UNIVERSITY OF CALIFORNIA

Los Angeles

Methods and models for the analysis of genetic variation across species using large-scale genomic data

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Tanya Ngoc Phung

2018

ABSTRACT OF THE DISSERTATION

Methods and models for the analysis of genetic variation across species using large-scale genomic data

by

Tanya Ngoc Phung
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2018
Professor Kirk Edward Lohmueller, Chair

Understanding how different evolutionary processes shape genetic variation within and between species is an important question in population genetics. The advent of next generation sequencing has allowed for many theories and hypotheses to be tested explicitly with data. However, questions such as what evolutionary processes affect neutral divergence (DNA differences between species) or genetic variation in different regions of the genome (such as on autosomes versus sex chromosomes) or how many genetic variants contribute to complex traits are still outstanding. In this dissertation, I utilized different large-scale genomic datasets and developed statistical methods to determine the role of natural selection on genetic variation between species, sex-biased evolutionary processes on shaping patterns of genetic variation on the X chromosome and autosomes, and how population history, mutation, and natural selection

interact to control complex traits. First, I used genome-wide divergence data between multiple pairs of species ranging in divergence time to show that natural selection has reduced divergence at neutral sites that are linked to those under direct selection. To determine explicitly whether and to what extent linked selection and/or mutagenic recombination could account for the pattern of neutral divergence across the genome, I developed a statistical method and applied it to human-chimp neutral divergence dataset. I showed that a model including both linked selection and mutagenic recombination resulted in the best fit to the empirical data. However, the signal of mutagenic recombination could be coming from biased gene conversion.

Comparing genetic diversity between the X chromosome and the autosomes could provide insights into whether and how sex-biased processes have affected genetic variation between different genomic regions. For example, X/A diversity ratio greater than neutral expectation could be due to more X chromosomes than expected and could be a result of mating practices such as polygamy where there are more reproducing females than males. I next utilized whole-genome sequences from dogs and wolves and found that X/A diversity is lower than neutral expectation in both dogs and wolves in ancient time-scales, arguing for evolutionary processes resulting in more males reproducing compared to females. However, within breed dogs, patterns of population differentiation suggest that there have been more reproducing females, highlighting effects from breeding practices such as popular sire effect where one male can father many offspring with multiple females.

In medical genetics, a complete understanding of the genetic architecture is essential to unravel the genetic basis of complex traits. While genome wide association studies (GWAS) have discovered thousands of trait-associated variants and thus have furthered our understanding of the genetic architecture, key parameters such as the number of causal variants and the

mutational target size are still under-studied. Further, the role of natural selection in shaping the genetic architecture is still not entirely understood. In the last chapter, I developed a computational method called InGeAr to infer the mutational target size and explore the role of natural selection on affecting the variant’s effect on the trait. I found that the mutational target size differs from trait to trait and can be large, up to tens of megabases. In addition, purifying selection is coupled with the variant’s effect on the trait. I discussed how these results support the omnigenic model of complex traits.

In summary, in this dissertation, I utilized different types of large genomic dataset, from genome-wide divergence data to whole genome sequence data to GWAS data to develop models and statistical methods to study how different evolutionary processes have shaped patterns of genetic variation across the genome.

The dissertation of Tanya Ngoc Phung is approved.

Sriram Sankararaman

Melissa Wilson-Sayres

Janet S Sinsheimer

Robert Wayne

Kirk Edward Lohmueller, Committee Chair

University of California, Los Angeles

2018

Dedicated to Michael and Newton

TABLES OF CONTENTS

List of Figures……………………………………………………………………………ix

List of Tables…………………………………………………………………………...xiii

Acknowledgements…………………………………………………………………….xv

Vita……………………………………………………………………………………….xviii

Chapter 1. Introduction…………………………………………………………………...1

Chapter 2. Determining the effect of natural selection on linked neutral divergence across species……………………………………………………………………………………….5

2.1 Abstract………………………………………………………………………..5

2.2 Introduction…………………………………………………………………6

2.3 Results…………………………………………………………………………..10

2.4 Discussion………………………………………………………………………22

2.5 Methods…………………………………………………………………………30

2.6 Figures…………………………………………………………………………..36

2.7 Tables…………………………………………………………………………...53

Chapter 3. Detecting mutagenic recombination using genome-wide divergence data……..61

3.1 Abstract………………………………………………………………………61

3.2 Introduction……………………………………………………………………..62

3.3 Methods…………………………………………………………………………63

3.4 Results…………………………………………………………………………..68

3.5 Discussion………………………………………………………………………71

3.6 Figures…………………………………………………………………………..73

3.7 Tables…………………………………………………………………………...79

Chapter 4. Complex patterns of sex-biased demography in canines………………………..85

4.1 Abstract……………………………………………………………………………85

4.2 Introduction………………………………………………………………………..86

4.3 Results……………………………………………………………………………..89

4.4 Discussion…………………………………………………………………………96

4.5 Methods……………………………………………………………………………98

4.6 Figures……………………………………………………………………………106

4.7 Tables……………………………………………………………………………116

Chapter 5. Inference of the mutational target size supports the omnigenic model for complex traits……………………………………………………………………………134

5.1 Abstract…………………………………………………………………………..134

5.2 Introduction………………………………………………………………………135

5.3 Results……………………………………………………………………………138

5.4 Discussion………………………………………………………………………..147

5.5 Methods…………………………………………………………………………..150

5.6 Figures……………………………………………………………………………155

5.7 Tables……………………………………………………………………………170

References…………………………………………………………………………...174

LIST OF FIGURES

Figure 2.1. Models of how genealogies are affected by selection at linked neutral sites.………36

Figure 2.2. Four-fold degenerate sites show similar levels of divergence as our putatively neutral noncoding sites…………………………………………………………………………...37

Figure 2.3. Human-primate divergence is reduced at putatively neutral sites near selected sites……………………………………………………………………………………………38

Figure 2.4. The correlation between human recombination rate and neutral divergence is stronger near genes……………………………………………………………………………39

Figure 2.5. Human-rodent divergence is reduced at putatively neutral sites near selected sites..40

Figure 2.6. Observed and modeled genome-wide distributions of human-chimp divergence and human-mouse divergence in 100 kb windows………………………………………………...41

Figure 2.7. Models incorporating background selection can generate patterns of neutral divergence that recapitulate the empirical correlations…………………………………………42

Figure 2.8. A two-locus model for the effect of background selection on divergence…………43

Figure 2.9. Background selection is predicted to affect neutral divergence across a range of split times and ancestral population sizes………………………………………………………..45

Figure 2.10. Variance of the total divergence attributable to the variance in levels of ancestral polymorphism…………………………………………………………………………………46

Figure 2.11. Relationship between divergence and functional content, human recombination, and McVicker’s B-values as a function of GERP score cutoff……………………………………47

Figure 2.12. Correlations between human-primate divergence and genomic features persist when filtering the 25% of the genome with the highest GERP scores……………………………....48

Figure 2.13. Correlations between human-rodent divergence and genomic features change when filtering the 25% of the genome with the highest GERP scores……………………………….49

Figure 2.14. Observed and modeled genome-wide distributions of human-chimp divergence and human-mouse divergence in 100kb windows when filtering sites whose GERP scores fall into the top 25% of the genome-wide distribution…………………………………………………50

Figure 2.15. Models incorporating background selection can recapitulate the empirical correlations after removing sites with GERP scores falling in the top 25% of the genome-wide distribution ………………………………………………………………………………51

Figure 2.16. Human-mouse mutational parameters used for the simulations…………………..52

Figure 3.1. Distribution of empirical population-scale recombination rate, ρ across 100kb-windows………………………………………………………………………………………73

Figure 3.2. Test dataset was generated with background selection……………………………..74

Figure 3.3. Test dataset was generated with mutagenic recombination……………………….....75

Figure 3.4. Test dataset was generated with both background selection and mutagenic recombination……………………………………………………………………………………76

Figure 3.5. Model with both background selection and mutagenic recombination can recapitulate the empirical human-chimp neutral divergence………………………………………………77

Figure 3.6. Biased gene conversion confounds signal for mutagenic recombination…………..78

Figure 4.1. X-linked and autosomal genetic diversity across canids ………………………….106

Figure 4.2. Model of demographic history including a bottleneck, redrawn from the fastsimcoal2 manual……………………………………………………………………………………..108

Figure 4.3. SFSs simulated under our best-fitting demographic models (blue bars) compared to the empirical SFSs on the autosomes (grey bars) in multiple canid populations………………109

Figure 4.4. Genetic diversity (π) computed from the SFSs generated by using inferred demographic parameters (blue bars) compared with uncorrected empirical genetic diversity on the autosomes (grey bars) in multiple canid populations……………………………………….110

Figure 4.5. Effective population size estimates for multiple canid populations………………111

Figure 4.6. Empirical SFSs (grey bars) compared to the SFSs simulated with a null C value of 0.75 (blue bars) and a C value that yielded the highest likelihood (yellow bars) for multiple canid populations……………………………………………………………………………...112

Figure 4.7. Empirical genetic diversity (π, grey bars) compared to diversity computed from SFSs simulated with a null C value of 0.75 (blue bars) and a C value that yielded the highest likelihood (yellow bars) for multiple canid populations …………………………………………………..113

Figure 4.8. Sex biased demography on recent time scales……………………………………..115

Figure 5.1. Trait-associated variants discovered from GWAS provide insights into the genetic architecture of complex traits…………………………………………………………………...155

Figure 5.2. τ could account for the negative correlation between effect size and allele frequency……………………………………………………………………………………..156

Figure 5.3. Overview of InGeAr…………………………………………………………….157

Figure 5.4. The mutational target size determines the number of causal variants……………..158

Figure 5.5. Both M and τ affect the number of GWAS hits and effect size …………………..159

Figure 5.6. Expected GWAS hits are similar for two different pairs of (τ, M) but the effect size can distinguish them and vice versa…………………………………………………………….160

Figure 5.7. Prior distribution of M and τ……………………………………………………161

Figure 5.8. Posterior distribution of M, τ, and their joint distribution…………………………162

Figure 5.9. Assess model fit for height……………………………………………………...163

Figure 5.10. Properties of causal variants, GWAS hits, and remaining undiscovered causal variants……………………………………………………………………………….164

Figure 5.11. Distribution of effect sizes for height…………………………………………….165

Figure 5.12. Posterior distribution of M, τ, and ρ when including pleiotropy in the inference..166

Figure 5.13. M and τ for other UKBiobank traits……………………………………………...167

Figure 5.14. Assess model fit for balding type 1………………………………………………168

Figure 5.15. Comparison of τ between InGeAr and Schoech et al…………………………….169

LIST OF TABLES

Table 2.1. Summary statistics of divergence at four-fold sites and at putatively neutral regions.53

Table 2.2. Correlation coefficients of human-primate divergence and functional content……...54

Table 2.3. Correlation coefficients of human-primate divergence and recombination rate……..55

Table 2.4. Correlation coefficients of human-rodent divergence and functional content……….56

Table 2.5. Correlation coefficients of human-rodent divergence and McVicker's B-values…...57

Table 2.6. Summary of parameters used for the coalescent simulations……………………..58

Table 2.7. Comparison of the mean and standard deviation of the empirical and simulated divergence……………………………………………………………………………..59

Table 2.8. GERP RS score cutoff……………………………………………………..60

Table 3.1. Parameters to infer for each model………………………………………………..79

Table 3.2. Maximum likelihood estimates and best likelihoods when test data was generated with just background selection…………………………………………………………..80

Table 3.3. Maximum likelihood estimates and best likelihoods when test data was generated with just mutagenic recombination…………………………………………………………81

Table 3.4. Maximum likelihood estimates and best likelihoods when test data was generated with both background selection and mutagenic recombination………………………………….82

Table 3.5. Maximum likelihood estimates and best likelihoods for each model for empirical human-chimp neutral divergence before accounting for biased gene conversion……………….83

Table 3.6. Maximum likelihood estimates and best likelihoods for each model for empirical human-chimp neutral divergence after accounting for biased gene conversion…………………84

Table 4.1. Data used in this study…………………………………………………………...116

Table 4.2. Dog-cat divergence, mutation rate, and estimates of male mutation bias…………..120

Table 4.3. Linked selection is stronger on the X chromosome than on the autosomes………..121

Table 4.4. Number of sites and number of variants in each threshold that remained after removing neutral sites potentially linked to selected sites…………………………………...122

Table 4.5. Demographic parameters that best fit the autosomal site frequency spectrum……..123

Table 4.6. SNP count Poisson log-likelihoods comparing the fit between the best fit demographic model to the observed SFS on the autosomes……………………………………124

Table 4.7. Likelihood ratio tests comparing models of sex-biased demography in multiple canid populations………………………………………………………………………………...127

Table 4.8. Individuals used in Q_FST analyses………………………………………………..133

Table 5.1. M and τ when considering variants with large effect size…………………………170

Table 5.2. M and τ when incorporating pleiotropy……………………………………………177

Table 5.3. Summary of the posterior distribution of M and τ for other traits…………………172

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor, Kirk Lohmueller. Kirk, thank you so much for giving me an opportunity to join your group even though I did not know anything about population genetics at the time. Thank you so much for being the most wonderful advisor that I could have asked for. Besides your brilliance, you have been so patient, understanding, approachable and easy to talk to. Thank you for letting me explore working in industry. Thank you for working with me to figure out a maternity plan that worked for me and my family. I am so lucky to have had the opportunity to work with you. I have learned so much from you, and I hope that we will continue to collaborate on projects together in the future.

I would like to acknowledge Melissa Wilson Sayres. Melissa, thank you so much for your mentorship for the past few years. You have taught me everything I know about sex chromosomes, I have also learned from you how to be a more compassionate and caring individual. Thank you so much for your willingness to work with me and I cannot wait to pursue projects together with you as a post-doc.

I would like to thank other faculty that have helped me along the way. Thank you, Janet, for teaching one of my most favorite class at UCLA, Theoretical Modelling in Genetics. Thank you, Bob, for your valuable input on Chapter 4 of my dissertation. Thank you, Bogdan, for working with me on Chapter 5 of my dissertation. Thank you, Sriram, for your valuable feedback along the way.

Thank you to all members of the Lohmueller lab. I specifically wanted to acknowledge Clare Marsden who was the first person I interacted with in Kirk's lab and one of the reasons why I decided to choose his lab. Clare, you are always so willing to help and has been such a wonderful resource, in both work-related and personal-related issues. You have always been there for me, in both the good and bad times. Your patience and willingness to listen have helped me overcome one of the most challenging periods of my life. So, I thank you for that.

Diego, thank you for all of your help during the first couple of years of my graduate studies. You are by far the nicest, most positive person I have ever met, and I always feel so much better after having a conversation with you. Bernard, you made my experience being a TA so much easier so thank you for that. Ying, thank you for being my office mate and for listening to me these past few years. I am truly happy for you that you can now establish your own group, and I wish you all the best. Christian, you have taught me so much about population genetics and thank you for helping me with my projects earlier on in my graduate studies. I am truly fortunate to have you as my officemate because you seem to have a solution to every problem. Jazlyn, it has been so fun seeing how much you have developed throughout your graduate career. I am certain that whatever you choose to do, you will excel at it. Annabel, thank you for teaching us all about whale's microbiome and otters. Your enthusiasm towards your research is truly contagious. Arun, thank you so much for all of your help on Chapter 5 of this dissertation. Eduardo, I wish you all the best in your future endeavor. Norris, thank you for all your hard work in pushing the linked selection project forward. And last but not least, Jesse, thank you for giving me an opportunity to mentor you when you were an undergraduate. Thank you for choosing the Lohmueller lab because it means I get to work with you longer. You were such a joy to mentor, and I have no doubt that you will be successful in graduate school and beyond.

I would like to thank the staff that has made my time in the Bioinformatics program smoother and more enjoyable: Pamela Hurley, Allison Taka, and Mandy McWeeney. In addition, thank you to David Tomita in the Biomathematics department for all of your help while I was under the System and Integrative Biology training program.

Thank you to my parents who were so courageous in uprooting our whole family to move to the United States so that I could have a better life than they did. Raising my own daughter now, I am even more thankful for what they have done for me.

Thank you to my daughter, Newton. Thank you for giving the motivation each and every day to accomplish my goals so I can spend time with you. You have transformed me into a more productive researcher. You have also taught me to value the present and not worry too much about the uncertainty that the future holds. It has been truly a joy to see you grow and develop your own personality.

Finally, thank you to my husband, Michael. You have been on this graduate school journey since day 0. Thank you for encouraging me to pursue graduate school, and for supporting me in so many ways. Thank you for driving me across the country (twice). Thank you for helping me with your all my IT needs by installing whatever I need and fixing whatever broken. Thank you for being the more patient, calmer, and more sensible partner. Thank you for being the best father to our little Newt. She is incredibly lucky to have you as her daddy.

Funding: This work as partially supported through the Systems and Integrative Biology Training Program (NIH T32-GM008185) and the Biomedical and Big Data Training Program (NIH-NCI National Cancer Institute T32CA201160). The contents of this dissertation are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

VITA

TANYA NGOC PHUNG

EDUCATION

2014	Bachelor of Science in Biology
	Massachusetts Institute of Technology, Cambridge, Massachusetts

FELLOWSHIPS

2017 – 2018	Biomedical and Big Data Training Program
	University of California, Los Angeles, Los Angeles, California
2015 – 2017	Systems and Integrative Biology Training Program
	University of California, Los Angeles, Los Angeles, California

PUBLICATIONS

Published work

Beichman, A.C., **Phung, T.N.**, and Lohmueller, K.E. (2017). Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories. G3 Genes Genomes Genet. *7*, 3605–3620.

Phung, T.N., Huber, C.D., and Lohmueller, K.E. (2016). Determining the Effect of Natural Selection on Linked Neutral Divergence across Species. PLOS Genet *12*, e1006199.

Aakre, C.D., Herrou, J., **Phung, T.N.**, Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. Cell.

Aakre, C.D., **Phung, T.N.**, Huang, D., and Laub, M.T. (2013). A bacterial toxin inhibits DNA replication elongation through a direct interaction with the β sliding clamp. Mol. Cell *52*, 617–628.

Liu, Z., Rader, J., He, S., **Phung, T.**, and Thiele, C.J. (2013). CASZ1 inhibits cell cycle progression in neuroblastoma by restoring pRb activity. Cell Cycle Georget. Tex *12*, 2210–2218.

Preprints

Phung, T.N., Wayne, R.K., Sayres, M.A.W., and Lohmueller, K.E. (2018). Complex patterns of sex-biased demography in canines. BioRxiv 362731.

Webster, T.H., Couse, M., Grande, B.M., Karlins, E., **Phung, T.**, Richmond, P.A., Whitford, W., and Sayres, M.A.W. (2018). Identifying, understanding, and correcting technical biases on the sex chromosomes in next-generation sequencing data. BioRxiv 346940.

Manuscripts in preparation

Phung, T.N., Huber, C.D., Lohmueller, K.E. Detecting mutagenic recombination using genome-wide divergence data.

Phung, T.N., Pasaniuc, B., Lohmueller, K.E. Inference of the mutational target size supports the omnigenic model for complex traits.

CHAPTER 1

Introduction

DNA is the genetic blueprint that determines an individual's trait such as eye color or height or susceptibility to certain diseases. DNA differs in different regions of the genome, between individuals of the same species, and between individuals from different species. Differences in DNA among individuals are termed genetic variation. Understanding what evolutionary processes shape genetic variation across the genome within and between species has been one of the main research goals in population genetics. This understanding is crucial to understand the evolutionary history of different species. In addition, understanding how and why DNA varies across the genome is key to unravel the genetic basis of human traits and diseases.

It has been observed that at sites that are expected to evolve under a neutral evolutionary model (neutral sites), genetic diversity is lower in regions of low recombination as compared to at regions of high recombination (Begun and Aquadro, 1992). Studies have attributed this pattern to the role of natural selection in reducing genetic diversity at regions of low recombination in multiple species (Cutter and Payseur, 2013; Lohmueller et al., 2011). Population geneticists have postulated two mechanisms to describe how natural selection could affect linked neutral sites: genetic hitchhiking and background selection (Cutter and Payseur, 2013). Though there is a consensus that linked selection has reduced genetic diversity within a population, whether linked selection has also shaped neutral divergence across the genome was still a topic of debate.

To address this long-debated question in population genetics, in Chapter 2 of my dissertation, I analyzed genome-wide divergence data from pairs of species with different split times (i.e. recent split time as in human and chimp and distant split time as in human and mouse). I also developed a coalescent framework to test whether a neutral model or a model with linked selection can recapitulate empirical patterns. This work highlights how widespread natural selection is across the genome.

An empirical observation that supports the role of natural selection in reducing linked neutral divergence explored in Chapter 2 is the positive correlation between neutral divergence and recombination. While Chapter 2 concluded that linked selection has reduced linked neutral divergence at regions of low recombination, previous research has proposed that mutagenic recombination could also generate a positive correlation between neutral divergence and recombination (Hellmann et al., 2003). To determine to whether and to what extent mutagenic recombination contributes to shaping patterns of divergence, in Chapter 3 of this dissertation, I developed a statistical method that models linked selection by reducing the ancestral population using estimates of linked selection (i.e. the B values as in McVicker et al. and models mutagenic recombination by a linear relationship between mutation rate and recombination rate. I also accounted for the potential effect of biased gene conversion by filtering the data to remove weak to strong mutations. I applied this method to human-chimp neutral divergence data to show whether and to what extent linked selection, mutagenic recombination, and biased gene conversion can explain the positive correlation between neutral divergence and recombination. This chapter confirms the role of natural selection and also illustrates the importance of correcting for biased gene conversion.

In Chapter 2 and Chapter 3 of this dissertation, I explored what evolutionary processes affect DNA differences between species. In addition to varying within and across species, DNA also varies across different regions of the genome, particularly between the autosomes and the sex chromosome. Understanding genetic variation between the X chromosome and the autosomes have yielded insights into how sex-biased processes such as sex-biased migration patterns or sex-biased mating practices have shaped the evolutionary history of humans (Hammer et al., 2008; Keinan et al., 2009). In recent years, due to the advent of next generation sequencing, many whole genomes have been sequenced in dogs to study their evolutionary history because of intense interest in dog domestication from both scientists and the public alike. Sex-biased mating practices are prevalent in both the wild and domesticated populations of canids. For example, male-biased migration and multiple paternity have been observed in wolves (Vonholdt et al., 2008). In dogs, the desire for popular sire has led to female-biased mating where one male fathers many offspring with multiple females (Ostrander and Kruglyak, 2000). However, existing population genomic studies of canid demographic history has not tested whether any evolutionary processes have been sex-biased. Given how comparing and contrasting genetic variation between the X chromosome and the autosomes has been fruitful in humans in understanding the role of evolutionary forces, in Chapter 4 of this dissertation, I utilized whole genome sequences of dogs and wolves to study whether their evolutionary history has been sex-biased. This work highlights how genetic data are used to validate mating patterns observed in field studies.

The work presented in Chapter 2, 3, and 4 of this dissertation has focused mainly on understanding how different evolutionary processes have impacted genetic variation across the genome. An equally important goal of studying genetic variation is to understand its role in contributing to a phenotypic trait. In fact, determining the number of trait-associate variants for a trait has been a key focus of medical genetics. As a result, many genome-wide association studies (GWAS) have been performed to understand what causal variants are contributing to a trait’s phenotype. Even though GWAS have identified thousands of trait-associated variants, these GWAS hits are only a subset of the total number of causal variants. Further, these causal variants are a subset of the total number of sites in the genome that, when mutated, would give rise to a trait-associated variant. We call this the mutational target size. The mutational target size has not been well studied. Understanding the mutational target size is important because we can make inferences about the number of causal variants giving rise to a trait. Further, understanding the role of natural selection is also important to fully understand the genetic architecture of complex traits. Natural selection has been hypothesized to affect the effect a variant has on a trait is the trait is affecting reproductive fitness (Eyre-Walker, 2010; Schoech et al., 2017). To contribute to better understand the genetic architecture of complex traits, I developed a statistical method called InGeAr to infer the mutational target size and infer the relationship between selection acting on a variant and a variant's effect on the trait. I applied InGeAr using summary statistics from UKBiobank GWAS for multiple complex traits. This chapter illustrates how summary statistics from GWAS can be used to infer the genetic architecture of complex traits and how our results provide support for the omnigenic model of complex traits.

CHAPTER 2

Determining the effect of natural selection on linked neutral divergence across species

2.1 Abstract

A major goal in evolutionary biology is to understand how natural selection has shaped patterns of genetic variation across genomes. Studies in a variety of species have shown that neutral genetic diversity (intra-species differences) has been reduced at sites linked to those under direct selection. However, the effect of linked selection on neutral sequence divergence (inter-species differences) remains ambiguous. While empirical studies have reported correlations between divergence and recombination, which is interpreted as evidence for natural selection reducing linked neutral divergence, theory argues otherwise, especially for species that have diverged long ago. Here we address these outstanding issues by examining whether natural selection can affect divergence between both closely and distantly related species. We show that neutral divergence between closely related species (e.g. human-primate) is negatively correlated with functional content and positively correlated with human recombination rate. We also find that neutral divergence between distantly related species (e.g. human-rodent) is negatively correlated with functional content and positively correlated with estimates of background selection from primates. These patterns persist after accounting for the confounding factors of hypermutable CpG sites, GC content, and biased gene conversion. Coalescent models indicate that even when the contribution of ancestral polymorphism to divergence is small, background selection in the ancestral population can still explain a large proportion of the variance in divergence across the genome, generating the observed correlations. Our findings reveal that, contrary to previous intuition, natural selection can indirectly affect linked neutral divergence between both closely and distantly related species. Though we cannot formally exclude the

possibility that the direct effects of purifying selection drive some of these patterns, such a scenario would be possible only if more of the genome is under purifying selection than currently believed. Our work has implications for understanding the evolution of genomes and interpreting patterns of genetic variation.

2.2 Introduction

Determining the evolutionary forces affecting genetic variation has been a central goal in population genetics over the past several decades. A large body of empirical and theoretical work has suggested that neutral genetic variation within a species (diversity) can be influenced by nearby genetic variants that are affected by natural selection (Cutter and Payseur, 2013). This can occur via two mechanisms. In a selective sweep, a neutral allele linked to a beneficial mutation will reach high frequency (Kaplan et al., 1989; Maynard Smith and Haigh, 1974). Selective sweeps reduce neutral genetic variation near regions of the genome that are directly affected by natural selection. The second process, background selection, also reduces neutral genetic variation (Charlesworth, 2012; Charlesworth et al., 1993; Hudson and Kaplan, 1995; Nordborg et al., 1996). Here, purifying selection that eliminates deleterious mutations also removes nearby neutral genetic variation. Many empirical studies have found strong evidence for the effects of background selection and selective sweeps affecting patterns of neutral genetic diversity (intra-species DNA differences) across the human genome. For example, several studies have reported a correlation between genetic diversity and recombination rate (Cai et al., 2009; Hellmann et al., 2003, 2005, 2008; Lohmueller et al., 2011; Nachman, 2001). This correlation can be driven by selective sweeps and background selection because these processes affect a larger number of base pairs in areas of the genome with a low recombination rate than with a high recombination rate. Additionally, other studies found reduced neutral genetic diversity surrounding genes (Cai et al., 2009; Enard et al., 2014; Hernandez et al., 2011; Lohmueller et al., 2011; McVicker et al., 2009; Payseur and Nachman, 2002), which is consistent with the idea that there is more selection occurring near functional elements of the genome.

While the evidence for natural selection reducing genetic diversity at linked neutral sites is unequivocal, the effect of natural selection on linked neutral divergence between species (inter-species DNA differences) is less clear. Elegant theoretical arguments have suggested selection does not affect the substitution rate at linked neutral sites (Birky and Walsh, 1988; Cutler, 1998). However, these theoretical arguments do not include mutations that arose in the common ancestral population, the population that existed prior to the split and formation of two descendant lineages. Such ancestral polymorphism has been shown to be a significant confounder in estimating population divergence times (Edwards and Beerli, 2000). When also including ancestral polymorphism, it becomes less clear whether selection affects divergence at linked neutral sites.

Based on coalescent arguments, neutral polymorphism in the ancestral population will be affected by linkage to selected sites the same way as genetic diversity within a population (Figure 2.1). Presumably, neutral divergence between closely related species, with lots of ancestral polymorphism, could be affected by selection. Indeed, McVicker et al. demonstrated that background selection could explain the variation in human-chimp neutral divergence across the genome (McVicker et al., 2009). Additionally, Cruickshank and Hahn (Cruickshank and Hahn, 2014) found that divergence between recently separated species pairs was reduced in regions of low recombination and in “islands of speciation”. They attributed at least some of these patterns to selection affecting linked neutral sites.

However, the reduction in neutral diversity in the ancestral population is thought to have a negligible effect and/or be undetectable when considering neutral divergence from species with a very long divergence time because there would be many opportunities for mutations to occur after the two lineages split (Figure 2.1) (Birky and Walsh, 1988; Hellmann et al., 2003). These neutral mutations that occur after the split would not be influenced by selection at linked neutral sites (Birky and Walsh, 1988) and would dilute the signal from the ancestral polymorphism. Thus, it is generally believed that selection at linked neutral sites should not affect divergence between distantly related species. An example of this argument was presented by Hellmann et al. (Hellmann et al., 2003). They argued that the positive correlation between human-baboon divergence and human recombination was due to mutagenic recombination, rather than selection affecting linked neutral sites, because of the long split time between humans and baboons (>20 million years). Reed et al. suggested that though it is unlikely background selection by itself could explain the entire correlation observed by Hellmann et al., background selection may still contribute to divergence (Hellmann et al., 2003; Reed et al., 2005). However, there has been little quantitative investigation of the effect that selection has on divergence at linked neutral sites among distantly divergent species when including ancestral polymorphism.

In addition to conflicting conceptual predictions about the expected effect of selection on divergence at linked neutral sites, empirical studies also have been ambiguous. While some studies found no evidence for a correlation between divergence and recombination such as in *Drosophila* (Begun and Aquadro, 1992; McGaugh et al., 2012) or in yeast (Noor, 2008), other studies have reported correlations between divergence and recombination in *Drosophila* (Begun et al., 2007; Kulathinal et al., 2008). Further, positive correlations between human-chimpanzee divergence and human recombination rate (Cai et al., 2009; Hellmann et al., 2005; Lohmueller et al., 2011), human-macaque divergence and human female recombination rate (Tyekucheva et al., 2008), or human-baboon divergence and human recombination rate (Hellmann et al., 2003) have been reported. Finally, even though there was evidence for a strong reduction in human-chimpanzee divergence and human-macaque divergence surrounding genes, McVicker et al. attributed the reductions seen for human-dog divergence to variation in mutation rates (McVicker et al., 2009; Tyekucheva et al., 2008). Thus, the degree to which divergence is affected by selection across species with different split times remains elusive.

Determining whether and how selection affects linked neutral divergence is critical to understanding the evolutionary forces influencing genetic variation and mutational processes. If selection in the ancestral population only has a limited effect on divergence, it would suggest correlations between recombination and divergence to be evidence of mutagenic recombination. This may further suggest the need to consider recombination rates when modeling variation in mutation rates across the genome (Arbeithuber et al., 2015; Francioli et al., 2015; Hellmann et al., 2003; Lercher and Hurst, 2002; Pratto et al., 2014). Because mutations rates have been difficult to estimate reliably in humans (Scally and Durbin, 2012; Ségurel et al., 2014), understanding the biological factors influencing them will be of paramount importance for obtaining improved estimates. If, on the other hand, selection can affect linked neutral divergence, reductions of linked neutral divergence surrounding genes would suggest an abundance of selection affecting linked neutral sites (Sella et al., 2009). Selection affecting linked neutral diversity and divergence is at odds with the neutral and nearly neutral theories (Akashi et al., 2012; Kimura, 1983; Ohta, 1973), which have been the prevailing views in molecular population genetics for the last several decades. It would also suggest the need to consider the effects of selection when estimating mutation rates from neutral divergence.

Here we aim to examine the effects of selection on linked neutral divergence for pairs of species with a range of split times. We first present evidence that neutral divergence is reduced at putatively neutral sites close to selected sites across a wide range of taxa, including those with split times as long as 75 million years ago. Factors such as hypermutable CpG sites, GC content, or biased gene conversion by themselves cannot explain these results. We then use coalescent simulations to explore whether models incorporating background selection in the ancestral population could generate the empirical patterns. We also present a theoretical argument as to how background selection can affect variation in neutral divergence across the genome, even for species with a long split time such as human and mouse. Finally, we show that purifying selection directly reducing divergence at putatively neutral sites cannot explain these findings unless a large fraction of the genome is directly under selection, or there is a substantial number of sites under selection in the human or mouse lineage that are not conserved across species. Even though we cannot formally reject the direct effects of purifying selection from driving some of these correlations, our empirical and simulation-based findings indicate that natural selection can indirectly affect neutral genetic divergence. In sum, the view that selection does not affect divergence at linked neutral sites between distantly diverged species should be re-considered.

2.3 Results

Obtaining putatively neutral divergence

We wished to test whether the genetic divergence at a linked neutral site is influenced by the indirect effects of natural selection. As such, we set out to obtain putatively neutral sites by removing sites that were potentially functional and under the direct effects of purifying selection. In particular, a site was considered putatively neutral if it was (1) located at least 5kb from the starting or ending position of an exon, (2) not located within a phastCons element that was calculated over different phylogenic scopes, (3) not alignable between human and zebrafish, and (4) not found within the top 10% of most conserved Genomic Evolutionary Rate Profiling (GERP) scores (Davydov et al., 2010). Criteria 2 and 3 remove sites that are likely to be conserved across species and therefore not neutral. We chose these filtering criteria following previous studies (Cai et al., 2009; Lohmueller et al., 2011; Moorjani et al., 2016). Additionally, we chose to remove the top 10% of sites having the most extreme GERP scores because previous work suggests <10% of the genome was under the direct effect of selection (Cooper et al., 2004; Davydov et al., 2010; Gulko et al., 2015; Meader et al., 2010; Mouse Genome Sequencing Consortium et al., 2002; Pollard et al., 2010; Rands et al., 2014; Schrider and Kern, 2015; Siepel et al., 2005; Ward and Kellis, 2012). The putatively neutral sites close to genes show comparable levels of divergence to four-fold degenerate sites (Figure 2.2, Table 2.1). As four-fold degenerate sites are often used as a neutral standard in molecular evolution, the fact that they show similar levels of divergence as our putatively neutral noncoding sites argues that our putatively neutral sites are unlikely to be under additional direct effects of selection.

Effects on putatively neutral divergence between humans and primates

To understand the evolutionary factors affecting linked neutral divergence between closely related species, we examined human-primate divergence, particularly human-chimp divergence and human-orangutan divergence. First, we explored the relationship between neutral human-primate divergence and functional content, defined as the proportion of sites within a 100kb-window that overlapped with an exon or a phastCons region. We hypothesized that if natural selection contributes to the reduction of divergence at linked neutral sites, its effect would be more pronounced at regions with greater functional content (Payseur and Nachman, 2002). This hypothesis predicts a negative correlation between functional content and neutral divergence. To test this, we divided the human genome into non-overlapping windows of 100kb and obtained putatively neutral divergence for each window as described above. We found a negative correlation between functional content and neutral divergence between pairs of closely related species (Spearman's $\rho_{human\text{-}chimp}$ = -0.235, $P < 10^{-16}$, Spearman's $\rho_{human\text{-}orang}$ = -0.204, $P < 10^{-16}$, Figure 2.3A, Figure 2.3B, Table 2.2).

We next examined the relationship between human-primate neutral divergence and broad-scale human recombination rate which we obtained from the deCODE genetic map (Kong et al., 2002). While recombination has not been conserved throughout evolutionary history, the recombination rate at the broad-scale level (i.e. 100kb) was shown to be correlated between human and chimp (Auton et al., 2012; Stevison et al., 2015). We found a positive correlation between neutral human-primate divergence and human recombination rate (Spearman's $\rho_{human\text{-}chimp}$ = 0.234, $P < 10^{-16}$, Spearman's $\rho_{human\text{-}orang}$ = 0.249, $P < 10^{-16}$, Figure 2.3C, Figure 2.3D, Table 2.3), which indicates that neutral human-primate divergence is reduced in regions of low recombination rate. Additionally, when we stratified windows into those that were near genes and those that were far from genes based on the proportion of sites in each window that overlapped with a RefSeq transcript, we found that the correlation between divergence and recombination is stronger for windows with a higher overlap with RefSeq transcripts (Figure 2.4). These observations indicate that neutral divergence is reduced at sites that are more tightly linked to those under the direct effect of selection, consistent with the hypothesis that natural selection indirectly reduces linked neutral divergence.

These two correlations are robust to the presence of multiple confounding factors. First, the correlations are robust to the choice of window size used for analysis as they persisted when using 50 kb windows (Table 2.2, Table 2.3). Second, some features of the genome such as hypermutable CpG sites or GC content are known to correlate with genic content (Bernardi, 2000; Polak et al., 2015; Supek and Lehner, 2015). To test whether these features confounded the correlations found in our data, we repeated our analyses removing potential CpG sites by omitting sites preceding a G or following a C (McVicker et al., 2009). The correlations persisted after filtering out CpG sites (Table 2.2, Table 2.3). We next computed partial correlations controlling for GC content. Similarly, we found that the correlations persisted (Table 2.2, Table 2.3).

Biased gene conversion is an additional evolutionary force that has been shown to influence patterns of divergence (Duret and Arndt, 2008; Galtier and Duret, 2007). In this process, double-strand breaks in the DNA in individuals heterozygous for AT/GC variants will be preferentially repaired with the GC allele, resulting in AT → GC substitutions occurring at a higher rate than GC → AT substitutions (Berglund et al., 2009; Duret and Arndt, 2008; Duret and Galtier, 2009). To control for the effects of biased gene conversion on this analysis, we filtered out sites that could be affected by removing any AT → GC substitutions genome-wide. The negative correlation between human-primate divergence and functional content did not change after controlling for biased gene conversion (Table 2.2). Though the positive correlation between human-primate divergence and human recombination decreased after this filter (from 0.234 to 0.108), it still remained significant (Table 2.3). Thus, the observed correlations are unlikely to be driven solely by choice of window size or mutational properties based on sequence composition. Because biased gene conversion appears to contribute to some of the correlation between divergence and recombination rate, subsequent analyses of this correlation use the divergence dataset filtered for biased gene conversion.

Effects on putatively neutral divergence between humans and rodents

We next explored the evolutionary forces affecting divergence between more distantly related pairs of species, specifically human-mouse and human-rat. These species were predicted to have diverged approximately 75 million years ago (Mouse Genome Sequencing Consortium et al., 2002) and, as such, current thinking would predict that natural selection would not affect linked neutral sites. Similar to what was seen for the closely related species, functional content is negatively correlated with neutral human-rodent divergence (Spearman's $\rho_{\textit{human-mouse}} = -0.184$, $P < 10^{-16}$, Spearman's $\rho_{\textit{human-rat}} = -0.149$, $P < 10^{-16}$, Figure 2.5A, 2.5B, Table 2.4). This negative correlation persisted when using 50kb windows and also after accounting for the confounding factors of hypermutable CpG sites, GC content, and GC-biased gene conversion (Table 2.4).

Since the broad-scale recombination rate at 100kb appears to have changed over the course of evolution of the species (Jensen-Seaman et al., 2004), we looked for other potential signatures of whether natural selection has affected linked neutral divergence. In particular, we examined the relationship between human-rodent divergence and the strength of background selection across the genome inferred from divergence within primates (McVicker et al., 2009). This strength of background selection is captured by the *B*-value, which represents the degree to which neutral variation at a given position is reduced by selection relative to neutral expectations. While McVicker et al. concluded that divergence between primates was indeed reduced due to background selection, they did not consider human-mouse divergence in their analyses and did not model background selection within the human-dog ancestor (McVicker et al., 2009). As such, there is no *a priori* reason why the *B*-values of McVicker et al. should be related to human-mouse divergence (McVicker et al., 2009).

Nevertheless, we found a positive correlation between human-rodent divergence and the

B-values from McVicker et al. (Spearman's $\rho_{human\text{-}mouse} = 0.445$, $P < 10^{-16}$, Figure 2.5C, Table 2.5, Spearman's $\rho_{human\text{-}rat} = 0.402$, $P < 10^{-16}$, Figure 2.5D, Table 2.5) (McVicker et al., 2009). The positive correlation between human-rodent divergence and *B*-values remained significant even after accounting for the confounding factors of CpG sites, GC content, and GC-biased gene conversion. Similarly, these correlations remained when using 50kb windows (Table 2.5). Taken together, the empirical correlations are consistent with the hypothesis that natural selection has contributed to reducing neutral divergence at linked sites even between species with a long split time such as human and mouse.

Models incorporating background selection in the ancestral population can generate the empirical correlations

To test whether a model including background selection in the ancestral population can explain the empirical observations regarding neutral human-primate divergence and neutral human-rodent divergence, we used a coalescent simulation approach. To a first approximation, the effect of background selection in a sample size of two chromosomes can be accounted for by scaling the ancestral population size with the strength of background selection (Charlesworth, 2012; Charlesworth et al., 1993, 1995; Comeron, 2014; Coop and Ralph, 2012; Corbett-Detig et al., 2015; Hudson and Kaplan, 1995; McVicker et al., 2009). Thus, we modeled the effect of background selection as a reduction in the ancestral population size using the *B*-values estimated in McVicker et al. (McVicker et al., 2009). Briefly, we first used *ms* (Hudson, 2002) to generate genetic variation in the ancestral population where the ancestral population has size N_aB. Then we simulated mutations that accumulated since the split between two species using a Poisson process. The total divergence was the sum of the mutations in the ancestral population and mutations accumulated since the split (see Methods). We modeled mutation rate variation by drawing a mutation rate for each window from a gamma distribution. We chose values for the parameters of the gamma distribution as well as the ancestral population size (N_a) such that the mean and standard deviation of the simulated divergence across the genome and the correlation coefficients between divergence and other functional properties were similar to those seen empirically (Figure 2.6, Table 2.6, Table 2.7, Methods).

