolution of immunity genes in Capsella. *eLife*, *8*, e43606. https://doi.org/10.7554/eLife.43606

Kutschera, V. E., Bidon, T., Hailer, F., Rodi, J. L., Fain, S. R., & Janke, A. (2014). Bears in a Forest of Gene Trees: Phylogenetic Inference Is Complicated by Incomplete Lineage Sorting and Gene Flow. *Molecular Biology and Evolution*, *31*(8), 2004–2017. https://doi.org/10.1093/molbev/msu186

Lee, B. T., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, C. M., Muthuraman, P., Nassar, L. R., Nguy, B., Pereira, T., Perez, G., Raney, B. J., Rosenbloom, K. R., Schmelter, D., Speir, M. L., … Kent, W. J. (2022). The UCSC Genome Browser database: 2022 update. *Nucleic Acids Research*, *50*(D1), D1115–D1122. https://doi.org/10.1093/nar/gkab959

Leffler, E. M., Gao, Z., Pfeifer, S., Segurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J. D., Sella, G., Donnelly, P., McVean, G., & Przeworski, M. (2013). Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science*, *339*(6127), 1578–1582. https://doi.org/10.1126/science.1234070

Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(Database), D19–D21. https://doi.org/10.1093/nar/gkq1019

Leydig, F. (1860). Naturgeschichte der Daphniden:(Crustacea cladocera). Laupp & Siebeck.

Li, D., Huang, J., Zhou, Q., Gu, L., Sun, Y., Zhang, L., & Yang, Z. (2022). Artificial Light Pollution with Different Wavelengths at Night Interferes with Development, Reproduction, and Antipredator Defenses of *Daphnia magna*. *Environmental Science & Technology*, *56*(3), 1702–1712. https://doi.org/10.1021/acs.est.1c06286

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*:1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lynch, M. (1987). The Consequences of Fluctuating Selection for Isozyme Polymorphisms in *Daphnia*. *Genetics*, *115*(4), 657–669. https://doi.org/10.1093/genetics/115.4.657

Lynch, M., Gutenkunst, R., Ackerman, M., Spitze, K., Ye, Z., Maruki, T., & Jia, Z. (2017). Population Genomics of *Daphnia pulex*. *Genetics*, *206*(1), 315–332. https://doi.org/10.1534/genetics.116.190611

Lynch, M., Wei, W., Ye, Z., & Pfrender, M. (2023). The Genome-wide Signature of Short-term Temporal Selection. *bioRxiv*, 2023.04.28.538790. https://doi.org/10.1101/2023.04.28.538790

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite—Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*(2), 584–595. https://doi.org/10.1111/1755-0998.13265

Marková, S., Dufresne, F., Manca, M., & Kotlík, P. (2013). Mitochondrial Capture Misleads about Ecological Speciation in the *Daphnia pulex* Complex. *PLOS ONE*, *8*(7), e69497. https://doi.org/10.1371/journal.pone.0069497

Martin, M., Ebert, P., & Marschall, T. (2023). Read-Based Phasing and Analysis of Phased Variants with WhatsHap. In B. A. Peters & R. Drmanac (Eds.), *Haplotyping* (Vol. 2590, pp. 127–138). Springer US. https://doi.org/10.1007/978-1-0716-2819-5_8

Mayer, W. E., Jonker, M., Klein, D., Ivanyi, P., Van Seventer, G., & Klein, J. (1988). Nucleotide sequences of chimpanzee MHC class I alleles: Evidence for trans-species mode of evolution. *The EMBO Journal*, *7*(9), 2765–2774. https://doi.org/10.1002/j.1460-2075.1988.tb03131.x

McCoy, R. C., Garud, N. R., Kelley, J. L., Boggs, C. L., & Petrov, D. A. (2014). Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population. *Molecular Ecology*, *23*(1), 136–150. https://doi.org/10.1111/mec.12591

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, *8*(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092

Novikova, P. Y., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., Guggisberg, A., Paape, T., Schmid, K., Fedorenko, O. M., Holm, S., Säll, T., Schlötterer, C., Marhold, K., Widmer, A., Sese, J., Shimizu, K. K., Weigel, D., Krämer, U., … Nordborg, M. (2016). Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, *48*(9), 1077–1082. https://doi.org/10.1038/ng.3617

Nunez, J. C. B., Rong, S., Damian-Serrano, A., Burley, J. T., Elyanow, R. G., Ferranti, D. A., Neil, K. B., Glenner, H., Rosenblad, M. A., Blomberg, A., Johannesson, K., & Rand, D. M. (2021). Ecological Load and Balancing Selection in Circumboreal Barnacles. *Molecular Biology and Evolution*, *38*(2), 676–685. https://doi.org/10.1093/molbev/msaa227

Ohta, T. (1971). Associative overdominance caused by linked detrimental mutations. *Genetical Research*, *18*(3), 277–286. https://doi.org/10.1017/S0016672300012684

Omilian, A. R., & Lynch, M. (2009). Patterns of Intraspecific DNA Variation in the *Daphnia* Nuclear Genome. *Genetics*, *182*(1), 325–336. https://doi.org/10.1534/genetics.108.099549

Pantel, J. H., Juenger, T. E., & Leibold, M. A. (2011). Environmental gradients structure *Daphnia pulex* × *pulicaria* clonal distribution: Hybrid *Daphnia* clonal distribution. *Journal of Evolutionary Biology*, *24*(4), 723–732. https://doi.org/10.1111/j.1420-9101.2010.02196.x

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, *25*(11), 2337–2360. https://doi.org/10.1111/mec.13557

Pennings, P. S., & Hermisson, J. (2006). Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. *Molecular Biology and Evolution*, *23*(5), 1076–1084. https://doi.org/10.1093/molbev/msj117

Pijanowska, J., & Kowalczewski, A. (1997). Predators can induce swarming behaviour and locomotory responses in *Daphnia*. *Freshwater Biology*, *37*(3), 649–656. https://doi.org/10.1046/j.1365-2427.1997.00192.x

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van Der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples*. BioRxiv, 201178.* https://doi.org/10.1101/201178

R Core Development Team. (2013). R: A language and environment for statistical computing.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

Revell, L. J. (2019). *learnPopGen*: An R package for population genetic simulation and numerical analysis. *Ecology and Evolution*, *9*(14), 7896–7902. https://doi.org/10.1002/ece3.5412

Richard J. (1896). Révision des Cladocères. Deuxième Partie. Anomopoda. Famille III.— Daphnidae (Vol. 2). Ann. Sci. Nat. Zool.

Ringelberg, J. (1999). The photobehaviour of *Daphnia spp*. As a model to explain diel vertical migration in zooplankton. *Biological Reviews of the Cambridge Philosophical Society*, *74*(4), 397–423. https://doi.org/10.1017/S0006323199005381

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Schiffels, S., & Wang, K. (2020). MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. In J. Y. Dutheil (Ed.), *Statistical Population Genomics* (Vol. 2090, pp. 147–166). Springer US. https://doi.org/10.1007/978-1-0716-0199-0_7

Scourfield, D. J. (1942). XIX.—The “Pulex” forms of *Daphnia* and their Separation into Two Distinct Series. *Annals and Magazine of Natural History*, *9*(51), 202–219. https://doi.org/10.1080/03745481.1942.9755477

Ségurel, L., Gao, Z., & Przeworski, M. (2013). Ancestry runs deeper than blood: The evolutionary history of *ABO* points to cryptic variation of functional importance. *BioEssays*, *35*(10), 862–867. https://doi.org/10.1002/bies.201300030

Ségurel, L., Thompson, E. E., Flutre, T., Lovstad, J., Venkat, A., Margulis, S. W., Moyse, J., Ross, S., Gamble, K., Sella, G., Ober, C., & Przeworski, M. (2012). The ABO blood group is a trans-species polymorphism in primates. *Proceedings of the National Academy of Sciences*, *109*(45), 18493–18498. https://doi.org/10.1073/pnas.1210603109

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction* (Vol. 1962, pp. 227–245). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_14

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Siewert, K. M., & Voight, B. F. (2017). Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Molecular Biology and Evolution*, *34*(11), 2996–3005. https://doi.org/10.1093/molbev/msx209

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Slater, G., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*(1), 31. https://doi.org/10.1186/1471-2105-6-31

So, M., Ohtsuki, H., Makino, W., Ishida, S., Kumagai, H., Yamaki, K. G., & Urabe, J. (2015). Invasion and molecular evolution of *Daphnia pulex* in Japan. *Limnology and Oceanography*, *60*(4), 1129–1138. https://doi.org/10.1002/lno.10087

Solberg, O. D., Mack, S. J., Lancaster, A. K., Single, R. M., Tsai, Y., Sanchez-Mazas, A., & Thomson, G. (2008). Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Human Immunology*, *69*(7), 443–464. https://doi.org/10.1016/j.humimm.2008.05.001

Soni, V., Vos, M., & Eyre-Walker, A. (2022). A new test suggests hundreds of amino acid polymorphisms in humans are subject to balancing selection. *PLOS Biology*, *20*(6), e3001645. https://doi.org/10.1371/journal.pbio.3001645

Spitze, K. (1993). Population structure in *Daphnia obtusa*: Quantitative genetic and allozymic variation. *Genetics*, *135*(2), 367–374. https://doi.org/10.1093/genetics/135.2.367

Standard, A. (2007). Standard guide for conducting acute toxicity tests on test materials with fishes, macroinvertebrates, and amphibians. *West Conshohocken, PA, United States. DOI: 0.1520/E0729-96. URL: Www. Atsm. Org*.

Suh, A., Smeds, L., & Ellegren, H. (2015). The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. *PLOS Biology*, *13*(8), e1002224. https://doi.org/10.1371/journal.pbio.1002224

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, *25*(1). https://doi.org/10.1002/0471250953.bi0410s25

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309. https://doi.org/10.1038/ng.3748

Těšický, M., & Vinkler, M. (2015). Trans-Species Polymorphism in Immune Genes: General Pattern or MHC-Restricted Phenomenon? *Journal of Immunology Research*, *2015*, 1–10. https://doi.org/10.1155/2015/838035

Thomas Lin Pedersen. (2022). patchwork: The Composer of Plots. *Https://Patchwork.Data-Imaginist.Com, Https://Github.Com/Thomasp85/Patchwork*.

Tucker, A. E., Ackerman, M. S., Eads, B. D., Xu, S., & Lynch, M. (2013). Population-genomic insights into the evolutionary origin and fate of obligately asexual *Daphnia pulex*. *Proceedings of the National Academy of Sciences*, *110*(39), 15740–15745. https://doi.org/10.1073/pnas.1313388110

Unckless, R. L., Howick, V. M., & Lazzaro, B. P. (2016). Convergent Balancing Selection on an Antimicrobial Peptide in *Drosophila*. *Current Biology*, *26*(2), 257–262. https://doi.org/10.1016/j.cub.2015.11.063

Van Gool, E., & Ringelberg, J. (2003). What goes down must come up: Symmetry in light-induced migration behaviour of *Daphnia*. *Hydrobiologia*, *491*(1–3), 301–307. https://doi.org/10.1023/A:1024406324317

Villanueva, R. A. M., & Chen, Z. J. (2019). ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 160–167. https://doi.org/10.1080/15366367.2019.1565254

Wang, B., & Mitchell-Olds, T. (2017). Balancing selection and trans-specific polymorphisms. *Genome Biology*, *18*(1), 231. https://doi.org/10.1186/s13059-017-1365-1

Wang, M., Zhang, L., Zhang, Z., Li, M., Wang, D., Zhang, X., Xi, Z., Keefover-Ring, K., Smart, L. B., DiFazio, S. P., Olson, M. S., Yin, T., Liu, J., & Ma, T. (2020). Phylogenomics of the genus *Populus* reveals extensive interspecific gene flow and balancing selection. *New Phytologist*, *225*(3), 1370–1382. https://doi.org/10.1111/nph.16215

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K.,

Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wills, C. (1975). Marginal overdominance in *Drosophila*. *Genetics*, *81*(1), 177–189. https://doi.org/10.1093/genetics/81.1.177

Wittmann, M. J., Bergland, A. O., Feldman, M. W., Schmidt, P. S., & Petrov, D. A. (2017). Seasonally fluctuating selection can maintain polymorphism at many loci via segregation lift. *Proceedings of the National Academy of Sciences*, *114*(46), E9932–E9941. https://doi.org/10.1073/pnas.1702994114

Wiuf, C., Zhao, K., Innan, H., & Nordborg, M. (2004). The Probability and Chromosomal Extent of trans-specific Polymorphism. *Genetics*, *168*(4), 2363–2372. https://doi.org/10.1534/genetics.104.029488

Wu, Q., Han, T.-S., Chen, X., Chen, J.-F., Zou, Y.-P., Li, Z.-W., Xu, Y.-C., & Guo, Y.-L. (2017). Long-term balancing selection contributes to adaptation in *Arabidopsis* and its relatives. *Genome Biology*, *18*(1), 217. https://doi.org/10.1186/s13059-017-1342-8

Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., Dai, Z., Lam, T. T., Guan, Y., & Yu, G. (2022). *Ggtree*: A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta*, *1*(4). https://doi.org/10.1002/imt2.56

Xu, S., Spitze, K., Ackerman, M. S., Ye, Z., Bright, L., Keith, N., Jackson, C. E., Shaw, J. R., & Lynch, M. (2015). Hybridization and the Origin of Contagious Asexuality in *Daphnia pulex*. *Molecular Biology and Evolution*, msv190. https://doi.org/10.1093/molbev/msv190

Ye, Z., Molinier, C., Zhao, C., Haag, C. R., & Lynch, M. (2019). Genetic control of male production in *Daphnia pulex*. *Proceedings of the National Academy of Sciences*, *116*(31), 15602–15609. https://doi.org/10.1073/pnas.1903553116

