UNIVERSITY OF CALIFORNIA

Los Angeles

Factors of Diversity:

Studies in Graduate Education, Recombination, and the Distribution of Fitness Effects

A dissertation submitted in partial satisfaction of the

Requirements for the degree Doctor of Philosophy

in Biology

By

Christina Alicia Del Carpio

2023

ABSTRACT OF THE DISSERTATION

Factors of Diversity:

Studies in Graduate Education, Recombination, and the Distribution of Fitness Effects

By

Christina Alicia Del Carpio

Doctor of Philosophy in Biology

University of California, Los Angeles, 2023

Professor Kirk Edward Lohmueller, Chair

Most, if not all, institutions benefit from diverse populations. The benefits of diversity are also seen in non-human populations, including on the levels of genomes. My three dissertation projects all touch on processes that can influence diversity in some capacity. My first chapter describes the effectiveness of the University of California, Los Angeles (UCLA) Competitive Edge (CE) bridge program in supporting PhD students from historically excluded and underrepresented groups (URG). Through surveys of 55+ first-year students, my study reveals CE's success in enhancing key aspects such as mentoring relationships, socialization, and overall preparedness. These are all crucial factors influencing student retention. Moving into the genetic sphere, my second chapter examines meiotic recombination, a process generating genetic diversity in sexually reproducing species through the decoupling of alleles. Focusing on the impact of

domestication on recombination rates, my study uses wolves and breed dogs as a model. I tested and rejected the hypothesis that domestication leads to increased recombination rates. My work shows intriguing patterns emerge. For example, border collies exhibit higher inferred recombination rates, while pugs have lower rates compared to wolves. Despite these differences, I estimated a stable recombination landscape in dogs and wolves. Additionally, my work created a genetic map for further exploration of the canid genome. My final chapter employs this genetic map to investigate the role of recombination in inferring the distribution of fitness effects (DFE) of new mutations. The DFE is a crucial aspect in population genetics because it reveals how selection impacts genetic diversity. However, Poisson Random Field (PRF) methods of DFE inferences assume sites are unliked. My work investigates how estimates of the DFE vary with linkage and recombination rates in wolves. I find that estimating the DFE in low recombination regions is similar to high recombination regions, despite differences in patterns of linked selection. Thus, my results suggest that DFE inference using PRF methods is not significantly biased by linked selection. Altogether, these results have implications for mechanisms that influence diversity in their relevant contexts.

The dissertation of Christina Alicia Del Carpio is approved.

Diana E. Azurdia

Nandita Garud

Paul Henry Barber

Pleuni S. Pennings

Kirk Edward Lohmueller, Committee Chair

University of California, Los Angeles

2023

Dedication

This one is finally for you, Mom

TABLE OF CONTENTS

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 1

Chapter 2

Chapter 3

Supplementary Materials

Chapter 1

Ch1_Supplementary_Information.pdf (File Format .pdf)

Chapter 2

Ch2_Supplementary_Information.pdf (File Format .pdf)

Chapter 3

Ch3_Supplementary_Information.pdf (File Format .pdf)

ACKNOWLEDGEMENTS

First, I wish to thank the members of my committee for supporting me in completing a nontraditional Biology dissertation. Kirk, thank you for tireless work in helping me get my research over the finish line. Diana, thank you for being my unofficial committee co-chair and expanding my vision of what my career can be. Paul, my first chapter wouldn't have been possible without you connecting me with Competitive Edge. Nandita, your feedback helped improved my science. Pleuni, your mentorship and advocacy helped me push through at the end of my PhD. Bob, thank you for believing in me from the start.

Thank you to all the members of the Lohmueller and Wayne labs who supported me throughout this 6+ year process. Gabe, I couldn't have survived the ups and downs of my first year without your support and our chats on Janss steps. Chris and Meixi, thank you for being there from the first day we sat outside Bob's door till the very end. Annabel, I'm most grateful for the ways you helped me remember grad school isn't everything with fantastic book recs and a trip the Hollywood Bowl. Jon, thank you for all the live coding sessions you did with me to make my analyses possible. Aina, your support through the difficulties of academic culture helped me feel seen and validated. Eduardo, thanks for your friendly and needed guidance through my final chapter. Jesse and Jazlyn, thanks for being my kayak buddies in Puerto Rico. Izabel, I'm glad you introduced me to GWF. Pedro, thank you for letting me be your mentor and for your dedicated work on our project.

I am grateful for the support of my many academic friends and collaborators beyond the Lohmueller and Wayne labs. Omar, mi capítulo finale no podia posible sin tu tutoría y colaboración. Jenny and Eric thank you for support of me both during and beyond GATP. Casey, your guidance and feedback were critical to completing my CE work. Morgan, thank you for always listening to me, and help me feel more secure and grounded at the end of my PhD. Tessa, I don't know that this PhD would have been possible without your friendship and advice from day one. David and Jennifer, thank you for being my roommates and cheering me up during one of my lowest points in grad school. Britt, thank you for supporting me through my many phases of growth since we were teenagers. Tawni, I could never have dreamed up a better work wife and friend than you.

My group chat with some of my closest grad school friends got me through so much. Evan, you were a great spotter (yawn) who always let me vent. Christiane, I always appreciated your calm and supportive vibes through everything grad school threw at us. Liz, you were an amazing roommate and I love that you still send so many memes that truly are me.

My gender group friends were such a rock in my life. J, thanks for constantly rooting for me against my nemeses in grad school. Natalie, you were always such a real one when it came to talking about the system. Ema, thank you for being the funny one in group and for convincing me to finally read *The Locked Tomb* series. Stephanie, thanks for all the last-minute outings we went on and for being my social science research lifeline.

Olivia and Athena, our weekly pandemic chats helped me survive to the other side of lockdown. Athena, I know I can always count on you to have sound life advice, like not moving the same month I finish this dissertation. Olivia, thank you for being emotionally vulnerable with me and always making me safe to be vulnerable too.

Becky and Kimberly, thank you for all the spoons you shared with me. Kimberly, you see the world in a unique way and I'm grateful for all that I learn from your perspective. Becky, thank you for staying such a close friend through so many years.

My Psi U brothers chat was a constant source of support. Josh, you were such a positive and energetic force. Mandy, you taught me to never utilize when I can use. Chris, thanks for pushing me to finally attend the World Board game Championships and being a part of that amazing 10 days of the year. Alyssa, your witty humor about academia helped me laugh instead of cry. Eng Seng, thank you for letting me tag along for part of The Great Adventure while I finished the first chapter of this document.

I am grateful that family members have always been my cheerleaders. Thank you to Titi Maria and Uncle Harry for supporting my education from a young age. Harry, thank you for letting me shadow you at the veterinarian clinical all those years ago. Titi, thank you for sharing your love and joy with me since my infancy. Thomas, I appreciate all the knowing looks and funny texts we've been able to share in recent years. My father has supported me by always wanting the best for me. This dissertation is dedicated to my mother because she has supported me in literally everything I have done in my life. Thanks, mom for buying me that baseball bat.

Tess, thank you for sticking with me through the final leg of this wacky thing. Your love and unwavering support helped me truly commit to the bit of finishing this PhD.

Lawrence, I sincerely could not have finished this dissertation without you. You have been such a load bearing source of support to me. I'm eternally grateful that you've become one of my best friends over the course of this dissertation. And thank you for fulfilling the critical role of wife who types up my manuscript.

Lastly, thank you to the two research assistants who got me through the late stages of my PhD with their sweet cuddles and absurd antics – Tuca and Jem.

Chapter 1 is a version of a manuscript in preparation for submission to CBE – Life Sciences.

Del Carpio, C. A. & Azurdia, D. E. (2023). UCLA's Competitive Edge Program Provides an Advantage to STEM Doctoral Students from Historically Excluded and Underrepresented Groups. *bioRxiv* (2023): 2023-09.
doi: https://doi.org/10.1101/2023.09.08.555984

Acknowledgements. The authors thank the following individuals for their valuable contributions to this research project. Beverly Yanuaria and Andrew Rameriz for their assistance in collecting program objectives, survey design, and communication to CE students. Dr. Rhiannon Little-Surowski for aiding in survey design and communication to students. Dr. Jaana Juvonen and Stephanie Dolbier for feedback on survey design and

data analysis. Dr. Casey Shapiro for substantial feedback on survey design, data analysis, and manuscript writing. Dr. Greg Payne and Dr. Lawrence Evalyn for valuable feedback on our manuscript. Dr. Gabe Hassler for guidance with statistical analyses. We would like to acknowledge the support of the UCLA Division of Graduate Education. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2034835 (awarded to Author Christina Del Carpio). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Lastly, the authors acknowledge our presence on the traditional, ancestral, and unceded territory of the Gabrielino/Tongva peoples.

Author Contributions. C.A.D.C. and D.E.A. conceived the study. C.A.D.C. collected the data, analyzed the data, and wrote the manuscript with input and edits from D.E.A.

Chapter 2 is a version of a manuscript in preparation for submission to Molecular Biology and Evolution.

Del Carpio, C. A., Perez, P. A., Cavassim, M. I. A., Wayne, R. K., and Lohmueller, K. E. (2023). Exploring the Dynamics of Recombination Rates across *Canis familiaris*.

Acknowledgements. We thank Jesse Garcia and Jonathan Mah for their help in simulating data using msprime. We thank Jazlyn Mooney, Nandita Garud, and Pleuni Pennings for helpful discussion on this manuscript. This material is based upon work

supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2034835 (awarded to C.A.D.C.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. C.A.D.C., M.I.A.C., and K.E.L. were supported by the National Institutes of Health (R35GM119856 to K.E.L.). Lastly, we acknowledge our presence on the traditional, ancestral, and unceded territory of the Gabrielino/Tongva peoples.

Author Contributions. C.A.D.C. and K.E.L. conceived the study. C.A.D.C. and M.I.A.C. analyzed the data with input from K.E.L. and R.K.W. P.A.P. conducted simulations and analyzed the simulation data with input from C.A.D.C. and K.E.L. C.A.D.C. wrote the manuscript with input and edits from all co-authors.

VITA

Education

Duke University, B.S. Biology, with Distinction, 2011

Publications

Del Carpio, C. A., Azurdia, D. E. (2023). UCLA's Competitive Edge Program Provides an Advantage to STEM Doctoral Students from Historically Excluded and Underrepresented Groups. *bioRxiv* (2023): 2023-09. doi: https://doi.org/10.1101/2023.09.08.555984

Del Carpio, C. A., Ford, A. T., Lowell, E. S. H., Ochoa, M. E., & Speck, H. P. (2023). How to diversify your department's seminar series. *Nature ecology & evolution*, 7(5), 637–639.

Lea, A. J., Vockley, C. M., Johnston, R. A., **Del Carpio, C. A.**, Barreiro, L. B., Reddy, T. E., & Tung, J. (2018). Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife*, 7, e37513.

Selected Presentations

2023 Conference talk, Society for the Advancement of Biology Education Research

2022 Conference talk, Population, Evolutionary, and Quantitative Genetics

2022 Conference workshop, Society for the Advancement of Chicanos/Hispanics and Native Americans in Science (SACNAS) Diversity in STEM Conference

2021 Conference talk, Evolution

2019 Poster, National Human Genome Research Institute (NGRI) Trainee Meeting

2018 Poster, SACNAS Diversity in STEM Conference

Selected Awards and Honors

- UCLA Student Leadership in Equity, Diversity, and Inclusion Award (2022-2023)
- National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP) Fellow (2017- 2018, 2020-2022)
- UCLA Life Sciences Division Excellence in Promoting Diversity and Inclusion (2021-2022)
- UCLA Ecology and Evolutionary Biology (EEB) Diversity, Equity, and Inclusion Award (2021-2022)
- UCLA Brown Memorial Award for Exceptional Graduate Students (2021-2022)
- UCLA EEB Special Faculty Award for Departmental Service (2021)
- National Institute of Health (NIH) Genomic Analysis Training Program Training Grant 2018-2019, 2019-2020)

Chapter 1:

UCLA's Competitive Edge Program Provides an Advantage to STEM Doctoral Students from Historically Excluded and Underrepresented Groups

In preparation for submission to CBE: Life Science Education

A Supplemental Appendix is available online as this dissertation's Supplementary Materials: Ch1_Supplementary_Information.pdf

Abstract

Universities benefit from recruiting and retaining diverse students, as it leads to more creative and rigorous problem solving. Efforts to improve the diversity, equity, and inclusion of graduate education have historically been focused on recruitment but are now shifting to retaining enrolled students. A low percentage of doctoral students, particularly those from historically underrepresented groups (URGs), complete their PhD within 10 years. To increase support of incoming doctoral students from URGs, UCLA established the Competitive Edge (CE) bridge program. CE provides six weeks of professional development and research training at the start of the PhD program. We surveyed 55+ first-year doctoral students (14 CE students, eight non-CE students from URGs, and 34 well-represented (WR) students) in STEM fields to understand CE's

effectiveness. We found that the CE program aided students in four areas that influence graduate student attrition: mentor/mentee relationship, socialization, finances, and preparedness. At the end of their first academic year, CE students reported that the program helped them in multiple areas relating to student success, such as mental wellbeing and sense of belonging. CE students reported a larger mean growth in seven of eight skills needed in graduate school compared to NonCE URG and WR students. Short answer responses revealed that NonCE students wished for more support in areas covered by the CE program, such as managing advising relationships and protecting mental health. Additionally, CE students received significant funding during the program. The CE program's successful model at UCLA can be adapted to improve support for underrepresented doctoral students at other institutions.

Introduction

Diversity among researchers is known to lead to a diversity of ideas and approaches that in turn leads to more creative and rigorous problem solving. Thus, it is in universities' best interests to recruit and retain diverse students. However, historic efforts in graduate education have been more focused on recruitment and only recently began to shift towards support of matriculated students. In the literature on graduate student education, there are multiple studies focused on the reasons students leave graduate school (*e.g.* Bowlin, Sweat, Watts, & Throne, 2017; Hardré, Liao, Dorri, & Beeson Stoesz, 2019; Maher, Wofford, Roksa, & Feldon, 2017; Meara, Griffin, & Robinson, 2017; Rockinson-Szapkiw, 2019), but the literature is more limited on how

institutions can counteract those factors to retain diverse talent. Some literature that focused on undergraduates explored the role of "bridge programs" to facilitate students in successfully applying to graduate schools (*e.g.* Mccoy, Winkle-wagner, & Winkle-wagner, 2020; Peteet et al., 2016), yet there are limited peer-reviewed publications on doctoral bridge programs that target admitted students. To meet the need for supporting incoming doctoral students from historically excluded and underrepresented groups (URGs), the University of California, Los Angeles (UCLA) has instituted the Competitive Edge (CE) bridge program. The program takes place over six weeks during the summer before typical doctoral programs begin. During this time, students receive professional and "soft" skills development, as well as research training. Additionally, CE gives students a chance to acclimate to graduate school and build community before the start of their Ph.D. program. In this article, we have surveyed 55+ first-year doctoral students in science, technology, engineering, and math (STEM) fields to gain insight into the effectiveness of CE as an approach to support diverse students in their transition to doctoral study.

Why students leave graduate school

Part of the urgency to increase support and retention of students comes from data showing that, after 10 years from starting their graduate program, less than 60% of all doctoral students in all fields complete their PhD (Sowell, Zhang, Redd, & King, 2008). Additional data shows seven year completion rates to be similarly low for Hispanic/Latino students, and the rates for Black/African-American students to be as low as ~50% (Sowell, Allum, & Okahana, 2015). It is also worth noting that other ethnic

and racial minorities were not analyzed by Sowell *et al.* (2015) because there were too few students in the doctoral student body to reach a significant sample size. Increasing retention of historically underrepresented groups is also an important piece in actively creating a balance of people from diverse backgrounds at all levels of the academia hierarchy.

A meta analysis of 79 graduate student attrition studies identified four key themes that repeatedly emerge as reasons students leave: **1)** advisor-advisee relationship, **2)** socialization, **3)** finances, and **4)** preparedness (Bowlin et al., 2017).

Advisor-Advisee Relationship: The impact of the advising relationship on student retention is unsurprising given its central role in the model of graduate school. For example, in most STEM fields, students work within a given lab under the supervision of a principal investigator (PI), particularly so in Biology. (Note the terms PI, mentor, and advisor are typically used interchangeably in STEM PhD programs.) The advisor is charged with overseeing the student's academic and professional growth, but also has a significant role in a student's degree progress and future career of the student through dissertation project assignments, funding, and letters of recommendation. Given the central role of an advisor to a STEM PhD student's work, it follows that having a constructive and supportive advisor-advisee relationship is crucial to a student's experience and chances for success (Weiss, 1981, Girves & Wemmerus, 1988; Lovitts, 2001; Ruud, Saclarides, George-Jackson, & Lubienski, 2018; National Academies of Sciences, Engineering, and Medicine, 2019).

Socialization: Student experiences are also impacted by socialization with peers, faculty, and staff. Peer interactions can provide a sense of support as well as be important in transferring heuristic knowledge about navigating graduate school (Gardner, 2007; Knight, Hall, & Green-Powell, 2014; Padilla, 1999; Tinto, 1982; Lovitts, 2001). Interactions with faculty (beyond just the advisor) contribute to a student's sense of belonging in the broader academic community. Further, faculty and staff interactions shape a student's perception of departmental and institutional support, which can impact retention (Astin, 2014; Bean, 1980; Golde, 2005; Knight et al., 2014; Zhou & Okahana, 2019)

Finances: Financial stability and support is also a core factor in student persistence and eventual completion. STEM PhD students are in a unique situation where their tuition is typically paid for them and they receive a stipend or salary to pay for personal expenses such as housing and meals. This funding may come in the form of employment as a teaching assistant (TA) or research assistant (RA) or as a stipend from a fellowship. But even with these sources of support, there can be great disparities in financial stability depending on amount of support and a student's particular circumstances. Unsurprisingly, the types and amount of funding play a role in doctoral students' decisions to remain in graduate school (Ampaw & Jaeger, 2012; Herman, 2008; Martinez, Ordu, Sala, & McFarlane, 2013)

Preparedness: Graduate study, particularly doctoral research, is a huge undertaking that differs significantly from undergraduate education. Many students enter graduate school without a robust understanding of the expectations thereof and are unaware that the skill set to succeed is very different from that needed in undergraduate studies (Brill, Balcanoff, Land, Gogarty, & Turner, 2014). Students from underrepresented backgrounds may be further disadvantaged with fewer personal connections to graduate students and professors who can share the expectations and "hidden curriculum" of graduate school.

While previous studies have identified these core reasons why graduate students leave, more evidence is needed for student support strategies that can counteract these factors and increase student retention.

Description of the CE program

CE is a bridge program at UCLA that aims to support URG doctoral students in STEM fields during their transition to graduate studies. Incoming students are nominated by their home department, and then a subset of nominees is selected to participate. The program began in 2008 with just a few students per year. However, since 2010, CE cohorts have ranged from 10 to 44 students, with recent cohorts averaging near 40 students. It should be noted that in 2019, the program expanded to include a few social science students, and in 2020 it also included humanities students. But given the history of the program and the primary constituency, this study focuses solely on STEM students. The program was initially funded by a National Science Foundation (NSF)

Alliances for Graduate Education and the Professoriate (AGEP) grant and is currently funded by UCLA's Division of Graduate Education.

CE students participate in an intensive six-week schedule with activities and programming that address a wide variety of skills and support for graduate students. In 2021, the program was conducted remotely on Zoom due to COVID-19 precautions. Students were expected to participate in the program full-time (40 hours a week); eight of those hours consisted of structured professional development / soft skills programming, and the rest of the time was spent performing research. The structured time included workshops on topics such as managing the advising relationship, writing grant and fellowship applications, and managing mental well-being. Students also participated in a journal club for their discipline that met four times over the course of the program. Additionally, there was a session with a panel of past CE student participants and another panel of faculty who were first-generation students. There was also built-in time for students to get to know each other and ask questions of program leaders. The schedule for the 2021 program can be found in supplemental materials (Table S1.1) Outside of structured events, students were expected to begin doing research guided by their advisor. The research aspect of the program could be remote, but some students did their research in person. Additionally, students worked with their advisor to write a research proposal by the end of the CE program that could be used to apply for a fellowship such as the NSF Graduate Research Fellowship Program. Additionally, students were funded during the program with a $6,000 stipend.

Objectives of the CE program

Through the various components, CE seeks to meet several broad learning objectives. Specifically, 1) students will learn how to conduct research, 2) students will learn how to communicate their research with others, 3) students will learn how to work with faculty members, and 4) students will be connected to others at UCLA. We recognized that these four objectives are well-designed to address three of the four primary reasons that students leave doctoral study without a PhD: preparedness, advisor-advisee relationship, and socialization. The unaddressed reason for student attrition is finances. While not a stated goal of the program, CE arguably addresses this area of concern with its stipend and its programming on applying to fellowships and how to budget. Additionally, every program component has specific learning objectives that aim to support students in these four areas (Table S2).

Our study

In this work, we use quantitative and qualitative survey data of 55+ first-year STEM doctoral students to analyze the impact of the CE bridge program on matriculated URG students. We consider multiple measures of student skills and experiences that align with both the program's objectives and known causes of graduate student attrition. Given the alignment between the CE program's objectives and known causes of graduate student attrition, we expect to see better acclimation and possibly retention outcomes for URG doctoral students who participate in the program. Evaluating the impacts of the CE program can determine the effectiveness of program elements as well as provide a basis for iterative improvements. Additionally, others seeking to create

better support for PhD students can learn from any strengths or weaknesses revealed by our analysis of the CE program.

Materials and Methods

This study was done with an approved UCLA Institutional Review Board protocol #21-000756.

Survey Question Development

Our survey instruments (Supplemental Text) were developed with feedback from the individuals administering the 2021 Competitive Edge (CE) program to ensure we measured metrics related to the program goals. The CE program administers its own feedback surveys focused on curricula assessment; these surveys have significantly fewer questions and a narrower scope. We looked at the responses of past CE administered surveys to gain insight into what parts of the program our own survey tool should focus on. In 2021, the program still administered their own feedback survey but it was conducted separately from the survey we use here.

Additionally, we surveyed faculty and staff who led program workshops on their individual learning objectives, and we consulted the program director on what the overall program objectives were. The majority of our instruments consist of questions written for this study that directly address those objectives. Furthermore, we used a previously published set of questions addressing the topic of sense of belonging

(Hermida 2017). The majority of our questions were framed as Likert scale questions asking respondent agreement with a statement ranging from strongly disagree (1) to strongly agree (5). Statements selected reflected the skills and attitudes we hypothesized that CE would increase. For example, one statement read “I can set expectations with my advisor.” Lastly, we collected demographic data with questions modeled closely after the Grad Student Experience in the Research University survey (SERU Consortium 2021).

Study Population

All students surveyed were first-year doctoral students in STEM fields at the University of California, Los Angeles (UCLA). Definition of STEM fields for this study included social science disciplines (Table S1.3). The study population consists of three subgroups referred from here on as 1) Competitive Edge (CE), 2) Non-Competitive Edge Under-Represented Group (NonCE URG), and 3) Non-Competitive Edge Well-Represented (NonCE WR). CE students were those that participated in the 2021 Competitive Edge program. To be eligible for CE, students must be a U.S. citizen, U.S. national, permanent resident, or undocumented student who qualifies for nonresident supplemental tuition exemptions under California law AB 540 (University of California Office of Admissions 2023). Thus we limited our results for our two NonCE cohorts to the same citizenship categories. Additionally CE students must have a background that is underrepresented in graduate education. Students in both NonCE groups did not participate in the program (that year or any other). For NonCE students, URG vs WR students were classified by UCLA’s Graduate Division’s definition of URG students.

Functionally, this designation separated White, Asian, and mixed-race White and Asian students as WR and all others as URG. UCLA identified URG students through self-reported information on graduate school applications. We identified URG students based on self-reported information on our survey. However, we note that this definition of URG does not apply to every STEM field. For example, Asian students are underrepresented in Ecology (Kou-Giesbrecht 2020). Nonetheless, given that the aims of the program are to address the needs of disadvantaged students, particularly those defined by the university as belonging to these racial and ethnic/racial categories, we find it prudent to evaluate the strengths of the program under those same categories. We present broad self-reported demographic information on our respondents in Table 1.1.

Survey Deployment

All surveys were administered via Google forms. Students were incentivized to participate via a raffle for $100 gift cards to Target. In July 2021, we requested the CE program leaders forward our email to the 36 STEM students who made up the 2021 cohort of CE. This was done separately from the CE program's own feedback surveys. We asked CE students to complete the pre-program questions and had 22 respondents. In June 2022, we again had CE program leaders contact the 2021 cohort students to fill out our end of year-one survey (n = 14).

To survey non-CE students, we had student affairs officers for all STEM PhD programs at UCLA contact current first-year students to take our end of year-one non-CE survey.

We also reached out to various graduate student identity-based affinity groups to contact students from racial and ethnic minority backgrounds. According to UCLA STEM Ph.D. program enrollment data, there were 639 first-year PhD students in fall 2021 (UCLA Department of Graduate Education 2023). Of those students, 55% were male, 44% female, and 1% non-binary. Almost a third (31.5%) were international students. UCLA does not release race/ethnicity data for international students. But within domestic students, 20.8% were classified as members of URGs. We received a total of 42 respondents for non-CE students, with eight grouped as from URG backgrounds and 34 grouped from WR backgrounds.