We first examined which models could generate the observed correlation between recombination and human-chimp divergence. Here we use the value of Spearman's ρ estimated from the data after filtering out sites that could be affected by biased gene conversion (ρ=0.108). When considering a model without background selection (i.e. B=1 for all windows), the average value of Spearman's ρ between human-chimp divergence and recombination rate was 0.042, and none of the 500 simulation replicates approached the value of Spearman's ρ seen empirically (Figure 2.7A, white histogram). On the other hand, when modeling background selection using the McVicker B-values, the average Spearman's ρ was 0.107 which was comparable to the Spearman's ρ computed from empirical human-chimp divergence with human recombination after accounting for biased gene conversion (Figure 2.7A, gray histogram).

We then tested whether a model incorporating background selection could generate a positive correlation between neutral human-rodent divergence and B-values as observed empirically. We modified our simulation approach to account for the difference in generation time between human and mouse (see Methods). When considering models without background selection (i.e. B=1 for all windows), the average value of Spearman's ρ was 0.012, and none of the 500 simulation replicates approached the value of Spearman's ρ seen empirically (Figure 2.7B, white histogram). However, when modeling background selection using the McVicker B-values, the average Spearman's ρ was 0.446 which was comparable to the Spearman's ρ computed from empirical human-mouse divergence and McVicker's B-values (Figure 2.7B, gray histogram).

In sum, our results suggest that for a given set of parameters, a model with background selection in the ancestral population can generate the correlations observed in the empirical data (i.e. a positive correlation between neutral human-primate divergence and human recombination and a positive correlation between neutral human-rodent divergence and B-values) whereas neutral coalescent models cannot.

Intuition for why background selection is a plausible explanation for the empirical correlations

Current thinking argues that natural selection affecting linked neutral sites is not a plausible explanation for the reduction in neutral divergence between pairs of species with a long split time such as human-mouse or human-rat. Here, we outline a theoretical analysis of a simple two-locus model to gain intuition about how the mutation rate (μ), strength of background selection (B), and ancestral population size (N_a) affect the degree to which background selection can affect divergence (Figure 2.8A).

If background selection has any effect on the variation in neutral divergence across the genome, this can only be due to its effect on divergence in the ancestral population, since deleterious mutations do not affect the fixation rate at linked neutral sites (Birky and Walsh, 1988). Recombination in the ancestral population results in a distribution of coalescent times within each locus, with an average coalescent time of $\bar{t}$. We assumed that the recombination rate within each locus is large enough, such that there is no variation in $\bar{t}$ for a fixed value of B, i.e. $\mathrm{Var}[\bar{t}|\mathrm{B}] \approx 0$. This is a reasonable assumption as long as the window size and recombination rate are not too small. Recombination events cause the sequence to be broken into independent segments, such that for a total ρ > 10 (where ρ denotes the population-scaled recombination rate, $4N_er$) the variance in $\bar{t}$ approaches zero (Wakely, 2008). For an average 100kb window in the human genome (r=10^{-8}/bp, N_e=10,000), ρ is 40 and thus this assumption holds true. Any difference in $\bar{t}$ between loci is then only attributable to differences in background selection: $\mathrm{E}[\bar{t}|B] = 2N_aB$. Further, variation in ancestral (d_a) and total (d_t) divergence results from a Poisson distributed number of mutations added to the genealogy, such that $Var[d_a|B] = E[d_a|B] = 4N_aB\mu L$ and $Var[d_t|B] = E[d_t|B] = E[d_a|B] + 2t_{split}\mu L$ where L is the sequence length of a locus. The law of total variance can be used to compute the variance in total divergence across loci with varying levels of background selection:

$$Var[d_t] = Var_B\big[E[d_t|B]\big] + E_B[Var[d_t|B]]$$

Thus, variance in total divergence can be decomposed into variance due to background selection and variance due to the mutational process. For simplicity, the first locus experiences no background selection ($B_1 = 1$), and the second locus experiences some fixed amount of background selection ($0 \leq B_2 \leq 1$). Under this model, we computed the variance due to background selection as:

$$Var_B\big[E[d_t|B]\big] = ((E[d_t|B = 1] - E[d_t|B = B_2])/2)^2\ .$$

We then computed the variance due to the mutational process as:

$$E_B[Var[d_t|B]] = (Var[d_t|B = 1] + Var[d_t|B = B_2])/2.$$

We assumed an old split time, such that the divergence that accumulated from present time to population split is similar to the human-mouse divergence (40%). Both loci have a sequence length (L) of 100kb. Our theoretical analysis of variance approach shows that with this old split time and assuming a low mutation rate of 1 x 10^{-9}/bp, more than 20% of the variation in the divergence can be explained by background selection in the ancestral population with the following conditions: ancestral population size > 600,000 and B < 0.2 (Figure 2.8A, panel 1, blue, purple, and pink lines). Note that under these conditions, the proportion of divergence that accumulated in the ancestral population can be as low as 0.3% (Figure 2.8B, panel 1). However, the proportion of the variance in divergence that is attributable to the ancestral population is larger than 20% (Figure 2.8C, panel 1), mainly due to background selection leading to differences in $\bar{t}$ between loci. With a larger mutation rate (2 x 10^{-8}/bp), background selection results in a stronger effect on variation in divergence even when ancestral population size is relatively small (>50,000; Figure 2.8A, yellow line). When assuming a moderately large population size of 200,000, and a moderate strength of background selection ($B = 0.75$), then as much as 50% of variance in divergence can be explained by background selection (Figure 2.8C, light green line). Nonetheless, the proportion of divergence that accumulated in the ancestral population in this case is still only 3.4%. Collectively, even for old split times, where the vast majority of divergence accumulated after the population split, with certain assumptions about the ancestral population size, mutation rate, and strength of background selection, the variance in the divergence could be explained by background selection.

Coalescent simulations predict background selection can reduce neutral divergence between species with long split times

Because the theoretical model described above ignores regions of low recombination and only considers one pair of loci at a time, we used coalescent simulations (similar to what we outlined above) to examine whether background selection could generate the positive correlation between estimates of background selection in primates and divergence between distantly related species using more realistic models. Since we were not particularly concerned with any specific species, we simplified these simulations by setting the mutation rate to 2.5 x 10^{-8}/bp.

We found that across all population sizes and split times examined, background selection generated a positive correlation between recombination and divergence as well as a positive correlation between divergence and *B*-values, even for pairs of species that split up to 100*N* generations ago (Figure 2.9, black lines and dashed lines). This correlation remained strong even when the proportion of the divergence due to ancestral polymorphism was small. For example, for a pair of populations with t_{split}=100*N* generations and an ancestral population of size 50,000, only 1.53% of the divergent sites are due to ancestral polymorphism (Figure 2.9, red lines). However, this model predicts a correlation of 0.211 between recombination and divergence and a correlation of 0.377 between recombination and *B*-values. Although ancestral polymorphism only contributes in a small way to the total divergence, the variance in the amount of ancestral polymorphism across the windows accounts for nearly 60% of the variance in divergence across different windows (Figure 2.10, black lines). In general, the correlations decreased as both the split time increased and the size of the ancestral population decreased (Figure 2.10). This behavior is expected as the contribution of the variance in levels of ancestral polymorphism to the variance in divergence decreases with increasing split time and decreasing ancestral population size (Figure 2.10).

Examining the direct effects of natural selection on observed correlations

While we have shown under a variety of models that natural selection can affect putatively neutral divergence and generate the correlations that we observe empirically, other selective scenarios could explain these patterns. An alternative explanation for the empirical correlations reported in Figure 2.3 and Figure 2.5 is that the filtering criteria we used to obtain neutral sites did not effectively remove all non-neutral sites. Therefore, the observed correlations could be due to the direct effects of purifying selection reducing genetic divergence. As sites under purifying selection may be located close to conserved functional elements and could conceivably result in low *B*-values, this is a potentially plausible explanation for our findings. As our current filters removed the 10% of the genome that was most likely under the direct effect of selection based upon the top 10% of GERP scores, we reasoned that additional sites under purifying selection would have elevated GERP scores relative to neutrality.

To test this hypothesis, we repeated our correlation analyses by first obtaining the neutral human-primate divergence and neutral human-rodent divergence using different GERP score cutoffs (i.e. 5% to 25%). When examining human and primate pairs, the correlation between neutral human-primate divergence and functional content decreased as a function of increasing GERP cutoff score (Figure 2.11A). Nevertheless, the negative correlation between neutral human-primate divergence and functional content remained significant even after removing any site whose GERP score fell within the top 25% of the distribution (Spearman's $\rho_{human\text{-}chimp} = -0.189$, $P < 10^{-16}$, Spearman's $\rho_{human\text{-}orang} = -0.122$, $P < 10^{-16}$, Figure 2.12A, Figure 2.12B). On the other hand, the relationship between neutral human-primate divergence and human recombination rate were not affected by varying GERP score cutoffs (Figure 2.11B, Figure 2.12C, Figure 2.12D).

When examining human and rodent pairs, we found that the negative correlation between human-rodent divergence and functional content decreased as a function of increasing GERP score cutoff. Further, the relationship became nonsignificant when filtering any site whose GERP score fell within the top 15th percentile (Figure 2.11C, Figure 2.13A, Figure 2.13B). The positive correlation between neutral human-rodent divergence and McVicker's *B* values decreased as a function of increasing GERP score cutoff, but remained significantly positive even after removing any sites whose GERP score fell within the top 25th percentile (Figure 2.11, Figure

2.13C, Figure 2.13D). Still, this latter pattern indicates that the direct effects of natural selection are unlikely to explain our findings, unless the selected sites are not in the upper 25% of the GERP score distribution.

To test whether background selection could explain these correlations when removing the 25% of the genome with the most conserved GERP scores, we used our coalescent simulation framework. These simulations match the empirical distribution of divergence across the genome (Figure 2.14, Table 2.7) and use the parameters given in Table 2.6. For human-chimp divergence, none of the 500 coalescent simulations resulted in a Spearman's ρ between divergence and human recombination rate as large as observed empirically after filtering sites affected by biased gene conversion (Figure 2.15A, white histogram). On the other hand, simulations including background selection in the ancestral population generated a Spearman's ρ between divergence and human recombination rate similar to what was observed empirically after filtering sites affected by biased gene conversion (Figure 2.15A, gray histogram). Similarly, for human-mouse divergence, while none of the 500 coalescent simulations using the neutral model could generate a Spearman's ρ between divergence and McVicker's *B*-values as large as the empirical correlation, models including background selection in the ancestral population could generate this correlation (Figure 2.15B).

2.4 Discussion

Here we have examined patterns of divergence between pairs of species with various degrees of divergence. We document several signatures that are consistent with the action of natural selection reducing divergence at linked neutral sites. First, for all pairs of species considered, we find that neutral divergence is lowest in regions of the genome with the greatest functional content (Figure 2.3 and Figure 2.5). This pattern may be expected if more selection occurs in regions of the genome with greater functional content. Second, human-primate neutral divergence strongly correlates with human recombination rate and the correlation persists after accounting for hypermutable CpG sites, GC content, and biased gene conversion. Regions of low recombination show lower levels of divergence, which is consistent with selection having a greater effect on linked neutral sites in regions of low recombination. The correlation between human-primate divergence and human recombination is higher in regions with greater overlap with RefSeq transcripts, indicative of a greater reduction in neutral divergence in regions near genes as opposed to far from genes (Figure 2.4). Third, human-rodent neutral divergence strongly correlates with the strength of background selection estimated for primates. These correlations persist after accounting for CpG sites, GC content, and biased gene conversion. Importantly, coalescent simulations including background selection can generate several of these correlations. However, neutral coalescent models without background selection do not.

One interesting observation made was that while most of our correlation analyses were robust to the confounding effect of biased gene conversion, the correlation between human-primate neutral divergence and recombination rate was affected significantly by biased gene conversion. This suggests that while some of the correlation between recombination and divergence can be driven by biased gene conversion, it cannot explain the entire correlation. This result also argues that when testing for a correlation between divergence and recombination, the effect of biased gene conversion should be taken into account.

While we found that models incorporating background selection predict correlations comparable to the empirical data, in principle, several other evolutionary processes may be able to generate these patterns. First, selective sweeps in the ancestral population could reduce divergence just like background selection. Given that we are unlikely to be able to survey patterns of polymorphism in the human-mouse ancestor in more than two lineages, it will be difficult or nearly impossible to distinguish between these two types of selection at linked neutral sites. Thus, one should interpret our use of *B*-values as reflecting a reduction in divergence due to the combined effects of both background selection and selective sweeps, as suggested in McVicker et al. (McVicker et al., 2009).

A second possibility is that the negative correlation between divergence and functional content as well as the positive correlation between divergence and *B*-values could be driven by variation in mutation rate across the genome. Indeed, McVicker et al. attributed a positive correlation between *B*-values and human-dog divergence to the effects of variable mutation rates (McVicker et al., 2009). However, for this mechanism to explain our results, it would require that mutation rates would have to be lower closer to genes and in regions of the genome thought to experience more background selection (i.e. in regions with lower *B*-values). There is some limited evidence of this effect in Arabidopsis where mutation rates are higher in regions of the genome with greater heterozygosity (Yang et al., 2015). However, the extent to which these results apply to mammalian genomes remains unclear. Further, other studies in humans do not support the view that mutation rates are systematically lower in regions of the genome more subjected to selection. Recent estimates of the *de novo* mutation rate have not found any evidence of a reduction close to genes (Francioli et al., 2015). Further, Palamara et al. found that their estimates of the mutation rate do not differ as a function of *B*-values (Palamara et al., 2015). Variation in mutation rate across the genome, while inflating the variance in divergence across the genome, would not be predicted to generate correlations between *B*-values and divergence as well as the correlation between functional content and divergence. Thus, we can rule it out as the sole explanation for the empirical patterns seen in our study.

Further, mutagenic recombination is unlikely to explain the empirical patterns in our study because the correlation between divergence and functional content does not depend on recombination rate. The negative correlation between divergence and functional content remained strong when controlling for variation in recombination rates (Table 2.2) suggesting our results are unlikely to be driven by mutagenic recombination. Nevertheless, our results do not rule out the possibility of mutagenic recombination and this topic certainly warrants further investigation.

Another possibility is that the reduction in neutral divergence near genes and in regions with lower *B*-values could be due to the direct effects of purifying selection removing variation from the population. Current evidence from a variety of comparative genomic studies suggests <10% of the genome is under purifying selection (Cooper et al., 2004; Davydov et al., 2010; Gulko et al., 2015; Meader et al., 2010; Mouse Genome Sequencing Consortium et al., 2002; Pollard et al., 2010; Rands et al., 2014; Schrider and Kern, 2015; Siepel et al., 2005; Ward and Kellis, 2012). We attempted to mitigate the direct effects of purifying selection by employing a conservative set of filters in order to obtain putatively neutral sites. When removing the 10% of the genome that is most conserved, using a variety of conservation metrics, the correlations persisted, suggesting they were not driven by the direct effects of selection. However, when we removed the top 15% of sites with the most conserved GERP score, the correlation between human-rodent divergence and functional content disappeared. This finding suggests that either the GERP scores themselves are affected by background selection, or, instead, that this correlation is driven, in part, by the direct effects of purifying selection. However, in order for direct purifying selection to explain the correlation, either more of the genome (at least 15%) would have to be under selection than suggested by current estimates (Cooper et al., 2004;

Davydov et al., 2010; Gulko et al., 2015; Meader et al., 2010; Mouse Genome Sequencing Consortium et al., 2002; Pollard et al., 2010; Rands et al., 2014; Schrider and Kern, 2015; Siepel et al., 2005; Ward and Kellis, 2012) or many of the sites in the top 15% most conserved GERP scores would have to be neutrally evolving. Additionally, the negative correlation between human-chimp divergence and functional content, the positive correlation between human-chimp divergence and recombination rate, and the positive correlation between human-mouse divergence and *B*-values, remained even after removing the 25% of the genome that is most conserved (Figure 2.11). This implies that even such a large amount of functional sites under selection cannot explain all of our results. Finally, an additional line of evidence suggesting that the putatively neutral sites close to genes are not subjected to the direct effects of purifying selection stems from the fact that they show similar levels of neutral divergence to four-fold degenerate sties (Figure 2.2, Table 2.1). Thus, our putatively neutral noncoding sites have levels of divergence comparable to those seen for sites solely subjected to background selection.

Additionally, our filters rely on functional annotations and conservation to remove functionally important sites directly under the effects of selection. It is formally possible that the direct effects of selection could generate the correlations seen in our study if there are sites under selection that were invisible to the conservation-based filters used in our study. This could occur if there are recently derived, lineage-specific functional elements under selection that cannot be picked up by conservation metrics, or if there are sequences subject to purifying selection in the ancestral population but subsequently became neutral and therefore were not conserved. While we cannot exclude such a scenario, current population genetic evidence provides, at most, limited support for such an explanation (Gulko et al., 2015; Palazzo and Gregory, 2014; Rands et al., 2014; Ward and Kellis, 2012).

One limitation in this study is that we made many assumptions regarding the parameters used in the simulations such as the ancestral population size, generation times, and mutation rates over the last 5-7 million years between human and chimp and 75 million years between human and mouse. There is much uncertainty surrounding all of these parameters (Geraldes et al., 2011; Hardison et al., 2003; Hodgkinson and Eyre-Walker, 2011; Kumar and Subramanian, 2002; Mouse Genome Sequencing Consortium et al., 2002; Smith et al., 2002). Overall, we used a set of parameters in which the simulated divergence dataset from the coalescent simulations matched closely with the mean and standard deviation of the empirical divergence dataset. This allowed us to assess whether a simple neutral model could result in the correlations as large as observed empirically or whether a model with background selection needed to be invoked. We utilized the coalescent simulations as a proof of concept and therefore, the parameters we used in these sets of simulations should not be taken as estimates of the true values. Estimation of these parameters (ancestral population size, mutation rate, split time, etc.) is beyond the scope of this study and certainly warrants further in-depth investigation.

Other studies have argued that background selection will not affect divergence between distantly related species because the genealogy in the ancestral population only comprises a small proportion of the total genealogy between one chromosome from each of the two species (Begun et al., 2007; Birky and Walsh, 1988; Hellmann et al., 2003). This means that ancestral polymorphism will only account for a small proportion of the total divergence between distantly related species. It was thought that the signature of selection reducing the genealogy in the ancestral population would be diluted by the mutations that occurred since the split. As such, there would be no detectable signature of background selection. Our theoretical results and simulations show the proportion of ancestral polymorphism actually is a poor predictor of the correlation between divergence and recombination as well as between divergence and *B*-values. For example, consider a pair of species that split *N* generations ago with an ancestral population size of 25,000. In this model, 40% of the divergence is attributable to *ancestral* polymorphism (Figure 2.9A). Now consider a second pair of species that split 100*N* generations ago where N_a=200,000. Here <5% of the divergence is due to ancestral polymorphism (Figure 2.9D). Previous intuition suggests the effect of background selection would be stronger in the first pair of species because they split more recently and ancestral polymorphism makes a greater contribution to divergence. However, our simulations show the exact opposite pattern (Figure 2.9A, 2.9D). The correlation between *B*-values and divergence is higher in the model with the more ancient split (Spearman's $\rho = 0.610$) than the one with the more recent split (Spearman's $\rho = 0.452$). Similar results are seen for the correlation between recombination rate and divergence. The reason for this discrepancy is that the main driver of these correlations is not the average amount of ancestral polymorphism, but rather the contribution to the variance in divergence due to the variance in ancestral polymorphism. Even when ancestral polymorphism makes only a small contribution to the overall average divergence, a substantial amount of the variance in total divergence across the genome can still be explained by variance in ancestral polymorphism, particularly if the ancestral population size is large. Our theoretical results suggest that the variance in the amount of background selection in different regions of the genome can account for a lot of the variance in total divergence, even for species that split long ago. In sum, our theoretical results and simulations suggest that previous intuition has understated the importance of even small amounts of ancestral polymorphism on the variability of genome-wide patterns of divergence between species.

Our results have important implications for understanding patterns of genetic variation and divergence across genomes. First, our findings add to the growing literature suggesting the importance of background selection at shaping genome-wide patterns of variability across species (Campos et al., 2014; Charlesworth, 2012; Comeron, 2014; Cutter and Choi, 2010; Cutter and Payseur, 2013; Flowers et al., 2012; Halligan et al., 2013; Hernandez et al., 2011; Lohmueller et al., 2011; McVicker et al., 2009; Slotte, 2014; Wilson Sayres et al., 2014; Wright and Andolfatto, 2008). Our new contribution to this literature is demonstrating that natural selection can affect neutral divergence, even between distantly related species. Second, our work suggests that estimators of mutational properties that rely on contrasting patterns of divergence across different parts of the genome that may be differentially affected by background selection may yield biased results. This effect has been studied within primates in greater detail in recent work (Narang and Wilson Sayres, 2015). Third, the fact that we detect evidence of background selection between distantly related species suggests that there is still some information about the distribution of coalescent genealogies across the genome. This distribution of coalescent genealogies can be exploited to obtain more reliable estimates regarding the human-mouse ancestral population size. While several methods exist to estimate ancestral demographic parameters from divergence (Gronau et al., 2011; Rannala and Yang, 2003; Siepel, 2009; Takahata, 1986; Wall, 2003), we suggest that these methods may be applicable for very distantly related species. Our finding that background selection can increase the variance in coalescent times across the genome suggests these methods as well as other statistical methods which seek to infer demographic history from the distribution of coalescent times across the genome, such as the PSMC approach (Li and Durbin, 2011), should account for the increased variance in coalescent times across the genome due to background selection. Not accounting for background selection could result in inferring spurious demographic events to account for the additional variance in coalescent times across the genome as has recently been suggested for positive selection (Schrider et al., 2016). Lastly, our results suggest a need for caution when using patterns of divergence to calibrate neutral mutation rates. Some of the variation in divergence across the genome may be due to varying coalescent times, further accentuated by selection, rather than differing mutation rates (Edwards and Beerli, 2000; Gillespie and Langley, 1979). Future work could explore the extent to which selection at linked neutral sites can explain the discrepancies between different types of estimates of mutation rates (Scally and Durbin, 2012; Ségurel et al., 2014).

2.5 Methods

Data sets

We obtained the pairwise (.axt) alignments between human/chimpanzee (hg18/panTro2), human/orang (hg18/ponAbe2), human/mouse (hg18/mm9), human/rat (hg18/rn4), and human/zebrafish (hg18/danRer15) from the UCSC genome browser (Kent et al., 2002). These alignments are the net of the best human chained alignments for each region of the genome (Kent et al., 2003). For quality control, we excluded sites that (1) were missing in either of the species in the alignment, (2) located at least 10Mbp from the starting or ending position of a centromere, (3) located at least 10Mbp from the ending position of a telomere, (4) not located in repetitive elements.

We obtained the coordinate positions for the exons, RefSeq transcripts, and different phastCons measures calculated from different phylogenetic scopes from the UCSC table browser (Karolchik et al., 2003) with the following specifications:

1. Exons: clade: Mammal, genome: Human, assembly: Mar. 2006 (NCBI36/hg18), group: Genes and Gene predictions, track: UCSC Genes, table: knownGene.

2. RefSeq transcripts: clade: Mammal, genome: Human, assembly: Mar. 2006 (NCBI36/hg18), group: Genes and Gene predictions, track: RefSeq Genes, table: refGene.

3. phastCons Vertebrates: clade: Mammal, genome: Human, assembly: Mar. 2006 (NCBI36/hg18), group: Comparative Genomics, track: Conservation, table: Vertebrate El (phastConsElements44way).

4. phastCons Primates: clade: Mammal, genome: Human, assembly: Mar. 2006 (NCBI36/hg18), group: Comparative Genomics, track: Conservation, table: Primate El (phastConsElements44wayPrimates).

5. phastCons Mammals: clade: Mammal, genome: Human, assembly: Mar. 2006 (NCBI36/hg18), group: Comparative Genomics, track: Conservation, table: Mammal El (phastConsElements44wayPlacental).

GERP scores were downloaded for hg18 from http://mendel.stanford.edu/SidowLab/downloads/gerp/. We used RS scores (range from -11.6 to 5.82) to obtain the conserved sites to remove. Table 2.8 summarizes the cutoffs we used.

For each window, we computed the recombination rate using the high resolution pedigree-based genetic map assembled by deCODE (Kong et al., 2010). The *B*-value for each window was obtained from McVicker et al. (McVicker et al., 2009). Four-fold divergence was calculated by counting the number of between species differences that overlapped four-fold sites, divided by the total number of four-fold sites within each window. Functional annotation was done following Lohmueller et al. (Lohmueller et al., 2011). Briefly, we translated the Consensus Coding Sequence (CCDS) genes from the UCSC Genome Browser into proteins and determined which nucleotide changes did not alter the encoded amino acid. If transcripts overlapped, we retained the longest one.

Correlation analyses

To calculate the divergence between each pair of species, we divided the human genome into 100kb non-overlapping windows. For each window, we computed the total number of sites that passed the filtering criteria which resulted in the total number of neutral sites in each 100kb window. To reduce variation, we only considered windows in which the total number of eligible sites was greater than 10,000 (for analyses using 50kb as window size, we only considered windows in which the total number of eligible sites was greater than 5,000). Then we computed the divergence by tabulating the number of sites that are different between the two species being compared. To account for multiple mutational hits for the distantly related species pairs (human-mouse and human-rat), we applied the Kimura two-parameter model (Kimura, 1980).

To compute Spearman's ρ, we used the *cor* function in R. We used the *pcor* function to calculate partial correlation (Kim, 2015).

Controlling for confounding factors

To filter out possible hypermuteable CpG sites, we excluded sites that were preceded by a C or were followed by a G in hg18 (McVicker et al., 2009). To control for the effects of biased gene conversion, we removed all AT→GC substitutions across the genome.

Coalescent simulations

We modeled background selection as a simple reduction in effective population size in the ancestral population (Charlesworth, 2012; Charlesworth et al., 1993, 1995; Comeron, 2014; Coop and Ralph, 2012; Corbett-Detig et al., 2015; Hudson and Kaplan, 1995; McVicker et al., 2009). This was done by scaling the ancestral population size N_a, by the *B*-values. We used the

B-values from McVicker et al. (McVicker et al., 2009). Each simulation replicate consisted of two parts. The first part modeled genetic variation in the ancestral population, and included the effects of background selection. For each window i, we simulated an ancestral recombination graph (ARG) with a population-scaled recombination rate $4N_aB_ir_i$, where N_a is the ancestral population size, B_i is the strength of background selection affecting window i, and r_i is the recombination rate for window i. Mutations were added to the genealogy assuming a population-scaled mutation rate $\theta=4N_aB_i\mu_{a,i}L_i$, where $\mu_{a,i}$ is the ancestral per-base pair mutation rate for window i and L_i is the number of successfully aligned neutral bases in window i. Simulations were done using the program *ms* (Hudson, 2002). Note, we included recombination in the ancestral population because it affects the variance in coalescent times across windows and this variance in coalescent times will in turn affect the variance in levels of divergence, which will ultimately affect the strength of the correlation between divergence and recombination. Thus, we aimed to capture this variance as accurately as possible. This part of the simulation generated the amount of divergence due to ancestral polymorphism, which we call d_a.

We then added the mutations that arose since (i.e. more recently than) the split. The divergence from the present time to the split time follows a Poisson distribution, where the rate parameter equals the expected divergence between two populations. For each window of the genome, d_s was simulated using the *rpois* function in R. Finally, the total divergence within a window is the sum of divergence generated in the ancestral population (d_a) and the divergence generated since the two species split (d_s).

For human chimp divergence (Figure 2.7A, Figure 2.15A**)**, $d_s = 2t_{split}\mu L$ where d_s is the expected divergence from the present time to the split time in the divergence model, t_{split} is the split time, μ is the mutation rate, and L is the length of each sequence. When computing both d_a

and d_s for human-chimp divergence in Figure 2.7A, we drew μ from a gamma distribution with shape = 16.82 and scale 1.7 X 10^{-10} (Table 2.6). In Figure 2.15A, we drew μ from a gamma distribution with shape = 15.68 and scale 1.8 X 10^{-10} (Table 2.6). These parameters were chosen to match the observed mean and standard deviation of the distribution of human-chimp divergence (after removing all AT to GC differences as such changes could be due to biased gene conversion) as well as the observed correlation coefficient between divergence and recombination rate (Figure 2.6A, 2.6B and Figure 2.14A, 2.14B). The split times and ancestral population sizes are roughly comparable to previous estimates from genetic data (Langergraber et al., 2012; Prado-Martinez et al., 2013; Siepel, 2009; Wall, 2003).

Due to the differences in generation times and mutation rates between the human and mouse lineages, we modified our approach for these simulations (Figure 2.7B, Figure 2.15B, Figure 2.16). First, here d_s =(t_{mouse} μ_{mouse}+ $t_{human}\mu_{human}$)L, where t_{mouse} is the number of generations on the lineage leading to the mouse from t_{split} till the present day, t_{human} is the number of generations on the lineage leading to human experienced from t_{split} till the present day, μ_{mouse} is the mutation rate along the mouse lineage, and μ_{human} is the mutation rate along the human lineage. There is much uncertainty surrounding these parameters. However, the following values are broadly consistent with what has been reported previously and match the observed mean and standard deviation of human-mouse divergence (Figure 2.6C, 2.6D, Figure 2.14C, 2.14D**,** and Table 2.7). First, we assumed t_{split} = 75 million years ago. We then assumed mice have 1 generation per year, giving t_{mouse} = 75 x 10^6 generations. We assumed humans have 25 years per generation, making t_{human} = 3 x 10^6. We then set μ_{mouse} = 3.8 x 10^{-9} per generation and μ_{human} =3.75 x 10^{-8} per generation (Figure 2.16). These estimates are broadly consistent with previous reports and allow for approximately twice as much divergence on the mouse lineage as compared to the human lineage (Mouse Genome Sequencing Consortium et al., 2002).

For the simulations in Figure 2.7B, we assumed that μ_a was equal to 2 x 10^{-8} per generation, which is the average of μ_{human} and μ_{mouse}. We accounted for variation in mutation rates across different regions of the genome by drawing μ_a from a gamma distribution (Voight et al., 2005). We kept the ratio of μ_a to μ_{mouse} constant across all windows of the genome. For example, μ_a / μ_{mouse} = 5.26. Then if $\mu_{a,\mathrm{i}}$ is the rate for the i^{th} region drawn from the gamma distribution, we set $\mu_{mouse,i}$ equal to $\mu_{a,\mathrm{i}}$ / 5.26. A similar procedure was used to find $\mu_{human,i}$. Note that for the simulations in Figure 2.15B, we used the average mutation rate of 2.7 X 10^{-8}, but we kept the ratio of μ_a to μ_{mouse} and the ratio of μ_a to μ_{mouse} to be the same as the simulations in Figure 2.7B. Increasing the variance in the mutation rate across regions increased the variance in divergence across windows of the genome and decreased the correlation between divergence and the *B*-values. We then examined different values of N_a and parameters of the gamma distribution that matched the observed mean and standard deviation of the distribution of human-mouse divergence as well as the observed correlation coefficient between divergence and *B*-values. The ancestral population size, shape, and scale parameters of the gamma distribution used for the simulations in Figure 2.7B and Figure 2.15B are reported in Table 2.6. The simulated human-mouse divergence using these parameters matched closely with the empirical human-mouse divergence (Figure 2.6C, 2.6D, Figure 2.14C, 2.14D, and Table 2.7).

2.6 Figures

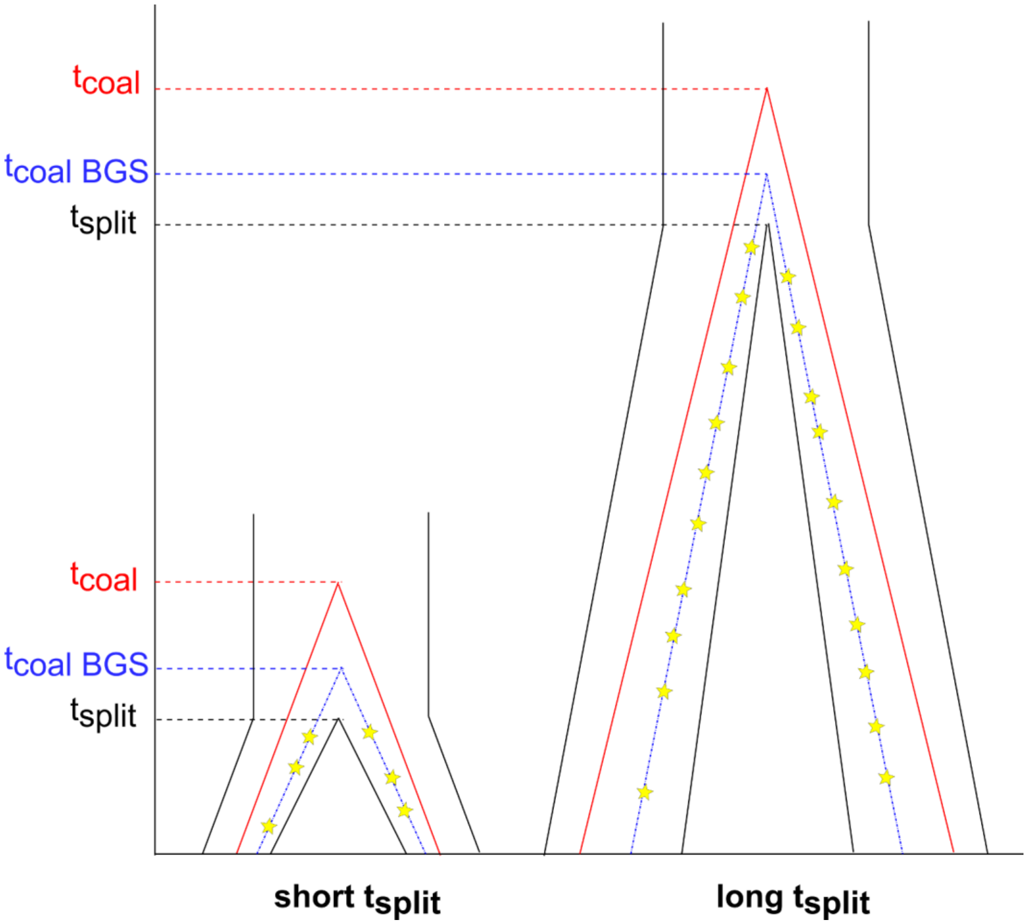


Figure 2.1. Models of how genealogies are affected by selection at linked neutral sites. The genealogies on the left represent species with a short split time such as human and chimpanzee. The genealogies on the right represent species with a long split time such as human and mouse. Red lines represent two lineages and their coalescent time. Blue lines represent two lineages and their coalescent time when there is selection at linked neutral sites in the ancestral population (abbreviated BGS). Yellow stars denote mutations accumulating on each of the two lineages after they split. Note that with the longer split time, the proportion of the genealogy attributed to the ancestral population decreases.

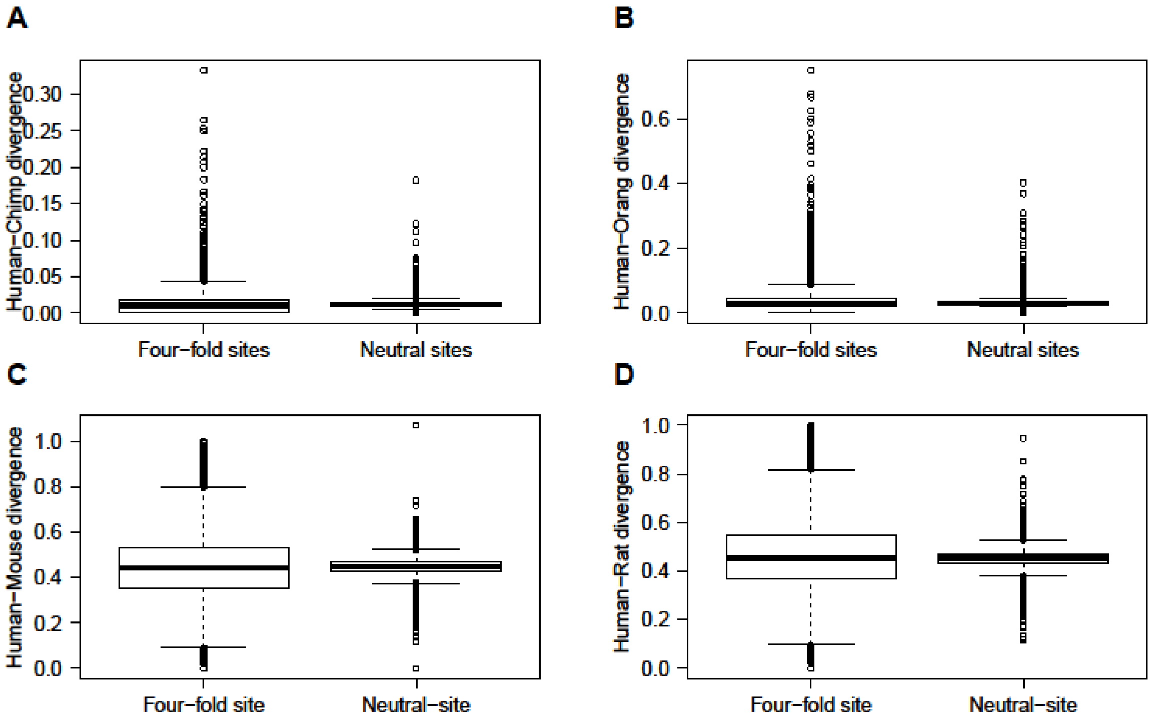


Figure 2.2. Four-fold degenerate sites show similar levels of divergence as our putatively neutral noncoding sites. Each point represents the divergence within a 100kb window. (A) Human-chimpanzee, (B) Human-orangutan, (C) Human-mouse, and (D) Human-rat.

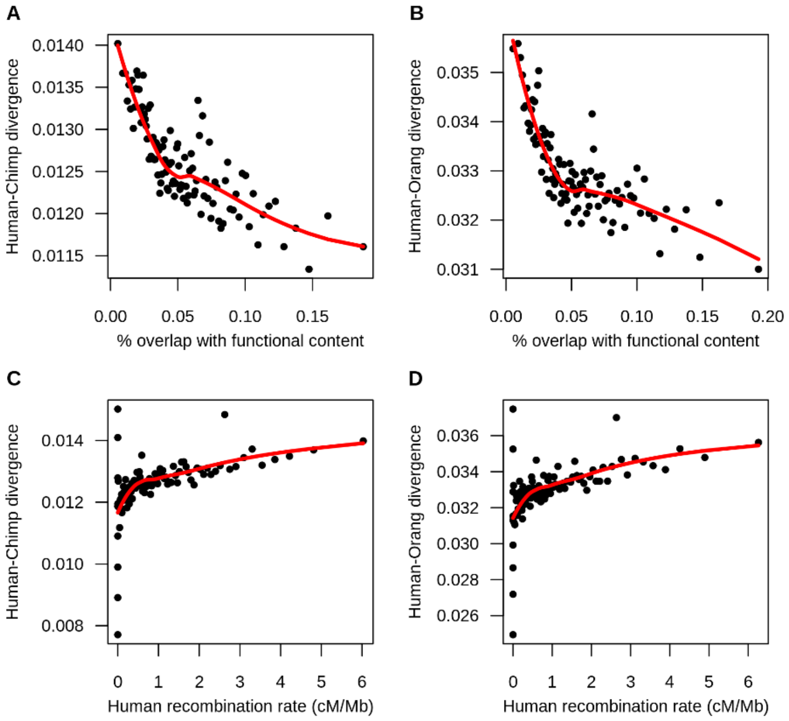


Figure 2.3. Human-primate divergence is reduced at putatively neutral sites near selected sites. (A) Neutral human-chimp divergence is negatively correlated with functional content. (B) Neutral human-orang divergence is negatively correlation with functional content. (C) Neutral human-chimp divergence is positively correlated with human recombination rate. (D) Neutral human-orang divergence is positively correlated with human recombination rate. Each point represents the mean divergence and functional content (A and B) or recombination rate (C and D) in 1% of the 100kb windows binned by functional content or recombination rate. Red lines indicate the loess curves fit to divergence and functional content (A and B) and divergence and recombination rate (C and D). The high variance of divergence at regions of low recombination rate is expected since the variance of divergence is inversely proportional to the recombination rate. Note that the last bin containing less than 1% of the windows was omitted from the plot. While the graph presents binned data, the correlations reported in the text are from the unbinned data.

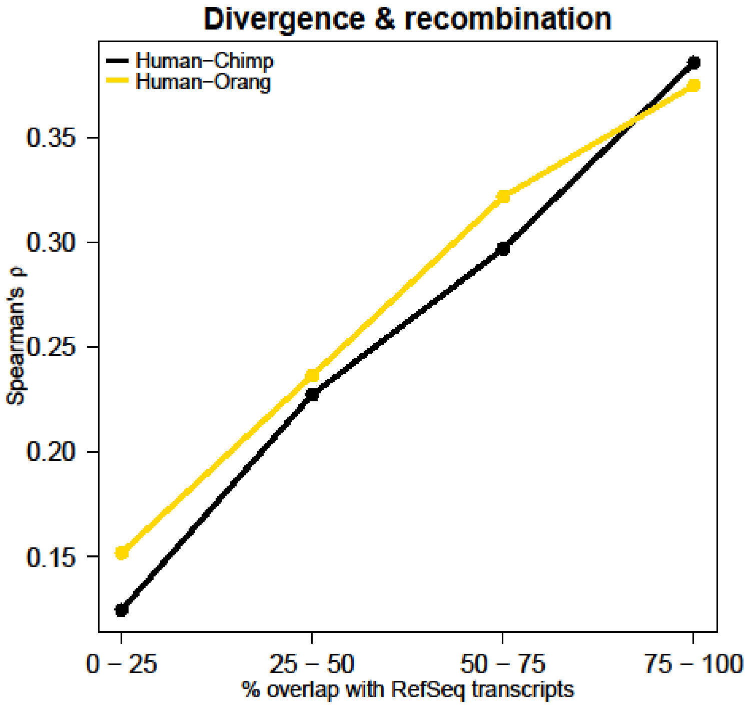


Figure 2.4. The correlation between human recombination rate and neutral divergence is stronger near genes. Correlation (Spearman's ρ) between neutral divergence and human recombination as a function of the amount of overlap with a RefSeq transcript. Black line denotes the correlations between human-chimpanzee neutral divergence and human recombination rate. Yellow line denotes the correlations between human-orangutan neutral divergence and human recombination rate.

