

**DYNAMICS OF NATURAL SELECTION ON HUMAN GENOMIC
VARIANTS**

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

by
Alyssa M. Pivrotto
August 2024

Examining Committee Members:

Jody Hey, Advisory Chair, Biology

Sudhir Kumar, Biology

Rob Kulathinal, Biology

Christina Bergey, External Reader, Rutgers University

©
Copyright
2024

by

Alyssa M. Pivrotto
All Rights Reserved

ABSTRACT

Evolutionary adaptation in humans is shaped primarily by the selection of beneficial alleles. Classical population genetic theory predicts that alleles under selection will experience a rapid increase in frequency. However, the effects of weakly deleterious and neutral alleles propelled to high frequency due to drift complicates the identification of sites under positive selection. Evolutionary probability uses vertebrate alignments and divergence times to estimate a site's evolutionary history, assigning probability values to each potential amino acid at that site. For each mutation, a probability value can be calculated indicating whether the mutation is favored or disfavored evolutionarily. Because sites under selection will increase in frequency more quickly than they would due to genetic drift alone, it is expected that both beneficial and deleterious mutations will be younger on average than neutral variants of the same frequency. In chapter two, this was found to be true for the disfavored, deleterious mutations which were younger on average. Notably, beneficial mutations were found to be older on average than neutral mutations of the same frequency. One possible model suggests that the enrichment of old, beneficial alleles segregating in modern humans can be explained due to linked, weakly deleterious variants hindering the fixation of beneficial mutations until recombination allows for the escape of the beneficial mutation. Assessing allele age estimation methods is crucial for understanding the potential selection a mutation is undergoing. While whole genome sequencing data is becoming increasingly accessible, a large amount of the currently available data for large population datasets exists in the form of whole exome sequencing data. In chapter three, the accuracy of three allele age estimators, Genealogical Estimation of Variant Age (GEVA), Relate, and time of coalescence is tested for accuracy for both

whole genome and whole exome sequencing datasets. Relate was found to outperform both other estimators of allele age for both a simple (Pearson: 0.64) and complex (Pearson: 0.68) demography model with the estimates based on whole exome data having an average drop in performance of 16 percent in comparison with the whole genome estimates. Beyond investigating segregating variants, phylogenetic methods such as evolutionary probability allow for the analysis of fixed candidate variants and the investigation into potential mechanisms by which these favored alleles arose. In chapter four, derived sites which have become fixed in modern humans where non-human primates all share the ancestral amino acid are identified. Utilizing the fixed, derived sites and the corresponding evolutionary probability values, it can be tested if adaptation occurs due to novel, low evolutionary probability mutations. A second hypothesis can also be tested where instead adaptation occurs due to a mutation to a more evolutionarily stable amino acid. It was found that while the majority of substitutions in modern humans are both arising by way of novel amino acids, however this is no evidence that these substitutions are driving phenotypic adaptation in modern humans.

To my family – thank you for your endless support.

Mom – thank you for listening to me recount every moment of my entire life
since I was five years old.

Dad – thank you for your weekly “Bad Dad Joke” Friday messages.

To my friends – thank you for many years of laughs and support.

All my love always.

ACKNOWLEDGMENTS

First, so much of my gratitude goes to my advisor, Dr. Jody Hey. Thank you for your advisement, support, and mentorship. When I began this program, I knew little about bioinformatics and even less about population genetics, so thank you for taking a chance on me. Over the last six years, you have consistently supported me throughout not only my research but also in my work teaching and developing the Bioinformatics Studio. I also have so much gratitude to members of the Hey Lab over the last six years:

Alexander Platt, Andrew Webb, Jared Knoblauch, Arun Sethuraman, Lanxin Liu, Vitor Pavinato, and Noah Peles.

Many thanks to my committee members, Dr. Sudhir Kumar and Dr. Rob Kulathinal. Your advisement and support over the last six years have been so appreciated. Your comments, advice, and questions during my committee meetings, coffee hours, and at conferences have greatly improved my work. Thank you to Dr. Christina Bergey for many helpful suggestions and comments about this work – I very much appreciate your time, kindness, and your scientific expertise.

To my cohort of fellow Bioinformaticians who started with me in the Fall of 2018, when we began this program six years ago all of you intimidated me, but you turned into some of my greatest friends and cheerleaders. Thank you to Jordan Zehr, Alexander Lucaci, Steven Weaver, Francisco McGee, and Stella Park – you have all made me a better scientist, a better leader, and a better person. Thank you to Jordan, Alex, and Francisco – the work we did to establish the Bioinformatics Studio remains some of my fondest memories and one of my proudest achievements of our time here at Temple University.

To all my colleagues in the Bioinformatics program, I am constantly in awe of the way you approach science – I am a better researcher having met you. I enjoyed our many coffee breaks, lunch sessions, conference expeditions, and birthdays. Thank you to Lisa, Sam, Hannah K., Chunan, Chong, Hannah V., John, Albert, Mina, Jason, Rohan, Oscar, Jordan, Alex, Steven, Francisco, and Stella. Thank you for the many laughs, supporting me when there were tears, and cheering me on along the way.

During Spring 2020, a very crazy thing happened, and we ended up alone in our apartments for weeks and months on end. If it wasn't for the camaraderie that I found within my friends at Temple, I cannot imagine where I would be now. Thank you to Amanda Wilson, Avery Selberg, Maddie McGhee, and Molly Schools for many early morning check-in meetings for months on end. My eight AM starts were always better with all of you.

Thank you to all the friends I met through TUGSA. Your strength, bravery, and positivity in the face of the strike at Temple University in Spring 2023 inspired me then and continues to inspire me every day. I am thankful to have met every single one of you.

Thank you to all my friends who have supported and encouraged me over the last six years. Thank you to Sabrina who has always believed in me and who has been one of my biggest cheerleaders. Thank you to Kimi who has cheered so loudly for me along the entire way and has made many a fretful trip to Philadelphia for me. Thank you to Avery for always reminding me that I can do anything and for threatening to move in with Rosie any time I decide I'd be better off running a food truck. Thank you to Maddie who somehow has a sense for when I'm stressed and always knows when a phone call is needed (and for traveling from Boston to Philadelphia to clean my apartment). Thank you

to Konrad for many mornings and afternoons writing from a coffee shop and for reminding me that the world isn't going to end (even when I think it is). Thank you to Andrew for encouraging me to take a break, always organizing games with friends, and never being competitive. Thank you to Molly and Jordan for many, many brunches over the years. Thank you to all my friends: Amanda, Andrew E., Andrew S., Avery, Bryn, Katie A., Katie M., Kimi, Konrad, Kylee, Logan, Maddie, Mary, Michaela, Molly S., Molly Z., Paige, Sammie, and Sabrina. Your never-ending support has meant the world to me.

Finally, so much of my thanks goes to my family. I have almost no words to thank you for the tremendous support you've provided me with throughout my life. Thank you for reading my papers even when they don't make sense. Thank you for listening to me ramble even when it goes on entirely too long. Thank you for everything that you do.

TABLE OF CONTENTS

	Page
ABSTRACT.....	iii
DEDICATION.....	v
ACKNOWLEDGMENTS	vi
LIST OF TABLES.....	xiii
LIST OF FIGURES	xiv
CHAPTER	
1. INTRODUCTION.....	1
2. ANALYSES OF ALLELE AGE AND IMPACT REVEAL HUMAN BENEFICIAL ALLELES ARE OLDER THAN NEUTRAL CONTROLS	11
Abstract.....	11
Introduction.....	11
Methods.....	16
Evolutionary Probabilities, Allele Frequencies, and Data Filtering	16
Allele Age Estimates.....	16
Rooting.....	17
Calculating ΔEP	18
Noncoding Variants as Neutral Controls	18
ANOVA	19
Rank Analysis	20
Recombination Analysis.....	20
F_{ST} Analysis.....	21
Heterozygosity Analysis	21

Dispersion Analysis	22
Gene Ontology Analysis	22
Comparison to Archaic Genomes	23
$\beta^{(2)}$ Values	23
Results and Discussion	24
Summary of Segregating and Fixed Derived Nonsynonymous Alleles	24
Deleterious Mutations are Younger on Average While Beneficial Mutations are Older on Average Than Neutral Mutations of the Same Frequency.....	25
Characterizing Old, Segregating, Positive ΔEP Alleles.	29
Examination of Modes of Balancing Selection: Population Structure and Overdominance.....	31
Models That can Account for a Period of Balancing Selection.....	32
Implications for the Adaptation of Human Populations.	36
3. ALLELE AGE ESTIMATORS DESIGNED FOR WHOLE GENOME DATASETS SHOW ONLY A MODEST DECREASE IN ACCURACY WHEN APPLIED TO WHOLE EXOME DATASETS.	39
Abstract.....	39
Introduction.....	40
Methods.....	43
Simulation Dataset	43
Convert to Exome Data.....	44
True Values	44
Allele Age Estimates.....	44
Statistics	46
Comparisons Across Sample Size	46
Simulating Sites With Background Selection.....	46

Results.....	47
Relate Outperforms With a Simple, Constant Population Size Model of Neutral Variation.....	47
All Three Estimators Show Relative Robustness to Demography and Selection.....	49
Rare Variants Show No Decrease in Accuracy of Allele Age Estimates in Comparison with Common Variation.	55
Discussion.....	56
4. SUBSTITUTIONS ON HOMININ BRANCH AROSE VIA NOVEL AMINO ACIDS BUT SHOW NO EVIDENCE OF ADAPTATION	61
Abstract.....	61
Introduction.....	62
Methods.....	64
Dataset Curation.....	64
Dealing with Low Coverage Regions.....	65
Assembling Gene List.....	66
Statistical Methods.....	67
Potential Distribution of Evolutionary Probabilities.....	67
Identification of Genes Under Selection Using ABSREL.....	67
Gene Enrichment Analysis	68
Results.....	68
Examination of Dataset Reveals That the Majority of Genes Only Have One Substitution.	68
The Majority of Substitutions Occur via Novel Amino Acids on the Hominin Branch.....	69
Comparison of Observed ΔEP to Expected, Simulated ΔEP Distributions Show Differences Which may be Indicative of Sites Under Selection.....	71

Commonly Identified Phenotypic Traits of Adaptation in Humans do not Show any Significant Difference in ΔEP Values.....	75
No Evidence of Enrichment of Substituted Sites for Genes Identified to be Under Positive Selection.	78
Discussion.....	79
REFERENCES	83
APPENDICES	
A. SUPPLEMENTAL MATERIAL FOR CHAPTER 2.....	98
B. SUPPLEMENTAL MATERIAL FOR CHAPTER 3.....	109
C. SUPPLEMENTAL MATERIAL FOR CHAPTER 4.....	119

LIST OF TABLES

Table	Page
1. ΔEP measures for fixed and polymorphic alleles.	25
2. Pearson's correlation coefficients for simple and complex model for WES and WGS datasets.	51
3. Statistics for sample populations.	66
4. Counts of fixed, negative ΔEP alleles between human and non-human primates.	78
5. Counts of fixed, positive ΔEP alleles between human and non-human primates.	78
6. ΔEP measures for fixed and polymorphic alleles.	98
7. Results of simulation-based tests of dispersion of positive ΔEP SNPs.	99
8. Gene ontology results	100
9. F_{ST} Values across ΔEP spectrum of values.	106
10. Statistical power for detecting excess heterozygosity.	107
11. Constant population simulation parameters.	110
12. Complex population simulation parameters	111
13. Summary statistics from three estimators for simple and complex model.	113
14. Selected genes identified by aBSREL.	122
15. GSEA results for positively selected genes from aBSREL.	124
16. Human substituted sites identified in positively selected genes.	125

LIST OF FIGURES

Figure	Page
1. EP for non-synonymous SNPs binned by allele frequency.	14
2. Mean derived-allele frequency binned by ΔEP values.	15
3. Illustrative example of mutational event.....	26
4. Mean allele age estimates from GEVA binned by allele frequency.....	27
5. Age rank as a function of ΔEP	28
6. Mean recombination rate per base per generation as a function of ΔEP for fixed and segregating alleles.....	35
7. Figurative example of the frequency trajectory of an allele under the staggered sweep (SS) or diploid fisher's geometric (DFG) model.....	37
8. Estimator comparison on WES data for simple model.....	50
9. Estimator comparison on WES data for a complex model.....	53
10. Correlation between true and estimated allele age increases with sample size.	54
11. Relate allele age estimates compared to true values under background selection.	55
12. Error in estimates across spectrum of frequency values for the simple model.....	56
13. Distribution of substitution counts per gene.	69
14. Distribution of ΔEP values for fixed coding changes in modern humans.	70
15. Simulated and Observed EP values for low EP ancestral amino acids.....	74
16. Simulated ΔEP values for invariant sites in primates.	75
17. Distribution of ΔEP values for each of the three gene sets and the control housekeeping gene set.....	77
18. Distributions of derived polymorphism frequency in UK10K.	108
19. Depiction of simulation pipeline.....	112
20. Estimator comparison on WGS data for simple model.	114
21. Estimator comparison on WGS data for complex model.	115

22. RMSLE across entire frequency spectrum of mutations.	116
23. Normalized RMSLE across frequency spectrum for the complex model	117
24. Comparison of Relate on samples of 100, 1000, and 10000 genomes	118
25. Distribution of ΔEP values for fixed coding changes in chimpanzees.	119
26. Distribution of difference in observed and simulated ΔEP values for loci where the ancestral and derived amino acids have a similar EP.....	120
27. Distribution of difference in observed and simulated ΔEP values for loci where the ancestral amino acid is preferred.....	121
28. Distribution of ΔEP values for substitutions in positively selected genes.....	126
29. Distribution of ΔEP values for fixed coding changes in archaic humans.....	127

CHAPTER 1

INTRODUCTION

The trajectories of beneficial alleles underlie the process of adaptation in humans. However, while alleles under positive selection will rise in frequency quickly towards fixation (1,2), alleles that are weakly deleterious or neutral will also rise to high frequencies and fix some of the time due to genetic drift (3,4). Thus, it is not sufficient to solely identify sites that have been fixed in modern humans to find sites that are candidates of adaptation. Complicating the issue further, there is very little evidence of classic sweeps of beneficial mutations in humans (5,6) with the majority of evidence supporting soft sweeps of standing variation to fixation instead (7).

There are millions of segregating variants in human populations with some portion of these mutations arising in coding regions potentially affecting the functionality of the protein. Of these segregating, coding sites, the majority are found at low frequency and are deleterious, however some portion of these are beneficial based on estimated proportions from fixed sites (8). As these sites are segregating, they present a unique opportunity to investigate the trajectory of beneficial mutations during the process of rising in frequency towards fixation. Mutations under selection, both negative and positive, should be younger on average than neutral alleles of the same frequency (1). Following the theoretical prediction that sites under selection will be younger, researchers can empirically test for potential sites undergoing selection as these sites will appear to be pulled from a different distribution of ages. This has previously been shown to be true empirically for deleterious variants (9), but an empirical analysis to examine both favored and disfavored had yet to be undertaken. Until recently these types of analyses have been

intractable due to data availability and method scaling. However, since the sequencing of the first human genome (10,11), the number of genomes sequenced has exponentially increased yielding more candidate sites arising and the potential to investigate both the selection and fitness effects of segregating mutations (12,13). In chapter two, the theory that sites under selection are younger on average than their neutral counterparts are tested in an analysis comparing the distribution of mutation age for neutral mutations with both positively and negatively selected nonsynonymous mutations.

To identify whether an allele is evolutionarily favored or disfavored, the likelihood of an amino acid at a given position is estimated by the evolutionary probability (EP) based on a posterior probability of each amino acid given a vertebrate alignment (14). Those amino acids that appear very often in the vertebrate tree at that site, or amino acids found in closely related species to the focal species of interest, will receive a higher probability. On the other hand, if an amino acid is not found anywhere on the tree, then it will receive a very low probability value. Previous work has found that there are many sites in the human genome where humans have a low probability amino acid that is segregating at moderate frequencies which may be candidates of adaptation (15,16).

Leveraging the ability to identify whether a site is either evolutionary likely or unlikely allows for further questions about adaptations in modern humans to be addressed. Previous studies have found “evolutionary forbidden” sites, where the evolutionary probability is less than 0.05, segregating at high frequencies (16) and some proportion of high frequency evolutionarily forbidden sites likely have been fixed during the evolutionary history of modern humans. In chapter four, the evolutionary probability

of substitutions in modern humans is used to test theories of the mechanism of adaptation. Sites where modern humans have a fixed amino acid that has a high evolutionary probability, but closely related non-human primates have a low EP amino acid at the site represent loci where humans are returning to a more evolutionarily stable amino acid for the protein. In the opposite case, sites where humans have the low EP amino acid and the ancestral allele shared by closely related species is the high EP amino acid are loci where humans may have experienced a sweep towards a novel amino acid. The derived alleles that have become fixed in modern humans where closely related non-human primates all share the ancestral allele represent candidates of adaptation. Utilizing evolutionary probability, the larger question of the evolutionary mechanism by which adaptation is occurring can be tested by investigating whether fixed sites are arising due to a return to a more evolutionarily likely amino acid or towards an unlikely amino acid.

In the identification of segregating beneficial alleles in chapter two, estimated allele ages were used to investigate selection by comparing the age of presumed neutral variants to presumed selected nonsynonymous variants. These estimators assume an infinite sites model meaning each locus would only have a single mutation and it's assumed that estimates are generated from whole genome sequencing data. These estimators have been shown to be robust to errors in the data, however a study examining the performance on missing data has yet to be performed. While segregating SNPs have commonly been identified through whole genome sequencing (WGS) for population studies, mutations are also often commonly identified using either genotype data or whole exome sequencing (WES). Genotype data focuses on calling the alleles at sites of common polymorphism which are selected beforehand using a targeted assay (17). Often

used in livestock and agriculture research, genotyping is a cheaper option, but does not allow for identification of non-common mutations and identifies the fewest number of sites of the three methods.