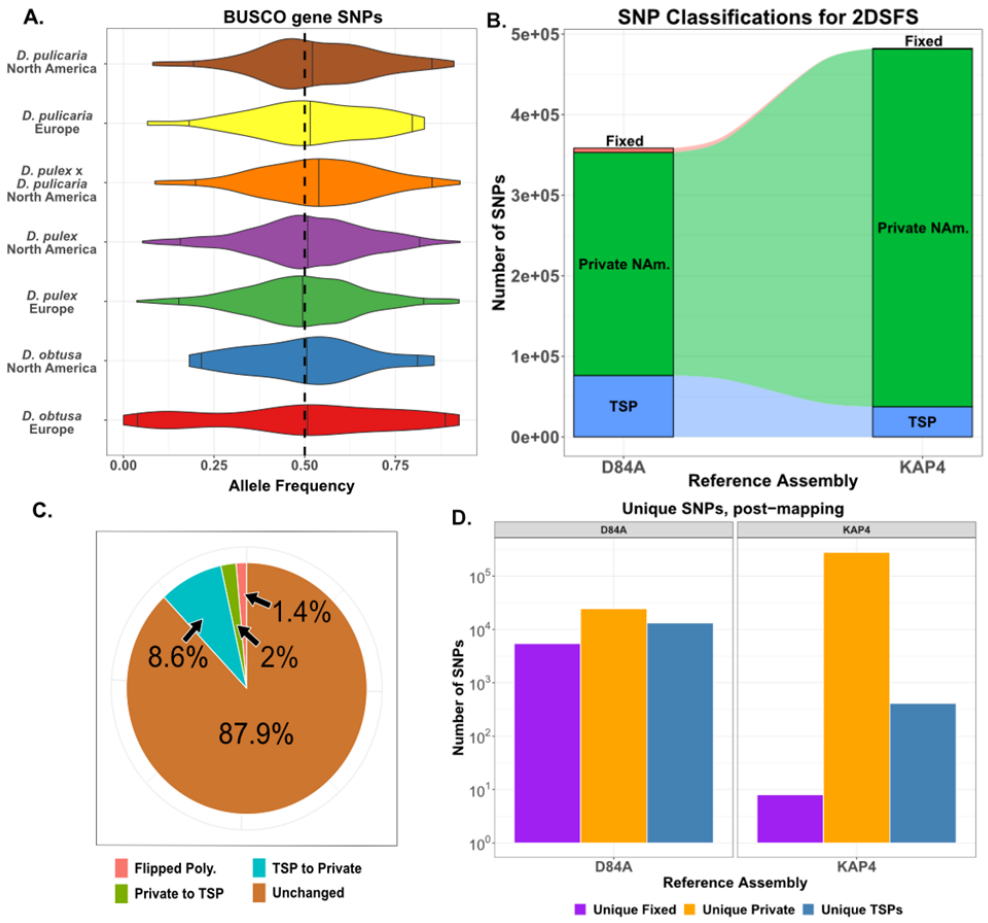
Ye, Z., Pfrender, M. E., & Lynch, M. (2023a). Evolutionary Genomics of Sister Species Differing in Effective Population Sizes and Recombination Rates. *Genome Biology and Evolution*, *15*(11), evad202. https://doi.org/10.1093/gbe/evad202

Ye, Z., Zhao, C., Raborn, R. T., Lin, M., Wei, W., Hao, Y., & Lynch, M. (2022). Genetic Diversity, Heteroplasmy, and Recombination in Mitochondrial Genomes of *Daphnia pulex* , *Daphnia pulicaria* , and *Daphnia obtusa*. *Molecular Biology and Evolution*, *39*(4), msac059. https://doi.org/10.1093/molbev/msac059

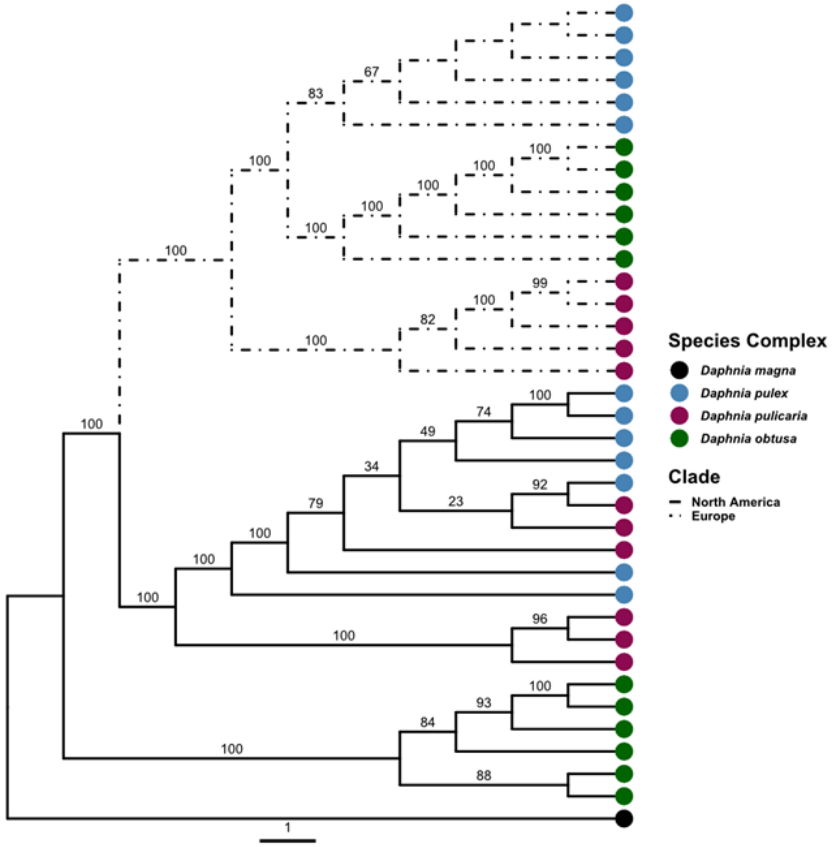
Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zheng, X., Gogarten, S. M., Lawrence, M., Stilp, A., Conomos, M. P., Weir, B. S., Laurie, C., & Levine, D. (2017). SeqArray—A storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, *33*(15), 2251–2257. https://doi.org/10.1093/bioinformatics/btx145

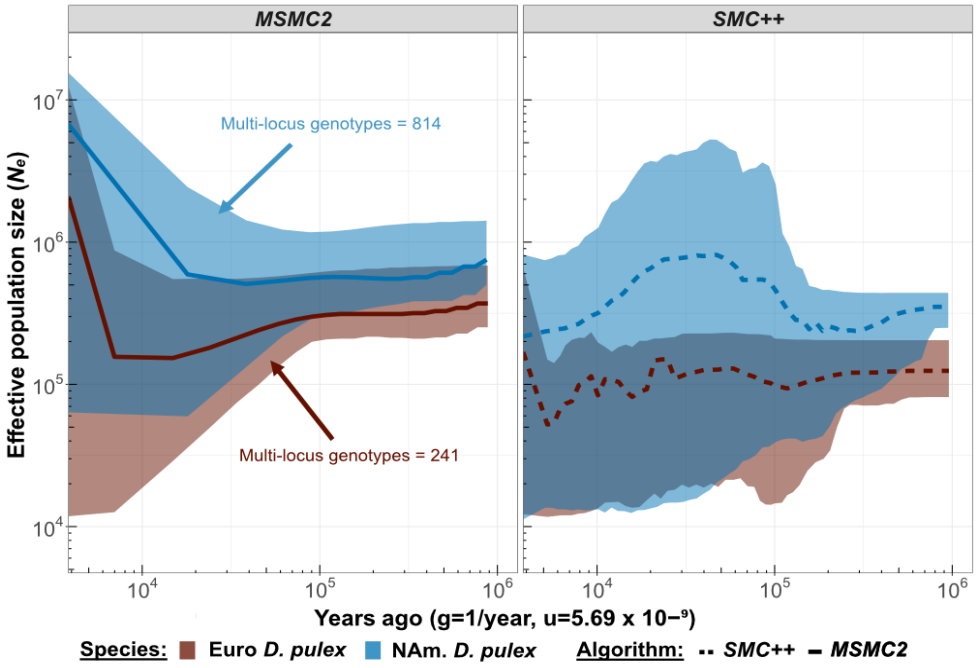
Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606



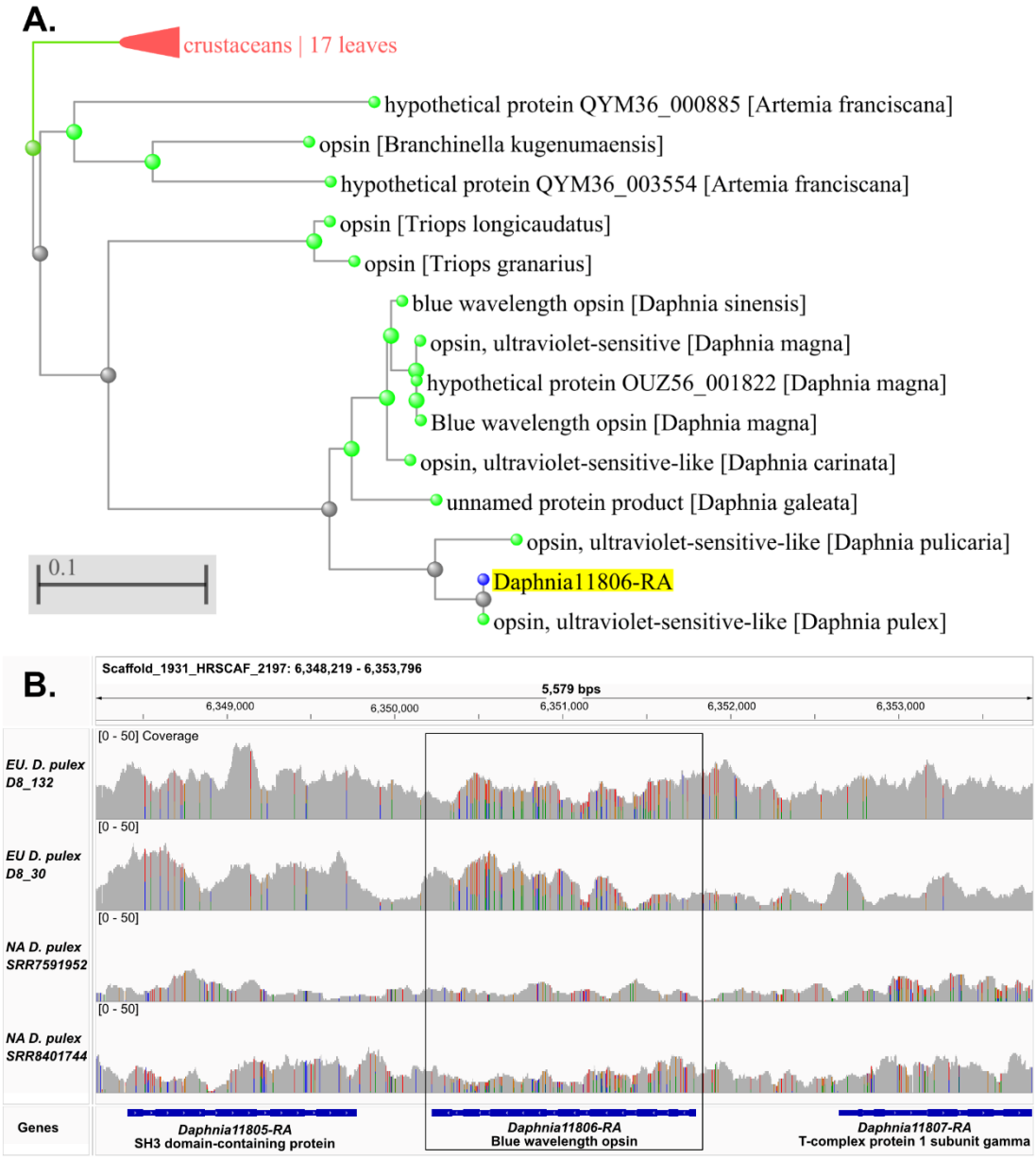
<u>Supplemental Figure 1.</u> No evidence of reference allele bias across the *Daphnia pulex* species complex. A) For each species, we extracted representative individuals (ranging from n=5-100 depending on the number of samples per species) and 1,000 biallelic heterozygous BUSCO gene SNPs (100 bootstraps) to gauge the severity of reference allele bias across the genome. We calculated the proportion of the alternative and reference dosage within a given individual for each site. The x-axis measures the proportion of alternative to reference dosage for each SNP and we show the 95% quantiles and median. **B)** Alluvial plot of the SNP classifications between assemblies of the European *D. pulex* (D84A) and the North American *D. pulex* (KAP4). **C)** Proportion of SNP classification changes when mapping to KAP4 exclusively. **D)** The number of classified SNPs that are exclusive to each assembly.



<u>Supplemental Figure 2.</u> Mitochondrial protein-coding tree for the *D. pulex* species complex. A maximum-likelihood tree with the "TN+F+I+G4" model output from IQTree2. The tree is rooted with *D. magna* as an outgroup. Bootstrap supports are listed as node labels.