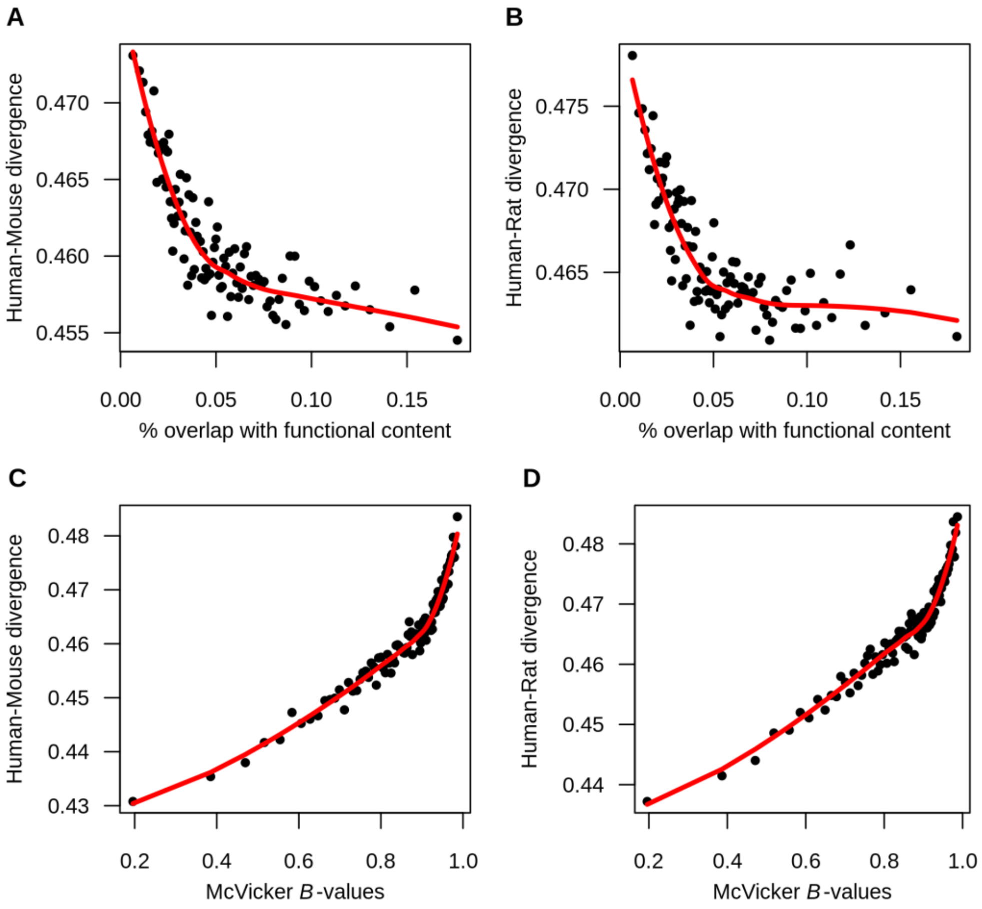


Figure 2.5. Human-rodent divergence is reduced at putatively neutral sites near selected sites. (A) Neutral human-mouse divergence is negatively correlated with functional content. (B) Neutral human-rat divergence is negatively correlated with functional content. (C) Neutral human-mouse divergence is positively correlated with McVicker's *B*-values. (D) Neutral human-rat divergence is positively correlated with McVicker's *B*-values. Each point represents the mean divergence and functional content (A and B) or *B*-values (C and D) in 1% of the 100kb windows binned by functional content or *B*-values. Red lines indicate the loess curves fit to divergence and functional content (A and B) and divergence and *B*-values (C and D). Note that the last bin containing less than 1% of the windows was omitted from the plot. While the graph presents binned data, the correlations reported in the text are from the unbinned data.

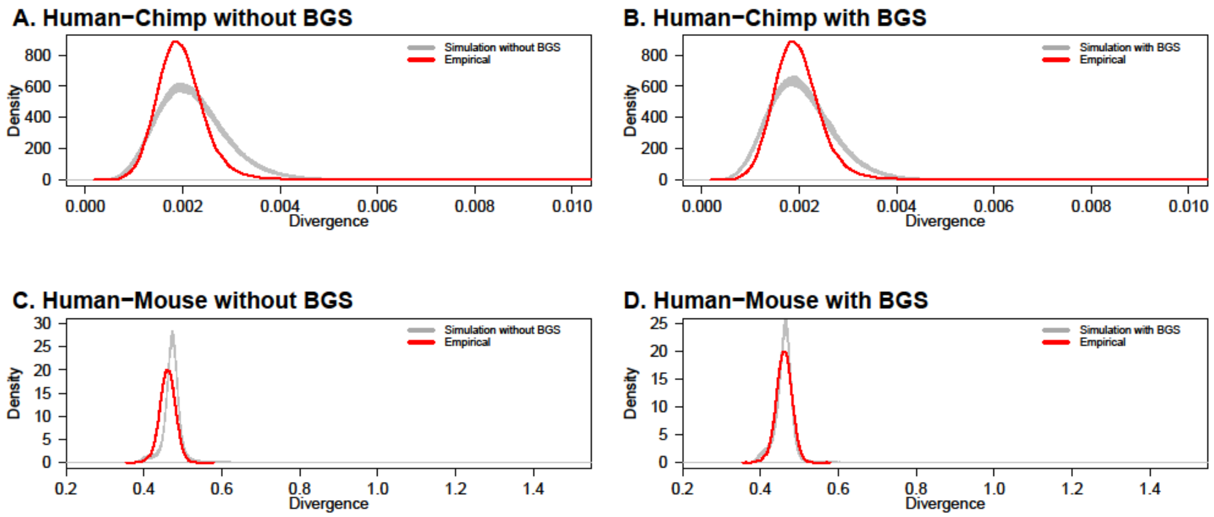


Figure 2.6. Observed and modeled genome-wide distributions of human-chimp divergence and human-mouse divergence in 100 kb windows. Gray lines denote 500 simulated genome-wide distributions of divergence. Red line denotes the observed distribution of neutral divergence. Note, the distribution of simulated divergence is comparable to that from empirical data. (A) Simulated human-chimp divergence without the effects of background selection (BGS). (B) Simulated human-chimp divergence with the effects of background selection. (C) Simulated human-mouse divergence without the effects of background selection. (D) Simulated human-mouse divergence with the effects of background selection. We filtered all AT→GC changes between the human and chimp sequences as they could be affected by biased gene conversion. Thus, the distribution of human-chimp divergence shown here is lower than the overall divergence.

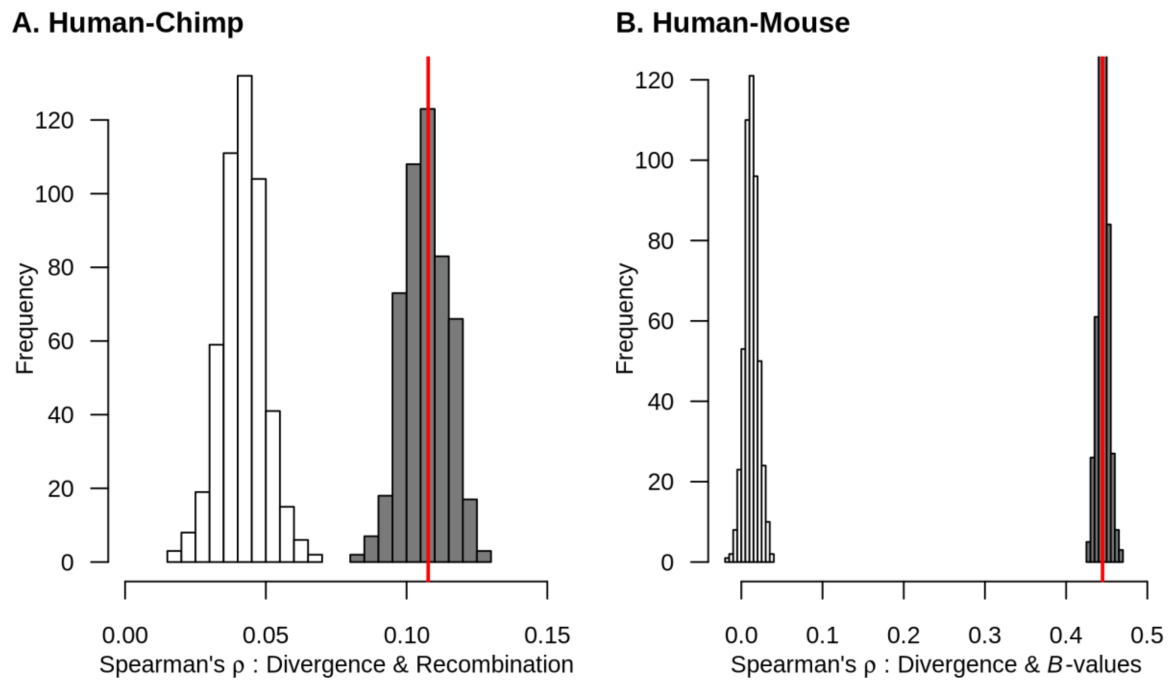


Figure 2.7. Models incorporating background selection can generate patterns of neutral divergence that recapitulate the empirical correlations. (A) Models of background selection predict a positive correlation between neutral human-chimp divergence and human recombination. Because our model does not include biased gene conversion, the empirical correlation was calculated omitting AT to GC sequence differences. (B) Models of background selection predict a positive correlation between neutral human-mouse divergence and McVicker’s *B*-values. White histogram denotes 500 simulations not including background selection. Gray histogram denotes 500 simulations incorporating background selection (see text). Red line represents the correlation computed from empirical data. Thus, plausible levels of background selection can match the observed correlations while neutral simulations cannot.

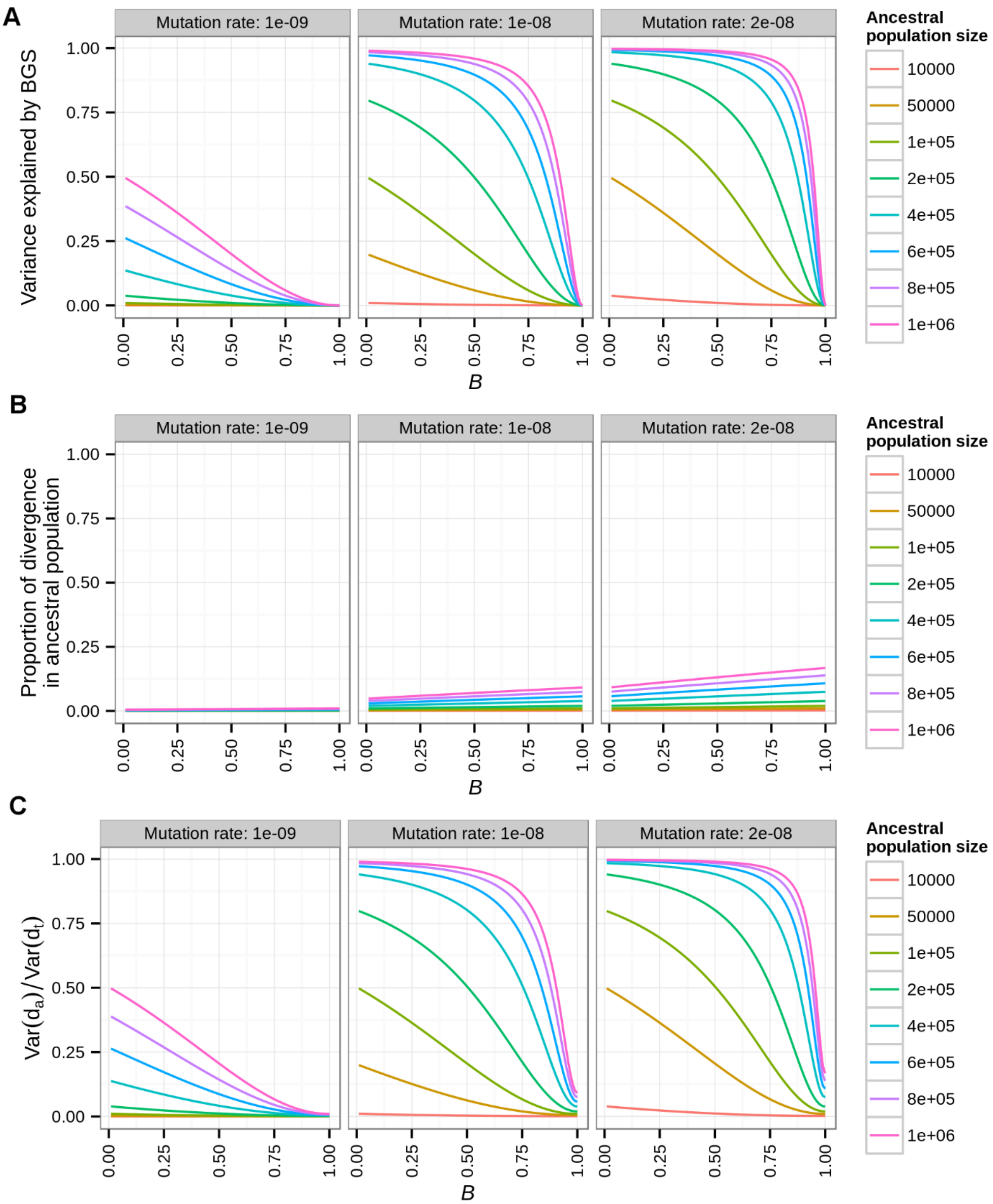


Figure 2.8. A two-locus model for the effect of background selection on divergence. (A) The variance in divergence between two loci explained by background selection (BGS) as a function of the strength of background selection at the second locus (B_2). (B) The expected proportion of

divergence due to polymorphism in the ancestral population as a function of B_2. (C) The variance in divergence between the two loci explained by polymorphism in the ancestral population as a function of B_2. Different columns denote different mutation rates. Colored lines denote different ancestral population sizes (N_a). Note that the variance in divergence attributable to background selection is greater than the expected proportion of divergence contributed by ancestral polymorphism.

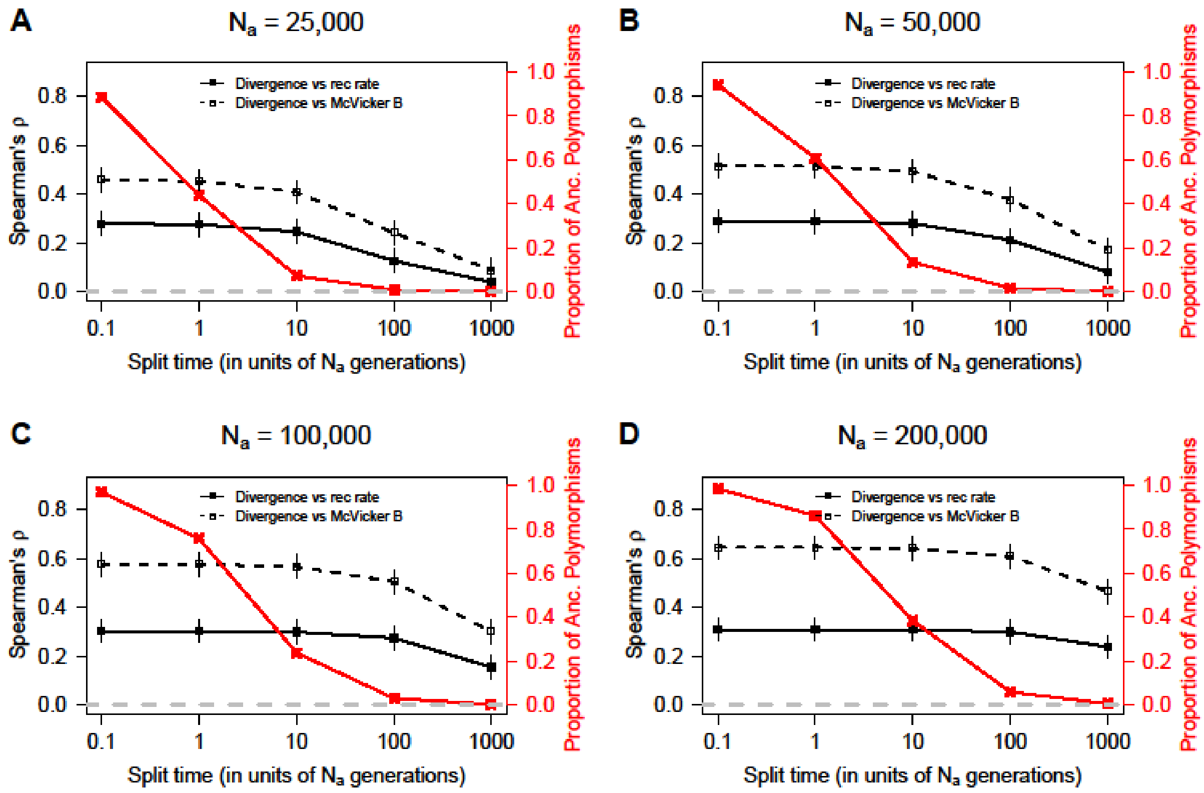


Figure 2.9. Background selection is predicted to affect neutral divergence across a range of split times and ancestral population sizes. Solid line shows the expected correlation coefficients (Spearman's ρ) between neutral divergence and recombination rate as a function of split time. Dashed line shows the expected Spearman's ρ between neutral divergence and McVicker's *B*-values as a function of split time. Red lines denote the proportion of the divergence due to polymorphism that arose in the ancestral population. Error bars denote ± one standard error of the mean. Panels A-D denote different ancestral population sizes (N_a). Note that the correlations are greater than 0 for a range of split times and ancestral population sizes, even when the proportion of divergence due to ancestral polymorphism is low.

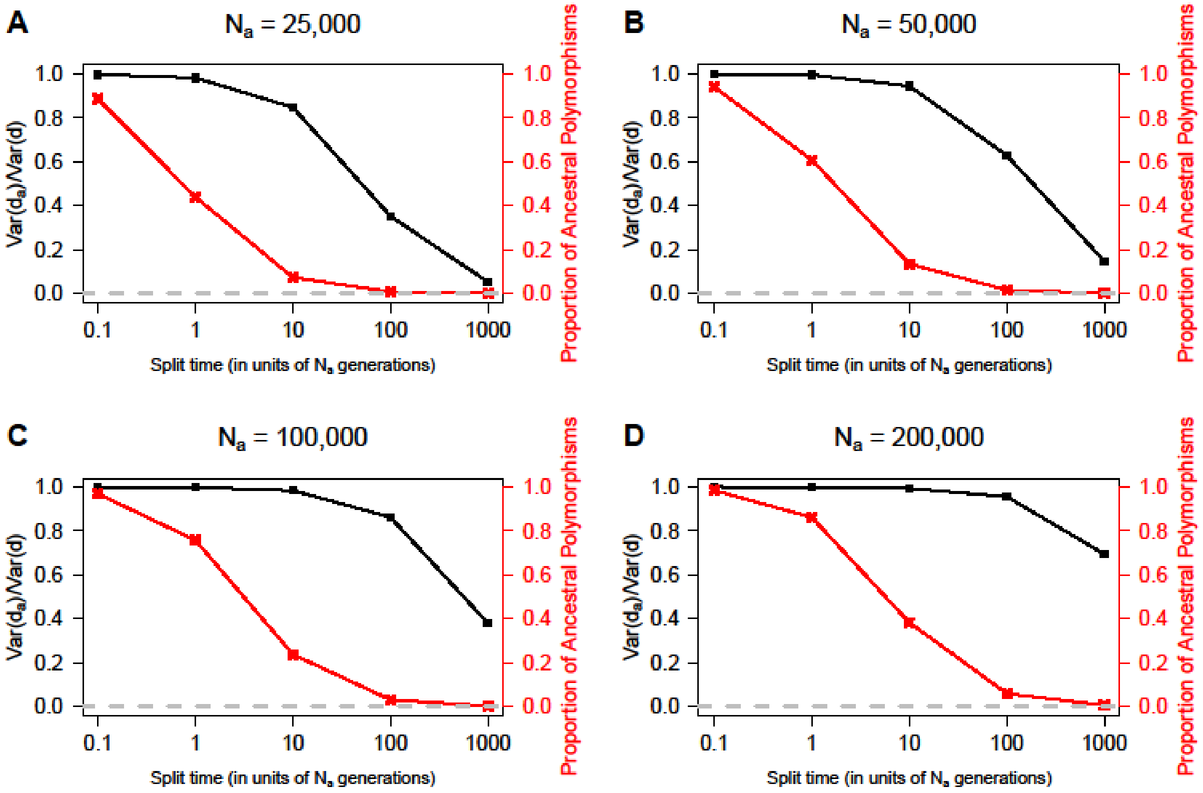


Figure 2.10. Variance of the total divergence attributable to the variance in levels of ancestral polymorphism. Black lines show the ratio of the variance of divergence in the ancestral population to the variance of the total divergence as a function of split time. Red lines denote the proportion of the divergence due to polymorphism that arose in the ancestral population. Panels A-D denote different ancestral population sizes (N_a).

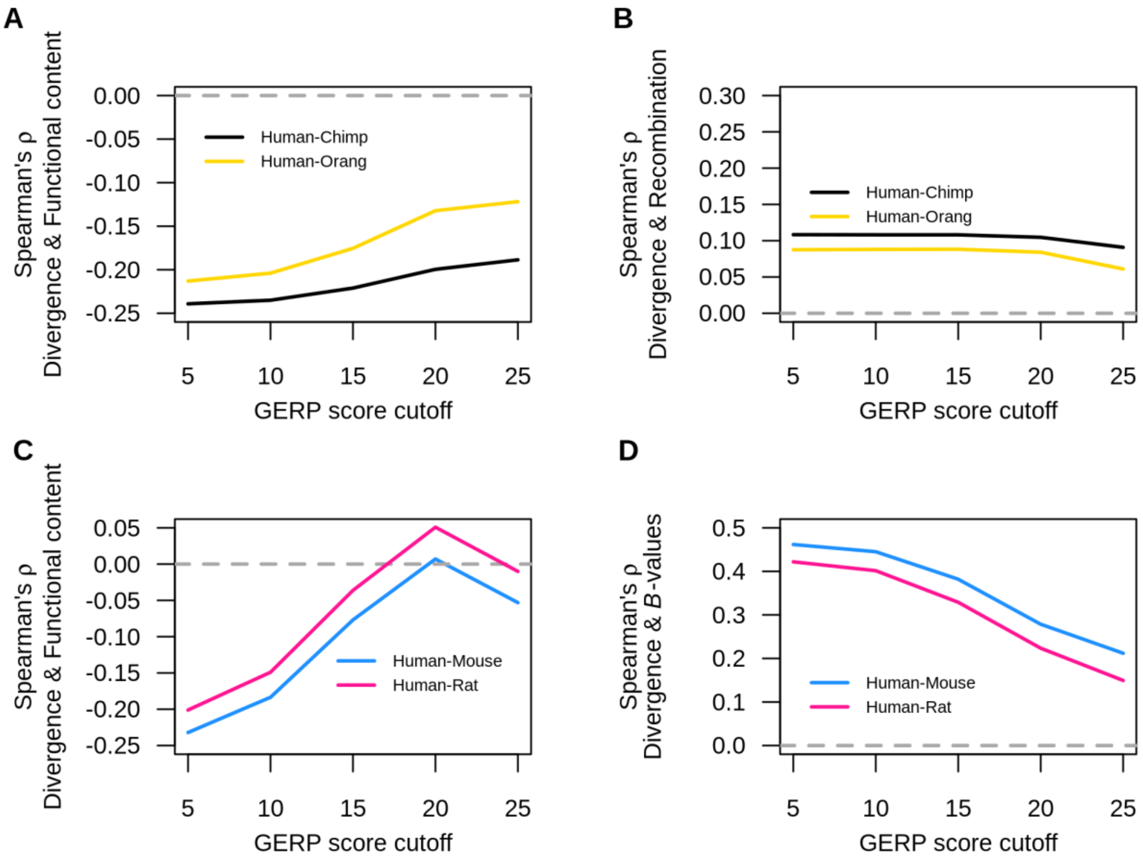


Figure 2.11. Relationship between divergence and functional content, human recombination, and McVicker's *B*-values as a function of GERP score cutoff. (A) Human-primate divergence versus functional content. (B) Human-primate divergence versus human recombination rate. (C) Human-rodent divergence versus functional content. (D) Human-rodent divergence versus McVicker's *B*-values.

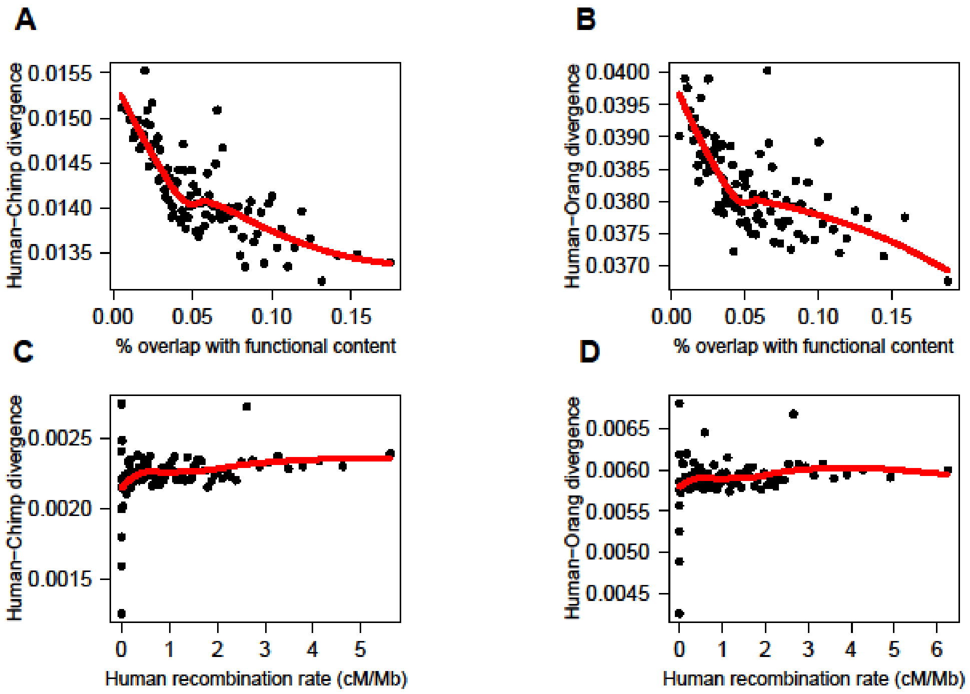


Figure 2.12. Correlations between human-primate divergence and genomic features persist when filtering the 25% of the genome with the highest GERP scores. (A) Neutral human-chimp divergence shows a negative correlation with functional content. (B) Neutral human-orang divergence shows a negative correlation with functional content. (C) Neutral human-chimp divergence shows a positive correlation with human recombination rate. (D) Neutral human-orang divergence shows a positive correlation with human recombination rate. Each point represents the mean divergence and functional content (A and B) or recombination rate (C and D) in 1% of the 100kb windows binned by functional content or recombination rate. Red lines indicate the loess curves fit to divergence and functional content (A and B) and divergence and recombination rate (C and D). Note that the last bin containing less than 1% of the windows was omitted from the plot. While the graph presents binned data, the correlations reported in the text are from the unbinned data.

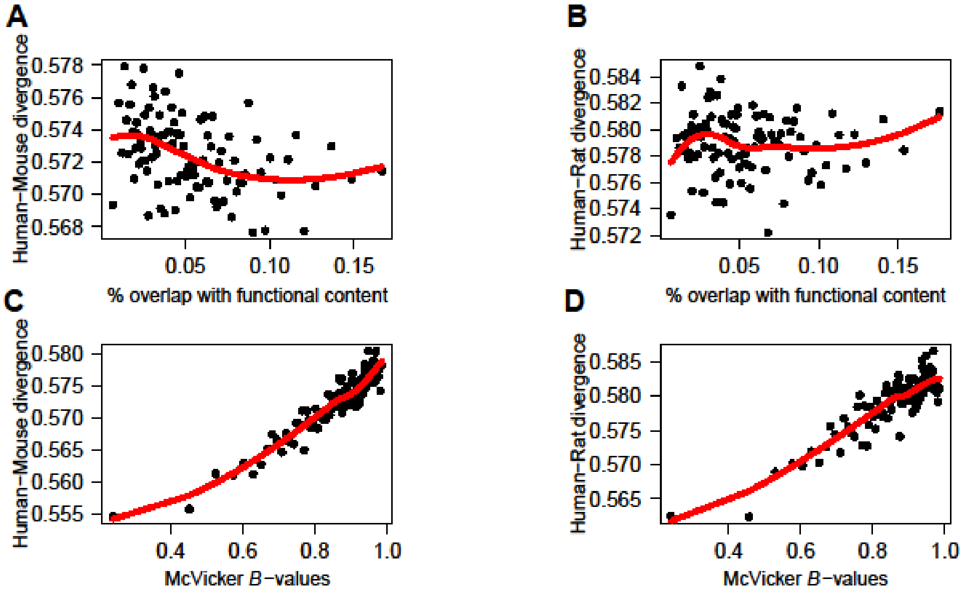


Figure 2.13. Correlations between human-rodent divergence and genomic features change when filtering the 25% of the genome with the highest GERP scores. (A) Neutral human-mouse divergence no longer correlates with functional content. (B) Neutral human-rat divergence does not correlate with functional content. (C) Neutral human-mouse divergence shows a positive correlation with McVicker's *B*-values. (D) Neutral human-rat divergence shows a positive correlation with McVicker's *B*-values. Each point represents the mean divergence and functional content (A and B) or *B*-values (C and D) in 1% of the 100kb windows binned by functional content or *B*-values. Red lines indicate the loess curves fit to divergence and functional content (A and B) and divergence and *B*-values (C and D). Note that the last bin containing less than 1% of the windows was omitted from the plot. While the graph presents binned data, the correlations reported in the text are from the unbinned data.

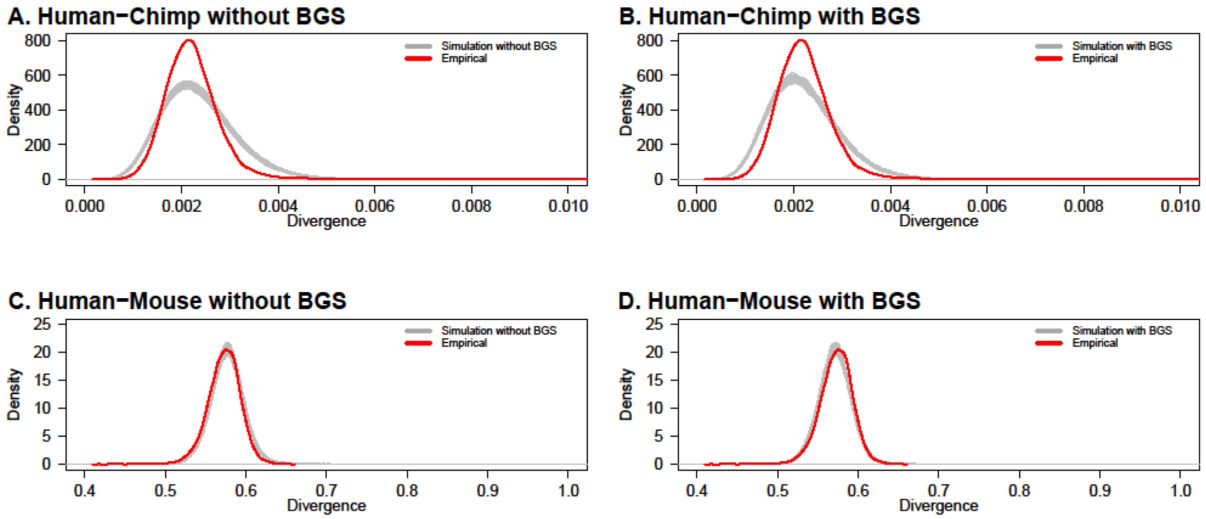


Figure 2.14. Observed and modeled genome-wide distributions of human-chimp divergence and human-mouse divergence in 100kb windows when filtering sites whose GERP scores fall into the top 25% of the genome-wide distribution. Gray lines denote 500 simulated genome-wide distributions of divergence. Red line denotes the observed distribution of neutral divergence. Note, the distribution of simulated divergence is comparable to that from empirical data. (A) Simulated human-chimp divergence without the effects of background selection (BGS). (B) Simulated human-chimp divergence with the effects of background selection. (C) Simulated human-mouse divergence without the effects of background selection. (D) Simulated human-mouse divergence with the effects of background selection. We filtered all AT→GC changes between the human and chimp sequences as they could be affected by biased gene conversion. Thus, the distribution of human-chimp divergence shown here is lower than the overall divergence.

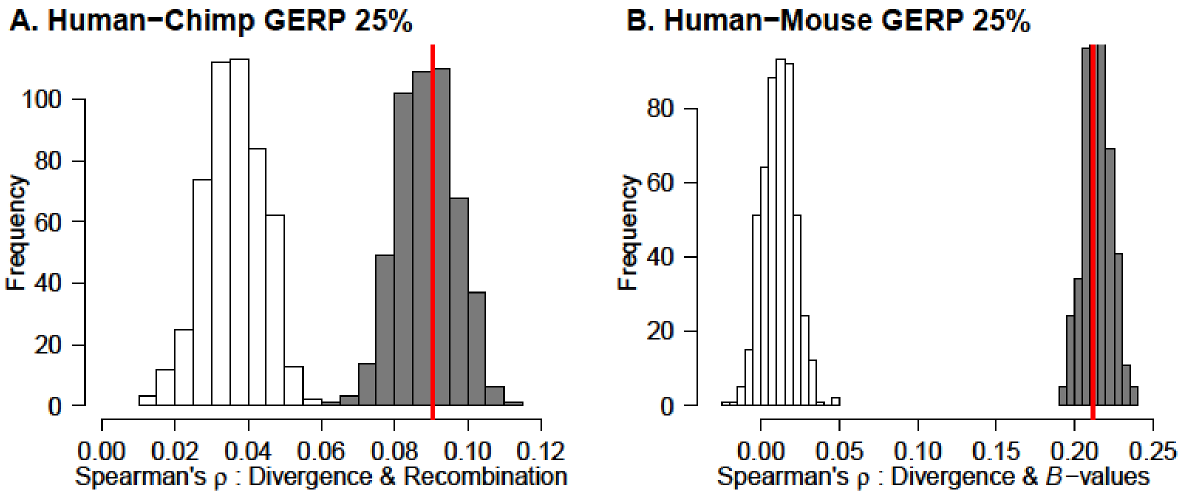


Figure 2.15. Models incorporating background selection can recapitulate the empirical correlations after removing sites with GERP scores falling in the top 25% of the genome-wide distribution. (A) Models of background selection predict a positive correlation between neutral human-chimp divergence and human recombination rate. Because our model does not include biased gene conversion, the empirical correlation was calculated omitting AT to GC sequence differences. (B) Models of background selection predict a positive correlation between neutral human-mouse divergence and McVicker's *B*-values. The white histogram denotes 500 simulations without including background selection. The gray histogram denotes 500 simulations incorporating background selection. Red lines represent the correlations computed from the empirical data. Thus, plausible levels of background selection can match the observed correlations when using the most stringent filtering criteria while neutral simulations cannot.

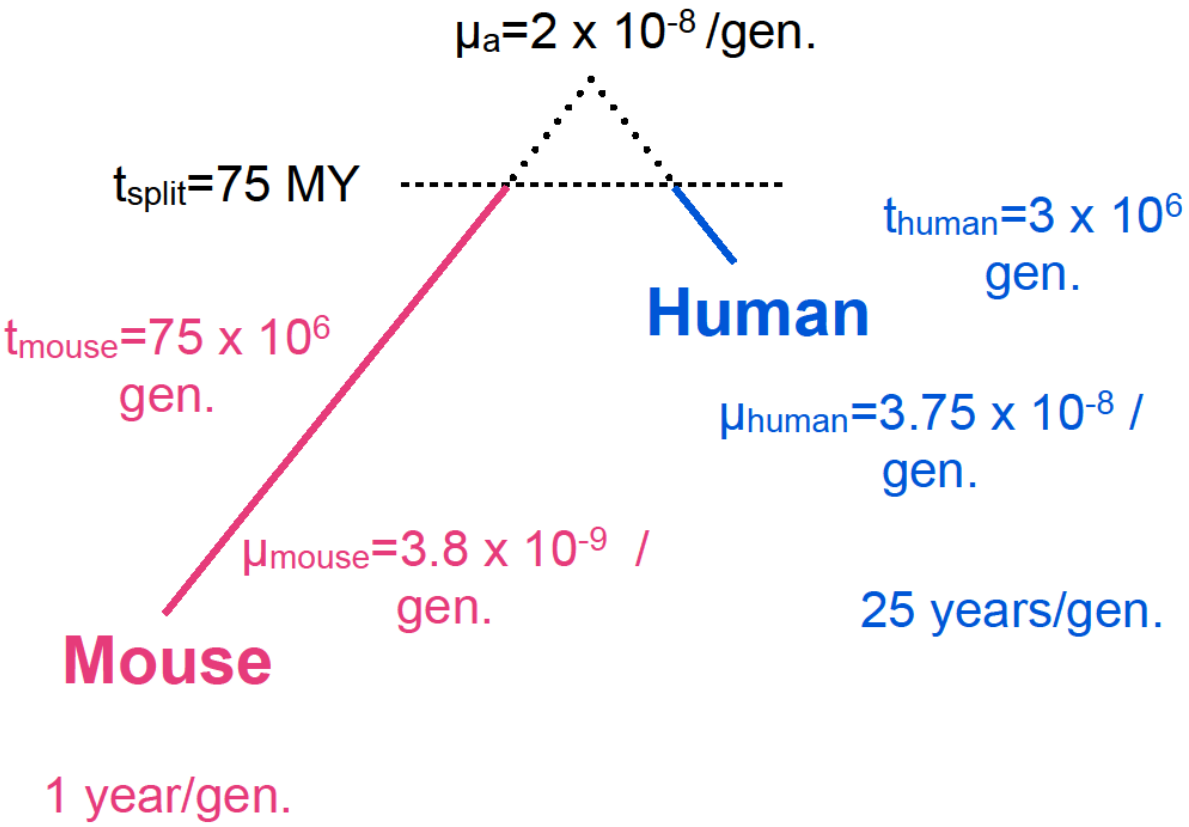

μ_a=2 x 10^{-8} /gen.
t_{split}=75 MY
t_{human}=3 x 10^6 gen.
Human
t_{mouse}=75 x 10^6 gen.
μ_{human}=3.75 x 10^{-8} / gen.
μ_{mouse}=3.8 x 10^{-9} / gen.
25 years/gen.
Mouse
1 year/gen.

Figure 2.16. Human-mouse mutational parameters used for the simulations.

2.7 Tables

Table 2.1. Summary statistics of divergence at four-fold sites and at putatively neutral regions. Mean of divergence and standard deviation of divergence were computed over 100kb windows at four-fold sites degenerate sites and at putatively neutral sites. Windows where the total number of eligible sites is equal to 0 were excluded from both sets. Further, for human-mouse and human-rat, we corrected for multiple mutations with Kimura 2-parameter models. Note that for human-mouse and human-rat, we also excluded windows with unrealistic divergence after the Kimura 2-parameter correction (i.e. windows where the divergence was greater than 1 were removed).

Species pair		Mean of divergence	Standard deviation of divergence
Human-chimp	Four-fold sites	0.013	0.016
	Putatively neutral sites	0.012	0.005
Human-orang	Four-fold sites	0.037	0.037
	Putatively neutral sites	0.033	0.012
Human-mouse	Four-fold sites	0.444	0.152
	Putatively neutral sites	0.447	0.038
Human-rat	Four-fold sites	0.456	0.156
	Putatively neutral sites	0.452	0.039

Table 2.2. Correlation coefficients of human-primate divergence and functional content.

Species pair	Spearman's ρ overall	Spearman's ρ post CpG filtering	Partial correlation controlling for GC content	Spearman's ρ post gBGC filtering	Partial correlation controlling for recombination	Spearman's ρ when using 50kb-windows
Human-chimp	-0.235**	-0.252**	-0.291**	-0.243**	-0.269**	-0.186**
Human-orang	-0.204**	-0.232**	-0.289**	-0.255**	-0.241**	-0.184**

**p-value < 2.2e-16

Table 2.3. Correlation coefficients of human-primate divergence and recombination rate.

Species pair	Spearman's ρ overall	Spearman's ρ post CpG filtering	Partial correlation controlling for GC content	Spearman's ρ post gBGC filtering	Spearman's ρ when using 50kb-windows
Human-chimp	0.234**	0.207**	0.242**	0.108**	0.182**
Human-orang	0.249**	0.221**	0.240**	0.088**	0.209**

**p-value < 2.2e-16

Table 2.4. Correlation coefficients of human-rodent divergence and functional content.

Species pair	Spearman's ρ overall	Spearman's ρ post CpG filtering	Partial correlation controlling for GC content	Partial correlation controlling for recombination	Spearman's ρ post gBGC filtering	Spearman's ρ when using 50kb-windows
Human-mouse	-0.184**	-0.202**	-0.243**	-0.194**	-0.339**	-0.135**
Human-rat	-0.149**	-0.169**	-0.206**	-0.156**	-0.337**	-0.103**

**p-value < 2.2e-16

Table 2.5. Correlation coefficients of human-rodent divergence and McVicker's *B*-values.

Species pair	Spearman's ρ overall	Spearman's ρ post CpG filtering	Partial correlation controlling for GC content	Spearman's ρ post gBGC filtering	Spearman's ρ when using 50kb-windows
Human-mouse	0.445**	0.450**	0.456**	0.419**	0.403**
Human-rat	0.402**	0.405**	0.413**	0.404**	0.366**

**p-value < 2.2e-16

Table 2.6. Summary of parameters used for the coalescent simulations. Note that shape and scale here refer to the shape and scale parameters used for the gamma distribution.