Both WGS and WES can identify increasing non-common variants as the number of samples sequenced increases. In WGS, DNA can be sequenced via long-read or short-read technology. Short-read sequencing first cleaves the DNA into fragments and then these fragments are sequenced into their base pairs (18). These fragments are often just 50-300 base pairs in length, and then require assembly where the fragments are assembled into contigs and then into scaffolds where in the final step the gaps in the sequence are filled to get the final genomic sequence (19–21). Long-read sequencing in comparison sequence long strings of base pairs without the fragmenting present in short-read technology (19). In this method, the base pair sequences can range in length from a kilobase to several Mbps (22). While long-read data has a higher error rate than short-read data, long-reads can sequence repeating elements and structural variation more accurately (19).

Exomes are sequenced using short-read sequencing technology except for the added step of targeting specific coding regions by hybridization (23). The hybridized coding DNA fragments can undergo sequencing yielding fragments in the target coding regions (23). As the cost of sequencing increasingly diminishes, the amount of genomic data has increased globally both in terms of whole genome datasets and whole exome datasets (24,25). Estimating the allele age of single nucleotide polymorphisms (SNPs) in a population is informative about whether an allele is undergoing selection. This is particularly of interest for rare variants as it has been found that mutations found at low

frequency contribute significantly to the phenotype of many diseases (23,26–28).

Understanding of the fitness of segregating variants contributes to the understanding of distribution of fitness effects (DFE) of new mutations which has direct implications to human disease (29).

The number of exomes sequenced has increased significantly over the last several years leading to more uncommon variants being discovered. Prior work has speculated that rare variants contribute to common disease phenotypes (30,31) and recent research has shown evidence of rare variants being important contributors to many phenotypes from development cognition (32) to the cardiovascular system (33). Tools to investigate increasingly rare variants in large datasets allow researchers to identify potential causative variants and also create an expanding dataset of information about a given mutation which can be applied in clinical settings (34). As rare variants have been found to have an impact on gene expression (27) and the contribution of many major diseases (13,28), it becomes increasingly important to have accurate tools to better understand the trajectory and fitness of uncommon variation. In chapter three, the accuracy of allele age estimators on both whole genome and whole exome sequencing data is ascertained. Testing the accuracy of common methods of allele age estimation using exome data allows for these tools to become accessible for use in large clinical datasets and in other instances where whole exome sequencing is preferred such as in non-model organism studies.

The estimate of the time a mutation enters a population has been a topic of research for several decades as the distribution of ages at a given frequency can be indicative of the selective pressures on the alleles in a population (35,36) along with

information about demography (37). An allele under either purifying or adaptive selection will rise to its given frequency much more quickly than a neutral allele of the same frequency (1). Thus, at a given frequency those alleles undergoing selection will be younger on average than their neutral counterparts of the same frequency. Early theoretical predictions on allele age estimation focused on estimating either when a mutation would become fixed in the population or when it would be lost (2,38). An extension to these early methods evolved to include segregating variants by leveraging allele frequency and population size as parameters to estimate the time on average it would take for a mutation to reach a given frequency (39). These methods are predicated on the allele being neutral in a population, but alleles under selection will be of an even greater interest as the drivers of functional changes leading to human disease. For deleterious alleles, a prior model was extended to include a dominance coefficient due to the mutation's age in the population being dependent on the heterozygote state to be retained in the population (40). This model considers just a constant population size, but in realistic populations this would not be accurate, and the model can be extended to a coalescent framework to include varying population sizes (41–45). At the end of the wave of early research on allele age, essentially two types of empirical methods emerged based on their theoretical predecessors: one, methods that leverage allele frequency, and two, methods that identify variation between samples in a population (43). In the last several years, these methods have been refined even further to focus almost exclusively on the latter methods of utilizing intra-allelic variation with genomic data.

Recently published methods either estimate the entire ancestral recombination graph (ARG) or specifically estimate the time of the branches that contain the mutation of

interest. Along the genome, the relationship between samples can be reconstructed identifying the time any two samples last shared a common ancestor, a coalescent event. The ARG expands on this information by also reconstructing the recombination events where two samples instead share a common ancestor due to recombination instead of a coalescent event (46,47). To identify these events in a tree, methods identify shared stretches of haplotype tracks between samples where a shared track is terminated by a recombination event. Between two samples that share a common ancestor and thus a shared tract of DNA, mutations will accumulate on that haplotype providing further information for estimating the age of the branch. Three recent estimators of allele age are Relate (48), Genealogical Estimator of Variant Age (GEVA) (49), and time of coalescence (t_c) (50). Even though these methods are tested for accuracy using whole genome data, based on the robustness of these models it is conceivable that these methods can be extended to estimate allele age from whole exome data. However, this requires careful examination of the assumptions and methodology of these estimators and can be empirically tested with an analysis of the accuracy of the estimates from whole exome data.

Relate employs a hidden Markov model (HMM) to calculate posterior probabilities to create a distance matrix of the relatedness of any two samples from the dataset (48). The HMM Relate uses is based on the method the Li and Stephens (51) except modified to explicitly identify the ancestral and derived states. Using the distance matrix, a tree is constructed using a clustering algorithm. Once a genealogical tree has been estimated, then the identified mutations along the haplotypes can be mapped onto the tree. To identify the times of the coalescent events and the subsequent branch lengths,

a Markov Chain Monte Carlo (MCMC) algorithm is utilized with a coalescent model as the prior. Relate has the added benefit of also estimating varying coalescent rate and changing mutation rates over time.

GEVA (49) first identifies the sets of haplotypes that share the focal allele (“concordant pairs”) and those sets of samples that do not have the focal allele (“discordant pairs”). For each set of pairs of haplotypes, a hidden Markov model (HMM) is used to extend the shared haplotype to either side of the focal allele until it identifies a breakpoint due to recombination. Since the mutation had to have arisen sometime prior to the time of the concordant pairs have last coalesced but after the time of the discordant pairs’ coalescence event, then the time of the mutation arose can be modeled by the distribution of each set of haplotypes estimated ages. The age of each set of haplotypes can be estimated as the posterior distribution using both the recombination distance and the number of mutations on the shared haplotype. GEVA outputs results for a mutation model, a recombination model, and a joint model of both recombination and mutation. For each of the models, the mean, median, and mode of the distribution of posterior probabilities for allele age.

For the estimator of time of coalescence (t_c), for a given focal mutation the shared samples that contain this focal mutation are identified and those samples that do not contain the focal mutation but appear in the sister clade to the ones that contain the focal mutation are identified (50). In the case of mutations that just occur a single time (singletons), these can be estimated based on just identifying the sister chromosome or clade to the focal mutation. Based on this focal site, looking at population of genome samples, the maximum tract of shared DNA or the maximum shared haplotype is found

by extending to both the left and right of the focal site. This longest shared tract will represent the focal sites sister clade or chromosome most closely related to those samples that contain the focal mutation. In considering just a single chromosome as the sister to the focal chromosome for simplicity, the shared length of haplotype tract shared between the two is broken when a mutational or recombination event occurs in one of the two genomes. This length of haplotype can be modeled as a function of the time of which these two samples last shared ancestry. A maximum likelihood model is used to estimate the age of the focal mutation using the length of the maximum shared haplotype along with the recombination and mutation rates.

All three methods of allele age rely on the distribution of mutations on a haplotype to estimate the time of the branch containing a focal variant of interest. While genome data provides the most complete information and exome data does not have the complete distribution of mutations as all non-coding mutations will be missing, these methods may be able to still provide relatively accurate estimates enough to be informative about selection to identify candidates for functional impact.

Classic predictions under population genetic theory can be tested such as that sites under selection will be younger on average than neutral sites of the same frequency. This is particularly important because sites under selection represent both deleterious sites that impact human disease and adaptations that have become fixed due to positive selection. For segregating sites, beneficial alleles can be directly investigated using evolutionary probability and estimators of mutation age to compare the distribution of allele ages of beneficial alleles to neutral ones. Leveraging this evolutionary method, larger questions about adaptation in modern humans can be investigated. Through the identification of

fixed, derived amino acids in modern humans where non-human primates have an ancestral amino acid, the likelihood of each amino acid at the locus can be calculated.

Using this method, identification of potential adaptive derived mutations can be identified and further examined.

CHAPTER 2

ANALYSES OF ALLELE AGE AND IMPACT REVEAL HUMAN BENEFICIAL ALLELES ARE OLDER THAN NEUTRAL CONTROLS

Abstract

A classic population genetic prediction is that alleles experiencing directional selection should swiftly traverse allele frequency space, leaving detectable reductions in genetic variation in linked regions. However, despite this expectation, identifying clear footprints of beneficial allele passage has proven to be surprisingly challenging. We addressed the basic premise underlying this expectation by estimating the ages of large numbers of beneficial and deleterious alleles in a human population genomic data set. Deleterious alleles were found to be young, on average, given their allele frequency. However, beneficial alleles were older on average than non-coding, non-regulatory alleles of the same frequency. This finding is not consistent with directional selection and instead indicates some type of balancing selection. Among derived beneficial alleles, those fixed in the population show higher local recombination rates than those still segregating, consistent with a model in which new beneficial alleles experience an initial period of balancing selection due to linkage disequilibrium with deleterious recessive alleles. Alleles that ultimately fix following a period of balancing selection will leave a modest ‘soft’ sweep impact on the local variation, consistent with the overall paucity of species-wide ‘hard’ sweeps in human genomes.

Introduction

Evolutionary adaptation depends upon the spread and fixation of beneficial alleles, however some neutral and slightly deleterious alleles also drift to high

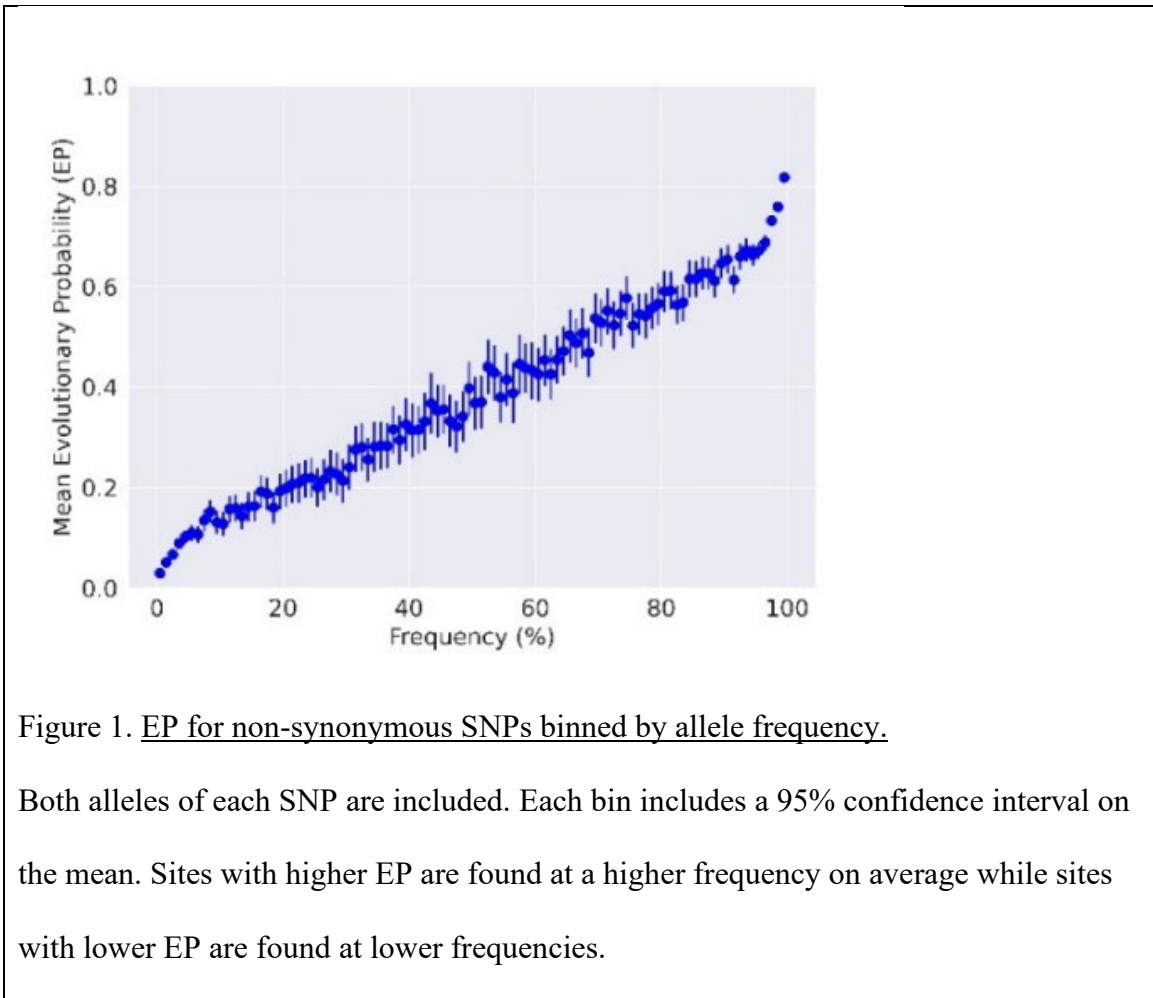
frequencies and become fixed, and so investigators have long sought ways to distinguish the fixation processes of adaptive alleles from those that are non-adaptive. Most methods are based on the classic population genetic prediction that beneficial alleles should move quickly through the range of allele frequencies (1,2) and leave a significant footprint on levels and patterns of linked variation (52). However, despite evidence that the fixation of beneficial alleles is common (53–55), investigators have found few instances where individual fixation events have left a clear footprint (5,6,56). In the human context, this has been particularly puzzling given that other methods suggest that there have been thousands of adaptive amino-acid substitutions in the human lineage since the common ancestor with chimpanzees (8,53,57–59). Consequently, much research in recent years has been devoted to understanding the fixation process of beneficial alleles and the kinds of impacts that may be left in contexts of multiple mutations (60,61), changing selection coefficients, selection at linked sites (62), and population structure (63–65).

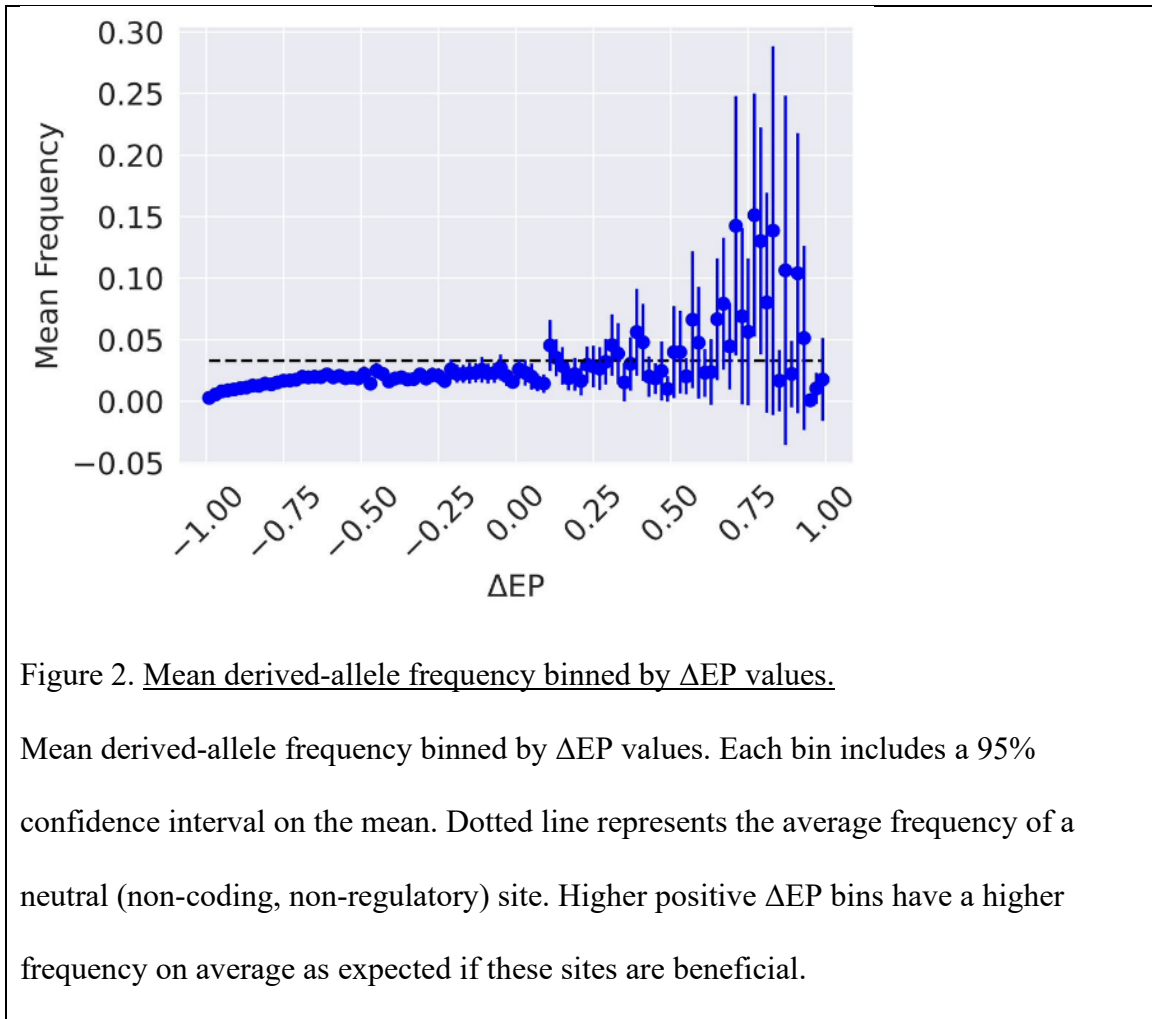
To better understand the allele frequency trajectories of beneficial alleles, we undertook a new kind of analysis that combines two unrelated advances of recent years, one that can identify a large number of segregating beneficial and deleterious alleles, and another that estimates allele age. Our initial goal was to test the fundamental population genetic prediction that alleles under directional selection should be younger, on average, than neutral alleles of the same frequency. This expectation was clearly affirmed for candidate deleterious alleles; however, the analysis revealed a striking pattern in which candidate beneficial alleles are older on average than neutral alleles.