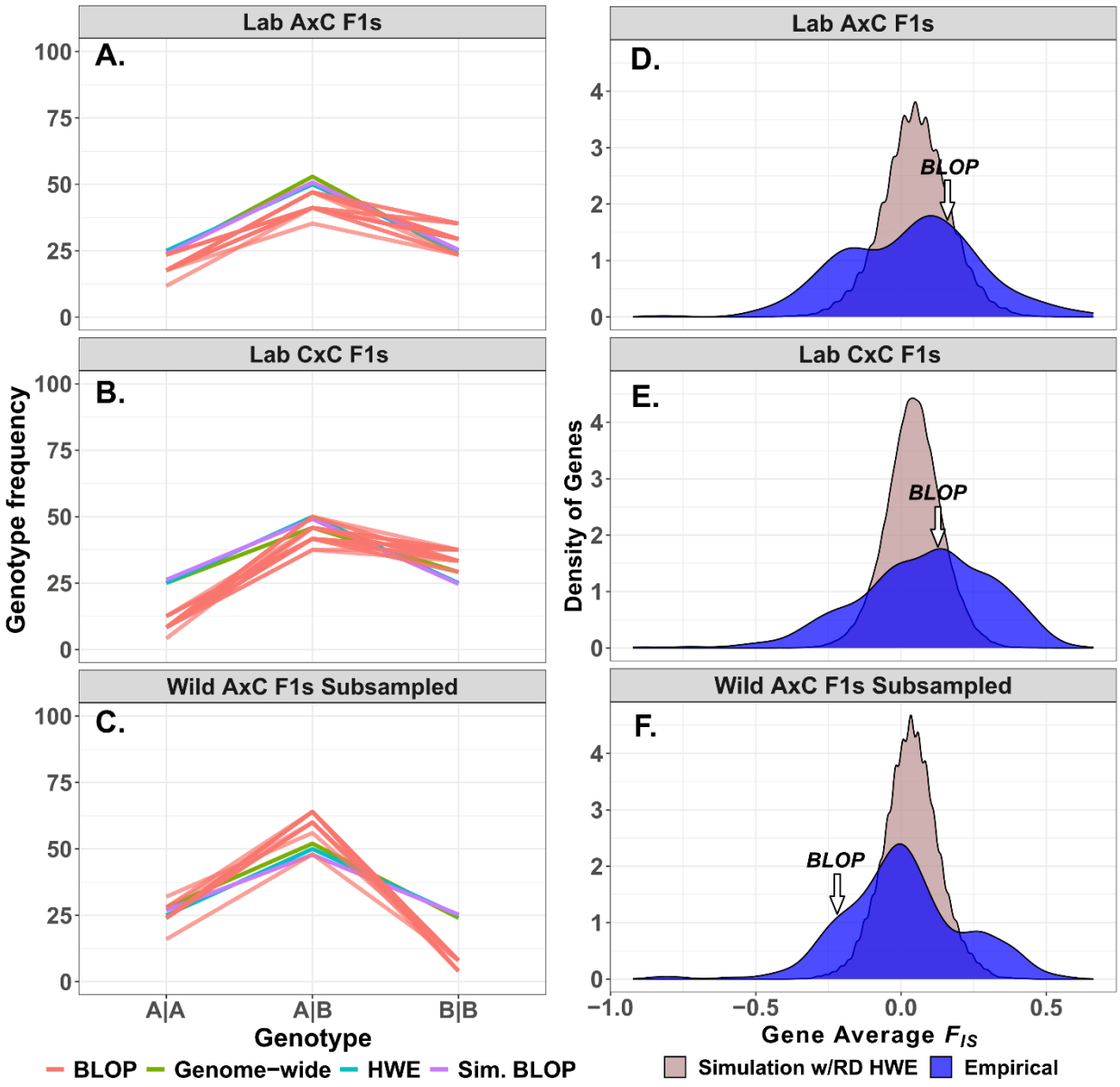


<u>Supplemental Figure 3.</u> Demographic reconstruction of North American and European *D. pulex* species. MSMC2 and SMC++ output for each multi-locus genotype sample. Each multi-locus genotype sample was run independently. The shaded ribbon shows the upper 95% quantiles and lower 5% quantiles from the run estimates.



<u>Supplemental Figure 4.</u> Blue wavelength opsin gene and orthologous proteins in Crustacea and within-species heterozygosity. A) This neighbor-joining protein tree was generated using *Blast's* tree widget. The query sequence is highlighted in yellow and has a blue tip symbol. The green tip symbols are related Crustacean protein sequences with the species name in brackets. **B)** We subsampled two representative individuals within the European (EU) and North American (NA) *D. pulex* species and are showing the coverage for each individual set to [0-50]. The vertical-colored bars are heterozygous regions (i.e., split-colored bars) and

homozygous alternative alleles (i.e., whole-colored bars), gray base pairs are the reference allele.



<u>Supplemental Figure 5.</u> Segregation patterns and F_{IS} of polymorphisms across lab and wild crossed *Daphnia* clones. A-C) Average segregation frequency of F1 genotypes expected based on a double heterozygous cross (i.e., AB x AB) using empirical read depth at each SNP. We produced crosses of AxC in the lab shown in panel A and CxC shown in panel B. Panel C shows the F1 genotypes subsampled based on their status belonging to superclones identified in Barnard-Kubow et al. 2022, reflecting a conservative sampling approach. “Genome-wide” is the segregation for SNPs based on the read depth. “HWE” is the segregation pattern expected for Hardy Weinberg equilibrium. “Sim. BLOP” is the segregation pattern expected for the SNPs within the blue opsin gene based on empirical read depth. “BLOP” is the empirical segregation of trans-specific polymorphisms within the blue wavelength opsin gene among F1 genotypes. **D-F)** Distribution of average gene F_{IS}. “HWE Simulation w/RD” is the expected F_{IS} for each gene based on the empirical read depth for each SNP within every gene and “Empirical” is the average is the F_{IS} across genes. The small arrow denotes where the gene average for the blue wavelength opsin falls along the empirical distribution.

Chapter 3

Forward genetic simulations and insight into seasonal demographic changes of overwintering *Drosophila melanogaster*

Connor S. Murray[1, *], Joaquin C. B. Nunez[1,2], Alan O. Bergland[1]

[1] Department of Biology, University of Virginia, Charlottesville, VA, USA

[2] Department of Biology, University of Vermont, 109 Carrigan Drive, Burlington, VT 05405, USA

ORCIDs:

Connor S. Murray: 0000-0002-8302-6585

Joaquin C. B. Nunez: 0000-0002-3171-8918

Alan O. Bergland: 0000-0001-7145-7575

Conflict of interest: The authors declare no conflicts of interest.

The work in this chapter is published in *Genetics*:

Nunez, J. C., Lenhart, B. A., Bangerter, A., Murray, C. S., Mazzeo, G. R., Yu, Y., Taylor L Nystrom, Courtney Tern, Priscilla A Erickson, Bergland, A. O. (2023). A cosmopolitan inversion facilitates seasonal adaptation in overwintering *Drosophila*. *Genetics*, iyad207.

https://doi.org/10.1093/genetics/iyad207

Abstract

Local demography is a useful tool for predicting shifts in population size in response to environmental changes. Studying local demography is important for predicting how population size will shift in response to environmental change. Allele frequencies are shifted by demography in sometimes massive ways across large swaths of the genome. Understanding how bottlenecks influence genetic diversity is key to better predict population dynamics and the maintenance of diversity in the face of changing environments. In this work, I examine the local seasonal consequences of overwintering on a population of *Drosophila melanogaster*, the common fruit fly, in Charlottesville, Virginia. I use forward genetic simulations and approximate Bayesian computation (ABC) to predict the population bottleneck size that the Charlottesville population endures every year and estimate the local N_e. Our results suggest massive population bottlenecks in the order of 98% and a local N_e of 2,200. Our study is useful because it illustrates how bottlenecks affect allele frequencies and can lead to insights regarding predictability of local extinctions in the face of seasonal change within wild *D. melanogaster* populations.

Introduction

The genome carries the markers of demographic events and selection encoded in its genetic diversity. Populations that historically went through a loss in population size, i.e., a bottleneck, will tend to lose genetic diversity, resulting in large-scale shifts in allele frequencies (Kirkpatrick & Jarne, 2000). Subsequently, loss of genetic diversity will reduce the ability for populations to adapt in the face of changing environments, will increase the rate of inbreeding, and can result in the accumulation of deleterious alleles (e.g., mutation meltdown: Lynch et al., 1995). These processes can ultimately result in extinction.

There are many species that endure repeated population expansion and contraction on a yearly or seasonal basis and yet remain extant (Bouzat, 2010). For example, many arthropods will go through cycles of summer “booms” and winter “busts” due to factors like competition, food availability, precipitation, and temperature (Alvarez et al., 2007; Franks et al., 2011; Hasselmann et al., 2015). Terrestrial arthropods alone make up as much biomass as that of humans and livestock combined (Rosenberg et al., 2023), so these population dynamics are a large impact on the biomass of Earth. Despite this great influence, we have a limited understanding of the scope of these boom-and-bust events across seasons within any given species or how these ubiquitous demographic events affect population-level genetic diversity and differentiation through short-timescales (Kovach & McCouch, 2008; Maron et al., 2015).

One popular approach to understand both short- and long-term demographic processes is to couple genetic simulations and summary statistics with approximate Bayesian computation (ABC). In this way, ABC can infer which simulated demographic parameters “best match” with the genetic results gained from next generation sequencing (NGS) datasets (Csilléry et al., 2010; Sunnåker et al., 2013). ABC has been used for years to investigate demographic parameters and is especially useful for time-series datasets (Csilléry et al., 2010; Saubin et al., 2023; Shafer et al., 2015). Pool-sequencing is a method that reduces cost while retaining information content from sequencing experiments and has performed well with ABC approaches

(Carvalho et al., 2023). However, ABC on pooled-sequencing data is still limited due to the lack of computational tools available for analyses.

Drosophila melanogaster, the common fruit fly, is a cosmopolitan human commensal that inhabits temperate environments and endures population boom-and-busts over the year (Behrman et al., 2015; Pool, 2015). While *D. melanogaster* has been studied for over a century, there is a limited understanding of the natural demographic events this species goes through in focal populations. Additionally, it remains unclear how genetic diversity and especially allele frequencies are affected by seasonal boom-and-busts over years. It has been thought that the long-term effective population size (N_e) of *D. melanogaster* is 1.5 - 2.5 million, yet the census size of the species could be in the billions (10^8-10^{20}; Buffalo, 2021; Karasov et al., 2010). Focal populations of *D. melanogaster* have been estimated to have an N_e of around 10,000 (Lange et al., 2022).

Despite being subject to large-scale demographic declines in the wild, many fly populations maintain genetic diversity through time, potentially through the action of balancing selection (Bergland et al., 2014). Populations of *D. melanogaster* across North America and Europe show hallmarks of large-scale allele frequency changes likely involved with seasonal adaptation, indicating that balancing selection could be widespread (Machado et al., 2021). These important findings also suggest that *D. melanogaster* do not extirpate and recolonize every year following the winter; rather, populations endure collapse and expand in size, at least in Charlottesville, Virginia (Bangerter, 2021; Machado et al., 2021). While we know a fair amount regarding the nature of worldwide *D. melanogaster* populations, we still need to refine our understanding of local population dynamics and study how genetic diversity and divergence change through time. Our research fills this gap by providing detailed estimates of population bottlenecks and insight into population-level genetic diversity, thus enhancing our understanding of how these processes operate on a molecular level.

In this work, we use a dense time-series (every 2 weeks) pooled-sequencing dataset of *D. melanogaster* from a Charlottesville, VA apple and peach orchard to understand the scope of population bottlenecks and estimate the local N_e from the 2016-2019 growing seasons (Nunez et al., 2024). Our approach uses forward genetic simulations with many N_{Max} and N_{Min} instantaneous models to understand the scope of population bottlenecks. Our results suggest that boom-and-bust bottlenecks are on the order of 98%, and we highlight the utility of ABC on pooled-sequencing data and the power for estimating the demographic events that wild populations endure (Gautier et al., 2013).

Materials and Methods

Fly sampling: New samples for this study were collected at an orchard in Charlottesville, VA (Carter Mountain Orchard, 37.99N, 78.47W) from 2016 to 2019. Collections from 2016 to 2018 were done using aspirators and netting every 2 weeks starting in mid-June when peaches come into season in central VA and ending in mid-December at the end of the fall apple season. The collection in 2019 was done at the beginning of the growing season in June. Because *D. melanogaster* is phenotypically similar to its sister taxa *D. simulans*, we determined species identity using the male offspring produced from isofemale lines set from wild-caught flies. *D. melanogaster* isofemale offspring were frozen in ethanol and stored at −20°C prior to sequencing.

DNA Extraction, sample preparation, and sequencing: Libraries were made using G1 male offspring from wild-caught isofemale lines. For pool-seq, we prepared 37 libraries (see number of pooled flies in Supplemental Table 1). Pool-seq sequencing, filtering, and mapping were done following the protocols outlined in (Kapun et al., 2021) using the DEST dockerized pipeline (https://github.com/DEST-bio/DEST_freeze1).

Pooled sequencing bioinformatics pipeline: Quality control, mapping, SNP calling, and dataset merging were done using the DEST dataset mapping pipeline (https://github.com/DEST-bio/DEST_freeze1) using the optimized settings for the PoolSNP caller (Kapun et al., 2020) and enforcing a global average minimum allele frequency of 1%. The DEST mapping pipeline accounts for potential contamination with *Drosophila simulans* in the pools using competitive mapping. We combined the Charlottesville pool-seq with the pool-seq samples from DEST to generate a new dataset that contains 283 pooled samples from 22 countries across 12 years 2003-2018. SNPs inside Repetitive elements, defined by the Interrupted Repeats, Microsatellite, RepeatMasker, SimpleRepeats, and WM_SDust tracks from UCSC Genome Browser (Morgulis et al., 2006) were removed from further analysis. Additional bioinformatic details can be found in our GitHub repository (https://github.com/Jcbnunez/Cville-Seasonality-2016-2019).