Species pair	Shape	Scale	N_a
Human-chimp GERP 10	$1.1X10^{8}$	$9.3X10^{-17}$	240000
Human-chimp GERP 25	$1.5X10^{9}$	$8.2X10^{-18}$	235000
Human-mouse GERP 10	$4.0X10^{3}$	$5.0X10^{-12}$	940000
Human-mouse GERP 25	$1.6X10^{3}$	$1.7X10^{-11}$	375000

Table 2.7. Comparison of the mean and standard deviation of the empirical and simulated divergence. Note that empirical mean and empirical standard deviation refer to the average neutral divergence calculated over 100kb windows.

Species pair	Empirical mean	Average of the mean of simulated divergence from 500 simulations		Empirical standard deviation	Average of the standard deviation of simulated divergence from 500 simulations	
		Without BGS	With BGS		Without BGS	With BGS
Human-chimp GERP 10	0.013	0.014	0.012	0.003	0.004	0.004
Human-chimp GERP 25	0.014	0.016	0.014	0.003	0.005	0.004
Human-mouse GERP 10	0.461	0.473	0.461	0.020	0.027	0.024
Human-mouse GERP 25	0.572	0.577	0.571	0.021	0.022	0.021

Table 2.8. GERP RS score cutoff. Note that we also removed sites whose RS score is equal to 0.

Proportion of the genome to be removed	Remove sites whose RS score is
5%	> 4.949
10%	> 4.078
15%	> 3.207
20%	> 2.336
25%	> 1.465

CHAPTER 3

Detecting mutagenic recombination using genome-wide divergence data

3.1 Abstract

Mutation is the ultimate source of genetic variation that is acted on by evolutionary processes and gives rise to disease. Currently, the reasons why mutation rate varies across the genome are not well understood. One possibility is that recombination could be mutagenic and as such, variation in the recombination rate contributes to variation in the mutation rate. In humans, a positive correlation between human-chimp neutral divergence and recombination rate has lent support to this hypothesis. However, this correlation could also be driven by the effects of natural selection, specifically background selection. To date, the degree to which recombination is mutagenic in the human genome remains unknown. To address this question, we developed a likelihood-based framework to test the extent to which background selection and/or the coupling between recombination and mutation contributes to the variation in human-chimp neutral divergence. Our method is built upon the idea that background selection primarily affects divergence in regions of low recombination while mutagenic recombination affects regions of high recombination. We model background selection by scaling the effective population size by the B-values from McVicker et al. (2009). We also estimate the degree of coupling between mutation and recombination. Application of this method to human-chimp divergence reveals that a model including background selection and mutation-recombination coupling fits the data substantially better than a model only including background selection. However, we observed that biased gene conversion can account for the effects seen from mutagenic recombination. When biased gene conversion is controlled for, a model including just background selection can

model the empirical data. Our work contributes to the growing literature on the importance of biased gene conversion and how it can confound many findings if not accounted for.

3.2 Introduction

Mutation is the ultimate source of genetic variation that is acted on by evolutionary processes and gives rise to disease. The rate at which mutation arises (i.e. mutation rate) is not constant but varies across the genome (Hodgkinson and Eyre-Walker, 2011). While many genomic features such as replication time, female recombination rate, or GC content correlate with mutation rate, there have been very few investigations into what causes mutation rate to vary across the genome (Eyre-Walker and Eyre-Walker, 2014; Hodgkinson and Eyre-Walker, 2011). In humans, there is some evidence that mutagenic recombination can explain why mutation rate varies across the genome (Hellmann et al., 2003). For example, since differences in DNA between human and chimpanzee (i.e. divergence) at noncoding regions are often used as a proxy for mutation rate, a positive correlation between human-chimp neutral divergence and recombination has been used as evidence for mutagenic recombination (Hellmann et al., 2003). However, this positive correlation could also be due to the effects of natural selection at linked neutral sites, specifically background selection (McVicker et al., 2009; Phung et al., 2016). It is unclear to what extent background selection and mutagenic recombination contributes to the positive correlation between neutral divergence and recombination.

In recent years, with the advance of next generation sequencing, trios (father, mother, and child) are sequenced and the rate of *de novo* mutations are estimated. Francioli et al. sequenced 250 Dutch trios and observed a positive correlation between the rate of *de novo* mutations and recombination rate, supporting mutagenic recombination (Francioli et al., 2015). However, Palamara et al. used the same data and found no correlation after correcting for biased gene conversion (Palamara et al., 2015). Therefore, it is still a topic of debate as to whether recombination is mutagenic.

Here, we wanted to investigate the correlation between human-chimp neutral divergence and recombination to assess whether recombination is mutagenic. We accounted for factors such as background selection and biased gene conversion. As such, we developed a likelihood-based approach based on a population genetic model. We applied our method on neutral human-chimp divergence data and found that a model including both background selection and mutagenic recombination best fits the empirical data. However, most of the effect of mutagenic recombination could be explained by biased gene conversion. Our results add to the growing literature about the important role of biased gene conversion and that biased gene conversion could lead to misinterpretation if it is not accounted for. Further, our method could be modified to be applied to study whether recombination is mutagenic in other species of interest.

3.3 Methods

Description of the data

The neutral human-chimp divergence dataset used in this studied was obtained from Phung et al. (Phung et al., 2016). To investigate the effect of biased gene conversion, we removed all AT → GC substitutions across the genome as in Phung et al. (Phung et al., 2016).

Overview of Methods

We aimed to model human-chimp neutral divergence across the genome by incorporating the effect of natural selection affecting linked neutral sites, specifically background selection, and the potential effect of mutagenic recombination. The intuition of our model is that background selection primarily affects neutral divergence at regions of low recombination whereas mutagenic recombination primarily affects neutral divergence at regions of high recombination. We divided the genome into non-overlapping 100kb windows and assumed independence between windows. For each window i, we computed the probability of observing the number of divergent sites in the i^{th} window given a set of parameters. We aimed to find the maximum likelihood estimates of the parameters that maximize the probability. We used population genetic theory to model the number of divergent sites in the i^{th} window, which is modelled as a Poisson process with the rate equal to the product of mutation rate μ and the genealogy underlying time to the most common ancestor (TMRCA) of two lineages (Wakely, 2008). The TMRCA is divided into two parts: before (i.e. more ancient) and after the split. We estimated the TMRCA before the split by simulating under a coalescent framework using *ms* (Hudson, 2002). To reduce Monte Carlo variance, we generated and averaged across 1,000 genealogies. The genealogy is affected by background selection. We modelled background selection in each window i as a reduction in the ancestral effective population size, $N_{ai} = N_a B_i$, where N_a is the ancestral effective population size and B_i is the effect of background selection. We captured the effect of background selection using the B-values from McVicker et al. as in Phung et al. (McVicker et al., 2009; Phung et al., 2016). To obtain the total TMRCA, T_{total}, we added the TMRCA before the split after converted it into units of generations to the $tsplit$ (see Methods). We modeled the effect of mutagenic recombination by a linear relationship between mutation rate and recombination rate, $\mu_i = \mu + \varphi r_i$, where μ_i is the mutation rate of window i, μ is the neutral mutation rate, r_i is the recombination rate, and φ is the coupling parameter between mutation rate and recombination rate.

For each set of parameters, we computed the log-likelihood of the observed divergence d_i given $T_{total_{i},j,}$, μ, and φ (see Methods). We averaged the log-likelihood over 1000 replicates to reduce Monte-Carlo variance: $L_i = \frac{1}{1000}\sum_{j=1}^{1000} Poisson\left((\mu + \varphi r_i)T_{total_{i},j} \middle| T_{total_{i},j}\right) \Pr(T_{total_{i},j})$.

Since we assumed independence between windows, we summed up the average log-likelihood for each window i for all n windows: $L = \sum_{i=1}^{n} L_i$. This is the genome-wide log-likelihood for each set of parameter values. We then picked the set of parameter values with the highest log-likelihood, which is the maximum likelihood estimates.

Simulating genealogies

We used the coalescent simulator *ms* to simulate the genealogy in each window i for a sample size of two chromosomes with population-scale recombination rate $\rho_i = 4N_aB_ir_i$, where N_a is the ancestral effective population size, B_i is the effect of background selection and r_i is the recombination rate for each window i (Hudson, 2002). For computational efficiency, we generated a look-up table where for each value of ρ, we generated and stored 1000 genealogies. Since Phung et al. (2016) used $N_a = 70000$ in the human-chimp coalescent for this divergence dataset, we computed ρ by using the same value for N_a to obtain an estimate of ρ (Figure 3.1). Since the highest value of ρ is around 1500, we generated the genealogies for a set of ρ ranging uniformly from 0 to 2000. For example, the command in *ms* used to simulate the genealogies for a ρ value of 10 is:

```
msdir/ms 2 100000 -r 10 100000 -L -seeds 1 2 3 > rho_10_genealogies.txt
```

Then, for each set of parameters, we calculated ρ and looked up the values for the 1000 genealogies. This strategy is computational efficient because we avoided having to re-simulate the genealogies for the same value of ρ.

Inference procedure

Since the number of parameters tested is vast, we narrowed down the parameter space by only considering the sets of parameters that when used to simulate a divergence dataset, would result in the mean divergence as comparable to the empirical divergence. Specifically, for each set of parameters, we computed the expected mean divergence, which is equal to $\frac{1}{n}\sum_{i=1}^{n}(\mu + \varphi r_i)2(2N_a B_i + tsplit)L$ where μ is the mutation rate, φ is the coupling parameter between mutation rate and recombination rate, r_i is the recombination rate in window i, N_a is the ancestral population size, B_i is the strength of background selection in window i, $tsplit$ is the split time between two species, and L is the length of the window. We kept the sets of parameters where the expected mean divergence is within 1% of the empirical divergence. The script used for this step can be found at https://github.com/tnphung/MutRec/blob/master/generate_constrain_grid.R.

Then, for each set of parameters that are kept after the initial filtering based on mean divergence, we computed population-scaled recombination rate, ρ, for each window, which is equal to $4N_a B_i rL$. The genealogies are obtained from the look-up table for each value of ρ as described above. Each genealogy, T_j from *ms* simulation is the genealogy in the ancestral population, before the split between two species and is in units of 2N. We then converted T_j to be in units of generation: $2T_j N_a B_i$. Then, $2T_j N_a B_i + tsplit$ is the genealogy that include both before and after the split between two species for one branch. Therefore, $2(2T_j N_a B_i + tsplit)$ is the total genealogy underlying a sample size of two chromosomes.

We then computed the Poisson probability of observing the number of divergent sites in that window given the set of parameters which is equal to $e^{-\lambda}\frac{\lambda^k}{k!}$, where the rate, λ is the product of the mutation rate μ and the genealogy underlying two samples. We accounted for mutagenic recombination by a linear relationship between μ and r_i: $\mu_i = \mu + \varphi r_i$. Then, the Poisson rate is:

$$\lambda_{ij} = (\mu + \varphi r_i)2\big(2T_j N B_i + tsplit\big)L$$

In sum, for each window i, we computed the Poisson probability of observing the number of divergent sites in window i given a set of parameters for one simulation replicate j. We accounted for Monte Carlo variance in the simulation by taking the average probability from 1000 replicates. We then calculated the log-likelihood by taking the logarithm of the probability. Since we assumed independence between windows, we summed up the likelihood from each window, which is then the likelihood of observing the number of divergent sites across the genome given a set of parameters. The scripts for the inference procedure can be found at: https://github.com/tnphung/MutRec/tree/master/infer_mut_rec.

Models tested

We applied our likelihood framework to test whether a model including only background selection, a model including only mutagenic recombination, or a model including both effects can best recapitulate the genome-wide human-chimp neutral divergence. To test the model with just background selection, we set φ to be 0 and inferred for the mutation rate μ and the ancestral population size N_a. We used the McVicker's B-values to capture the effect of background selection (McVicker et al., 2009). To test the model with just mutagenic recombination, we set B to be 1 to indicate that there is no background selection effect. We then inferred for μ, N_a, and the coupling parameter between recombination rate and mutation rate φ. Finally, to test whether the model including both background selection and mutagenic recombination can recapitulate the data, we inferred for μ, N_a, and φ, and used the McVicker's B-values to capture background selection. In all models tested, we set the split time between human and chimpanzee to be 200000 generations. Table 3.1 summarized the parameters being inferred in each model.

Simulations to evaluate performance

We utilized coalescent simulation to simulate divergence datasets under different models as implemented in Phung et al. (2016). We first simulated divergence in the ancestral population using *ms* in each window using the population-scaled recombination rate $4N_aB_ir_i$ and the population-scaled mutation rate $4N_aB_i\mu$. We then added the divergent sites that arose since the split and accumulated following a Poisson distribution where the rate is equal to $2tsplit(\mu + \varphi r_i)L$. We simulated three divergence datasets. To simulate a divergence dataset that was affected by background selection by itself, we set φ to be equal to 0. To simulate a divergence dataset that was affected by mutagenic recombination by itself, we set background selection in each window B_i to be 1. Table 3.1 listed the values we used to simulate each divergence dataset. The scripts used to simulate divergence data can be found at: https://github.com/tnphung/MutRec/tree/master/simulate_divergence_dataset.

3.4 Results

Simulations to validate inference method

To test whether our inference method can distinguish between different models, we simulated three test datasets. In the first dataset, only background selection is included. In the second dataset, only mutagenic recombination is included. Similarly, in the third dataset, both background selection and mutagenic recombination are included (see Methods). For each test dataset, we applied the inference procedure to test which model (i.e. background selection, mutagenic recombination, or both) would yield the maximum likelihood and visually fit best to the test data.

When the test dataset was generated with only background selection, we observed that both the model with just background selection and the model with both background selection and mutagenic recombination have similar likelihood (Table 3.2). Both models visually fit the test data (Figure 3.2A, 3.2C). It is expected that the model with both background selection and mutagenic recombination would fit the data as well as the model with just background selection. This is because the model with just background selection is nested within the model incorporating both effects. Given two models with similar likelihood, we would select the simpler model with fewer parameters, which is the model with just background selection. On the other hand, while a model with just mutagenic recombination fits the pattern of neutral divergence at regions of high recombination (>2cM/Mb) reasonably well, the fit at regions of low recombination (<2cM/Mb) is poor (Figure 3.2B). This observation validates our intuition that background selection primarily affects patterns of divergence at regions of low recombination while mutagenic recombination primarily affects patterns of divergence at regions of high recombination.

When the test dataset was simulated using a model with just mutagenic recombination, the model with mutagenic recombination resulted in the highest likelihood and visually fit well to the test data (Table 3.3 and Figure 3.3). The inferred parameters are also close to the true values (Table 3.3). On the other hand, both the model with background selection and the model incorporating both effects resulted in a poor fit to the test data (Figure 3.3A and 3.3C).

When the test dataset was simulated with both background selection and mutagenic recombination, we observed that the model including both effects yielded the highest likelihood and fitted well to the test data (Table 3.4 and Figure 3.4C). The inferred parameters are also close to the true values (Table 3.4). However, both the model with just background selection or the model with just mutagenic recombination yielded poor fits to the test data (Figure 3.4A and 3.4B). Interestingly, a model with just background selection fitted the test data reasonably well at regions of low recombination (<2cM/Mb) but failed to recapitulate the pattern of divergence at regions of high recombination (>2cM/Mb) (Figure 3.4A). On the contrary, a model with just mutagenic recombination fitted the test data well at regions of high recombination (>2cM/Mb), but the divergence at regions of low recombination is higher than the test data (Figure 3.4B). These results suggest that our method can distinguish between these different models and can infer parameters close to the true values.

Estimating the degree of coupling between mutation and recombination in human

To understand the extent of background selection and recombination-mutation coupling (if any) can generate the observed correlation pattern between human-chimp neutral divergence and human recombination rate, we applied our inference procedure to human-chimp neutral divergence data and tested three models: a model with background selection, a model with mutagenic recombination, and a model with both background selection and mutagenic recombination (see Methods). We observed that a model with just background selection can recapitulate the pattern of neutral divergence well at regions of low recombination (<2cM/Mb), but could not predict the higher divergence at regions of high recombination (>2cM/Mb) (Figure 3.5A). On the contrary, the model with just mutagenic recombination fail to recapitulate the reduction in neutral divergence, presumably due to linked selection, at regions of low recombination (Figure 3.5B). However, the model including both effects yielded the highest likelihood and the best visual fit to the data (Table 3.5 and Figure 3.5C), suggesting that both background selection and mutagenic recombination contribute to the pattern of human-chimp neutral divergence. We found that the coupling parameter between mutation rate and recombination rate is around 0.03 (Table 3.5), indicating that a one unit increase in recombination rate (in units of morgan per base pair) results in a 3 percent increase in the mutation rate (in units of base pair per generation).

Biased gene conversion is an evolutionary force that has been shown to contribute to some of the correlation pattern between neutral divergence and recombination. We accounted for this confounding factor by removing divergent sites that could have been affected by biased gene conversion as in Phung et al. (2016). Specifically, we removed any AT → GC substitutions. We repeated the inference procedure using this dataset where sites that could be affected by biased gene conversion are removed. We found that a model including just background selection resulted in the highest likelihood and is sufficient to recapitulate the empirical divergence (Table 3.6, Figure 3.6). When removing the potential effect of biased gene conversion, we did not observe the increase in neutral divergence at regions of high recombination. Therefore, a model of just background selection can fit the data reasonably well (Figure 3.6A).

3.5 Discussion

We developed a likelihood-based method to model neutral divergence across the genome by incorporating the effect of natural selection on linked neutral sites and the effect of mutagenic recombination. To the best of our knowledge, this is the first method that explicitly tests whether background selection or mutagenic recombination or both effects can best recapitulate the empirical correlation observed between neutral divergence and recombination. Previous studies have attributed the positive correlation between human-chimp neutral divergence and recombination to the effect of linked selection or to mutagenic recombination (Hellmann et al., 2003; McVicker et al., 2009; Phung et al., 2016). Even though previous work has shown that a model including background selection can generate a correlation comparable to the empirical data, these studies have not explicitly rule out the effect of mutagenic recombination. Here we applied our method to show that a model including both effects best recapitulates the empirical human-chimp neutral divergence across the genome. However, we presented evidence that the signal of mutagenic recombination could be explained by biased gene conversion. Our finding supports previous research that used sequencing data of trios to study whether recombination is mutagenic in humans (Palamara et al., 2015). Our study used genome-wide divergence data; however, we came to a similar conclusion to Francioli et al. (2015) and Palamara et al (2015). Specifically, the evidence supporting mutagenic recombination is seemingly due to biased gene conversion. Our research contributes to the growing literature on the importance of biased gene conversion in affecting divergence. For example, Smith et al. investigated variation in mutation rate using *de novo* mutations and observed that the correlation between divergence and mutation rate could be due to biased gene conversion (Smith et al., 2018)

One limitation of the method developed in this study is that we assume that the relationship between neutral divergence and mutation is linear. Other relationship such as quadratic could be possible, especially if this method is to be applied to other species where the relationship between divergence and recombination could be more complex. Further, we applied this method to human-chimp neutral divergence to detect whether recombination is mutagenic in humans. The advantage of applying to human dataset is that some parameters needed in the coalescent simulation are known such as the split time between human and chimpanzee and the amount of background selection. Even though we anticipate that this method could be applied to other species, some of these parameters may not be readily available for species besides humans and additional steps are required to optimize those values first.

3.6 Figures

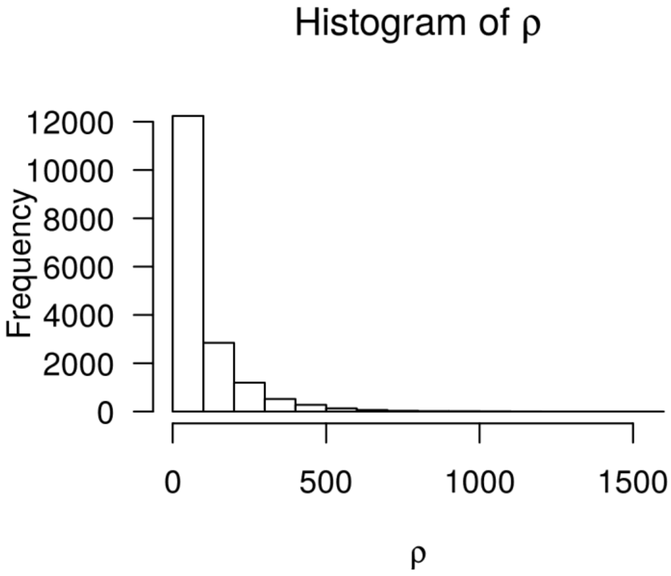


Figure 3.1. Distribution of empirical population-scale recombination rate, $\boldsymbol{\rho}$ across 100kb-windows. Here, the population-scale recombination rate for each 100kb window is: $\rho_i = 4N_aB_ir_i$, where N_a is the ancestral effective population size, B_i is the effect of background selection and r_i is the recombination rate for each window i. Following Phung et al., we set $N_a = 70000$ (Phung et al., 2016). We used the B-values from McVicker et al. (McVicker et al., 2009) and recombination rate from the high resolution pedigree-based genetic map assembled by deCODE (Kong et al., 2010).

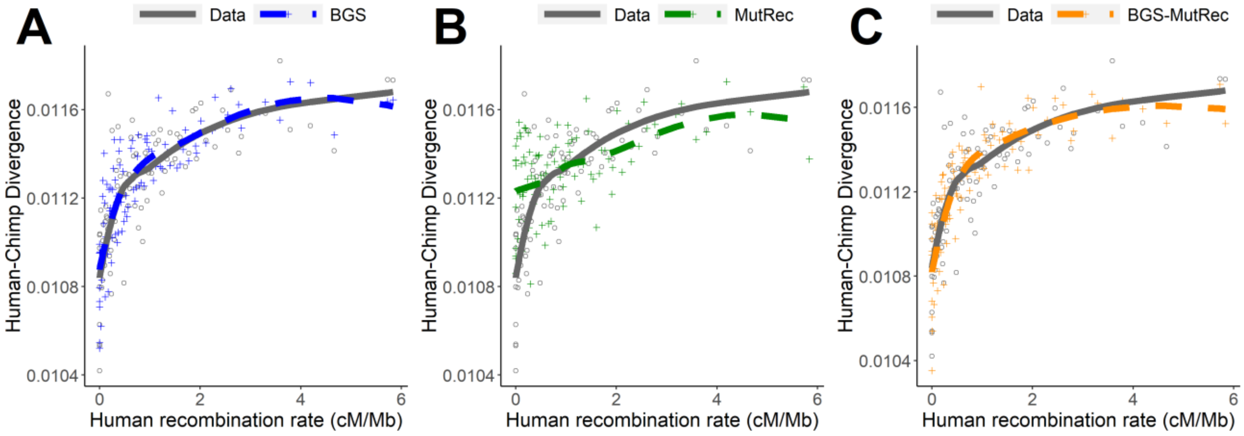


Figure 3.2. Test dataset was generated with background selection. Each point represents the mean divergence and recombination rate in 1% of the 100kb windows binned by recombination rate. Each line represents the loess curve fit to the neutral divergence and recombination rate. Gray points and gray line represent the test dataset. (A) Blue points and blue line were generated using the MLEs from a model with just background selection. (B) Green points and green line were generated using the MLEs from a model with just mutagenic recombination. (C) Orange points and orange line were generated using the MLEs from a model with both background selection and mutagenic recombination. Both (A) and (C) showed an excellent fit to the test data.

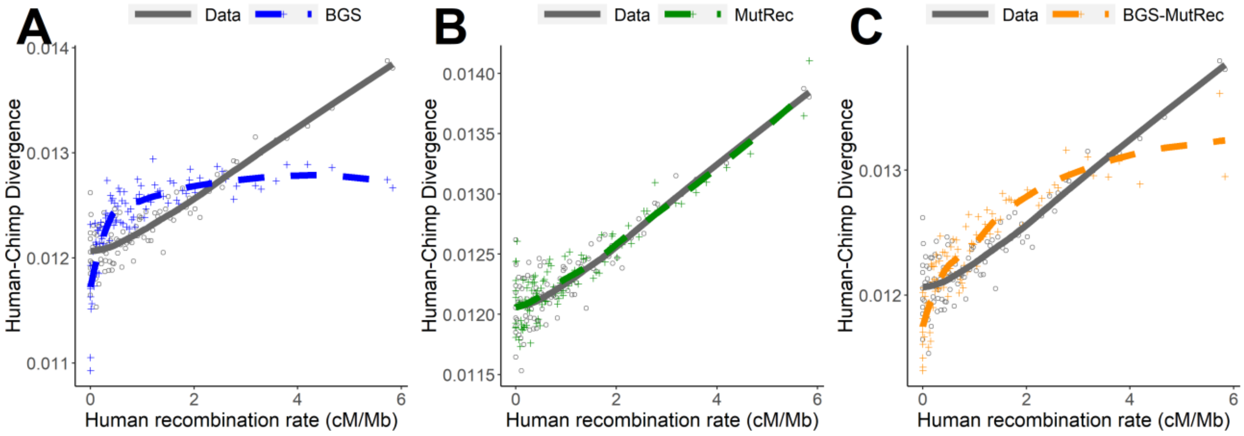


Figure 3.3. Test dataset was generated with mutagenic recombination. Each point represents the mean divergence and recombination rate in 1% of the 100kb windows binned by recombination rate. Each line represents the loess curve fit to the neutral divergence and recombination rate. Gray points and gray line represent the test dataset. (A) Blue points and blue line were generated using the MLEs from a model with just background selection. (B) Green points and green line were generated using the MLEs from a model with just mutagenic recombination. (C) Orange points and orange line were generated using the MLEs from a model with both background selection and mutagenic recombination. (B) showed an excellent fit to the test data.

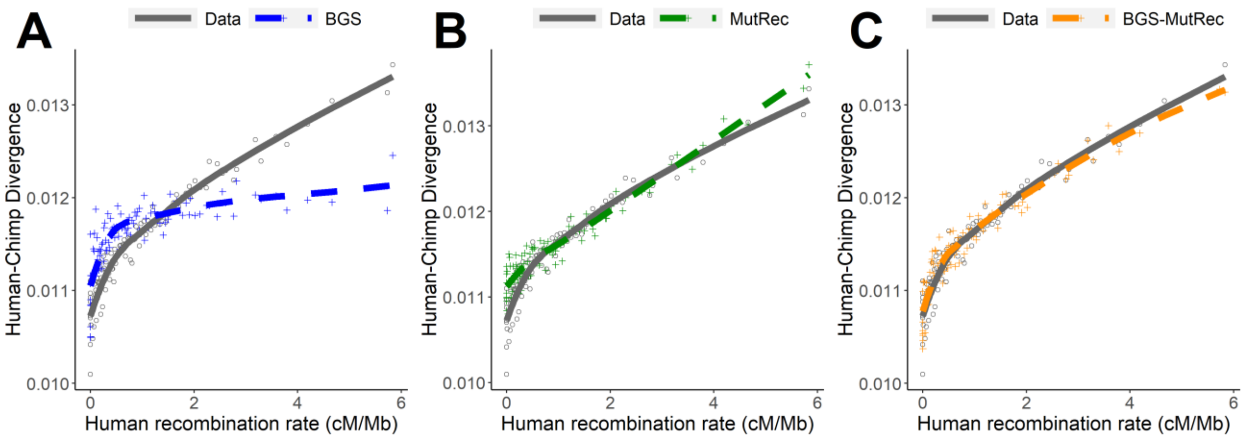


Figure 3.4. Test dataset was generated with both background selection and mutagenic recombination. Each point represents the mean divergence and recombination rate in 1% of the 100kb windows binned by recombination rate. Each line represents the loess curve fit to the neutral divergence and recombination rate. Gray points and gray line represent the test dataset. (A) Blue points and blue line were generated using the MLEs from a model with just background selection. (B) Green points and green line were generated using the MLEs from a model with just mutagenic recombination. (C) Orange points and orange line were generated using the MLEs from a model with both background selection and mutagenic recombination. (C) showed an excellent fit to the test data.

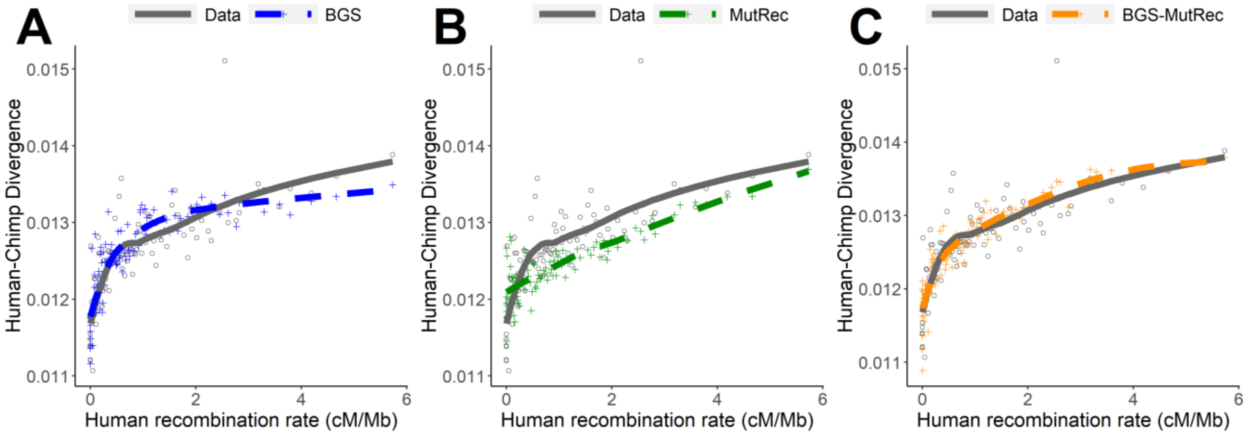


Figure 3.5. Model with both background selection and mutagenic recombination can recapitulate the empirical human-chimp neutral divergence. Each point represents the mean divergence and recombination rate in 1% of the 100kb windows binned by recombination rate. Each line represents the loess curve fit to the neutral divergence and recombination rate. Gray points and gray line represent the test dataset. (A) Blue points and blue line were generated using the MLEs from a model with just background selection. (B) Green points and green line were generated using the MLEs from a model with just mutagenic recombination. (C) Orange points and orange line were generated using the MLEs from a model with both background selection and mutagenic recombination. (C) showed an excellent fit to the empirical data.

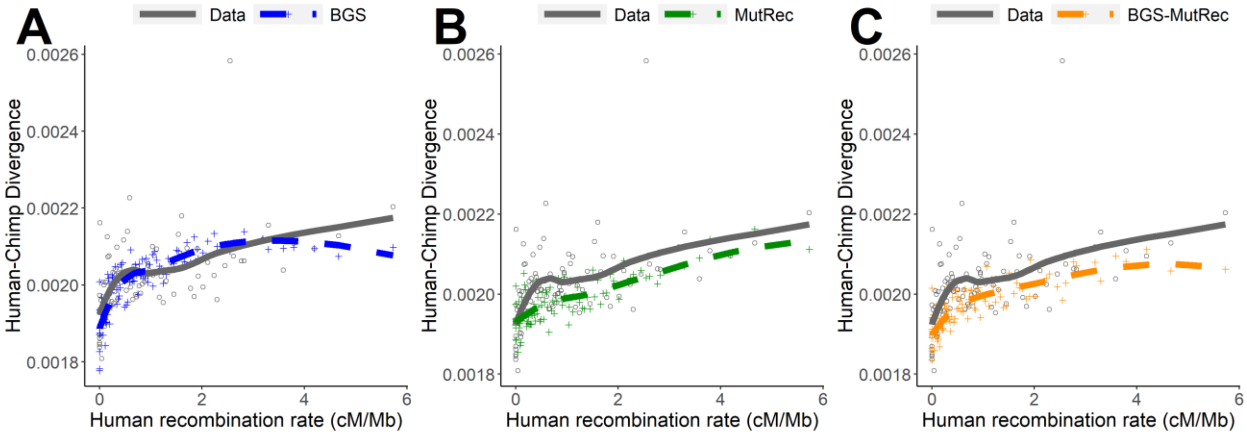


Figure 3.6. Biased gene conversion confounds signal for mutagenic recombination. Each point represents the mean divergence and recombination rate in 1% of the 100kb windows binned by recombination rate. Each line represents the loess curve fit to the neutral divergence and recombination rate. Gray points and gray line represent the test dataset. (A) Blue points and blue line were generated using the MLEs from a model with just background selection. (B) Green points and green line were generated using the MLEs from a model with just mutagenic recombination. (C) Orange points and orange line were generated using the MLEs from a model with both background selection and mutagenic recombination. (A) showed a good fit to the empirical data, indicating that when biased gene conversion is accounted for, a model with just background selection can recapitulate the empirical human-chimp neutral divergence.

3.7 Tables

Table 3.1. Parameters to infer for each model

Models	Parameters to infer	B-values used
Background selection	μ	McVicker's B-values
	N	
Mutagenic recombination	μ	B = 1
	N	
	φ	
Background selection and mutagenic recombination	μ	McVicker's B-values
	N	
	φ	

Table 3.2. Maximum likelihood estimates and best likelihoods when test data was generated with just background selection

Models	B-values	Parameters to infer	True value	Maximum likelihood estimates	Best log-likelihood
Background selection	McVicker's B	μ	2e-8	2e-8	-39725.129
		N_a	50000	50000	
Mutagenic recombination	B = 1	φ	N/A	0.018	-40246.541
		μ		1.96e-8	
		N_a		44000	
Background selection and mutagenic recombination	McVicker's B	φ	N/A	0.001	-39725.333
		μ		2e-8	
		N_a		50000	

Table 3.3. Maximum likelihood estimates and best likelihoods when test data was generated with just mutagenic recombination

Models	**B-values**	**Parameters to infer**	**True value**	**Maximum likelihood estimates**	**Best log-likelihood**
Background selection	McVicker's B	μ	N/A	2.06e-8	-42750.425
		N_a		62000	
Mutagenic recombination	B = 1	φ	0.05	0.052	-41097.746
		μ	2e-8	1.98e-8	
		N_a	50000	51000	
Background selection and mutagenic recombination	McVicker's B	φ	N/A	0.02	-42693.294
		μ		2.07e-8	
		N_a		58000	

Table 3.4. Maximum likelihood estimates and best likelihoods when test data was generated with both background selection and mutagenic recombination

Models	B-values	Parameters to infer	True values	Maximum likelihood estimates	Best log-likelihood
Background selection	McVicker's B	μ	N/A	1.96e-8	-40149.845
		N_a		59000	
Mutagenic recombination	B = 1	φ	N/A	0.071	-40307.766
		μ		1.97e-8	
		N_a		42000	
Background selection and mutagenic recombination	McVicker's B	φ	0.05	0.048	-39753.398
		μ	2e-8	2e-8	
		N_a	50000	50000	

Table 3.5. Maximum likelihood estimates and best likelihoods for each model for empirical human-chimp neutral divergence before accounting for biased gene conversion

Models	B-values	Parameters to infer	Maximum likelihood estimates	Best log-likelihood
Background selection	McVicker's B-values	μ	1.6e-8	-50184
		N_a	120000	
Mutagenic recombination	B-values = 1	φ	0.071	-51565
		μ	1.6e-8	
		N_a	90000	
Background selection and mutagenic recombination	McVicker's B-values	φ	0.034	-49902
		μ	1.8e-8	
		N_a	90000	

Table 3.6. Maximum likelihood estimates and best likelihoods for each model for empirical human-chimp neutral divergence after accounting for biased gene conversion

Models	B-values	Parameters to infer	Maximum likelihood estimates	Best log-likelihood
Background selection	McVicker's B-values	μ	2.9e-9	-33909
		N_a	90000	
Mutagenic recombination	B-values = 1	φ	0.01	-34885
		μ	2.7e-9	
		N_a	80000	
Background selection and mutagenic recombination	McVicker's B-values	φ	0.003	-34817
		μ	3.3e-9	
		N_a	60000	

CHAPTER 4

Complex patterns of sex-biased demography in canines

4.1 Abstract

Studies of genetic variation have shown that the demographic history of dogs has been complex, involving multiple bottleneck and admixture events. However, existing studies have not explored the variance in the number of reproducing males and females, and whether it has changed across evolutionary time. While male-biased mating practices, such as male-biased migration and multiple paternity, have been observed in wolves, recent breeding practices could have led to female-biased mating patterns in breed dogs. In addition, breed dogs are thought to have experienced the popular sire effect, where a small number of males father many offspring with a large number of females. Here we use genetic variation data to test how widespread sex-biased mating practices in canines are during different time points. Using whole genome sequence data from 33 dogs and wolves, we show that patterns of diversity on the X chromosome and autosomes are consistent with a higher number of reproducing males than females over ancient evolutionary history in both dogs and wolves, suggesting that mating practices did not change during early dog domestication. In contrast, since breed formation, we found evidence for a larger number of reproducing females than males in breed dogs, consistent with the popular sire effect. Our results confirm that the canines demography has been complex, with unique and opposite sex-biased processes occurring at different times. The signatures observed in the genetic data are consistent with documented sex-biased mating practices in both the wild and domesticated populations, suggesting that these mating practices are pervasive.

4.2 Introduction

Dogs were the first animals known to be domesticated and have lived alongside humans and shared our environment ever since (Hemmer 1990). There is tremendous interest in understanding their genetics and evolutionary history (Freedman et al. 2014; Freedman et al. 2016; Freedman and Wayne 2017; Ostrander et al. 2017). Many studies have shown that dogs have a complex evolutionary history; they experienced a population size reduction (i.e. bottleneck) associated with domestication and additional breed-specific bottlenecks associated with breed formation during the Victorian era (Boyko 2011). In addition to bottleneck events, dogs experienced admixture with wolves during the domestication process (vonHoldt et al. 2011). Studies have disagreed about the process of domestication, including when, where, and how many times dogs were domesticated (Larson et al. 2012; Thalmann et al. 2013; Freedman et al. 2014; Drake et al. 2015; Frantz et al. 2016; Botigué et al. 2017). However, despite the extensive work on understanding dog demographic history, existing studies have not explored the population history of males and females across dog domestication. Departures from an equal number of reproducing males and females are called sex-biased demographic processes, and leave signatures in the genome (reviewed in Wilson Sayres 2018 (Wilson Sayres 2018)). Previous ecological and field studies suggested that mating practices have been sex-biased in canines. In the wild populations, vonHoldt et al. (2008) observed that in some cases, Yellowstone male wolves would migrate to an existing wolf pack to mate with the alpha female when the alpha male dies (vonHoldt et al. 2008). The migration into an existing wolf pack is therefore male-biased. An additional source of male biased migration may come from male wolves called “Casanova wolves”. These wolves leave their natal packs and visit a nearby wolf pack around mating season to mate with the subordinate females (Westfall 2010). Lastly, there has also been evidence of multiple paternity in Ethiopian wolves and foxes (Sillero-Zubiri et al. 1996; Baker et al. 2004). In the domesticated populations, it is thought that more females contributed to breed formation than males, indicating female-biased processes (Sundqvist et al. 2006). In addition, recent reproductive practices, such as the popular sire effect, which involves a small number of males reproducing with a large number of females can lead to female-biased demography (Ostrander and Kruglyak 2000). Despite these observations of mating practices suggesting the numbers of reproducing males and females has been unequal during canid evolution, it is unclear how pervasive these processes are, and which have had the dominant effect on shaping patterns of diversity.

To test how widespread sex biased demography has been throughout canid evolution, we calculated and compared measures of genetic diversity on the X chromosome to those on the autosomes. This ratio has been termed Q in Emery et al. (2010) and we will use this notation throughout (Emery et al. 2010). In male-heterogametic sex-determining systems (XX/XY) with equal numbers of reproducing males and females, there are three copies of the X chromosome for every four copies of the autosomal genome. Therefore, in a constant size population without any natural selection or sex-biased processes, Q is expected to be 0.75 (reviewed in Webster and Wilson Sayres 2016 (Webster and Wilson Sayres 2016)). Specifically, $Q = N_X/N_A \cong 0.75$. Deviations from this expected ratio could be indicative of sex-biased processes. If $Q < 0.75$, there are fewer copies of the X chromosome than expected, suggesting a larger number of reproducing males than reproducing females, indicative of male-biased processes. If $Q > 0.75$, there are more copies of the X chromosome than expected, suggesting a larger number of reproducing females than reproducing males, indicative of female-biased processes.

Studies comparing measures of genetic diversity between the X chromosome and autosomes have resulted in many insights into the evolutionary history of humans. Hammer et al. (2008) computed Q by fitting a model of demographic history to the ratio in the mean of genetic diversity within the X chromosome and autosomes: $Q_{\pi} = \pi_X / \pi_A$ (Hammer et al. 2008). They found that Q_{π} is greater than 0.75 in all human populations examined, suggesting female-biased processes that have led to more reproducing females than males during human evolutionary history (Hammer et al. 2008). Later, Keinan et al. (2009) (Keinan et al. 2009) computed Q by calculating the ratio in fixation index, F_{ST}, between the X chromosome and the autosomes: $Q_{FST} = \frac{\ln(1-2F_{ST}^{A})}{\ln(1-2F_{ST}^{X})}$. They found that Q_{FST} is less than 0.75 only when comparing a non-African population to an African population (Keinan et al. 2009). This result suggests that there was a male-biased migration out of Africa, where there were more reproducing males than females. Even though these two studies came to different conclusions regarding the sex ratio in human history, a later study reconciled these seemingly disparate findings by demonstrating that Q can detect bias in sex ratios at different timescales, depending on whether it is calculated from genetic diversity (Q_{π}) or the fixation index (Q_{FST}) (Emery et al. 2010). Specifically, Q_{π} can detect sex bias in ancient timescales, which is before or immediately after the split between populations, whereas Q_{FST} detects sex-biased demography on recent timescales, after the populations split from each other (Emery et al. 2010). Emery et al. (2010) reconciled results from Hammer et al. (2008) and Keinan et al. (2009) by showing that evolutionary processes within human history are consistent with an earlier female bias followed by a male bias during the migration of some humans out of Africa (Emery et al. 2010). Additionally, direct comparisons of the two studies were complicated by linked selection on the X chromosome (Hammer et al. 2010; Arbiza et al. 2014). In addition to humans, comparing the genetic diversity between the X chromosome and autosomes has also been used to study sex-biased processes in many other species (Wilson Sayres 2018).