For nonsynonymous single nucleotide polymorphisms (SNPs) in a whole-genome sequencing study of over 3600 individuals from the United Kingdom (13), we identified

candidate alleles under selection using the evolutionary probability (EP) of amino acids residing at each position in 17,209 autosomal genes calculated from a multi-species protein sequence alignment (16). EP estimates are based on alignments of a large number of vertebrate genomes and do not depend on the alleles currently segregating in a population or their frequency. The use of EP estimates for identifying alleles under selection is well supported by simulation (15), and they are increasingly used to identify nonsynonymous changes that are candidates for adaptive changes (66–70). As shown in Figure 1, EP values correlate with allele frequency, with common alleles tending to have higher EP values as expected if high EP alleles are favored by selection more than are low EP alleles.

We rooted non-synonymous variants using the inferred ancestral sequence from Ensembl (71) and a maximum likelihood estimator. We defined ΔEP as the derived allele EP minus the ancestral allele EP. The large majority of derived alleles are at low frequency, as expected from basic theory (72), and we observed that mean derived allele frequency increases for sites with higher positive ΔEP (Figure 2), as expected if they are favored by natural selection (73,74).





To consider the ages of alleles predicted to be under directional selection, we used a large control set of non-coding, non-regulatory SNPs. These will necessarily have experienced similar mutational and recombinational processes, as well as the same demographic history, that non-synonymous SNPs have experienced, and they offer the ideal landscape upon which to inquire of the impact of selection on allele age.

Methods

Evolutionary Probabilities, Allele Frequencies, and Data Filtering

Non-synonymous SNP sites in the UK10K dataset were identified with their corresponding transcript ID using the hg19 RefGene annotations in the UCSC table browser (75), that are based on NCBI RefSeq annotations (76), and the UK10K VCF (Variant Calling Format) files (13). For each two-allele polymorphism, the transcript IDs and site locations were used to retrieve the EP values for both the reference and alternative alleles. EP values were estimated using the method described in previous literature (14,16) using posterior probabilities from a multispecies alignment with associated divergence times. Mutations excluded from this dataset include those with uncurated transcript IDs that have not been verified. Frequency data for the reference and alternative allele at each site was extracted directly from the VCF file. Analyses thought to be sensitive to high mutability in CpG regions were limited to SNPs that did not occur as part of a CpG site. These included analyses that utilized allele ages (Figures 4, 5A, and 5B) as mutation rate was used as a parameter in estimating these values.

Allele Age Estimates

To get approximate allele age estimates, we used both the time of most recent coalescence (t_c) estimator (50) from the Hey Lab and the Genealogical Estimation of Variant Age (GEVA) estimator (49). To estimate t_c , for each of the autosomal chromosome VCF files, first the singletons were phased by placing each singleton on the longer of the two haplotypes. Following this step, the time of coalescence was estimated (runtc.py) using the following parameters: k-range, mutation rate of $1e-8$, and recombination map as HapMap Phase II genetic map for hg19 (77). To obtain GEVA

estimates, the VCF file for each autosomal chromosome was first parsed and converted into a binary file with corresponding marker and site files containing information per variant. GEVA values were obtained for all positive EP SNPs with more than two copies of the derived allele. GEVA estimates were obtained using the default parameters of effective population size of 10000, mutation rate of $1e^{-8}$, and the provided Hidden Markov Model (HMM) probability files. The output estimated age files were then filtered using the provided program in R (<https://github.com/pkalbers/geva>).

GEVA estimates were obtained for all positive Δ EP sites in the sampled genes (2729 in total). Because of time constraints large random samples of sites were used for non-coding, non-regulatory sites (71628 in total) and negative Δ EP sites (19053 in total). To generate figures with binned Δ EP values, the number of sampled noncoding, non-regulatory sites range from 800 to 2500 sites with estimated ages. For the negative Δ EP bins have approximately 1000 to 4000 sites with estimated ages, while the positive Δ EP bins have 60 to 600 with estimated ages.

Rooting

Two methods of rooting were used, a parsimony-based approach using Ensembl (78) and a maximum likelihood approach using RAxML (79). For the parsimony-based rooting method, estimates of the hg19 ancestral states were retrieved from Ensembl (71) and included for each position in the dataset. For all analyses of allele age, SNPs were limited to those where the ancestral allele state matched the reference allele. For maximum likelihood rooting a primate alignment was extracted for each RefSeq annotated gene from an Ensembl alignment whole genome alignment (<http://ftp.ensembl.org/pub/release-104/maf/ensembl->

[compara/multiple_alignments/12_primates.epo/12_primates.epo.10_1.maf.gz](#)) (78). The phylogeny for each gene was estimated using RAxML-NG using the model GTR+ Γ . At positions in each gene where there was a non-synonymous mutation in the UK10K dataset, the human sequence base in the alignment was replaced with a missing value, N. Using this newly constructed primate alignment with the modified human sequence to reflect UK10K mutations, RAxML-NG was run again to estimate the base pair values at the base of the edge of the human sequence. The output generated posterior probability estimates for each of the four nucleotides at each non-synonymous SNP site. Using the posterior probabilities, the most likely ancestral state was predicted as the base pair with the highest probability. Downstream analyses were filtered by those sites where a single base pair has a probability above 0.9 indicating a higher certainty for the ancestral state.

Calculating ΔEP

Values of ΔEP were calculated by finding the difference between the derived EP value and the ancestral EP value for a position given an estimated ancestral state for that position. The ΔEP metric indicates the difference from neutrality at a given site between the ancestral and derived allele. Sites where the amino acid mutated from an unlikely state evolutionarily to a more likely state yielded a positive ΔEP value, and in the reverse, sites where the amino acid mutated from a more likely state to less likely state yielded a negative ΔEP value.

Noncoding Variants as Neutral Controls

To account for allele frequency in our analyses of age across the spectrum of ΔEP values, a method to report age in relation to similar frequency control variants was needed. To assess whether an allele was young or old, each allele was compared to a

large control set of alleles of the same frequency. For this purpose, we used the ages of noncoding, non-regulatory alleles, treating them as a neutral control set. Candidate SNPs for the control set were first identified from intergenic regions using annotations from SNPeff Human Genome build GRCH37 Ensembl release 75 (80,81). This set was then filtered to remove those in regulatory regions, identified as falling into at least one of three data sets available from the UCSC Genome Browser: Candidate cis-Regulatory Elements by ENCODE (82); RefSeq Functional Elements (83); and curated regulatory annotations in the ORegAnno database (84).

Noncoding alleles in non-regulatory regions were assembled into bins of a similar frequency. Of the variants that have identified ancestral states matching the reference allele, noncoding, non-regulatory variants were split into bins of approximately 75,000 variants per frequency bin. At the lower end of frequency bins ($k = 1, 2, 3, 4, 5, 6, 7, 8$), same k value variants were kept together even if this resulted in bins larger than a size of 75,000 variants. In higher frequency bins, several k values were binned together to yield bins of an approximate size of 75,000 noncoding, nonregulatory variants.

ANOVA

To test the hypotheses that neutral derived allele ages have the same average allele age as either beneficial or deleterious alleles we used two-way ANOVA, with selected vs control as one effect, and allele frequency bin as a second effect. We first applied the Box-Cox transformation (85) to GEVA estimates of allele age for each treatment and allele frequency group.

Rank Analysis

To account for differences in allele ages between different frequency bins and to compare variants across the genome, we implemented a ranking system to assign each variant a rank within their own null frequency distribution. Initially, null distributions of noncoding, nonregulatory variant ages were constructed as described above. For each non-synonymous variant remaining in the filtered dataset, the corresponding frequency bin was identified based on the k value of the derived allele at that site. Within the null distribution of ages that correlated to the frequency bin for the focal non-synonymous mutation, the position of the focal mutation's age within the null distribution was found. Based on that position, the rank within the null distribution was calculated as the position divided by the length of the null distribution (approximately 75000 variants). This yielded a corresponding rank for each non-synonymous variant based on its own specific null distribution of ages from similar frequency variants.

Recombination Analysis

To identify the changes in recombination across the genome, we found associated recombination rate values for every segregating and fixed non-synonymous site in the UK10K dataset. With all segregating and fixed non-synonymous sites identified using the rooting method described above, the recombination rate at that location was extracted from the genetic map file for the specific demographic in the dataset. In this case, a UK population specific recombination map (86) was used. With each site's associated recombination rate, comparisons were made between both fixed and segregating sites across the spectrum of ΔEP values.

F_{ST} Analysis

We examined the relationship between F_{ST} and ΔEP . In 1000 Genomes data (12), F_{ST} was calculated (87) for SNPs also found in the UK10K sample for three comparisons: pooled African samples versus pooled European and Asian samples, pooled European versus pooled Asian samples, and Great Britain sample versus Italian sample. Only SNPs with at least 10 copies of the derived allele in the pooled contrast populations were considered. Table 9 in Appendix A shows mean F_{ST} as a function of ΔEP for each contrast.

To test whether F_{ST} was higher for older positive ΔEP SNPs than for control SNPs of the same allele frequencies, the F_{ST} for each positive ΔEP SNP with age rank greater than 0.5 was placed in the ranking of F_{ST} for all control SNPs of the same derived allele frequency. A single classification Wilcoxon test was conducted on each contrast to test whether there was an excess of positive ΔEP SNPs with F_{ST} ranking above 0.5.

Heterozygosity Analysis

A test was conducted for the hypothesis that positive ΔEP SNPs have higher heterozygosity than control SNPs of the same allele frequency. For each positive ΔEP SNP, the rank position of the observed count of the number of heterozygotes was determined by placing the observed count into a sorted list of heterozygote counts for controls SNPs with the same derived allele frequency. In case of ties, the rank position was a random value of all possible ranks with the same heterozygote count. To test the hypothesis that positive ΔEP SNPs have a mean rank above 0.5, a one-sided z-test was conducted.

A power analysis was conducted by simulating data sets of the same size and distribution of allele frequencies as the actual data. For a given selection coefficient s , where the fitness of a heterozygote is $1+s$, genotype frequencies were simulated using the observed allele count for each ΔEP SNPs in the data. Heterozygous counts were then placed in corresponding rankings of null distributions of heterozygous counts that were simulated for each of the observed allele frequencies of positive ΔEP SNPs. A z-test was conducted for each of 1000 simulated data sets for each selection coefficient. The results are shown in Table 10 in Appendix A.

Dispersion Analysis

To assess whether positive ΔEP SNPs are evenly distributed among the genes for which we have EP values, we simulated tree-sequence (88) samples of 7242 UK chromosomes using STDPOPSIM (89) under an Out-of-Africa model (90) for each of the autosomes. Then for each autosome mutations were simulated for each gene on that chromosome, using each gene's actual length and map position, at the same mean density as observed for positive ΔEP SNPS. The variance in simulated density of SNPs was recorded for each of 200 simulations for each autosome.

Gene Ontology Analysis

To test whether positive ΔEP SNPs appeared more often in specific molecular, biological, and cellular classes (GO database released 2022-07-01, DOI: 10.5281/zenodo.6799722), PANTHER pathways (91) and protein classes (version 17.0, released 2022-02-22), and Reactome Pathways (Reactome database version 77, released 2021-10-01), a PANTHER Overrepresentation Test (Release 20221013) was used (92,93). The analyzed set of genes were identified by counting the number of positive

Δ EP SNPs per gene. The number of positive Δ EP SNPs was normalized by gene length, and all genes with more than one positive Δ EP were retained. A final subset of 73 genes were used in the PANTHER GO term analysis. For the reference list, the gene database for *Homo sapiens* was used. Analyses were conducted with a Fisher's Exact test with a False Discovery Rate correction. Results are detailed in Table 8 in Appendix A.

Comparison to Archaic Genomes

In order to identify whether a large proportion of our sites of interest arose prior to the speciation between modern humans and archaic humans, we examined for each site whether it was also present in any one of four archaic genomes (94–97). For each category: nonsynonymous – Δ EP, nonsynonymous + Δ EP, and neutral noncoding sites, the number of shared loci with at least one archaic genome is reported along with percent of shared sites over the number of all sites in that category.

Not only was there interest in knowing whether these sites arose prior to the speciation event, but some subset of these sites potentially could be found in both modern human genomes and archaic human genomes due to gene flow between the two species. Sites were identified as appearing in introgression regions based on S^* values generated from the CEU dataset from 1000 Genomes (98) (available at <https://data.mendeley.com/datasets/y7hyt83vxt/1>). Sites annotated as matching in either Neanderthal or Denisovan would be included as introgression sites for our analysis.

$\beta^{(2)}$ Values

For $\beta^{(2)}$ scores (99), the CEU standardized scores generated from 1000 Genomes data was used (available at <https://zenodo.org/record/7842447>). For each site in our analysis, we identified from this published dataset the Beta2 score if available. A Mann-

Whitney U test was done to analyze the difference between the Beta2 values of the $-\Delta EP$ and $+\Delta EP$ distributions.

Results and Discussion

Summary of Segregating and Fixed Derived Nonsynonymous Alleles

With many rooted segregating and fixed SNPs, we can examine some basic expectations of positive and negative directional selection on non-synonymous mutations (Table 1, Appendix A Table 6). First, if adaptation operates primarily at the margins of optimality, then more non-synonymous variants will be harmful than beneficial, and the magnitude of effect for deleterious mutations should be greater on average than for beneficial mutations (100). We observe both patterns, with many more negative ΔEP alleles overall, and the mean absolute magnitude of ΔEP is much greater for negative ΔEP SNPs than for positive ΔEP SNPs (0.830 versus 0.274). Comparing fixed and segregating sites, it is expected that derived positive ΔEP alleles with a frequency of 1.0 will have larger ΔEP values than those in which both ancestral and derived alleles occur in the sample, which is confirmed (0.418 for fixed vs. 0.274 for segregating). The same prediction for negative ΔEP SNPs, with fixed alleles having a higher mean value than polymorphic alleles, was also confirmed (-0.685 vs. -0.830).

Table 1. ΔEP measures for fixed and polymorphic alleles.

Measure	Negative ΔEP		Positive ΔEP	
	Fixed (% or CI)	Polymorphic (% or CI)	Fixed (% or CI)	Polymorphic (% or CI)
# SNPs	23,456 (10.4%)	202,105 (89.6%)	3,308 (40.4%)	4,890 (59.6%)
Mean Derived Frequency	1.000	0.023 (0.022, 0.025)	1.000	0.073 (0.068, 0.080)
Mean Ancestral EP	0.725 (0.722, 0.728)	0.847 (0.843, 0.848)	0.154 (0.150, 0.159)	0.243 (0.240, 0.247)
Mean Derived EP	0.040 (0.039, 0.04)	0.017 (0.016, 0.018)	0.571 (0.565, 0.579)	0.517 (0.512, 0.522)
Mean ΔEP	-0.685 (-0.689, -0.682)	-0.830 (-0.834, -0.828)	0.418 (0.408, 0.427)	0.274 (0.267, 0.280)

Non-synonymous changes were rooted using the Ensembl ancestral sequence estimate (71). 95% confidence intervals on the mean, determined by bias-corrected bootstrap (101), are given in parentheses. See Table 7 in Appendix A for values based on maximum likelihood rooting.

Deleterious Mutations are Younger on Average While Beneficial Mutations are Older on Average Than Neutral Mutations of the Same Frequency.

Both positively and negatively selected alleles are expected to be younger on average than neutral alleles of the same frequency (1,9,39,43). We used the Genealogical Estimation of Variant Age (GEVA) method (49) to estimate the descendent node time, or coalescent time, for genes carrying the derived allele (Figure 3). We used RUNTC (50) to estimate t_c , the time of the ancestral node of the edge carrying the mutation (Figure 3). Rooted bi-allelic SNPs at non-coding, non-regulatory sites were used for a control set, identified hereafter as “neutral.” The t_c estimator is not a function of allele frequency,

and GEVA makes only limited use of allele frequency in the setting of priors for the recombinational landscape.