Forward genetic demographic simulations: To test if overwintering bottlenecks influence patterns of genetic differentiation through time, and to infer minimum and maximum population sizes during boom-and-bust cycles that are consistent with our data, we performed genetic simulations. First, we performed a coalescent-based neutral simulation of a single population with $\theta_\pi = 0.001$ using *msprime* (Baumdicker et al., 2022) in *Python 3.8*. This neutral background was used as a burn-in within the forward genetics software, *SLiM 3* (Haller & Messer, 2019). *SLiM 3* was used to simulate cyclic population crashes while varying the population size maximum (N_{Max}) and the population size minimum (N_{Min}) under a model of the instantaneous change in population size (Figure 1a). For each parameter combination, the simulated population had a constant size at N_{Max} from generations 1–16, 19–33, and 36–50 and the bottlenecks occurred at generations 17–18 and 34–35 where the population size was set to N_{Min} (Figure 1b). The generation decisions were made to emulate those that occur in wild populations where there is roughly 15 generations of growth. A Variant Call Format (VCF) file of

50 simulated diploid individuals was output at the end of each generation to track allele frequency changes. AF were simulated to mimic pooled sequencing using *poolSeq v0.3.5* (Taus et al., 2017) with a mean coverage of 60. Pairwise FST was calculated using *poolfstat v2.1.1* (Gautier et al., 2022). Every parameter combination was simulated 100 independent times with different seeds. Parameter estimation was performed using Approximate Bayesian Computation (ABC) using the local linear regression method (loclinear) with a tolerance threshold of 5% using *abc v2.1* (Csilléry et al., 2012) in *R*. The summary statistics used were the medians of within year F_{ST}, between year F_{ST}, and the correlation (R^2) of PC1, LD1, and LD2 values relative to the simulation year (Figure 1c and d). These latter three statistics are, respectively, the principal component (PC) projections of dimensions 1 (i.e. PC1), and the first and second linear discriminants (i.e. LD1–2) of a discriminant analysis of principal components (DAPC; Jombart, 2008), using the simulated year as a grouping prior. For PCA, we used a matrix of AF (columns) and samples (rows) in the *PCA()* function from *FactoMineR*. The first and second PC values from each sample were extracted and used in a simple linear regression with simulation year (Years 1–3) to calculate correlations (i.e. PC1 ~ Year, PC2 ~ Year). We repeated this step for both the first and second linear discriminant (LD) axes as well. First, a matrix of AF and samples was used in the *dapc* function in *adegenet v2.1.10* (Jombart, 2008) with simulation year as a grouping prior. After extracting LD1 and LD2 values, we ran a linear regression with the LD values and simulation year. In this way, we were able to measure how the severity of yearly bottlenecks affects both PC and LD space due to shifts in AF across samples. A leave-one-out analysis was performed on the input summary statistics to understand how each contributes to the estimates of N_{Max} and N_{Min}.

Data and scripts availability. The *R*, *SLiM*, and bash scripts used for all analyses are deposited on our GitHub repository: https://github.com/Jcbnunez/Cville-Seasonality-2016-2019/tree/main/CODE/5.Simulation_Demography.

The data used for our analyses are on Zenodo: https://doi.org/10.5281/zenodo.7271502.

Results

Seasonal boom-bust demography: To quantify the general strength of a winter bottleneck, we conducted forward genetic simulations designed to emulate the boom-bust cycle and sampling scheme for the Charlottesville samples (Figure 1A). We simulated 50 generations (~3 years) of a population with similar genetic properties as *D. melanogaster*. We subjected these populations to yearly cycles of population size change of variable magnitude (booms-and-busts), as well as a null model of constant population size (Figure 1A & B).

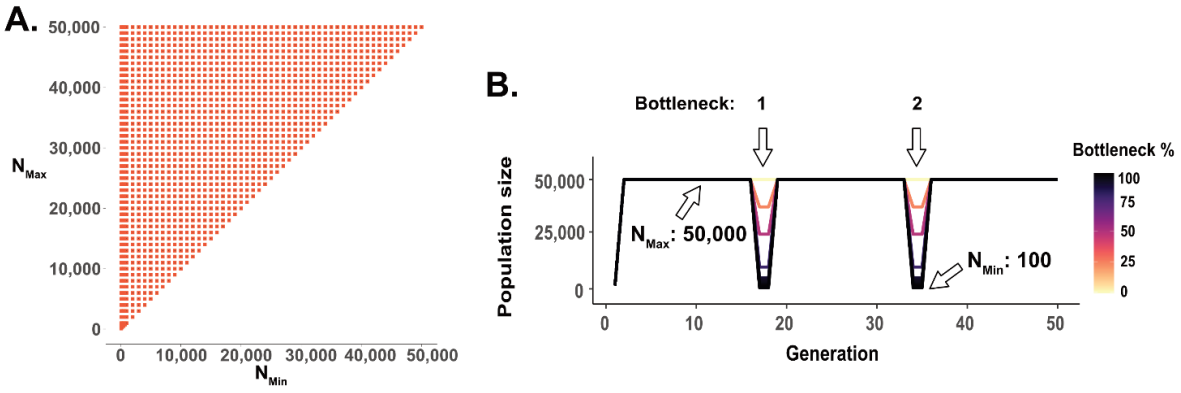
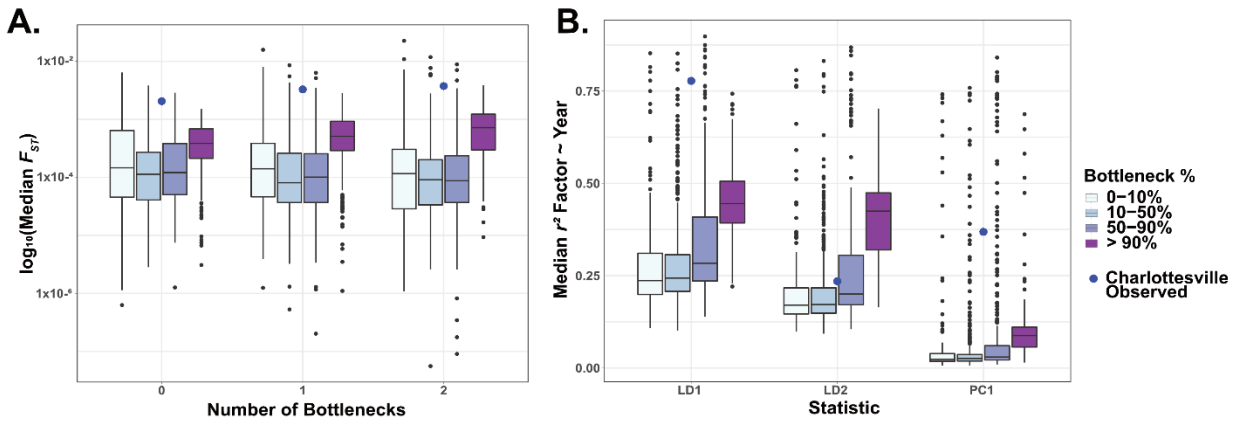


Figure 1: Simulated boom-and-bust demography. A) Simulations were conducted with variable maximum (N_{Max}) and minimum (N_{Min}) population sizes, each point indicates that 100 independent seed simulations were run for each N_{Max} and N_{Min} combination. **B)** Cartoon model of the simulated overwintering demography, illustrating population dynamics during bottlenecks. We ran instantaneous population size changes during the overwintering generations.

We calculated a variety of summary statistics (see Materials and Methods), including pairwise F_{ST} between each yearly pool sample. We show that the median pairwise F_{ST} decays over time increasingly with bottleneck severity (Figure 2A). Comparing simulated to empirical

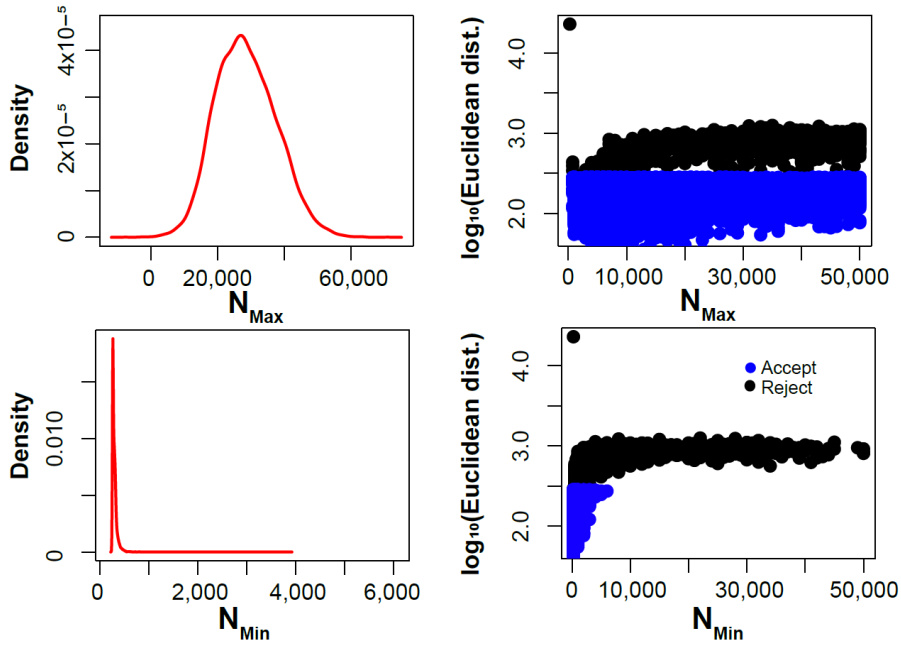
data, we found that the Charlottesville population has F_{ST} values in the upper-tail distributions of most of the simulated events. We also used the correlation (r^2) of principal component 1 (PC1), linear discriminant 1 (LD1), and LD2 values relative to the simulation year. In all three of these multidimensional statistics, we see an increase in r^2 with bottleneck severity (Figure 2B). The Charlottesville population again is in the upper-tail extremes for LD1 and PC1, but for LD2 the empirical value is closer to the 50-90% bottleneck median values.



<u>Figure 2:</u> Summary statistics calculated across variable bottleneck severity. A) The median pairwise-F_{ST} within and between simulation years. **B)** The median r^2 of principal component (PC) and linear discriminant (LD) axes across simulation year. The blue dot marks the location of observed values within the Charlottesville population.

We used the six summary statistics within approximate Bayesian computation (ABC) to determine the set of parameters that most closely fit the Charlottesville data. Our ABC results provide support for the hypothesis of yearly population expansions and contractions and suggest that the magnitude of winter collapse in Charlottesville is on the order of 98% of the maximum summer size (median N_{Min} = 283 [97.5% CI: 260; 406], median N_{Max} = 27,584 [13,217; 46,746], median N_e [effective population size, i.e. the harmonic mean of N] = 2,234 [1,926; 3,240]). The Euclidean distances between the empirical and simulated values of N_{Min} had tight

95% confidence intervals, from 260-406 compared to the much larger distribution in N_{Max} of 13,000-47,000 (Figure 3).



<u>Figure 3:</u> Posterior probabilities and Euclidean distances for demographic parameters. The *loclinear* model of approximate Bayesian computation (ABC) was employed for parameter estimation, and posterior probabilities for N_{Max} and N_{Min} parameters are shown on the left. On the right, Euclidean distances quantify the disparities between cumulative simulation results and observed Charlottesville population statistics. Simulations were only accepted by ABC if they were close enough to the observed data under a tolerance threshold of 5%.

We were next interested to see which of the summary statistics contribute most to the estimates of N_{Max} and N_{Min} and so we performed a leave-one-out analysis compared to the inclusion of every statistic. In this way, we can individually assess how ABC distributes its 95% confidence intervals between both N_{Max} and N_{Min}. This analysis revealed that N_{Max} was affected only by exclusion of LD1 and LD2. N_{Min} was massively affected by the exclusion of LD1 and LD2 and sensitive to the exclusion of F_{ST} estimates and PC1 (Figure 4). Overall, we show that the

most important statistics are both LD1 and LD2 on ABC, because their exclusion results in massive shifts in the estimated N_{Max} and N_{Min} values.

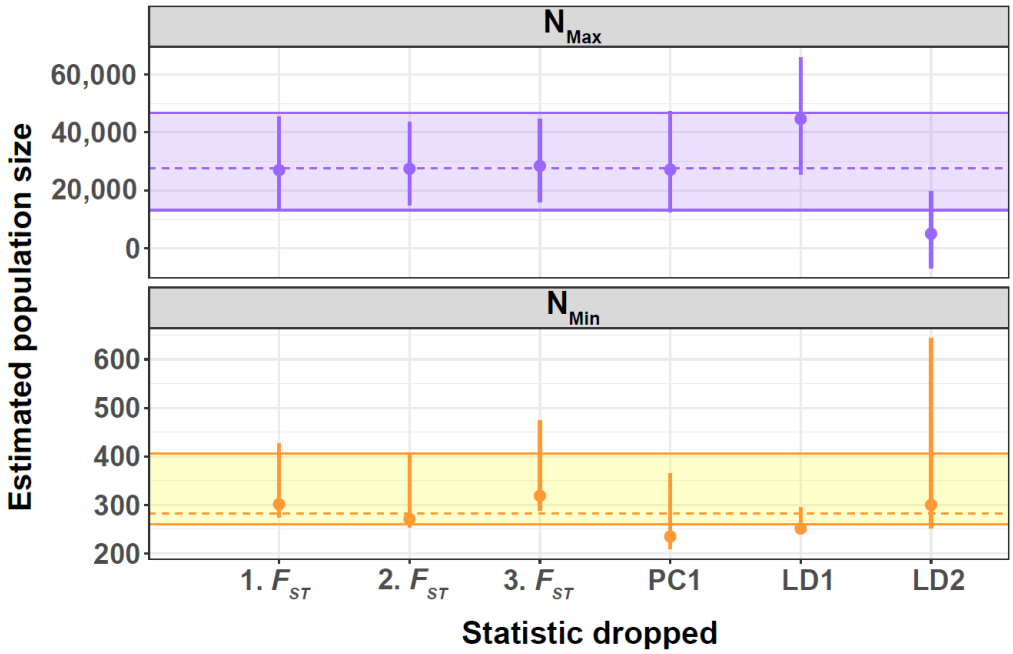


Figure 4: Leave-one-out analysis on ABC fit. A leave-one-out analysis was performed, excluding each statistic individually (x-axis), with 95% confidence intervals (lines) and median values (points) illustrating the resulting parameter estimates for N_{Max} and N_{Min}. Shaded colored intervals represent the 95% confidence interval for the *loclinear* ABC output when all statistics were employed, and the dotted line signifies the median value. Deviations from the shaded region indicate the influence of individual statistics on estimated population sizes for N_{Max} and N_{Min}.