Survey Response Analysis

To quantitatively analyze our results, we asked CE students to rate how the program improved their experiences across five categories during their first year of doctoral study. The five categories we selected are: **1)** research skills, **2)** sense of belonging, **3)** self confidence, **4)** overall well being, and **5)** interactions with advisors. These categories were hypothesized by us, the CE staff, and contributors to be highly impacted by the program. Furthermore, these categories are highly correlated with success in graduate school (Maher et. al 2020, Holmes et. al 2019, Martinez et. al 2013, Brill et. al 2014).

For additional quantitative insights, we focused on skills relevant to student success that were targeted by the CE program. We asked CE and NonCE students to rate themselves on eight skills: **1)** interacting with faculty, **2)** science communication, **3)**

mental wellbeing strategies, **4)** connection to resources, **5)** conducting research, **6)** evaluating journal articles, **7)** financial literacy, and **8)** fellowship application writing. These eight areas were chosen because one or more components of CE focuses on building these skills. For example, students attend a journal club where they learn both how to evaluate literature in their field and how to communicate the findings of those articles. Furthermore, these skills are critical to success in graduate school.

To assess these five experience categories and eight skills, we asked one or more Likert scale questions relating to each one. If multiple questions were incorporated into the score for a specific category or skill, we averaged the answers to all the relevant questions. Additionally, for the eight skills, we calculated students' growth by subtracting how they rated themselves prior to starting CE or doctoral study (pre-score) from self ratings at the end of their first year of graduate school (post-score).

Our end of year-one surveys concluded with three open ended questions. For the CE students, we asked separately about positive impacts and negative impacts of CE on their doctoral study. We also asked for any additional comments about their first year not addressed by the rest of the survey. For the two NonCE groups, we asked about any programs or experiences that positively impacted and then any that negatively impacted their first year. We also asked for additional comments not addressed elsewhere. These qualitative answers were categorized using an inductive coding method (Thomas 2003). To do this, one author read through all the responses before listing possible categories. They then re-read every response and assigned one or more

categories to each response. They consolidated all those themes into 13 broader themes and re-coded all responses with those 13 themes. Those were then collapsed into five final themes: **1)** community support, **2)** financial resources, **3)** mental health, **4)** mentorship/advising, and **5)** skills development (Table S1.4). The first author then labeled all responses with those classifications. The second author reviewed them and agreed they were consistent with the theme definitions. For the additional comment responses, the text segments were classified as positively or negatively impacting student experiences and combined with the appropriate group of responses. We then tabulated the number of times a given category was present in each cohort's positive and negative responses.

Statistical Analysis

Quantitative analysis and plotting of results was carried out in RStudio (RStudio Team 2020). For all statistical tests, an alpha of 0.05 was used. Statistical tests used the T test for comparing differences between two groups. For the statistical comparisons we carried out, we also calculated a power analysis. For the power analysis, we simulated 1,000 data sets per condition for T tests comparing CE vs NonCE URG and CE vs NonCE WR students. We tested effect sizes varying from 0.2 to 2. Effect sizes were defined as the mean of the CE cohort minus the mean of NonCE Cohort then divided by the standard deviation of the NonCE cohort. The means and standard deviations of our simulated data were based on the same parameters from our observed data for NonCE cohorts. To examine the power of each condition, we calculated the proportion of simulations that correctly resulted in a statistically significant T test result of < 0.05.

Results

What experiences relating to student success did CE help students with?

Importantly, we asked CE students to rate at the end of their first year how much the program impacted their experiences in 5 key areas. The areas we focused on were **1)** research skills, **2)** sense of belonging, **3)** self-confidence, **4)** overall wellbeing, and **5)** interactions with advisors. Students rated if CE positively influenced their experience with a Likert scale 1 (strongly disagree) to 5 (strongly agree) (Fig. 1.1). We present our areas of focus here in ascending order of mean student response. Research skills had the lowest mean response of 3.7 (between neutral and somewhat agree). CE students predominantly agreed that the program improved their sense of belonging with an average response of 3.9 (between neutral and somewhat agree). All but one student agreed that CE improved their self-confidence with a mean response of 4.2 (somewhat agree). CE students agreed that CE improved their overall wellbeing with a mean response of 4.2 (somewhat agree). CE student respondents unanimously agreed that the program aided their interactions with their advisors (mean = 4.5). CE students reported generally having improved experiences across these five areas relating to student success that they attribute to the CE program. These responses speak to the high value of the CE program.

What structured components of CE were most helpful?

We asked CE students to select up to three of the most helpful structured components of the program for them. The majority of students reported benefiting from the mental health strategies workshop focused on “Resiliency and Managing Negative Thoughts” (71.4%). They also identified one of the two workshops focused on managing their relationship with their mentor (42.8%). And students did also report benefits from the writing skills workshop on grant and fellowship applications (35.7%).

CE students reported improvement in skills related to success in graduate school

Students rated themselves before doctoral study and at the end of year-one in eight skills: **1)** interacting with faculty, **2)** science communication, **3)** mental wellbeing strategies, **4)** connection to resources, **5)** conducting research, **6)** evaluating journal articles, **7)** financial literacy, and **8)** fellowship application writing. We calculated students’ change (post score - pre score) in our eight skills for our three cohorts (CE, NonCE URG, and NonCE WR) (Fig. 1.2). CE students reported a larger mean increase in skills than both NonCE cohorts for all skills except financial literacy. For financial literacy, NonCE URG students indicated the most increase in this skill.

We were most interested in statistically testing if CE students reported more improvements than each of the NonCE cohorts. We did so using a T test to make two pairwise comparisons (CE vs NonCE URG and CE vs NonCE WR) for each skill. We found two statistically significant differences. For CE vs NonCE URG students, CE students reported a larger growth in their connection to resources (T test, p-value =

0.030). For CE vs NonCE WR students, CE students indicated larger growth in working with faculty (T test, p-value = 0.045). All other tests resulted in p-values > 0.05 (Table S5).

Additionally, we used a power analysis to determine the probability that we could accurately detect a difference between our cohorts with our sample sizes plus the observed mean and standard deviation of our control NonCE groups (Fig. 1.S1). For comparisons of NonCE students we used the averages of their mean and standard deviation per skill. For URG students the mean was 0.60 and standard deviation was 1.0. WR students had a mean of 0.66 and a standard deviation of 0.80. For our power analysis of comparisons between CE and NonCE URG students, we find that a large effect size > 1.25 would be needed to correctly find a statistical difference in at least 80% of simulated data sets. For CE vs NonCE WR students, a large effect size of 0.75 would be needed to reach 80% accurate statistical tests. Given the means and standard deviations of the NonCE groups, our power analysis suggests we could accurately detect statistical significance in 80% of cases where our CE cohort had a mean response that is > 1.85 units above NonCE URG students or > 1.22 units above NonCE WR students. Our two statically significant cases have a smaller difference in mean than those thresholds, but still may be in the less than 80% of cases where we could detect a true difference.

What did students say?

Students were asked to identify factors that influenced their first year of graduate school and separate them by those that had positive vs negative impacts. For CE students, there were 9 responses (64.2% of cohort) for positive factors and 5 responses (35.7% of cohort) for negative factors. Among NonCE URG students, 2 responded (18.2% of cohort) with positive and the same 2 (18.2% of cohort) with negative factors. NonCE WR students had 15 responses (26.3% of cohort) for positive and 15 responses (26.3% of cohort) for negative factors.

We coded student responses based on our inductively created categories of **1)** community support, **2)** skills development, **3)** mentorship/advising, **4)** mental health, and **5)** financial resources. A given response could be coded as discussing more than one category. We totaled the number of responses for each category subdivided by cohort and positive vs negative factors (Table 1.2). The themes are listed above and in the table by order of their popularity ranging from community support with 31 responses to financial resources with 6 responses for all cohorts and positive and negative factors combined.

The most common theme was community support which appeared in over 65% of responses. All but one CE student who responded to the positive question commented on how CE made them feel supported. One CE student wrote: *"Because of CE, I felt like I belonged at UCLA."* And 3 of the 4 negative CE responses about community support noted that they think the online delivery of the program hindered their community

building within CE. For example: *"...the program being completely virtual made it difficult to connect with other students."* NonCE students generally referred to student groups and structured university activities that connected them to others at UCLA. One NonCE WR student noted they found community through a student association: *"The engineering graduate student association plans a lot of events which helps us get to know each other and meet new people."* While another NonCE WR student found connection to be lacking: *"I didn't feel very connected to my department or the school as a whole…"* Reflecting the importance of community for diversity, a NonCE URG student poignantly wrote, *"I feel like I don't belong in a lot of the spaces I have been a part of thus far."*

Skills development was also a popular topic, but referred to both technical and "soft skills" such as achieving work life balance. NonCE WR students in particular commented on skills they wish they had been taught, including interacting with an advisor. For example, *"Interacting with advisors has been a difficult journey for me."* They also noted programs that were not effective, such as a departmental coding "bootcamp." CE students also reiterated points about a software workshop that wasn't particularly relevant to them because their field favors a similar but different software program: *"I would have appreciated an R [software] tutorial day versus a Tableau [software] training."*

Mentorship and advising proved to be a salient topic for students. CE students noted that it was particularly helpful to begin working with their advisor over the summer. One

student wrote, *“I really appreciated the opportunity to do a small research project and start working with my advisor before the beginning of the school year.”* And another added*, “I was also able to publish two manuscripts within my first year because I was able to begin working on them during [CE].”* In contrast, both NonCE URG students and one WR respondent specifically wished for help with navigating their advising relationship. One URG student wrote, *“I need help navigating the relationship with my advisor…”* and the WR student similarly stated, *“I wish there were more tips and guidance about how to find and interact with an advisor.”* Furthermore, when NonCE WR and URG students cited mentorship as playing a positive role in their first year, they only credited mentorship from sources other than their advisor such as student organizations, senior graduate students, or “younger faculty.”

Mental health was highlighted as impacting first year experiences in nearly 20% of student responses. Two of the 10 CE students focused on how mental health oriented workshops benefited them and the one negative response expressed a desire for more workshops on mental health related topics including “burnout.” In contrast, only two of the 19 NonCE WR students remarked on having support in maintaining their mental health. And four NonCE WR remarked that they were struggling with their mental health. For example, one NonCE WR student wrote: “*It has been very hectic trying to balance coursework with research expectations and finding a project I want to work on.*” and another added, “*I'm super depressed.*”

Lastly, a factor multiple students identified was financial resources. One CE student commented, "*I am very thankful for this program in their financial support…*" But another CE student wrote, "*I was unable to do research in-person [due to constraints with my access to health insurance]. That significantly hindered my ability to make progress in my project.*" While the CE program was conducted online, students had the option to carry out CE research in person. But some mechanism relating to accessing health insurance hindered this student's experience in the program. For the NonCE cohorts, no URG students mentioned finances. For the four NonCE WR students who mentioned financial resources, three commented specifically on housing. One was grateful for university subsidized family housing. But the other two NonCE WR students commented on the strains of the Los Angeles housing market. For example, "*Looking for affordable housing within a 20 minute walking distance to campus has been very difficult… [Commuting] while living not within walking distance to campus makes it hard to get enough rest and maintain well-being.*"

Discussion

The research described here highlights ways a diversity-oriented STEM summer bridge program can benefit students during their first year of doctoral study. In particular, quantitative and qualitative analyses of responses from CE students and control cohorts identified data trends supporting a positive impact of CE on research skills and psychosocial traits important for success in graduate school. From the perspective of diversifying STEM graduate programs, we present the benefits reported by CE students in the context of four major causes of graduate student attrition.

Advisor-Advisee Relationship / Working with Faculty

A positive effect of CE on students' advisor-advisee relationships was evident throughout our analyses. Notably, CE students unanimously agreed that the program improved their interactions with their advisor (Fig. 1.1). When asking students to rate their skill of interacting with faculty, CE students reported a higher mean growth in working with faculty compared to NonCE URG and WR students (Fig. 1.2). The positive impact was likely driven by two workshops which focused entirely on how to manage interactions with advisors. CE students identified the "Mentoring Up" workshop as the second most helpful CE program component.

Additionally, CE students repeatedly commented on the benefit of starting research training with their advisor over the summer. Summer may be a particularly advantageous time to begin working with an advisor because faculty generally have more availability due to limited teaching and service requirements. While many faculty may take vacation during the summer, and in some disciplines may conduct field work over this time, the CE program requires students to have a faculty member who commits to mentoring the student during the program. So the students are paired with advisors who are explicitly available during the six week program. Thus, compared to NonCE cohorts, CE students in our study had additional training in how to manage their advising relationship and additional time to navigate the interaction.

It is worth noting that multiple NonCE students commented on difficulties interacting with their advisors. One URG and one WR student explicitly wished for more support in navigating this relationship. These observations emphasize the need for this type of support for STEM PhD students and highlight the significance of our finding that all CE respondents viewed the program as specifically assisting them in interacting with their advisors.

Socialization / Connection to others at UCLA

Receiving social support may be particularly salient to CE students because of their minoritized racial/ethnic identities as well as first-generation status. All but one CE student identified with a URG racial or ethnic identity (Table 1.2). There is strong evidence that it is important for mentees from URG backgrounds to have mentors with cultural awareness (Thomas 2001, Osula & Irvin 2009, Womack *et al.* 2020). Additionally, 50% of CE students reported being first-generation with regards to college degrees and 78.6% with regards to advanced degrees. In contrast, 17.5% of NonCE WR students were first-generation college students and 45.6% first-generation for post-college education. The lack of familial experience in advanced degree studies may translate to fewer avenues for first-generation students to learn how to navigate the non-technical parts of graduate school. Additionally, we view socialization as linked to mental health. Social connections can offer emotional support, companionship, and a sense of belonging, which can help individuals cope with stress, reduce feelings of loneliness, and enhance overall well-being. Notably, multiple aspects of our analyses show that CE students benefited from the social aspects of CE.

Most CE students agreed that the program improved their sense of belonging, their self-confidence and their overall well-being during their first year (Fig. 1.1). These findings are also reflected in CE students reporting a larger increase in their mental wellbeing strategies and their connection to resources compared to both NonCE groups (Fig. 1.2). When asked for the most helpful components of CE, students' top answer was a workshop that highlighted social strategies to build self-efficacy (belief that one can do what is necessary to achieve their goals). These results all support a conclusion that CE student socialization was aided by the program.

In all cohorts' short answer responses, community support was the most common theme. CE students explicitly commented on how the program made them feel more connected at UCLA. While many NonCE URG and WR students described community support they had found, some described a lack of belonging. Relatedly, students from the CE and NonCE WR cohorts commented on mental health topics. CE students reaffirmed how the self-efficacy workshop helped them as well as a workshop on resilience and negative thoughts. Two NonCE WR students remarked on the support they found for their mental health, but four described ways they needed more support.

The very nature of the CE program gave this cohort of students access to URG peers and invested faculty and staff, which created a natural space for these students to feel connected and supported from the start of their program. Also, the workshops the CE students highlighted focused on non-technical or so-called "soft skills." These types of

skills are not usually taught explicitly in graduate education. Instead, most traditional elements of graduate education, such as coursework and advising from PIs, is focused on developing the critical technical skills needed for the research field. Notably, CE student open-ended responses specifically commented that the program's focus beyond technical skills was a significant benefit of the program. This speaks to students' desire and need to develop skills that contribute to graduate student success beyond knowledge of their field of study. Thus it is not surprising that CE students valued support in areas that are not typically a focus of STEM doctoral study.

Finances

Finances may be particularly relevant to the >70% CE students who self-reported being from low-income and working class socioeconomic backgrounds. This contrasts with ~36% of NonCE URG students and ~26% of NonCE WR students with these backgrounds. For the duration of the program, CE students are paid a $6,000 stipend for living expenses and possibly to offset some of their moving costs . For reference, UCLA graduate student teaching assistants in Fall 2023 would be paid a minimum of only $3,608 for the same time period (UAW Local 2865 2023). The pay above the university minimum that CE provides may be a significant boost to financial stability for CE students during the critical period of transition to their doctoral study. Such an effect was noted by one CE student's open-ended response stating gratitude for the program's financial support. Additionally, CE students are encouraged to move to Los Angeles to conduct their research in person. As the program occurs in the summer, this may give CE students greater potential to find housing at a time when there is less

competition from incoming students for housing, including UCLA subsidized housing, near campus. Lastly, CE students have the potential for better financial support in the future from a boost in writing skills from the CE workshop on fellowship applications (discussed in more detail under preparedness).

Preparedness / Doing and communicating research

In addition to the topics described above, students need to know and develop skills that are important for success in graduate school. We find evidence in multiple of our analyses that CE students developed important skills, including doing and communicating research.

We asked CE students if they thought that the program improved their research skills. The mean response of students was between neutral and somewhat agree (Fig. 1.1). This moderate response likely reflects the limited opportunity for students to delve in depth into research in a program of only 6 weeks. Even so, the most common response was “somewhat agree” (4) suggesting that many students believe their overall research skills benefited from the program.

When asked about the most helpful program components, CE students identified the grant and fellowship writing workshop as one of the top three. Better proficiency in writing can help students more easily secure funding for themselves through fellowships and their research through grants. Additionally, many doctoral programs also have milestones such as a dissertation proposal which has similarities to grant applications.

Being able to effectively communicate one's research plans will likely help these students meet these milestones more easily. Lastly, the final milestone in doctoral study is of course completion of the dissertation itself. The boost in writing skills that CE students gain during their first year has the potential for large long-term advantages.

Lastly, we found a striking trend that CE students reported a larger mean improvement in seven of the eight skills relating to student success compared to NonCE students (Fig. 1.2). The eight skills we focused on were chosen both because the CE program specifically seeks to support students in these areas and because the skills are important for one or more metrics of doctoral student success. Specifically we found statistically significant differences in connection to resources relative to NonCE URG students and interactions with faculty relative to NonCE WR students. However, we note that our statistical power was limited (Fig. S1.1) likely due to our small sample sizes. So we argue that the overall trend of CE students indicating more growth in almost all skills we measured is worth consideration.

Limitations and future directions

Small sample size limits this study, especially statistical inferences. The number of CE respondents was 14 students, or about one-third of eligible CE students. We chose to focus on a single cohort of CE students because the leadership, instructors, and program topics have varied between years. These variations introduce confounding differences between cohorts that would complicate analysis of data from larger sample sizes obtained by surveying multiple cohorts. The 11 NonCE URG student respondents

was also a small sample. This reflects the inherently smaller proportion of the overall student body that URG students comprise by definition. This was also a lower response rate than CE students, likely because NonCE students have no invested interest in the program.

It is worth noting that students are not randomly selected for the CE program. They must be nominated by their department and then selected by a committee. Because of the non-random nature of selecting students to participate in CE there are some uncontrolled variations between our study groups.

We also acknowledge that in our skills comparisons between CE and NonCE groups, there are many paths for students to acquire and develop the skills we studied. While the CE program directly targets these skills and the quantitative data show a trend of more CE students improving than NonCE URG students, CE is not the only way that gap could have developed. However, because the trend occurs across all but one skill, the parsimonious interpretation is that CE conferred an advantage to participating students over their NonCE peers.

In order to explore impacts of CE further, we suggest future studies look across multiple cohorts of CE students. With sufficient sample size, the confounding differences of program and environmental differences could be more easily controlled for in statistical analyses. For example, an ordered logistic regression could incorporate program year as a variable in the statistical model. Additionally, other demographic information such

as gender and first generation status could be tested as model components as well.

Lastly, a longitudinal approach should incorporate long term metrics of success for CE students. A particular metric of interest would be attrition, given the alignment between CE's goals and the causes of graduate student attrition. Future studies could also track other outcomes such as publication record and time to degree. While the primary focus of CE is to assist students during their acclimation to graduate school, it would be prudent to explore the potential long-term effects of beginning doctoral study with a "competitive edge".

Conclusion

The Competitive Edge (CE) program at the University of California, Los Angeles (UCLA) seeks to support first year doctoral students from historically excluded and underrepresented groups (URGs). Survey results of CE and NonCE first year PhD students at UCLA suggest that CE achieved the goal of better preparing program participants. Specifically, we found ways that the CE program addressed four major causes of graduate student attrition: 1) advisor-advisee relationship, 2) socialization, 3) finances, and 4) preparedness. *Advising Relationship:* A higher percentage of CE students than their nonCE peers reported improvements in managing interactions with their advisor. Furthermore, all CE respondents indicated that the program specifically helped those interactions. *Socialization:* The majority of CE students agreed that the program improved their sense of belonging and overall well-being. *Finances:* CE

students received a significant stipend and were grateful for that financial support. They also reported benefiting greatly from a workshop on fellowship writing, which could improve future funding prospects. *Preparedness:* In seven of eight key skills for graduate students, proportionally more CE students improved compared to their NonCE peers. Based on the program's impact in these areas, we anticipate the program having long term effects on participants' retention and success in graduate school, a hypothesis that warrants longitudinal studies of multiple CE student cohorts. Given the positive results of the CE program at UCLA, the program's model could be used to build or improve upon institutional support for doctoral students from URGs at other institutions.

Figures

Figure 1.1

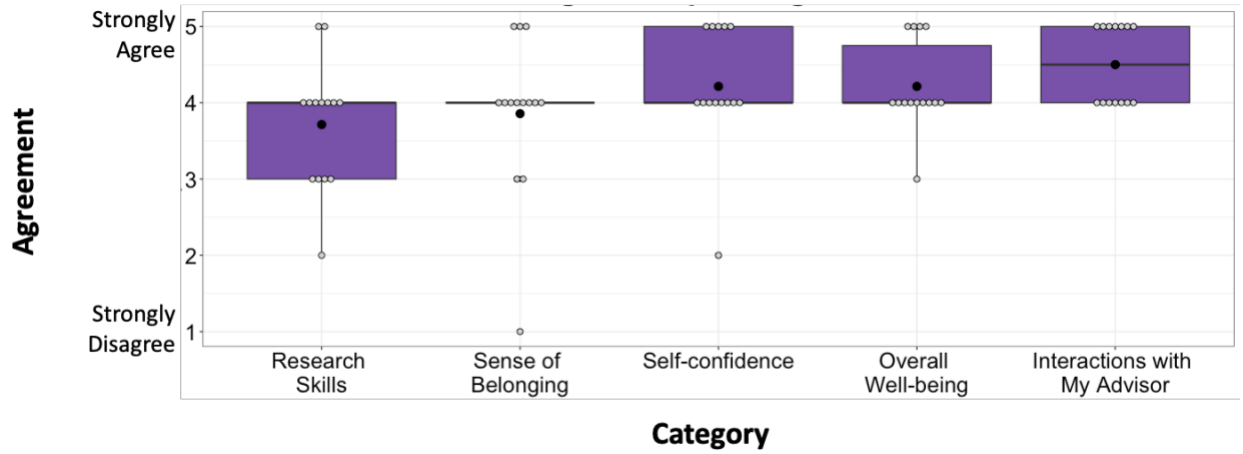

Figure 1.1. CE improves student experience in categories relevant to success in graduate school. Box plots of CE students' level of agreement with the statements: "During the 2021-2022 academic year, the Competitive Edge program improved my…" for five categories that the program focuses on improving (x-axis). Responses (y-axis) were on a Likert scale from 1 (Strong Disagree) to 5 (Strongly Agree). Light gray dots represent individual responses while black dots represent mean responses. Thick black lines represent median values.