Given how examining patterns of genetic diversity on the X chromosome and the autosomes has facilitated our understanding of sex-biased demography in other species and what has been observed regarding sex-biased mating practices in canines, we wanted to test how widespread these mating practices are throughout different time points during canine evolutionary history. We utilized whole-genome sequences of 21 dogs and 12 wolves. Using the estimator of the effective sex ratio based on nucleotide diversity, we found that Q_{π} is less than 0.75 in both dogs and wolves, indicative of an ancient male bias either in the shared ancestral population, or immediately after their split. We then inferred the effective sex ratio in a population genetic model, demonstrating that a population size reduction by itself cannot generate the empirical patterns. Rather, a male-biased sex ratio was needed in conjunction with a population size reduction to recapitulate empirical patterns. Finally, using the estimator of the effective sex ratio based on the fixation index, we showed that while the demographic history in wolves has remained male-biased in recent history, the demographic history in dogs has changed from male-biased in the ancient timescale to female-biased in recent times. These results add to our current understanding about the canine demographic history and suggest the need to incorporate sex-biased demography in future studies.

4.3 Results

Description of the data

We collected a dataset of 33 female canid whole genomes that include 4 German Shepherds, 5 Tibetan Mastiffs, 12 dog individuals from a variety of breeds, 6 Arctic Wolves, and 6 Grey Wolves (Table 4.1). The German Shepherd and Tibetan Mastiff data were sequenced by

Gou et al. (2014) (Gou et al., 2014) and the *fastq* files were downloaded from NCBI SRA. We combined 12 high coverage (>15X) whole genome sequences of female dogs from multiple breeds that were included in Marsden et al. (2016) (Marsden et al. 2016) because we were interested in how results differ between using a group of one breed versus using a group consisting of multiple breeds. We named this pooled group the "Pooled Breed Dogs". The Arctic Wolf data were sequenced by Robinson et al. (Submitted). These Arctic Wolves were located in Northern Canada (north of the Arctic circle). The longitudinal and latitudinal locations for these Arctic Wolves are included in Table 4.1. We also used high coverage (>15X) whole genome sequences of female Grey Wolves from Marsden et al. (2016) (Marsden et al. 2016). Since these Grey Wolves originated from Europe, Asia, and Yellowstone, we named this population the "Pooled Grey Wolves". Details about coverage and accession numbers for the individuals in this study are summarized in Table 4.1.

Estimating the effective sex ratio based on genetic diversity

Previous work has shown that dogs experience male mutation bias, where the mutation rate is higher in males compared to females due to more germline cell divisions in males at reproduction (Li et al. 2002; Lindblad-Toh et al. 2005; Wilson Sayres and Makova 2011). Male mutation bias has a significant impact on measurements of genetic diversity because it can inflate raw metrics of genetic diversity on the autosomes compared to on the X chromosome (reviewed in Webster and Wilson Sayres 2016 (Webster and Wilson Sayres 2016)). To confirm that male mutation bias exists in our data, we computed male mutation bias for each population using dog-cat divergence (see Methods). We observed that the level of male mutation bias is around 2, which is consistent with previous reports (Lindblad-Toh et al. 2005; Wilson Sayres and Makova 2011) (Table 4.2). Therefore, we controlled for male mutation bias in all estimates of genetic variation by normalizing autosomal and X chromosome diversity by dog-cat divergence in the corresponding regions.

Natural selection is thought to be more efficient at reducing genetic diversity on the X chromosome than on the autosomes because males have only one X chromosome which is exposed directly to selection (reviewed in Webster and Wilson Sayres 2016 (Webster and Wilson Sayres 2016)). To control for natural selection affecting the X chromosome more than the autosomes, we used regions of the genome in which mutations would be putatively neutral by removing sites that are functional. Specifically, we removed genic and conserved sites (see Methods).

To understand whether any evolutionary process has been sex-biased over ancient timescales, we computed Q_{π}. We found that in both dog and wolf populations, Q_{π} is significantly less than 0.75 (Figure 4.1, No cM cutoff), suggesting a male-biased sex ratio, with more males reproducing relative to females.

Q_{π} of less than 0.75 could occur due to the effect of natural selection on linked neutral sites. Specifically, natural selection could have reduced diversity in linked neutral regions on the X chromosome more than on the autosomes, as seen in humans (Keinan et al. 2009; Hammer et al. 2010; Arbiza et al. 2014). Further, it is possible that there is more constraint on noncoding regions near genes on the X chromosome than on the autosomes (Narang et al. 2016). To measure how neutral diversity is affected by linked selection, we compared diversity on the X chromosome and autosomes in regions near genes versus putatively unconstrained regions 0.4 cM away from the nearest gene. Diversity increased more with increasing distance from genes on the X chromosome than on the autosomes, consistent with natural selection reducing diversity more on the X chromosome than on the autosomes near genes (Table 4.3).

To test whether stronger linked selection acting on the X chromosome relative to the autosomes could cause Q_π to be less than 0.75, we expanded our filtering criteria to remove sites that are near genes, defined by genetic distance (see Methods). Since we did not know *a priori* what the minimum genetic distance would be required to obtain sites that are not affected by selection, we included several thresholds. We removed sites whose genetic distance to the nearest genes is less than 0.2 cM, 0.4 cM, 0.6 cM, 0.8 cM, and 1 cM. We observed that even after removing sites whose genetic distance to the nearest genes are less than 1 cM, Q_π is still less than the expected 0.75 in both dog and wolf populations, except for the German Shepherd (Figure 4.1). In the German Shepherd, when using the thresholds of 0.8 cM and 1 cM, Q_π approaches 0.75. However, since there are significantly fewer sites and variants left after removing sites whose genetic distance to the nearest genes is less than 0.8 cM or 1 cM, we could not exclude the possibility that we are underpowered to detect any signal in the data (Table 4.4). Nonetheless, these results suggest that while linked selection may partially account for Q_π of less than 0.75, especially in the German Shepherd, linked selection by itself cannot explain why Q_π is less than 0.75 across all dog and wolf populations. In sum, our results suggest that there has been male-biased sex ratios in both dogs and wolves over ancient evolutionary timescales.

Inference of sex-biased demographic processes under population genetic models

Pool and Nielsen (2007) demonstrated that a Q_π of less than 0.75 could be explained by a reduction in population size even with an equal number of breeding males and females (Pool and Nielsen 2007). To test whether population bottlenecks can explain the reduction in diversity on the X chromosome, we fitted a demographic model that includes a bottleneck using the autosomal site frequency spectrum (SFS) (Figure 4.2) and asked whether the best fitting demographic model on the autosomes could also account for the level of diversity on the X chromosome when using an N_X/N_A ratio of 0.75. If a demographic model including a bottleneck by itself can generate a Q_π of less than 0.75, we would expect that scaling the population size of the X chromosome to be three-quarters that of the autosomes should result in a Q_π comparable to the empirical data. Additionally, we then employed a composite likelihood framework to directly infer the N_X/N_A ratio from the SFS while accounting for the complex non-equilibrium demography.

First, we fitted a demographic model that includes a bottleneck using the SFS on the autosomes using *fastsimcoal2* (Excoffier et al. 2013) for each population considering regions of greater than 0.4 cM, 0.6 cM, 0.8 cM, and 1 cM from genes. We reasoned that we would not be able to exclude the role of selection when not removing sites near genes or using too small of a threshold (i.e. 0.2 cM). We also corrected for male mutation bias using mutation rates that we inferred from dog-cat divergence in the same windows (see Methods; Table 4.2). The inferred demographic parameters that resulted in the best likelihood of the data are presented in Table 4.5. To test whether the inferred demographic parameters can recapitulate the autosomal data, we used *fastsincoal2* to generate the expected SFSs. In all populations except the German Shepherds, across all thresholds examined, we observed that the SFSs generated using the inferred demographic parameters visually match with the empirical autosomal SFSs (Figure 4.3). The differences in log-likelihood between the simulated SFSs and the empirical SFSs are also small (Table 4.6), confirming our visual inspection of the fit of the demographic models. In addition, autosomal genetic diversity (π) computed from the demographic model is comparable to the empirical estimates of π (Figure 4.4). Thus, these lines of evidence demonstrate that the inferred demographic parameters can recapitulate the empirical data on the autosomes, except for the more stringent filtering on the German Shepherd.

To understand whether the demographic model including a bottleneck that was fitted to the autosomal data could account for the level of diversity on the X chromosome, we used the inferred demographic parameters to simulate the SFSs for the X chromosome. To account for the differences in population size between the X chromosome and the autosomes, we adjusted the population size on the X chromosome by a constant value which we called C, where $N_X = CN_A$. If a bottleneck by itself without any sex biased demography can generate a Q_π of less than 0.75, we expected that using a C value of 0.75 would recapitulate the empirical data. If a bottleneck model by itself is not sufficient to generate a Q_π of less than 0.75, and sex-biased processes need to be invoked, we expected that rescaling the population size on the X chromosome to be three-quarters of the population size on the autosomes would not fit well. Rather, a different value of C would yield a better fit.

To assess whether a null C value of 0.75 or a different C value yielded a better fit to the empirical SFSs on the X chromosome, we searched over a grid of C values. We found the maximum likelihood value of C for each population and filtering threshold. To do this, for each C on a grid of C values, we first calculated the population size on the X chromosome, which is $N_X = CN_A$. We then used *fastsimcoal2* to simulate an SFS and assess the fit by comparing the Poisson log-likelihood to the SFS on the X chromosome (see Methods). For each population and for each threshold, we found a set of C values that maximizes the likelihood of the data (Figure 4.5, Table 4.7).

With the exception of the German Shepherd at the most stringent filtering thresholds (>0.8 cM and >1 cM), we inferred that C is less than 0.75 for all population and filtering thresholds. When using a filtering threshold of 0.4 cM from genes, we found that C ranges from 0.61 to 0.68. The full model, where we inferred C for each comparison, fits the observed X chromosome SFS significantly better than a model where C is constrained to be 0.75 (Likelihood Ratio Tests > 30, p-value < 10^{-8}; Table 4.7). Further, the null C value of 0.75 does not visually fit the SFSs on the X chromosome (Figure 4.6, blue bars), suggesting that we can reject an equal number of reproducing males and females. Third, we observed that diversity on the X chromosome from simulating with a null C value of 0.75 overestimated the empirical X chromosome diversity (Figure 4.7, blue bars). These results suggest that a model including both a bottleneck and a male-bias sex ratio can generate Q_{π} of less than 0.75 and recapitulate the observed SFSs and genetic diversity. Only in the German Shepherd population when using the most stringent threshold (>0.8 cM and >1 cM), can a demographic history including a bottleneck by itself generate a Q_{π} of less than 0.75.

Female-biased sex ratio within dogs in recent history

Since estimates of sex ratios from levels of genetic diversity are sensitive to ancient sex-biased processes (prior to or immediately after the split between two species), we wanted to determine whether the pattern of male-biased contributions remained constant throughout the evolutionary history of canines (Emery et al. 2010). To study sex-biased demography on recent timescales, we computed Q_{FST} for each pair of populations (see Methods). In the dog to dog comparison, we computed Q_{FST} between German Shepherds and Tibetan Mastiffs, between German Shepherds and Pooled Breed Dogs, and between Tibetan Mastiffs and Pooled Breed Dogs. We observed that Q_{FST} is greater than 0.75 for all three pairs and across all thresholds, suggesting a female-biased sex ratio within the dog populations in recent history (Figure 4.8). This is consistent with fewer reproducing males than females in the population since the formation of different dog breeds. In the wolf to wolf comparison, we computed Q_{FST} between Arctic Wolves and Pooled Grey Wolves. In contrast to the breed dogs, we found that Q_{FST} is less than 0.75 when using the thresholds of >0.4 cM and >0.6 cM, suggesting that a male-biased sex ratio has been maintained within the wolf populations in recent history (Figure 4.8). However, we noted that when using a more stringent threshold (>0.8 cM or >1 cM), Q_{FST} within wolves approaches 0.75 or greater than 0.75 (Figure 4.8). We could not exclude the possibility that we are unable to detect a true signal in the data due to significantly fewer sites and variants left after the more stringent filtering (Table 4.4). Overall, these results indicate that while the process within wolves has probably maintained a male-bias from ancient to recent history, the process within dogs has changed to female-bias, potentially because of breeding practices that have led to female-biased processes such as the popular sire effect.

4.4 Discussion

In this study, we used two different statistics to estimate the ratio of reproducing males to females in canines and found that the demographic history of dogs and wolves has been sex-biased, but not always in the same direction. Estimating the sex ratio based on the levels of genetic diversity (Q_{π}) from the X chromosome and autosomes showed a male-biased sex ratio in both dogs and wolves on an ancient timescale, which cannot be explained by linked selection or a population size reduction on its own (Figure 4.1 and Figure 4.5). Instead, in both dogs and wolves, there has been a larger number of reproducing males than females. In wolf packs, the alpha male and female are the dominant reproducers, but subdominant reproduction is common and may involve multiple fathers for a single litter (vonHoldt et al. 2008). Multiple paternity is a unique aspect of canid reproduction and may help drive a male bias in reproduction, as offspring of a single litter can only have a one mother, but may have multiple fathers and litter size may be as large as 16 individuals (Stahler et al. 2013). In addition, wolves migrating to existing wolf packs are predominantly male-biased (vonHoldt et al. 2008). Further, “Casanova wolves” who stay near a wolf pack during mating season to mate with the non-alpha females could also cause male-biased mating patterns (Westfall 2010). Multiple paternity and male-biased migration likely occurred in early dogs, but under more recent controlled breeding, valuable sires would be the only father of a litter. Hence the controlled nature of breeding in modern dog breeds, and the focus on a subset of “popular” sires could drive the female bias in reproduction. The population sire effect also reduces the effective size of breeds and effects such as inbreeding further skew evolution in modern breeds.

In addition, we observed that determining the amount of bias based on the absolute value of Q_π by itself can lead to overestimation, because the reduction of diversity on the X chromosome due to a population size reduction is not accounted for. For example, in Tibetan Mastiff, when using a threshold of 0.6 cM to remove linked neutral sites, a Q_π of 0.52 suggests an N_X/N_A ratio of 0.52. However, we inferred a *C* value of 0.57 (confidence interval: 0.56-0.6) using our modelling framework, indicating that the sex ratio is higher than when just examining the absolute value of Q_π. This difference exists because the estimate of Q_π could be affected by a population size reduction differentially influencing diversity on the X and autosomes (Pool and Nielsen 2007), but our inference framework accounts for this effect. Our findings suggest that inferring the sex ratio in a model-based framework should yield a more accurate estimate than the absolute Q_π (Hammer et al. 2008).

Our results add to the growing literature on the complex demographic history of dogs (reviewed in Freedman et al. 2016 (Freedman et al. 2016) and Ostrander et al. 2017 (Ostrander et al. 2017)). In addition to multiple episodes of bottleneck and admixture events, we now present evidence for sex-biased demographic processes. Furthermore, we provide evidence that sex-biased processes within dogs have changed throughout evolution, switching from a male-bias in ancient timescales to a female-bias in recent timescales, reflecting how modern breeding practices influence the sex ratio. To the best of our knowledge, this is the first genomic study of sex-biased demography in dogs. Some limitations in this study provide avenues for future work. First, our study was limited by the availability of high coverage (>15X coverage) whole-genome sequences of female individuals at the time of analysis. Future studies could utilize more female individuals and a variety of populations to understand whether there are differences in sex-biased processes between breeds. Second, future work could extend our modelling framework by including more complex demographic scenarios such as migration events to better capture the autosomal data, especially the German Shepherds. Finally, future studies could examine whether processes such as admixture with wolves or introgression has been sex-biased.

4.5 Methods

Whole-genome sequence processing

We followed Genome Analysis Toolkit's (GATK) documentation for variant discovery best practices (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). Scripts used for processing whole-genome sequencing for each of the following steps can be found at https://github.com/tnphung/NGS_pipeline.

Data pre-processing for variant calling

First, we converted all fastq files to raw unmapped reads using Picard FastqToSam. Second, we marked Illumina adapters using Picard MarkIlluminaAdapters. Third, we mapped to the reference dog genome (canFam3) using bwa-mem(Li 2013). Fourth, we marked duplicates using Picard MarkDuplicates. We then recalibrated base quality scores using GATK where we performed three rounds of recalibration to obtain analysis-ready reads in BAM file format.

Variant calling with GATK

We used GATK Haplotype caller for variant calling (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). We first generated a gVCF file for each individual. We then performed joint-genotyping for all 33 individuals in our study.

Filtering to obtain high quality sites

To obtain sites that are high confidence, we retained sites whose depth (annotated as DP in VCF file format) is between 50% and 150% of the mean depth across all sites. In addition, we only kept sites that were genotyped in all 33 individuals (i.e. the total number of alleles in called genotypes, AN, is equal to 66).

Variant filtering

We obtained variant sites from the VCF files by using GATK SelectVariant s(McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). We then filtered these variants by applying GATK Hard Filter (QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0). In addition, we only selected biallelic SNPs and removed any clustered SNPs defined by having 3 SNPs within 10bp.

Filtering nucleotide sites

Filtering out the pseudoautosomal regions (PARs) of the X chromosome

Previous work showed that the PARs in canines span the first 6.59Mb of the X chromosome (Young et al. 2008). Therefore, we filtered out the PARs by removing any site that overlaps with the first 6.59Mb of the X chromosome. In humans, it was shown that genetic diversity does not drop abruptly at the PAR boundary (Cotter et al. 2016). Rather, genetic diversity decreases gradually over the PAR boundary and reaches nonPAR diversity past the PAR boundary (Cotter et al. 2016). One concern is that filtering out the PARs is not sufficient to avoid any inflation of X-linked variation. However, if this is the case, we would expect Q_π we calculated to be higher than the actual Q_π. Therefore, Q_π less than 0.75 is not caused by not sufficiently filtering sites on the nonPARs.

Filtering sites that could be under the direct effect of selection

To control for the effects of direct selection, we removed sites that are potentially functional and therefore are more likely to be affected by purifying or positive selection. Specifically, we removed sites that overlap with a gene transcript as defined by Ensembl (gene transcripts include both exons and introns). We also removed sites that are conserved across species. To obtain conserved sites, we downloaded phastConsElements100way for hg19 from the UCSC Genome Browser and used liftOver command line tool to convert hg19 coordinates to canFam3 coordinates.

Filtering out sites that could be affected by linked selection

To control for the effect of natural selection on linked neutral sites, we employed a filtering criterion to remove sites near genes as defined by genetic distance to the nearest genes. We used the genetic distance map based on patterns of linkage disequilibrium from Auton et al. (2013) because this genetic map includes information for the X chromosome whereas the pedigree map from Campbell et al. (2016) does not have information on the X chromosome

(Auton et al. 2013; Campbell et al. 2016). For each site that is outside of genes and conserved regions, we found its nearest gene in terms of physical distance. We then converted physical distance to genetic distance using the genetic map from Auton et al. (2013) (Auton et al. 2013). Since we did not know *a priori* what the minimum genetic distance is required to remove sites near genes to control for linked selection, we used multiple thresholds. Specifically, we removed sites whose genetic distance to the nearest gene is less than 0.2 cM, less than 0.4 cM, less than 0.6 cM, less than 0.8 cM, and less than 1 cM.

Identifying sites that are alignable between dog and cat

Since we controlled for mutation rate variation by normalizing the uncorrected genetic diversity by dog-cat divergence, we identified regions of the genome that are alignable between dog and cat. We downloaded the pairwise alignment between dog and cat from the UCSC Genome browser (Kent et al., 2002). We then generated BED files whose coordinates represent regions of the genome that are alignable between dog and cat.

In summary, for our empirical analyses, we used regions of the genome that are (1) not affected directly by selection, (2) not affected by linked selection using multiple thresholds, (3) high in quality (see the section on filtering to obtain high quality sites above), and (4) alignable between dog and cat.

Computing $\boldsymbol{Q_\pi}$

Computing uncorrected average pairwise differences between sequences (π)

We computed genetic diversity, π, defined as the average number of differences between pairs of sequences (Tajima 1983):

$\pi = \frac{n}{n-1}\sum_{i}^{\text{all sites}} p_i(1-p_i)$ where p_i is the allele frequency and n is the number of alleles. For each region of the genome that satisfies the filtering criteria above, we computed π for the X chromosome and autosomes. To obtain the mean in diversity, π/site, we calculated: $\pi/\text{site} = \frac{\sum_i^{\text{regions}} \pi}{\sum_i^{\text{regions}} \text{total sites}}$.

Computing dog-cat divergence

For each region of the genome that satisfies the filtering criteria above, we tabulated the number of DNA differences between dog and cat. To obtain the mean in divergence, we calculated $\frac{\text{divergence}}{\text{site}} = \frac{\sum_i^{\text{regions}} \text{number of divergent sites}}{\sum_i^{\text{regions}} \text{total sites}}$.

Computing male mutation bias

We computed male mutation bias (α) using divergence on the X chromosome and on the autosomes as follows(Link et al. 2017): $\alpha = \frac{4-3\frac{X}{A}}{3\frac{X}{A}-2}$.

Computing corrected diversity

To control for variation in mutation rates across chromosomes, we normalized diversity by dog-cat divergence by dividing π/site by divergence/site.

Constructing 95% confidence interval by bootstrapping

We generated bootstrap replicates of the BED file that we used to compute genetic diversity and divergence by randomly selecting a fragment from the BED file with replacement. For each bootstrap replicate, the number of fragments chosen was equal to the number of fragments in the original BED file. We generated 1000 bootstrap replicates. For each of the 1000 bootstraps on the X chromosome, we computed uncorrected π, dog-cat divergence, and corrected π. We did the same calculations for each of the 1000 bootstraps on the autosomes. We then divided corrected π on the X chromosome by corrected π on the autosomes to obtain Q_π. We calculated 95% confidence interval using 1000 bootstrapped values of corrected π_X, 1000 bootstrapped values of corrected π_A, and 1000 bootstrapped values of corrected Q_π by selecting the values at the 2.5 and 97.5 percentiles.

Computing $\boldsymbol{Q_{FST}}$

Computing F_{ST}

We computed Weir and Cockerham's F_{ST} for each pair of populations using the *SNPRelate* package implemented in *R* (Zheng et al. 2012). For dog-to-dog comparison, we computed F_{ST} for German Shepherds and Tibetan Mastiffs, German Shepherds and Pooled Breed Dogs, and Tibetan Mastiff and Pooled Breed Dogs. For wolf-to-wolf comparison, we computed F_{ST} for Arctic Wolves and Grey Wolves. Since the number of individuals differs between populations, we subsampled such that there were four individuals in each population (Table 4.8). We computed F_{ST} for the X chromosome and for the autosomes.

Computing Q_{FST}

We computed Q_{FST} using: $Q_{FST} = \frac{\ln(1-2F_{ST}^A)}{\ln(1-2F_{ST}^X)}$ (Keinan et al. 2009; Emery et al. 2010).

Constructing 95% confidence interval by bootstrapping

Since the input to *SNPRelate* to calculate F_{ST} is a VCF file format, we generated 1000 bootstrapped VCF files by randomly selecting variants from the VCF file with replacement. The number of variants selected for each bootstrapped VCF is equal to the number of variants in the empirical VCF file. For each bootstrapped VCF, we computed F_{ST} and Q_{FST} as explained above. From the 1000 values of bootstrapped Q_{FST}, we then calculated 95% confidence interval by selecting the values at the 2.5 and 97.5 percentiles.

Modeling framework to estimate the $\boldsymbol{N_X/N_A}$ ratio ($\boldsymbol{C}$)

Obtaining the site frequency spectrum (SFS)

We computed the folded SFSs using Equation 1.2 of Wakely's An Introduction to Coalescent Theory (Wakely, 2008), reproduced as follows:

$$\eta_i = \frac{\xi_i + \xi_{n-i}}{1 + \delta_{i,n-i}} \quad 1 \ \leq i \leq [n/2]$$

where ξ_i is the number of sites where the alternate allele is present at *i* copies, $\delta_{i,n-i}$ is equal to 0 when $i \ \neq n - i$ and is equal to 1 when $i = n - i$. For each population and for each threshold to remove linked neutral sites (>0.4 cM, >0.6 cM, >0.8 cM, and >1 cM), we computed the folded SFSs for the X chromosome and autosomes.

Computing mutation rates

We utilized dog-cat divergence to infer the mutation rates for the X chromosome and autosomes. Specifically, $\mu = \frac{D}{2t_{split}}$, where D is the divergence/site between dog and cat (see Computing dog-cat divergence section above) and t_{split} is the split time between dog and cat in unit of generation. We used 54 million years as the split time between dog and cat and a generation time of 3 years per generation(Hedges et al. 2006; Hedges et al. 2015). The estimates of mutation rates are in the same order of magnitude as estimate from ancient DNA (Table 4.4) (Botigué et al. 2017).

Inferring demographic parameters

We inferred demographic parameters from the autosomal data (SFSs on the autosomes) using a maximum likelihood framework as implemented in *fastsimcoal2* (Excoffier et al. 2013). We specified a bottleneck demographic model and inferred four parameters: N_{ANC} which is the population size in the ancestral population, N_{BOT} which is the population size during the bottleneck, N_{CUR} which is the population size in the current day, and T_{BOT} which is the duration between the end of the bottleneck and current day (Figure 4.2). Further, we repeated the inference of the previous four parameters for values $\mathrm{BOT}_{\mathrm{DUR}}$ (the duration of the bottleneck) ranging from 75 to 100 generations (Figure 4.2) and chose the value that yielded the highest likelihood. We implemented this procedure for each population and for thresholds of >0.4 cM, >0.6 cM, >0.8 cM, and >1 cM to remove linked sites. The demographic parameters that maximized the likelihood are summarized in Table 4.5.

Inferring N_X/N_A ratio (C)

To account for differences in population size between the X chromosome and autosomes, we scaled the population size on the X chromosome to that on the autosomes by a constant factor we called C, where $N_X = CN_A$. To find the maximum likelihood estimate of C, we searched over a grid for values of C, including 0.75, to find a value that resulted in the highest likelihood. Because the number of SNPs at particular frequencies contains substantial information about demography, we used a Poisson likelihood for the number of SNPs in each entry of the SFS to compute the Poisson log-likelihood as in Beichman et al. (2017) (Beichman et al. 2017).

Accessing fit of MLEs of C to π

We computed diversity from the simulated SFSs under the demographic models fit to the autosomes using the MLEs of *C* (Table 4.7) and compared that to the empirical uncorrected diversity.

Data, codes, and materials

All scripts can be found at https://github.com/tnphung/SexBiased. SRA numbers for *fastq* files for published genomes are listed in Table 4.1. SRA numbers for fastq files for Arctic Wolf individuals will be deposited to SRA before publication. Post base quality score calibration (BQSR) BAM files and VCF files will be deposited on Dryad before publication.

4.6 Figures

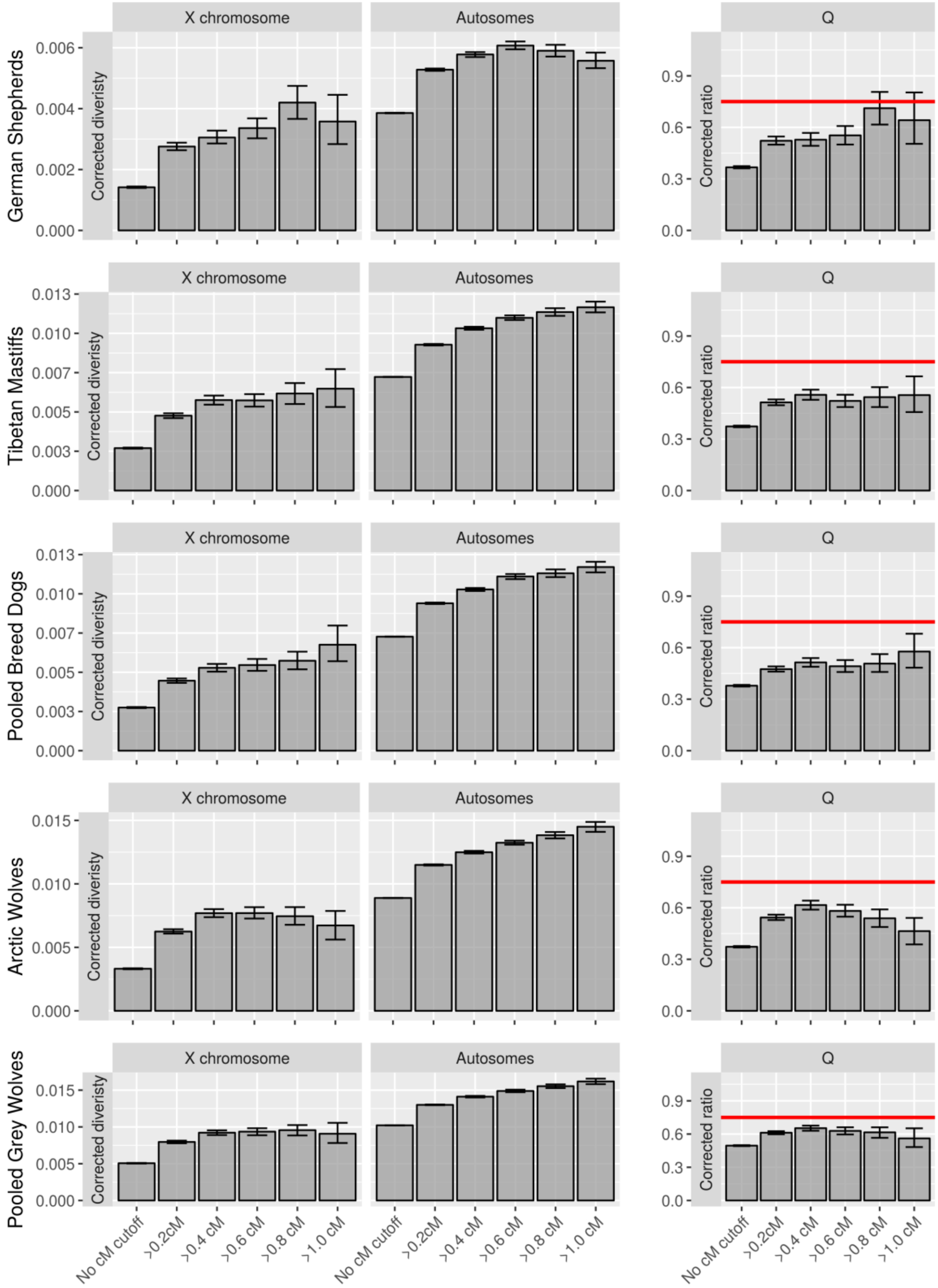

X chromosome
Autosomes
Q
German Shepherds
Tibetan Mastiffs
Pooled Breed Dogs
Arctic Wolves
Pooled Grey Wolves
Corrected diveristy
Corrected ratio
0.000
0.002
0.004
0.006
0.003
0.005
0.007
0.010
0.013
0.015
0.0
0.3
0.6
0.9
No cM cutoff
>0.2cM
>0.4 cM
>0.6 cM
>0.8 cM
>1.0 cM

Figure 4.1. X-linked and autosomal genetic diversity across canids. Genetic diversity measured as the average pairwise differences between sequences (π) corrected for mutation rate variation using divergence (see Methods) on the X chromosome and autosomes in multiple canid populations. Q denotes the ratio of π on the X chromosome to that of the autosomes. The horizontal red line denotes the null expectation of 0.75. Bins along the x-axis denote different filtering based on genetic distances from genes. Error bars denote 95% confidence intervals obtained through bootstrapping (see Methods).

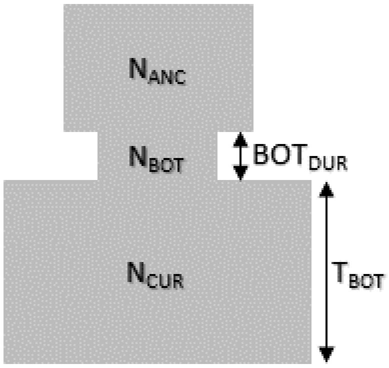


Figure 4.2. Model of demographic history including a bottleneck, redrawn from the *fastsimcoal2* manual. Four parameters were inferred from the site frequency spectrum: N_{ANC} is the ancestral population size, N_{BOT} is the population size at during the bottleneck, N_{CUR} is the population size in current day, and T_{BOT} is the duration between the end of the bottleneck and present day which is in units of generations. In *fastsimcoal2*, the population size is in unit of haploid individuals. In addition, BOT_{DUR} represents the duration of the bottleneck, also in units of generations. We tried multiple values of BOT_{DUR} and chose the value that yielded the highest likelihood. The values of N_{ANC}, N_{BOT}, N_{CUR}, T_{BOT}, and BOT_{DUR} that resulted in the best likelihood are summarized in Table 4.5.

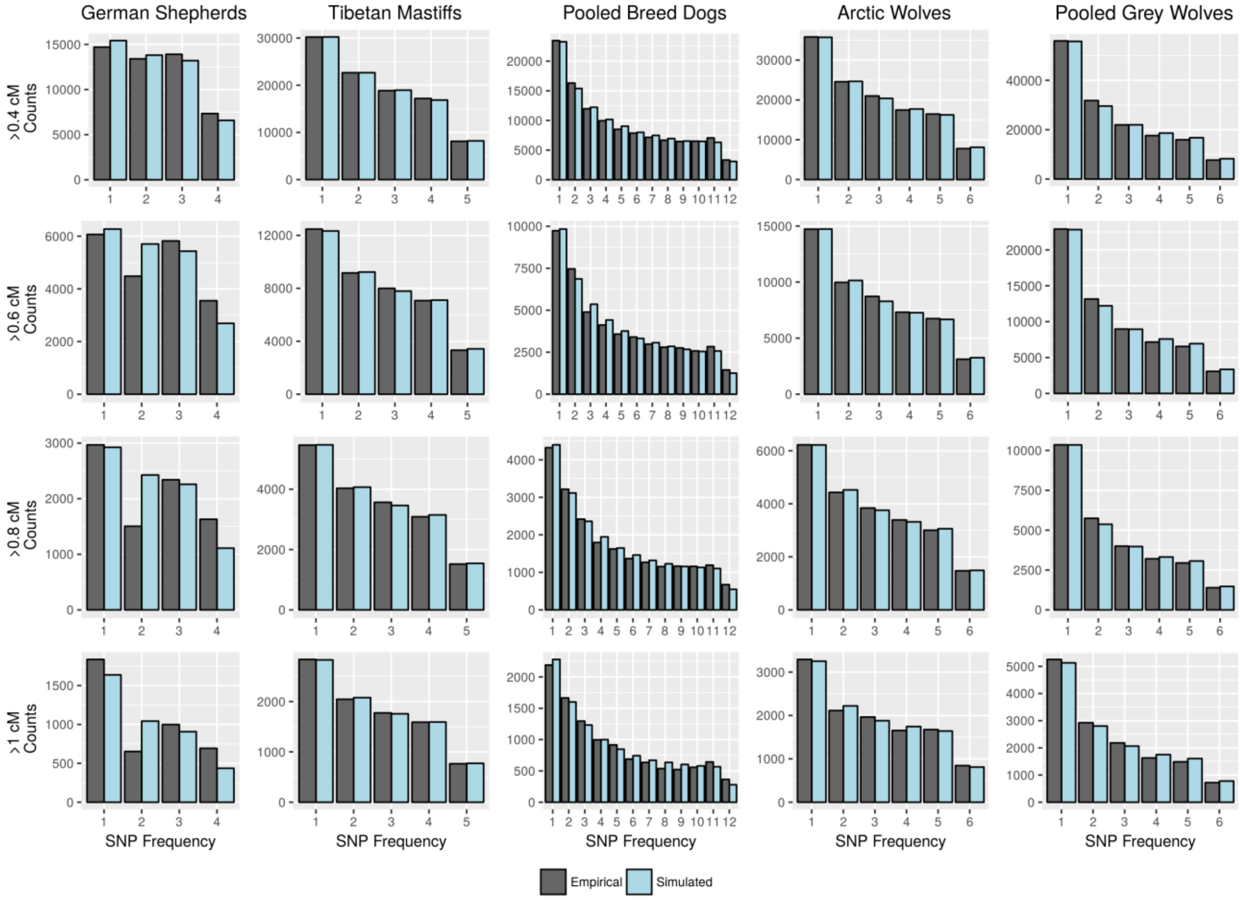


Figure 4.3. SFSs simulated under our best-fitting demographic models (blue bars) compared to the empirical SFSs on the autosomes (grey bars) in multiple canid populations. Here, multiple thresholds were used to remove sites potentially linked to selected sites (>0.4 cM, >0.6 cM, >0.8 cM, and >1 cM). For each canid population and for each threshold, the simulated SFSs were generated using the demographic parameters that resulted in the highest likelihood for the autosomal SFS.

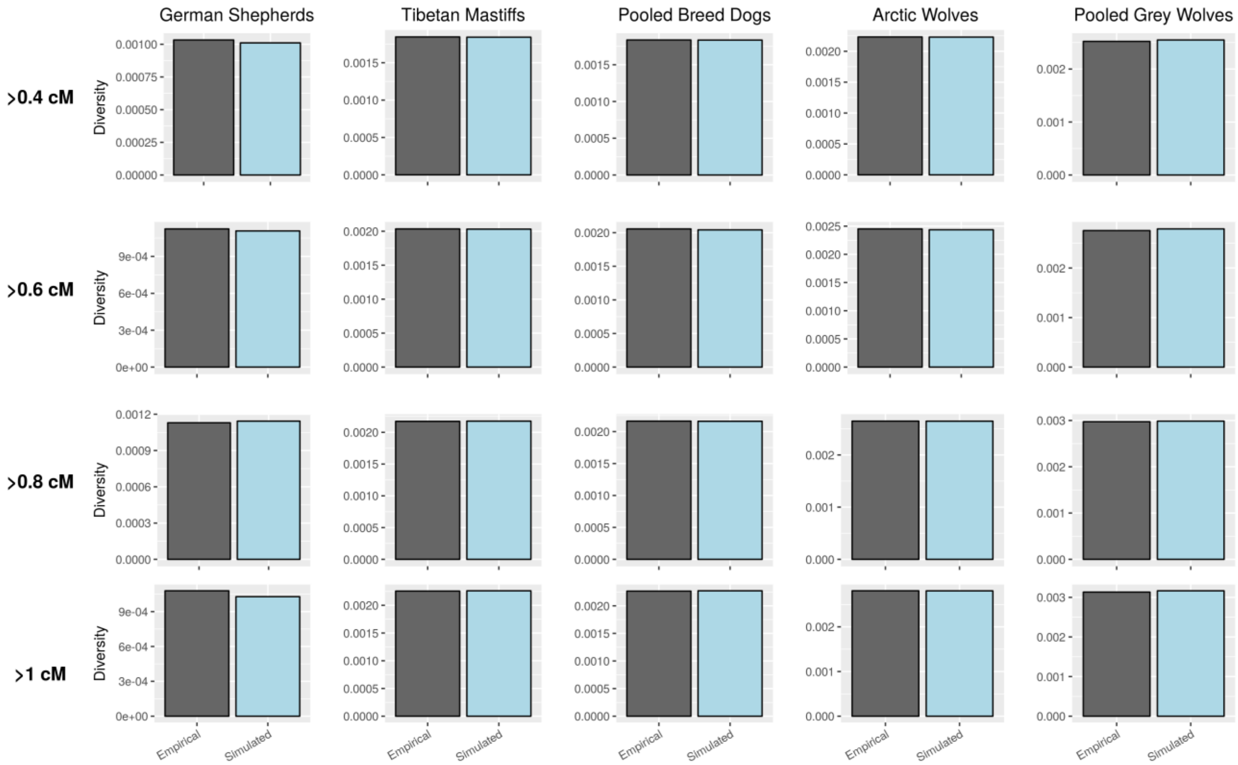


Figure 4.4. Genetic diversity (π) computed from the SFSs generated by using inferred demographic parameters (blue bars) compared with uncorrected empirical genetic diversity on the autosomes (grey bars) in multiple canid populations. Here, multiple thresholds were used to remove sites potentially linked to selected sites (>0.4 cM, >0.6 cM, >0.8 cM, and >1 cM). Note the excellent fit of the demographic models to the empirical data.