Discussion

In this work, we use forward genetic simulations and approximate Bayesian computation (ABC) to estimate the degree of bottleneck that the *Drosophila melanogaster* population endures during the winter seasons of 2016-2019 in Charlottesville, VA. Our results show

population reductions of 98% and a minimum size of ~300 and a maximum size of ~30,000. Empirically, we knew that winter bottlenecks in *D. melanogaster* were large based on the reduction of seasonal temperatures during the late-fall and winter months, however, we did not have an estimate of the population dynamics. We could not have obtained this from censusing due to the scarcity of flies during winter months. Our genomic approach fills this gap by providing an estimate of the population dynamics during winter bottlenecks, enhancing our understanding of how these seasonal changes impact *D. melanogaster* populations and their genetic diversity.

Our work was motivated by the fact that populations endure winter collapse, yet little is known about how fruit flies overwinter and survive the harsh winter months. We had evidence that at our experimental location, Carter's Mountain Peach and Apple Orchard, populations are not going through complete extirpation and recolonization every year (Bangerter, 2021). Rather, the *D. melanogaster* population appears to be tracking alleles over time, through the action of selection differentially favoring specific combinations of phenotypes. This is called adaptive tracking, which is a form of balancing selection that results in the maintenance of genetic diversity over time and space. It could explain the relatively small increase in F_{ST} over three years for Charlottesville (see blue points in Figure 2A) and the small contribution of F_{ST} on approximate Bayesian computation (ABC) results in N_{Min} and N_{Max} (Figure 4). Adaptive tracking has been observed across global populations (Kapun et al., 2021; Nunez et al., 2024).

Our results show that pairwise F_{ST} does not change drastically over three years (Figure 2A). This stability could be explained by the population having many diapausing individuals or winter morphs (Collett & Jarman, 2007; Rossi-Stacconi et al., 2016). Founder effects might be strong, with the surviving flies forming the predominant genetic background for the summer months (Boulétreau-merle & Fouillet, 2002). Additionally, we excluded large chromosomal inversions like *In2Lt*, which are known to be driven by seasonal evolution (Machado et al., 2021; Nunez et al., 2024). Therefore, while we show limited change in genetic diversity in

Charlottesville, there could nonetheless be large shifts in inversion frequencies or other linked regions associated with seasonal selection.

Furthermore, considering phenotypic plasticity in response to seasonal changes might provide deeper insights into these dynamics. Some *Drosophila* species develop winter-morphs, a plastic response to low temperatures that enables flies to survive over the winter (Kirkpatrick et al., 2018). The classic *Drosophila* winter morph phenotype is characterized by large wings, dark pigmented bodies, and a switch towards diapause strategies, particularly in the invasive *D. suzukii* species (Erickson et al., 2020; Shearer et al., 2016). This phenotypic plasticity could be crucial for future modeling of population dynamics (Bale & Hayward, 2010). For instance, understanding how different morph proportions affect genetic diversity and allele frequencies over an overwinter bottleneck, or explicitly modeling a winter-adapted versus a summer-adapted morph could offer valuable perspectives.

In this work we employed a novel technique using ABC methods paired with pool-seq to reveal intersections between demography and genetics. ABC methods have not been widely explored in tandem with pool-seq data as tools are limited relative to whole-genome approaches. Furthermore, the coupling of pool-seq with ABC has only recently been explored with forward genetic simulations like SLiM (Haller & Messer, 2019). We chose to simulate pool-seq noise into our SLiM dataset using a static average coverage of 60. Yet, we can examine sources of error outside of just variation in coverage. For instance, pool-seq has variation in depth of coverage, unequal individual and pool contributions, miscellaneous sequencing errors, and contamination with the sister species, *D. simulans.* Carvalho et al., (2023), examined the concern with pool-seq data and found that simulating errors associated with pool-seq approaches are necessary to resolve more complex demographic situations but otherwise are robust for more simple models like done with our work. Collin et al., (2021), also examined the incidence of pool-seq bias on estimates and found high repeatability with their largescale SNP dataset. In general, while only a limited set of tests have been performed with both ABC and

simulated pool-seq data, they tend to perform well even with inherent statistical noise associated with pool-seq coverage. Even with these caveats in mind, we believe our simulations provide insight into the changes in natural *Drosophila* genetic diversity caused by bottlenecks.

In conclusion, our simulation-based analyses validate that winter bottlenecks are massive, and shift patterns of diversity accordingly. We highlight the utility of ABC approaches with pooled-sequencing datasets (Carvalho et al., 2023) and motivate future modeling studies of adaptation to overwinter census decline in wild-derived *Drosophila*. We aim to develop simulations that incorporate seasonal adaptation models that could help inform how large any given years' bottlenecks could be, which could be useful in species management as environments change and predict how populations could evolution (Hoban, 2014).

Author contributions. C.S.M, J.C.B.N, and A.O.B contributed to the simulation-based study design. CSM ran the simulations and analyzed the data, J.C.B.N analyzed the empirical data within the Charlottesville, VA population. C.S.M and A.O.B contributed to edits of this manuscript.

Acknowledgments. We would like to thank members of the Bergland lab for their thoughtful input during the development of this simulation study, as well as Keric Lamb and Antoine Perrier for their helpful comments about simulations. The authors acknowledge Research Computing at the University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication. URL: https://rc.virginia.edu.

Funding information. A.O.B. was supported by grants from the National Institutes of Health (R35 GM119686), and by start-up funds provided by the University of Virginia. C.S.M. was supported by an Expand National Science Foundation Research Traineeship program at UVA.

References

Alvarez, N., Hossaert-McKey, M., Restoux, G., Delgado-Salinas, A., & Benrey, B. (2007). Anthropogenic effects on population genetics of phytophagous insects associated with domesticated plants. *Evolution*, *61*(12), 2986–2996. https://doi.org/10.1111/j.1558-5646.2007.00235.x

Bale, J. S., & Hayward, S. A. L. (2010). Insect overwintering in a changing climate. *Journal of Experimental Biology*, *213*(6), 980–994. https://doi.org/10.1242/jeb.037911

Bangerter, A. (2021). Dense seasonal sampling of an orchard population uncovers population turnover, adaptive tracking, and structure in multiple *Drosophila* species. University of Virginia. https://doi.org/10.18130/wyr9-fz68

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

Behrman, E. L., Watson, S. S., O'Brien, K. R., Heschel, M. S., & Schmidt, P. S. (2015). Seasonal variation in life history traits in two *Drosophila* species. *Journal of Evolutionary Biology*, *28*(9), 1691–1704. https://doi.org/10.1111/jeb.12690

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., & Petrov, D. A. (2014). Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila*. *PLoS Genetics*, *10*(11), e1004775. https://doi.org/10.1371/journal.pgen.1004775

Boulétreau-merle, J., & Fouillet, P. (2002). How to overwinter and be a founder: Egg-retention phenotypes and mating status in *Drosophila melanogaster*. *Evolutionary Ecology*, *16*(4), 309–332. https://doi.org/10.1023/A:1020216230976

Bouzat, J. L. (2010). Conservation genetics of population bottlenecks: The role of chance, selection, and history. *Conservation Genetics*, *11*(2), 463–478. https://doi.org/10.1007/s10592-010-0049-0

Buffalo, V. (2021). Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox. *eLife*, *10*, e67509. https://doi.org/10.7554/eLife.67509

Carvalho, J., Morales, H. E., Faria, R., Butlin, R. K., & Sousa, V. C. (2023). Integrating Pool-seq uncertainties into demographic inference. *Molecular Ecology Resources*, *23*(7), 1737–1755. https://doi.org/10.1111/1755-0998.13834

Collett, J. I., & Jarman, M. G. (2007). Adult female *Drosophila pseudoobscura* survive and carry fertile sperm through long periods in the cold: populations are unlikely to suffer substantial bottlenecks in overwintering. *Evolution*, *55*(4), 840–845. https://doi.org/10.1111/j.0014-3820.2001.tb00820.x

Collin, F., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., Marin, J., & Estoup, A. (2021). Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using *DIYABC* Random Forest. *Molecular Ecology Resources*, *21*(8), 2598–2613. https://doi.org/10.1111/1755-0998.13413

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

Csilléry, K., François, O., & Blum, M. G. B. (2012). *abc*: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Erickson, P. A., Weller, C. A., Song, D. Y., Bangerter, A. S., Schmidt, P., & Bergland, A. O. (2020). Unique genetic signatures of local adaptation over space and time for diapause, an ecologically relevant complex trait, in *Drosophila melanogaster*. *PLOS Genetics*, *16*(11), e1009110. https://doi.org/10.1371/journal.pgen.1009110

Franks, S. J., Pratt, P. D., & Tsutsui, N. D. (2011). The genetic consequences of a demographic bottleneck in an introduced biological control insect. *Conservation Genetics*, *12*(1), 201–211. https://doi.org/10.1007/s10592-010-0133-5

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., & Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), 3766–3779. https://doi.org/10.1111/mec.12360

Gautier, M., Vitalis, R., Flori, L., & Estoup, A. (2022). *F* -Statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package *poolfstat*. *Molecular Ecology Resources*, *22*(4), 1394–1416. https://doi.org/10.1111/1755-0998.13557

Haller, B. C., & Messer, P. W. (2019). *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

Hasselmann, M., Ferretti, L., & Zayed, A. (2015). Beyond fruit-flies: Population genomic advances in non-*Drosophila* arthropods. *Briefings in Functional Genomics*, *14*(6), 424–431. https://doi.org/10.1093/bfgp/elv010

Hoban, S. (2014). An overview of the utility of population simulation software in molecular ecology. *Molecular Ecology*, *23*(10), 2383–2401. https://doi.org/10.1111/mec.12741

Jombart, T. (2008). *adegenet*: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Kapun, M., Barrón, M. G., Staubach, F., Obbard, D. J., Wiberg, R. A. W., Vieira, J., Goubert, C., Rota-Stabelli, O., Kankare, M., Bogaerts-Márquez, M., Haudry, A., Waidele, L., Kozeretska, I., Pasyukova, E. G., Loeschcke, V., Pascual, M., Vieira, C. P., Serga, S., Montchamp-Moreau, C., … González, J. (2020). Genomic Analysis of European *Drosophila melanogaster* Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Molecular Biology and Evolution*, *37*(9), 2661–2678. https://doi.org/10.1093/molbev/msaa120

Kapun, M., Nunez, J. C. B., Bogaerts-Márquez, M., Murga-Moreno, J., Paris, M., Outten, J., Coronado-Zamora, M., Tern, C., Rota-Stabelli, O., García Guerreiro, M. P., Casillas, S., Orengo, D. J., Puerma, E., Kankare, M., Ometto, L., Loeschcke, V., Onder, B. S., Abbott, J. K., Schaeffer, S. W., … Bergland, A. O. (2021). *Drosophila* Evolution over Space and Time (DEST): A New Population Genomics Resource. *Molecular Biology and Evolution*, *38*(12), 5782–5805. https://doi.org/10.1093/molbev/msab259

Karasov, T., Messer, P. W., & Petrov, D. A. (2010). Evidence that Adaptation in *Drosophila* Is Not Limited by Mutation at Single Sites. *PLoS Genetics*, *6*(6), e1000924. https://doi.org/10.1371/journal.pgen.1000924

Kirkpatrick, D. M., Leach, H. L., Xu, P., Dong, K., Isaacs, R., & Gut, L. J. (2018). Comparative Antennal and Behavioral Responses of Summer and Winter Morph *Drosophila suzukii* (Diptera: Drosophilidae) to Ecologically Relevant Volatiles. *Environmental Entomology*, *47*(3), 700–706. https://doi.org/10.1093/ee/nvy046

Kirkpatrick, M., & Jarne, P. (2000). The Effects of a Bottleneck on Inbreeding Depression and the Genetic Load. *The American Naturalist*, *155*(2), 154–167. https://doi.org/10.1086/303312

Kovach, M., & McCouch, S. (2008). Leveraging natural diversity: Back through the bottleneck. *Current Opinion in Plant Biology*, *11*(2), 193–200. https://doi.org/10.1016/j.pbi.2007.12.006

Lange, J. D., Bastide, H., Lack, J. B., & Pool, J. E. (2022). A Population Genomic Assessment of Three Decades of Evolution in a Natural *Drosophila* Population. *Molecular Biology and Evolution*, *39*(2), msab368. https://doi.org/10.1093/molbev/msab368

Lynch, M., Conery, J., & Bürger, R. (1995). Mutational Meltdowns In Sexual Populations. *Evolution*, *49*(6), 1067–1080. https://doi.org/10.1111/j.1558-5646.1995.tb04434.x

Machado, H. E., Bergland, A. O., Taylor, R., Tilk, S., Behrman, E., Dyer, K., Fabian, D. K., Flatt, T., González, J., Karasov, T. L., Kim, B., Kozeretska, I., Lazzaro, B. P., Merritt, T. J. S., Pool, J. E., O'brien, K., Rajpurohit, S., Roy, P. R., Schaeffer, S. W., … Petrov, D. A. (2021). Broad geographic sampling reveals the shared basis and environmental correlates of seasonal adaptation in *Drosophila*. *eLife*, *10*, 1–21. https://doi.org/10.7554/eLife.67577