Figure 1.2

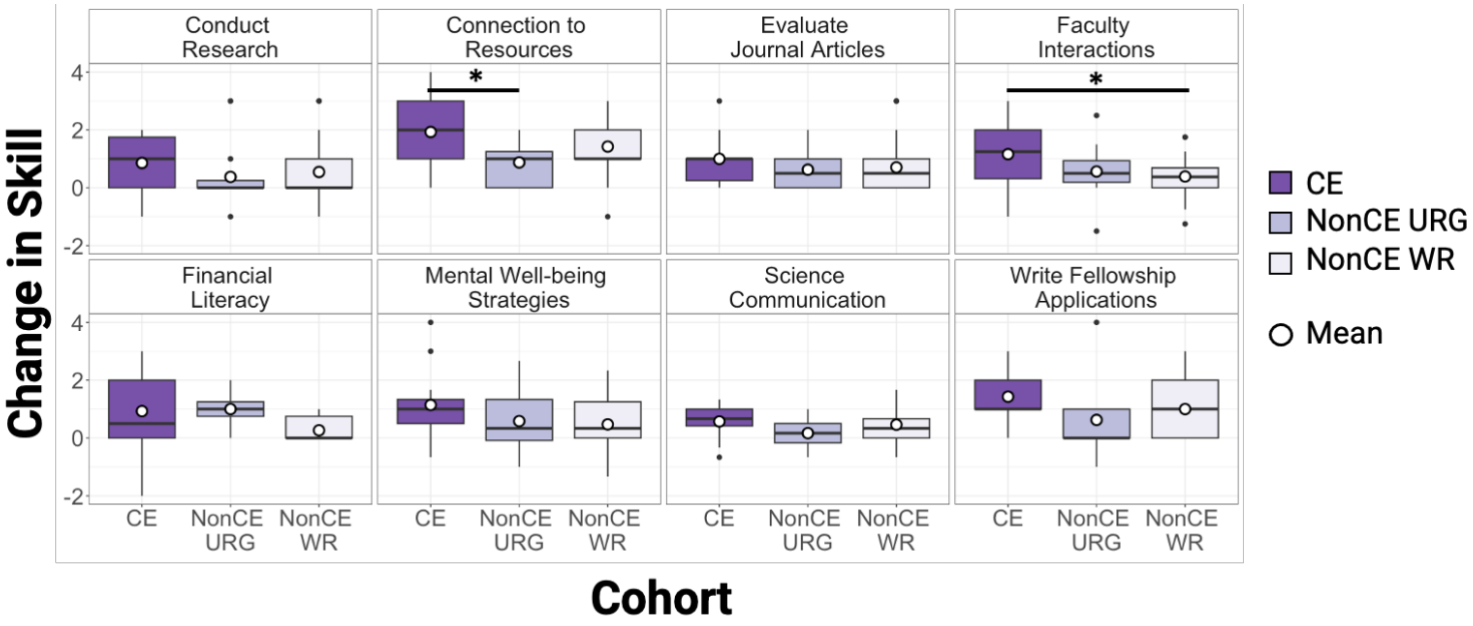

Figure 1.2. Self-reported change in skills related to success in graduate school. Box plots showing the self-reported change in skills related to success in doctoral studies during the first year of grad school (scale -4 to 4). Cohorts are shown in different colors (CE = dark purple, NonCE URG = purple, and NonCE WR = light purple). Skills we measured are displayed at the top of each panel. White dots represent mean responses and thick black lines represent median values. Black dots represent outliers. Horizontal lines with an asterisk represent T tests that were statistically significant for p-values < 0.05. For all skills but financial literacy, CE students report a higher mean improvement in these skills.

Tables

Table 1.1

Demographic breakdown of responses by cohort. Numbers represent the total number of responses for a specific cohort and demographic category. Percentages in parentheses show what percentage that number makes up of a given cohort.

	Cohort			
Race / Ethnicity	**CE**	**NonCE URG**	**NonCE WR**	**All Cohorts**
Asian	0 (0.0%)	0 (0.0%)	40 (70.2%)	40 (48.8%)
Black / African American	1 (7.1%)	1 (9.1%)	0 (0.0%)	2 (2.4%)
Filipino	1 (7.1%)	0 (0.0%)	0 (0.0%)	1 (1.2%)
Hispanic / Chicano / Latinx	10 (71.4%)	4 (36.4%)	0 (0.0%)	14 (17.1%)
Middle Eastern/North African	0 (0.0%)	1 (9.1%)	0 (0.0%)	1 (1.2%)
Mixed Race Asian and White	0 (0.0%)	0 (0.0%)	5 (8.8%)	5 (6.1%)
Mixed Race Not Asian and White	1 (7.1%)	5 (45.5%)	0 (0.0%)	6 (7.3%)
White	1 (7.1%)	0 (0.0%)	12 (21.1%)	13 (15.9%)

Gender				
Female	8 (57.1%)	6 (54.5%)	35 (61.4%)	49 (59.8%)
Male	4 (28.6%)	4 (36.4%)	21 (36.8%)	29 (35.4%)
Non-binary or Androgynous	2 (14.3%)	1 (9.1%)	1 (1.8%)	4 (4.9%)
First Generation*				
Undergraduate	7 (50.0%)	5 (45.5%)	10 (17.5%)	22 (26.8%)
Graduate or Professional School	11 (78.6%)	7 (63.6%)	26 (45.6%)	44 (53.7%)
Socioeconomic Class				
Low Income / Working Class	10 (71.4%)	4 (36.4%)	15 (26.3%)	29 (35.4%)
Middle Class	4 (28.6%)	4 (36.4%)	24 (42.1%)	32 (39.0%)
Upper-Middle Class / Upper Class	0 (0.0%)	3 (27.3%)	17 (29.8%)	20 (24.4%)

*Students whose parents did not complete the degree indicated

Table 1.2.

Counts of student responses to open ended questions classified by cohort, question, and themes. All students were asked to name factors that contributed positively to their first year of doctoral study. Then they were asked about negative factors. Column titles divide response counts by the three cohorts and/or the question prompt of positive vs negative factors. Five common themes were identified in these responses with inductive coding. These themes are listed in column 1. The final row shows the total number of responses for a given column. A single response could be categorized with discussing multiple themes, so the total responses in a column may be smaller than the total number of times a given topic was discussed by that group.

	CE		NonCE URG		NonCE WR		All Cohorts		Total
Themes	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	
Community Support	8	4	1	1	12	5	21	10	31
Skills Development	5	2	1	0	4	6	10	8	18
Mentorship/Advising	3	0	1	2	3	1	7	3	10
Mental Health	2	1	0	0	2	4	4	5	9
Financial Resources	1	1	0	0	2	2	3	3	6
Response Totals	9	5	2	2	15	14	26	21	47

References

Ampaw, F. D., & Jaeger, A. J. (2012). Completing the Three Stages of Doctoral Education: An Event History Analysis. Research in Higher Education, 53(6), 640–660. https://doi.org/10.1007/s11162-011-9250-3

Astin, A. W. (2014). Student involvement: A developmental theory for higher education. College Student Development and Academic Life: Psychological, Intellectual, Social and Moral Issues, (July), 251–263.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. Research in Higher Education, 12(2), 155–187. https://doi.org/10.1007/BF00976194

Bowlin, L., Sweat, K., Watts, S., & Throne, R. (2017). Agency, Socialization, and Support: A Critical Review of Doctoral Student Attrition. Online Submission.

Brill, J. L., Balcanoff, K. K., Land, D., Gogarty, M., & Turner, F. (2014). Best Practices in Doctoral Retention: Mentoring. Higher Learning Research Communications, 4(2), 26. https://doi.org/10.18870/hlrc.v4i2.186

Devine, K., & Hunter, K. (2016). Doctoral Students' Emotional Exhaustion and Intentions to Leave Academia. International Journal of Doctoral Studies, 11, 035–061. https://doi.org/10.28945/3396

Gardner, S. K. (2007). "I heard it through the grapevine": Doctoral student socialization in chemistry and history. Higher Education, 54(5), 723–740. https://doi.org/10.1007/s10734-006-9020-x

Girves, J. E., & Wemmerus, V. (1988). Developing Models of Graduate Student Degree Progress. The Journal of Higher Education, 59(2), 163–189. https://doi.org/10.1080/00221546.1988.11778320

Golde, C. M. (2005). The role of the department and discipline in doctoral student attrition: Lessons from four departments. Journal of Higher Education, 76(6), 669–700. https://doi.org/10.1080/00221546.2005.11772304

Hardré, P. L., Liao, L., Dorri, Y., & Beeson Stoesz, M. A. (2019). Modeling American graduate students' perceptions predicting dropout intentions. International Journal of Doctoral Studies, 14, 105–132. https://doi.org/10.28945/4161

Herman, C. (2008). Obstacles to success – doctoral student attrition in South Africa Research on doctoral attrition. Africa, 40–52.

Hermida, A. (2017). Everyday Oppression: The Challenges of Belonging for Underrepresented Doctoral Students at a Predominantly White Institution. [Doctoral dissertation, University of Minnesota]. Core. https://core.ac.uk/display/211352927?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1

Knight, L., Hall, T., & Green-Powell, P. (2014). An Analysis of Historically Black Colleges and Universities Student Retention and Attrition Efforts, 1(8), 123–138.

Kou-Giesbrecht, Sian. "Asian Americans: the forgotten minority in ecology." *The Bulletin of the Ecological Society of America* 101.3 (2020): e01696.

Lovitts, B. E. (2001). Leaving the Ivory Tower: The Causes and Consequences of Departure from Doctoral Study. Lanham, Maryland: Rowman & Littlefield Publishers, Inc.

National Academies of Sciences, Engineering, and Medicine. 2019. The Science of Effective Mentorship in STEMM. Washington, DC: The National Academies Press. https://doi.org/10.17226/25568.

Maher, M. A., Wofford, A. M., Roksa, J., & Feldon, D. F. (2017). Exploring Early Exits: Doctoral Attrition in the Biomedical Sciences. Journal of College Student Retention: Research, Theory & Practice, 152102511773687. https://doi.org/10.1177/1521025117736871

Martinez, E., Ordu, C., Sala, M. R. D., & McFarlane, A. (2013). Striving to obtain a school-work-life balance: The full-time doctoral student. International Journal of Doctoral Studies, 8, 39–59. https://doi.org/10.28945/1765

Mccoy, D. L., Winkle-wagner, R., & Winkle-wagner, D. L. M. R. (2020). Bridging the Divide : Developing a Scholarly Habitus for Aspiring Graduate Students Through Summer Bridge Programs Participation Bridging the Divide : Developing a Scholarly Habitus for Aspiring Graduate Students Through Summer Bridge Programs Participation, 56(5), 423–439.

Meara, K. O., Griffin, K. A., & Robinson, T. (2017). Sense of Belonging and Its Contributing Factors in Graduate Education. International Journal of Doctoral Studies, 12, 251–279. https://doi.org/10.28945/3903

Osula, B., & Irvin, S. M. (2009). Cultural awareness in intercultural mentoring: A model for enhancing mentoring relationships. International Journal of Leadership Studies, 5(1), 37-50.

Padilla, R. V. (1999). College Student Retention: Focus on Success. Journal of College Student Retention: Research, Theory & Practice, 1(2), 131–145. https://doi.org/10.2190/6w96-528b-n1kp-h17n

Peteet, B., Bridge, E., Ethnic, P., Ronald, T., Postbaccalaureate, E. M., Program, A., … Preparation, I. (2016). How graduate school bridge programs can help increase diversity in STEM subject admission, 1–4.

Rockinson-Szapkiw, A. J. (2019). Toward understanding factors salient to doctoral students' persistence: The development and preliminary validation of the doctoral academic-family integration inventory. International Journal of Doctoral Studies, 14, 237–258. https://doi.org/10.28945/4248

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Ruud, C. M., Saclarides, E. S., George-Jackson, C. E., & Lubienski, S. T. (2018). Tipping Points: Doctoral Students and Consideration of Departure. Journal of College Student Retention: Research, Theory and Practice, 20(3), 286–307. https://doi.org/10.1177/1521025116666082

SERU Consortium. gradSERU Survey Design . (2021). Retrieved March 8, 2021, from https://cshe.berkeley.edu/seru/about-seru/seru-surveys/gradseru-survey-design

Sowell, R., Allum, J., & Okahana, H. (2015). Doctoral Initiative on Minority Attrition and Completion. https://doi.org/10.1145/1401890.1402023

Sowell, R., Zhang, T., Redd, K., & King, M. F. (2008). Analysis of baseline program data from the Ph.D. completion project.

Thomas, D. A. (2001). Race matters. Harvard Business Review. April.

Tinto, V. (1982). Limits of Theory and Practice in Student Attrition. The Journal of Higher Education, 53(6), 687–700. https://doi.org/10.1080/00221546.1982.11780504

UAW Local 2865. (2023). Academic Student Employees (ASE) Contract. Retrieved May 26, 2023 from https://uaw2865.org/ase-contract/

UCLA Department of Graduate Education. (2023). Retrieved May 20, 2023. https://go.grad.ucla.edu/

University of California, Office of Undergraduate Admissions. (2023). AB 540 Nonresident Tuition Exemption. Retrieved June 5, 2023, from https://admission.universityofcalifornia.edu/tuition-financial-aid/tuition-cost-of-attendance/ab-540-nonresident-tuition-exemption.html

Weiss, C. S. (1981). The Development of Professional Role Commitment Among Graduate Students. Human Relations, 34(1), 13–31.

Womack, V. Y., Wood, C. V., House, S. C., Quinn, S. C., Thomas, S. B., McGee, R., & Byars-Winston, A. (2020). Culturally aware mentorship: Lasting impacts of a novel intervention on academic administrators and faculty. PloS one, 15(8), e0236983.

Zhou, E., & Okahana, H. (2019). The Role of Department Supports on Doctoral Completion and Time-to-Degree. Journal of College Student Retention: Research, Theory and Practice, 20(4), 511–529. https://doi.org/10.1177/1521025116682036

Chapter 2: Exploring the dynamics of recombination rates across *Canis familiaris*

In preparation for submission to Molecular Biology and Evolution

A Supplemental Appendix is available online as this dissertation's Supplementary Materials: Ch2_Supplementary_Information.pdf

Abstract

Meiotic recombination is a notable mechanism that, in sexually reproducing species, permits the proper alignment and segregation of homologous chromosomes while also generating novel combinations of alleles in populations. Recombination rates are known to vary on the level of species, populations, sexes, and individuals. Thus, it is a trait that can be acted upon by forces of evolution such as natural selection and genetic drift. In fact, it has been hypothesized that domesticated animals may have higher recombination rates due to selection. Here, we investigate the possible changes in recombination rate due to the process of domestication in canids. We estimated rates of recombination using patterns of linkage disequilibrium (LD) from high-coverage whole genome sequence data from a population of North American gray wolves and four breeds of domestic dogs. Performing inference using the software pyrho, we find similar estimated recombination rates in wolves (mean r = 4.2e-9/bp/generation) and two dog breeds: labs (mean r = 3.0e-9/bp/generation) and Tibetan mastiffs (mean r = 4.3e-9 /bp/generation). Interestingly, we infer that border collies (mean r = 1.6e-8

/bp/generation) have ~4X higher recombination rate than wolves, and pugs (mean r = 1.6e-10/bp/generation) have a recombination rate ~27X lower than wolves (Kruskal-Wallis test, p value < 2.2e-16). Simulations suggest that this difference in mean recombination rate between our populations is not driven solely by differences in demography, presence of runs of homozygosity, or variation in mutation rate. However, we also find that the landscape of recombination along the genome is concordant across populations. Our findings indicate that there is still much more to discover regarding the recombination landscape of canids, which will further enhance our understanding of evolutionary patterns associated with recombination.

Introduction

Meiotic recombination is observed across sexually reproducing species and serves as a mechanism to create genetic diversity by decoupling alleles. The patterns and rates of recombination are known to vary at many levels including between species (Kawakami et al., 2017; Singhal et al., 2015; Wall & Stevison, 2016), populations (Chan, Jenkins, & Song, 2012; Kawakami et al., 2017)**,** and individuals (Kong et al., 2010)**.** Because of this variation and the impact recombination can have on phenotypes, it is a trait that can evolve under selection and/or drift. For example, natural selection has been shown to influence relative recombination rates in two populations of *Drosophila* (Samuk, Manzano-Winkler, Ritz, & Noor, 2020).

Rates of recombination are known to vary throughout the genome with spikes in 1 to 2KB regions known as hotspots. In many vertebrates, the gene PRDM9 has been

shown to direct the location of these hotspots (Baker *et al.* 2015; Cavassim *et al.* 2020; Grey, Baudat, & de Massy, 2018; Oliver et al., 2009; Parvanov, Petkov, & Paigen, 2010). Despite recombination being directed by this shared gene, sister taxa do not necessarily share hotspot locations. A striking example exists in humans and chimpanzees where overlap in recombination rates is present at the broad scale, but not on a fine scale (Auton et al., 2012). Additionally, many lineages have experienced loss of function mutations in the *PRDM9* gene including canids (Muñoz-Fuentes, Rienzo, & Vilà, 2011). Additionally, within vertebrates the whole *PRDM9* gene has been lost multiple times (Cavassim *et al.* 2022). Despite a non-functional or absent *PRDM9*, recombination hotspots have still been found in many species including but not limited to birds (e.g. Kawakami et al., 2017; Singhal et al., 2015), flies (e.g. Samuk, Manzano-Winkler, Ritz, & Noor, 2020), and dogs (Auton et al., 2013; Campbell, Bhérer, Morrow, Boyko, & Auton, 2016; Wong et al., 2010). These studies in taxa without functioning PRDM9 have found that hotspots tend to overlap with transcriptional start site (TSS) elements and CpG islands. Despite hotspots being enriched near TSSs, they are not necessarily conserved in the same places in closely related species (e.g. Kawakami et al., 2017; Samuk et al., 2020)

Domestic dogs (*Canis familiaris*) and their sister wolves (*Canis lupus*) serve as an interesting case study of evolutionary patterns of recombination in species lacking a functional *PRDM9*. This is partially due to the potential impacts of domestication. The breeding of dogs is believed to be the earliest domestication event (Larson *et al.* 2012). Domestication is a form of artificial selection that is a faster and goal-oriented form of

evolution relative to natural selection. Through this repeated process of selecting multiple traits, domestication can result in extreme phenotypic differences among breeds (Clutton-Brock 1992; Diamond 2002). However, it has long been hypothesized that domestication may have also been unknowingly selected for higher recombination rates (Gornall, 1983). The logic behind this hypothesis is that increased recombination can amplify the genetic diversity underlying the extreme phenotypic diversity in species such as domestic dogs. And while this has been suggested, to the authors' knowledge, most empirical studies on this topic have been conducted primarily in plants (Dreissig, Mascher, Heckmann, & Purugganan, 2019; Gornall, 1983; Ross-Ibarra, 2004; Fuentes *et al.* 2022). One study in mammals that compared dogs and wolves did not find support for the hypothesis that domesticates have increased recombination rates, but this study was limited to comparing fewer than 20 gene regions(Munoz-Fuentes et al., 2015). Given the already available research on genome-wide recombination rates in village dogs (Auton et al., 2013) and breed dogs (Campbell et al., 2016; Wong et al., 2010), this is a prime opportunity to further test this hypothesis in mammals.

Here we explore how recombination rates differ between North American gray wolves and some domestic dogs to examine how domestication may have influenced recombination rates on an evolutionary time scale. Specifically, we focus on the patterns of recombination between four dog breeds (border collies, labs, pugs, and Tibetan mastiffs) and their sister taxa of wolves. We find that recombination rates of two of our focal breeds (labs and Tibetan mastiffs) resemble those inferred in wolves while rates for pugs and border collies vary significantly in opposite directions compared to

wolves. We explore how runs of homozygosity, mutation rates and demographic history impacts our inferences of recombination. Our surprising finding of variation between dog breeds suggests potential differentiation in genomic parameters between dog breeds and highlights challenges in inferring recombination rates from genomic data.

Results

Demographic inferences

To infer recombination across the genome in the five candid populations, we used the program pyrho which leverages patterns of linkage disequilibrium (LD), or the correlations in alleles at different loci. Pyrho accounts for the fact that demographic history also influences LD patterns across the genome by allowing users to input a model of population history. Thus, we first used SMC++ to estimate the demography of our five canid populations. (Fig. 2.1). The demographic history estimated for all dog breeds shows a bottleneck at approximately 20,000 generations ago, consistent with the domestication process leading to a much smaller effective population size relative to wolves (Freedman *et al.* 2014, Mooney *et al.* 2021). We then used these demographic models in our LD-based inference of recombination rates.

Recombination maps

We inferred the per base-pair recombination rate, *r*, across the genome in wolves and four dog breeds using pyrho. As seen in the map of recombination across chromosome one, the recombination rate varies between some populations (Fig. 2.2, and Fig. S2.1 for other chromosomes). We observe that relative to wolves, labs and Tibetan mastiffs

have similar ranges of estimated recombination rates while the rates in border collies and pugs deviate from those inferred in wolves. Estimated recombination rates are the highest for border collies throughout the genome and lowest for pugs. We observe similar patterns when considering the chromosome-wide average rates for all populations (Fig. 2.3). Additionally we found that the chromosomal averages of recombination rate are significantly different between at least one of these five populations (Kruskal-Wallis test, p value < 2.2e-16).

Comparing rates between species and populations

After observing the global differences in recombination rates between the canid populations, we also explored how recombination rates vary on more localized scales. For example, while the pugs have the lowest global recombination rate, we asked whether the areas of the pug genome with the highest recombination rates are the same as the regions of the wolf genome with the highest recombination rate. To compare the intensity of recombination rates in given windows of the genome, we calculated the average recombination rate for non-overlapping windows of three different sizes: 100KB, 1MB, and 5MB. Due to the difference in scale between pug and wolf recombination rates, we calculated the rank for each window within a given population.

We then compared the ranks of the windows between our four dog breeds and wolves (Fig. 2.4 Panel A, 2.S2 Panel A, S2.3 Panel A). We find similar trends across our window sizes, and focus here on the correlations at the scale of 5MB windows. Among

breed dogs, we find the recombination landscape along the genome is significantly correlated with that of wolves. For example, the highest correlation to wolves is the Tibetan mastiff recombination rate ranks (R^2 = 0.709, p-value = 1.2e-114). Border collies (R^2 = 0.474, p-value = 1.2e-60) and labs (R^2 = 0.55, p-value = 9.6e-75) had similar levels of correlation with wolves. In contrast, the pugs had the lowest correlation with wolves (R^2 = 0.157, p-value = 2.6e-17). These results suggest that the relative intensity of recombination on a localized 5MB and smaller scale is fairly concordant across the genome between wolves and domestic dogs. Pugs are an outlier, with a notably lower correlation.

To validate our LD-based recombination maps, we also compared the ranks of recombination for wolves and the four breed dogs to estimates of recombination from a pedigree based genetic map derived from mixed-breed dogs (Campbell *et al.* 2013) (Fig. 2.4 Panel B, S2.2 Panel B, S2.3 Panel B). Genetic maps derived from pedigrees estimate recombination events directly between parents and offspring, albeit on a broader scale than LD based methods. In contrast, LD-based methods, as those used in this study, can estimate the impact of recombination on longer evolutionary time scales but at higher resolution (McVean et. al 2004, Myers et. al 2005, Dapper and Payseur 2017). By comparing our estimates of *r* based on LD patterns to those derived from pedigrees, we can validate our LD-based inferences and test the concordance of recombination rates over different timescales.

We found moderate levels of correlation between the pedigree-based genetic map and the LD-based map from all of our populations (border collie: R^2 = 0.425, p-value = 1.1e-49; lab: R^2 = 0.442, p-value = 2.1e-52; pug: R^2 = 0.138, p-value = 1.7e-14; Tibetan mastiff: R^2 = 0.612, p-value = 7.4e-84; wolf: R^2 = 0.595, p-value = 4.4e-80). Again, pugs have the lowest correlation. This moderate correlation suggests the relative intensity of recombination is conserved at the broad-scale across the dog populations we evaluated. Some of the differences observed are likely attributed to the differences in time scale and spatial specificity of LD-based vs pedigree based methods. The further deviation of pugs specifically is consistent with our finding that the ranking of pug recombination rates differed the most from wolves compared to our three other breeds. Overall, these results suggest general agreement in recombination rate inferences between our LD based inferences and those made from pedigree based observations, validating our LD-based estimates.

Simulations

Given the observed differences in average recombination rates of two dog breeds relative to wolves, we used population genetic simulation to explore the robustness of the LD-based inference of recombination rate to various assumptions. We used pugs as a case study for this exploration via simulated genomic data.

Demography

One potential explanation for the apparent differences in *r* inferred using pyrho between pugs and wolves is that pyrho cannot accurately infer *r* due to the complex demographies of these populations (see Fig. 1). To directly test the role demography

plays in our inference of recombination rates using pyrho, we performed coalescent simulations of pug and wolf populations with identical mutation and recombination rates, but with differing demographic histories. Specifically, the demographies used for simulations were those we inferred from our empirical genomic data with *SMC++*.

We simulated 10 replicate populations for both wolves (simWolf) and pugs (simPug) with a constant recombination rate of 1e-8 /bp/generation, similar to the rate estimated in the literature (Campbell *et al.* 2016). We inferred the demography of each simulated dataset using SMC++ and then inferred *r* using the same pyrho pipeline as used for the empirical analyses. For each replicate, we calculated a mean recombination rate for the 40 chromosomes simulated. Among those 40 chromosome-wide recombination rates, we calculated the median value resulting in one value per replicate. We compared median values of each replicate(Fig. 2.5). The range and mean values for our simulated wolf and simulated pug replicates are very similar to each other (simWolf: range 4.3e-9 to 5.2e-9 and mean 4.6e-9 ; simPug: range 4.8e-09 to 6.0e-9 and mean 5.3e-9 per BP per generation). However, it is noteworthy that the estimated values of *r* are ~2-fold smaller than the true values used to generate the data. We next conducted additional simulations to explore why there might be an underestimate.

As demographic inference is imperfect and this uncertainty could contribute to the underestimation of *r*, we analyzed the recombination rates of these simulated genomes using the true demography they were simulated under (instead of inferring the demography using SMC++). We find that recombination rates are again similar between

simulated pugs and simulated wolves (Fig. 2.S5). The mean values are similar (simWolf = 8.23 e-9 /bp/generation and simPug = 6.96 e-9 /bp/generation). These results are still below the true recombination rate, but closer than the estimates using inferred demographic models.