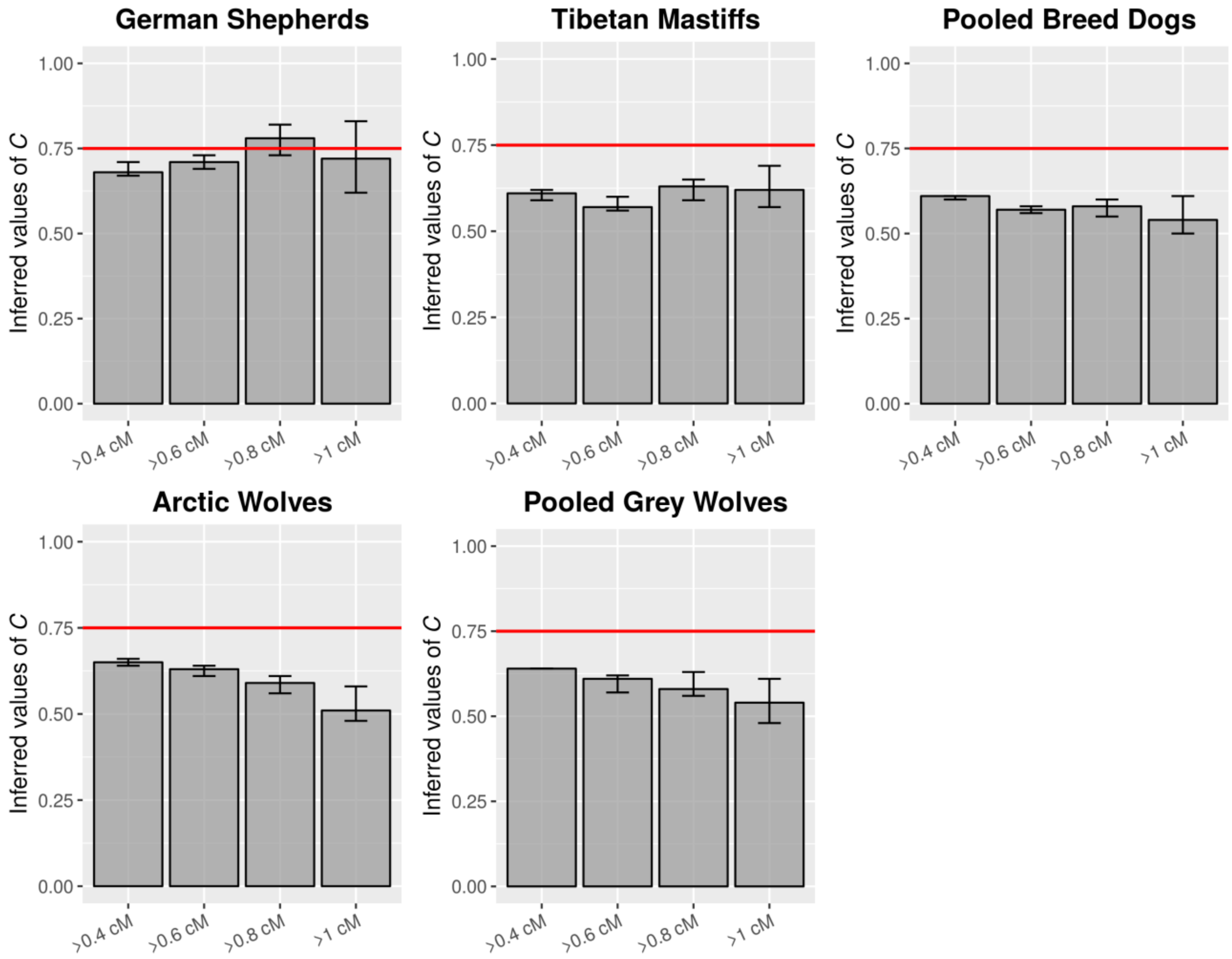


Figure 4.5. Effective population size estimates for multiple canid populations. Maximum likelihood estimates (MLEs) of the effective population size on the X chromosome relative to that of the autosomes ($C = N_X/N_A$) are shown for German Shepherds, Tibetan Mastiffs, Pooled Breed Dogs, Arctic Wolves and Pooled Grey Wolves with increasing distance from genes. Error bars denote approximate asymptotic 95% confidence intervals obtained as the parameter values within 2 log-likelihood unites of the MLE.

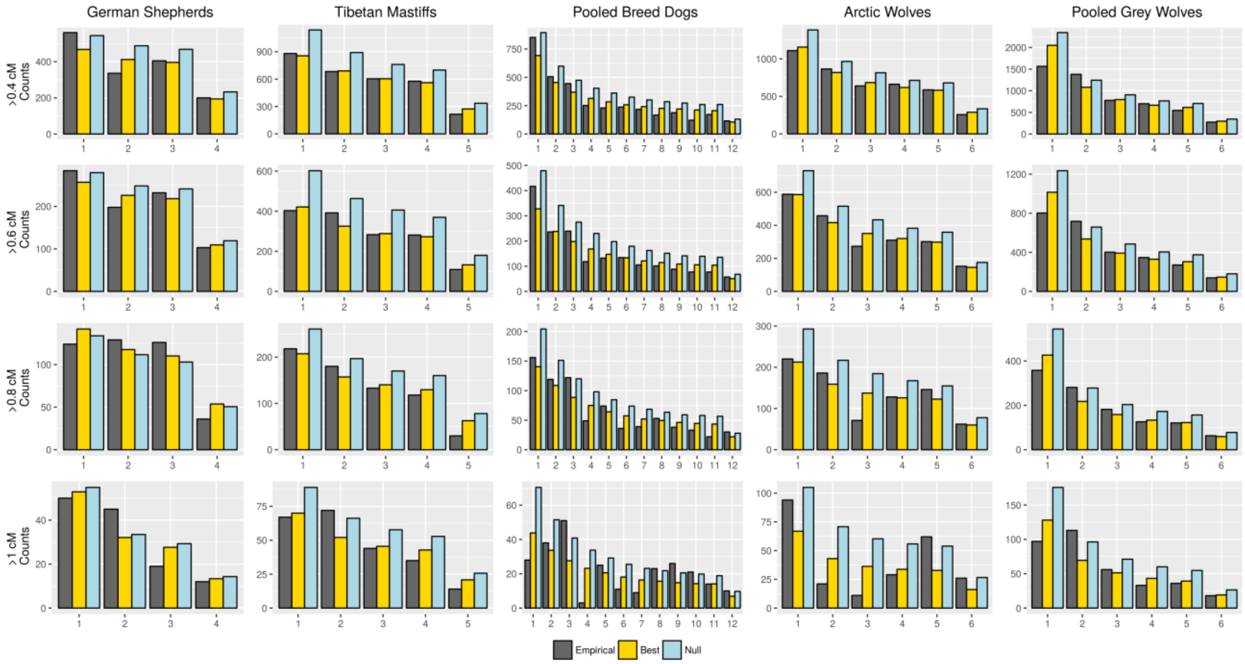
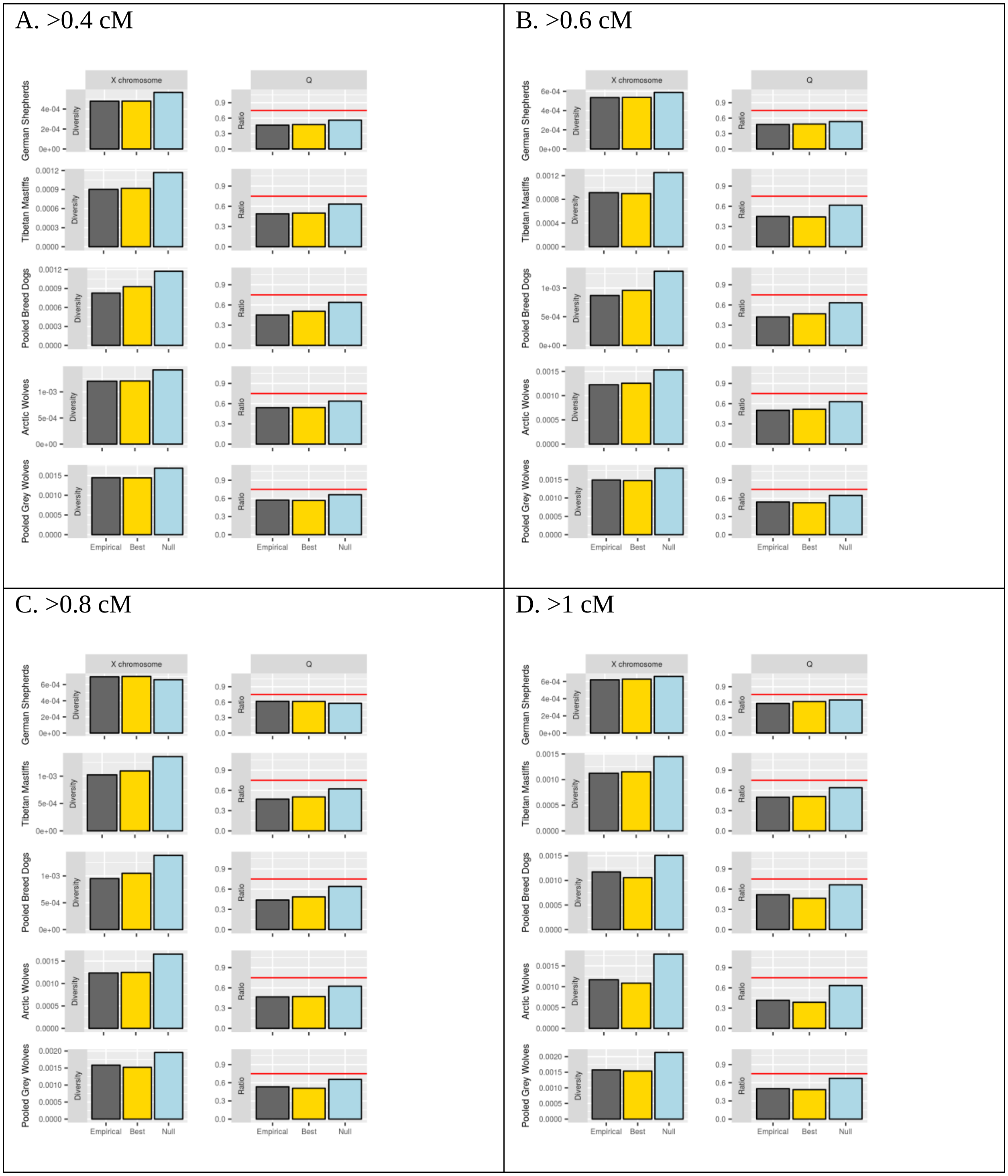


Figure 4.6. Empirical SFSs (grey bars) compared to the SFSs simulated with a null *C* value of 0.75 (blue bars) and a *C* value that yielded the highest likelihood (yellow bars) for multiple canid populations. Here, multiple thresholds were used to remove sites potentially linked to selected sites (>0.4 cM, >0.6 cM, >0.8 cM, and >1 cM). Note the superior fit of the maximum likelihood value of *C* (yellow) compared to *C*=0.75 (blue).



A. >0.4 cM

B. >0.6 cM

C. >0.8 cM

D. >1 cM

Figure 4.7. Empirical genetic diversity (π, grey bars) compared to diversity computed from SFSs simulated with a null *C* value of 0.75 (blue bars) and a *C* value that yielded the highest likelihood (yellow bars) for multiple canid populations. Here, multiple thresholds were used to remove sites potentially linked to selected sites (>0.4 cM, >0.6 cM, >0.8 cM, and >1 cM).

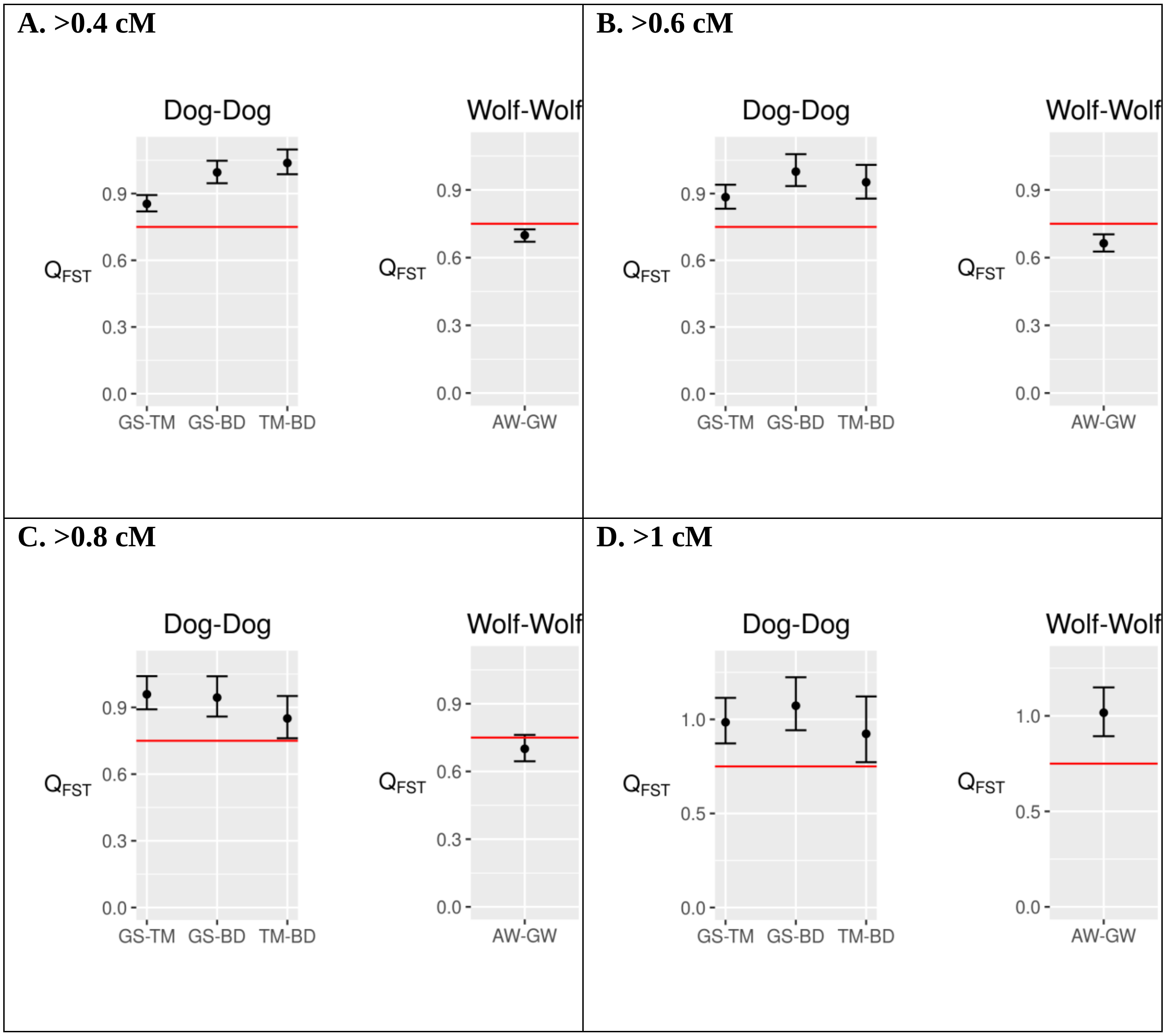


Figure 4.8. Sex biased demography on recent time scales. Estimates of the sex ratio for a pair of populations computed using F_{ST} (see Methods) using a threshold of >0.4 cM (A), >0.6 cM (B), >0.8 cM (C) and >1 cM (D) to remove sites putatively linked to selected sites. The horizontal red line denotes the null expectation of 0.75. Error bars denote 95% confidence intervals obtained through bootstrapping (see Methods). Abbreviations: GS (German Shepherds), TM (Tibetan Mastiffs), BD (Pooled Breed Dogs), AW (Arctic Wolves), GW (Pooled Grey Wolves).

4.7 Tables

Table 4.1. Data used in this study. Coverage for the German Shepherds, Tibetan Mastiffs, and Arctic Wolves was obtained from the software *qualimap* rounded to the nearest integer. Coverage for the Pooled Breed Dogs and Pooled Grey Wolves was obtained from Marsden et al. (2016).

Species	Population	In House ID	Coverage	Source ID	SRA	Breed or Location
Dogs	German Shepherds	GS1	16	DKD301	SRR1122359	
		GS3	16	DKS303	SRR1124304	
		GS4	16	DKS304	SRR1130247	
		GS5	15	DKS305	SRR1130350	
	Tibetan Mastiffs	TM1	17	DQZA81	SRR1138369	
		TM2	17	DQZA80	SRR1138368	
		TM3	17	DQZA33	SRR1138365	

		TM4	15	DQZA12	SRR1138361	
		TM5	14	DQZA06	SRR1138360	
	Pooled Breed Dogs	BD1	21	ddair_RS74411	SRS932161	Airedale
		BD2	16	ddbas_RS80704	SRS932158	Basenji
		BD3	16	ddbdt_RS86407	SRS932144	Border Terrier
		BD4	15	ddbrt_RS86399	SRS833812	Black Russian Terrier
		BD5	29	ddjrt_RS86400	SRS932151	Jack Russell Terrier
		BD6	21	ddjrt_RS86404	SRS932147	Jack Russell Terrier
		BD7	18	ddkbt_RS74408	SRS932164	Kerry Blue Terrier
		BD8	15	ddlab_RS86398	SRS932152	Labrador Retriever
		BD9	16	ddpwc_RS73323	SRS732550	Pembroke Welsh Corgi

		BD10	17	ddpwc_RS86409	SRS732551	Pembroke Welsh Corgi
		BD11	18	ddsct_RS86393	SRS932157	Scottish Terrier
		BD12	17	ddwhw_RS86397	SRS932153	West Highland White Terrier
Wolves	Arctic Wolves	AW12	37			Lat: 61.49534 Long: -105.433287
		AW13	43			Lat: 77.22 Long: -85.42
		AW16	38			Lat: 72.538139 Long: -110.534228
		AW18	43			Lat: 73.44 Long: -121.925
		AW19	39			Lat: 71.191287 Long: -85.508735
		AW20	41			Lat: 69.62 Long: -93.9
		GW1	15	gwcwx_XinXJ24		Xinjiang

	Pooled Grey Wolves	GW2	22	gwcwz_RKW3916		China
		GW3	18	gwibe_XXWIB98	SRS661495	Iberia
		GW4	20	gwirw_RKW3073	SRS661488	Iran
		GW5	18	gwprt_LOBO423	SRS661492	Portugal
		GW6	21	gwynp_RKW1547	SRS661496	Yellowstone National Park

Table 4.2. Dog-cat divergence, mutation rate, and estimates of male mutation bias

	X chromosome		Autosomes		α (male mutation bias)
Threshold	Divergence	Mutation rate	Divergence	Mutation rate	
No cM cutoff	0.162	4.49 X 10^{-9}	0.169	4.69 X 10^{-9}	1.293
>0.2 cM	0.159	4.42 X 10^{-9}	0.172	4.79 X 10^{-9}	1.600
>0.4 cM	0.157	4.35 X 10^{-9}	0.179	4.96 X 10^{-9}	2.169
>0.6 cM	0.159	4.42 X 10^{-9}	0.185	5.14 X 10^{-9}	2.445
>0.8 cM	0.166	4.61 X 10^{-9}	0.191	5.31 X 10^{-9}	2.335
>1 cM	0.174	4.82 X 10^{-9}	0.194	5.38 X 10^{-9}	1.900

Table 4.3. Linked selection is stronger on the X chromosome than on the autosomes

Population	$X_{0.4\ cM}/X_{no\ cM\ cutoff}$	$A_{0.4\ cM}/A_{no\ cM\ cutoff}$
German Shepherds	2.155	1.499
Tibetan Mastiffs	2.132	1.428
Pooled Breed Dogs	1.921	1.414
Arctic Wolves	2.321	1.405
Pooled Grey Wolves	1.820	1.381

Table 4.4. Number of sites and number of variants in each threshold that remained after removing neutral sites potentially linked to selected sites. We tabulated the number of sites from the BED files used to compute the statistics in our analyses. We intersected regions of the genome that are putatively neutral (i.e. excluding genic and conserved sites, and removing sites whose genetic distance is less than a threshold), high quality, and alignable between dog and cat. Similarly, for the number of variants, these variants are found in regions of the genome that are putatively neutral.

Threshold	Number of sites		Number of variants	
	X chromosome	Autosomes	X chromosome	Autosomes
No cM cutoff	35,643,675	822,468,024	323,495	9,728,949
>0.2 cM	3,275,430	67,760,411	50,734	967,762
>0.4 cM	1,126,087	17,844,986	22,055	284,222
>0.6 cM	554,218	6,680,919	11,522	117,436
>0.8 cM	219,129	2,767,274	5,532	53,149
>1 cM	68,831	1,353,692	2,461	27,742

Table 4.5. Demographic parameters that best fit the autosomal site frequency spectrum

Population	Threshold	NANC	NBOT	TBOT	NCUR	BOTDUR
German Shepherds	>0.4 cM	196,509	413	5,370	16,235	95
	>0.6 cM	225,733	10,321	10,919	16,467	80
	>0.8 cM	183,196	24,905	14,513	29,907	95
	>1 cM	129,203	429	12,750	128,490	80
Tibetan Mastiffs	>0.4 cM	232,750	973	990	38,581	95
	>0.6 cM	25,3450	855	3,441	69,086	90
	>0.8 cM	265,327	1,089	6,341	70,543	95
	>1 cM	269,344	804	5,092	113,518	95
Pooled Breed Dogs	>0.4 cM	211,231	24,472	9,325	71,007	100
	>0.6 cM	22,0051	1,948	807	76,739	100
	>0.8 cM	223,743	1,851	49	50,760	90
	>1 cM	232,926	2,785	308	26,283	110
Arctic Wolves	>0.4 cM	274,829	1,591	7,073	100,099	105
	>0.6 cM	28,9198	1,098	4,811	139,307	95
	>0.8 cM	308,722	1,202	4,938	78,489	95
	>1 cM	347,330	712	17,798	154,684	100
Pooled Grey Wolves	>0.4 cM	280,098	77,382	10,090	370,540	100
	>0.6 cM	296,003	78,930	10,921	376,718	105
	>0.8 cM	304,071	79,003	13,571	453,092	95
	>1 cM	322,149	65,689	7,988	368,592	95

Table 4.6. **SNP count Poisson log-likelihoods comparing the fit between the best fit demographic model to the observed SFS on the autosomes**

German Shepherds			
Threshold	**Model**	**Poisson LL**	**Δ LL (Model – Data)**
>0.4 cM	Data to Data	417,038.4	0
	Best fit demographic model	416,954.7	-83.7
>0.6 cM	Data to Data	150,103.2	0
	Best fit demographic model	149,819.8	-283.4
>0.8 cM	Data to Data	56,509.68	0
	Best fit demographic model	56,198.87	-310.81
>1 cM	Data to Data	25,269.04	0
	Best fit demographic model	25,104.24	-164.8

Tibetan Mastiffs			
Threshold	**Model**	**Poisson LL**	**Δ LL (Model – Data)**
>0.4 cM	Data to Data	868,148.7	0
	Best fit demographic model	868,144.7	-4
>0.6 cM	Data to Data	322,729.2	0
	Best fit demographic model	322,723.6	-5.6
>0.8 cM	Data to Data	127,926.2	0

	Best fit demographic model	127,923.8	-2.4
>1 cM	Data to Data	59,175.79	0
	Best fit demographic model	59,175.38	-0.41

Pooled Breed Dogs

Threshold	**Model**	**Poisson LL**	**Δ LL (Model – Data)**
>0.4 cM	Data to Data	956,019.4	0
	Best fit demographic model	955,906.6	-112.8
>0.6 cM	Data to Data	361,320.8	0
	Best fit demographic model	361,228.5	-92.3
>0.8 cM	Data to Data	141,427.9	0
	Best fit demographic model	141,395.2	-32.7
>1 cM	Data to Data	65,644.59	0
	Best fit demographic model	65,603.31	-41.28

Arctic Wolves

Threshold	**Model**	**Poisson LL**	**Δ LL (Model – Data)**
>0.4 cM	Data to Data	1,110,073	0
	Best fit demographic model	1,110,054	-19
>0.6 cM	Data to Data	411,281.3	0

	Best fit demographic model	411,265.3	-16
>0.8 cM	Data to Data	163,246.9	0
	Best fit demographic model	163,243.6	-3.3
>1 cM	Data to Data	76,600.19	0
	Best fit demographic model	76,591.97	-8.22

Pooled Grey Wolves			
Threshold	**Model**	**Poisson LL**	**Δ LL (Model – Data)**
>0.4 cM	Data to Data	1,404,387	0
	Best fit demographic model	1,404,239	-148
>0.6 cM	Data to Data	520,155	0
	Best fit demographic model	520,085	-70
>0.8 cM	Data to Data	210,418.7	0
	Best fit demographic model	210,399.2	-19.5
>1 cM	Data to Data	98,597.3	0
	Best fit demographic model	98,578.51	-18.79

Table 4.7. Likelihood ratio tests comparing models of sex-biased demography in multiple canid populations. Likelihood ratio tests of the amount of sex-biased demography are shown when removing any sites whose genetic distance to the nearest genes is less than 0.4 cM, 0.6 cM, 0.8 cM, and 1 cM.

Threshold: >0.4 cM				
Population	***C***	**Log-likelihood**	**Likelihood ratio test**	**p-value**
German Shepherds	Null (*C* = 0.75)	7460.957	34.779	3.69 X 10^{-9}
	Best (*C* = 0.68)	7478.347		
Tibetan Mastiffs	Null (*C* = 0.75)	16050.84	208.972	2.30 X 10^{-47}
	Best (*C* = 0.61)	16155.32		
Pooled Breed Dogs	Null (*C* = 0.75)	16875.26	249.627	3.13 X 10^{-56}
	Best (*C* = 0.61)	17000.07		
Arctic Wolves	Null (*C* = 0.75)	23035.46	133.341	7.62 X 10^{-31}

	Best (*C* = 0.65)	23102.13		
Pooled Grey Wolves	Null (*C* = 0.75)	30767.69	188.719	6.05 X 10^{-43}
	Best (*C* = 0.64)	30862.05		

Threshold: >0.6 cM				
Population	***C***	**Log-likelihood**	**Likelihood ratio test**	**p-value**
German Shepherds	Null (*C* = 0.75)	3,568.57	6.037	0.014
	Best (*C* = 0.68)	3,571.58		
Tibetan Mastiffs	Null (*C* = 0.75)	6,892.14	164.747	$1.04 \text{ X } 10^{-37}$
	Best (*C* = 0.61)	6,974.51		
Pooled Breed Dogs	Null (*C* = 0.75)	7,303.47	209.331	$1.92 \text{ X } 10^{-47}$
	Best (*C* = 0.61)	7,408.14		
Arctic Wolves	Null (*C* = 0.75)	10,198.7	109.011	$1.61 \text{ X } 10^{-25}$
	Best (*C* = 0.65)	10,253.2		
Pooled Grey Wolves	Null (*C* = 0.75)	13,898.57	136.982	$1.22 \text{ X } 10^{-31}$

	Best (*C* = 0.64)	13,967.06		

Threshold: >0.8 cM				
Population	***C***	**Log-likelihood**	**Likelihood ratio test**	**p-value**
German Shepherds	Null (*C* = 0.75)	1541.624	0.602	0.438
	Best (*C* = 0.78)	1541.925		
Tibetan Mastiffs	Null (*C* = 0.75)	2710.71	42.255	8.01 X 10^{-11}
	Best (*C* = 0.63)	2731.837		
Pooled Breed Dogs	Null (*C* = 0.75)	2516.035	82.291	1.17 X 10^{-19}
	Best (*C* = 0.58)	2557.18		
Arctic Wolves	Null (*C* = 0.75)	3187.933	81.815	1.49 X 10^{-19}
	Best (*C* = 0.59)	3228.84		

Pooled Grey Wolves	Null (*C* = 0.75)	4906.004	68.493	1.27 X 10^{-16}
	Best (*C* = 0.58)	4940.25		

Threshold: >1 cM				
Population	***C***	**Log-likelihood**	**Likelihood ratio test**	**p-value**
German Shepherds	Null (*C* = 0.75)	322.378	0.635	0.426
	Best (*C* = 0.72)	322.696		
Tibetan Mastiffs	Null (*C* = 0.75)	673.92	12.178	4.84 X 10^{-4}
	Best (*C* = 0.62)	680.009		
Pooled Breed Dogs	Null (*C* = 0.75)	527.835	38.649	5.07 X 10^{-10}
	Best (*C* = 0.54)	547.16		
Arctic Wolves	Null (*C* = 0.75)	648.812	53.136	3.11 X 10^{-13}

	Best (*C* = 0.51)	675.38		
Pooled Grey Wolves	Null (*C* = 0.75)	1110.162	38.736	4.85 X 10^{-10}
	Best (*C* = 0.54)	1129.53		

Table 4.8. Individuals used in Q_{FST} analyses

Population	In House ID
German Shepherds	GS1
	GS2
	GS3
	GS4
Tibetan Mastiffs	TM1
	TM2
	TM3
	TM4
Pooled Breed Dogs	BD1
	BD2
	BD3
	BD4
Arctic Wolves	AW1
	AW2
	AW3
	AW4
Pooled Grey Wolves	GW1
	GW2
	GW3

CHAPTER 5

Inference of the mutational target size supports the omnigenic model for complex traits

5.1 Abstract

Genetic variants associated with complex traits contain a wealth of information about the architecture of the trait as well as the evolutionary forces impacting the particular phenotype. However, utilizing these data for evolutionary inferences is challenging due to the complex ascertainment scheme and limited power to detect rare variants inherent in genome wide association studies (GWAS). Here we circumvent this problem by combining explicit realistic population genetic models of demography and selection together with quantitative genetic models of GWAS data. Specifically, we develop an Approximate Bayesian Computational framework to estimate the number of sites in the genome, M, that, if mutated, would give rise to a variant affecting the phenotype. Our method also infers the effect of purifying selection by estimating the coupling parameter between a mutation's effect on the trait and its effect on fitness, sometimes called τ. Our approach models the limited power of GWAS for detecting rare variants, thus improving the accuracy of the method. We applied our new method to 21 quantitative traits using publicly available GWAS summary statistics from the UKBiobank. Surprisingly, we found that the coupling parameter between a mutation's effect on the trait and its effect on fitness is similar across 11 traits examined (around 0.25), indicating that many complex traits, including those not typically thought to be associated with reproductive fitness, are affected by selection. For 10 out of 21 traits, the mutational target size is on the order of tens of megabases (Mb). For example, M is inferred to be around 25Mb for systolic blood pressure, around 50Mb for body mass index, and around 95Mb for height. Interestingly, 5 traits all have target sizes of around 25-30 Mb. Both height and forced vital capacity (FVC) have large target

sizes of greater than 80Mb. The finding that disparate traits show similar target sizes suggests that the same peripheral genes could be affecting many traits. We propose that this finding combined with the large mutational target sizes inferred for all traits examined supports the omnigenic model.

5.2 Introduction

The genetic architecture of a complex trait is defined as the number, frequency, and effect size of the trait-associated variants combined with the interactions among these factors (Timpson et al., 2018). Understanding the genetic architecture is essential to unraveling the genetic basis of human traits and diseases. Studying genetic architecture has been made possible by data from genome-wide association studies (GWAS). GWAS have discovered tens of thousands of susceptibility variants that are associated with complex traits and helped to further our understanding of the genetic architecture (Altshuler et al., 2008; MacArthur et al., 2017; Stranger et al., 2011). For example, most trait-associated variants discovered from GWAS are common (i.e. have a minor allele frequency > 0.2; Figure 5.1A). However, it does not necessarily mean that common variants explain all of the genetic basis of the trait. Rather, GWAS has been done using limited sample sizes and is therefore underpowered at detecting associations with rare variants (Visscher et al., 2017). An additional factor is that rare variants are not assayed or imputed as well as common variants using standard genotyping arrays (Visscher et al., 2017). Data from GWAS also show that the effect size is negatively correlated with allele frequency in many traits (Figure 5.1B) (Park et al., 2011; The UK10K Consortium, 2015). However, in principle, the allele frequency of a variant should not be related to the effect size, unless there is a relationship between an allele’s effect on the trait and its effect on reproductive fitness. If a variant has a large effect on a disease and the disease either directly or indirectly affects reproductive fitness, the variant is evolutionarily deleterious and is subject to purifying selection, keeping it at low frequency (Gibson, 2011). Eyre-Walker proposed to model the relationship between a variant's effect on the trait and its effect on fitness (i.e. the selection coefficient) using a parameter called τ (Eyre-Walker, 2010). When $\tau = 0$, the variant's effect on the trait is independent of its selection coefficient and consequently there is no relationship between effect size and allele frequency (Figure 5.2, orange lines). When $\tau = 0.5$, there is a positive relationship between effect size and selection, resulting in a negative correlation between effect size and allele frequency (Figure 5.2, green lines). Recent work has shown support for a non-independent relationship between effect size and selection (i.e. $\tau \geq 0$) in many traits (Schoech et al., 2017). These results suggest negative selection has played an important role in shaping the genetic architecture of complex traits (Schoech et al., 2017; Zeng et al., 2018).

An aspect of the genetic architecture that is understudied is how many causal variants are contributing to a trait. Here we define a causal variant as any variant in the genome that has a non-zero effect size on the trait. Because not all causal variants can be detected in a GWAS, the observed GWAS hits are drawn from a larger pool of causal variants. Furthermore, not every position in the genome that could give rise to a causal variant necessarily contains a causal variant, because some sites have not been mutated in the population. Thus, the causal variants are drawn from a pool of sites in the genome that could be mutated to give rise to variants affecting a trait. We call this parameter the mutational target size, M. Agarwala et al. (2013) developed a population genetic framework to understand the genetic architecture of type 2 diabetes (Agarwala et al., 2013). Using a model with a mutational target size of 1.25Mb, they found that the simulated disease model is consistent with the empirical data (Agarwala et al., 2013). Simons et al. (2018) also developed a model to explore how genetic architecture is affected by evolutionary processes (Simons et al., 2018). In applying their method to height and body mass index, they found that the mutational target size for height and body mass index to be 5Mb and 1Mb, respectively. Since these studies have only examined a small number of trait traits, how the mutational target size differs between traits remained to be explored. In addition, Simons et al. (2018) did not explore the relationship between effect size and selection in their analyses. Further, while their model is based on a multivariate stabilizing selection model, it does not include realistic distributions of selection coefficients, mutations, linkage disequilibrium, or demography (Simons et al., 2018).

Here, we developed a population genetic model for the genetic architecture of complex trait and infer the mutational target size and the coupling parameter between a variant's effect on the trait effect and its effect on fitness. We improved on existing methods by using summary statistics from GWAS as our empirical data. We call our method to infer the genetic architecture InGeAr. InGeAr accounts for the incomplete statistical power of GWAS by modelling the power to detect a variant to match the number of GWAS variants to the empirical data. We first applied InGeAr to height and found a large mutational target size for height, 95Mb. We also examined how the number of causal variants, GWAS hits, and variants remained to be detected differ in terms of their effect sizes, selection coefficients, and additive genetic variance. We then applied InGeAr to multiple complex human diseases using publicly available GWAS summary statistics from the UKBiobank. We found that the mutational target size varies from trait to trait but is consistent on the order of ten of megabases. Curiously, we found that the coupling parameter between effect size and selection is similar across traits. We suggest that our results provide support for the omnigenic model of complex traits. Understanding the mutational target size, the number of causal variants for each trait, and the relationship between effect size and selection will lead to improved knowledge of the underlying biology and suggest future strategies for mapping further risk variants for these traits.

5.3 Results

GWAS simulation

To simulate a genome-wide association study, we first simulate causal variants for a given mutational target size (Figure 5.3). We use the forward-in-time simulation software SLiM and specify an evolutionary model with realistic parameters for the demography, selection, mutation rate, and recombination rate (Haller and Messer, 2017). We use a European demographic history from the Gravel model and a distribution of fitness effects for non-coding variants from Torgerson et al. (Gravel et al., 2011; Torgerson et al., 2009). We model the effect size of each variant following Equation 1 of Eyre-Walker (2010), which is a function of the selection coefficient s and τ (Eyre-Walker, 2010):

$$\alpha_i = \delta s_i^{\tau}(1+\varepsilon),$$

where α_i is the effect size of variant on the trait i, s_i is the selection coefficient which is an output from the SLiM simulations, τ is the coupling parameter between the effect size and selection coefficient, ε is the error term which is drawn from a normal distribution with mean 0 and a standard deviation of 0.5, and δ randomly takes a value of -1 and 1. To achieve the desired heritability for each trait and to match the effect size from the simulation to the empirical data, we adjust the effect size based on the heritability as was done in Lohmueller (2014): $\alpha_{\text{adjusted}} = \alpha C$ where C is computed from the heritability, h^2 (see Methods) (Lohmueller, 2014). To recapitulate the fact that GWAS is underpowered to detect all causal variants, we employ a rejection sampling scheme where SNPs are retained in proportion to their power to be detected. Specifically, we calculate the power to detect a variant based on its effect size, heritability, allele frequency, and sample size (see Methods). We use the same sample size for each trait as in the UKBiobank study (around 500,000 individuals). We then draw a random value between 0 and 1. If this randomly drawn value is less than the calculated power to detect a variant, we consider this variant a GWAS hit.

The mutational target size determines the number of causal variants

To understand how the mutational target size affects the number of causal variants, we simulated data using the mutational target sizes of 100kb, 1Mb, and 10Mb (Figure 5.4). As expected, there are more causal variants when the mutational target size is larger (Figure 5.4A). We also observed that most causal variants segregate at very low frequency (<0.5%) (Figure 5.4A), reflecting the fact that most variants in human populations are at low frequency (Gravel et al., 2011). For a constant heritability, as M increases, the effects on the trait are distributed across more variants. Thus, the effect size of a causal variant is smaller for a larger value of M as compared to a smaller value of M (Figure 5.4B).

Both M and τ determine the number of GWAS hits and their effect sizes

To understand whether and how the number of GWAS hits and the effect size are determined by the mutational target size (M) and the coupling parameter between a variant's effect on the trait and its selection coefficient (τ) at different allele frequencies, we examined the number of GWAS hits and their effect sizes stratified by allele frequencies from the simulation generated using different values of M and τ (Figure 5.5). For all analyses, we used the sample size of $N = 50000$ individuals to be comparable to that from the UKBiobank.

When there is no relationship between effect size and selection coefficient ($\tau = 0$), the power to detect a variant is lowest for rare variants (<0.5%) (Figure 5.5, first row, first column, orange line). As M increases, the power to detect these rare variants (<0.5%) decreases further

(Figure 5.5, first column, orange lines). As M becomes very large (i.e. 10Mb), there is little power to detect even variants segregating at frequency <5%. Since there are more causal variants for larger value of M, the effects on the trait is distributed to more variants when M increases (Figure 5.4B). Therefore, as M increases, causal variants segregating at low frequency have tiny effect size and GWAS are underpowered to detect them. When M is small (i.e. 100kb), rare variants can be detected because the effect size of each causal variant is, on average, higher than those for larger values of M.

As τ increases, power to detect rare variant in GWAS also increases. For instance, at the mutational target size of 1Mb, when $\tau = 0$, there is little power to detect rare variants. For a greater value of τ ($\tau = 0.2$), however, power is higher (Figure 5.5, second row, first column, comparing orange line and green line). Therefore, most GWAS hits are common when $\tau = 0$ whereas most GWAS hits are rare when $\tau = 0.2$ (Figure 5.5, second row, comparing orange line and green line). For $\tau = 0.2$ and a larger mutational target size ($M = 10$Mb), there is little power to detect rare variants, resulting in most GWAS hits to be common (Figure 5.5, third row, green line). At the same time, when $M = 10$Mb, when τ is larger ($\tau = 0.6$), the majority of GWAS hits are rare (Figure 5.5, second and third rows, blue lines).

As expected, when $\tau = 0$, there is no relationship between effect size and allele frequency for all values of M (Figure 5.5, third column, orange lines). Similarly, when $\tau > 0$, there is a negative relationship between effect size and allele frequency. The negative correlation is stronger for larger value of τ (Figure 5.5, third column, green and blue lines). This is due to the coupling between a mutation's effect on the trait and its effect on fitness (Eyre-Walker, 2010).

These observations suggest that M and τ can affect different summary statistics of the GWAS data differently. For instance, for a given value of M, such as $M = 10$Mb, with a τ value of 0.2, most GWAS hits are common. Whereas with a τ value of 0.6, most GWAS hits are rare (Figure 5.5, second column, second and third rows, comparing green and blue lines). Similarly, for the same value of τ such as τ of 0.2, M of 100kb resulted in most GWAS hits being rare (Figure 5.5, first row, second column, green line) whereas M of 10Mb resulted in most GWAS hits to be common (Figure 5.5, third row, second column, green line).

Using summary statistics from UKBiobank GWAS, we observed that there is a positive relationship between the number of GWAS hits and allele frequency and a negative relationship between effect size and allele frequency (Figure 5.1). We observed that there are values of M and τ that can result in a positive relationship between the number of GWAS hits and allele frequency, but not the negative relationship between effect size and allele frequency (Figure 5.5). For example, for $M = 1\text{Mb}$ and $\tau = 0$, there is a positive correlation between the number of GWAS hits and allele frequency (Figure 5.5, second row, second column, orange line), but there is no relationship between effect size and allele frequency (Figure 5.5, second row, third column, orange line). In addition, we found that in some scenarios, one summary statistic shows a similar pattern for two pairs of M and τ but the other statistic can distinguish between these pairs. For example, when $\tau = 0$ and $M = 100\text{kb}$ or $\tau = 0.1$ and $M = 100\text{kb}$, the expected number of GWAS hits is similar between these two parameter combinations (Figure 5.6A). However, the effect size stratified by allele frequency differs for these two pairs of M and τ values. Similarly, when $M = 400\text{kb}$ and $\tau = 0.4$ or $M = 1\text{Mb}$ and $\tau = 0.4$, the effect size stratified by allele frequency is similar between these two pairs (Figure 5.6B). However, the expected number of hits can distinguish between these two parameter combinations. Specifically, there are more rare variants detected when $M = 1\text{Mb}$. These observations led us to perform inference under a model that includes both M and τ. In addition, we use both the number of GWAS hits and the effect size stratified by allele frequency as summary statistics of the observed GWAS data.

Overview of the Inference of Genetic Architecture (InGeAr) Method

We infer the mutational target size, M and the coupling parameter between effect size and selection, τ in an Approximate Bayesian Computation framework. We call our Inference of Genetic Architecture method InGeAr. For each value of M and τ that are drawn from prior distributions, we first simulate GWAS as described above. We then apply a rejection sampling algorithm to decide whether to accept or reject M and τ (Figure 5.3). Since the simulations described above demonstrated that both the number of GWAS hits at different allele frequencies and the effect sizes at different allele frequencies are complementary statistics, we employ them both as summary statistics. Specifically, we divide the GWAS hits into six allele frequency bins: <0.5%, 0.5-1%, 1-5%, 5-10%, 10-20%, >20%. In the first rejection step, we compute the scaled difference in effect size between the simulated GWAS and the empirical GWAS across all 6 allele frequency bins:

$$\text{difference} = \sum_{\text{bin}_i=1}^{6} \frac{|\alpha_i^{\text{empirical}} - \alpha_i^{\text{simulated}}|}{\alpha_i^{\text{empirical}}}$$

For each allele frequency bin, the difference in effect size between the empirical and simulated effect sizes divided by the empirical effect size represents how far away the simulated effect size is from the empirical one. For example, if this value is 0.1, it means that the simulated effect size is within 10% of the empirical effect size. Since there are six allele frequency bins, we choose 0.6 as the threshold to accept or reject a pair of M and τ values. We repeat the procedure until 10,000 pairs of M and τ are accepted. We then compute the sum of squared difference between the observed the number of GWAS hits in each frequency bin compared to the values from our simulations. We retain the 1,000 pairs of tau and M whose sum of squared difference is in the bottom 10% of the distribution (Figure 5.3). These 1,000 pairs of M and τ form the posterior distribution.