Maron, M., McAlpine, C. A., Watson, J. E. M., Maxwell, S., & Barnard, P. (2015). Climate-induced resource bottlenecks exacerbate species vulnerability: A review. *Diversity and Distributions*, *21*(7), 731–743. https://doi.org/10.1111/ddi.12339

Morgulis, A., Gertz, E. M., Schäffer, A. A., & Agarwala, R. (2006). A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *Journal of Computational Biology*, *13*(5), 1028–1040. https://doi.org/10.1089/cmb.2006.13.1028

Nunez, J. C. B., Lenhart, B. A., Bangerter, A., Murray, C. S., Mazzeo, G. R., Yu, Y., Nystrom, T. L., Tern, C., Erickson, P. A., & Bergland, A. O. (2024). A cosmopolitan inversion facilitates seasonal adaptation in overwintering *Drosophila*. *Genetics*, *226*(2), iyad207. https://doi.org/10.1093/genetics/iyad207

Pool, J. E. (2015). The Mosaic Ancestry of the *Drosophila* Genetic Reference Panel and the *D. melanogaster* Reference Genome Reveals a Network of Epistatic Fitness Interactions. *Molecular Biology and Evolution*, msv194. https://doi.org/10.1093/molbev/msv194

Rosenberg, Y., Bar-On, Y. M., Fromm, A., Ostikar, M., Shoshany, A., Giz, O., & Milo, R. (2023). The global biomass and number of terrestrial arthropods. *Science Advances*, *9*(5), eabq4049. https://doi.org/10.1126/sciadv.abq4049

Rossi-Stacconi, M. V., Kaur, R., Mazzoni, V., Ometto, L., Grassi, A., Gottardello, A., Rota-Stabelli, O., & Anfora, G. (2016). Multiple lines of evidence for reproductive winter diapause in the invasive pest *Drosophila suzukii*: Useful clues for control strategies. *Journal of Pest Science*, *89*(3), 689–700. https://doi.org/10.1007/s10340-016-0753-8

Saubin, M., Tellier, A., Stoeckel, S., Andrieux, A., & Halkett, F. (2023). Approximate Bayesian Computation applied to time series of population genetic data disentangles rapid genetic changes and demographic variations in a pathogen population. *Molecular Ecology*, mec.16965. https://doi.org/10.1111/mec.16965

Shafer, A. B. A., Gattepaille, L. M., Stewart, R. E. A., & Wolf, J. B. W. (2015). Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: In silico evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular Ecology*, *24*(2), 328–345. https://doi.org/10.1111/mec.13034

Shearer, P. W., West, J. D., Walton, V. M., Brown, P. H., Svetec, N., & Chiu, J. C. (2016). Seasonal cues induce phenotypic plasticity of *Drosophila suzukii* to enhance winter survival. *BMC Ecology*, *16*(1), 11. https://doi.org/10.1186/s12898-016-0070-3

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. *PLoS Computational Biology*, *9*(1), e1002803. https://doi.org/10.1371/journal.pcbi.1002803

Taus, T., Futschik, A., & Schlötterer, C. (2017). Quantifying Selection with Pool-Seq Time Series Data. *Molecular Biology and Evolution*, *34*(11), 3023–3034. https://doi.org/10.1093/molbev/msx225

Appendix

Re-evaluating the evidence for a universal genetic boundary among microbial species.

Connor S. Murray[1], Yingnan Gao[1], Martin Wu[1]

[1] Department of Biology, University of Virginia, Charlottesville, VA, USA

Arising from Jain et al. *Nature Communications* https://doi.org/10.1038/s41467-018-07641-9 (2018)

Published in *Nature Communications*: Murray, C.S., Gao, Y. & Wu, M. Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nature Communications* 12, 4059 (2021). https://doi.org/10.1038/s41467-021-24128-2

A fundamental question in studying microbial diversity is whether there is a species boundary and if the boundary can be delineated by a universal genetic discontinuity. To address this question, Jain et al. computed the pairwise average nucleotide identity (ANI) of 91,761 microbial (bacterial and archaeal) genomes (the 90K genome dataset) and found that the ANI values from the 8 billion comparisons follow a strong bimodal distribution with a wide gap between 83 and 95%[1]. As a result, the authors concluded that a clear genetic discontinuum and species boundary were evident from the unprecedented large-scale ANI analysis, and claimed that “it (the 95% ANI threshold) represents an accurate threshold for demarcating almost all currently named prokaryotic species”. We argue that the paper’s conclusion of a universal genetic boundary among named species in the current NCBI taxonomy is questionable and resulted from the substantial biased sampling in genome sequencing, and caution against being overly confident in using 95% ANI for microbial species delineation as the high benchmarks reported in the paper were inflated by using highly redundant genomes.

To demonstrate our point, we first show that biased within-species sampling can generate the bimodal distribution of ANI observed in the original paper even when the speciation rate is constant and the genetic diversity is continuous. We simulated continuous genetic diversity using a 3000-tip phylogenetic tree that diversifies at a constant rate, which we estimated from a genome tree of 3000 bacterial genomes that represents the phylogenetic diversity in the 10,616 NCBI RefSeq complete bacterial genomes (the 10K genome dataset). We then calculated ANI between tips using a function that accurately captures the relationship between ANI and the branch length in the real data (Supplementary Fig. 1). Because the diversification rate is constant, the branch lengths follow an exponential distribution expected from a Poisson process. As expected, the frequency of ANI declined monotonically when ANI increased (Fig. 1a). However, when only 30 tips (1% of the tips) were sampled with a protocol that emulated within-species sampling bias (each tip has two very closely related genomes

sequenced), the ANI distribution became bimodal (Fig. 1b). Although our simulation using a simplistic model does not disprove the existence of a universal genetic boundary, it demonstrates the possibility that limited within-species sampling bias alone can create the bimodal distribution when genetic diversity is continuous.

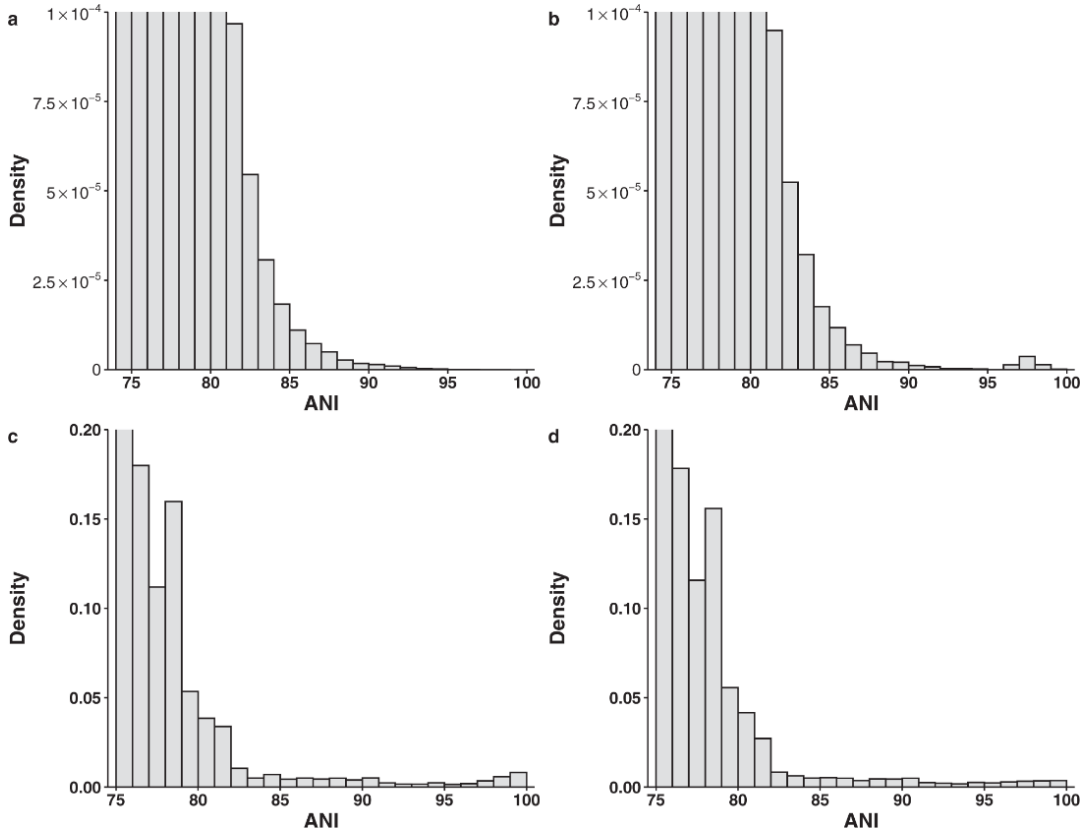


Figure 1: Top panel: from the phylogenetic simulations. a Comparisons between 3000 taxa simulated using a phylogenetic tree with 3000 tips and a constant rate of diversification. **b** Comparisons in the same dataset except that 30 of 3000 taxa each have two very closely related genomes sequenced. Bottom panel: from genomes subsampled from the 90K genome dataset. **c** Two genomes were randomly selected from each of the 397 named species with ≥10 genomes. **d** Two phylogenetic representative genomes were selected for each of the 397 named species with ≥10 genomes.

In the original 90K genome dataset, 33% of named species have been sequenced at least twice. As cultivation bias is widespread and strains of medical and economic interest are heavily favored in our genome sequencing efforts, next we show that there is substantial within-species sampling bias in the genome datasets. For the model organism Escherichia coli, its 602 complete genomes in the 10K genome dataset only represent 22% of the diversity captured by the 16S rRNA gene in the GreenGene database (Supplementary Fig. 2). For species with at least 100 genomes, on average the first 75.8% (range: 50.9–97.7%) of dropped genomes contribute <5% of the genetic diversity of the species as measured by the branch length (Supplementary Fig. 3). As demonstrated in our phylogenetic simulation, these highly redundant genomes can create the bimodal distribution even when genetic diversity is continuous. Contrary to the authors' claim and from a purely statistical point of view, randomly subsampling (e.g., sampling five genomes from the same species as done in the original study[1]) from a biased dataset will not correct for the pre-existing sampling bias. When two genomes were sampled randomly from each of the 397 named species with ≥10 genomes in the 90K dataset, the bimodal distribution was evident (Fig. 1c). However, when we reduced the within-species sampling bias by selecting two phylogenetically representative genomes from the same dataset, the distribution flattened near the end (Fig. 1d), further demonstrating that the bimodal distribution can be caused by the widespread sampling bias within species. In fact, the within-species sampling bias has increased over time when more strains were sequenced based on their medical and economic relevance and not on their phylogenetic positions (Supplementary Fig. 4), thereby producing the consistent bimodal distributions of ANI over different periods of time as observed in the original study[1].

The authors indicated that only 0.2% of the 8 billion ANI values are between 83 and 95% and suggested that such a wide gap is evidence of a clear genetic boundary in microbial genomes. In our phylogenetic simulation of 3000 genomes with continuous genetic diversity,

only 0.11% of ANI values span the same region. This is expected because without biased sampling, the vast majority of all pairwise comparisons for a genome will be with distantly related genomes, whose number will always be much larger than the number of closely related taxa. As a result, the fraction of ANI values in the intermediate range [83–95%] will always be marginal for a decently sized tree (>50 tips) and will decrease when the number of genomes increases (Supplementary Fig. 5). Our result shows that the low density of ANI values in the gap region does not necessarily indicate a genetic boundary. Instead, it could simply reflect the hierarchical structure of the phylogenetic relationship. Consistent with our finding, a recent analysis of ~150,000 bacterial and archaeal genomes shows the interspecies ANIs between closest representatives within a genus are nearly evenly distributed between 78 and 95% ANI and there is no genetic discontinuum in this region[2].

Previous studies have advocated using an ANI cutoff to demarcate bacterial species[3,4,5,6]. Based on the unprecedented large-scale ANI analysis of the study, Jain et al. claimed that using the 95% ANI criterion led to both high recall and precision rates (>98.5%) in demarcating named prokaryotic species in the current NCBI taxonomy[1]. However, both benchmarks were calibrated with ~78k named genomes that were highly redundant. For example, the dataset contained more than 5000 *E. coli* genomes. As such, the recall and precision rates can be misled by the overly sampled genomes. To better assess the performance of using 95% ANI for species demarcation, we queried the 3000 representative genomes against the 10K dataset. We found the recall and precision rates were much lower, at 73.4% and 83.3% respectively. The large impact of the extreme sampling bias on the benchmarks is best illustrated by the authors' own finding that excluding the *E. coli* vs. *Shigella spp*. comparison alone substantially increased their overall precision from 93.1 to 98.7%. Plotting of the intraspecific ANI for intensely sequenced species in the 90K dataset (61 named species with ≥100 genomes, representing 6 phyla) shows that a universal boundary at 95% ANI

clearly does not exist in these species, as ANI values drop well below 95% with few exceptions (Fig. 2). The high density above the 95% threshold could be an artifact caused by comparing highly redundant genomes within the same species.