Next, we hypothesized that recent population size would have the biggest impact on patterns of LD in dogs and, in turn, our inferences of *r*. Thus we repeated our analysis of our simulated data assuming a constant population size equal to the most recent N_e from the true demographies (15,820 for pug and 34,630 for wolf) of these simulated populations. Again, the mean genome-wide recombination rate per simulated population does not differ significantly between pugs and wolves (Fig. S2.5). However, the mean estimates of *r* (simWolf = 9.22 e-9 /bp /generation and simPug = 9.03 e-9 /bp /generation) are much closer to the true recombination rate. This finding suggests that recent population sizes have a strong impact on the inference of recombination rate. Assuming a constant population size that is the same as the true most recent population size results in more accurate inference of *r* using pyrho. While our real populations differ in more ways than just demographic history, these simulations give insights into the influence different demographic models have on our analysis. Furthermore, these simulations suggest that different demographic histories the populations do not artificially generate apparent decreased estimates of recombination rates in pugs.

ROHs

As a result of inbreeding during domestication and breed formation, dogs are known to have large portions of their genomes in runs of homozygosity (ROHs; Sams and Boyko 2019). To examine whether ROHs can impact inference of recombination rates using pyrho, we altered our simulated pug genomes to include ROHs. We added ROHs in two different ways. First, we added ROHs in individuals independent of ROHs in other individuals. We then re-inferred *r* using our pyrho analysis pipeline. Adding in ROHs resulted in a wider range (5.0e-9 to 7.2e-8 /bp/generation) but similar mean (1.3e-08/bp/generation) of recombination rates per simulated pug population (Fig. 2.5). As an alternate strategy, we added ROHs in the same genomic location across all 15 individuals (Fig. S2.5). We randomly selected an individual from our original ROH simulations and duplicated their ROHs in all individuals such that all ROHs were at identical locations in all individuals. The simulations with overlapping ROHs lead to similar inferences of *r* as our other model that introduced ROHs (range = 5.05 e-9 to 6.72 e-8 /bp /generation, mean = 1.21 e-8 /bp /generation). While the presence of ROHs can influence inferences of recombination rate from LD patterns, for our simulated populations, it mostly increased the variance of inferred recombination rates. Thus ROHs in dogs do not appear to be a driver of the recombination rate differences we observed between pugs and wolves.

Mutation rate

Importantly, pyrho uses the demographic history inferred from SMC++ to accurately disentangle *r* from $\rho=4Nr$. SMC++ is calibrated by the mutation rate (μ) when inferring

N_e from the patterns of genetic variation in the genome. Thus, inference of *r* could be sensitive to the underlying μ used in the analysis. For our analyses and simulations thus far, we assumed the same mutation rate of 4.5e-9 /bp/generation for wolves and pugs estimated from a different population of North American wolves (Koch *et al.* 2019). To explore the role of mutation rate on our results, we repeated our simulations for pugs with varying values of μ. We modified μ only for pugs because the assumed mutation rate is more likely to be accurate for wolves since it was estimated from a population of North American gray wolves.

For pugs, we simulated datasets with three different true values of μ: 4.5e-10, 2.25e-8 and 4.5e-8 per bp per generation; and then inferred *r* assuming μ=4.5e-9 /bp/generation (Fig. 2.5, Fig. S2.5). Thus we had conditions where we assumed a value of μ that was 10X, 0.5X, and 0.1X of the true μ, respectively. Importantly, this analysis assumes the incorrect μ in the SMC++ inference, allowing us to test how misspecification of the mutation rate impacts the pyrho inference of *r*.

When the assumed mutation rate is 10-fold larger than the true mutation rate, inferred recombination rates are overestimated (2.59 e-8 /bp /generation mean value). Because the assumed μ is too large, our simulated data has fewer SNPs than would be expected with this assumed mutation rate. Thus to reconcile the number of SNPs observed with this assumed larger mutation rate, SMC++ infers the population size to be smaller than the true population size (Fig. S2.5). This smaller population size is then used by pyrho when converting rho to *r,* resulting in the inferred values of *r* being larger than the true

values. As expected, we do find that when we inferred the demography using a a tue assuming a μ 10X larger than the true μ, inferred population sizes were ~10 fold smaller than those in the other misspecified μ scenarios (Fig. S2.4).

For the case of assuming a μ that 0.1X the true value, *r* was slightly underestimated (5.92 e-9 /bp/generation mean value) (Fig. S2.5). For an assumed mutation rate that is 0.5X of the true μ, *r* also was overestimated (1.43 e-8 /bp/generation mean value), but not as significantly as in the case of the assumed μ being 10X the true μ. In these cases of assuming a smaller μ than truth, the number of SNPs observed is larger than expected with the assumed μ, thus SMC++ infers larger than true population sizes (Fig. S2.4). While these large population sizes would drive down estimates of *r* through rho as discussed above, the larger number of SNPs increases power for inferring higher values of *r*. Thus, there are two competing forces that might offset each other here, potentially explaining why assuming a smaller mu may not consistently drive our inferred *r* in one direction relative to the true *r*.. While these inferred values of *r* are somewhat influenced by varying three different true values of μ, they do not suggest that changes only in pug mutation rates can fully explain the low inferred recombination rates in pugs.

Inference of recombination assuming different demographic models

The simulations described above provided some guidance as to how assumed demographic histories impact recombination rate inferences using pyrho. Thus, we performed additional inferences of *r* on the empirical data considering different

demographic models. We summarized recombination rates by the average rate per each chromosome under different demographic models (Fig. 2.6.). We inferred rates under the three demographic models: **1)** demography inferred from SMC++ (left panel A), **2)** the SMC++ inferred demography with recent demography replaced with estimates of population size previously inferred (Mooney et al 2021) with the software IBDNe (Browning and Browning 2015) (middle panel A), and **3)** and a constant N_e of the smallest dog N_e and the largest N_e for wolves previously inferred population size of similar populations from IBDNe (right panel A). In the two most realistic models, those generated with SMC++, we observe that the mean, per-chromosome wolf recombination rate is 11+ times (model 1) and 25+ times (model 2) larger than the median rates in pugs (Fig. 2.6. Left and middle Panel B). Under these two models, the range of average chromosome rates does not overlap between wolves and pugs. Only when we inferred under model 3, the demography of a constant population size, the mean chromosome-wide values are very similar (wolves = 3.0e-8 and pugs = 5.5e-8) (Fig. 2.6. Right panel B). Only under such an extreme model will the recombination rates be the same. There is no evidence for such a difference in N_e (e.g Gray et al 2009, Freedman et al 2016, Mooney et al 2021). Additionally, we considered additional demographic models, such as the one in the middle panel of Figure 2.6, but using a smaller ancestral population size for pugs, and found broadly similar patterns to those shown here (Fig. S2.6). In sum, these results support that the demographic model used can influence our estimates of *r*; however, analysis under likely demographies recapitulates the pattern of diminished dog recombination rate relative to wolves.

Discussion

In this work we used patterns of linkage disequilibrium to examine patterns of meiotic recombination in four dog breeds (border collies, labs, pugs, and Tibetan mastiffs) and a north american population of wolves. Estimates of recombination rates for certain breeds are similar to those of wolves, while for other breeds, such as the border collie and pug, estimated recombination rates appear to be significantly different from those in wolves. However, the local patterns of recombination rate variation along the genome appear to be more concordant across populations. We use simulations to test the role of various potential confounders of estimation of recombination rate from LD, such as complex demographic history, the presence of ROHs in dogs, or a misspecification of mutation rates. While some of these confounders appear to influence the average estimates of *r*, they cannot explain all of the observed patterns. Altogether our results raise the possibility that the intensity of recombination rates may have shifted in breed dogs following domestication.

To assess the validity of our estimates of the local recombination rate along the genome, we compared them to a previously generated genetic map from pedigrees. All populations, except for pugs, showed a moderate to high correlation (R^2 ranging from 0.425 to 0.612) with the inferred pedigree-based maps. Interestingly, while Campbell *et al.* (2016) found an R^2 of 0.740 between their pedigree-based map and an LD based map in village dogs (Auton et al 2013), they found an R^2 of 0.562 when comparing their pedigree-based map to another pedigree-based map. Auton *et al.* (2013) compared their LD based genetic maps of village dogs to a pedigree based map from

microsatellite data of mixed-breed dogs (Wong *et al.* 2010), and found a correlation of R^2 = 0.76 at the 5 MB scale). The correlations between LD-based and pedigree based genetic maps reported in the literature for different species seem to vary from very high in humans (R^2=0.966 at the 5Mb scale; Myers 2005) to more intermediate for flycatchers (R2 of 0.38 at the 200kb scale; Kawakami *et al.* 2017) and Drosophila (R2=0.4-0.6 at a finer scale; Chan *et al.* 2012; note that recombination rates and polymorphism rates are higher in flycatcher and Drosophila, thus making the smaller scale comparisons relevant for these species). Thus, the concordance between our LD-based maps of recombination in wolves and three of four breed dogs with the pedigree-based genetic map is generally concordant with what has been observed for different species in the literature.

We compared the concordance local patterns of inferred recombination across populations. Variation in local recombination rate is documented at the population level, particularly on fine scales, in species such as fruit flies and birds (Chan, Jenkins, & Song, 2012; Kawakami et al., 2017; Samuk, Manzano-Winkler, Ritz, & Noor, 2020). In many vertebrate species, differences in recombination rate hotspot locations can be attributed to shifting binding motifs in the protein PRDM9 (Baker, Walker, Kajita, Petkov, & Paigen, 2014; Grey, Baudat, & de Massy, 2018; Oliver et al., 2009; Parvanov, Petkov, & Paigen, 2010). However, domestic dogs present an interesting case study because of canids' lack of a functional PRDM9 protein (Muñoz-Fuentes, Rienzo, & Vilà, 2011). We instead focus on broader scale patterns across species.

We found high correlations between the rankings of 5 MB (ordered by average weighted recombination of a given window) between wolves and our three non-Pug populations (*R* range = 0.69 to 0.84). Pugs in contrast have an *R* of only 0.40 with wolves. However, the lower *R* for pugs could be due to the limited range of inferred recombination rate rates making the ranks of recombination rates for genomic windows less meaningful. The concordance of the recombination landscape between dogs and wolves is comparable to what has been seen when comparing sister taxa of chimp and bonobo. Specifically, estimates of *r* in great ape populations find a spearman rank *R* of ~0.7 between sister taxa of chimps and bonobos (Stevison *et al.* 2015). Thus, overall, broadly speaking, the landscape of recombination rates seems to be fairly similar across dogs and wolves.

When considering the average rate of recombination genome-wide, wolves, labs, and Tibetan mastiffs all share similar ranges and median values of chromosome wide *r*. In contrast, border collies exhibit drastically higher and pugs exhibit notably reduced inferred recombination rates. Differences in average genome-wide recombination rates between populations have been noted previously. Specifically, Samuk et al found an 8% difference in average recombination rates between populations of *Drosophila pseudoobscura* (Samuk *et al.* 2020). We find a 196% difference in average recombination rates between pugs and border collies . And compared to wolves, pugs have an 186% difference and border collies 117%. Thus, the differences in estimated recombination rates in canids appear extreme relative to previous comparisons in flies.

Given the surprising variation of average estimated recombination rates across canids, we examined several possible explanations. The first possibilities focus on technical issues including data quality and biases of our analysis pipeline. We then consider differences in biological factors, such as demographic history and mutation rates.

First, we considered variation in data quality. Quality of samples measured as the number of individuals sampled and mean coverage were comparable for all populations, ranging from 16X to 41.6X (Table 2.1). We then examined genetic diversity measured as Watterson's θper called site. Our findings reflect that wolves have the most diversity, which is expected given the impact of breed formation on decreasing genetic diversity. Border collies, labs, and Tibetan mastiffs have similar values of θ to one another while θ in pugs is about half as much.

We next investigated the level of observed LD for SNPs 50KB +/- 5KB apart to understand how LD varied between our populations. Because our methods of investigating recombination were based on LD patterns, we expected our inferences of recombination rates to correlate inversely with overall levels of LD. Given pugs' low estimated recombination rates, we expected for pugs to have more LD than our four other canid populations. Indeed, that is what we found (Table 2.1). We then normalized LD by Watterson's θ per site. Assuming the same mutation rate (μ) across populations, Watterson's θ provides a proxy for the effective size of each population. Pugs have the highest ratio of r^2/θ out of all 4 populations, suggesting that pugs have an usually high amount of LD compared to their effective population size. This result supports the pyrho

analysis, where the exceptionally high levels of LD relative to the demographic model lead to pugs having a lower recombination rate. Overall, these metrics suggest that estimated differences in recombination rates are not caused simply by differences in data quality.

We next used simulations to test inferred recombination rates were driven by runs of homozygosity (ROHs) and demography. For these simulations, we focused on wolves and pugs as one of the differing dog breeds. Our simulations used known recombination rates, so we could understand how our inference varied from ground truth.

Our simulations for the two populations initially used the demographic histories inferred in SMC++ from the empirical genomic data. Recombination rates estimated for the wolves and pugs were not notably different from each other. In both the case of pugs and wolves, we did estimate recombination rates lower than ground truth suggesting that pyrho may be underestimating *r* in both dogs and wolves.

We explored this underestimation of true recombination rates by analyzing recombination rates while assuming different demographic histories. When recombination rates were inferred assuming a constant population size of a given population's most recent N_e, estimates of r were closer to the true value. This relation is driven in part by the inverse relationship of N_e and *r* given a constant value of rho. While this resulted in more accurate inferences of recombination rates, it did not lead to differences in the estimates of *r* between pugs and wolves, and thus does not explain

the recombination patterns observed in the empirical data. However, since these simulations showed that recent population sizes had a strong relationship with inferences of recombination rate, we examined the recent N_e for each canid population. Notably border collies, have the smallest estimate of recent N_e which may lead to larger estimates of recombination rates (Table 2.1). Pugs, however, do not have the highest recent N_e.

Next, we examined how more elaborate demographic histories impacted recombination inference using pyrho. We re-inferred *r* using pyrho on the empirical data assuming previously published demographic histories inferred for dog and wolf populations (instead of the demography inferred by SMC++). While the assumed demographic histories affected the inferred values (Fig. 5, Fig. S2.5), we found the rates were higher in wolves. Unlikely demographic models could cause our inferences of recombination rate between wolves and pugs to converge. Only when assuming a constant N_e of our overall largest population size for wolves and smallest for pugs, we did see a change in the difference of estimate of *r* with a decrease for wolves and increase for pugs. But These values we selected were on the extreme ends of those inferred for these populations and thus unlikely to be accurate reflections of the overall demographic history of these populations. Furthermore, inferring *r* under these demographic models results in average estimates of *r* that nearly double what is observed in other mammalian species (e.g. Dumont and Payseur 2008). Thus, while we found that the inference of average *r* using pyrho is sensitive to the assumed demographic history, the

demographic models we were able to examine do not recapitulate the large differences we estimate in recombination rates for wolves versus pugs and border collies.

Another potential explanation for some of the variation in the inferred recombination rates we see across breeds comes from the practices of breed formation and maintenance. One way this happens is through the demographic history and inbreeding. Purebred dogs that are registered with the American Kennel Club may have pedigrees dating as far back as 1875 (American Kennel Club 2017). The nonrandom mating of breed dogs includes the maintenance of specific lines within a given breed. Additionally, dog breeders commonly preserve sperm for future use in artificial insemination meaning a dog can sire pups after its death. A pedigree study of purebred dogs found in labs that only 8% of males are sires and the most popular sire had nearly 2,000 offspring (Calboli et al 2008). Dams are also limited to a subset of a breed population, but the percentage of females used as dams is about double compared to sires. While these breeding practices are done typically to meet the breed standards and the aesthetic preferences of a given breeder, loci of the genome unrelated to those phenotypes are also being subjected to this selection and drift, potentially affecting LD patterns. We found that large ROHs generated from recent inbreeding are unlikely to bias estimates of the average recombination rate (Fig. 2.5, Fig. S2.5).

Another way that breed formation could affect LD is through selective sweeps. As expected with the intense artificial selection of dogs, there has been an increase in deleterious variation near loci under selection due to selective sweeps (Marsden *et al.*

2015). Selective sweeps are known to bias patterns of LD in different ways such as increasing LD during a sweep (Sabeti *et al.* 2002) and then eliminating LD once an advantageous allele has reached fixation (Przeworski 2002, Kim and Neilson 2004, Stephan, Song, and Langley 2006, McVean 2007). Additionally in the case of soft sweeps, LD patterns are affected (Pennings and Hermission 2006). While sweeps alter patterns of LD in the canid genome, we do not believe it to have biased our overall findings of variation in recombination rates in some breed dogs. The patterns of decreased recombination in pugs and increased recombination in border collies occurs across the entire genome. Selective sweeps impact only a small portion of the genome. Thus masking of sites undergoing selective sweeps may yield finer resolution of the differences in recombination, it is highly unlikely that sweeps alone could be biasing our results of these genome wide differences. However, it is possible that dog breeders have unintentionally selected for breeds that have recombination rates or related genomic patterns different from wolves in some but not all breeds. Interestingly, previous work by Mooney *et. al* (2023) analyzing these data sets found that pugs had the largest weighted F_{st} of the four dog breeds relative to the wolf population (Table 2.1). This suggests that pugs may be our most genetically distant dog population to wolves.

Another potential factor that could affect the estimates of the average *r* is the mutation rate assumed in the inference. Specifically, a value of μ is used to calibrate estimates of N_e via SMC++ based on observed genetic diversity. As we have shown, estimates of N_e can influence our estimates of *r* primarily because pyrho calculates *r* from an inferred

value of rho=4Nr that is calibrated by N_e. Thus, through the intermediate term of N_e, our assumptions about mutation rate can potentially impact our inferences of recombination rates. For our analyses, we generally assumed the mutation rate of our dog and wolf populations to be the same as that measured in a different North American wolf population (Koch *et al.* 2019). Previous work has found in mammals that phylogenetic distance is a strong predictor of differences in mutation spectra (Beichmen et al 2023) suggesting that dogs and wolves likely do have mutation rates more similar to one another than compared to more distantly related taxa. Despite a likely similarity in mutation patterns relative to other taxa, differences in the mutation spectra have been observed between human populations (Harris 2015; Harris and Pritchard 2017). Thus, we simulated pug genomes under a true mutation rate that did not match the value of μ assumed in our inferences of *r*. Regardless of the direction, our analysis of simulated data showed an increased recombination rate estimates for pugs, which is the opposite direction of the observed differences in the empirical data. Thus differences in inferred recombination rates are likely not explained by solely a shift in mutation rate between dogs and wolves.

In any event, our study suggests more direct methods are required to falsify the hypothesis that recombination rates have shifted in canid evolution. Pedigree-based genetic maps in pugs would provide direct estimates of the recombination rate that could then be compared to those in other breeds. Similarly, trio-based whole-genome sequencing studies could be used to infer whether μ differs from that inferred in wolves (Lindsay *et al.* 2019, The 1000 Genomes Project 2011, Suárez-Menéndez *et al.* 2023,

Wang *et al.* 2022, Bergeron *et al.* 2023). If direct estimates of these parameters in pugs appear similar to those of other dog breeds, it would indicate that misspecification of some key parameter, perhaps in the demographic model, is accounting for the increased LD and decreased *r* in pugs.

Our work builds on recent studies (Wall and Stevison 2016, Dapper and Payseur 2017, Samuk and Noor 2022, and Raynaud *et al.* 2023) suggesting care is required when inferring recombination from patterns of LD. While Dapper and Payseur focused on inference of hotspots, our work suggests that the average rate of recombination appears to be highly sensitive to the assumed demographic history of the population. We saw this when pyrho systematically underestimated the recombination rate in simulated data under certain demographies but not others (Fig. 2.5, Fig. S2.5). This same trend was observed in the analysis of the empirical data, where the estimates of *r* also depended on the assumed demographic model (Fig. 2.6). Our simulations suggest that assuming a constant population size equivalent to the recent effective population size may provide more reliable estimates of the average *r* than using a demographic history inferred from patterns of polymorphism. Further work is required to test how general this conclusion is across demographic histories.

While inference of average recombination rates is challenging for the reasons discussed above, we found that the local landscapes of recombination along the genome are fairly similar across dog and wolf populations. Further, three of 4 dog breeds showed similar or lower average recombination rates compared to wolves. Only the border collie

appeared to have potentially higher rates of recombination than the wolves. Thus, overall, our study provides little support for Gornall's hypothesis (Gornall, 1983) that recombination is higher in domesticated species.

Materials and Methods

Genomic data

We analyzed previously published data on 60 canid genomes comprising 15 wolves (Robinson *et al.* 2019) and 45 breed dogs. We examined four breeds consisting of: 10 border collies (Plassais *et al.* 2019), 10 labradors (Plassais *et al.* 2019), 15 pugs (Marchant *et al.* 2017), and 10 Tibetan mastiffs (Phung *et al.* 2019). The genomes were previously sequenced at high coverage, and reads were previously processed, filtered, and aligned to the canFam3.1 dog reference genome as described in Marsden *et al.* 2016. To call single nucleotide polymorphisms (SNPs), we used GATK (McKenna *et al.* 2010) and retained only biallelic SNPs while excluding indels for these analyses. For more detailed description of genotype and variant calling, see Mooney *et al.* 2023.

Recombination rate inference and demographic inference

To infer the fine-scale recombination maps, we used these unphased high coverage polymorphism data from unrelated individuals by analyzing patterns of linkage disequilibrium (LD). We used pyrho (Spence and Song 2019) for recombination inference. pyrho accepts non-equilibrium demographic histories (Spence and Song 2019). Therefore, to account for the changes in N_e through time, we used the software SMC++ (Terhorst, Kamm, and Song 2017) and, for each population, we estimated their

demographic histories based on the joint inference of all autosomes (38 autosomes in total). We set the mutation rate () to 4.5e-9 based on previous estimates in wolves (Koch *et al.* 2019). When computing a lookup table in pyrho, we used the manual recommendation of calculating statistics of LD and rho based on a population size that was 50% larger than our sample size and then down sampled to our population size. For the final step of inferring *r* in pyrho, we used a window size and block penalty of 50.

Comparisons of the genetic maps

To compare the concordance of different maps, we first binned the genome into non-overlapping windows. We used three window sizes: 100 kb, 1MB, and 5MB. Windows were created only for segments of the chromosome sequenced in both data sets being compared. Then the estimates of *r* within that window were multiplied by the number of basepairs with the same *r* value. Then the sum of the *r* times length values were divided by the total window size. Our formula for *n* values of *r* within a given window was

$$\frac{(r_1 * length_1) + (r_2 * length_2) + ... (r_n * length_n)}{window\ size}$$

This yielded a mean weighted value of *r* for each window. For each population, we calculated a rank for every window. For a given window we then compared the ranks for pairs of populations. We compared the four breed dog populations against the wolf population. We also compared the inferred recombination maps from the breed dogs and wolves against previously published mixed-breed dog recombination maps inferred from a pedigree (Campbell *et al.* 2016). We calculated recombination rates per window from the results of Campbell *et al.* using the same formula above. For each pairwise comparison, we fit a linear model with the lm function in RStudio.

Simulated wolf and pug data

To test the impact of multiple variables such as demography, mutation rate, and runs of homozygosity (ROHs) on pyrho inferences, we performed coalescent simulations of wolves (simulated Wolf) and pugs (simulated Pug). We selected pugs as a case study for dogs because their inferred recombination rates differed from wolves. We used msprime (Kelleher, Etheridge, & McVean 2016; Baumdicker *et al.* 2022) to simulate genetic variation data. For both populations, we simulated 40 contigs of 20 MB each with a population size of 15 diploid individuals. We choose these parameters to approximate the size and number of chromosomes in the wolf and dog genome as well as the sample size of our study populations.

We simulated the data under a constant recombination rate or 1×10^{-8} bp/generation. This rate is close to the mean rate previously described in mixed breed dogs in Campbell *et al* 2016. Unless otherwise stated, we performed 10 replicates of the

simulation for each set of parameters. After generating the simulated canid genomes, we followed our analysis pipeline described above to infer recombination rates. Below we detail the different scenarios we explored with the simulations.

Simulations with changing demography

To understand how demography affects the inference of recombination rates using pyrho, we simulated genomes under our inferred demography from SMC++ for wolves and pugs. All other parameters were held constant per above and we followed our analysis pipeline to estimate recombination rates on the simulated data.

To determine how the demographic model assumed in pyrho affects our analysis of recombination rates, we re-analyzed this same simulated data, but supplied pyrho with the true demography used to simulate the data, rather than a demography inferred from the simulated data. We also repeated the pyrho inference assuming a constant size of the true most recent population size for a given population.

ROHs

Dogs have notable runs of homozygosity (ROHs) as a result of inbreeding during breed formation (Sams and Boyko 2019). To test their impact on our analysis of recombination rates as inferred from LD patterns, we altered the simulated dog genomes to include ROHs. We based the sizes and proportion of runs on previously published data (Sams and Boyko 2019). To randomly select locations in the genome, we used the shuffle function from the program bedtools (Quinlan and Hall 2010). Then using a custom R

script, we altered VCFs for a random individual at each location to a genotype of 0/0. We then proceeded with our analysis pipeline where we inferred demography using SMC++ and then inferred *r* using pyrho.

The simulation described above placed ROHs at random locations in each individual. We also tested the impact of ROHs overlapping between all individuals in a given population. To do this, we randomly selected one individual and duplicated the ROHs previously inserted into that individual into the original VCFs of the 14 other individuals. We then inferred recombination rates using the pipeline described above.