Application of InGeAr to infer M and τ for height

We applied our method to infer the mutational target size and the coupling parameter between effect sizes and selection coefficients to height because height has been of immense interest in medical genetics. We utilized GWAS summary statistics and independent GWAS hits to remove the effect of linkage disequilibrium from Kichaev et al. (Kichaev et al., 2017).

We drew a value of M and a value of τ from uniform distributions (Figure 5.7). We found that the median inferred value of M is 95Mb (95% credible interval of 54Mb-131Mb) and the inferred value of τ is 0.27 (95% credible interval of 0.22-0.31) (Figure 5.8A, 5.8B). We also observed a positive correlation between M and τ (Figure 5.8C). This positive correlation between M and τ suggests that previous methods that only inferred τ while ignoring M could be problematic as the parameters depend on each other. To assess whether the inferred values of M and τ can recapitulate the empirical data, we drew M and τ from the posterior distributions and simulated 10,000 replicates. We found that the simulated data from the posterior tau and M matched reasonably well with the empirical data (Figure 5.9).

We next use our model to understand how the currently identified GWAS hits differ from the causal variants that remain to be discovered. To do this, we examined the variants from each of the 10,000 simulations where M and τ are drawn from the posterior distributions as above (Figure 5.10). Interestingly, we observed that across all frequency bins, most of the causal variants have not been detected by GWAS, leaving most causal variants remaining to be discovered (Figure 5.10A). We hypothesized that many of these remaining undiscovered causal variants likely have small effect sizes. To understand the property of the effect size of causal variants, GWAS hits, and remaining undiscovered causal variants, we calculated the number of variants in each category stratified by effect size ranging from small ($\alpha \leq 0.001$) to medium ($0.001 < \alpha \leq 0.01$) to large ($\alpha > 0.01$). Consistent with this expectation, most causal variants that remain to be discovered have small and medium effect (Figure 5.10B). GWAS has done well at finding variants with larger effects as we found that >95% of GWAS hits have large effect size (Figure 5.10B). However, about 60% of causal variants have weak effect and about 25% of causal variants have medium effect (Figure 5.10B). In addition, we observed that GWAS hits are more evolutionarily deleterious compared to the entire set of causal variants (Figure 5.10C). While about 85% of causal variants and variants remaining to be discovered are weakly deleterious ($|s| < 10^{-4}$), about 70% of GWAS hits have intermediate strength of selection ($10^{-4} \leq |s| < 10^{-2}$) (Figure 5.10C).

Much attention in complex trait genetics has been devoted determining which categories of variants can account for much of the heritability of traits and why previously identified GWAS hits do not account for all of this heritability (Manolio et al., 2009). To understand which category of variants in terms of selection can explain more of the additive genetic variance, we calculated the proportion of additive genetic variance explained stratified by strength of selection. We found that most of the additive genetic variance can be explained by causal variants with intermediate selection strength (i.e. $10^{-4} \leq |s| < 10^{-2}$) (Figure 5.10D) .

So far, we have considered all causal variants in our inference for the mutational target size. To explore how the inferred mutational target size would differ when only variants with large effect size are considered, we modified our inference procedure at the rejection step and only compared variants with large effect size. Since the cutoff value to determine whether a variant has a large effect size is arbitrary, we examined the distribution of effect size for height and chose two values: 0.02 and 0.05 (Figure 5.11). In both the empirical and simulated GWAS, we only considered causal variants with effect size greater than 0.02 or 0.05. Since there are fewer GWAS hits when considering variants with large effect size, we observed that the inferred mutational target size is smaller (Table 5.1). This is unsurprising, as the mutational target size inferred here is that which would give rise to a causal variant of strong effect, which by definition will be smaller than the mutational target that will give rise to a variant with any non-zero effect. The coupling parameter between trait's effect and selection remains approximately the same, suggesting some robustness of this parameter to different thresholds of the effect sizes of GWAS variants (Table 5.1).

The role of pleiotropy on the genetic architecture of height

The Eyre-Walker model (2010) that we used in our inference quantifies the coupling between a mutation's effect on the trait and its effect on fitness, τ. This coupling parameter τ is assumed to be the same for all of the variants, yielding the same average relationship between a mutation's effect on fitness and its effect on the trait for all causal variants. In principle, this may not be the case. Some causal variants may affect the trait in proportion to how they affect fitness, but other variants may have discordant effects. This effect has been captured in a model of pleiotropy proposed by Uricchio et al. (Uricchio et al., 2016). In this model, pleiotropy is capture through ρ where $\rho < 1$ indicates that more variants have independent effects on fitness and the trait. Following Uricchio et al. (2016), the effect size α of variant i is equal to s^{τ} as before with probability ρ. Otherwise, the effect size is equal to s_r^{τ} where s_r is the selection coefficient from another variant picked at random. We modified InGeAr to also infer for ρ by sampling a value of ρ from a prior distribution and performing the simulation of GWAS and rejection algorithm as described above. We found that the posterior distribution of ρ includes 1 (median = 0.982; 95% credible interval of 0.920-0.999) (Figure 5.12, Table 5.2). This suggests that many variants have proportional effects on fitness and the trait. Given that $\rho \gg 0$, it is unsurprising that the posterior distribution of τ is similar between the inference with and without ρ (Figure 5.12, Table 5.2). Interestingly, we found that the inference including rho leads to a modest decrease in the mutational target size (from 95Mb to 77Mb) (Figure 5.12, Table 5.2). Overall, this finding suggests that if many of the causal variants for height are pleiotropic, their effects on height are still largely proportion to their effects on reproductive fitness.

Application of inGeaR to other traits

We applied InGeAr to 20 other quantitative Biobank traits that were also analyzed by Kichaev et al. (Kichaev et al., 2017). We found that applying the same acceptance criteria as outlined above, in 10 out of 20 traits, there are values of M and τ that passed the acceptance criteria. For these traits, we found that the mutational target size differs across traits (Figure 5.13A, Table 5.3). It ranges from 3Mb for age at menarche to 95Mb for height (Figure 5.13A, Table 5.3). Most traits have a relatively large mutational target size, on the orders of tens of megabases (Figure 5.13A, Table 5.3). Curiously, the coupling parameter between effect size and selection is similar (median value of τ ranges from 0.214 to 0.301) (Figure 5.13B, Table 5.3).

As mentioned above, for 10 out of 20 traits examined here, setting the threshold to be 0.6 in the step to accept or reject M and τ when comparing the effect size between the empirical data and simulated data works well. However, we found that for the other 10 traits, the total difference between the empirical effect size and simulated effect size is greater than 0.6. As a result, no pairs of M and τ were accepted. These results suggest that there is no value of M and τ where the fit to the empirical data in the effect size is close enough across all bins of allele frequency. When we increased the threshold to be high enough, the simulation typically overestimates the effect size in intermediate and common frequency bins but underestimates the number of GWAS variants in intermediate bins (Figure 5.14). These observations could indicate that the model is not appropriate for these traits, and one may need to incorporate other forces such as positive selection.

5.4 Discussion

In this work, we developed a method called InGeAr to infer the mutational target size (M) and the extent to which a mutation's effect on a trait is coupled with its selection coefficient (τ) for complex traits. An advantage of our method over existing work is that InGeAr jointly infers for both M and τ while existing approaches only estimate τ (Schoech et al., 2017). Since M and τ are highly correlated, inferring for one but not the other could be problematic (Figure 5.8C). Another advantage of InGeAr is that it only requires summary statistics from GWAS data, which are easier to obtain. Existing approaches to estimate τ require individual genotype data which could be more difficult to obtain (Schoech et al., 2017). Therefore, we anticipate that InGeAr can be easily applied to study other traits in humans or other species of interest.

We found that the mutational target sizes differ between traits (Figure 5.13A). Interestingly, except for age of menarche where the inferred mutational target size is around 3Mb, the mutational target sizes for the other traits are very large, on the orders of tens of megabases (Figure 5.13A). Curiously, Simons et al. estimated that the mutational target sizes for height and BMI are 5Mb and 1Mb, respectively (Simons et al., 2018). The mutational target sizes inferred from Simons et al. (2018) is an order of magnitude smaller than those inferred here. There are several important differences between InGeAr and the method develop in Simons et al. (2018) that could explain the discrepancies. First, Simons et al. used a different GWAS datasets for height and BMI. Specifically, Simons et al. (2018) used GWAS data for height from Wood et al. that consisted of 697 GWAS hits and GWAS data for BMI from Locke et al. that consisted of 97 GWAS hits for BMI (Locke et al., 2015; Wood et al., 2014). We used GWAS data from UKBiobank and further narrowed down the GWAS hits to contain only those that are independent to account for linkage disequilibrium (Kichaev et al., 2017). Our dataset consisted of 2452 and 965 GWAS hits for height and BMI, respectively. Since the sample size in the UKBiobank dataset is larger ($N = 500k$) and therefore there are more GWAS hits, we expect the mutational target size to be larger. Second, in our model, we used the distribution of fitness effect for noncoding variants as in Torgerson et al., resulting in most causal variants being weakly deleterious (Torgerson et al., 2009). On the other hand, Simons et al. (2018) assumed causal variants with stronger selection than in our study. Considering only variants with strong selection could lead to the inferred mutational target size to be smaller, as we inferred when restricting to large effect variants (Table 5.1). Importantly, it is not clear whether the model proposed by Simons et al. where many causal variants of the trait are under strong purifying selection is consistent with existing studies of the distribution of fitness effects across the genome. Third, Simons et al. (2018) conclude that there is substantial pleiotropy of causal variants and that this may reduce the mutational target size. Importantly, utilizing the model of pleiotropy, presented by Uricchio et al 2016, we found little evidence for substantial pleiotropy for height ($\rho \approx 1$, Figure 5.12, Table 5.2). This difference is likely due to differences in how pleiotropy is modeled. Within the model from Uricchio et al. that is used in this study, $\rho \ll 1$ only if the trait effects and selection coefficients are uncoupled for a large portion of the variants, while the remaining variants have effects in the same direction (i.e. variants with larger effects on the trait are more evolutionarily deleterious). There could be extensive pleiotropy even if $\rho \ll$

1, as long as variants have effects in the same direction for all traits and fitness. Our models have the advantage of including an explicit population genetic model of genetic variation. As such, we leverage the extensive existing studies on the distribution of fitness effects for new mutations, mutation rates, and demographic parameters that have been fit to human polymorphism data. Thus, our models are consistent with salient features of human genetic variation data and use reasonable parameter values. We believe that this increases the utility and interpretability of our inferences.

In addition to inferring the mutational target size, InGeAr also infers the extent to which the magnitude of the effect of a variant is affected by selection (τ). Our work confirmed previous findings that there is a coupling between a variant's effect on a trait and its selection coefficient (Schoech et al., 2017). However, we observed that the median for inferred τ using InGeAr differs from the mean τ in previous work. Nevertheless, we found that the credible interval overlaps with the confidence interval (Figure 5.15). Curiously, we found that τ is similar across all traits examined here, perhaps suggesting a common mechanism for the relationship between a mutation's effect on a trait and its effect on fitness that is shared across many traits.

We suggest that our findings that the mutational target sizes are large for most traits and that τ is similar across all traits examined is consistent with the omnigenic model. The omnigenic model posits that variants in a large proportion of the genome affect most traits (Boyle et al., 2017). Our results are consistent with this prediction. For example, we found that the mutational target size for height is around 95Mb, which is approximately 3% of the genome. A second prediction of the omnigenic model is that most of the heritability for traits is explained by the weak effects of peripheral genes. The fact that τ is similar across traits supports this prediction. Suppose that for some traits that are affecting reproductive fitness causing τ to be greater than 0, a number of peripheral genes are contributing to these traits' phenotypes. These same peripheral genes are also contributing to other traits that may not be affecting reproductive fitness, but causing τ to be greater than 0.

Our work contributes to the current literature aiming to unravel the genetic architecture of complex traits. We contributed by developing a method to infer under-studied but important parameters of the genetic architecture, specifically the mutational target size and the coupling between trait's effect and selection. Our work provides support for the newly developed omnigenic model of complex traits. However, our method accounts for linkage disequilibrium (LD) by using GWAS hits that have been previously shown to be independent. We suggest that future work could explicitly incorporate LD into the inference framework.

5.5 Methods

Obtaining summary statistics from genome-wide association studies

Downloading GWAS summary statistics for UKBiobank traits

We downloaded GWAS summary statistics for UKBiobank traits from https://data.broadinstitute.org/alkesgroup/UKBB/.

Selecting independent variants

One summary statistic of the GWAS data used by InGeAr is the number of variants associated with the trait at different freuqencies. When computing this quantity, we need to account for the fact that there may be many GWAS hits within a region of the genome in high LD with each other all showing a significant association, despite there being only one causal variant. To ameliorate this issue, we only used variants that are putatively independent by utilizing the list of independent variants from Supplementary Table 6 from Kichaev et al. (Kichaev et al., 2017).

Obtaining allele frequency for each variant

For each GWAS variant, we obtained its minor allele frequency using the 1000 Genome Project Phase 3 for the European (EUR) population.

Simulating causal variants

For each 100kb region, we simulated causal variants using the forward genetic simulation program SLiM 2 (Haller and Messer, 2017). We simulated 10,000 such 100kb region. We used a mutation rate (mu) of 1.5X10^{-8} and a recombination rate of r= 1X10^{-8}. We specified the distribution of fitness effects for noncoding region using the parameters from Torgerson et al. (Torgerson et al., 2009). We simulated 503 individuals because there are 503 European (EUR) individuals in the 1000 Genome Project Phase 3 from which we used to calculate minor allele frequency from. We outputted the same number of EUR individuals as in the 1000 Genome Project because we wanted the allele frequency calculation from the simulation to be comparable to that from empirical data.

InGeAr

For each trait, we aimed to find a posterior distribution of M and τ. We drew a value for M and a value for τ from uniform prior distributions. τ is drawn uniformly between 0 and 1. Since the number of causal variants differ for each trait, we used different prior distributions for M for different traits. For each pair of M and τ, we performed the inference procedure as described below.

Determining the number of causal variants for each pair of M and τ

For each value of the mutational target size, M, we drew simulated 100kb fragments randomly from the 10,000 simulated fragments. The number of replicates drawn is determined by the mutational target size. For example, if M is equal to 5Mb, we drew 50 simulation replicates

(50Mb/100kb = 50). This is computationally efficient because we could avoid performing SLiM simulations for every value of M. This step returns a list of causal variants, each with an allele frequency and selection coefficient s.

Assigning effect size

For each causal variant, we assigned its effect size on the trait following Equation 1 from Eyre-Walker (2010), which is reproduced below:

$$\alpha_i = \delta s_i^\tau (1 + \varepsilon),$$

where α_i is the effect size of variant i, δ randomly takes value of +1 or -1 with equal probability, s_i is the selection coefficient of variant i which is an output of SLiM simulation, and ε is the error term that is drawn from a normal distribution with mean 0 and a standard deviation of 0.5 as in Eyre-Walker (2010) and Lohmueller (2014).

Computing scaled effect size

For each set of causal variants, we aimed to achieve the desired heritability for a particular trait. As such, we scaled the effect size for each variant with a scaling constant as in Lohmueller (2014):

$$\alpha_i^{\text{scaled}} = \alpha_i C,$$

where C is a normalizing constant for the effect sizes so that:

$$V_A = \sum_{i \text{ variants}} 2p_i(1 - p_1)(\alpha_i C)^2 \approx h_C^2$$

Therefore,

$$C = \sqrt{\frac{h_C^2}{\sum_{i \text{ variants}} 2p_i(1 - p_i)\alpha_i^2}}$$

We first computed the normalizing constant C and then rescaled the effect size for each variant with C, which is the scaled effect size for each causal variant for each trait.

Obtaining GWAS variants from causal variants

Since GWAS does not have statistical power to detect all causal variants, we retained a portion of all the simulated causal variants as GWAS variants. Whether a variant is kept or rejected is determined by its power of being observed in a GWAS. As such, for each variant, we computed the probability it would reach genome-wide significance (P<5X10^{-8}) in the UKBioBank:

$$\text{Variant } i^{th}\text{power} = \Phi(\Phi^{-1})\left(\frac{a}{2}\right) + \lambda_i\sqrt{N} + 1 - \Phi(-\Phi^{-1})\left(\frac{a}{2}\right) + \lambda_i\sqrt{N},$$

where a is the genome-wide significance P-value threshold (5X10^{-8}), N is the number of individuals which is equal to the number of individuals in the UKBiobank for each trait, and λ_i is equal to:

$$\lambda_i = |\alpha_{\text{scaled}}|\sqrt{2p_i(1 - p_i)},$$

where p_i is the minor allele frequency of variant i^{th}.

Then, we retain a subset of the causal variants as GWAS hits . To do this, we draw a random value uniformly between 0 and 1. If the random drawn value is less than the calculated GWAS power for that variant, the variant is kept. Otherwise, it is discarded. This procedure resulted in a set of causal variants that should recapitulate the variants observed in a GWAS.

Rejecting or accepting a pair of τ and M

From a set of causal variants from the previous step, we computed the number of expected GWAS hits and the effect size stratified by allele frequency. Next we employed a two-step procedure to accept or reject values of tau and M from the prior distribution. First, we computed how far the average effect size from the empirical data the relative difference between the average ffect size between the empirical GWAS and the simulated GWAS for each allele frequency bin:

$$\sum_{\mathrm{bin}_i=1}^{6} \frac{|\alpha_i^{\mathrm{emp}} - \alpha_i^{\mathrm{sim}}|}{\alpha_i^{\mathrm{emp}}}$$

We accepted a pair of τ *and* M if this score is less than 0.6. In the second step, we computed the sum of squared differences between the empricial GWAS and the simulated GWAS using the number of variants:

$$\text{sum of squared difference} = \sum_{i=1}^{\mathrm{bins}} (\text{number of variants}_i^{\mathrm{emp}} - \text{number of variants}_i^{\mathrm{sim}})^2$$

We then keep any pair of τ *and* M in the bottom 10% of the distribution. The retained values of τ *and* M comprise the posterior distribution.

Assessing how well *M* and *τ* fit the empirical data

To assess the fit, for each trait, we simulated 10,000 replicates using a value of M and τ drawn from the posterior distribution. We then obtained the number of causal variants, the number of GWAS variants, and the effect size following the procedure as described above for the inference.

Codes availability

All codes can be found on Github at: https://github.com/tnphung/Genetic_Architecture

5.6 Figures

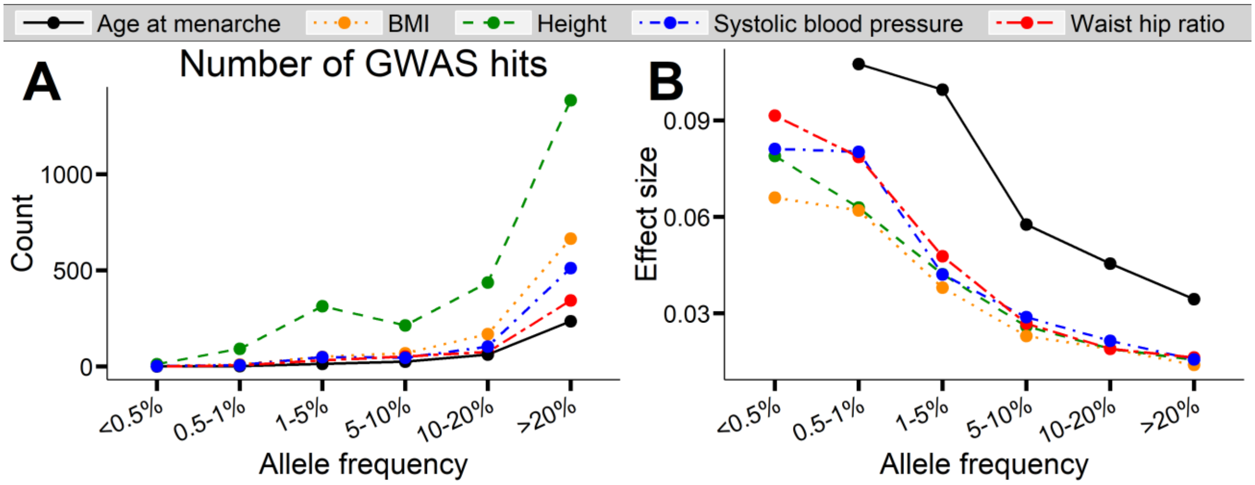


Figure 5.1. Trait-associated variants discovered from GWAS provide insights into the genetic architecture of complex traits. (A) Number of GWAS hits stratified by minor allele frequency. (B) Effect size stratified by minor allele frequency. Minor allele frequency was divided into six bins: <0.5%, 0.5-1%, 1-5%, 5-10%, 10-20%, and >20%. Summary statistics for a representative five traits (Age at menarche, BMI, Height, Systolic blood pressure, and Waist hip ratio) from UKBiobank GWAS were obtained from Kichaev et al. (Kichaev et al., 2017).

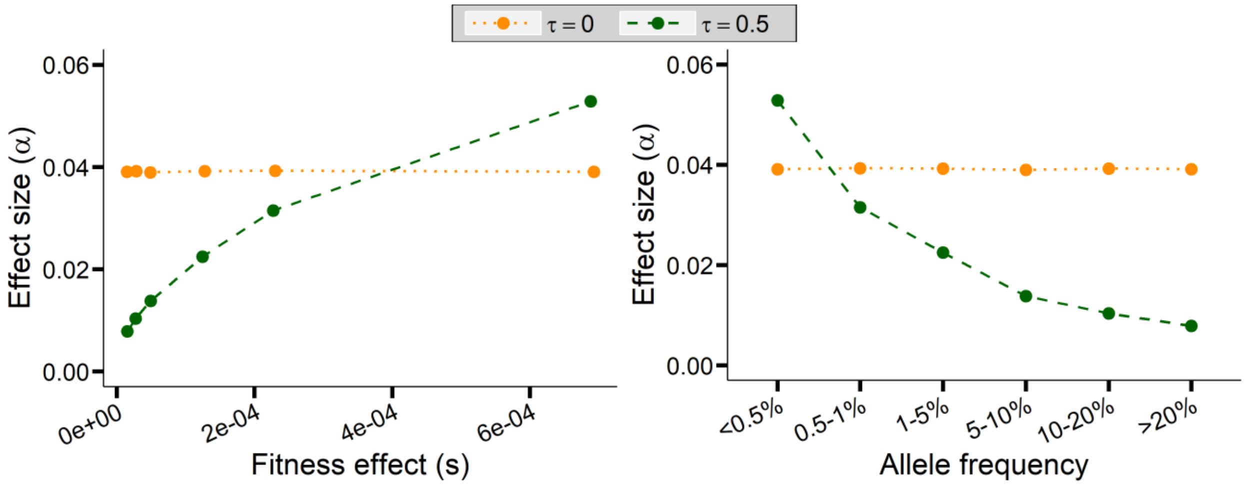


Figure 5.2. τ could account for the negative correlation between effect size and allele frequency. Data were simulated with $\tau = 0$ (orange line) or with $\tau = 0.5$ (green line). When $\tau = 0$, there is no relationship between effect size (α) and fitness effect (s) (left plot). Similarly, we observe no relationship between effect size (α) and allele frequency when $\tau = 0$ (right plot). However, then $\tau > 0$ (i.e. $\tau = 0.5$), a positive relationship between effect size (α) and fitness effect (s) results in a negative correlation between effect size (α) and allele frequency.

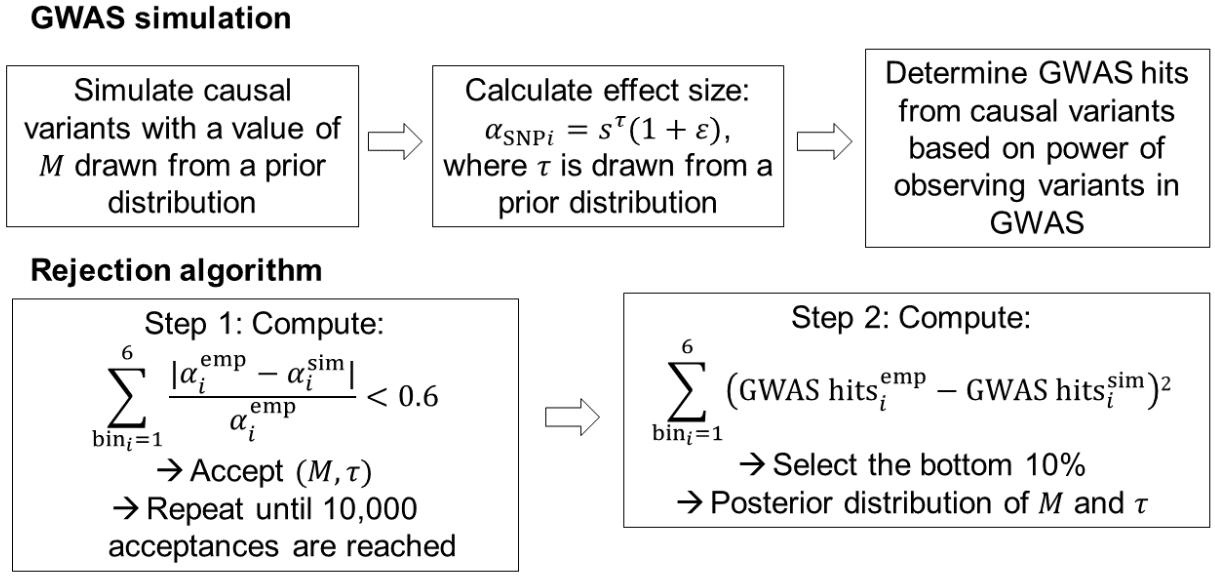

GWAS simulation
Simulate causal variants with a value of M drawn from a prior distribution
Calculate effect size: $\alpha_{\mathrm{SNP}i} = s^{\tau}(1+\varepsilon)$, where τ is drawn from a prior distribution
Determine GWAS hits from causal variants based on power of observing variants in GWAS
Rejection algorithm
Step 1: Compute:
$\sum_{\mathrm{bin}_i=1}^{6} \frac{|\alpha_i^{\mathrm{emp}} - \alpha_i^{\mathrm{sim}}|}{\alpha_i^{\mathrm{emp}}} < 0.6$
→ Accept (M, τ)
→ Repeat until 10,000 acceptances are reached
Step 2: Compute:
$\sum_{\mathrm{bin}_i=1}^{6} \left(\mathrm{GWAS\ hits}_i^{\mathrm{emp}} - \mathrm{GWAS\ hits}_i^{\mathrm{sim}}\right)^2$
→ Select the bottom 10%
→ Posterior distribution of M and τ

Figure 5.3. Overview of InGeAr.

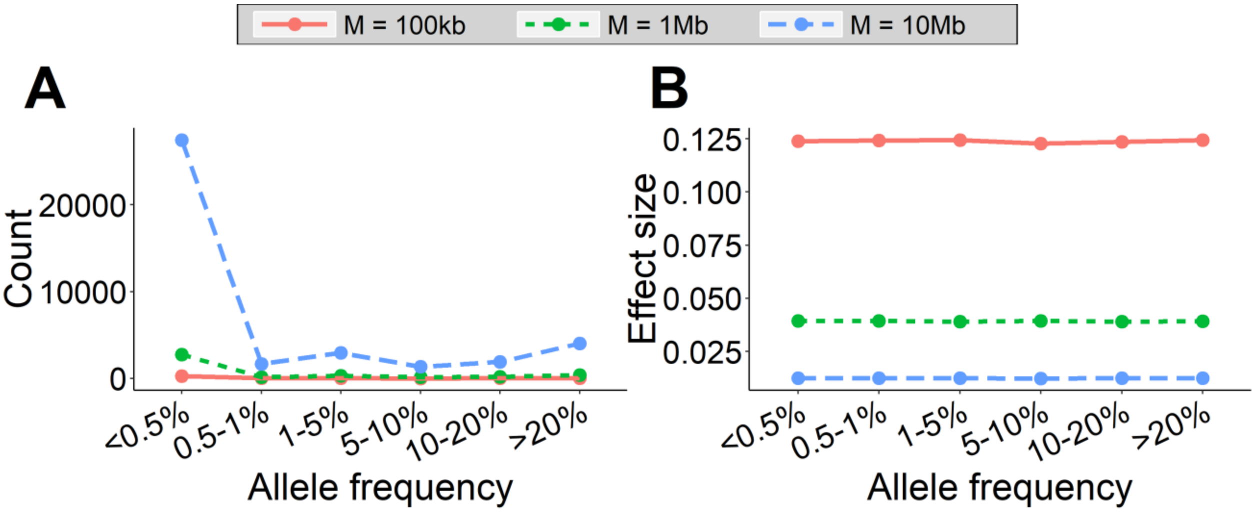


Figure 5.4. The mutational target size determines the number of causal variants. Data were simulated with three values of the mutational target size, ranging from small ($M = 100$kb, red lines) to intermediate ($M = 1$Mb, green lines) to large ($M = 10$Mb, blue lines). A τ value of 0 was used in these simulations. (A) The number of causal variants stratified by minor allele frequency. (B) Effect size stratified by minor allele frequency. Minor allele frequency was divided into six bins: <0.5%, 0.5-1%, 1-5%, 5-10%, 10-20%, and >20%.

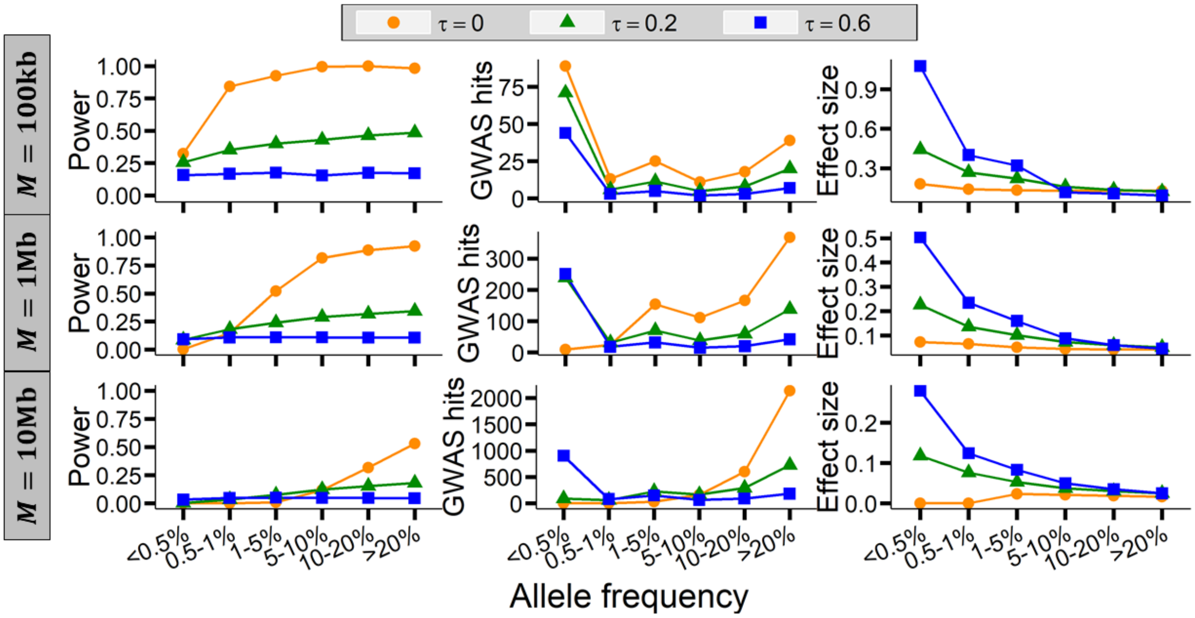


Figure 5.5. Both M and τ affect the number of GWAS hits and effect size. Data were simulated with three values of the mutational target size, ranging from small ($M = 100$kb, first row) to intermediate ($M = 1$Mb, second row) to large ($M = 10$Mb, third row). Three values of τ were used, ranging from small ($\tau = 0$, orange lines) to intermediate ($\tau = 0.2$, green lines) to large ($\tau = 0.6$, blue lines). The first column plots the power to detect a causal variant in a GWAS stratified by minor allele frequency. The second column plots the number of GWAS hits stratified by minor allele frequency. The third column plots the effect size of GWAS hits stratified by minor allele frequency. Minor allele frequency was divided into six bins: <0.5%, 0.5-1%, 1-5%, 5-10%, 10-20%, and >20%.

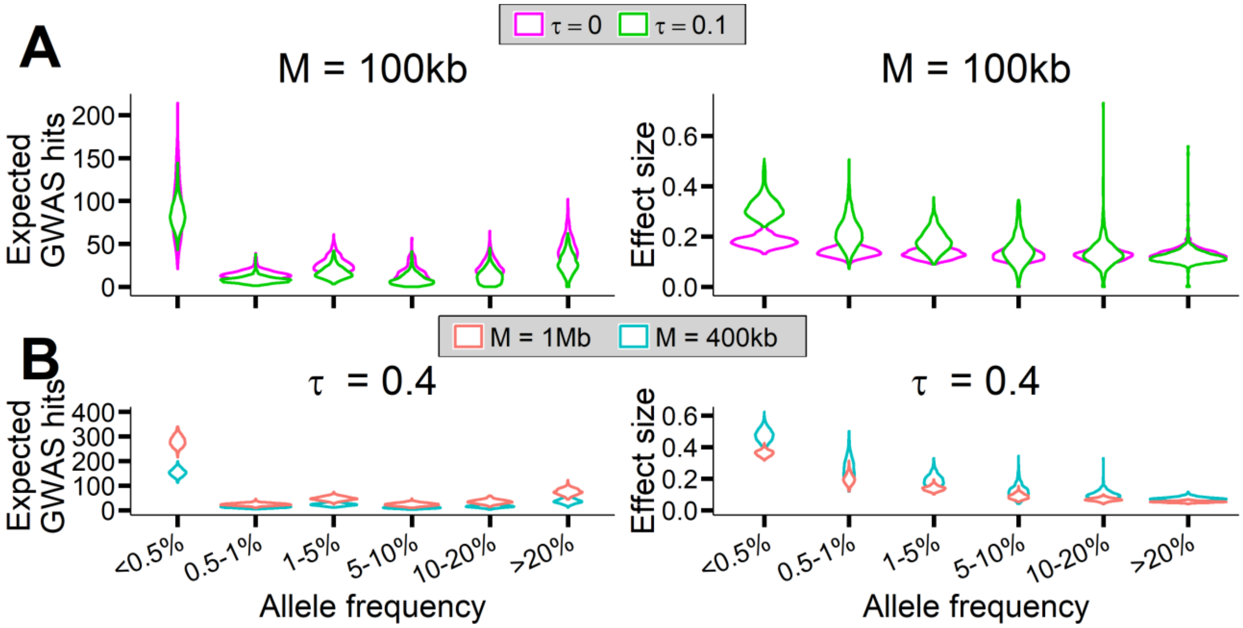


Figure 5.6. Expected GWAS hits are similar for two different pairs of (τ, M) but the effect size can distinguish them and vice versa. (A) Data were simulated with the mutational target size of 100kb. Two values of τ were used: $\tau = 0$ (magenta violin plots) and $\tau = 0.1$ (green violin plots). (B) Data were simulated with two values of the mutational target size: $M = 1$Mb (red violin plots) and $M = 400$kb (blue violin plots). Plots on the left column represent the number of expected GWAS hits stratified by minor allele frequency. Plots on the right column represent the effect size of GWAS hits stratified by minor allele frequency. For each pair of (τ, M), 500 simulations were performed. The number of expected GWAS hits and effect size of expected GWAS hits for each simulation are plotted as violin plots to show the distribution.

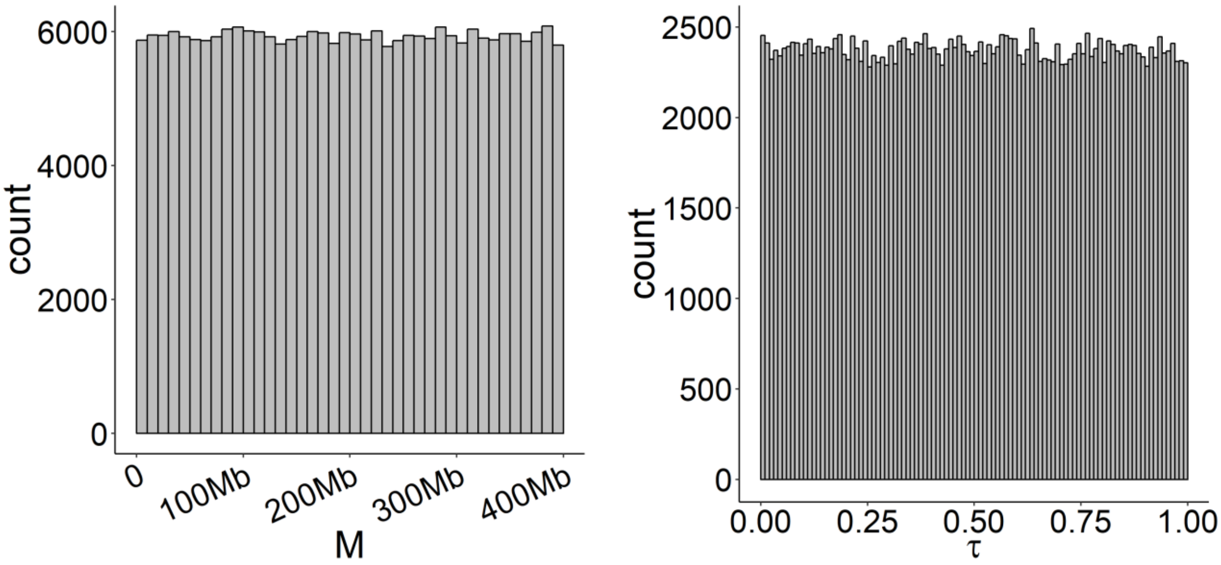


Figure 5.7. Prior distribution of M and τ. The mutational target size (M) was drawn from a uniform distribution that ranges from 0 to 400Mb in increment of 100kb (left plot). τ was drawn from a uniform distribution that ranges from 0 to 1 (right plot).

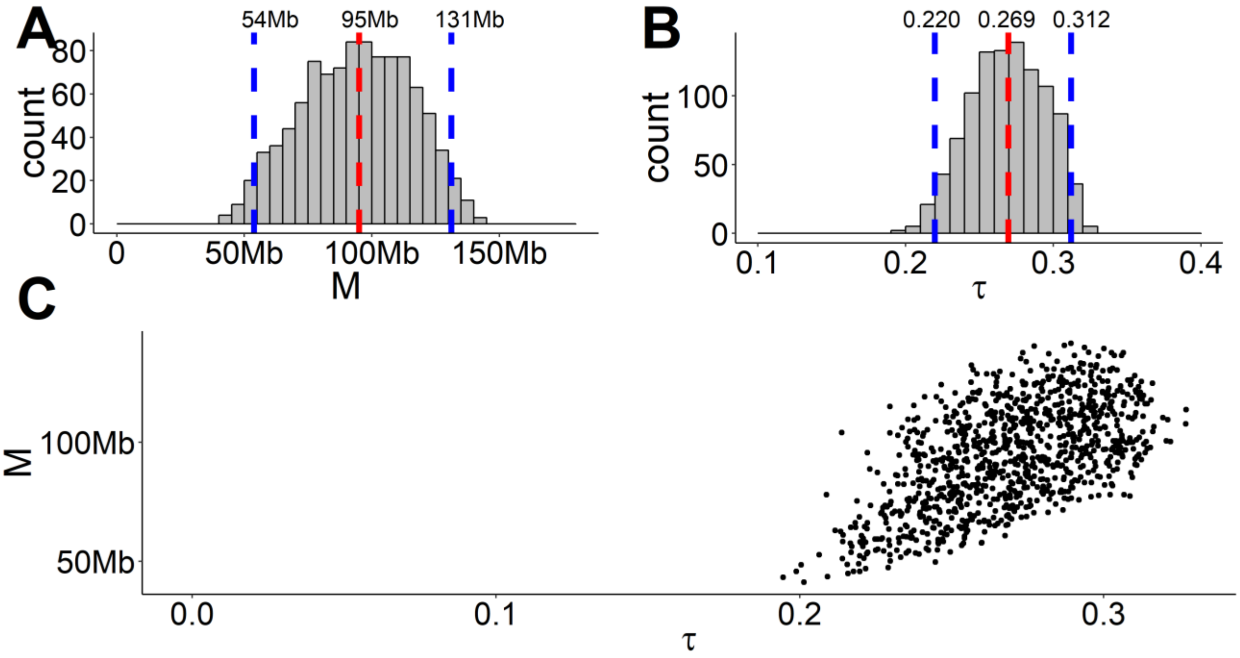


Figure 5.8. Posterior distribution of M, τ, and their joint distribution. (A) Posterior distribution of the mutational target size (M). (B) Posterior distribution of the coupling parameter between effect size and selection (τ). Red lines represent the median value of the posterior distribution. Blue lines represent the 95% confidence interval. (C) The joint posterior distribution of M and τ.