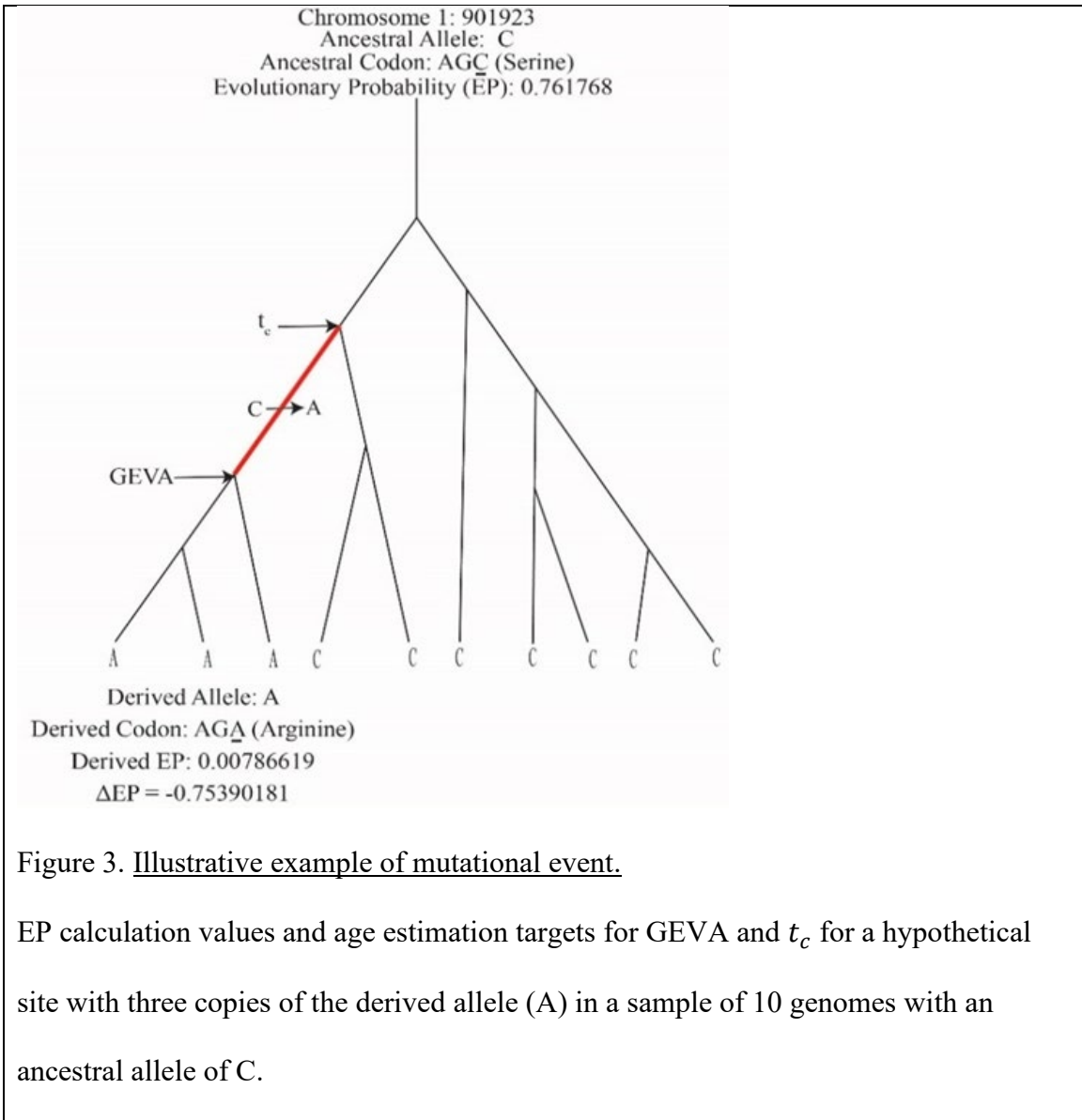
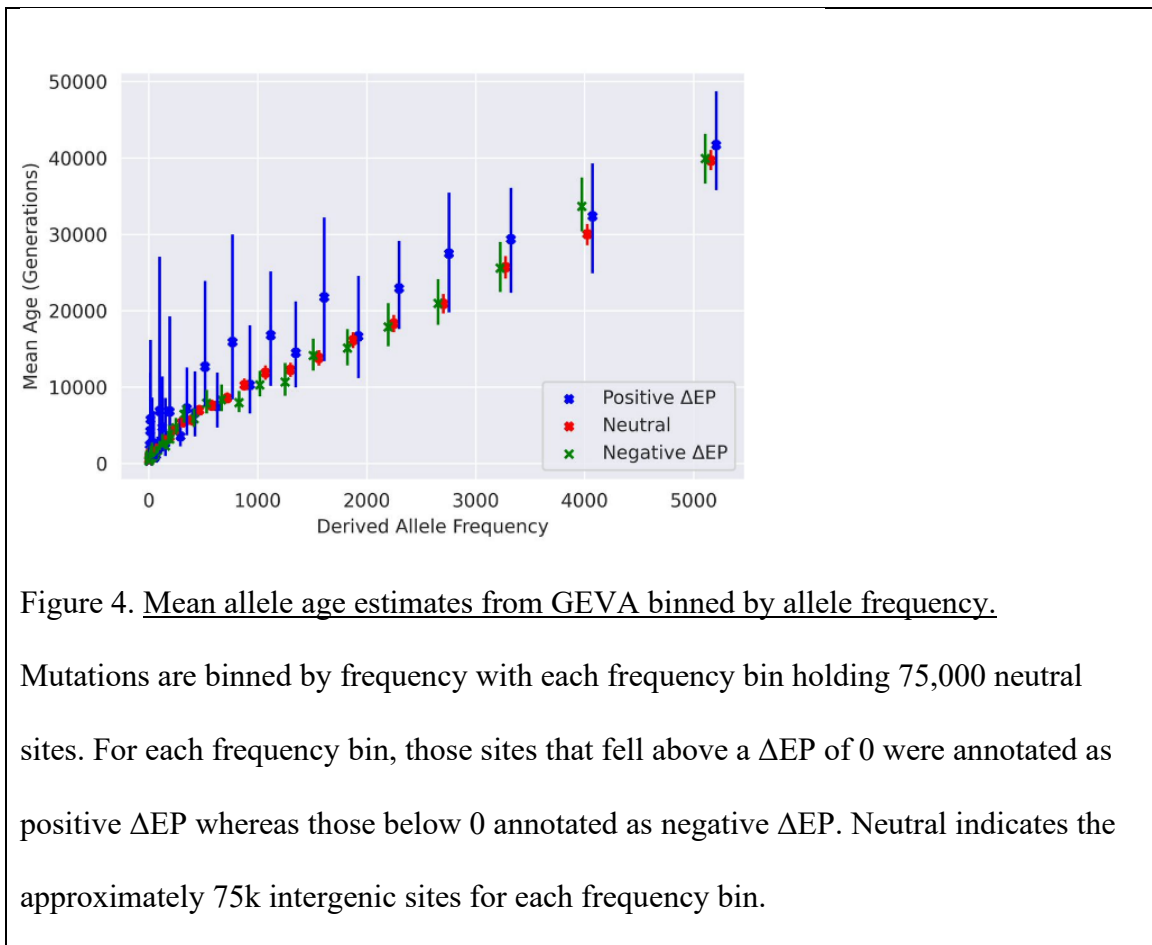


Figure 3. Illustrative example of mutational event.

EP calculation values and age estimation targets for GEVA and t_c for a hypothetical site with three copies of the derived allele (A) in a sample of 10 genomes with an ancestral allele of C.

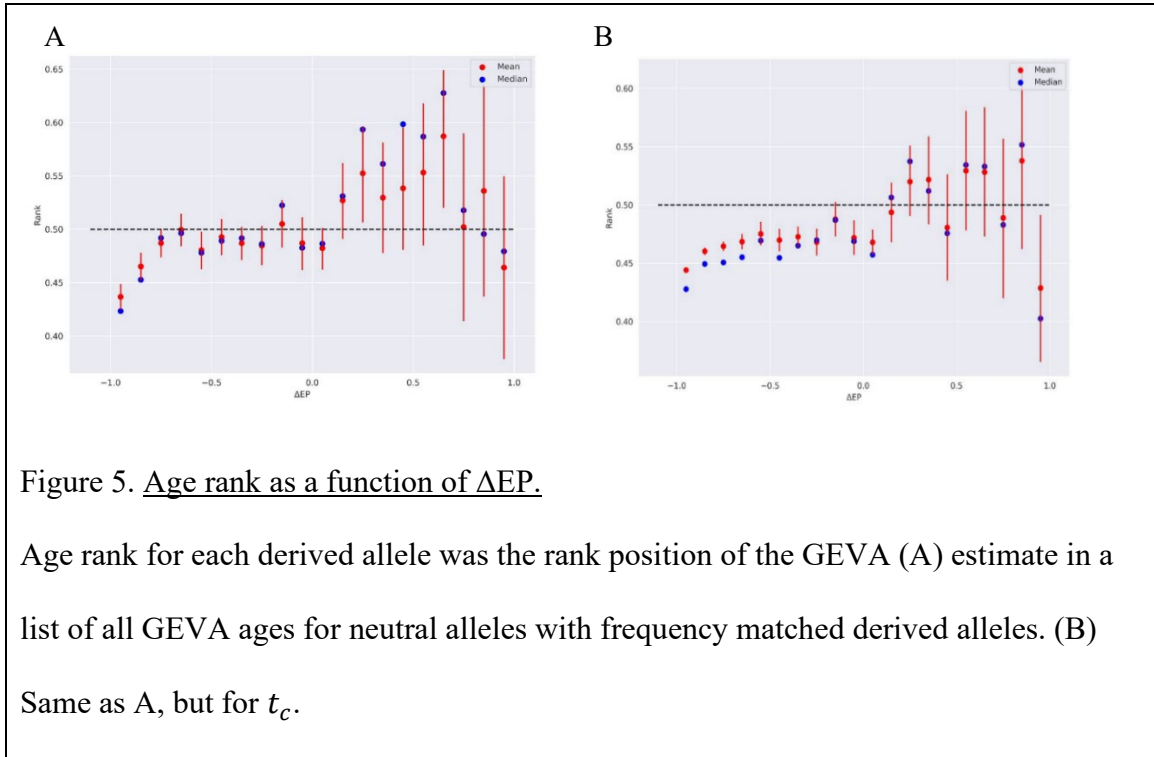
Allele frequency is a strong predictor of allele age, and as expected, the mean derived-allele age rises with frequency for all three classes of SNPs (Figure 4). For both positive and negative ΔEP SNPs, an analysis of variance (ANOVA) was conducted to test the hypothesis that selected derived alleles have the same mean age as control SNPs.

In both cases the null hypothesis was strongly rejected ($p = 4.17 \times 10^{-12}$ for negative ΔEP SNPs and 3.04×10^{-27} for positive ΔEP SNPs). However, unlike derived negative ΔEP alleles, which were younger on average than control alleles, as predicted, the positive ΔEP SNPs are older on average. Surprisingly, across most frequency intervals, derived positive ΔEP alleles exhibit mean ages thousands of generations older than the neutral control set.



To isolate the relationship between ΔEP and allele age independently of allele frequency, we placed each allele's age estimate into an ordered list of ages for neutral alleles of the same frequency. Non-synonymous alleles in the top half of the distribution (ranked higher than 0.5) are thus older than the median age of those neutral alleles. As

shown in Figures 5A and 5B, the ranked ΔEP values show a clear trend, with negative ΔEP values falling consistently below 0.5 (i.e., with ages less than neutral alleles of the same frequency) and positive ΔEP alleles have mean age ranks consistently above 0.5.



Because the set of non-coding, non-regulatory controls necessarily experienced the same demographic context as the selected alleles, explanations of older ages for candidate beneficial alleles that depend upon interactions of selection and demography are largely ruled out, at least for models in which the beneficial alleles are indeed under directional selection. Nor can models in which these alleles are sometimes neutral and sometimes favored help explain the observation, as such alleles would still be expected to be younger on average than our control set. This pattern, in which alleles are maintained longer than alleles that are not subject to selection, is simply not consistent with positive directional selection, but rather suggests some form of balancing selection (101).

Characterizing Old, Segregating, Positive ΔEP Alleles.

Overall, a large proportion of positive ΔEP alleles are older than neutral controls. For t_c there were 3511 positive ΔEP alleles, 1354 of which had age ranks greater than 0.5 (38.6%). For GEVA there were 1390 positive ΔEP alleles (fewer than for t_c as GEVA cannot be applied to alleles that occur only once), 741 of which had age ranks greater than 0.5 (53.3%). We considered the possibility that the elevated ages of segregating positive ΔEP alleles were a kind of sampling artifact, as would occur if they represented the tail of a distribution of ages for all favored alleles, including those that became fixed (which do not appear as SNPs and for which we do not have age estimates). This explanation does not apply to alleles under strong directional selection, for which the mean and variance in sojourn times are low. On the other hand, weakly selected favored alleles will have a large mean and variance in sojourn times (102), and a large sample of such alleles would have some that, by chance, had been segregating for a long time. However, if the old segregating positive ΔEP alleles were only very weakly favored, and if they constitute the minority of alleles that were held back by the chance effects of genetic drift, then they would make up only a small fraction of all positive ΔEP alleles, including both fixed and segregating. We do not observe this in the data, with segregating alleles constituting a large fraction (0.596, Table 1) of all positive ΔEP alleles.

Balancing selection can take many forms (103), but whatever the mode of selection for these alleles, it does not appear to be the kind of long-term balancing selection that causes trans-species polymorphisms like those found in immune-related genes (104,105). Of the positive ΔEP alleles, none of the GEVA values, and only 2.5% of

the t_c values, are over 200,000 generations, which would correspond approximately to the human chimpanzee divergence time, assuming a 29-year generation time (106). Most positive ΔEP sites, including those with age ranks greater than 0.5 (i.e., older than neutral alleles of the same frequency) also do not fit a conventional model of balancing selection in that the derived allele frequency is usually low (Figure 4, Appendix A Figure 18). For t_c the mean frequency of positive ΔEP sites with age ranks greater than 0.5 is 0.039, while for GEVA it is 0.091.

When we seek these alleles in archaic humans, we find that relatively few positive ΔEP alleles identified in the UK10K sample (241; 4.0%) occur in a sample of 4 archaic genomes. The same analysis for negative ΔEP alleles found a smaller proportion of shared alleles (2030; 1.4%), whereas an intermediate value of noncoding sites (401741; 3.0%) was observed among the sample of archaic genomes. For genomic regions identified as introgression from archaic humans, only 13 positive ΔEP alleles (0.2% of all positive ΔEP sites) and 180 negative ΔEP alleles (0.1% of all negative ΔEP sites) were found.

We applied an alternative method for identifying balancing selection to positive ΔEP alleles that is based on the number of nearby polymorphisms that have risen to a similar frequency as the candidate allele (99). We find that the test statistic, β , is significantly higher for positive ΔEP sites compared to negative ΔEP sites (p -value = $1.588e-6$), however the magnitude of these differences is small at just an average β value of 1.09 for positive ΔEP sites and 0.55 for negative ΔEP sites. Because most of the positive ΔEP sites in our study are found at low to moderate frequencies, and because the elevated ages, relative to neutral sites, are on the order of hundreds or thousands of

generations, it is likely that there has not been sufficient time for genetic drift to bring flanking sites into the configuration that the β statistic is designed to be sensitive to.

Examination of Modes of Balancing Selection: Population Structure and Overdominance.

We observed significant clumping of positive ΔEP SNPs among the genes included in the study. For every autosome, the observed variance in SNP density was significantly greater than that generated by population genetic simulation (Appendix A Table 7). Gene ontology analyses for genes rich in positive ΔEP SNPs revealed enrichment in several categories (Appendix A Table 8), most notably blood coagulation and several disease pathways.

One mechanism that could give rise to new balanced polymorphisms is if the selection regime arose because of the human population structure that favored ancestral alleles in some populations and derived alleles in other populations (as suggested in a recent analysis (107)). To examine the possibility that population structure is facilitating a large amount of balancing selection, we examined F_{ST} in the 1000 genomes data (12). Analysis of F_{ST} values in 1000 Genomes data for alleles from the UK10K samples with positive ΔEP and age ranks greater than 0.5 found no sign that these alleles show greater population structure than control alleles (Appendix A Table 9). In three comparisons, the hypothesis that F_{ST} was higher for positive ΔEP alleles that are older than expected could not be rejected by single classification Wilcoxon test in pooled African samples versus pooled European and Asian samples ($p = 0.1804$), pooled European versus pooled Asian samples ($p = 0.5298$), and Great Britain sample versus Italian sample ($p = 0.7854$).

Another possibility is if heterozygous positive ΔEP sites have higher fitness than homozygotes for both the ancestral and the derived alleles. To test this in a way that combined the signal from all positive ΔEP alleles, we asked whether positive ΔEP alleles had higher heterozygote counts than neutral alleles of the same allele frequencies. Analyzing SNPs with at least 100 derived allele copies, we observed equal proportions of positive ΔEP sites with more heterozygotes than the neutral class, compared to fewer; and we found a mean rank for heterozygote count for positive ΔEP sites of 0.501. A one-sided z -test of the null hypothesis that the mean rank was equal to or less than 0.5 did not approach statistical significance ($p = 0.48$). This is consistent with previously published results which failed to find evidence of overdominance at deletion sites thought to be under balancing selection (108). To assess our ability to detect heterozygote advantage using counts of heterozygotes, a power analysis was conducted using simulations that mirrored the actual data set, assuming genotypes are sampled under heterozygote advantage after selection has been acted. The analyses revealed that over a wide range of weak to moderate selection coefficients where the selective advantage is less than 1% (i.e., $s < 0.01$), that an excess of heterozygotes is unlikely to be detected given the UK10K sample size (Appendix A Table 10).

Models That can Account for a Period of Balancing Selection.

The absence of very old, derived alleles among positive ΔEP sites suggests that the balancing selection that occurs undergoes a change of character, such that balancing selection occurs for a period of time and is then followed by directional selection or no selection (i.e. genetic drift alone) leading to a loss of one or other of the alleles. If that

were not the case, then we would not expect the absence of very old alleles in this data set. To address this, we consider two models that both provide mechanisms for balancing selection and that both predict that balancing selection will be a temporary phase in the process of the fixation of beneficial alleles.

One theory to explain many positive ΔEP alleles with elevated ages includes two selection stages, including first a period of balancing selection under heterozygote advantage, after which positive directional selection carries the allele to fixation. Under this “staggered sweep” model, balancing selection occurs when a favorable allele arises on a chromosome that carries one or more recessive deleterious alleles at nearby locations, and it lasts until recombination moves the allele onto other haplotypes not having linked deleterious alleles (109). A heterozygote for this chromosomal region is initially favored because of the new allele’s dominance and the harmful allele’s recessivity, such that the net positive selection coefficient on heterozygotes is strong enough to counter the effects of genetic drift. The model is supported by the fact that individual humans, and human populations, carry very large numbers of deleterious alleles, the large majority of which are expected to be mostly recessive in their effects. Considering, for example, just loss-of-function alleles for which diploid European genomes are estimated to carry about 100 (mostly in the heterozygous state), then the odds that a new beneficial mutation arises near to, and in-phase, with a deleterious allele, may be quite high (110).

Testing the staggered sweep model is difficult because local linkage estimates, as well as t_c and GEVA estimates, all depend on a common estimate of the genetic map. However, we can avoid this complication, and partially test the staggered sweep model,

by comparing local recombination rates near positive ΔEP alleles that are fixed to those that are segregating. If segregating alleles are under balancing selection because of linkage to deleterious alleles, and the fixed alleles include those that had escaped by recombination, we expect segregating alleles to show lower local recombination rates than fixed positive ΔEP alleles. As predicted, the recombination rates of genomic regions near fixed positive ΔEP alleles were significantly higher than for segregating alleles (Mann Whitney U test $p=6.0 \times 10^{-19}$, Figure 6).