Misspecified mutation rate

The mutation rate used in our simulations and analyses was measured in a wolf pedigree (Koch *et al.* 2019) distinct from the wolves analyzed here. Thus, we investigated the effect of having a true mutation rate in our simulated data that did not match the mutation rate used in the inferences of recombination. For pugs, we simulated data under the following mutation rates: 4.5e-10 per bp per generation (1/10th of our assumed μ for recombination rate inference), 2.25e-8 per bp per generation (5X our assumed μ), and 4.5e-8 per bp per generation (10X our assumed μ). For wolves, we tested a mutation rate of 4.5e-10 per bp per generation (1/10th of our assumed μ for recombination rate inference). We then continued with our analysis pipeline assuming the published wolf mutation rate of 4.5e-9 per bp per generation.

Data Availability

Detailed scripts and set parameters are available at GitHub

https://github.com/cad17/canid_recombination#canid_recombination

Data from Robinson *et al.* 2019 are available on SRA under PRJNA512209; from Marchant *et al.* 2017 are available on European Nucleotide Archive (ENA) under PRJEB17926; and from Plassais *et al.* 2019 are available on SRA under PRJNA448733.

Software availability:

SMC++: https://github.com/popgenmethods/smcpp

msprime: https://tskit.dev/msprime/docs/stable/intro.html

pyrho: https://github.com/popgenmethods/py

Figures

Figure 2.1

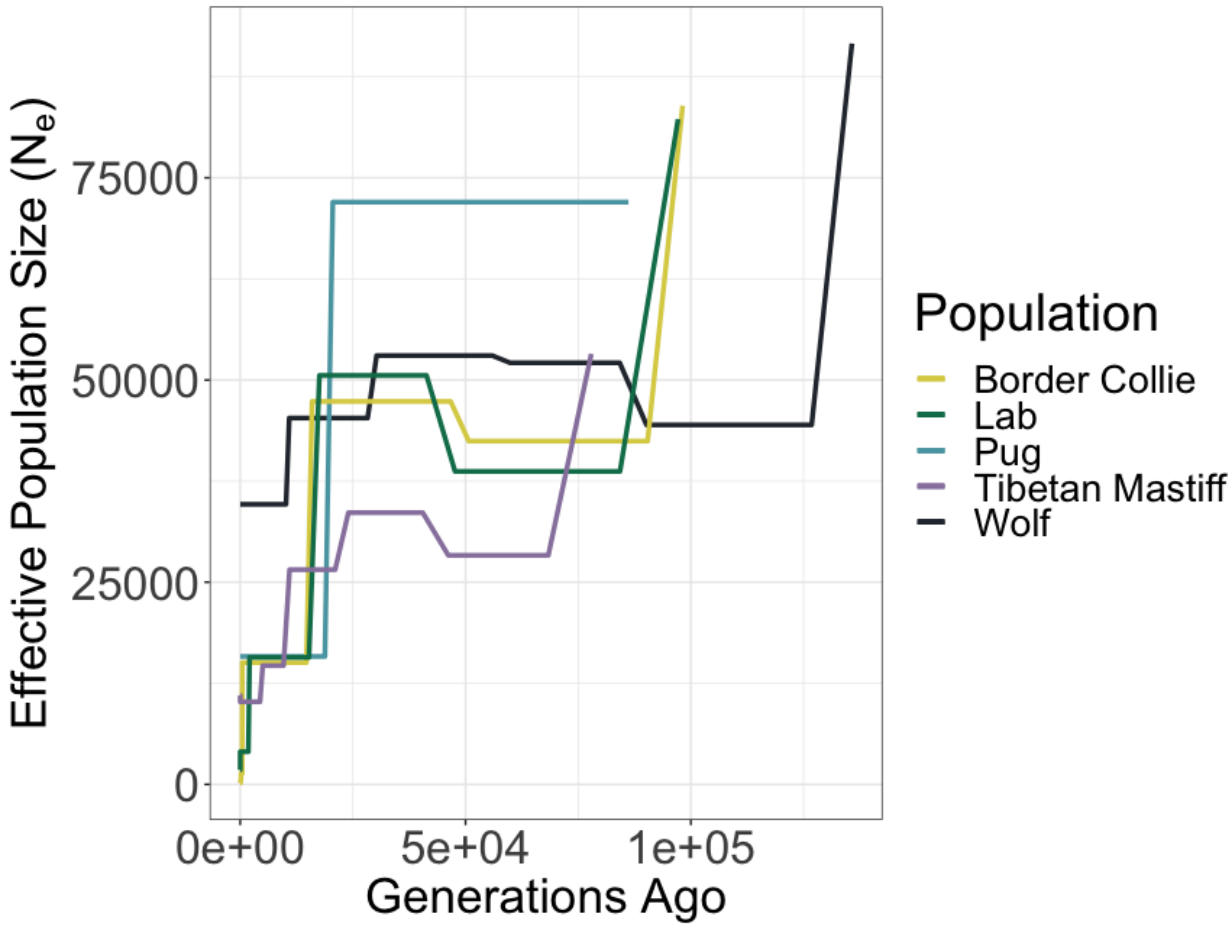

Fig 2.1. Population size histories inferred with SMC++ for wolves (black), border collies (yellow), labs (green), pugs (blue), and Tibetan mastiffs (purple). The x axis is measured as generations ago with present time on the left and older time on the right. All dog populations show a bottleneck around 15,000 generations ago, likely corresponding with domestication.

Figure 2.2

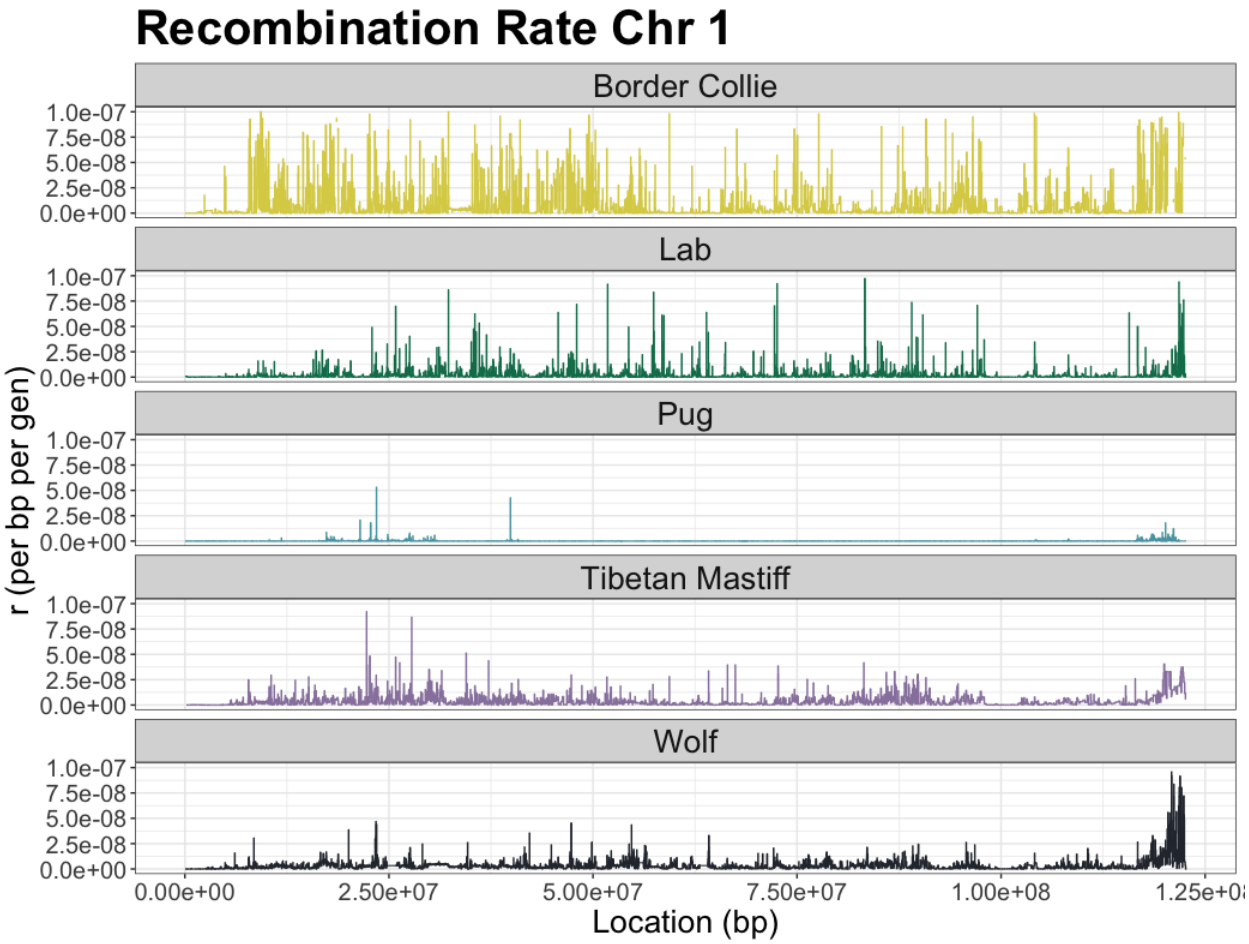

Fig 2.2. Inferred recombination rates of chromosome one for border collies (yellow), labs (green), pugs (blue), Tibetan mastiffs (purple), and wolves (black). The y-axis is the same for all populations. Some values for border collies that exceed the y-axis are not shown here. Relative to the other populations, border collies consistently have high recombination rates while pugs have consistently low recombination. All groups show an increase in recombination at the telomeric (right) end of the chromosome similar to observations in previous studies. See Figure S2.1 for similar plots for other chromosomes.

Figure 2.3

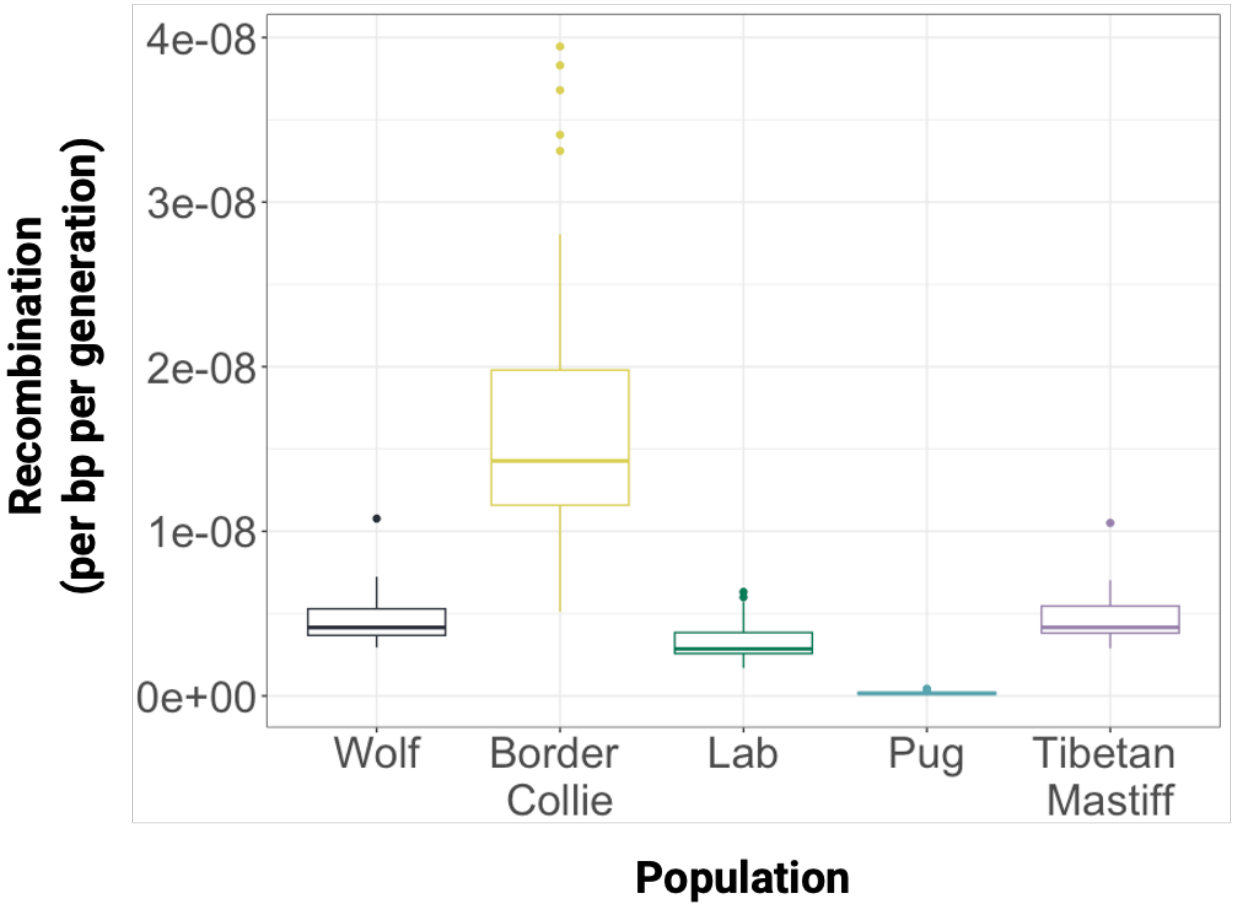

Fig 2.3. Weighted average recombination rates inferred for all 38 autosomes of wolves (black), border collies (yellow), labs (green), pugs (blue), and Tibetan mastiffs (purple). Each data point represented is the weighted average recombination rate for a single chromosome. Thick horizontal lines represent median values. Boxes represent second and third quartiles. Whiskers represent first and fourth quartiles. Dots represent outliers. Ranges and median values are similar between wolves, labs, and Tibetan mastiffs. Border collies and pugs have median values that do not overlap with the ranges of other populations.

Figure 2.4

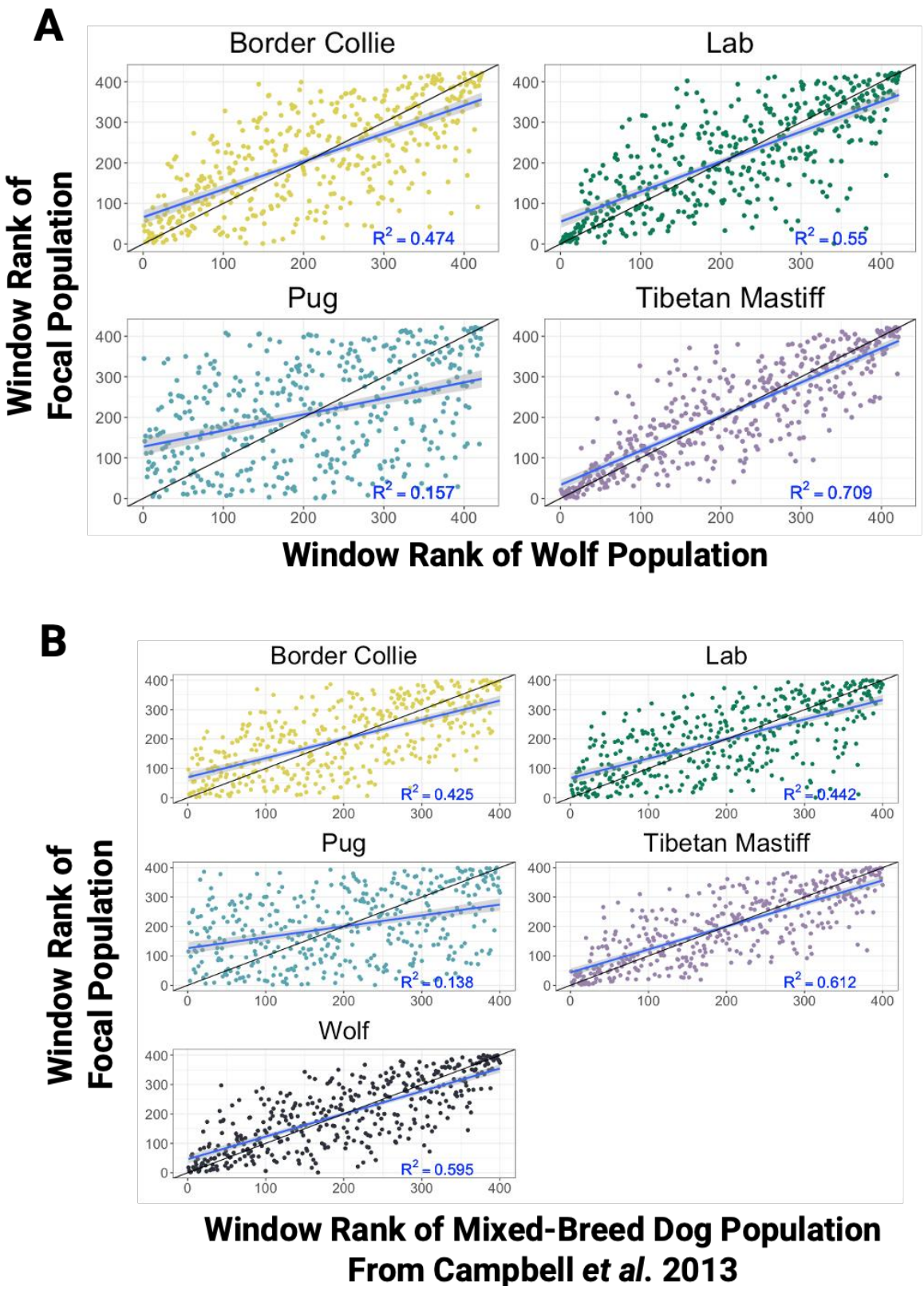
A
Border Collie
Lab
Pug
Tibetan Mastiff
R² = 0.474
R² = 0.55
R² = 0.157
R² = 0.709
Window Rank of Focal Population
Window Rank of Wolf Population
B
Border Collie
Lab
Pug
Tibetan Mastiff
Wolf
R² = 0.425
R² = 0.442
R² = 0.138
R² = 0.612
R² = 0.595
Window Rank of Focal Population
Window Rank of Mixed-Breed Dog Population
From Campbell et al. 2013

Fig 2.4. Scatter plot of Ranks of inferred recombination rates in 5MB genomic windows for each dog breed vs wolves (panel A) and those dog breeds plus wolves vs a pedigree-based genetic map (panel B, Campbell *et al.* 2013). Populations are color coded (border collie = yellow, lab = green, pug = blue, Tibetan mastiff = purple, Wolf = black). Lower ranks correspond to higher recombination rates. Each point represents a non-overlapping window in the canid genome and plots the rank of that specific window between the two populations. For panel A, the y-axis corresponds to the window's rank in a given dog breed, and the x-axis is the same window's rank for wolves. In panel B, the y-axis corresponds to the window's rank in breed dogs or wolves, and the x-axis is the same window's rank from the pedigree-based map. The blue line represents the trend line of the data with gray shading representing the 92.5 confidence intervals. The R^2 value of the linear regression is printed on each grid. The black line is the y=x line. All trend lines have a positive slope suggesting some correlation between the "hotter" areas of the dog genomes and the "hotter" areas of the wolf genome.

Figure 2.5

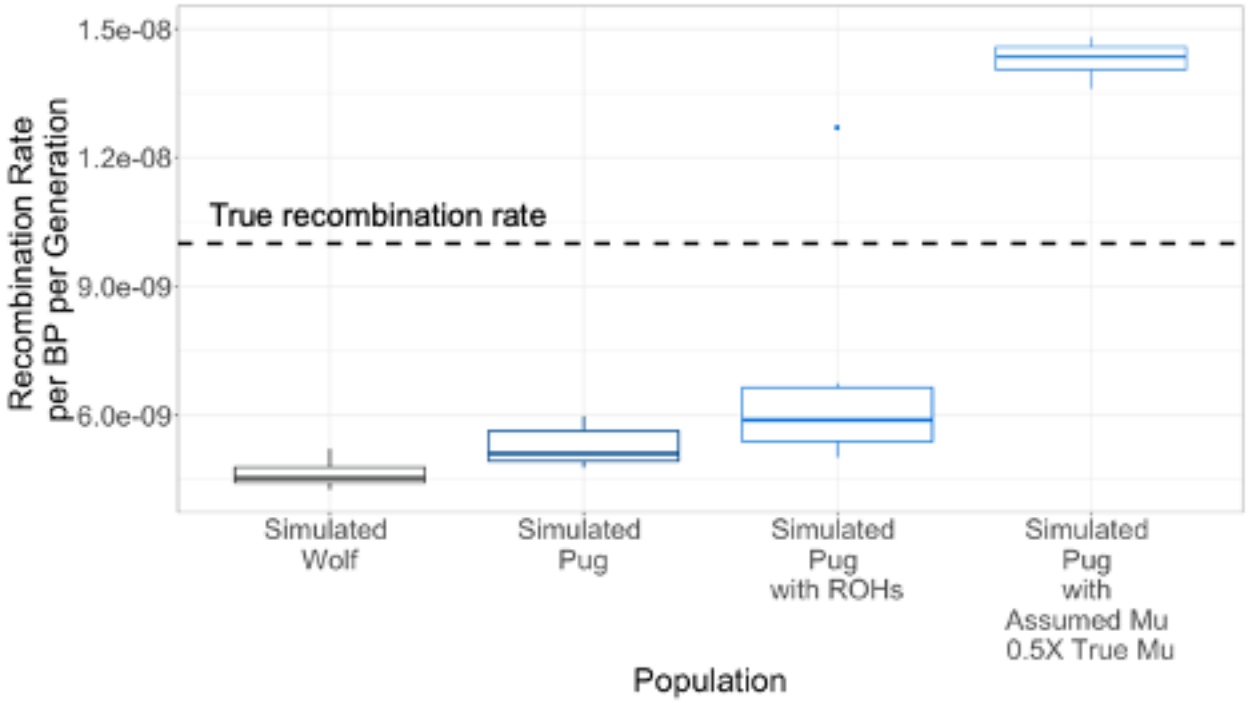

Fig 2.5. Recombination rates estimated from simulated canid populations with different parameters. The first condition (Simulated Wolf) was created using the demographic model inferred from wolves with SMC++. The second condition (Simulated Pug) used the demography inferred from pugs. The third condition (Simulated Pug with ROHs) added ROHs similar to those seen in dogs to the simulated pug data. Lastly, we simulated under the pug demography with a true mutation rate of 2.25×10^{-8} /bp/generation and then inferred *r* assuming a mutation rate that was 4.5×10^{-9} /bp/generation (Simulated Pug with Assumed Mu 0.5X True Mu). For each condition we simulated 10 replicates. The points shown for each condition represent the median chromosome-wide estimate of recombination for a given replicate. Note that one outlier of 7.2×10^{-8} is omitted for the ROH population for plot scaling purposes. The ranges of recombination rates estimated for these four conditions are all overlapping.

Figure 2.6

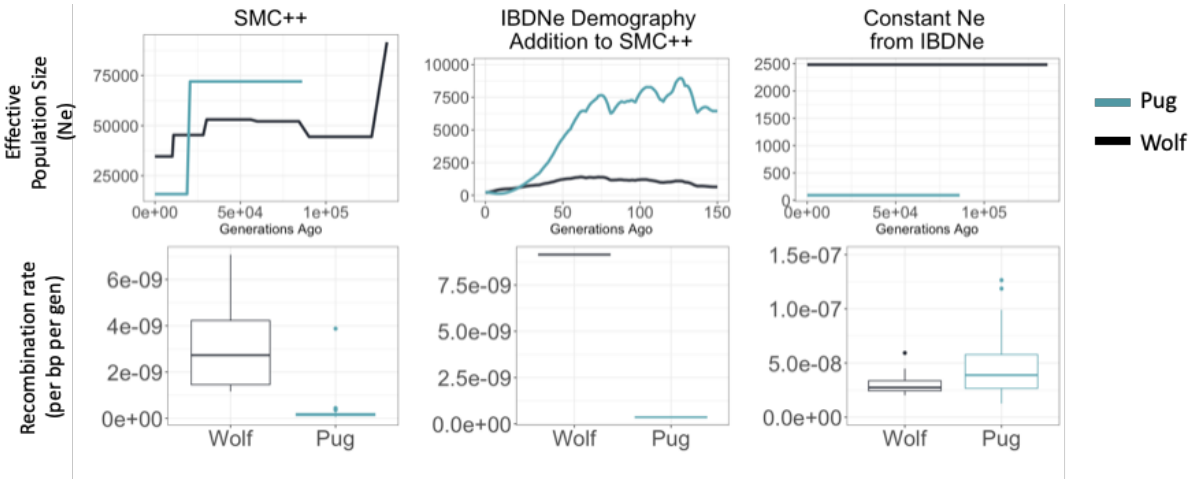

Fig 2.6. Sensitivity of inferred recombination rates to the assumed demographic model. Demographic models used for inferring recombination rates (top plots) and corresponding inferences of recombination rates (bottom plots). These results are generated from the same sequence data for wolves (black) and pugs (blue) using the different demographic models shown. For demographic plots, the x-axis is measured as generations ago with present time on the left and older time on the right. The y-axis represents estimated effective population sizes. For recombination plots, each data point represented is the weighted average recombination rate for a single chromosome. Thick horizontal lines represent median values. Boxes represent second and third quartiles. Whiskers represent first and fourth quartiles. Dots represent outliers. SMC++ plots (left) follow the recommended pyrho analysis pipeline. The middle plots represent when we use both our population sizes from SMC++ and recent population sizes inferred in related populations using the program IBDNe (Mooney *et al.* 2021). On the right, we show results using a fabricated demographic model can lead to similar

inferences of recombination rates in our wolf and pug populations.