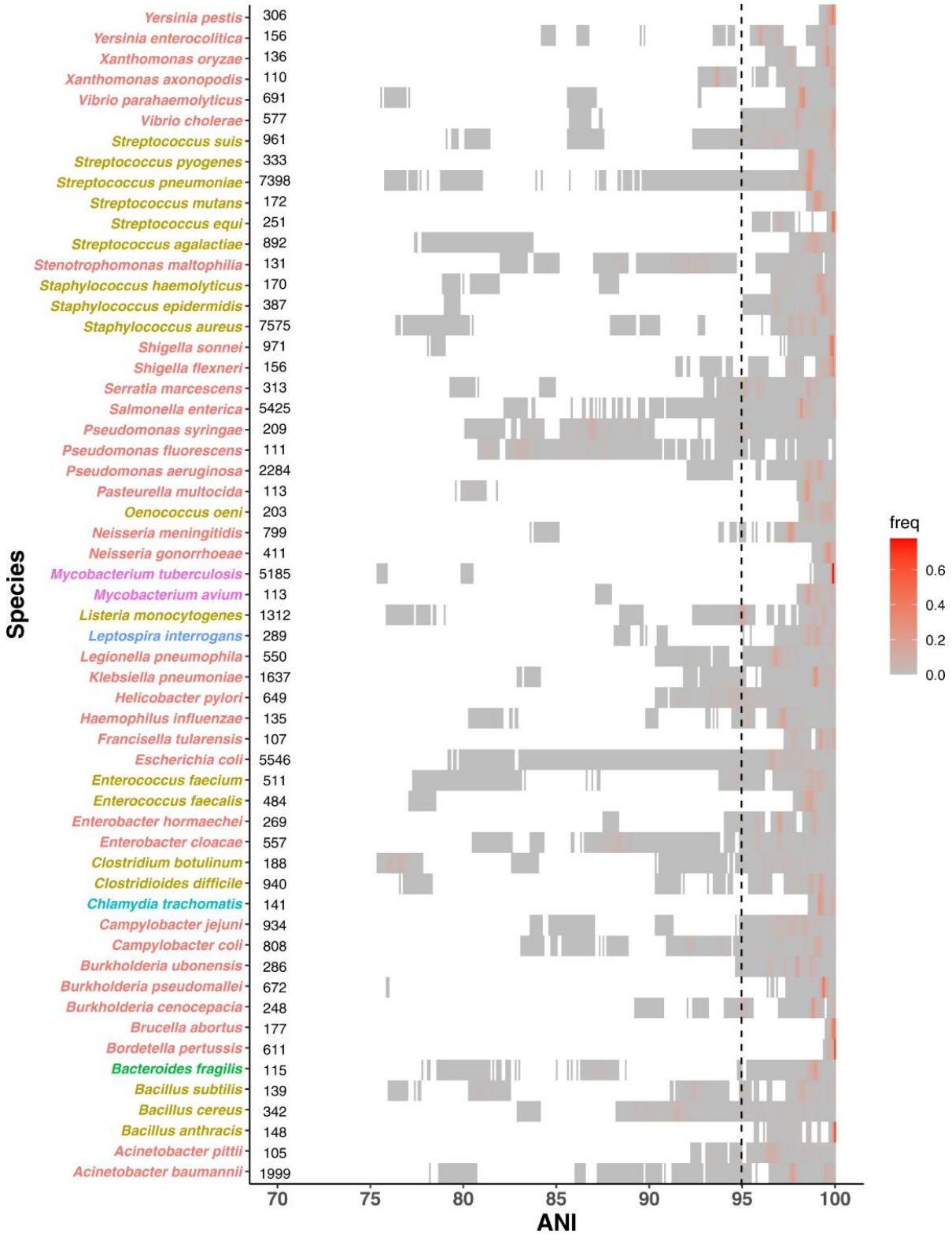


Figure 2: Each block represents a bin of ANI values between pairs of genomes within the same species and is colored by the ANI density in the bin. Species are colored by phylum. The

numbers next to the species names are the numbers of genomes in the 90K dataset for the species. The vertical dotted line represents the 95% ANI threshold.

Although having no cultivation bias, metagenome assemblies do favor abundant strains over rare ones. This bias can lead to the appearance of genetic clusters when diverse but rare strains in the community are excluded from the assembled genomes. In addition, a rare strain in one environment can be abundant in another. Furthermore, even if a genetic cluster does exist in one environment, it does not necessarily delineate the genetic boundary of the species in general. For example, various *E. coli* strains spanning a continuum of diversity live in the gut, water, and soil[7]. Analyzing gut metagenomic data alone will most likely reveal a tight genetic cluster of *E. coli*. However, this genetic cluster does not represent the genetic boundary of *E. coli* as a species because it excludes the environmental *E. coli* strains. It can be argued that *E. coli* is an exception, but it is also well known that many bacterial species are “generalists” that live in a wide range of habitats. One potential solution is to narrow down our species definition to accommodate the local genetic clusters, but doing so will require substantially overhauling the current taxonomy and change the subject of this debate, the named species in the current taxonomy. Unless low abundance strains are readily recovered (e.g., through read recruitment to a reference genome) and metagenomic sequences from different types of environments are compared, there are also potential pitfalls associated with demonstrating the genetic boundary of currently named species using metagenome-assembled genomes[8]. Interestingly, several metagenomic studies have revealed genetic continuum in nature[9,10,11].

There is much evidence against the existence of a universal genetic boundary for microbial species. First, the molecular substitution rate is highly variable across species. Secondly, selection and recombination are thought to be the main cohesive forces driving the formation of genetic clusters. Although recombination rate can be influenced by sequence similarity, there is no correlation between the recombination rate and ANI in bacteria[12], as

recombination can also be affected by physical and ecological barriers. Microbes living in narrow ecological niches and with limited dispersal rate (e.g., obligate intracellular bacteria) may develop genetic clusters. On the other hand, free living microbes exploring different habitats and mixing by dispersal are more likely to exhibit a genetic continuum[13]. Selection is unlikely to produce a universal genetic boundary either, as microbial species are unique in nature, with each species subject to its own evolutionary and ecological forces[14].

In summary, our study shows that the genetic boundary perceived in the original paper can be explained by persistent within-species sampling bias from historic and current genome sequencing efforts. A more balanced analysis of the present genomic data shows that although genetic clusters may exist in individual species, we find no evidence of a universal genetic boundary among named microbial species in the most recent NCBI taxonomy.

Methods

Genome datasets. Two genome datasets were used in this study. The first is the 90K genome dataset from the original paper1. It contains both complete and draft bacterial and archaeal genomes. The second dataset consists of 10,616 complete bacterial genomes downloaded from the NCBI RefSeq database on September 6, 2018 (10K genome dataset). From each genome in the 10K dataset, we identified 31 universal protein-coding marker genes using AMPHORA2[15] and constructed a bacterial genome tree based on the concatenated and trimmed protein sequence alignment of the marker genes using FastTree16. Treemmer (version 0.3)[17] was used to choose 3000 representative genomes that maximized the phylogenetic diversity in the 10K genome dataset.

Average nucleotide identity (ANI). The ANI values for the 90K genome dataset were downloaded from the original study. For the 10K genome dataset, the 3000 representative genomes were compared against the full 10K dataset using FastANI (version 1.2).

Modeling the relationship between branch length and ANI. The 3000 representative genomes were used to model the empirical relationship between ANI and branch length. The median of ANI was calculated across binned branch lengths (bin width: 0.05 substitution/site) to use as the actual data to fit the relationship between ANI and branch length l through the function, where k, s and α are shape parameters to be estimated. Minimization of the sum of squares error was performed using the optim function in *R*. The best fit parameters for our data are $\alpha = 0.075$, $k = 73.94$, and $s = 0.63$. Branch lengths >2.5 substitutions/site were removed because of the lack of data points.

Simulation of continuous genetic diversity and biased within-species sampling. The *rtree* function in the *ape* package in *R* was used to simulate a random phylogenetic tree of 3000 tips, with its branch lengths following an exponential distribution with a constant rate of 19.2, estimated from the genome tree of the 3000 representative genomes. Using the formula described above, the ANI value between a pair of genomes was computed from the branch length between them. To simulate biased sampling within species, a random tip was chosen and two descendants were added to that tip, with the branch length from the tip to the descendant sampled from the same exponential distribution, but its value restricted to the bottom 1% of the distribution. This procedure was repeated on the remaining 2999 tips until n tips were processed. Each simulation was run with ten replicates.

Assessing the within-species sampling bias. Species with ≥10 genomes were selected from the 10K genome dataset. For each species, a subtree compiling the respective genomes was extracted from the full phylogeny of 10,616 genomes and *Treemmer* was used to iteratively remove one tip of the tip-pair with the shortest branch length until three tips remained. The remaining total branch length of the tree was divided by the total branch length of the initial tree to calculate the relative tree length at each iteration. *Rickettsia japonica* and *Chlamydia muridarum* were removed from this analyses because their genomes have identical marker sequences and branch lengths equal to zero.

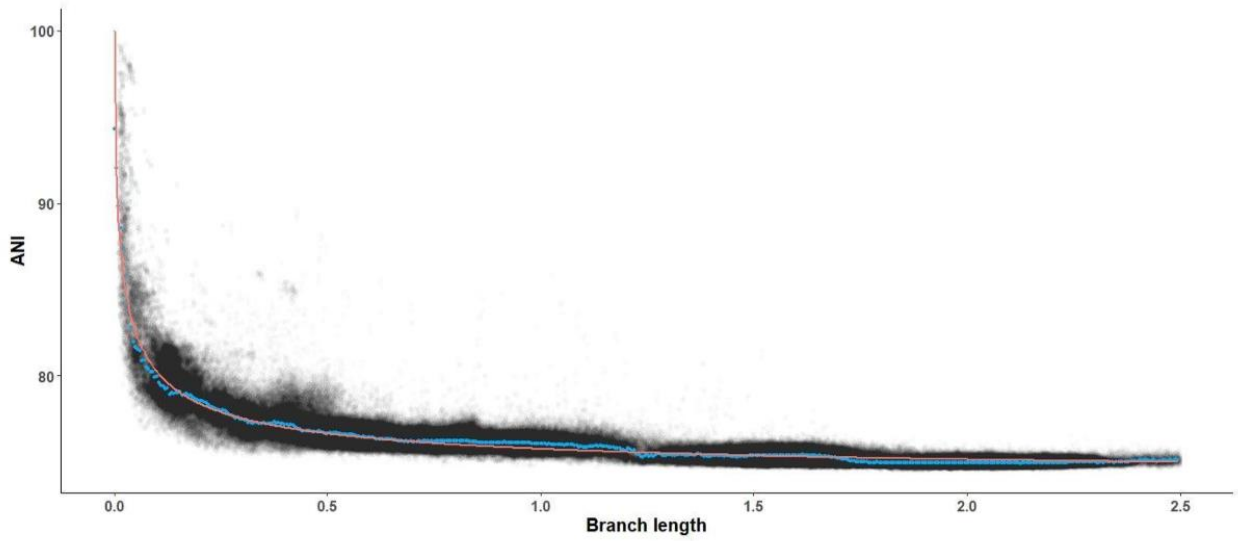
To further evaluate the within-species sampling bias, we tested how much known genetic diversity is recovered by the complete genomes, using *E. coli* as an example. We extracted 868 unique 16S rRNA gene sequences with no ambiguous bases from 602 complete *E. coli* genomes in the 10K genome dataset and BLAST searched them against 8655 *E. coli* 16S rRNA gene sequences from the GreenGene 13.8 database. A match was defined as a pair of sequences with 100% identity for their entire sequences. The matched GreenGene 16S rRNA sequences were then mapped to the 99% OTUs (operational taxonomic units) of the GreenGene database. The total branch length covered by the mapped OTUs in the 16S rRNA tree of 44 *E. coli* 99% OTUs was calculated to estimate the coverage of *E. coli* diversity by complete genomes.

Subsampling of two genomes. For named species with ≥10 genomes in the 90K dataset, two genomes were sampled from each species either randomly or by selecting the pair with the lowest ANI value. Among all pairwise comparisons within the species, the pair with the lowest ANI best represents the phylogenetic diversity of the species.

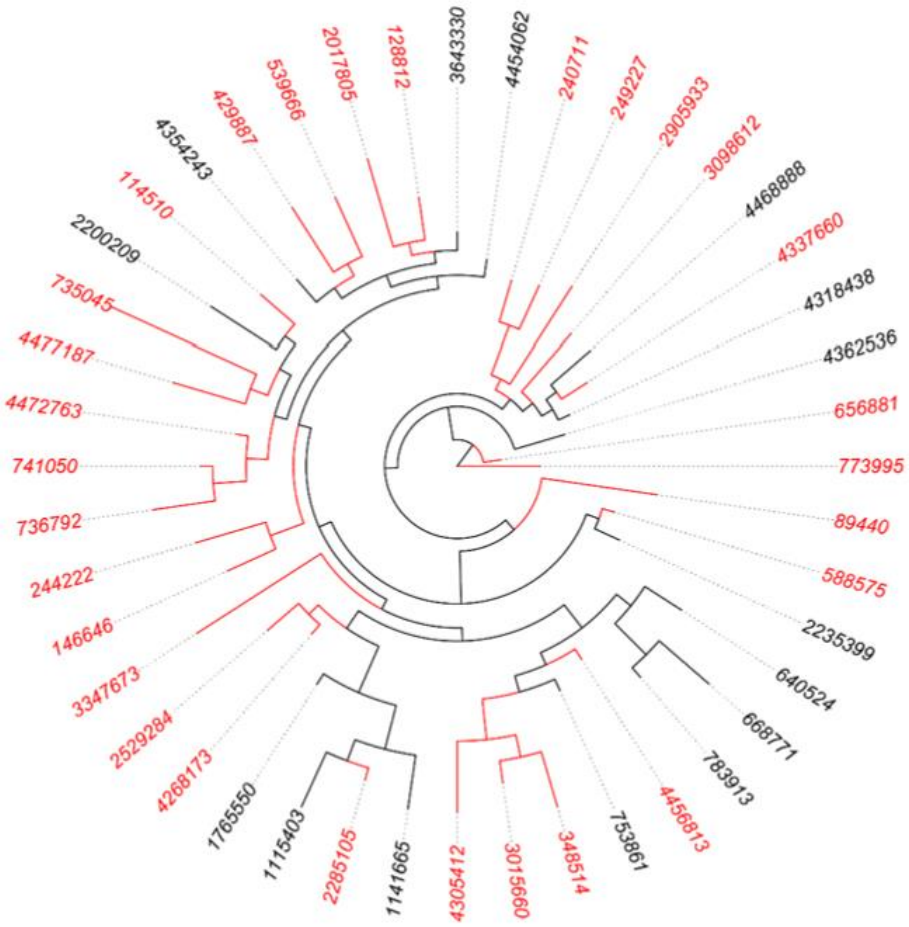
Benchmark the performance using 95% ANI for species demarcation. Using the 3000 representative genomes as the query, we ran FastANI against the full 10K genome dataset. For each query genome, the subject genome with the maximum ANI value was used to benchmark the performance of using the 95% ANI threshold to demarcate bacterial species. A true positive is a query-subject pair belonging to the same species and having an ANI ≥ 95%. A false positive is a genome pair of different species with an ANI ≥ 95%, and a false negative is a genome pair of the same species with an ANI < 95%. Precision was calculated by: the number of true positive/(number of true positive + number of false positive) and recall was calculated by: the number of true positive/(number of true positive + number of false negative).