Another explanation that also invokes heterozygote advantage is a diploid version of Fisher's geometric model (denoted hereafter as DFG) in which mutations that carry the phenotype in the direction of the optimum may be favored when heterozygous under codominance and yet disfavored in homozygotes if that phenotype is more extreme and further away from the optimum (111). Under this model, balancing selection may be a common phase during an adaptive walk toward increasing fitness, with balanced alleles ultimately being lost when new alleles under simple positive directional selection arise and become fixed. The staggered sweep model and the DFG model differ most clearly in

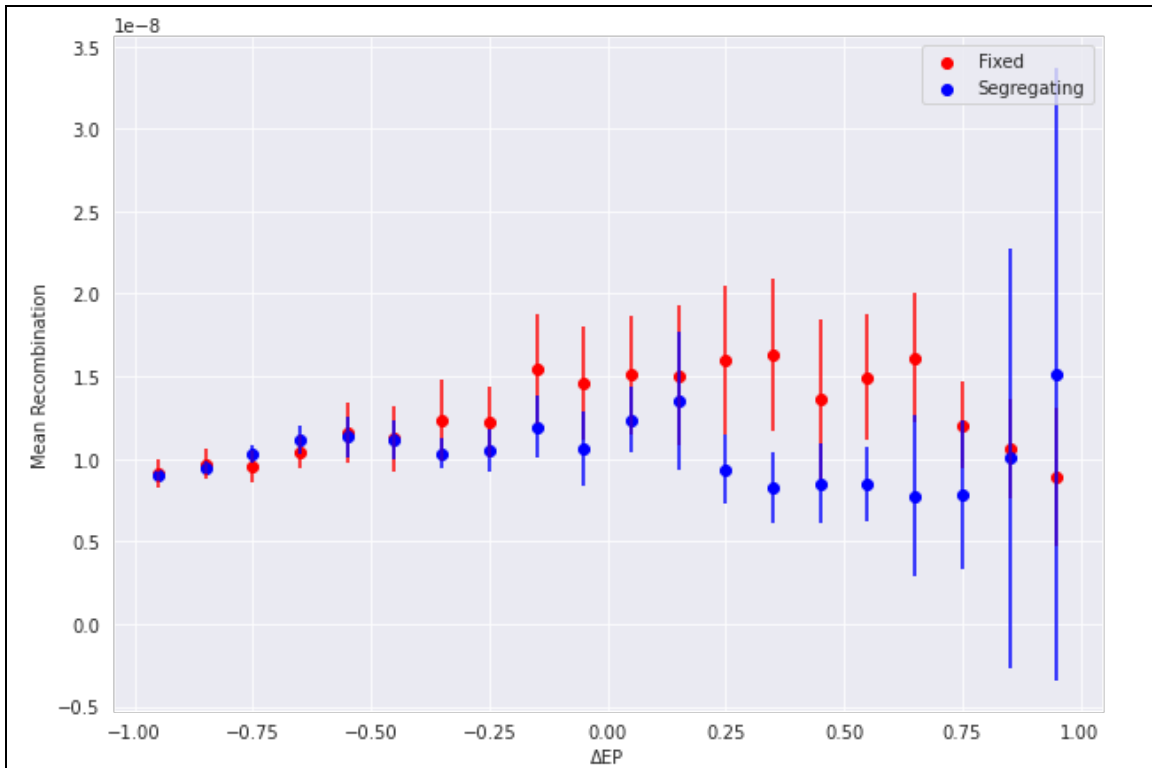


Figure 6. Mean recombination rate per base per generation as a function of ΔEP for fixed and segregating alleles.

Fixed derived alleles were identified based on the ancestral state (Ensembl and RAxML) and the hg19 reference base at a site. The average recombination rate for each ΔEP bin for both segregating and fixed derived sites was calculated using a population specific genetic map.

that the former has the period of balancing selection as a phase before the fixation of the allele, whereas the latter has the balanced allele being replaced by a new allele that is simply favored by directional selection. The former model predicts that some, perhaps many, selective sweeps are actually ‘soft’ sweeps caused by the fixation of a relatively old allele. In contrast, the DFG model predicts that when a selective sweep occurs, it is a conventional sweep by a new favored allele (i.e., a ‘hard’ sweep). Both models predict

partial sweeps around new alleles that arise in a balancing selection fitness scheme (Figure 7).

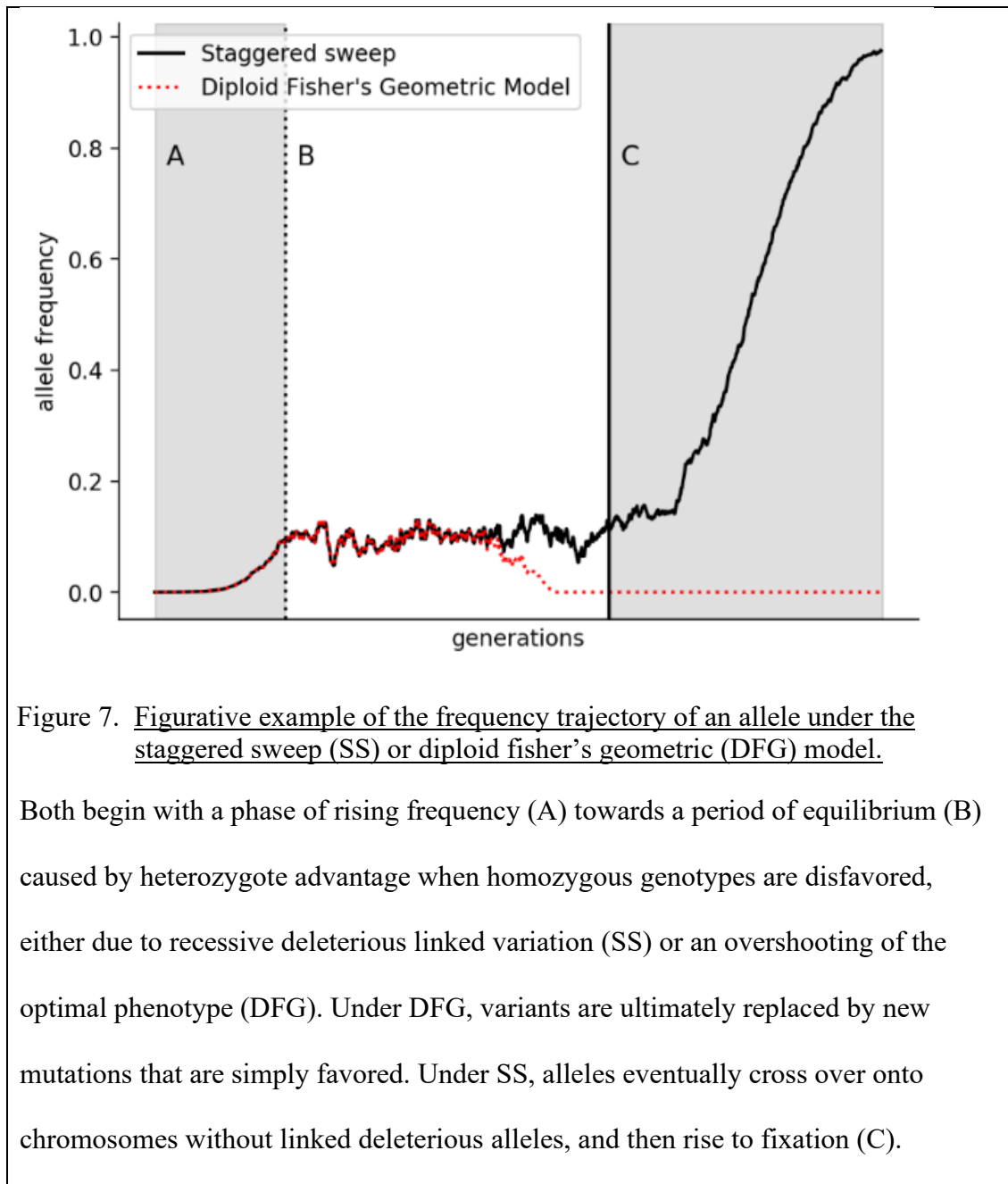
Implications for the Adaptation of Human Populations.

We find that the majority of candidate derived beneficial alleles in a human population are segregating, rather than fixed, and yet the mean ages of these alleles are older than those for derived control alleles. These relatively old SNPs do not appear to fit a classical balancing selection model in that most of them are at low frequency, and have age estimates almost always less than the age of the hominin branch. The overall pattern suggests that when fixation of beneficial alleles does occur, it often follows an initial period of balancing selection.

We did not find evidence that ΔEP alleles are maintained due to commonly considered mechanisms of balancing selection such as population structure or heterozygote advantage, although the power to detect these factors was low, unless selection has been quite strong. Instead, we found support for the staggered sweep model in which beneficial alleles arise on the same haplotype as a deleterious mutation which delays them from fixing. Under a staggered sweep model, we predict that there should be differences in recombination rates between segregating and fixed alleles allowing for some alleles to escape selection from nearby deleterious which we find to be true for moderate positive ΔEP sites.

If many beneficial alleles have a lengthy period of balancing selection, before proceeding to fixation, then a significant fraction of adaptive fixations experienced by the human species (not just individual populations) will have occurred as a ‘soft’ sweep

rather than a ‘hard’ sweep. This would help explain why there are few unambiguous cases of complete hard sweeps in large population genomic data sets (5,6).



An additional implication of these findings is that the process of adaptation by human populations may be slower than basic population genetic models predict. If a

significant fraction of ultimately beneficial fixed alleles undergoes a period of balancing selection, then at least at these sites, the process of adaptation is slowed and limited, not for lack of mutation, but rather by the process causing the period of balancing selection.

CHAPTER 3

ALLELE AGE ESTIMATORS DESIGNED FOR WHOLE GENOME DATASETS SHOW ONLY A MODEST DECREASE IN ACCURACY WHEN APPLIED TO WHOLE EXOME DATASETS.

Abstract

Personalized genomics in the healthcare system is becoming increasingly accessible as the costs of sequencing decreases. With the increase in number of genomes, larger numbers of rare variants are being discovered allowing for significant work to be done to identify their functional impacts in relation to disease phenotypes. One potential way to characterize these variants is to estimate the time the mutation entered the population. However, allele age estimators such as Relate, Genealogical Estimator of Variant Age, and time of coalescence, were developed based on the assumption that datasets include the entire genome with relatively low amounts of missing data. We examined the performance of each of these estimators on simulated exome data under a neutral constant population size model and found that each provides usable estimates of allele age from whole-exome datasets. To test the robustness of these methods, analyses were undertaken to simulate data under a population expansion model and background selection. Relate performs the best amongst all three estimators with Pearson coefficients of 0.64 and 0.68 (neutral constant and expansion population model) with a 17 percent and 15 percent drop in accuracy between whole genome and whole exome estimations. Of the three estimators, Relate is best able to parallelize to yield quick results with little resources, however, even Relate is only able to scale to thousands of samples making it unable to match the hundreds of thousands of samples being currently released. While more work is needed to expand the capabilities of current methods of estimating allele age, these methods estimate the age of mutations with a modest decrease in performance.

Introduction

With the rapid advancement in sequencing technology and availability, genomic data has become integral to investigating the genetic cause of many major diseases. A major benefit of whole genome sequencing is the identification of mutations across the entire genome including non-coding regions, however it remains the costlier option in comparison with whole-exome sequencing (112–114). Whole-exome sequencing (WES) targets the coding regions of the genome allowing for investigation into functional genomic mutations, which can more easily be associated with phenotypic changes and found more commonly in databases of disease-causing variants (115–118). This is more easily accomplished for monogenic diseases where just one gene contributes to a disease phenotype, however, improved methods have made understanding and identifying potential underlying mutations to polygenic diseases possible (119). In particular, rare variants have been found to be associated with complex human diseases and phenotypes (34) making variants at low frequency of particular interest to researchers.

As increasing numbers of genomes are sequenced, researchers are finding more rare variants, some of which have been shown to be causative of human disease (120). It is becoming more common for healthcare providers to perform exome sequencing to better understand the underlying cause of a patient's condition (121–123). However, while these studies can identify potentially causative mutations through association studies, the history of these mutations are not as easily able to be examined except through large population studies (124). Currently, there are initiatives to expand current genomic resources with corresponding medical information (24,125,126). In the UK, researchers have released nearly 500,000 exome datasets corresponding to other

participant information collected over the last fifteen years (24). Other similar biobanks operated primarily by healthcare systems are also primarily utilizing genotype and exome data in their electronic health record collection (125,126) with the United States launching the “All of Us” initiative which just recently was expanded to include 250,000 whole genome sequences (127).

One way to better understand the history of a potential causative variant is to estimate when a mutation was introduced to a population. Combining the approximate age of an allele along with other information about allele frequency globally allows for a better understanding of the selective pressure the mutation has undergone (128). Studies that attempt to understand the selective pressure on rare causative variants of disease rely on small whole-genome studies which focus on just a handful of individuals who carry a particular mutation of interest (129). Many large population studies have worked to expand the database of WES sequences combating the problem of sample size and yielding very large datasets. This leads to the obvious avenue of measuring the accuracy of previously used methods on the growing amount of data.

Previous research has attempted to estimate the age of alleles in a dataset containing thousands of exomes using a coalescent method where researchers found that most deleterious variants arose recently in human history (130). In more recent years, improved methods to estimate allele age have been published that yield more accurate estimates utilizing haplotype-based estimates of genealogies (49,50) and ancestral recombination graphs (48). These more recent methods of allele age estimation have been shown to estimate the age of a mutation from WGS with relative accuracy, however little to no work has been done to establish accuracy of these methods when utilizing the

methods on WES data. Each of the three estimators of allele age (Genealogical Estimator of Variant Age, Relate, time of coalescence), rely on mutation accumulation on shared haplotypes to accurately estimate the branch length of the branch containing the focal mutation. With WES data there will be numerous missing mutations in the non-coding regions likely causing the estimates to skew younger than the true age due to the fewer mutations appearing to accumulate. However, if there is a sufficient number of mutations in the coding regions, the error in estimates may be largely mitigated. Additionally, comparison between the ages for different mutations will retain the same relative difference in ages allowing for comparison type analyses to be performed.

In the methodology for Relate (48), the local genealogy underlying a focal site is reconstructed based on clustering from a distance matrix of mutation variation between haplotype pairs computed from a hidden Markov model (HMM). Then using a coalescent prior, Relate employs a Monte Carlo Markov Chain (MCMC) algorithm to estimate the time of the branch containing the focal mutation. Similarly, Genealogical Estimator of Variant Age (GEVA) (49) also employs an HMM to identify shared and non-shared haplotype pairs, but then subsequently samples from those pairs before calculating the posterior estimate of the age of the mutation from a joint clock model leveraging recombination and mutation. Time of coalescence (t_c) (50) however calculates the maximum shared haplotype (*msh*) which can be bounded by either a recombination or mutation event. Allele age can then be estimated based on a likelihood function leveraging the length of the shared haplotype.

In this study, three estimators of allele age (GEVA, relate, t_c) are assessed to test for accuracy in estimates of mutation age from whole exome sequence data. Initial

simulations are undertaken with a constant population size of neutral mutations. To test the robustness of these methods, an African origin demography model with a population size expansion is simulated with neutral mutations. In addition to simulations of neutral mutations, both a constant size population and an expansion population model are simulated with background selection. For each simulation, the true age of each mutation is compared to the estimates of the age generated from each program. Because rare variants are so integral to understanding human disease (28), we will additionally assess whether these estimators of allele age can be used successfully on rare variants found in exome datasets.

Methods

Simulation Dataset

Using stdpopsim version 0.2.0 (89), mutations were generated under both a simple, constant population model (Model id: PiecewiseConstant) and a complex demography model (Model id: OutOfAfrica_3G09) (131) with no selection. Both simulations were performed using msprime version 1.2.0 as the simulator (132) generating neutral mutations. From the simple model, a population of 7242 genomes was sampled, and from the complex model, 7242 CEU genomes were sampled. Full parameters for each simulation model can be found in the supplement (Tables 11 and 12 in Appendix B). From each simulation, segregating sites were output in VCF format using tskit version 0.5.0 (88). The output from the simulation will serve as the whole genome dataset for chromosome 22 for downstream analyses. Entire simulation pipeline described in Figure 19 in Appendix B.

Convert to Exome Data

For each of the simulations (simple and complex), the simulated chromosome 22 files were filtered for the coding regions. A bed file containing the coding regions for chromosome 22 was generated from a bed file containing all autosome coding regions from the UK Biobank (available at: <https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=3803>) (24). Using the BCFtools version 1.10.2-27-g9d66868 (133) filter function, the VCF outputs from stdpopsim were filtered just for the segregating variants that fell within the regions in the supplied bed file. These filtered VCF files serve as the WES data in the analyses. Each of these original simulation runs are annotated as WGS or WES, and then either SIMPLE for constant population size or COMPLEX for the population expansion model hereafter.

True Values

From the tree output file from stdpopsim (89), corresponding Nodes, Sites, Edges, and Mutations files were extracted using tskit (88). Using the extracted tree and genomic information from these output files, for each site, the position, the time of the mutation, and the time corresponding to both the parent and child node of the branch containing the mutation was extracted.

Allele Age Estimates

Genealogical Estimate of Variant Age (GEVA) Estimates (49). The chromosome column of each of the four VCF files (WES_Simple, WES_Complex, WGS_Simple, WGS_Complex) first were converted to number format using BCFtools version 1.10.2-27-g9d66868 (133) as required by GEVA. Each VCF file was then

converted to a binary file using the GRCh38 HapMap genetic map (77) supplied from stdpopsim (89) annotated with the additional map position column in centimorgans (cM). To run the estimator program for all sites of a frequency of 2 copies or more, a list of all sites ($k \geq 2$) was extracted and split into batches of 300 variants due to memory constraints as recommended in the documentation. The estimator was run with default parameters of 10000 for the effective population size and a mutation rate of $1e-8$. Estimates from the Joint Clock model that passed the supplied heuristic filter were used. This includes the mode of the posterior distribution of the concordant and discordant sampled haplotype pairs. Conversion and estimator programs available at <https://github.com/pkalbers/geva>.