Tables

Table 2.1

Summary of genetic variation data across populations.

	Border Collie	Lab	Pug	Tibetan Mastiff	Wolf
Median chr wide r	1.40E-08	3.00E-09	2.00E-10	4.00E-09	4.00E-09
N	10	10	15	10	15
Median individual average read depth	23.7	29.3	46.1	16	37.3
Recent Ne	171	1799	714	11,007	1605
Watterson's Theta per site	8.98E-04	9.14E-04	6.34E-04	1.08E-03	1.44E-03
Weighted Fst v Wolves*	0.296	0.311	0.421	0.237	NA
r2 of SNPs 50kb +/- 0.5KB Apart	0.27	0.26	0.44	0.2	0.15
r2 / Watterson's Theta per site	300.53	284.51	694.37	185.48	103.82

*Data from Mooney et. al 2023

References

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., … McVean, G. (2012). A fine-scale chimpanzee genetic map from population sequencing. *Science*, *336*(6078), 193–198. https://doi.org/10.1126/science.1216872

Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J. K., … Boyko, A. R. (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genetics*, *9*(12). https://doi.org/10.1371/journal.pgen.1003984

Baker, C. L., Kajita, S., Walker, M., Saxl, R. L., Raghupathy, N., Choi, K., ... & Paigen, K. (2015). PRDM9 drives evolutionary erosion of hotspots in Mus musculus through haplotype-specific initiation of meiotic recombination. *PLoS genetics*, *11*(1), e1004916.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., ... & Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229.

Beichman, A. C., Robinson, J. A., Lin, M., Moreno-Estrada, A., Nigenda-Morales, S., & Harris, K. (2023). Evolution of the Mutation Spectrum Across a Mammalian Phylogeny, *Molecular Biology and Evolution,* 40(10), msad213

Browning, S. R., & Browning, B. L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, *97*(3), 404-418.

Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., ... & Zhang, G. (2023). Evolution of the germline mutation rate across vertebrates. *Nature*, *615*(7951), 285-291.

Calboli, F. C. F., Sampson, J., Fretwell, N., & Balding, D. J. (2008). Population structure and inbreeding from pedigree analysis of purebred dogs. *Genetics*, *179*(1), 593–601. https://doi.org/10.1534/genetics.107.084954

Campbell, C. L., Bhérer, C., Morrow, B. E., Boyko, A. R., & Auton, A. (2016). A pedigree-based map of recombination in the domestic dog genome. *G3: Genes, Genomes, Genetics*, *6*(11), 3517–3524. https://doi.org/10.1534/g3.116.034678

Cavassim, M. I. A., Baker, Z., Hoge, C., Schierup, M. H., Schumer, M., & Przeworski, M. (2022). PRDM9 losses in vertebrates are coupled to those of paralogs ZCWPW1 and ZCWPW2. *Proceedings of the National Academy of Sciences*, *119*(9), e2114401119.

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLoS Genetics*, *8*(12). https://doi.org/10.1371/journal.pgen.1003090

Clutton-Brock, J. (1992). The process of domestication. *Mammal review*, *22*(2), 79-85.

Dapper, A. L., & Payseur, B. A. (2017). Connecting theory and data to understand recombination rate evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160469.

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature*, *418*(6898), 700-707.

Dreissig, S., Mascher, M., Heckmann, S., & Purugganan, M. (2019). Variation in Recombination Rate Is Shaped by Domestication and Environmental Conditions in Barley. *Molecular Biology and Evolution*, *36*(9), 2029–2039. https://doi.org/10.1093/molbev/msz141

Dumont, B. L., & Payseur, B. A. (2008). Evolution of the genomic rate of recombination in mammals. *Evolution*, *62*(2), 276-294.

Freedman, A. H., Gronau, I., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Silva, P. M., ... & Novembre, J. (2014). Genome sequencing highlights the dynamic early history of dogs. *PLoS genetics*, *10*(1), e1004016.

Freedman, A. H., Lohmueller, K. E., & Wayne, R. K. (2016). Evolutionary history, selective sweeps, and deleterious variation in the dog. *Annual Review of Ecology, Evolution, and Systematics*, *47*, 73-96.

Fuentes, R. R., de Ridder, D., van Dijk, A. D., & Peters, S. A. (2022). Domestication shapes recombination patterns in tomato. *Molecular biology and evolution*, *39*(1), msab287.

Gornall, R. J. (1983). Recombination systems and plant domestication. *Biological Journal of the Linnean Society*, *20*(4), 375–383. https://doi.org/10.1111/j.1095-8312.1983.tb01598.x

Gray, M. M., Granka, J. M., Bustamante, C. D., Sutter, N. B., Boyko, A. R., Zhu, L., ... & Wayne, R. K. (2009). Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, *181*(4), 1493-1505.

Grey, C., Baudat, F., & de Massy, B. (2018). PRDM9, a driver of the genetic map. *PLoS Genetics*, *14*(8), e1007479. https://doi.org/10.1371/journal.pgen.1007479

Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences*, *112*(11), 3439-3444.

Harris, K., & Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *Elife*, *6*, e24284.

Kawakami, T., Mugal, C. F., Suh, A., Nater, A., Burri, R., Smeds, L., & Ellegren, H. (2017). Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. *Molecular Ecology*, *26*(16), 4158–4172. https://doi.org/10.1111/mec.14197

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, *12*(5), e1004842.

Kim, Y., & Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, *167*(3), 1513-1524.

Koch, Evan, Rena M. Schweizer, Teia M. Schweizer, Daniel R. Stahler, Douglas W. Smith, Robert K. Wayne, and John Novembre. 2019. "De Novo Mutation Rate Estimation in Wolves of Known Pedigree." *Molecular Biology and Evolution*, July. https://doi.org/10.1093/molbev/msz159.

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., … Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, *467*(7319), 1099–1103. https://doi.org/10.1038/nature09525

Larson, G., Karlsson, E. K., Perri, A., Webster, M. T., Ho, S. Y., Peters, J., ... & Lindblad-Toh, K. (2012). Rethinking dog domestication by integrating genetics,

archeology, and biogeography. *Proceedings of the National Academy of Sciences*, *109*(23), 8878-8883.

Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T., & Hurles, M. E. (2019). Similarities and differences in patterns of germline mutation between mice and humans. *Nature communications*, *10*(1), 4053.

Marchant, T. W., Johnson, E. J., McTeir, L., Johnson, C. I., Gow, A., Liuti, T., ... & Schoenebeck, J. J. (2017). Canine brachycephaly is associated with a retrotransposon-mediated missplicing of SMOC2. *Current Biology*, *27*(11), 1573-1584.

Marsden, Clare D., Diego Ortega-Del Vecchyo, Dennis P. O'Brien, Jeremy F. Taylor, Oscar Ramirez, Carles Vilà, Tomas Marques-Bonet, Robert D. Schnabel, Robert K. Wayne, and Kirk E. Lohmueller. 2016. "Bottlenecks and Selective Sweeps during Domestication Have Increased Deleterious Genetic Variation in Dogs." *Proceedings of the National Academy of Sciences of the United States of America* 113 (1): 152–57.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, *et al.* 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, *304*(5670), 581-584.

McVean, Gil, Philip Awadalla, and Paul Fearnhead. 2002. “A Coalescent-Based Method for Detecting and Estimating Recombination from Gene Sequences.” *Genetics* 160 (3): 1231–41.

McVean, G. (2007). The structure of linkage disequilibrium around a selective sweep. *Genetics*, *175*(3), 1395-1406.

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, *310*(5746), 321-324.

Mooney, J. A., Marsden, C. D., Yohannes, A., Wayne, R. K., & Lohmueller, K. E. (2023). Long-term small population size, deleterious variation, and altitude adaptation in the ethiopian wolf, a severely endangered canid. *Molecular Biology and Evolution*, *40*(1), msac277.

Mooney, J. A., Yohannes, A., & Lohmueller, K. E. (2021). The impact of identity by descent on fitness and disease in dogs. *Proceedings of the National Academy of Sciences*, *118*(16), e2019116118.

Muñoz-Fuentes, V., Rienzo, A., & Vilà, C. (2011). Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. *PLoS ONE*, *6*(11), 1–7. https://doi.org/10.1371/journal.pone.0025498

Muñoz-Fuentes, V., Marcet-Ortega, M., Alkorta-Aranburu, G., Forsberg, C. L., Morrell, J. M., Manzano-Piedras, E., … Vila, C. (2015). Strong artificial selection in domestic mammals did not result in an increased recombination rate. *Molecular Biology and Evolution*, *32*(2), 510–523. https://doi.org/10.1093/molbev/msu322

Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., … Ponting, C. P. (2009). Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genetics*, *5*(12). https://doi.org/10.1371/journal.pgen.1000753

Parvanov, E. D., Petkov, P. M., & Paigen, K. (2010). Prdm9 controls activation of mammalian recombination hotspots. *Science*, *327*(5967), 835. https://doi.org/10.1126/science.1181495

Phung, T. N., Wayne, R. K., Wilson, M. A., & Lohmueller, K. E. (2019). Complex patterns of sex-biased demography in canines. *Proceedings of the Royal Society B*, *286*(1903), 20181976.

Pennings, P. S., & Hermisson, J. (2006). Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS genetics*, *2*(12), e186.

Plassais, J., Kim, J., Davis, B. W., Karyadi, D. M., Hogan, A. N., Harris, A. C., ... & Ostrander, E. A. (2019). Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature communications*, *10*(1), 1489.

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics*, *160*(3), 1179-1189.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842.

Raynaud, M., Gagnaire, P. A., & Galtier, N. (2023). Performance and limitations of linkage-disequilibrium-based methods for inferring the genomic landscape of

recombination and detecting hotspots: a simulation study. *Peer Community Journal*, *3*.

Robinson, J. A., Räikkönen, J., Vucetich, L. M., Vucetich, J. A., Peterson, R. O., Lohmueller, K. E., & Wayne, R. K. (2019). Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances*, *5*(5), eaau0757.

Ross-Ibarra, J. (2004). The Evolution of Recombination under Domestication: A Test of Two Hypotheses. *American Naturalist*, *163*(1), 105–112. https://doi.org/10.1086/380606

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., ... & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832-837.

Sams, A. J., & Boyko, A. R. (2019). Fine-scale resolution of runs of homozygosity reveal patterns of inbreeding and substantial overlap with recessive disease genotypes in domestic dogs. *G3: Genes, Genomes, Genetics*, *9*(1), 117-123.

Samuk, K., Manzano-Winkler, B., Ritz, K. R., & Noor, M. A. F. (2020). Natural Selection Shapes Variation in Genome-wide Recombination Rate in Drosophila pseudoobscura. *Current Biology*, *30*(8), 1517-1528.e6. https://doi.org/10.1016/j.cub.2020.03.053

Samuk, K., & Noor, M. A. (2022). Gene flow biases population genetic inference of recombination rate. *G3*, *12*(11), jkac236.

Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., Hooper, D. M., … Przeworski, M. (2015). Stable recombination hotspots in birds, *350*(6263).

Spence, Jeffrey P., and Yun S. Song. 2019. “Inference and Analysis of Population-Specific Fine-Scale Recombination Maps across 26 Diverse Human Populations.” *Science Advances* 5 (10): eaaw9206.

Stephan, W., Song, Y. S., & Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, *172*(4), 2647-2663.

Stevison, L. S., Woerner, A. E., Kidd, J. M., Kelley, J. L., Veeramah, K. R., McManus, K. F., ... & Wall, J. D. (2016). The time scale of recombination rate evolution in great apes. *Molecular biology and evolution*, *33*(4), 928-945.

Suárez-Menéndez, M., Bérubé, M., Furni, F., Rivera-León, V. E., Heide-Jørgensen, M. P., Larsen, F., ... & Palsbøll, P. J. (2023). Wild pedigrees inform mutation rates and historic abundance in baleen whales. *Science*, *381*(6661), 990-995.

Terhorst, Jonathan, John A. Kamm, and Yun S. Song. 2017. “Robust and Scalable Inference of Population History from Hundreds of Unphased Whole Genomes.” *Nature Genetics* 49 (2): 303–9.

The 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. *Nat Genet* 43, 712–714 (2011). https://doi.org/10.1038/ng.862

Wall, J. D., & Stevison, L. S. (2016). Detecting recombination hotspots from patterns of linkage disequilibrium. *G3: Genes, Genomes, Genetics*, *6*(8), 2265–2271. https://doi.org/10.1534/g3.116.029587

Wang, R. J., Peña-Garcia, Y., Bibby, M. G., Raveendran, M., Harris, R. A., Jansen, H. T., ... & Hahn, M. W. (2022). Examining the effects of hibernation on germline mutation rates in grizzly bears. *Genome biology and evolution*, *14*(10), evac148.

Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., Bontrop, R. E., … Altshuler, D. (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, *308*(5718), 107–111. https://doi.org/10.1126/science.1105322

Wong, A. K., Ruhe, A. L., Dumont, B. L., Robertson, K. R., Guerrero, G., Shull, S. M., … Neff, M. W. (2010). A comprehensive linkage map of the dog genome. *Genetics*, *184*(2), 595–605. https://doi.org/10.1534/genetics.109.106831

Chapter 3: Impact of Recombination on Inference of the Distribution of Fitness Effects

A Supplemental Appendix is available online as this dissertation's Supplementary Materials: Ch3_Supplementary_Information.pdf

Abstract

Quantifying the distribution of fitness effects (DFE) is a foundational concept in population genetics. Current approaches to inferring the DFE of new mutations depend on the assumption of sites being independent and unlinked. In truth, sites on the same chromosome are often linked unless decoupled by meiotic recombination. It is not fully understood how unmodeled linkage biases DFE inferences. In this work, we find a limited impact of linkage and recombination rate on inferring the DFE. We use wolves as a model due to the relatively stable recombination landscape across canids. This stability stems from a non-functional gene (PRDM9) that is central to evolutionary changes in recombination hotspots in vertebrates. We inferred the DFE based on the whole wolf genome and three subsets of the genome with different ranges of estimated recombination rates: 0 to 1.9 x 10^{-9} (low *r*), >1.9 x 10^{-9} to 4.3 x10^{-9} (med *r*), and >4.3 x 10^{-9} to 2 x 10^{-8} (high *r*) per bp per generation. To condition our DFE models, we estimated the demographic history of these 4 datasets. We found subtle differences in the ratio of ancestral to current population sizes (ν) and time since population size change (τ). There was wider variation in estimated ancestral population sizes (N_a) (range: ~45 K to ~81 K). These differences in N_a estimates may be due to biases from selection affecting the amount of variation at linked neutral sites. We inferred the DFE of

these datasets conditioned on their respective demographic models. We found similar DFE estimates between low and high recombination regions. Despite regions of moderate recombination having overall different DFEs from high and low recombination regions, we found a similar proportion of neutral mutations in all three categories. Thus we find support that inference of the DFE using PRF methods is not notably biased by the effects of background selection.

Introduction

Understanding the distribution of fitness effects (DFE) of new mutations is a foundational topic in population genetics because it describes how selection can shape genetic variation (Eyre-Walker and Keightley 2007). Additionally, the DFE is important for estimating genetic load, modeling introgression, and understanding the strength of selection on disease loci.

A challenge in quantifying the DFE is the effect of demography on patterns of genetic variation that may bias DFE inference and selection, in turn biasing demographic inferences (see review of Johri *et al.* 2021). A two-step process of inferring DFEs can circumvent this issue by using site frequency spectra (SFS) of synonymous variants in exons to infer demographic history and then infer the DFE from the SFS of the nonsynonymous mutations conditioned on the demography (*e.g.* Kim *et al.* 2017, Keightley and Eyre-Walker 2007, Boyko *et al.* 2008, and Li *et al.* 2010). However, this approach of inferring the DFE makes the assumption that sites are independent and unlinked. Specifically, DFE inference using Poisson Random Field (PRF) models

assume mutations are independent of each other (*i.e.* free recombination). This is because the PRF framework is based on the emergence of a Poisson distribution of new mutations each generation, and those allele frequencies are changed only by selection and drift (Hartl *et al.* 1994, Sawyer and Hartl 1992). Intermediate or low levels of recombination will violate this assumption of independence, potentially biasing estimates of the DFE.

It is also well known that there are differences in effective population size along the genome of organisms and that these variations differ among organisms (e.g Charlesworth 2009, Gossman *et al.* 2011, Jiménez-Mena *et al.* 2016). Gossman *et al.* (2011) find that estimates of N_e along the genome are positively correlated to recombination rate in *Drosophila melanogaster* and negatively correlated with the density of selected sites in humans and *Arabidopsis thaliana*.

Mounting evidence supports a widespread prevalence of slightly deleterious alleles in the human genome (Bustamante *et al.* 2005, Eyre-Walker and Keightley 2007, Lohmueller *et al.* 2011, Kim *et al.* 2017). Purifying selection removing these slightly deleterious sites can also cause a decrease of genetic diversity at linked neutral sites through background selection (Charlesworth *et al.* 1995, Charlesworth 2012). The impact of background selection is greatest when recombination rates are low (Charlesworth *et al.* 1995). This is because lower levels of recombination create higher rates of linkage disequilibrium (LD) leading to larger portions of a chromosome sharing the same genealogy as nearby selected sites. In regions of low recombination,

background selection may also cause an increase in rare variants of synonymous sites (Lohmueller *et al.* 2011, Good *et al.* 2014). Additionally, it has been suggested that some linked selection can impact estimates of adaptive evolution and maybe the DFE (Messer and Petrov 2013). However, it has not yet been fully explored to what degree low recombination rates impact estimates of the DFE.

Canids are a useful model for determining the effects of linkage and recombination rate variation on DFE inference because they lack a functional PRDM9 gene (Muñoz-Fuentes, Rienzo, & Vilà, 2011, Cavassim *et al.* 2022). PRDM9 is known in vertebrates to influence location and intensity of recombination hotspots (Baker *et al.* 2015; Cavassim *et al.* 2022; Grey, Baudat, & de Massy, 2018; Oliver *et al.,* 2009; Parvanov, Petkov, & Paigen, 2010). In dogs, recombination hotspots typically occur near transcription start sites (Auton *et al.,* 2013; Campbell *et al.* 2016). All together this suggests that recombination patterns are more stable between canid species versus vertebrate taxa with functional PRDM9 genes.

In this study, we infer the DFE for the wolf genome as well as three subsets of the genome with different ranges of estimated recombination rates. We use fit∂a∂i, a PRF based method, to estimate the DFE. We observed a reduction in diversity in neutral sites in regions of low recombination, consistent with background selection. Despite this, the inferred DFE for low recombination regions of the genome closely resembled that of high recombination regions and the genome-wide DFE. Thus, we find strong

support that PRF-based inferences of the DFE are not biased by violation of the assumption of unliked variants.

Materials and Methods

Genomic data

We inferred recombination rates and DFE on previously published genomes of 15 wolves (Robinson *et al.* 2019). These genomes were sequenced at high coverage, with individuals' read depth averaging to >35X. Sites were filtered and aligned to the canFam3.1 dog reference genome as described in Marsden *et al.* 2016. Single nucleotide polymorphisms (SNPs) were called using GATK (McKenna *et al.* 2010) with only biallelic SNPs being retained. Indels were excluded for these analyses. More detailed information on genotype and variant calling of this data can be found in Mooney *et al.* 2023.

Inferring recombination rates

We used the recombination rate map previously inferred for this data using linkage disequilibrium data (LD) as previously described in chapter 2 (see Materials and Methods: Recombination rate inference and demographic inference).

Dividing the genome by recombination rate

We divided our genome-wide data (all *r*) into 3 different bins of inferred recombination rates: 0 to 1.9 x 10^{-9} (low *r*), >1.9 x 10^{-9} to 4.3 x10^{-9} (med *r*), and >4.3 x 10^{-9} to 2 x 10^{-8} (high *r*) per bp per generation. We excluded regions with recombination rates inferred above 2 x 10^{-8} per bp per generation. These bins were selected based on dividing the

genome wide recombination map into 3 ranges with roughly equal portions of the whole genome.

We divided the genome into non-overlapping 1 MB windows. For each window we calculated the weighted mean of the recombination rates. Details of how weighted averages were calculated are given in chapter 2 (see Materials and Methods: Comparisons of the genetic maps). We binned the 1 MB windows of the genome by recombination rates. The data binned by subsets of recombination rates all contained roughly ⅓ the number of synonymous and nonsynonymous SNPs found across all exons (Table S3.1). The values of the synonymous sequence length (L_S) varied somewhat by recombination bin. The ratio of synonymous SNPs / L_S ranged from 3.76 x 10^{-3} (low *r*) to 6.55 x 10^{-3} (high *r*). For each bin, we then created synonymous and nonsynonymous SFSs.

Computing Site Frequency Spectra

One individual was removed due to first-degree relatedness (i.e., parent-child and siblings). To compensate for missing data, we projected our data down to a sample size of 13. Using ∂a∂i (Gutenkunst *et al.* 2009), we generated folded SFSs for the observed synonymous sites and separately for the observed nonsynonymous sites. Folded SFSs were used to avoid the effects of mis-specifying ancestral alleles (Hernandez *et al.* 2007).

Calculating synonymous and nonsynonymous sequence lengths

All exonic sites were classified as 0-, 2-, 3-, or 4-fold degenerate sites based on the canFam3 reference genome. The nonsynonymous sequence length (L_{NS}) was calculated by the formula:

$$L_{NS} = (\#\text{ of 0-fold sites}) + (\tfrac{2}{3} * \#\text{ of 2-fold sites}) + (\tfrac{1}{3} * \#\text{ of 3-fold sites})$$

Conversely the synonymous sequence length (L_S) was calculated as:

$$L_S = (\#\text{ of 4-fold sites}) + (\tfrac{2}{3} * \#\text{ of 3-fold sites}) + (\tfrac{1}{3} * \#\text{ of 2-fold sites})$$

Additionally, we assumed a mutation 10 X the exon rate in putatively methylated CpG sites (Kong *et al.* 2012, Bird 1980). This resulted in a calculation of the L_{NS}:L_S ratio of 2.21 for canids (Amorim *et al.* 2023). Given that ratio of 2.21 nonsynonymous mutations for every 1 synonymous mutation, we calculated the L_S as 1 / 3.31 of all callable sites.

Demographic Inference

We inferred demographic models using the synonymous SFS from each of the 3 recombination rate bins and the genome-wide synonymous SFS. We focused on the synonymous SFS because these sites are putatively neutral and thus largely impacted by demographic history rather than direct selection (Kim *et al.* 2017). We assumed a two-epoch demographic model with an instantaneous size change. We inferred the best-fit parameters of ν (ratio of ancestral to current population sizes) and τ (time since

population size change) for the synonymous SFS data using ∂a∂i. We assumed an exon mutation rate (μ_{exon}) of 5.39 x 10^{-9} per bp per generation (Koch *et al.* 2019). The best model was the one with the largest multinomial log-likelihood.

We calculated the ancestral population size (N_a) using the equation:

$$N_a = \Theta_{syn} /(4 * \mu_{exon} * L_s)$$

Θ of synonymous sites (Θ_{syn}) was calculated from the synonymous SFS.

DFE inference

We inferred DFEs for the nonsynonymous mutations in each of the three recombination rate bins as well as for all nonsynonymous mutations using the fit∂a∂i module (Kim *et al.* 2017). We conditioned our DFE models on the demographic histories inferred from the synonymous SFS. By conditioning our DFE inference on a model of demography, we can remove the majority of the impact of demographic history (Kim *et al.* 2017). For our DFE models, we assumed a gamma distribution and selected the best model as the one with the largest Poisson log likelihood.

For plotting, we discretized the gamma DFE into 4 bins of selection coefficients (|s|), assuming all mutations are deleterious. These bins included $s < 1 \times 10^{-4}$ (nearly neutral), 1×10^{-4} to 1×10^{-3} (weakly deleterious), 1×10^{-3} to 1×10^{-2} (moderately

deleterious), and > 1 x 10^{-2} (strongly deleterious). We report values of s by dividing 2Ns by the ancestral population size inferred from the synonymous SFS.

Results and Discussion

SFSs

We characterized the synonymous and nonsynonymous proportional SFS and a normalized count SFS for all four datasets (Fig. 3.1). For the proportional SFS we divided the number of SNPs with a given frequency by the total number of SNPs for that dataset. For the normalized count SFS, we divided the number SNPs with a given count in the sample by the total number of called sites (including invariant sites) for that dataset. This normalization provides for a direct comparison between datasets controlling for the fact that each dataset contains a different proportion of the genome.

For the proportional SFS, we find similar synonymous and nonsynonymous SFSs for all datasets (Fig. 3.1 Left). However, we note that regions of low recombination have a subtle skew towards singletons for synonymous sites relative to med and high *r* regions (Proportion of SFS - Low *r*: 0.201, Med *r*: 0.194, High *r*: 0.193). Background selection is known to cause an increase of rare variants of synonymous sites in regions of low recombination (Lohmueller *et al.* 2011, Good *et al.* 2014).