Data availability. Genome sequences were downloaded from the NCBI RefSeq Database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq). FastANI values were downloaded from the server listed in the original paper. All the other data are available for download at https://github.com/wu-lab-uva/FastANI-Rebuttal.

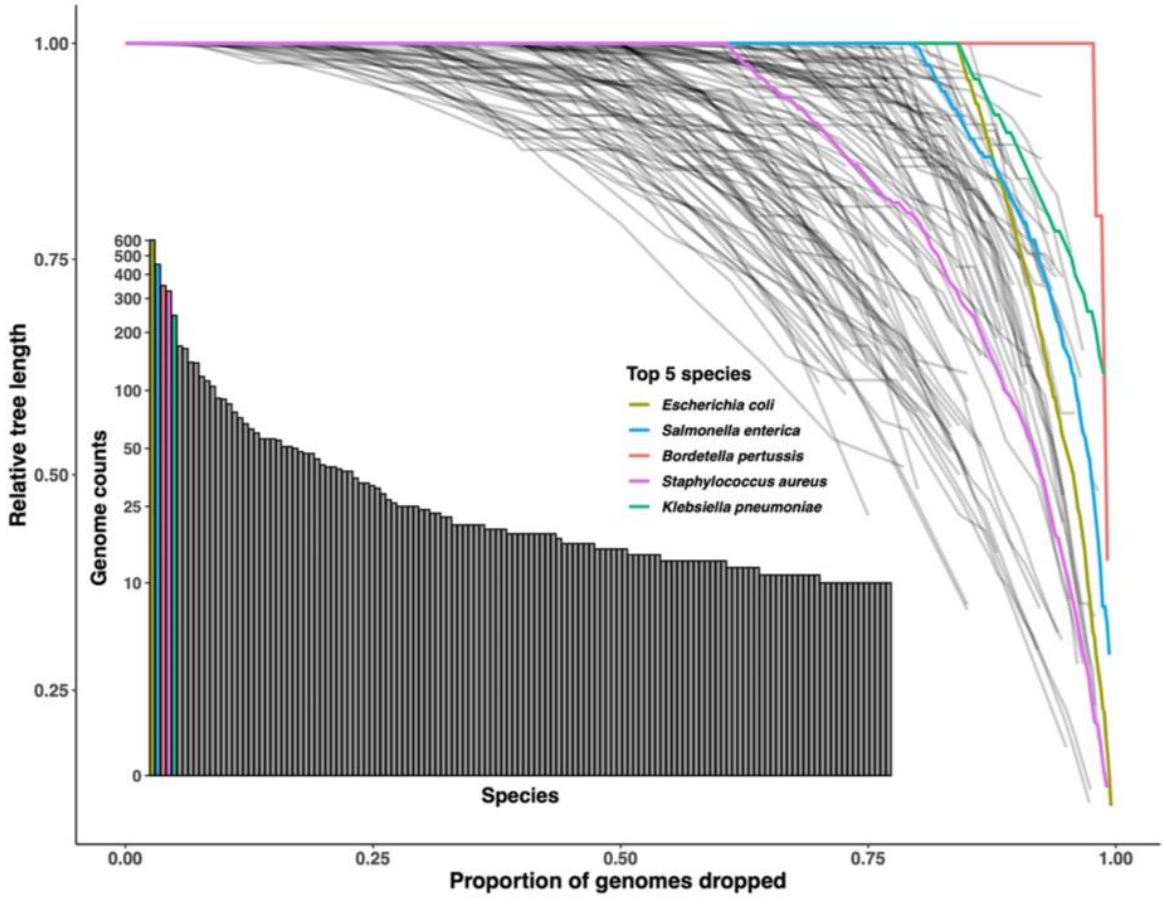
Code availability. All *R* code is available for download at https://github.com/wu-lab-uva/FastANI-Rebuttal.



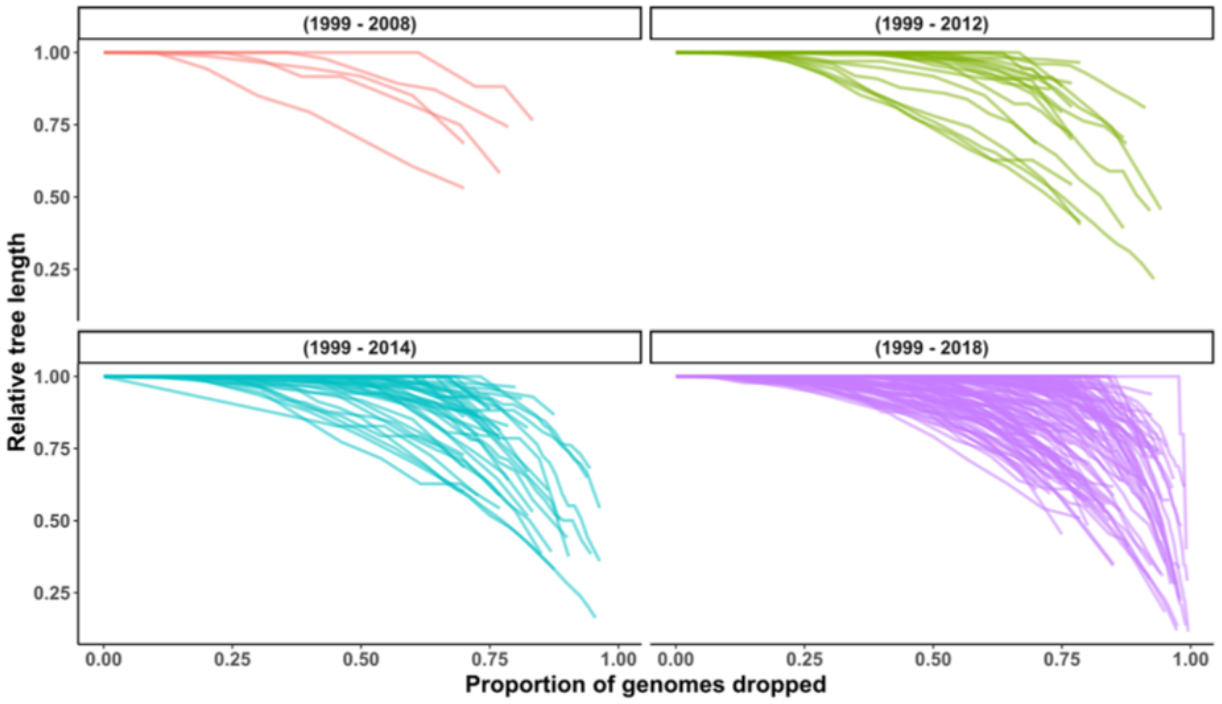
Supplementary Figure 1. The relationship between the branch length and ANI. Shaded black points are pairwise comparisons between the 3,000 representative genomes. Blue points are the median ANI for each branch length bin (bin width: 0.05 substitution/site). Orange line is the fitted curve. Branch lengths greater than 2.5 were removed because of the lack of data points.



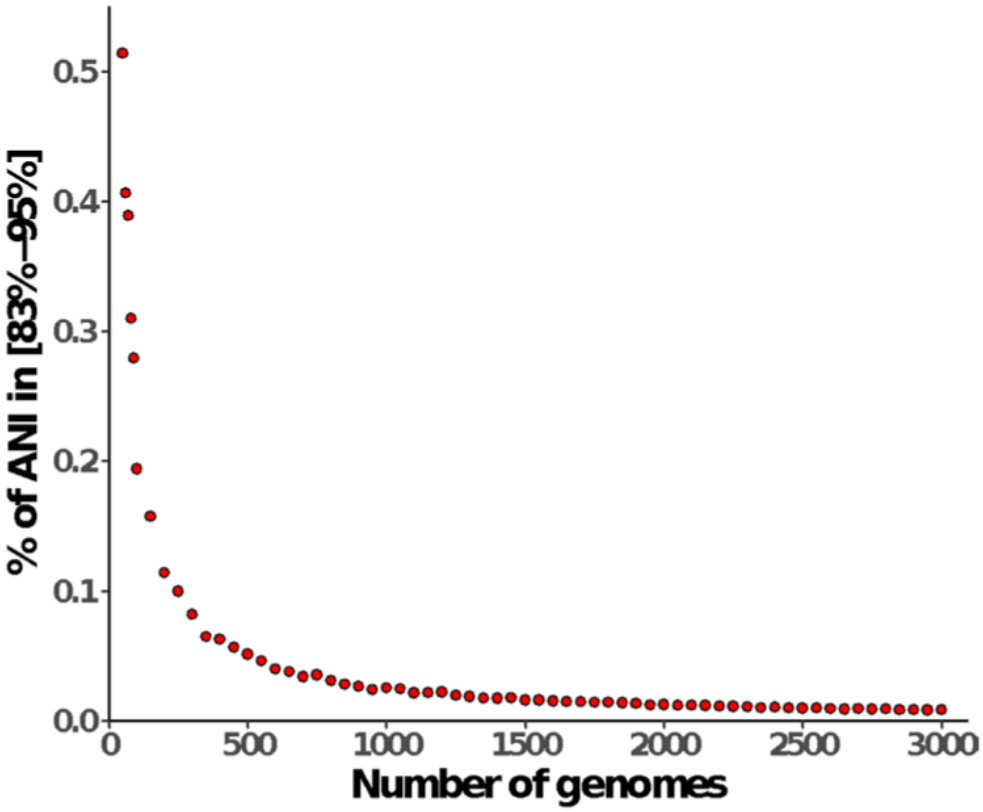
Supplementary Figure 2. Incomplete and biased coverage of *E. coli* diversity by complete genomes. Of 44 *E. coli* 99% OTUs in the phylogeny of the 16S rRNA gene, only 15 (in black) are represented by the 602 complete *E. coli* genomes in the 10K genome dataset, and only 22% of the total branch length in the phylogeny is covered by these 15 OTUs. The branch length was square-root transformed to better show the short branches in the figure.



Supplementary Figure 3. Widespread within-species sampling bias revealed by the Treemmer analysis. Treemmer iteratively drops genomes contributing the least amount of diversity within a phylogenetic tree until only three genomes remain. For each species, the relative tree length (RTL) is plotted against the proportion of genomes dropped at each iteration. Colored lines are the 5 most sequenced species in the 10K genome dataset. Shaded black lines are species with ≥10 genomes. The bar chart shows species ranked by the number of genomes sequenced.



Supplementary Figure 4. Relative tree length (RTL) decay for species with ≥ 10 sequenced genomes in the NCBI RefSeq database in four different time periods.



Supplementary Figure 5. The percentage of ANI values in the [83%-95%] range is marginal and has a negative relationship with the number of genomes. A random phylogeny with an exponential distribution of branch lengths was used to simulate continuous genetic diversity across genomes. The pairwise ANI values were calculated from the phylogenetic distances between genomes (see materials and methods). The simulations were run with phylogenies of different sizes ranging from 50 to 3,000 tips.

Author contributions. M.W. conceived the design of the response, C.S.M., Y.G. and M.W. conducted the analysis, M.W. wrote the first draft and C.S.M., Y.G. and M.W. edited the final version.

Competing interests. The authors declare no competing interests.

Additional information. The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-24128-2.

References

1. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 5114 (2018).
2. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 38, 1079–1086 (2020).
3. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl Acad. Sci. U. S. A.* 102, 2567–2572 (2005).
4. Kim, M., Oh, H. S., Park, S. C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 64, 346–351 (2014).
5. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl Acad. Sci. U. S. A.* 106, 19126–19131 (2009).
6. Varghese, N. J. et al. Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* 43, 6761–6771 (2015).
7. Luo, C. et al. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl Acad. Sci. U. S. A.* 108, 7200–7205 (2011).

8. Olm, M. R. et al. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems* 5, e00731–19 (2020).
9. Hallam, S. J. et al. Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc. Natl Acad. Sci. U. S. A*. 103, 18296–18301 (2006).
10. Caro-Quintero, A. & Konstantinidis, K. T. Bacterial species may exist, metagenomics reveal. *Environ. Microbiol.* 14, 347–355 (2012).
11. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176, 649–662 (2019).
12. Bobay, L.-M. & Ochman, H. Biological species are universal across life's domains. *Genome Biol. Evol.* 9, 491–501 (2017).
13. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. B: Biol. Sci*. 361, 1929–1940 (2006).
14. Palmer, M., Venter, S. N., Coetzee, M. P. A. & Steenkamp, E. T. Prokaryotic species are *sui generis* evolutionary units. *Syst. Appl. Microbiol*. 42, 145–158 (2019).
15. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2*. Bioinformatics* 28, 1033–1034 (2012).
16. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5, e9490 (2010).
17. Menardo, F. et al. Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinforma.* 19, 164 (2018).