Time of Coalescence (t_c) Estimates (50). Singletons (variants with just one copy in a population) in each of the four VCF files (WES_Simple, WES_Complex, WGS_Simple, WGS_Complex) first were phased by placing the mutation on the longer of the two possible haplotypes for each singleton in the population. A modified version of the time of coalescence estimator program (modified version described in supplement). The estimator was run with groups based on their k (copy frequency number) value using the flag `--k-range` to speed up the estimation pipeline and was run with default values of $1e-8$ for mutation rate and 10000 for effective population size. The same genetic map as used for the simulations and other estimates for GRCh38 was used with the `--map` flag. This was done on all variants across the frequency spectrum. Phasing and estimator programs available at <https://github.com/jaredgk/runtc>.

Relate Estimates(48). Using Relate version 1.1.9, each of the four VCF files (WES_Simple, WES_Complex, WGS_Simple, WGS_Complex) were converted to

haps/sample format. Using Relate's supplied parallelization script, each simulation was run using 12 threads with a mutation rate of $1e-8$ and an effective population size of 10000. The same genetic map as used for the simulations and other estimates for GRCh38 was used with the --map flag. Relate is available at <https://myersgroup.github.io/relate/index.html>.

Statistics

For every combination of estimator and simulation, four statistics were calculated: bias, Pearson's R, Spearman's R, and Root Mean Square Log Error (RMSLE) for each set of mutational ages in comparison with the true age of the mutation as calculated from the simulation.

Comparisons Across Sample Size

For both the simple and complex models described above, a spectrum of sampled genomes was extracted to compare accuracy across sample sizes. Sample sizes of 100 genomes, 1000 genomes, and 10000 genomes were used in addition to the 7242 genomes sampled in the first set of simulations. The same parameters as described above for each simple and complex model were used. Following the same pipeline as above (Appendix B Figure 19), the Tree file for each simulation was extracted and converted to a VCF which was then filtered for exon sites. Using just Relate, allele ages for the WES versions of the simulation samples were estimated.

Simulating Sites with Background Selection

Following the pipeline described in Figure 19 in Appendix B, stdpopsim version 0.2.0 (89) was utilized to simulate mutations under background selection using SLiM version 4.1 (134) and the Gamma distribution of fitness effects based on Kim et al (135).

Relate was then used to estimate the ages of mutations from both the original entire chromosome output from the simulation (WGS) and the filtered exome version (WES) with the same parameters as described above.

Results

Relate Outperforms With a Simple, Constant Population Size Model of Neutral Variation.

To better understand the accuracy of current methods of estimating allele age on whole-exome data, mutations were simulated under both a simple, constant population size model and a more complex model based on Gutenkunst's out-of-Africa demography (131) with no selection. Simulations were generated using msprime (132) from stdpopsim (89) using HapMap GRCh38 genetic map (77). Complete simulation parameters can be found in the methods and Tables 11 and 12 in Appendix B. The SNP data directly generated from these simulations serve as the whole-genome dataset (WGS). For the simulated whole-exome data (WES), the generated simulated data was filtered based on the exome regions sequenced from the UK Biobank (24). For each of the simulations (simple and complex, WES and WGS), the ages of mutations were estimated with Relate (48), Genealogical Estimation of Variant Age (GEVA) (49), and time of coalescence (t_c) (50).

Each method provided estimates for a different subset of variants depending on the filtering mechanism implemented. Relate estimated the age of all mutations that passed the filtering threshold with all sites identified as non-mapping or flipped being excluded. t_c estimated the age of all mutations including both singletons and derived

mutations found in all but one genome. GEVA estimates the age of mutations of a frequency of 2 copies or more, and additionally several other sites were excluded due to proximity to the end of the chromosome.

Relate estimated allele age the quickest, completing the estimation of allele age for the WES dataset in less than one hour for both demography models, and for the WGS in less than 12 hours for the simple model and around 13.5 hours for the complex model (Appendix B Table 13). Both GEVA and t_c took much longer as neither can easily parallelize. In the case of both estimators, sites can be separated into batches to allow for each batch to be run separately, however if forced to run sequentially both estimators would take on the magnitude of days to complete even for the much smaller WES dataset. Both GEVA and Relate (when run in parallel with multiple threads) required large amounts of memory and required dedicated computational resources.

For each method, the estimated age of each variant was compared to the variants' corresponding true value for variants under both the simple and complex model. True mutation time was extracted from the tree files from the simulation output. For the simple model, comparing the true age of the mutation to the estimated age from WES data for each of the three methods, a Pearson's correlation coefficient of 0.64 was found for the Relate method with both GEVA and t_c following with correlations of 0.45 and 0.26 respectively (Figure 8). All three estimators underestimated mutations that were exceptionally old, with t_c having the largest discrepancy. Each estimator did perform better estimating the ages of the mutations from WGS dataset with correlations of 0.77 for Relate (WES: 0.64), 0.65 for GEVA (WES: 0.45), and 0.36 for t_c (WES: 0.26) (Appendix B Figure 20). GEVA saw the biggest drop in performance between WGS

estimates and WES with 31% drop (Table 2). t_c had the next biggest drop in performance at a 29% drop followed by Relate with the lowest drop in performance at a 17% deficit.

All Three Estimators Show Relative Robustness to Demography and Selection.

To investigate whether estimators of mutational age are robust to more complicated demography models, a more complex out-of-Africa model was simulated (131). Again, ages of all polymorphic mutations were estimated using the three estimators and compared with true values. A similar pattern of estimates from the simple model emerged with all three methods underestimating old mutations. Relate remained the most accurate with a Spearman's correlation coefficient of 0.68 with GEVA and t_c with correlations of 0.45 and 0.37 respectively (Figure 9). Again, Relate and GEVA were better able to estimate the ages of the mutations from WGS data with correlations of 0.80 for Relate (WES: 0.68) and 0.63 for GEVA (WES: 0.45) (Appendix B Figure 21). However, t_c performed similarly with Pearson's correlation coefficients of 0.37 for both WES and WGS datasets with the smallest drop in performance at only 0.05% (Table 2). GEVA had the largest drop in performance again at 28% with Relate having a 15% drop in performance.

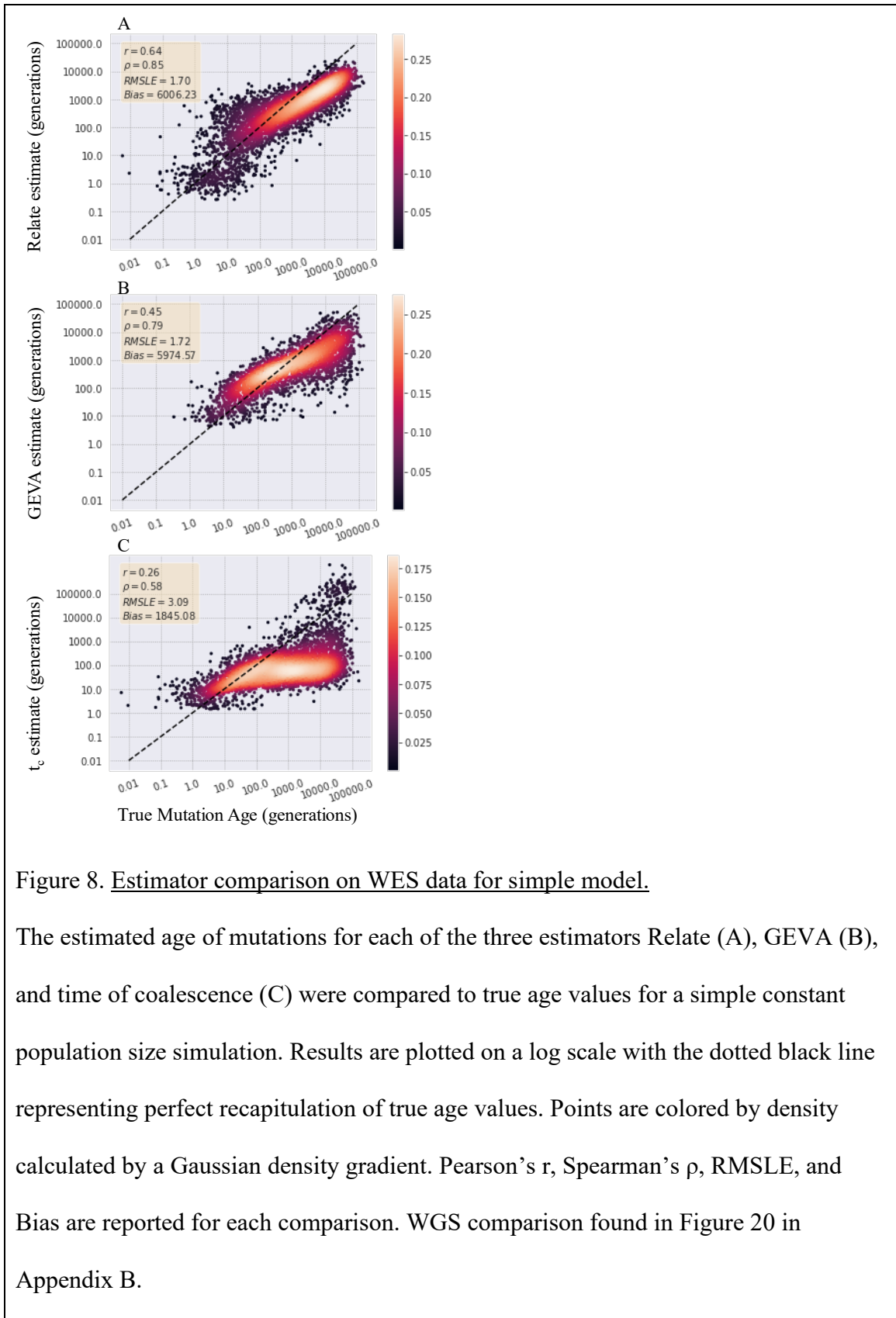


Table 2. Pearson’s correlation coefficients for simple and complex model for WES and WGS datasets.

	Simple Model		Complex Model	
	WES	WGS	WES	WGS
Relate (12 threads)	0.6394	0.7743	0.6799	0.7994
GEVA	0.4467	0.6476	0.4510	0.6290
t_c	0.2564	0.3623	0.3691	0.3693

Each of the three allele age estimators (Relate, GEVA, t_c) were run on neutral simulations of the simple constant population size model and the out-of-Africa expansion population size model for both the whole exome sequences (WES) and the whole genome sequences (WGS). For each comparison between estimator and true value, the Pearson’s correlation coefficient was calculated.

Further testing the robustness of allele age estimation on exome data, the estimator Relate was used to estimate the age of mutations in a sample size of 100 genomes, 1000 genomes, and 10000 genomes. Relate was used as it was the most tractable for the larger sample size of 10000 genomes, however estimation of mutations in a sample of 100000 genomes was also attempted but required over 15 TB of storage and high memory requirements making it not easily tractable even with dedicated computing resources. With increasing sample size, the Pearson coefficient between estimated age of mutations and true age increased (Figure 10, Appendix B Figure 24). This is expected as Relate depends on the surrounding mutations around a focal site to

estimate the age of the mutation of interest, and as the sample size increases as does the number of mutations present.

In nature, mutations are not just experiencing genetic drift, but often also natural selection, particularly negative selection because of either itself is deleterious or a nearby, linked mutation is deleterious. To simulate this, keeping with the simple and complex model, each model is simulated using SLiM with background selection with a distribution of selection coefficients pulled from a gamma distribution. Relate was then used to estimate the ages of the mutations both using filtered exon dataset (WES) and the entire simulated chromosome (WGS). Relate performed slightly worse on data generated from a simulation with background selection, relative to its performance with no selection, for both the simple and complex models (Table 2, Figure 11). Relate had an 18.6% decrease in accuracy for the simple model with the WES dataset in comparison with the WGS and a 26.7% decrease in accuracy with the WES dataset for the complex model. This is on average a 22% decrease in accuracy, higher than the 16% average decrease in accuracy on mutations under a simulation with no selection.

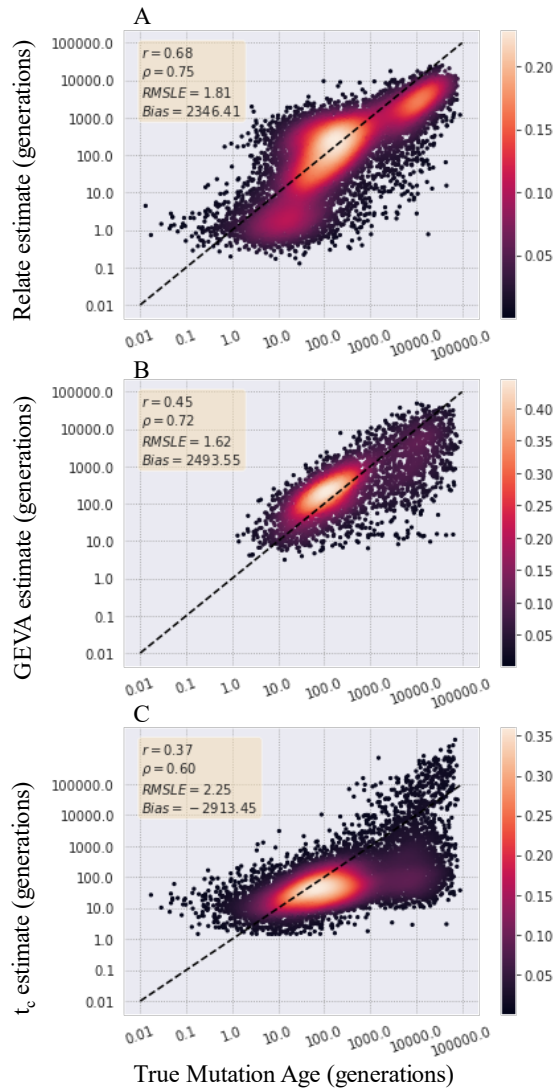


Figure 9. Estimator comparison on WES data for a complex model.

Estimated age of mutations for each of the three estimators Relate (A), GEVA (B), and time of coalescence (C) were compared to true age values for a complex demography simulation. Results are plotted on a log scale with the dotted black line representing perfect recapitulation of true age values. Points are colored by density calculated by a Gaussian density gradient. Pearson's r , Spearman's ρ , RMSLE, and Bias are reported for each comparison. WGS comparison found in Figure 21 in Appendix B.

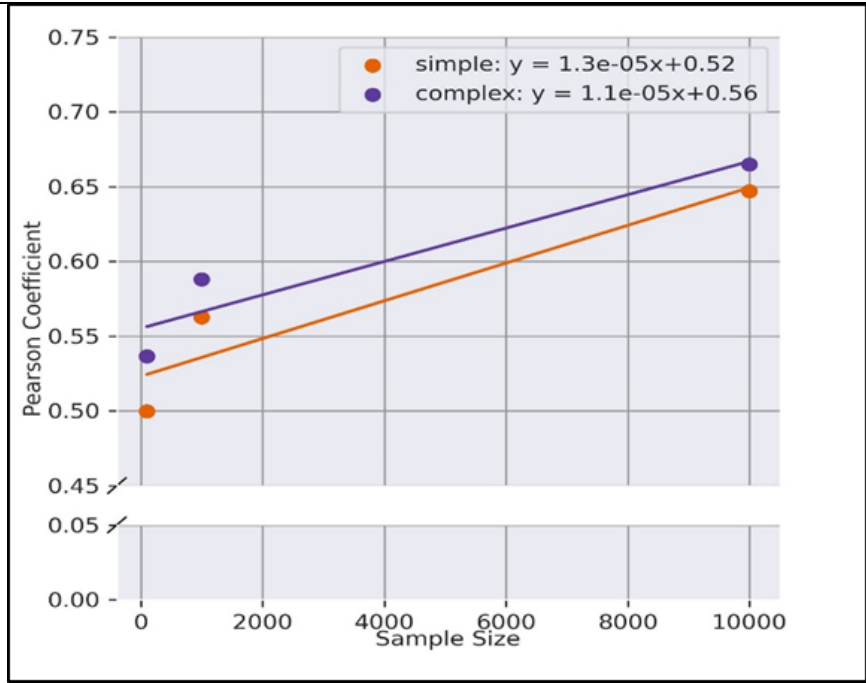


Figure 10. Correlation between true and estimated allele age increases with sample size. Samples of 100, 1000, and 10000 genomes are sampled from simulations. The true age of mutations is compared with the estimated age of mutations estimated with Relate for each of the three sampled set of mutations. A regression line is fit for both the simple (orange) and complex (purple) simulation estimates.

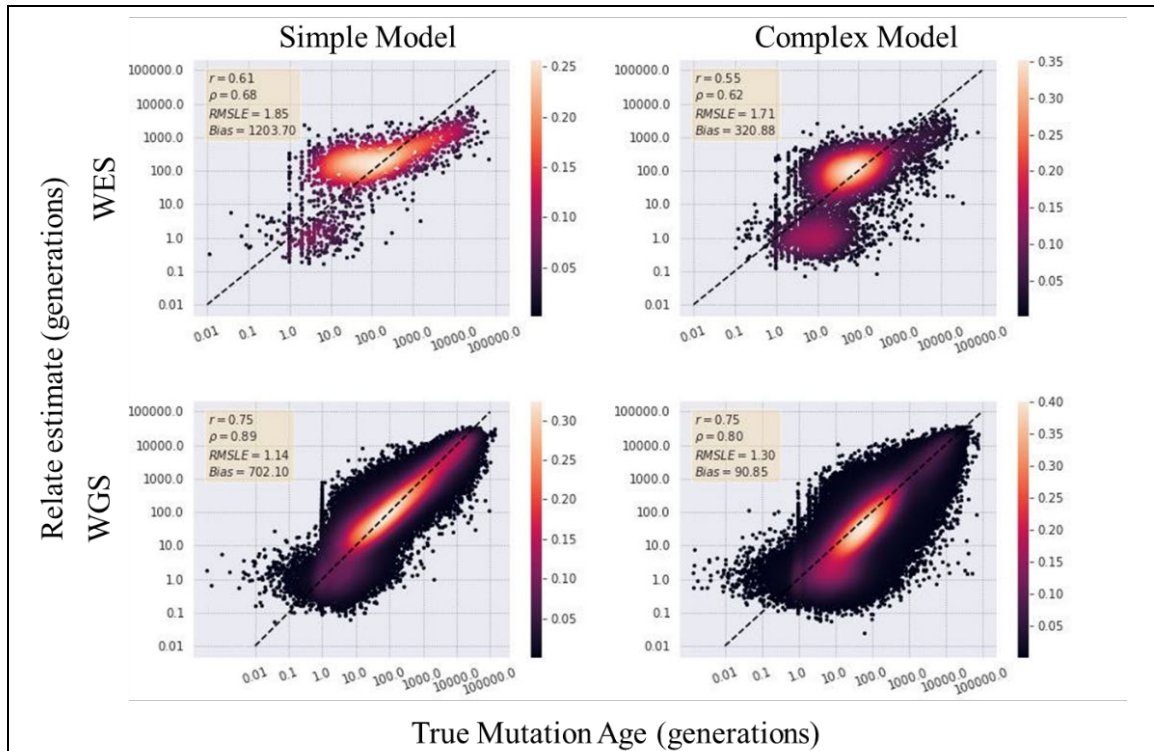


Figure 11. Relate allele age estimates compared to true values under background selection.