By contrast, for our normalized count SFSs, we observe notable variation in proportions of each frequency class across the recombination rate bins for both synonymous and nonsynonymous sites (Fig. 3.1 Right). We find a positive correlation of genetic diversity

and recombination rate, with low *r* regions having the least diversity and high *r* regions having the most diversity. This result is consistent with the observation that background selection decreases the diversity of linked neutral sites particularly in regions of low recombination (Charlesworth *et al.* 1995, Charlesworth 2012). The genome wide SFSs most closely resemble the med *r* regions suggesting that genome wide estimates of the SFS reflect a more moderate ratio of SNPs to all sites than regions with extreme (low or high) recombination.

The results of our observed synonymous and nonsynonymous SFSs provide validation for our approach of dividing the genome into three subsets of recombination rate bins. Our study focuses on the influence of recombination rate on inferring DFEs with a hypothesis that regions of low recombination may bias DFE inferences due to background selection. The observed SFS results are consistent with the notion that our three bins of recombination rates may experience differing levels of background selection. Thus, these genomic regions with different recombination rates can inform us as to whether background selection impacts DFE inference.

Demographic inference

We inferred demographic models using the synonymous SFS for our four datasets (low, med, high, and all *r*). For all the datasets we inferred two-epoch demographic models with similar maximum likelihood estimates of the parameter values (Table 3.1). The estimates of the ν and τ parameters differed slightly across the recombination rate bins. And while values of Θ were similar between the datasets binned by recombination rate,

their values of N_a ranged from ~45K to ~82K. For all the demographic models, we found that the observed SFS and the expected SFS of the model matched well, suggesting a satisfactory fit of the model (Fig. 3.2). The inferred different N_a for each recombination rate bin is consistent with findings that effective population size (N_e) is positively correlated with recombination rate (Gossman *et al.* 2011).

The differences in parameters for the low, med, and high *r* datasets suggests that either 1) the effects of linked selection are affecting the demographic inferences differently or 2) multiple combinations of demographic history parameters fit each dataset similarly well. To test the second possibility, we used ∂a∂i to evaluate how the demographic parameters of ν and τ inferred from one dataset fit another dataset. For example, we evaluated how a demographic model with the values of ν and τ inferred from the med *r* observed synonymous SFS fit the data of the low *r* observed synonymous SFS. We then also considered a Θ parameter to examine the fit of the model on the non-scaled synonymous SFS. For the example of the med *r* model fit to the low *r* data, we calculated Θ based on the N_a of the med *r* model and the L_S of the low *r* dataset. We used these parameters to calculate Θ because N_a is a property of the model while L_S is a property of the data. We did this for the 3 bins of data with a subset of recombination rates (low, med, and high *r*). We tested all 6 pairwise combinations of observed synonymous SFS and demographic parameters inferred from a different dataset. We evaluated the model fit by comparing the multinomial and Poisson log likelihoods (Tables 3.2 and 3.3).

The log likelihood tables represent fitting the demographic model from one data set (labeled data and listed as column headings) to another dataset (labeled data and listed as row names) (Tables 3.2 and 3.3). The log likelihood of the example described above of fitting the med *r* demographic model to the low *r* data is found in the cell of row 1 and column 2. The diagonal cells highlighted in bold represent the log likelihood of the models inferred from a given data set and how they fit that data.

Fitting different demographic models to our low, med, and high *r* datasets resulted in a small range of multinomial log likelihoods (-61.7 to -57.3) (Table 3.2). The model that fit the low *r* data best was the low *r* model. And the data that fit the high *r* data best was the high *r* model. By contrast, the med *r* data was fit equally well by the med and high *r* models. The fit of these model and data combinations can be visualized by comparing the synonymous proportional SFS of the observed data and the model's expected SFS (Fig. 3.3 Left). Given that the synonymous proportional SFS of the observed low, med, and high data were very similar, this result is not surprising. Because all the models were fit to similar data (with the exception of the very subtle rare variance skew in the low *r* data set), the models all produce similar proportional SFSs.

By contrast, there is wide variation of the Poisson log likelihood of these model and data combinations (-57.3 to -2183.4) (Table 3.4). When measuring model fit by Poisson log likelihood, each dataset is best fit by its own model. This is shown in Table 3.4 where the bolded diagonal of matched models and datasets has Poisson log likelihoods larger (and thus better) than any of the mismatched models and datasets. This pattern is also

reflected in the wide range of expected count SFSs from the different models (Fig. 3.3 Right).

This difference in Poisson log likelihoods is likely driven by the differences in the Θ values for scaling the SFS. The scaled SFS is used to evaluate the Poisson log likelihood. For the example of fitting the med *r* model to the low *r* data, we calculated a Θ of ~4.4 K calculated with the low *r* L_S of ~ 3 M and the med *r* N_a of ~ 65 K. By contrast, the Θ of the low *r* data we inferred was ~3.1 K based on the low *r* L_S of ~ 3 M and the low *r* N_a of ~ 45 K. These differences are also consistent with the observation that the normalized count SFSs of the low, med, and high *r* datasets vary noticeably (Fig. 3.1 Right).

The substantial variations in Poisson log likelihoods between our six model and data combinations indicate significant differences in the demographic models we derived. In turn this reflects that there are large and meaningful differences of genetic diversity between our datasets with low, med, and high *r* rates. Although these regions of the genome have a shared true demographic history, methods to estimate N_a for these subsets of the genome are consistent with how N_e varies with recombination rate because of linked selection (Gossman *et al.* 2011). Given the differences of genetic diversity between these bins of recombination rates, we might expect to see variation in the inferred DFEs of these bins.

DFE Inference

We inferred DFEs for nonsynonymous mutations from our 4 datasets of varying recombination rate ranges. The DFE models are conditioned on the best fitting demographic model for that dataset's synonymous SFS. We found similar DFEs for all datasets except the med *r* data (Fig. 3.4). Parameters inferred and statistics measured for these DFEs are shown in Table 3.4. Both the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters of the inferred DFE of the med *r* data vary noticeable from the other data. Notably, the DFE for med *r* data was skewed towards strongly deleterious new mutations compared to the low, high, and all *r* data (Proportion of new mutations with |s| > 0.01 low *r*: 0.48, med *r*: 0.67, high *r*: 0.47, and all *r*: 0.54).

To examine if there was a statistically significant difference in the med *r* DFE versus other DFEs inferred from the other recombination rate bins, we evaluated the fit of all our DFE models via the models' Poisson log likelihoods and the expected SFS of each model. The best DFE model for each dataset was selected by the largest Poisson log likelihood of models inferred using fit∂a∂i. The Poisson log likelihood of the DFE models for the low, med, and high *r* were within a small range (-59.0 to -56.8). The all *r* Poisson log likelihood was somewhat smaller than the others (-71.5). Adequate fit of all our models was supported by a close match between the observed nonsynonymous SFSs and the respective models' expected SFS (Fig. 3.5). Given that all our datasets have mostly similar log likelihoods and good matches between observed and expected SFSs, there is strong support that each DFE model is well fit to the data it was inferred from. This means there may be other DFE models that fit our datasets better or similarly well

to those we inferred. In other words, while our med *r* DFE fits that data well, there may be other models that fit it just as well or better. And if there is no difference in the true DFE of all our datasets, perhaps the models we inferred from the low and/or high *r* datasets meet that criterion.

To test if any of our inferred DFEs for the low, med, and high *r* data fit all datasets well, we considered how the DFE parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) of one subset of recombination rate data fit a different subset of recombination rate data. For example, to test the fit of the med *r* data's inferred DFE on the low *r* data, we conditioned the DFE on the $\boldsymbol{\nu}$, τ, and Θ values inferred from the low *r* data. We created the DFE model using the $\boldsymbol{\alpha}$ and unscaled $\boldsymbol{\beta}$ values of the med *r* data. To scale the $\boldsymbol{\beta}$ value, we used the N_a inferred from the low *r* data. We calculated the scaled $\boldsymbol{\beta}$ value as $\boldsymbol{\beta}$ * $2N_a$. To quantify the DFE model fit, we calculated the multinomial and Poisson log likelihoods of the model's expected nonsynonymous SFS compared to the observed nonsynonymous SFS. We repeated this for all 6 directional pairwise comparisons of low, med, and high *r* datasets.

The log likelihood tables for the DFE (Table 3.5 and 3.6) are structured the same as those described for the demographic models (Table 3.2 and 3.3). The example listed above of fitting the med *r* DFE to the low *r* data can be found in cell row 1 x column 2. The bolded diagonal represents the DFE model's log likelihood when fit to the data it was inferred from.

We found that trying to fit the DFE parameters from one recombination rate data set to another resulted in a range of multinomial log likelihoods (-86.1 to -55.7) (Table 3.5). The DFE model with the largest, and thus best, multinomial log likelihood for each dataset was the DFE inferred from that data. And notably, the DFE parameters inferred from med *r* data had the smallest multinomial log likelihood for the low and high data. This suggests that the med *r* DFE skewed towards strongly deleterious new mutations was a worse fit for the low and high *r* data than the other two models (low *r* and high *r*). By contrast, the med *r* parameters had the largest multinomial log likelihood for all parameter sets applied to the med *r* data supporting that the med *r* model fits that data the best. The DFE model fits as measured by the multinomial log likelihoods is visualized by comparing the proportional SFS of the observed nonsynonymous sites and the models' expected SFS (Fig. 3.6 left). The models with larger log likelihoods have expected SFSs more similar to the observed data.

We also examined the Poisson log likelihood of fitting DFE models to different recombination rate bins. We found a wider range of Poisson log likelihoods than the multinomial (-58.8 to -113.9) (Table 3.6). Like the comparisons of the multinomial log likelihoods, we found that the DFE model with the largest, and thus best, Poisson log likelihood for each dataset was the DFE inferred from that data. This is seen in the fact that the bolded diagonal values (*e.g.* low *r* data and low *r* model) have much larger values than mismatches (*e.g.* low *r* data and med *r* model). These differences are much more apparent in the Poisson log likelihoods than the multinomial log likelihoods. This is because the Poisson log likelihood has greater potential to reject worse models due to

the increased data of the number of mutations. The model fits quantified by Poisson log likelihoods are visualized in comparisons of the count SFS of nonsynonymous sites for the observed data and the expected SFS of the model (Fig. 3.6 Right). The worse fit of the expected count SFS vs the expected proportional SFS can be seen in the med *r* model applied to the low *r* data (Fig. 3.6 Top Left vs Right).

This difference of multinomial and Poisson log likelihoods supports that the DFE models have statistically significant differences. While these DFE models may be statistically distinct, it is important to consider whether they are different from one another in biologically meaningful ways.

One potential reason we see this skew towards highly deleterious mutations in the med *r* data could be driven by the medium recombination rate regions having genes with different functions than other bins of the genome. If the med *r* regions of the wolf genome are enriched for genes with essential function, for example homeobox genes, these regions would also be enriched towards strong negative selection (larger estimates of *s*).

Yet, it is important to consider how different these DFE inferences are in the context of variation of the DFE across taxa. Kyriazis *et al.* (2023) reviewed published DFE inference of several species and shows that less than 20% of new mutations in mice, *Drosophila*, yeast, and *Arabidopsis* are strongly deleterious. That contrasts with all our inferred DFE models in wolves with more than 45% of new mutations being strongly

deleterious. So while we infer that wolves have a higher proportion of strongly deleterious mutations in regions with moderate recombination rates, this result is still more similar to our findings in the low, med, and all *r* genomic regions in wolves than to many other species. Additionally, we should not over interpret the biological significance of the med *r* DFE skew towards strongly deleterious variations as this class of mutations is more difficult to measure in small sample sizes (Kim *et al.* 2017). In contrast, estimating the density of nearly neutral sites among new mutations is less likely to be impacted by sample size. And in the case of nearly neutral sites, we inferred highly similar proportions for the four datasets varied by recombination rate (Low $r = 0.34$, Med $r = 0.29$, High $r = 0.32$, and All $r = 0.32$).

Importantly, a major motivation for this work was the question of how increased background selection in regions of low recombination (Charlesworth *et al.* 1995, Charlesworth 2012) might bias DFE inferences. In our data we observed the impact of background selection notably in the positive correlation of recombination rate and genetic diversity as seen in the normalized count SFSs (Fig. 3.1). Additionally, we see a subtle skew towards rare variants in the low *r* data which can be caused by background selection (Lohmueller *et al.* 2011, Good *et al.* 2014) (Fig. 3.1). Despite the prevalence of background selection in our low *r* data set and lack of background selection in the high *r* data (Fig. 3.1); the estimated DFE of the low, high, and all *r* data are substantially similar to each other (Fig. 3.3, Table 3.4). If background selection was strongly biasing inferences of the DFE, we would have expected to find larger variation in the low *r* DFE

relative to all other data sets. Thus we find strong support that background selection does not bias inferences of the DFE using PRM methods.

Future Directions

This work focused on inferring the DFE of wolves because canids lack a functional copy of PRDM9, a gene that influences vertebrates' recombination location and intensity (Muñoz-Fuentes *et al.* 2011; Cavassim *et al.* 2022, Baker *et al.* 2015; Grey, Baudat, & de Massy, 2018; Oliver et al., 2009; Parvanov *et al.* 2010). Prior work (see chapter 2 Fig 2.4A) has shown significant correlation in the recombination landscape of wolves and breed dogs. However, inferred global recombination rates vary in the domestic dog breeds of border collies and pugs relative to their sister taxa of wolves. Additionally, the estimates of the genome wide DFE of these two breeds is similar to estimates in wolves (Amorim *et al.* 2023). Given this combination of differences in the genome-wide recombination rates but similar DFE estimates, these two dog breeds could be an interesting comparison to the results of this work. Furthermore, the question of how recombination does or does not influence estimates of the DFE with PRF methods should be considered in other species with high quality genetic maps. Lastly, inferring the DFE for simulations of genetic data with different recombination rates and background selection could further our understanding of how these two mechanisms affect estimates of the DFE.

Conclusion

The DFE is a foundational concept in population genetics, thus methods to accurately estimate the DFE are of critical importance to the field. Inferring the DFE with PRF based methods depends on the assumption of free recombination (Hartl *et al.* 1994, Sawyer and Hartl 1992). A particular concern is the increased influence of background selection in regions of low recombination (Charlesworth *et al.* 1995). In this work we support that violation of this assumption does not bias DFE inferences. We inferred the DFE of the North American gray wolf for the whole genome as well as subsets of the genome with different ranges of recombination rates. We observed a decrease of diversity in neutral sites in regions of low recombination consistent with background selection. However, our inferred DFE for genomic regions with low recombination is notably similar to regions with high recombination and the genome wide DFE. Thus, we found strong evidence that PRF based inferences of the DFE are not biased by violation of assuming sites are unlinked via free recombination. Our findings provide valuable support for the accuracy of contemporary methods of inferring the DFE (e.g. Kim *et al.* 2017, Boyko *et al.* 2008, Keightley and Eyre-Walker 2007).

Figures

Figure 3.1

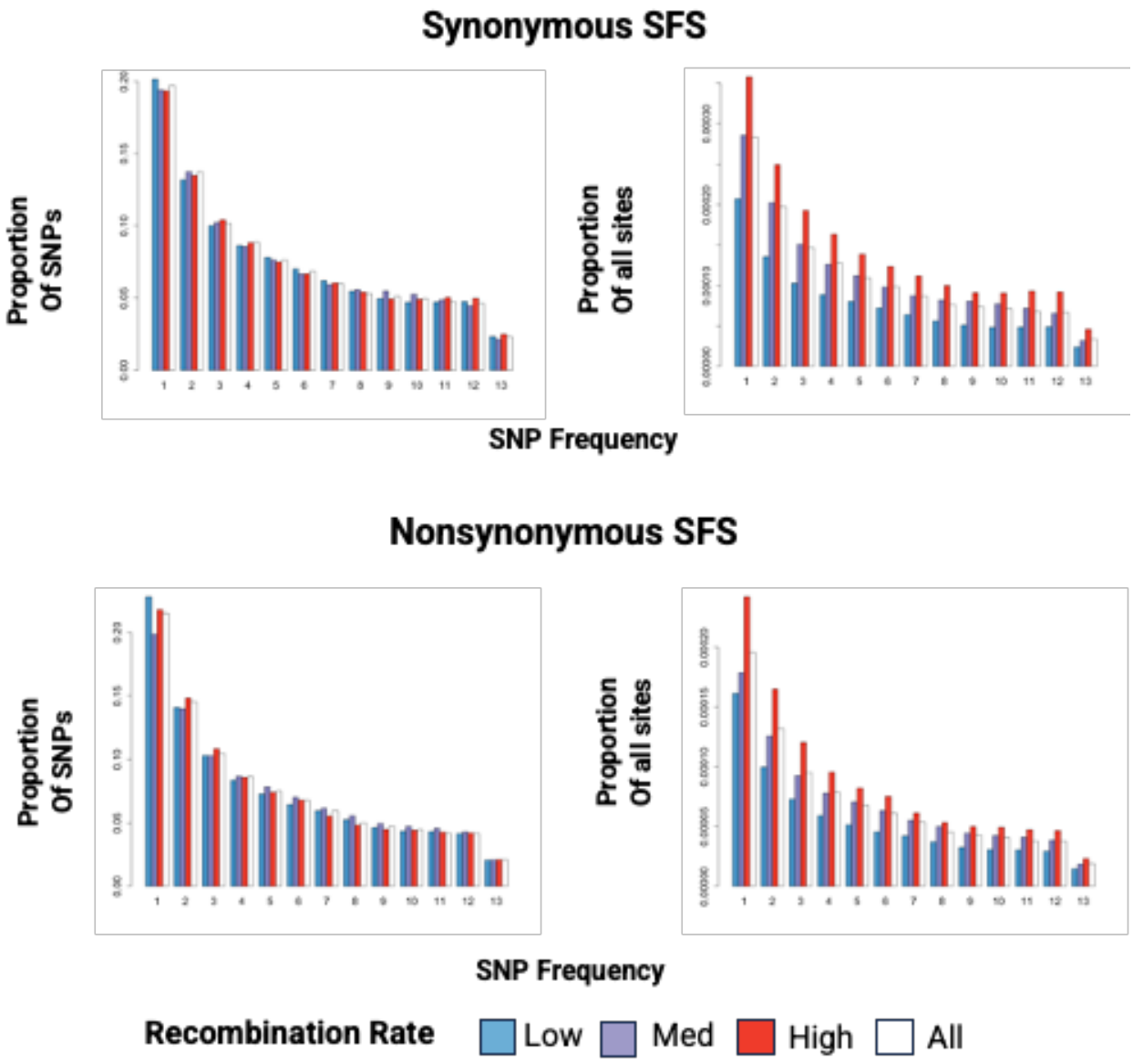

Fig 3.1. Observed Synonymous and Nonsynonymous SFSs. Bar plots showing the proportion of exonic SNPs with a given frequency by recombination rates. SNPs are separated by synonymous sites (top) and nonsynonymous sites (bottom). Left panels show SNPs as a proportion of all SNPs of that type for a given recombination rate

range. Right panels show SNPs as a proportion of all callable sites in a specific range of recombination rates. 1MB windows of the genome were classified into three subsets of r values: 0 to 1.9 x 10^{-9} (low), 1.9 x 10^{-9} to 4.3 x10^{-9} (med), and 4.3 x 10^{-9} to 2 x 10^{-8} (high). SFSs for those three bins and all recombination rates are color coded: low (blue), med (purple), high (red), and all (white).

Figure 3.2

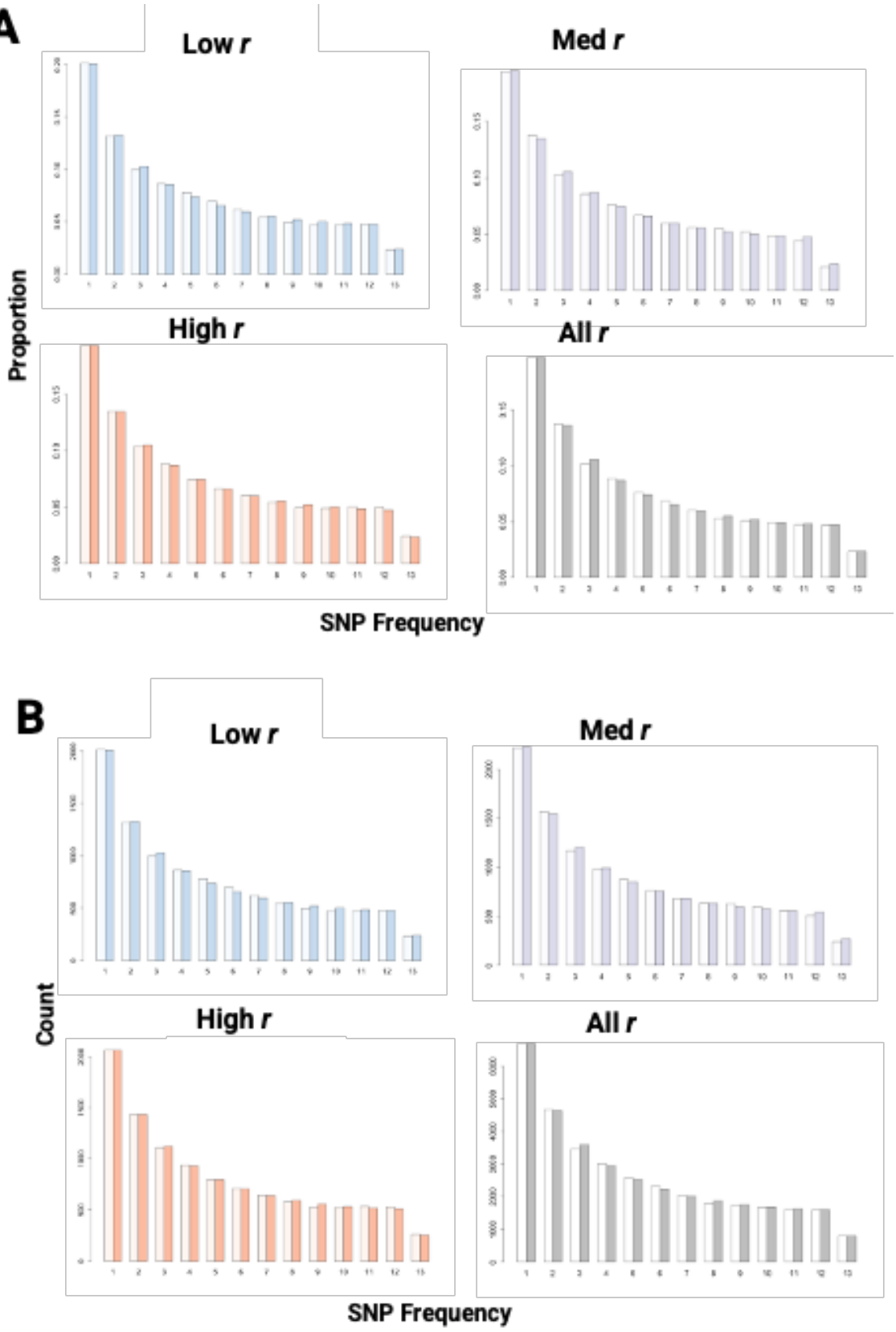
A
Low r
Med r
High r
All r
Proportion
SNP Frequency
B
Low r
Med r
High r
All r
Count
SNP Frequency

Fig 3.2. Observed and Expected SFSs for Demographic Models. Bar plots showing the observed (data, white) and expected (model, color) SFSs of synonymous sites by recombination rate. The count (panel A) and proportion (panel B) SFSs are shown. 1MB windows of the genome were classified into three subsets of *r* values: 0 to 1.9×10^{-9} (low), 1.9×10^{-9} to 4.3×10^{-9} (med), and 4.3×10^{-9} to 2×10^{-8} (high). SFSs for those three bins and all recombination rates are color coded: low (blue), med (purple), high (red), and all (white). Expected SFSs are derived from a two-epoch demographic model fit to the observed data.