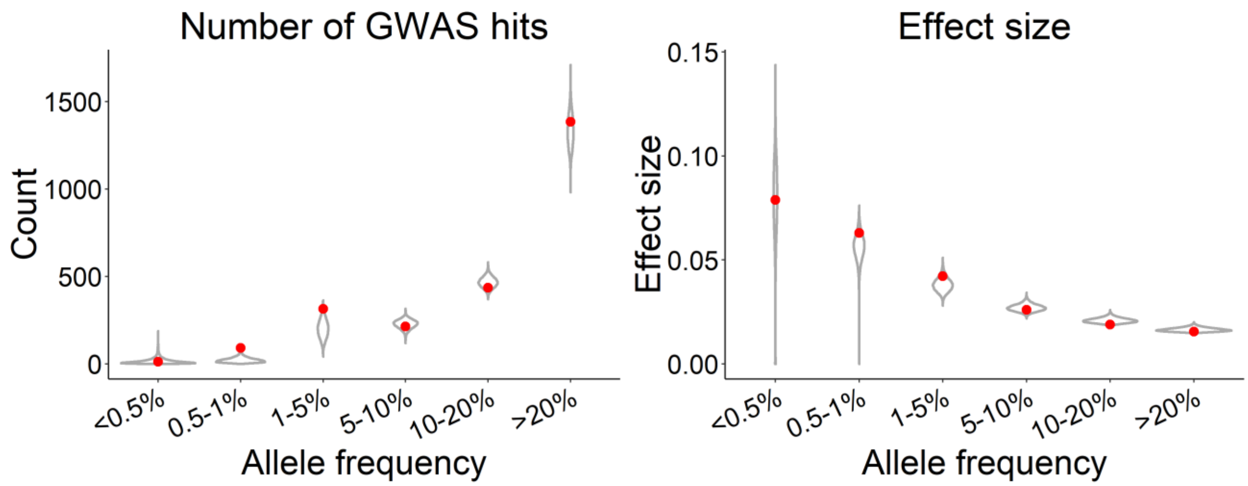


Figure 5.9. Assess model fit for height. Data were simulated using a value of M and a value of τ that are drawn from posterior distributions. 10,000 simulations were performed and plotted as violin plots. Left plot: The number of GWAS hits stratified by minor allele frequency. Right plot: The effect size of GWAS hits stratified by minor allele frequency. Minor allele frequency was divided into six bins: <0.5%, 0.5-1%, 1-5%, 5-10%, 10-20%, and >20%. Red points represent values from the empirical data for height. Note the excellent fit between the simulated data and empirical data.

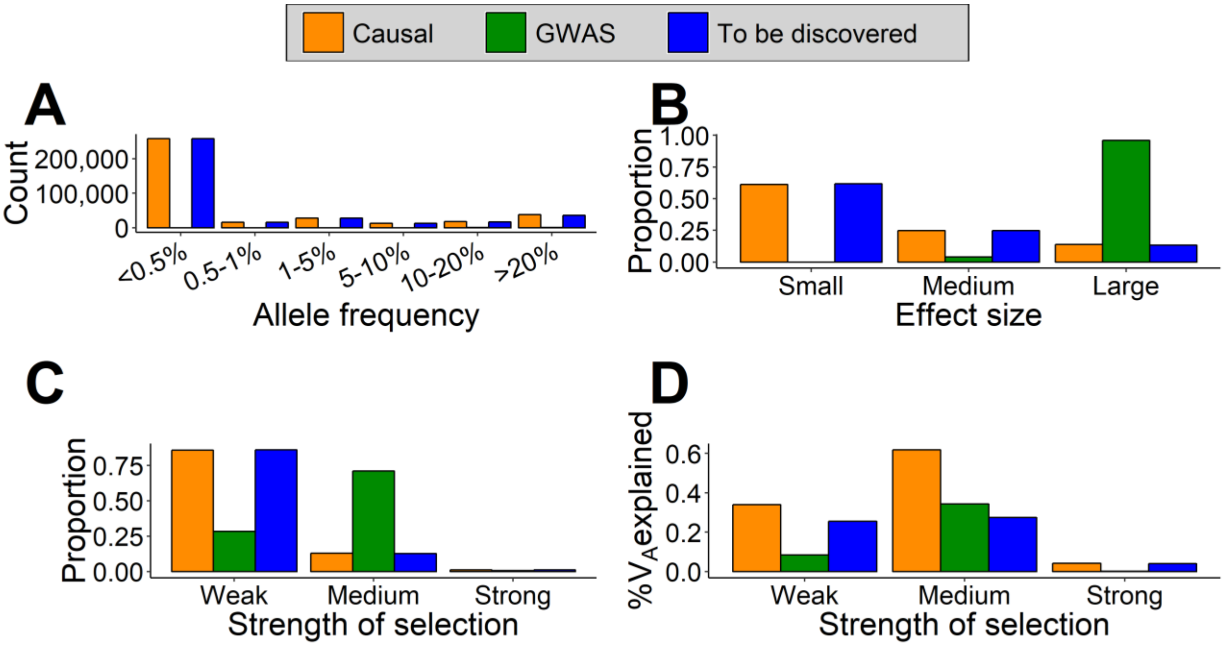


Figure 5.10. Properties of causal variants, GWAS hits, and remaining undiscovered causal variants. (A) Number of causal variants (orange), GWAS hits (green), and remaining undiscovered causal variants (blue) stratified by minor allele frequency. Minor allele frequency was divided into six bins: <0.5%, 0.5-1%, 1-5%, 5-10%, 10-20%, and >20%. (B) Proportion of causal variants (orange), GWAS hits (green), and remaining undiscovered causal variants (blue) stratified by effect size. Effect size was divided into three bins: small ($\alpha \leq 0.001$), medium ($0.001 < \alpha \leq 0.01$), and large ($\alpha > 0.01$). (C) Proportion of causal variants (orange), GWAS hits (green), and remaining undiscovered causal variants (blue) stratified by strength of selection. Selection strength was divided into three bins: weak ($|s| \leq 10^{-4}$), medium ($10^{-4} < |s| \leq 10^{-2}$), and strong ($|s| > 10^{-2}$). (D) Proportion of additive variance explained by causal variants (orange), GWAS hits (green), and remaining undiscovered causal variants (blue) stratified by strength of selection. Selection strength was divided into three bins: weak ($|s| \leq 10^{-4}$), medium ($10^{-4} < |s| \leq 10^{-2}$), and strong ($|s| > 10^{-2}$).

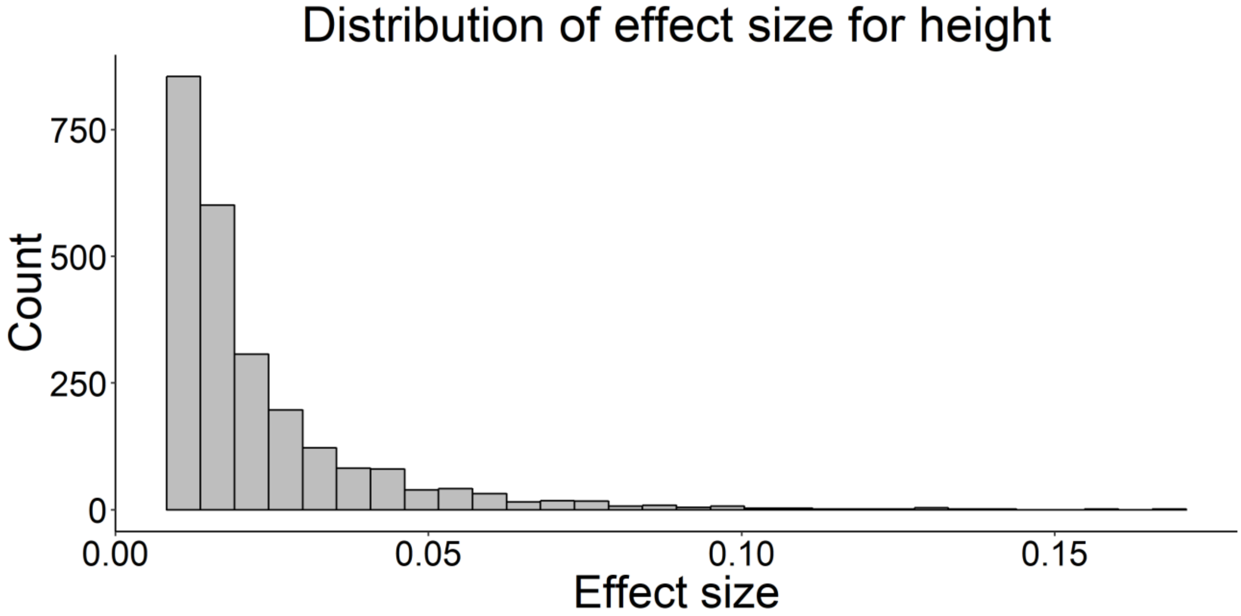


Figure 5.11. Distribution of effect sizes for height. Effect size for GWAS of height from UKBiobank was obtained from Kichaev et al. (Kichaev et al., 2017).

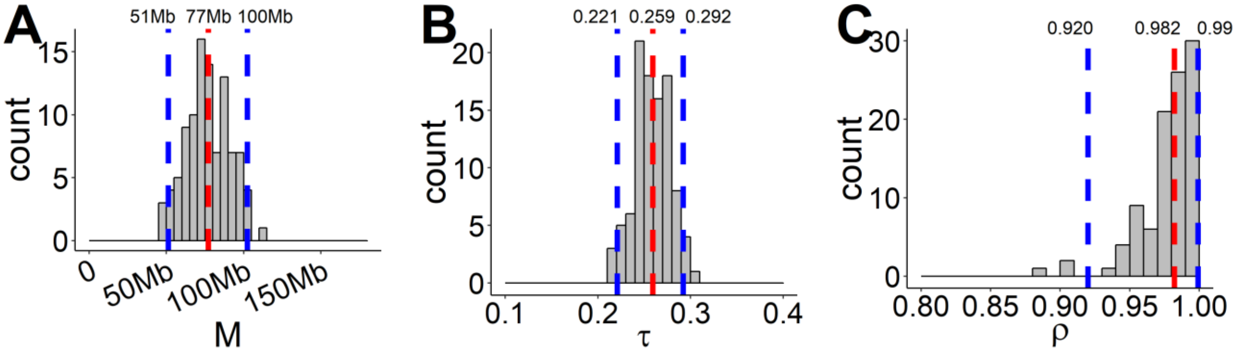


Figure 5.12. Posterior distribution of M, τ, and ρ when including pleiotropy in the inference. (A) Posterior distribution of the mutational target size (M). (B) Posterior distribution of the coupling parameter between effect size and selection (τ). (C) Posterior distribution of the parameter that captures pleiotropy (ρ). Red lines represent the median value of the posterior distribution. Blue lines represent the 95% confidence interval.

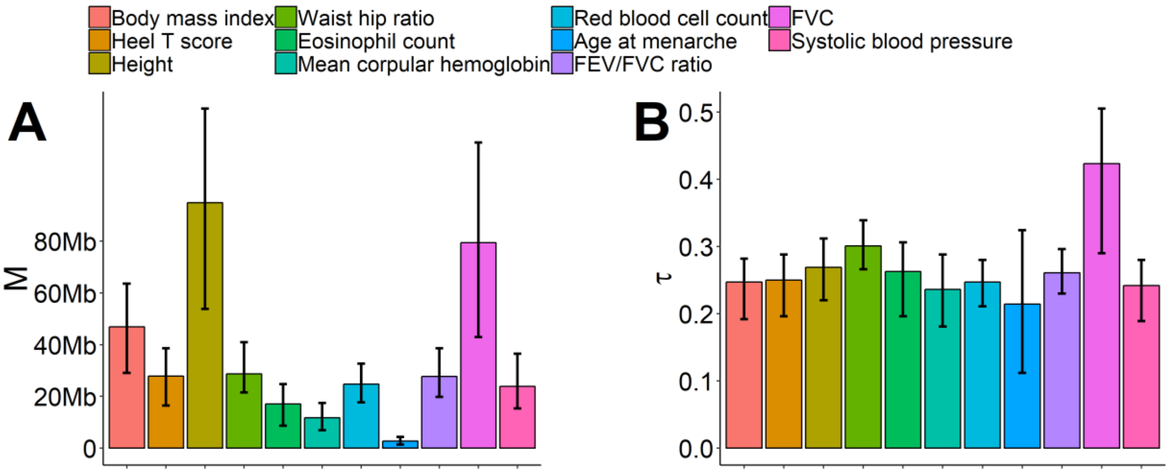


Figure 5.13. M and τ for other UKBiobank traits. Summary statistics from GWAS for these traits were obtained from Kichaev et al. (Kichaev et al., 2017).

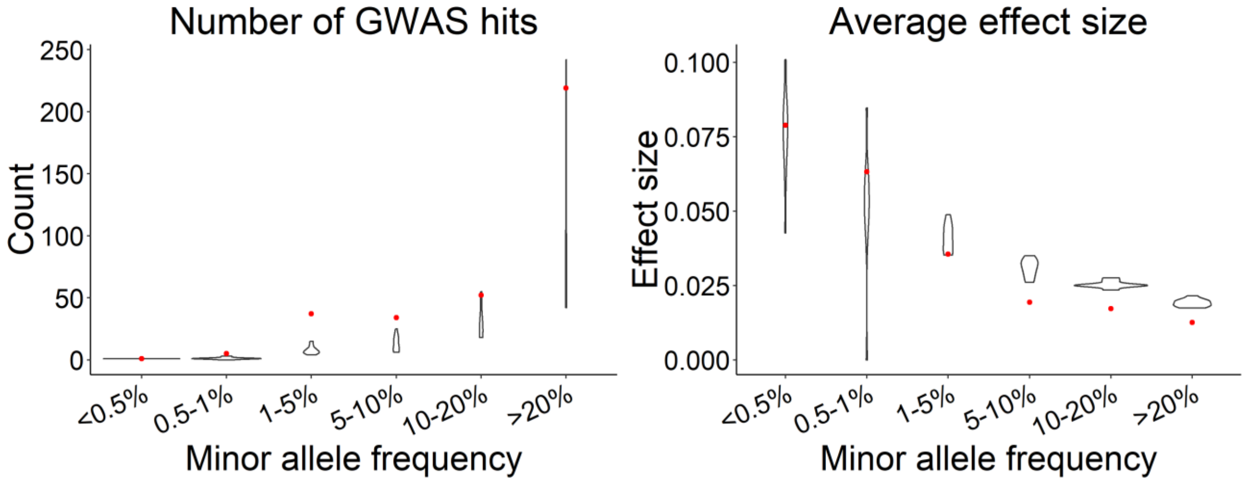


Figure 5.14. Assess model fit for balding type 1. Data were simulated using a value of M and a value of τ that are drawn from posterior distributions. 10,000 simulations were performed and plotted as violin plots. Left plot: The number of GWAS hits stratified by minor allele frequency. Right plot: The effect size of GWAS hits stratified by minor allele frequency. Minor allele frequency was divided into six bins: <0.5%, 0.5-1%, 1-5%, 5-10%, 10-20%, and >20%. Red points represent values from the empirical data for height. Note that here the simulated data underestimated the number of GWAS hits for variants segregating at intermediate allele frequency. In addition, the simulated data overestimated the effect size for common variants.

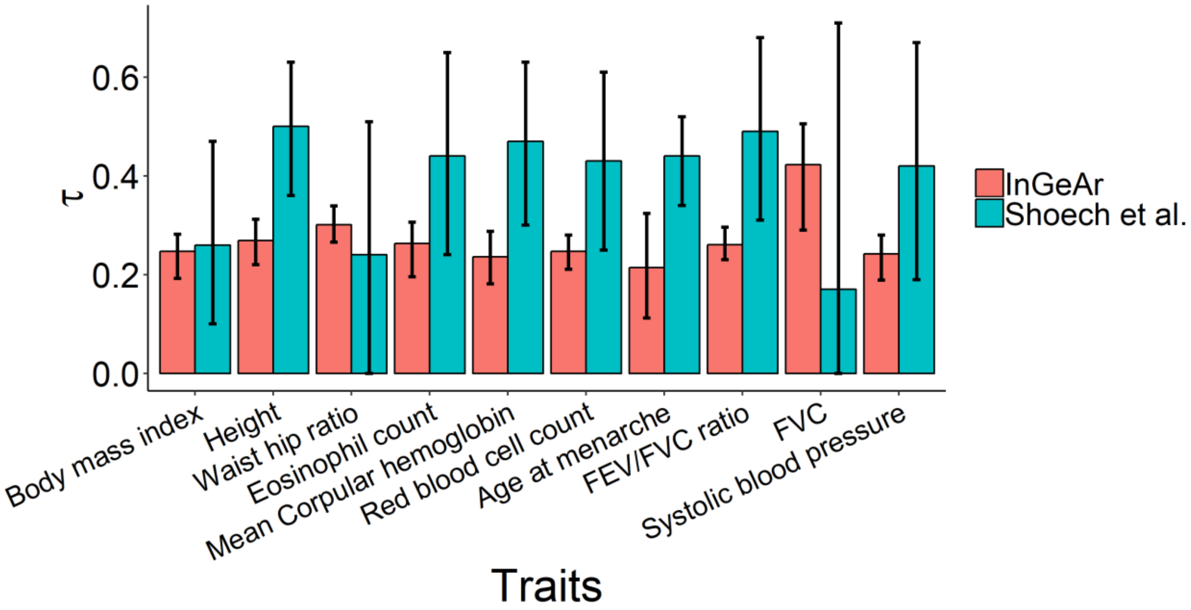


Figure 5.15. Comparison of τ between InGeAr and Schoech et al. Schoech et al.'s τ was obtained from Supplementary Table 6 (Schoech et al., 2017). InGeAr's τ is from present study. Note that while the inferred values differ between these two studies, the confidence interval overlapped.

5.7 Tables

Table 5.1. M and τ when considering variants with large effect size.

	M	τ
All variants	95Mb (54Mb – 131Mb)	0.268 (0.220 – 0.312)
Only variants with effect size ≥ 0.02	45Mb (21Mb – 72Mb)	0.253 (0.187 – 0.313)
Only variants with effect size ≥ 0.05	15Mb (7Mb – 40Mb)	0.169 (0.111 – 0.259)

Table 5.2. M and τ when incorporating pleiotropy.

	M	τ	ρ
All variants	77Mb (51Mb – 102Mb)	0.259 (0.221 – 0.292)	0.982 (0.920 – 0.999)
Only variants with effect size ≥ 0.02	40Mb (10Mb – 94Mb)	0.232 (0.114 – 0.343)	0.890 (0.482 – 0.993)
Only variants with effect size ≥ 0.05	24Mb (5Mb – 118Mb)	0.245 (0.081 – 0.539)	0.710 (0.104 – 0.966)

Table 5.3. Summary of the posterior distribution of M and τ for other traits.

Traits	M		τ	
	Median	95% credible interval	Median	95% credible interval
Height	95Mb	54Mb – 130Mb	0.269	0.220 – 0.312
Heel T score	28Mb	16Mb – 39Mb	0.250	0.196 – 0.288
Eosinophil count	17Mb	9Mb – 25Mb	0.263	0.196 – 0.306
Red blood cell count	25Mb	18Mb – 33Mb	0.247	0.211 – 0.280
Body mass index	47Mb	29Mb – 64Mb	0.247	0.192 – 0.282
Mean corpular hemoglobin	12Mb	7Mb – 17Mb	0.236	0.181 – 0.288
Age at menarche	3Mb	1Mb – 4Mb	0.214	0.112 – 0.324
FEV/FVC ratio	28Mb	20Mb – 39Mb	0.261	0.230 – 0.296
Systolic blood pressure	24Mb	15Mb – 36Mb	0.242	0.189 – 0.280

Waist hip ratio	29Mb	22Mb – 41Mb	0.301	0.266 – 0.339

REFERENCES

Agarwala, V., Flannick, J., Sunyaev, S., and Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. Nat. Genet. *45*, 1418–1427.

Akashi, H., Osada, N., and Ohta, T. (2012). Weak selection and protein evolution. Genetics *192*, 15–31.

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881–888.

Arbeithuber, B., Betancourt, A.J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc. Natl. Acad. Sci. *112*, 2109–2114.

Arbiza, L., Gottipati, S., Siepel, A., and Keinan, A. (2014). Contrasting X-Linked and Autosomal Diversity across 14 Human Populations. Am. J. Hum. Genet. *94*, 827–844.

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E.M., Bowden, R., Aneas, I., Broxholme, J., et al. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science *336*, 193–198.

Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J.K., Hayward, J.J., Cohen, P.E., Greally, J.M., Wang, J., et al. (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLoS Genet *9*, e1003984.

Baker, P.J., Funk, S.M., Bruford, M.W., and Harris, S. (2004). Polygynandry in a red fox population: implications for the evolution of group living in canids? Behav. Ecol. *15*, 766–778.

Begun, D.J., and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature *356*, 519–520.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., et al. (2007). Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. PLoS Biol *5*, e310.

Beichman, A.C., Phung, T.N., and Lohmueller, K.E. (2017). Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories. G3 Genes Genomes Genet. *7*, 3605–3620.

Berglund, J., Pollard, K.S., and Webster, M.T. (2009). Hotspots of Biased Nucleotide Substitutions in Human Genes. PLoS Biol. *7*, e26.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. Gene *241*, 3–17.

Birky, C.W., and Walsh, J.B. (1988). Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. U. S. A. *85*, 6414–6418.

Botigué, L.R., Song, S., Scheu, A., Gopalan, S., Pendleton, A.L., Oetjens, M., Taravella, A.M., Seregély, T., Zeeb-Lanz, A., Arbogast, R.-M., et al. (2017). Ancient European dog genomes reveal continuity since the Early Neolithic. Nat. Commun. *8*, 16082.

Boyko, A.R. (2011). The domestic dog: man's best friend in the genomic era. Genome Biol. *12*, 216.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186.

Cai, J.J., Macpherson, J.M., Sella, G., and Petrov, D.A. (2009). Pervasive Hitchhiking at Coding and Regulatory Sites in Humans. PLoS Genet *5*, e1000336.

Campbell, C.L., Bhérer, C., Morrow, B.E., Boyko, A.R., and Auton, A. (2016). A Pedigree-Based Map of Recombination in the Domestic Dog Genome. G3 GenesGenomesGenetics *6*, 3517–3524.

Campos, J.L., Halligan, D.L., Haddrill, P.R., and Charlesworth, B. (2014). The relation between recombination rate and patterns of molecular evolution and variation in Drosophila melanogaster. Mol. Biol. Evol. *31*, 1010–1028.

Charlesworth, B. (2012). The effects of deleterious mutations on evolution at linked sites. Genetics *190*, 5–22.

Charlesworth, B., Morgan, M.T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics *134*, 1289–1303.

Charlesworth, D., Charlesworth, B., and Morgan, M.T. (1995). The pattern of neutral molecular variation under the background selection model. Genetics *141*, 1619–1632.

Comeron, J.M. (2014). Background selection as baseline for nucleotide variation across the Drosophila genome. PLoS Genet. *10*, e1004434.

Coop, G., and Ralph, P. (2012). Patterns of neutral diversity under general models of selective sweeps. Genetics *192*, 205–224.

Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S., and Sidow, A. (2004). Characterization of Evolutionary Rates and Constraints in Three Mammalian Genomes. Genome Res. *14*, 539–548.

Corbett-Detig, R.B., Hartl, D.L., and Sackton, T.B. (2015). Natural Selection Constrains Neutral Diversity across A Wide Range of Species. PLOS Biol. *13*, e1002112.

Cotter, D.J., Brotman, S.M., and Sayres, M.A.W. (2016). Genetic Diversity on the Human X Chromosome does not Support a Strict Pseudoautosomal Boundary. Genetics genetics.114.172692.

Cruickshank, T.E., and Hahn, M.W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. *23*, 3133–3157.

Cutler, D.J. (1998). Clustered Mutations Have No Effect on the Overdispersed Molecular Clock: A Response to Huai and Woodruff. Genetics *149*, 463–464.

Cutter, A.D., and Choi, J.Y. (2010). Natural selection shapes nucleotide polymorphism across the genome of the nematode Caenorhabditis briggsae. Genome Res. *20*, 1103–1111.

Cutter, A.D., and Payseur, B.A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. *14*, 262–274.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLOS Comput Biol *6*, e1001025.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Drake, A.G., Coquerelle, M., and Colombeau, G. (2015). 3D morphometric analysis of fossil canid skulls contradicts the suggested domestication of dogs during the late Paleolithic. Sci. Rep. *5*, 8299.

Duret, L., and Arndt, P.F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. *4*, e1000071.

Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. *10*, 285–311.

Edwards, S.V., and Beerli, P. (2000). Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evol. Int. J. Org. Evol. *54*, 1839–1854.

Emery, L.S., Felsenstein, J., and Akey, J.M. (2010). Estimators of the Human Effective Sex Ratio Detect Sex Biases on Different Timescales. Am. J. Hum. Genet. *87*, 848–856.

Enard, D., Messer, P.W., and Petrov, D.A. (2014). Genome-wide signals of positive selection in human evolution. Genome Res. *24*, 885–895.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., and Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. PLOS Genet *9*, e1003905.

Eyre-Walker, A. (2010). Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl. Acad. Sci. *107*, 1752–1756.

Eyre-Walker, A., and Eyre-Walker, Y.C. (2014). How Much of the Variation in the Mutation Rate Along the Human Genome Can Be Explained? G3 Genes Genomes Genet. *4*, 1667–1670.

Flowers, J.M., Molina, J., Rubinstein, S., Huang, P., Schaal, B.A., and Purugganan, M.D. (2012). Natural Selection in Gene-Dense Regions Shapes the Genomic Pattern of Polymorphism in Wild and Domesticated Rice. Mol. Biol. Evol. *29*, 675–687.

Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Genome of the Netherlands Consortium, van Duijn, C.M., Swertz, M., Wijmenga, C., et al. (2015). Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. *advance online publication*.

Frantz, L.A.F., Mullin, V.E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., Linderholm, A., Mattiangeli, V., Teasdale, M.D., Dimopoulos, E.A., et al. (2016). Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science *352*, 1228–1231.

Freedman, A.H., and Wayne, R.K. (2017). Deciphering the Origin of Dogs: From Fossils to Genomes. Annu. Rev. Anim. Biosci. *5*, 281–307.

Freedman, A.H., Gronau, I., Schweizer, R.M., Vecchyo, D.O.-D., Han, E., Silva, P.M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B., et al. (2014). Genome Sequencing Highlights the Dynamic Early History of Dogs. PLOS Genet. *10*, e1004016.

Freedman, A.H., Lohmueller, K.E., and Wayne, R.K. (2016). Evolutionary History, Selective Sweeps, and Deleterious Variation in the Dog. Annu. Rev. Ecol. Evol. Syst. *47*, 73–96.

Galtier, N., and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. TIG *23*, 273–277.

Geraldes, A., Basset, P., Smith, K.L., and Nachman, M.W. (2011). Higher differentiation among subspecies of the house mouse (Mus musculus) in genomic regions with low recombination. Mol. Ecol. *20*, 4722–4736.

Gibson, G. (2011). Rare and common variants: twenty arguments. Nat. Rev. Genet. *13*, 135–145.

Gillespie, J.H., and Langley, C.H. (1979). Are evolutionary rates really variable? J. Mol. Evol. *13*, 27–34.

Gou, X., Wang, Z., Li, N., Qiu, F., Xu, Z., Yan, D., Yang, S., Jia, J., Kong, X., Wei, Z., et al. (2014). Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. Genome Res. *24*, 1308–1315.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Project, T. 1000 G., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. *108*, 11983–11988.

Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. Nat Genet *43*, 1031–1034.

Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet. *47*, 276–283.

Haller, B.C., and Messer, P.W. (2017). SLiM 2: Flexible, Interactive Forward Genetic Simulations. Mol. Biol. Evol. *34*, 230–240.

Halligan, D.L., Kousathanas, A., Ness, R.W., Harr, B., Eory, L., Keane, T.M., Adams, D.J., and Keightley, P.D. (2013). Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet *9*, e1003995.

Hammer, M.F., Mendez, F.L., Cox, M.P., Woerner, A.E., and Wall, J.D. (2008). Sex-Biased Evolutionary Forces Shape Genomic Patterns of Human Diversity. PLOS Genet. *4*, e1000202.

Hammer, M.F., Woerner, A.E., Mendez, F.L., Watkins, J.C., Cox, M.P., and Wall, J.D. (2010). The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat. Genet. *42*, 830–831.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O’Connor, M., Kolbe, D., et al. (2003). Covariation in Frequencies of Substitution, Deletion, Transposition, and Recombination During Eutherian Evolution. Genome Res. *13*, 13–26.

Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinforma. Oxf. Engl. *22*, 2971–2972.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. *32*, 835–845.

Hellmann, I., Ebersberger, I., Ptak, S.E., Pääbo, S., and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. *72*, 1527–1535.

Hellmann, I., Prufer, K., Ji, H., Zody, M.C., Paabo, S., and Ptak, S.E. (2005). Why do human diversity levels vary at a megabase scale? Genome Res. *15*, 1222–1231.

Hellmann, I., Mang, Y., Gu, Z., Li, P., de la Vega, F.M., Clark, A.G., and Nielsen, R. (2008). Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res. *18*, 1020–1029.

Hemmer, H. (1990). Domestication: The Decline of Environmental Appreciation (Cambridge University Press).

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., Project, 1000 Genomes, Sella, G., and Przeworski, M. (2011). Classic Selective Sweeps Were Rare in Recent Human Evolution. Science *331*, 920–924.

Hodgkinson, A., and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. *12*, 756–766.

Hudson, R.R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.

Hudson, R.R., and Kaplan, N.L. (1995). Deleterious background selection with recombination. Genetics *141*, 1605–1617.

Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.F., Thomas, M.A., Haussler, D., and Jacob, H.J. (2004). Comparative recombination rates in the rat, mouse, and human genomes. Genome Res *14*.

Kaplan, N.L., Hudson, R.R., and Langley, C.H. (1989). The “hitchhiking effect” revisited. Genetics *123*, 887–899.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. (2003). The UCSC Genome Browser Database. Nucleic Acids Res *31*.

Keinan, A., Mullikin, J.C., Patterson, N., and Reich, D. (2009). Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat. Genet. *41*, 66–70.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, and D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. *100*, 11484–11489.

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M., Scoech, A., Pasaniuc, B., and Price, A. (2017). Leveraging polygenic functional enrichment to improve GWAS power. BioRxiv 222265.

Kim, S. (2015). ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Commun. Stat. Appl. Methods *22*, 665–674.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111–120.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution (Cambridge: Cambridge University Press).

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat. Genet. *31*, 241–247.

Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

Kulathinal, R.J., Bennett, S.M., Fitzpatrick, C.L., and Noor, M.A. (2008). Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. Proc. Natl. Acad. Sci. U. S. A. *105*, 10051–10056.

Kumar, S., and Subramanian, S. (2002). Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. *99*, 803–808.

Langergraber, K.E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J.C., Muller, M.N., et al. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc. Natl. Acad. Sci. *109*, 15716–15721.

Larson, G., Karlsson, E.K., Perri, A., Webster, M.T., Ho, S.Y.W., Peters, J., Stahl, P.W., Piper, P.J., Lingaas, F., Fredholm, M., et al. (2012). Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proc. Natl. Acad. Sci. *109*, 8878–8883.

Lercher, M.J., and Hurst, L.D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. TIG *18*, 337–340.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.

Li, W.H., Yi, S., and Makova, K. (2002). Male-driven evolution. Curr Opin Genet Dev *12*.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature *438*, 803–819.

Link, V., Aguilar-Gómez, D., Ramírez-Suástegui, C., Hurst, L.D., and Cortez, D. (2017). Male Mutation Bias Is the Main Force Shaping Chromosomal Substitution Rates in Monotreme Mammals. Genome Biol. Evol. *9*, 2198–2210.

Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.

Lohmueller, K.E. (2014). The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. PLOS Genet. *10*, e1004379.

Lohmueller, K.E., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A.F., Grarup, N., et al. (2011). Natural Selection Affects Multiple Aspects of Genetic Variation at Putatively Neutral Sites across the Human Genome. PLoS Genet *7*, e1002326.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Marsden, C.D., Vecchyo, D.O.-D., O'Brien, D.P., Taylor, J.F., Ramirez, O., Vilà, C., Marques-Bonet, T., Schnabel, R.D., Wayne, R.K., and Lohmueller, K.E. (2016). Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc. Natl. Acad. Sci. *113*, 152–157.

Maynard Smith, J., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genet. Res. *23*, 23–35.

McGaugh, S.E., Heil, C.S., Manzano-Winkler, B., Loewe, L., Goldstein, S., Himmel, T.L., and Noor, M.A. (2012). Recombination modulates how selection affects linked sites in Drosophila. PLoS Biol. *10*, e1001422.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. *5*, e1000471.

Meader, S., Ponting, C.P., and Lunter, G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. *20*, 1335–1343.

Moorjani, P., Amorim, C.E.G., Arndt, P.F., and Przeworski, M. (2016). Variation in the molecular clock of primates. BioRxiv 036434.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Nachman, M.W. (2001). Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. TIG *17*, 481–485.

Narang, P., and Wilson Sayres, M.A. (2015). Long-term natural selection affects patterns of neutral divergence on the X chromosome more than the autosomes. BioRxiv 023234.

Narang, P., Sayres, W., and A, M. (2016). Variable Autosomal and X Divergence Near and Far from Genes Affects Estimates of Male Mutation Bias in Great Apes. Genome Biol. Evol. *8*, 3393–3405.

Noor, M.A. (2008). Mutagenesis from meiotic recombination is not a primary driver of sequence divergence between Saccharomyces species. Mol. Biol. Evol. *25*, 2439–2444.

Nordborg, M., Charlesworth, B., and Charlesworth, D. (1996). The effect of recombination on background selection. Genet. Res. *67*, 159–174.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature *246*, 96–98.

Ostrander, E.A., and Kruglyak, L. (2000). Unleashing the canine genome. Genome Res. *10*, 1271–1274.

Ostrander, E.A., Wayne, R.K., Freedman, A.H., and Davis, B.W. (2017). Demographic history, selection and functional diversity of the canine genome. Nat. Rev. Genet. *18*, 705–720.

Palamara, P.F., Francioli, L.C., Wilton, P.R., Genovese, G., Gusev, A., Finucane, H.K., Sankararaman, S., Sunyaev, S.R., de Bakker, P.I.W., Wakeley, J., et al. (2015). Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. Am. J. Hum. Genet. *97*, 775–789.

Palazzo, A.F., and Gregory, T.R. (2014). The Case for Junk DNA. PLOS Genet *10*, e1004351.

Park, J.-H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J., Fraumeni, J.F., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc. Natl. Acad. Sci. *108*, 18026–18031.

Payseur, B.A., and Nachman, M.W. (2002). Gene density and human nucleotide polymorphism. Mol. Biol. Evol. *19*, 336–340.

Phung, T.N., Huber, C.D., and Lohmueller, K.E. (2016). Determining the Effect of Natural Selection on Linked Neutral Divergence across Species. PLOS Genet *12*, e1006199.

Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J.A., et al. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature *518*, 360–364.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Pool, J.E., and Nielsen, R. (2007). Population Size Changes Reshape Genomic Patterns of Diversity. Evolution *61*, 3001–3006.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. Nature *499*, 471–475.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G.V., and Camerini-Otero, R.D. (2014). Recombination initiation maps of individual human genomes. Science *346*, 1256442.

Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. PLOS Genet *10*, e1004525.

Rannala, B., and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics *164*, 1645–1656.

Reed, F.A., Akey, J.M., and Aquadro, C.F. (2005). Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Res. *15*, 1211–1221.

Sayres, W., and A, M. (2018). Genetic Diversity on the Sex Chromosomes. Genome Biol. Evol. *10*, 1064–1078.

Scally, A., and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. *13*, 745–753.

Schoech, A., Jordan, D., Loh, P.-R., Gazal, S., O'Connor, L., Balick, D.J., Palamara, P.F., Finucane, H., Sunyaev, S.R., and Price, A.L. (2017). Quantification of frequency-dependent genetic architectures and action of negative selection in 25 UK Biobank traits. BioRxiv 188086.

Schrider, D.R., and Kern, A.D. (2015). Inferring Selective Constraint from Population Genomic Data Suggests Recent Regulatory Turnover in the Human Brain. Genome Biol. Evol. *7*, 3511–3528.

Schrider, D., Shanku, A.G., and Kern, A.D. (2016). Effects of linked selective sweeps on demographic inference and model selection. BioRxiv 047019.

Ségurel, L., Wyman, M.J., and Przeworski, M. (2014). Determinants of Mutation Rate Variation in the Human Germline. Annu. Rev. Genomics Hum. Genet. *15*, 47–70.

Sella, G., Petrov, D.A., Przeworski, M., and Andolfatto, P. (2009). Pervasive natural selection in the Drosophila genome? PLoS Genet. *5*, e1000495.

Siepel, A. (2009). Phylogenomics of primates and their ancestral populations. Genome Res. *19*, 1929–1941.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*.

Sillero-Zubiri, C., Gottelli, D., and Macdonald, D.W. (1996). Male philopatry, extra-pack copulations and inbreeding avoidance in Ethiopian wolves (<Emphasis Type="Italic">Canis simensis</Emphasis>). Behav. Ecol. Sociobiol. *38*, 331–340.

Simons, Y.B., Bullaughey, K., Hudson, R.R., and Sella, G. (2018). A population genetic interpretation of GWAS findings for human quantitative traits. PLOS Biol. *16*, e2002985.

Slotte, T. (2014). The impact of linked selection on plant genomic variation. Brief. Funct. Genomics *13*, 268–275.

Smith, N.G.C., Webster, M.T., and Ellegren, H. (2002). Deterministic Mutation Rate Variation in the Human Genome. Genome Res. *12*, 1350–1356.

Smith, T.C.A., Arndt, P.F., and Eyre-Walker, A. (2018). Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. PLOS Genet. *14*, e1007254.

Stahler, D.R., MacNulty, D.R., Wayne, R.K., vonHoldt, B., and Smith, D.W. The adaptive value of morphological, behavioural and life-history traits in reproductive female wolves. J. Anim. Ecol. *82*, 222–234.

Stevison, L.S., Woerner, A.E., Kidd, J.M., Kelley, J.L., Veeramah, K.R., McManus, K.F., Project, G.A.G., Bustamante, C.D., Hammer, M.F., and Wall, J.D. (2015). The Time Scale of Recombination Rate Evolution in Great Apes. Mol. Biol. Evol. msv331.

Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics *187*, 367–383.

Sundqvist, A.-K., Björnerfeldt, S., Leonard, J.A., Hailer, F., Hedhammar, Å., Ellegren, H., and Vilà, C. (2006). Unequal Contribution of Sexes in the Origin of Dog Breeds. Genetics *172*, 1121–1128.

Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature *521*, 81–84.

Tajima, F. (1983). Evolutionary Relationship of DNA Sequences in Finite Populations. Genetics *105*, 437–460.

Takahata, N. (1986). An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet. Res. *48*, 187–190.

Thalmann, O., Shapiro, B., Cui, P., Schuenemann, V.J., Sawyer, S.K., Greenfield, D.L., Germonpré, M.B., Sablin, M.V., López-Giráldez, F., Domingo-Roura, X., et al. (2013). Complete Mitochondrial Genomes of Ancient Canids Suggest a European Origin of Domestic Dogs. Science *342*, 871–874.

The UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

Timpson, N.J., Greenwood, C.M.T., Soranzo, N., Lawson, D.J., and Richards, J.B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat. Rev. Genet. *19*, 110–124.

Torgerson, D.G., Boyko, A.R., Hernandez, R.D., Indap, A., Hu, X., White, T.J., Sninsky, J.J., Cargill, M., Adams, M.D., Bustamante, C.D., et al. (2009). Evolutionary Processes Acting on Candidate cis-Regulatory Regions in Humans Inferred from Patterns of Polymorphism and Divergence. PLOS Genet. *5*, e1000592.

Tyekucheva, S., Makova, K.D., Karro, J.E., Hardison, R.C., Miller, W., and Chiaromonte, F. (2008). Human-macaque comparisons illuminate variation in neutral substitution rates. Genome Biol. *9*, 1–13.

Uricchio, L.H., Zaitlen, N.A., Ye, C.J., Witte, J.S., and Hernandez, R.D. (2016). Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. Genome Res.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinforma. *43*, 11.10.1-33.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

Voight, B.F., Adams, A.M., Frisse, L.A., Qian, Y., Hudson, R.R., and Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. U. S. A. *102*, 18508–18513.

Vonholdt, B.M., Stahler, D.R., Smith, D.W., Earl, D.A., Pollinger, J.P., and Wayne, R.K. (2008). The genealogy and genetic viability of reintroduced Yellowstone grey wolves. Mol. Ecol. *17*, 252–274.

vonHoldt, B.M., Pollinger, J.P., Earl, D.A., Knowles, J.C., Boyko, A.R., Parker, H., Geffen, E., Pilot, M., Jedrzejewski, W., Jedrzejewska, B., et al. (2011). A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res. *21*, 1294–1305.

Wakely, J. (2008). Coalescent Theory: An Introduction (W. H. Freeman).

Wall, J.D. (2003). Estimating ancestral population sizes and divergence times. Genetics *163*, 395–404.

Ward, L.D., and Kellis, M. (2012). Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. Science *337*, 1675–1678.

Webster, T.H., and Wilson Sayres, M.A. (2016). Genomic signatures of sex-biased demography: progress and prospects. Curr. Opin. Genet. Dev. *41*, 62–71.

Wilson Sayres, M.A., and Makova, K.D. (2011). Genome analyses substantiate male mutation bias in many species. BioEssays News Rev. Mol. Cell. Dev. Biol. *33*, 938–945.

Wilson Sayres, M.A., Lohmueller, K.E., and Nielsen, R. (2014). Natural Selection Reduced Diversity on Human Y Chromosomes. PLoS Genet *10*, e1004064.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

Wright, S.I., and Andolfatto, P. (2008). The Impact of Natural Selection on the Genome: Emerging Patterns in Drosophila and Arabidopsis. Annu. Rev. Ecol. Evol. Syst. *39*, 193–213.

Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., Hurst, L.D., and Tian, D. (2015). Parent-progeny sequencing indicates higher mutation rates in heterozygotes. Nature *advance online publication*.

Young, A.C., Kirkness, E.F., and Breen, M. (2008). Tackling the characterization of canine chromosomal breakpoints with an integrated in-situ/in-silico approach: The canine PAR and PAB. Chromosome Res. *16*, 1193–1202.

Zeng, J., Vlaming, R. de, Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. *50*, 746.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinforma. Oxf. Engl. *28*, 3326–3328.