The allele age of mutations from simulations of sampled genomes under a model which incorporates background selection are estimated using Relate. Upper panels have estimates from whole-exome datasets while lower panels contain estimates from the entire chromosome of mutations.

Rare Variants Show No Decrease in Accuracy of Allele Age Estimates in Comparison with Common Variation.

To address the accuracy of allele age estimators for rare alleles in WES datasets in comparison to more common variants, an analysis of the root mean square log error (RMSLE) averaged across frequency bins was undertaken. Variants were binned into 1% frequency bins and for each estimator method the RMSLE was calculated for that bin

normalized by the average true age of the bin (Figure 12). While it was found that the RMSLE was much lower for rare variants without the normalization for true age (Appendix B Figure 22), once normalized it was found that all methods did better at estimating the age of more common variants in comparison with rare variants (less than a 1% frequency) (Figure 12A). Similar trends in error rates were found in the complex model (Appendix B Figure 23). While GEVA is unable to estimate the age of singletons, both Relate and t_c were able to estimate the age of mutations appearing once in the population as well as other low frequency mutations (Figure 12B).

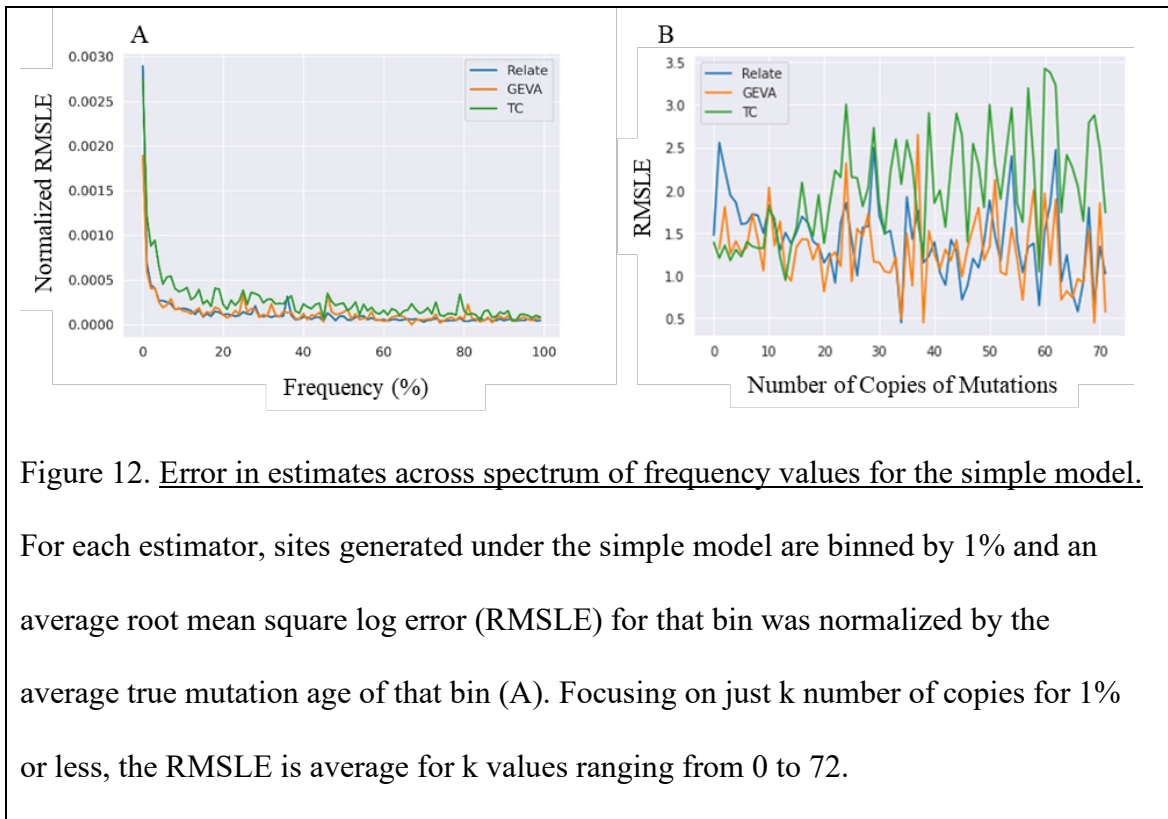


Figure 12. Error in estimates across spectrum of frequency values for the simple model.

For each estimator, sites generated under the simple model are binned by 1% and an average root mean square log error (RMSLE) for that bin was normalized by the average true mutation age of that bin (A). Focusing on just k number of copies for 1% or less, the RMSLE is average for k values ranging from 0 to 72.

Discussion

As sequencing becomes more cost efficient and thus more accessible (113,114), personalized genomics will also become increasingly accessible. In personalized

genomics, a large portion of resources is spent identifying causative mutations for disease (116,117), and in doing so, it becomes important to not only understand the functional impacts of the mutation but also the demographic history. Currently, there are many methods to characterize a mutation (115,116,136,137), with one subset being the estimation of the time of the mutation entered the population (43,48–50). These methods are modeled and tested with whole genome sequence data. However, since WES data remains the more cost-efficient option making it more likely to be used in routine clinical diagnostics (121,122), a comparison of several methods of allele age estimation to ascertain whether the methods can accurately be applied to whole exome sequence data is required.

Of the three estimators, all had the highest accuracy in identifying the ages of mutations from WGS data for both the simple and complex demography models with Relate performing the best across all estimators. For WES data, we find that of all three methods Relate again performs best with the highest correlation and lowest bias with a loss of between 15 to 17 percent in comparison of the estimates on the WGS data. However, all three methods slightly underestimate the oldest of alleles with t_c underestimating them to the highest degree with the RMSLE tripling between 100 generations to 10000 generations. T_c greatly underestimates the ages of very old variants revealing that even in the modified version that considers the circumstances where a haplotype ends in non-coding regions. This modification does not totally mitigate the issue of haplotype tracks ending in intergenic or intronic regions with instead t_c estimating much longer shared haplotype tracts than the true shared haplotypes. As a trend, each method slightly overestimated the youngest of alleles as the bias values flip in

signs from the youngest (less than 100 generations) to the oldest (greater than 100 generations). This is expected, as coalescent events that happen further back in time allow for more mutations to accumulate along that branch. Thus, in estimating allele age from a dataset missing some substantial proportion of the mutations means that the estimators are observing fewer mutations on a branch and estimating it shorter than its true length. The overestimation of the youngest variants and underestimation of the oldest variants has been previously identified in comparison of whole genome sequence data estimates from both Relate and GEVA (138,139). In the comparison examining the robustness of the methods on simulations expanding beyond just a neutral model to also include deleterious mutations, Relate had a larger decrease in accuracy in the estimates between exome and genome data in comparison with the neutral mutation model (simple: 19%, complex: 27%) (Figure 11).

All three estimators assume an infinite-site model where only a single mutation is able to arise at each site, however in nature this is not always the case (140,141). For simulation data, this is not a concern as sites where a site had more than one mutation occur can be excluded or not allowed to occur. However, for empirical data, repeated mutations may occur especially in regions of the genome where the mutation rate is particularly high such as in CpG sites (142). If the infinite-site model is violated with repeated mutation at a given loci with multiple common ancestral haplotypes, then haplotypes with differing genealogical histories would then be pooled together artificially inflating the estimated ages. Relate deals with this issue by excluding sites where mutations cannot be mapped onto the local genealogy (48) while t_c takes a composite of the estimated age of each copy of allele allowing for repeated mutations at a single locus

not to skew the estimated ages too high (50). However, it is likely that the majority of human variants have arisen by way of single mutations, with violations of the infinite-site model by multiple mutations constituting a very small proportion of sites, given the few known examples (143–145).

While common variants have long been identified as potential pools of variants affecting human phenotypes (146,147), rare variants are often less well understood and harder to be studied and characterized. Investigation of rare alleles using methods such as mutation age estimators allows for these mutations to be better characterized in a population. This is particularly important because these mutations are much less likely to be found in published large population data such as 1KGP (12) leading to very few opportunities to learn more about their history except through individualized studies. When normalized for the true age, rare variants are less able to be accurately aged, however since many rare variants are extremely young the overall error in estimates is small for these sites especially in comparison to common variants which may be extremely old and have a wide range of variation in estimated values for alleles of the same frequency.

While this analysis identifies some limitations of utilizing allele age estimators on whole exome data, it is clear that these methods can also be leveraged to better understand the history of focal mutations. Expanding these methods of characterization for mutations remains an important question as clinicians and researchers are still attempting to understand mechanisms of many disease phenotypes (147). While these estimators, in particular Relate, demonstrate reasonable accuracy in estimates based on whole exome data, users of these methods must also acknowledge the biases that appear in estimating the most extreme frequency variation. Utilization of these tools are still able

to offer some valuable insights into the history of potentially causative alleles to human disease especially if comparison type analysis utilizing the rank of allele age is considered. As the number of sequenced exomes continue to increase as the field of personalized genomics becomes more widespread, these methods will likely play an important role in advancing our understanding of the genetic basis of disease.

CHAPTER 4

SUBSTITUTIONS ON HOMININ BRANCH AROSE VIA NOVEL AMINO ACIDS BUT SHOW NO EVIDENCE OF ADAPTATION

Abstract

Since chimpanzees and modern humans diverged around 7 million years ago, thousands of substitutions have become fixed across both lineages. While several genes and functional processes have been identified to be positively selected, i.e., adaptive, it has been difficult to identify specific sites that contribute to adaptations in human phenotypic traits. Additionally, there is a lack of evidence for new mutations sweeping to fixation via classic hard sweeps. Using sites where humans have a fixed allele different from non-human primates, we leverage evolutionary probability to identify the mechanism by which adaptation occurs. Evolutionary probability varies with the evolutionary history of amino acids at a given locus with a low probability for novel amino acids and a high probability for more commonly found amino acids for that site. In one scenario, evolutionary adaptation is occurring through positive selection on uncommon amino acids while in an alternative scenario, adaptation creates a more evolutionarily stable protein due to a mutation back to a more commonly found amino acid in the vertebrate tree for that site. We found that most substituted sites in the human lineage have a derived, fixed allele forming a low probability amino acid whereas all other non-human primate lineages tend to possess ancestral, high probability amino acids. While we show that the majority of substitutions in modern humans occur by way of novel amino acids, there was no evidence that these sites are driving the adaptation for phenotypic changes found on the hominin branch.

Introduction

Previous work has estimated two substitutions per gene in modern humans with 10 to 20 percent of these substitutions found to be adaptive (3,8,57,58), however, most studies have identified regions under adaptation rather than specific sites driving those adaptations. There is evidence for more recent selection on segregating traits in modern humans (148,149), and some evidence suggests regulatory regions are involved in human evolution (150,151). This leaves an obvious question about whether specific sites in protein coding regions can be associated with adaptation of specific traits.

Chimpanzees have been found to have a larger number of genes to be under selection in comparison with modern humans due to natural selection being able to work more efficiently due to differences in historical effective population sizes (152). This contrasts with the higher substitution rate observed on the hominin branch (152). Between chimpanzees and modern humans there are millions of fixations between the two species that have occurred since speciation around 7 million years ago (57). This corresponds to thousands of protein coding sites where a derived allele has become fixed in either chimpanzees or modern humans (153). While there is evidence that both regulatory and protein coding substitutions have an effect on phenotypic differences between humans and non-human primates (154), there is a method to gain insight into the evolutionary history of protein coding changes.

An evolutionary probability value indicates whether an amino acid at a given site is likely or unlikely based on the occurrence of that amino acid across the vertebrate tree at that site. Thousands of sites where a evolutionarily unlikely site had become fixed in modern humans (14) indicates that adaptation may work best in environments where a

novel amino acid is preferred by selection. However, the opposite could be true if there is less evolutionary constraint on a region allowing for several amino acids to be common on the vertebrate tree yielding multiple likely amino acids. Evolutionary probability (EP) is calculated by a posterior probability based on a vertebrate alignment with divergence times to assign probability values for every possible amino acid at a given locus (14). A high EP amino acid corresponds to a residue found often in the alignment at the position with a high probability of that residue being found in closely related species. Likewise, a low EP amino acid would be rare in the alignment at the position. Utilizing EP to identify sites where modern humans and non-primate humans display a transition from one extreme EP value to another allows the opportunity to investigate the mechanism by which adaptations arise.

With thousands of genetic protein coding substitutions identified between humans and non-human primates, specific impacts on phenotype can be investigated through functional and gene ontology analyses. There have been several genes associated with adaptation in modern humans since the divergence from other primates (154) especially genes associated with neural development (155,156). Other adaptations of interest that have been investigated include manipulation and bipedalism (157), cell immunity (57), and sexual reproduction (158). While studies have identified genes under selection for several key phenotype differences between humans and non-human primates, the specific mutations associated with the adaptation are less well understood. However, EP provides the unique opportunity to identify which sites in the gene are creating transitions in amino acids from low to high probability, or vice versa, allowing for researchers to investigate

not only the sites under selection but also the mechanism by which adaptation is occurring.

Two separate hypotheses for the mode of adaptation arising in modern humans can be tested. First, if adaptation arises by way of novel amino acids, then genes undergoing selection in modern humans will have a larger enrichment of negative delta EP sites consistent with the mutation from a more stable amino acid to a new less evolutionarily favored but newly adaptive amino acid. Second, if adaptation arises by way of returning to a more evolutionarily stable amino acid, then the genes undergoing selection will have a larger enrichment of positive delta EP sites consistent with the new derived mutation being evolutionarily favored over the ancestral site shared with other primates. Using primate phylogeny of modern humans, archaic humans, chimpanzees, gorillas and orangutans, sites where the modern human amino acid differs from the amino acid of non-human primates can be identified. For each of these sites the evolutionary probability is calculated and compared for genes undergoing selection and genes more likely to be conserved across the primate phylogeny.

Methods

Dataset Curation

Polymorphism data from across the primate tree was assembled from the 1000 Genome Project (12), the Great Ape Project (159), and four genomes from archaic human individuals (94–97). The polymorphism data includes genomic data from *Gorilla gorilla gorilla*, *Gorilla beringei graueri*, *Pan paniscus*, *Pan troglodytes*, *Pongo abelii*, *Pongo pygmaeus*, *Homo sapiens neanderthalensis*, *Homo sapiens ssp. Denisova*, and *Homo*

sapiens sapiens (1kGP). Sample sizes for each population can be found in Table 3. As the analysis was limited to sites for which there are available evolutionary probability (EP) values, first the VCF files for each species was filtered for those sites found in coding regions with EP values (list of REFseq ids used can be found in the supplemental file). Because the analysis requires identification of sites where each species is fixed for the reference, this required identification of sites where polymorphisms were not able to be called due to low coverage. Description of how these areas of the genome were dealt with is described in the next section. With a VCF files for each species both filtered for coding regions and with the inclusion of missing sites due to low coverage, for each chromosome the polymorphism data for each species was merged using BCFtools (133) with the flag that calls all sites not included in the VCF file as the reference. The frequency of each allele at a given site was extracted for each site without missing data for each species. In addition to frequency, tables of nonsynonymous and synonymous sites were assembled and EP values for the amino acids found at each nonsynonymous site were calculated.

Dealing With Low Coverage Regions

To account for sites in each primate population where sufficient coverage was not available to accurately call polymorphic sites, sites in these regions were annotated as missing so ensure these sites were not mischaracterized as locations where the population was fixed for the reference. For each primate population, the noncallable bed file for that population was downloaded. Those sites that fell within both the coding regions (from the REFseq BED file) and the noncallable BED file for the population, a line was inserted into the VCF file for the site with every individual in the population having a missing

genotype. This ensures when the primate VCF files are merged all sites not included in the VCF file can be called as the reference genotype.

Table 3. Statistics for sample populations.

Species	Total Sampled Individuals	Subpopulations
Modern Human	2504	Africa: 661, East Asia: 504, South Asia: 489, Europe: 503, Admixed America: 347
Archaic Human	4	<i>Neanderthal</i> : 3, <i>Denisovan</i> : 1
<i>Pan</i> (Chimpanzee)	70	<i>Pan paniscus</i> : 10, <i>Pan troglodytes</i> : 60
<i>Gorilla</i>	31	<i>Beringei graueri</i> : 3, <i>Gorilla gorilla</i> : 28
<i>Pongo</i> (Orangutan)	10	<i>Pongo pygmaeus</i> : 5, <i>Pongo abelii</i> : 5

Counts of each species used in the analysis are listed along with the breakdown of the number of individuals for each subpopulation included for a given species.

Assembling Gene List

Genes associated with three phenotypic adaptations in modern humans were considered: bipedalism, neurodevelopment, and integumentary system development. As a comparison set, a list of housekeeping genes, genes with a low Gini coefficient (160), are used as the control set. For the phenotype of bipedalism, genes that are associated with gene ontology (GO) (161,162) terms “skeletal system development” and “cartilage development” were used. For the phenotype of changes in neurodevelopment, any genes falling into the GO category for cognition were included. To test for adaptation in the integumentary system, genes that fall in the GO categories “skin development” and “epidermal development” along with genes in the epidermal differential complex (163).