Figure 3.3

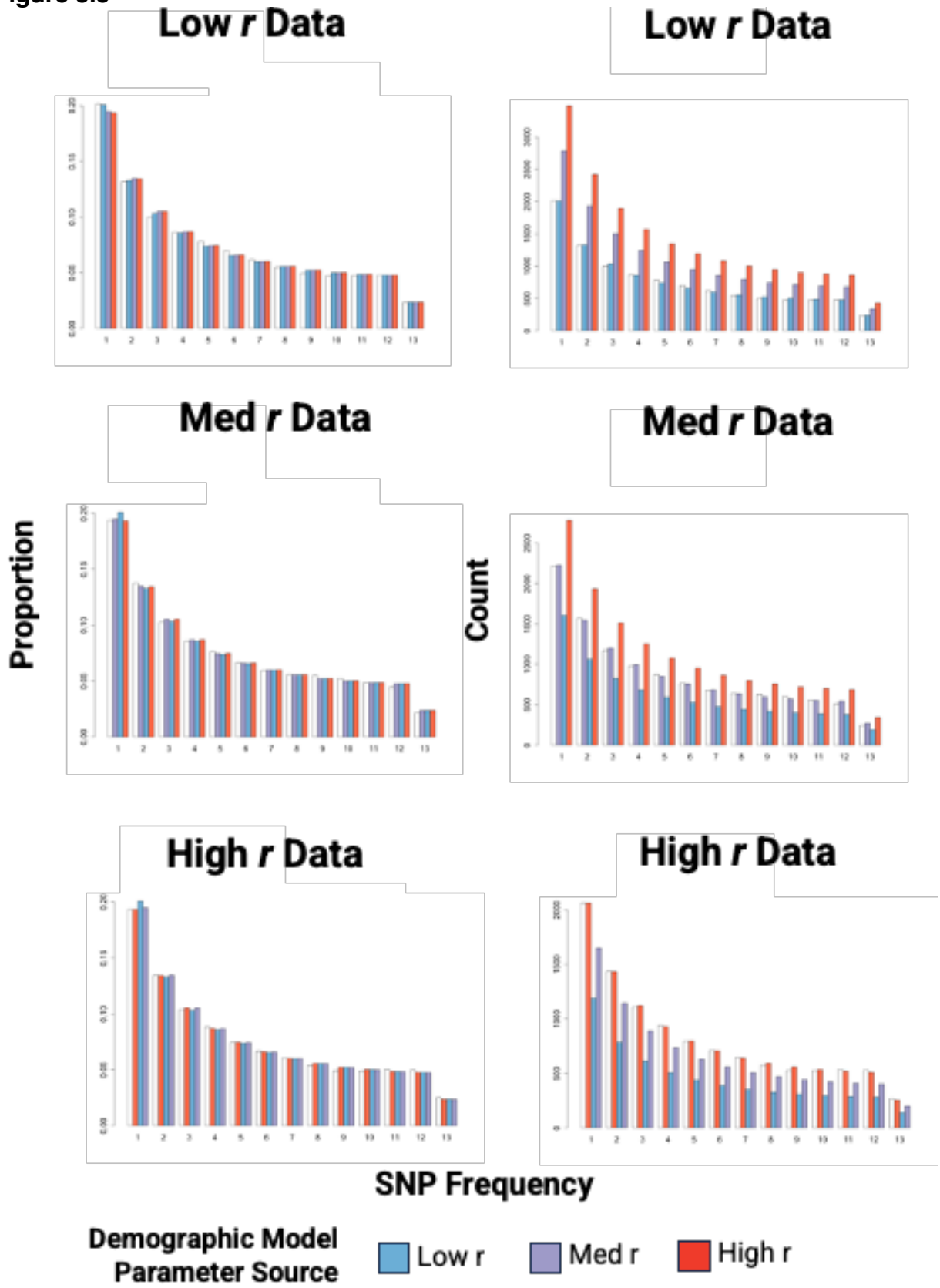
Low r Data
Low r Data
Med r Data
Med r Data
High r Data
High r Data
Proportion
Count
SNP Frequency
Demographic Model Parameter Source
Low r
Med r
High r

Fig 3.3. Observed and Expected SFSs for Various Demographic Models. Bar plots showing the observed (data, white) and expected (model, color) SFSs of synonymous sites for various demographic models. The proportion (left) and count (right) SFSs are shown. Genes were classified into three subsets of *r* values: 0 to 1.9 x 10^{-9} (low), >1.9 x 10^{-9} to 4.3 x10^{-9} (med), and >4.3 x 10^{-9} to 2 x 10^{-8} (high). Rows are separated by recombination rate of the data. For each set of data, we tested 3 demographic model parameters of nu and tau. The parameters tested came from two-epoch demographic models derived from 1 of the 3 data sets binned by recombination rates (see Table 3.1). Expected SFSs for those 3 parameter sets are color coded: low *r* (blue), med *r* (purple), high *r* (red).

Figure 3.4

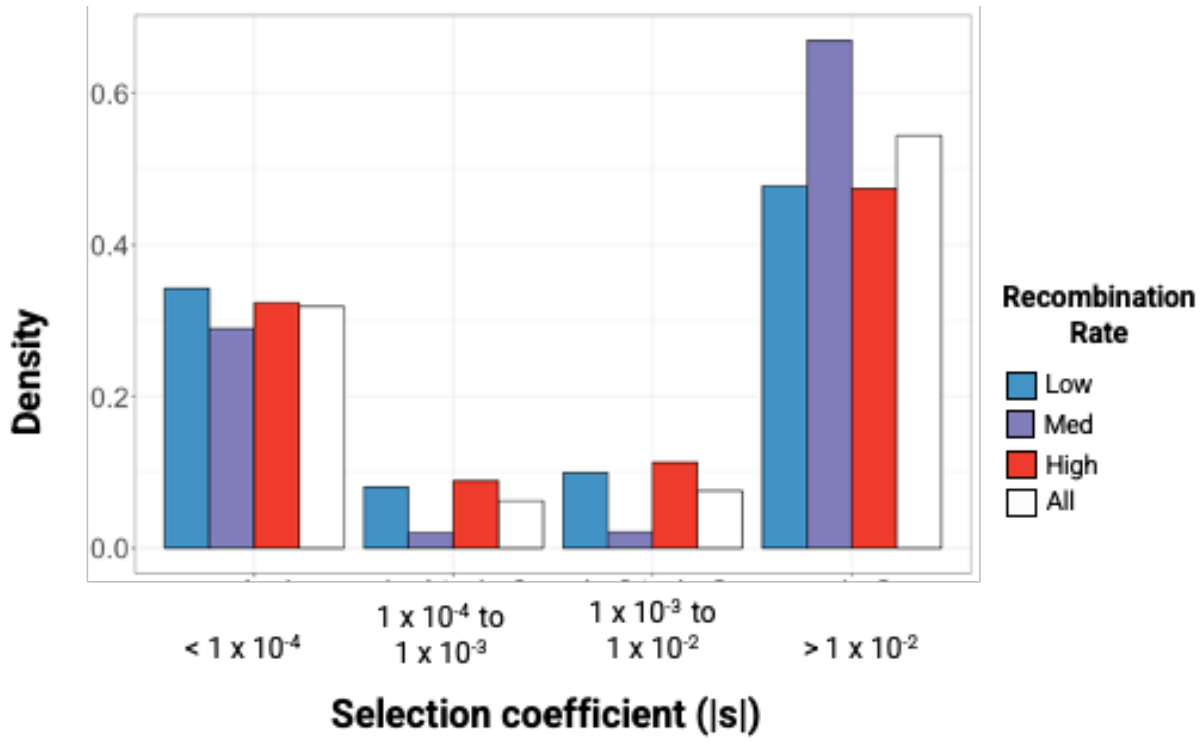

Fig 3.4. Inferred DFE for nonsynonymous mutations in regions of the genome with different recombination rates. Bar plot showing proportion of nonsynonymous mutations inferred to have a specific range of selection coefficient (s) compared by recombination rates. 1MB windows of the genome were classified into three subsets of *r* values: 0 to 1.9 x 10^{-9} (low), >1.9 x 10^{-9} to 4.3 x10^{-9} (med), and >4.3 x 10^{-9} to 2 x 10^{-8} (high). DFEs for those three bins and all recombination rates are color coded: low (blue), med (purple), high (red), and all (white). Selection coefficients ranges include < 1 x 10^{-4} (nearly neutral), 1 x 10^{-4} to 1 x 10^{-3} (weekly deleterious), 1 x 10^{-3} to 1 x 10^{-2} (moderately deleterious), and > 1 x 10^{-2} (strongly deleterious).

Figure 3.5

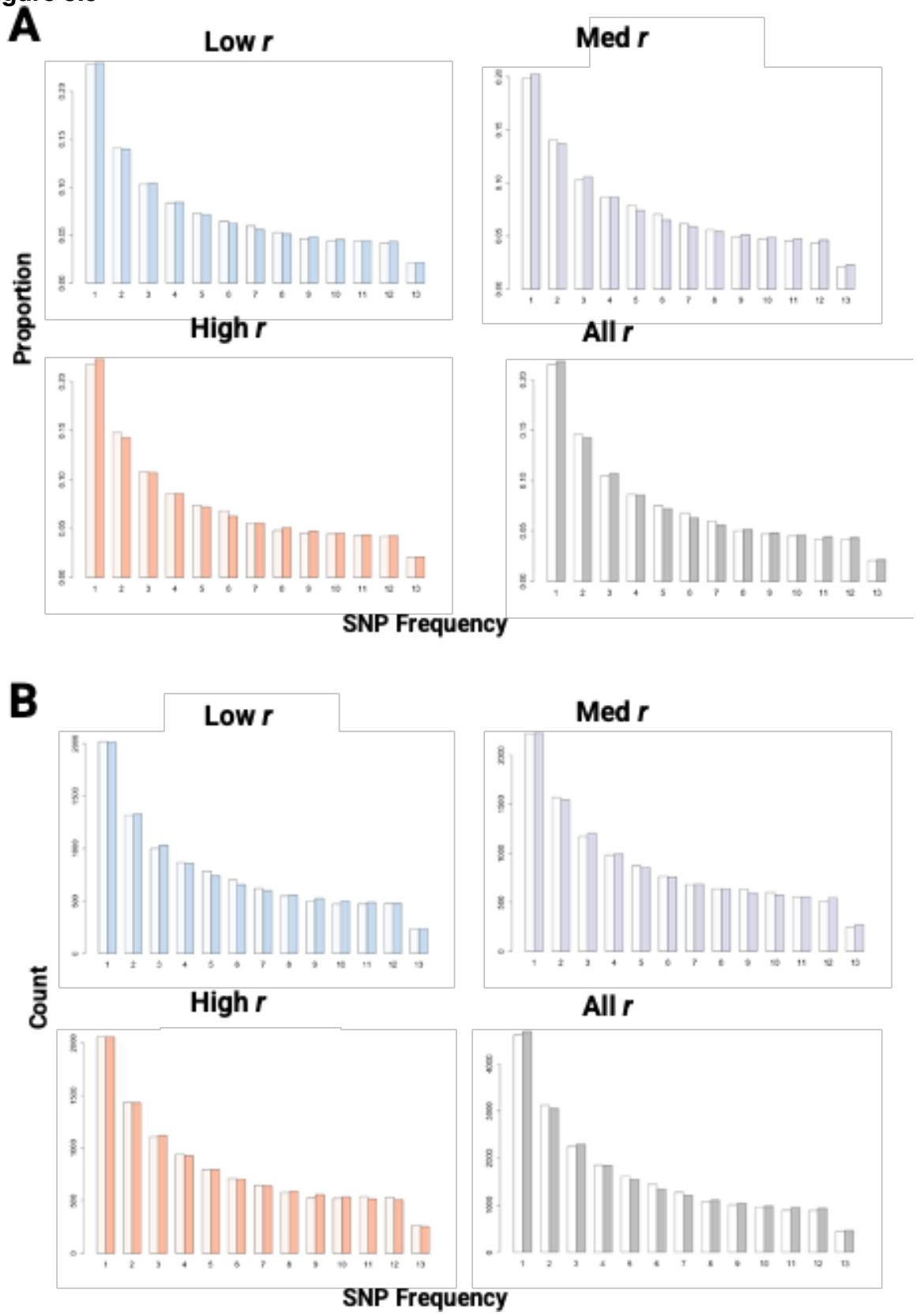
A
Low r
Med r
High r
All r
Proportion
SNP Frequency
B
Low r
Med r
High r
All r
Count
SNP Frequency

Fig 3.5. Observed and Expected SFSs for DFE Models. Bar plots showing the observed (data, white) and expected (model, color) SFSs of nonsynonymous sites by recombination rate. The count (panel A) and proportion (panel B) SFSs are shown. 1MB windows of the genome were classified into three subsets of *r* values: 0 to 1.9 x 10^{-9} (low), 1.9 x 10^{-9} to 4.3 x10^{-9} (med), and 4.3 x 10^{-9} to 2 x 10^{-8} (high). SFSs for those three bins and all recombination rates are color coded: low (blue), med (purple), high (red), and all (white). Expected SFSs are derived from models of DFEs conditioned on a two-epoch demographic model and the observed data.

Figure 3.6

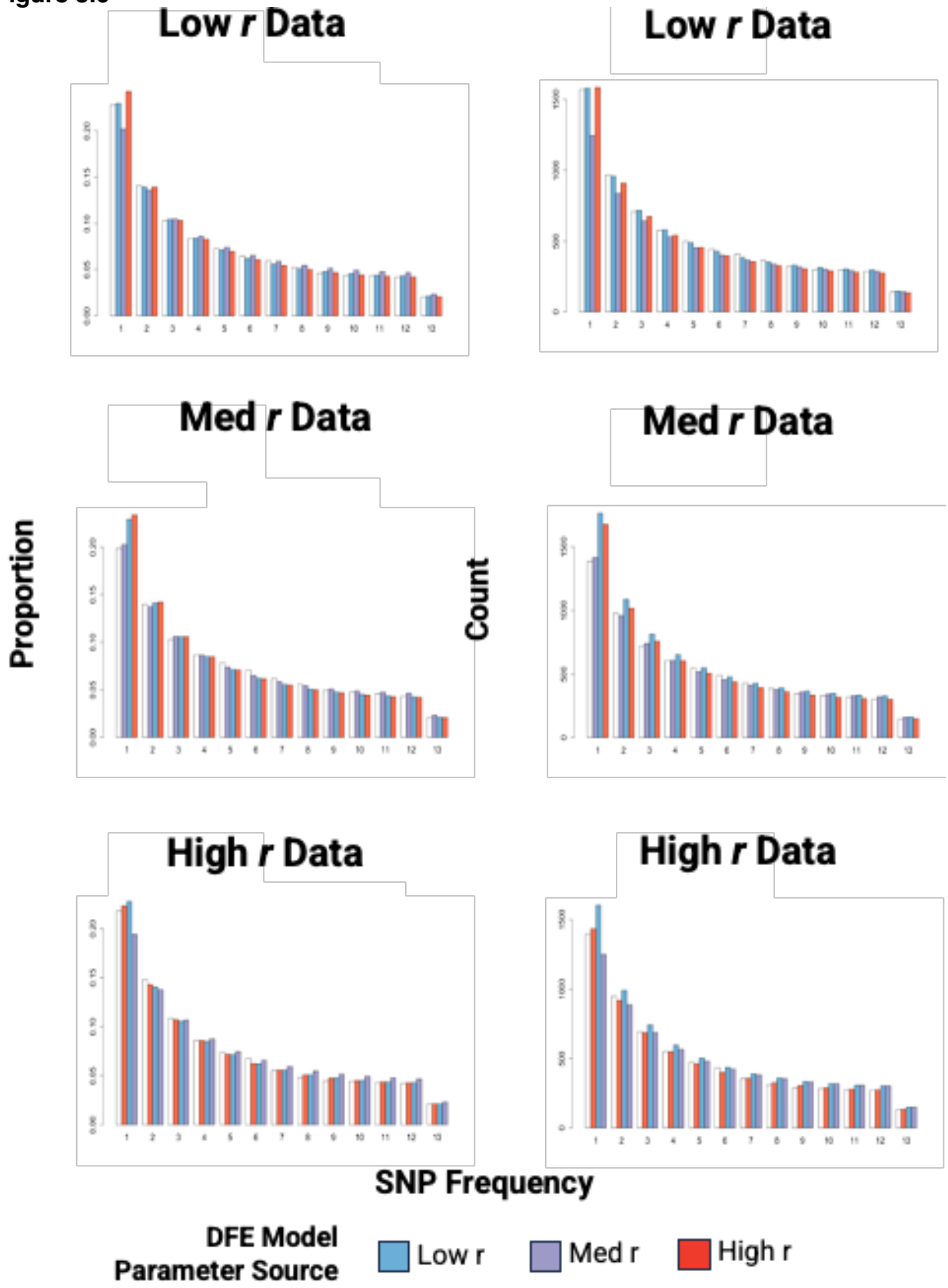
Low r Data
Low r Data
Med r Data
Med r Data
High r Data
High r Data
Proportion
Count
SNP Frequency
DFE Model Parameter Source
Low r
Med r
High r

Fig 3.6. Observed and Expected SFSs Conditioned on Various DFE Models. Bar plots showing the observed (data, white) and expected (model, color) SFSs of synonymous sites for various DFE models. The proportion (left) and count (right) SFSs are shown. Genes were classified into three subsets of *r* values: 0 to 1.9 x 10^{-9} (low), >1.9 x 10^{-9} to 4.3 x10^{-9} (med), and >4.3 x 10^{-9} to 2 x 10^{-8} (high). Rows are separated by recombination rate of the data. For each set of data, we tested 3 DFE model parameters of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The parameters tested came from DFE models derived from 1 of the 3 data sets divided by recombination rates. Expected SFSs for those 3 parameter sets are color coded: low *r* (blue), med *r* (purple), high *r* (red).

Tables

Table 3.1

Parameters inferred and statistics measured from demographic models for synonymous SFS data.

r Values	ν	τ	Na	Θ	Poisson Log Likelihood
Low	0.32313	0.03313	45321	3089	-59.6
Med	0.20271	0.01616	64652	3522	-60.0
High	0.19515	0.01581	81654	3296	-57.3
All	0.18375	0.01319	62697	10373	-69.8

Table 3.2

Multinomial log likelihood of demographic parameters inferred from synonymous SFSs (model, column headers) fit to the synonymous SFS data (data, row headers) for regions of the genome in 3 different recombination rate ranges.

Model / Data	**Low *r***	**Med *r***	**High *r***
Low *r*	**-59.6**	-61.2	-61.7
Med *r*	-61.6	**-60.0**	-60.0
High *r*	-59.3	-57.4	**-57.3**

Table 3.3

Poisson log likelihood of demographic parameters inferred from synonymous SFSs (model, column headers) fit to the synonymous SFS data (data, row headers) for regions of the genome in 3 different recombination rate ranges.

Model / Data	**Low *r***	**Med *r***	**High *r***
Low *r*	**-59.6**	-785.7	-2183.4
Med *r*	-714.3	**-60.0**	-386.4
High *r*	-1587.1	-317.9	**-57.3**

Table 3.4

Parameters inferred and statistics measured from DFE models.

r Values	α	Scaled *β*	*β*	Poisson Log Likelihood
Low	0.092	1.70E+06	1.88E+01	-56.8
Med	0.029	9.24E+19	7.14E+14	-57.8
High	0.106	1.20E+06	7.33E+00	-59.0
All	0.078	4.90E+07	3.90E+02	-71.5

Table 3.5

Multinomial log likelihood of DFE parameters inferred from SFSs (model, column headers) fit to the SFS data (data, row headers) for regions of the genome in 3 different recombination rate ranges.

Model / Data	**Low *r***	**Med *r***	**High *r***
Low *r*	**-55.7**	-72.7	-58.0
Med *r*	-78.5	**-57.5**	-86.1
High *r*	-57.1	-75.6	**-58.4**

Table 3.6

Poisson log likelihood of DFE parameters inferred from SFSs (model, column headers) fit to the SFS data (data, row headers) for regions of the genome in 3 different recombination rate ranges.

Data \ Model	**Low *r***	**Med *r***	**High *r***
Low *r*	**-56.8**	-113.9	-67.1
Med *r*	-113.5	**-57.8**	-88.5
High *r*	-86.3	-75.7	**-59.0**

References

Amorim, C. E. G., Marsden, C. D., Mah, J. C., Lin, M., Del Carpio, C. A., Guardado, M., Robinson, J., Kim, B. Y., Lohmueller, K. E. (2023). The Impact of Domestication on the Selective Effects of Mutations in the Canid Genome [Unpublished Manuscript]. Department of Biology, California State University Northridge.

Baker, C. L., Kajita, S., Walker, M., Saxl, R. L., Raghupathy, N., Choi, K., ... & Paigen, K. (2015). PRDM9 drives evolutionary erosion of hotspots in Mus musculus through haplotype-specific initiation of meiotic recombination. *PLoS genetics*, *11*(1), e1004916.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, *8*(7), 1499-1504.

Boyko A. R., Williamson S. H., Indap A. R., Degenhardt J. D., Hernandez R. D., *et al.* (2008) Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. PLOS Genetics 4(5): e1000083. https://doi.org/10.1371/journal.pgen.1000083

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., & Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–1157. https://doi.org/10.1038/nature04240

Cavassim, M. I. A., Baker, Z., Hoge, C., Schierup, M. H., Schumer, M., & Przeworski, M. (2022). PRDM9 losses in vertebrates are coupled to those of paralogs ZCWPW1

and ZCWPW2. *Proceedings of the National Academy of Sciences*, *119*(9), e2114401119.

Charlesworth, D., Charlesworth, B., & Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics*, *141*(4), 1619–1632. https://doi.org/10.1093/genetics/141.4.1619

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195-205.

Charlesworth B. (2012). The effects of deleterious mutations on evolution at linked sites. *Genetics*, *190*(1), 5–22. https://doi.org/10.1534/genetics.111.134288

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature reviews. Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Good, B. H., Walczak, A. M., Neher, R. A., & Desai, M. M. (2014). Genetic diversity in the interference selection limit. *PLoS genetics*, *10*(3), e1004222.

Gossmann, T. I., Woolfit, M., & Eyre-Walker, A. (2011). Quantifying the variation in the effective population size within a genome. *Genetics*, *189*(4), 1389-1402.

Grey, C., Baudat, F., & de Massy, B. (2018). PRDM9, a driver of the genetic map. *PLoS Genetics*, *14*(8), e1007479. https://doi.org/10.1371/journal.pgen.1007479

Gutenkunst R. N., Hernandez R. D., Williamson S. H., Bustamante C. D. (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics* 5(10): e1000695. https://doi.org/10.1371/journal.pgen.1000695

Hartl D. L., Moriyama E. N., Sawyer S. A., 1994. Selection intensity for codon bias. *Genetics* 138: 227–234.

Hernandez R. D., Williamson S. H., Bustamante C. D., 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* 24: 1792–1800.

Jiménez-Mena, B., Hospital, F., & Bataillon, T. (2016). Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. *Conservation genetics resources*, *8*, 35-41.

Johri, P., Riall, K., Becher, H., Excoffier, L., Charlesworth, B., & Jensen, J. D. (2021). The impact of purifying and background selection on the inference of population history: problems and prospects. *Molecular biology and evolution*, *38*(7), 2986-3003.

Keightley P. D., Eyre-Walker A., 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177: 2251–2261.

Kim, B. Y., Huber, C. D., & Lohmueller, K. E. (2017). Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics*, *206*(1), 345–361. https://doi.org/10.1534/genetics.116.197145

Koch, Evan, Rena M. Schweizer, Teia M. Schweizer, Daniel R. Stahler, Douglas W. Smith, Robert K. Wayne, and John Novembre. 2019. “De Novo Mutation Rate Estimation in Wolves of Known Pedigree.” *Molecular Biology and Evolution*, July. https://doi.org/10.1093/molbev/msz159.

Kong A., Frigge M. L. , Masson G., Besenbacher S., Sulem P., Magnusson G., Gudjonsson S. A., Sigurdsson A., Jonasdottir A., Jonasdottir A., Wong W. S., Sigurdsson G., Walters G. B., Steinberg S., Helgason H., Thorleifsson G., Gudbjartsson D. F., Helgason A., Magnusson O. T., Thorsteinsdottir U., Stefansson K. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*. 2012 Aug 23;488(7412):471–5. doi: 10.1038/nature11396

Li Y., Vinckenbosch N., Tian G., Huerta-Sanchez E., Jiang T., et al., 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* 42: 969–972.

Lohmueller K. E., Albrechtsen A., Li Y., Kim S. Y., Korneliussen T., *et al.* (2011) Natural Selection Affects Multiple Aspects of Genetic Variation at Putatively Neutral Sites across the Human Genome. *PLOS Genetics* 7(10): e1002326. https://doi.org/10.1371/journal.pgen.1002326

Marsden, C. D., Ortega-Del Vecchyo, D., O'Brien, D. P., Taylor, J. F., Ramirez, O., Vilà, C., ... & Lohmueller, K. E. (2016). Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences*, *113*(1), 152-157.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Messer, P. W., & Petrov, D. A. (2013). Frequent adaptation and the McDonald–Kreitman test. *Proceedings of the National Academy of Sciences*, *110*(21), 8615-8620.

Mooney, J. A., Agranat-Tamir, L., Pritchard, J. K., & Rosenberg, N. A. (2023). On the number of genealogical ancestors tracing to the source groups of an admixed population. *Genetics*, *224*(3), iyad079. https://doi.org/10.1093/genetics/iyad079

Muñoz-Fuentes, V., Rienzo, A., & Vilà, C. (2011). Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. *PLoS ONE*, *6*(11), 1–7. https://doi.org/10.1371/journal.pone.0025498

Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., … Ponting, C. P. (2009). Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genetics*, *5*(12). https://doi.org/10.1371/journal.pgen.1000753

Parvanov, E. D., Petkov, P. M., & Paigen, K. (2010). Prdm9 controls activation of mammalian recombination hotspots. *Science*, *327*(5967), 835. https://doi.org/10.1126/science.1181495

Robinson, J. A., Räikkönen, J., Vucetich, L. M., Vucetich, J. A., Peterson, R. O., Lohmueller, K. E., & Wayne, R. K. (2019). Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances*, *5*(5), eaau0757.

Sawyer, S. A., & Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genetics*, *132*(4), 1161–1176. https://doi.org/10.1093/genetics/132.4.1161