Statistical Methods

A Chi-Square analysis between each gene set of interest with the housekeeping set of genes as a control is undertaken. Both the numbers of nonsynonymous and synonymous sites are counted in order to account for differences in the number of genes per set and the total length of a gene.

Potential Distribution of Evolutionary Probabilities

The set of sites where modern humans have a derived substituted amino acid, and all other non-human primates have the ancestral amino acid was used as the set of interest. As a comparison null distribution, a set of invariant sites where modern humans and non-human primates all share the same amino acid is also assembled. For each site in each set, all possible mutations for that site are simulated to calculate the evolutionary probability for each possible mutation and the average for all potential mutations is recorded. As an example, if in the ancestral state has a codon AAG coding for lysine, this a mutation in this codon in each of the three base pairs could mutate to seven possible amino acids (due to redundancy in the third position). At this position in the protein, the EP can be calculated for each of those possible amino acids that can potentially occur due to a single mutation in the ancestral codon. The EP values are then averaged for this amino acid position.

Identification of Genes Under Selection Using ABSREL

Multiple sequence alignments (MSAs) of primate species were assembled from Ensembl release 106 (71,164,165). Ensembl used the EPO pipeline to align primate genomes. For this analysis, genes were retained if orthologous genes could be identified in each of the primate genomes. Genes must also be able to be associated with a REFseq

ID for which there are EP values leaving 4438 genes. To identify genes under positive selection on the hominin branch, aBSREL (166) from the HYPHY suite of selection analyses was utilized (167). For each gene tested, the gene tree was estimated using RAxML (79) using a general time reversible model with a gamma distribution (GTR+ Γ). aBSREL was run with the default parameters using the hominin branch as the foreground branch to be tested. Using an error filter BUSTED-E (168), sites that had extreme omega values were removed from the estimation for genes under selection.

Gene Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) (169,170) was performed to identify overlaps between positively selected genes and the Hallmark Gene Set. The identified gene set overlaps were compared to those identified using Panther (92) Gene Ontology statistical overrepresentation test (93).

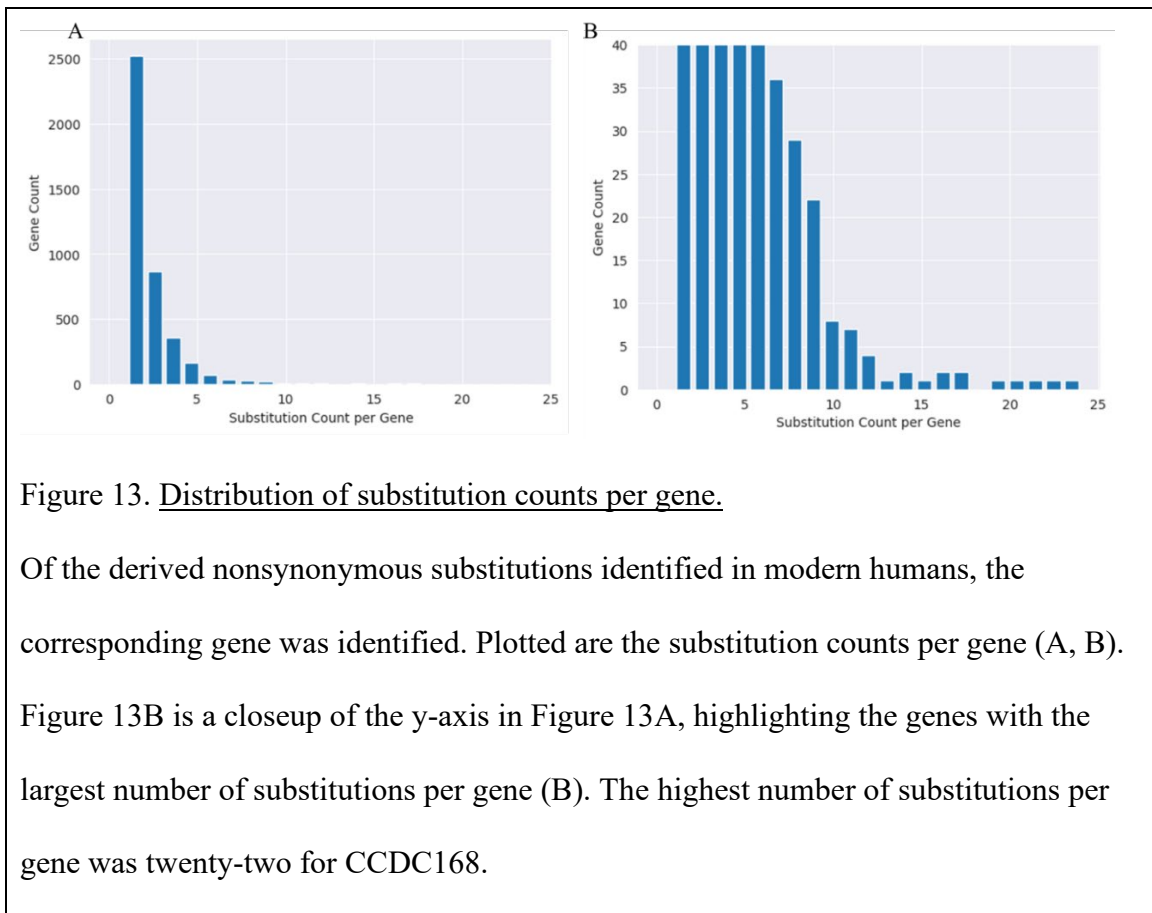
Results

Examination of Dataset Reveals That the Majority of Genes Only Have One Substitution.

Sites that differ between non-human primates and modern humans are identified across the genome using the 1k Genome Project dataset along with the Great Ape Project dataset. Since coverage varied across each of the sequencing projects and it is crucial that loci that are truly fixed for the reference are identified, sites that fall into low coverage and noncallable regions are excluded from the analyses. Criteria used for filtering of the sites exclude sites where modern humans are polymorphic above a threshold (MAF > 0.001) for a given allele, and chimpanzees, orangutans, and gorillas are not all fixed for the ancestral allele. Both chimpanzees and orangutans have two species included (*Pan paniscus* & *Pan troglodytes*, *Pongo abelii* & *Pongo*

pygmaeus), so the site was retained as long as one of the two populations was fixed for the ancestral allele within the species allowing for the other species to have low coverage at that site.

In modern humans, 17680 substituted sites were identified in the protein coding regions, with 7287 of these sites creating a nonsynonymous change and 10393 yielding a synonymous change. For the 18819 genes utilized in the analysis for which there are EP calculations, the majority of genes had just a single substitution for the entire gene (Figure 13A). However, the largest number of substitutions per gene was 22 substitutions for CCDC168, which is a protein coding gene for a transmembrane protein (Figure 13B).



The Majority of Substitutions Occur via Novel Amino Acids on the Hominin Branch.

For most nonsynonymous substitutions, humans have the low EP amino acid while all other non-human primates have the high EP amino acid (Figure 14). This is

indicative of the non-human primates having the ancestral allele at this site with humans having a derived substitution yielding a novel amino acid for the given site. A smaller percentage of substitutions occurred between two low EP amino acids yielding ΔEP values close to zero. One explanation for these sites is that these sites are in regions of the genome where mutations are more tolerable, and several different amino acids can exist at the focal site. The smallest proportion of sites in humans are substitutions where humans have the high EP amino acid with non-human primates have the low EP amino acid.

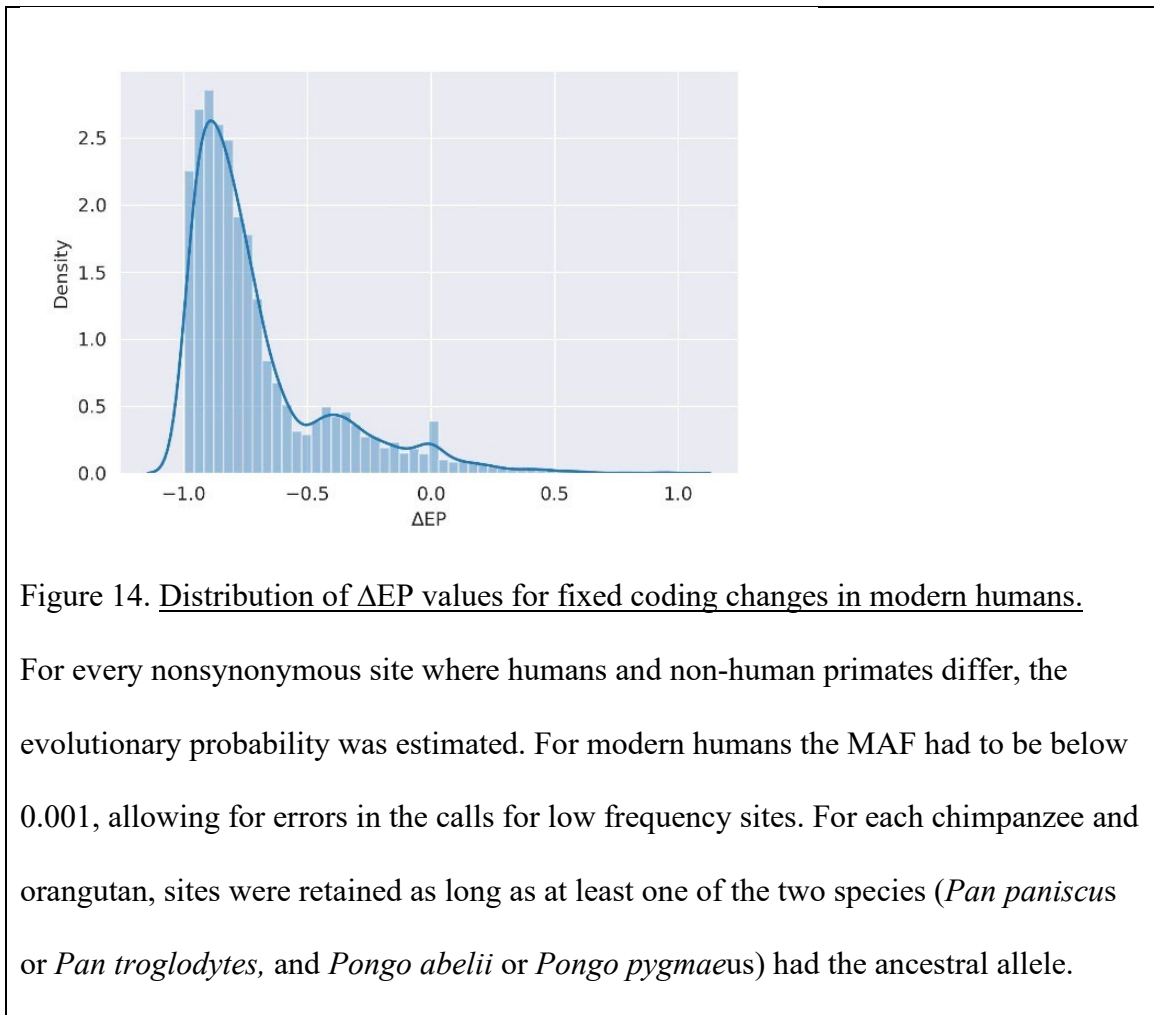


Figure 14. Distribution of ΔEP values for fixed coding changes in modern humans.

For every nonsynonymous site where humans and non-human primates differ, the evolutionary probability was estimated. For modern humans the MAF had to be below 0.001, allowing for errors in the calls for low frequency sites. For each chimpanzee and orangutan, sites were retained as long as at least one of the two species (*Pan paniscus* or *Pan troglodytes*, and *Pongo abelii* or *Pongo pygmaeus*) had the ancestral allele.

It is expected then that substitutions in chimpanzees should follow the same pattern with most fixations on the chimp branch yielding very low EP amino acids in comparison to high EP ancestral amino acid. This is found to be the case with the vast majority of substitutions on the chimpanzee lineage being towards low EP amino acids (Appendix C Figure 25). All substituted sites where chimpanzees have a derived allele while modern humans and other primates all have the ancestral allele were identified yielding 17255 total sites in the coding region with 6983 of these sites being nonsynonymous and 10272 yielding synonymous changes. However, it should be noted that on the chimpanzee branch there is a slightly larger proportion of high EP amino acid substitutions, still as a minority of sites. This is likely due to bias in evolutionary probability estimates as chimpanzees are included in the vertebrate tree when calculating these probability values.

Comparison of Observed ΔEP to Expected, Simulated ΔEP Distributions Show Differences Which may be Indicative of Sites Under Selection.

To identify whether the observed distributions of ΔEP values of substitutions in modern humans reflects what is expected, an expected distribution is generated. This distribution is based on the ancestral state at each amino acid position and all corresponding possible mutations at that position. For each of these potential amino acids the probability of that amino acid at that loci are calculated and averaged for the entire site. With these values, the observed ΔEP of the derived substituted amino acid in modern humans can be compared with the expected average ΔEP value from all possible

mutations. Depending on the ancestral state probability, several different hypotheses can be tested.

First, a comparison between sites where both the ancestral and substitute, derived amino acid are both found with similar probability yielding a ΔEP near zero is undertaken. Sites were restricted to just those where the ancestral amino acid has a probability of at least 0.2 to ensure that these are not cases where both the ancestral and derived are not found anywhere else in the phylogeny. The expectation is that if a mutation is tolerated at these sites due to lower conservation rates, then random mutations should yield amino acids with similar EP values. However, the results show that random mutations simulated at these sites yield more forbidden amino acids occurring by chance than appear observed substituted in humans (Appendix C Figure 26). One explanation for the increased number of the substituted sites found with a high EP value in modern humans is that these sites may have rose in frequency to fixation due to selection at these sites, however, these sites fixing due to genetic drift alone cannot be ruled out.

Another comparison that can be tested with simulated expectation data is whether the observed data matches what is expected when the ancestral amino acid is preferred at a site. Simulating the potential mutations for sites where the substituted derived amino acid has a low EP (negative ΔEP sites), yields a similar distribution to the observed ΔEP values with no significant difference between the distributions (Appendix C Figure 27).

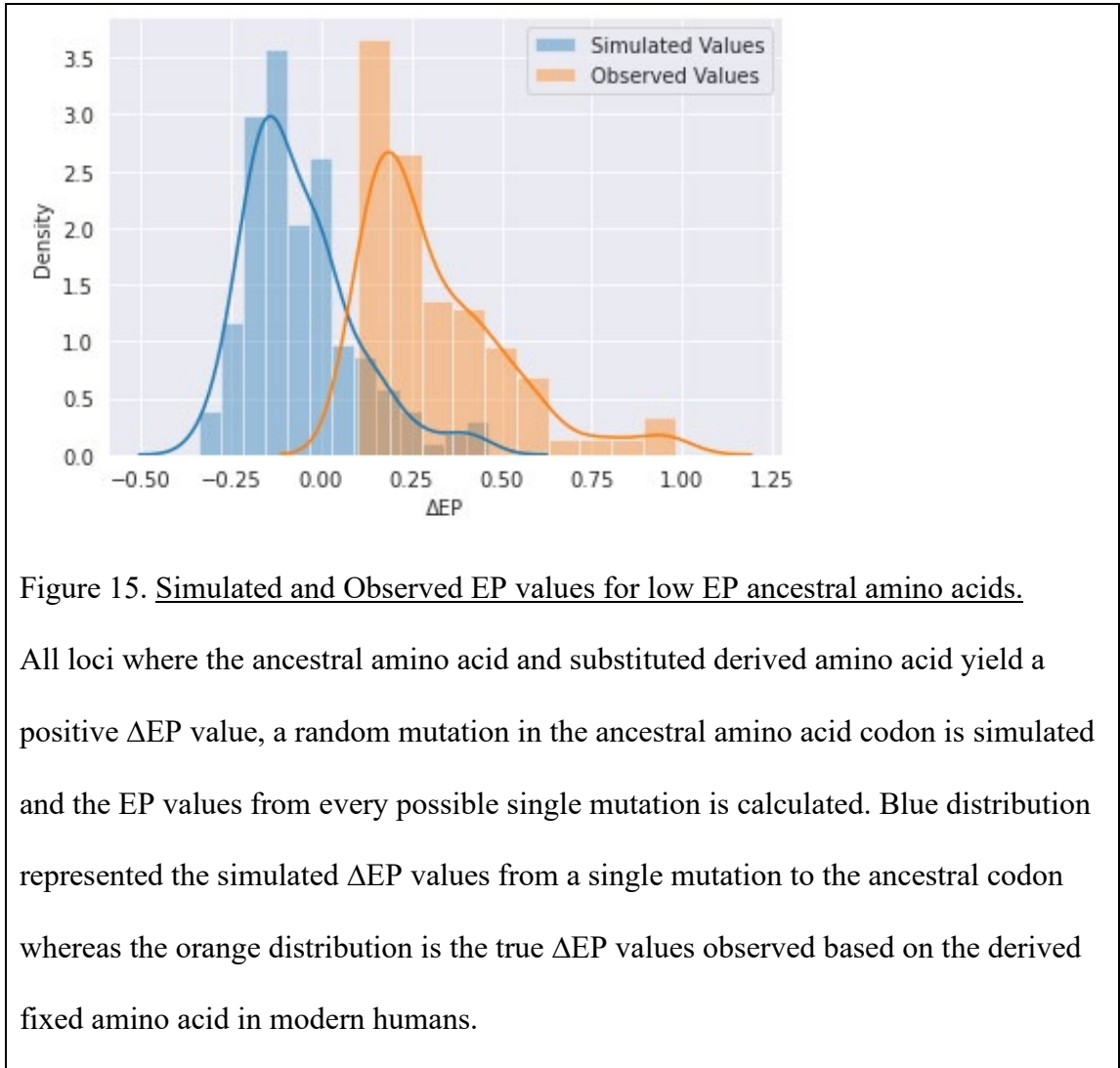
Additionally, an analysis was undertaken simulating potential mutations for loci where the ancestral amino acid is less likely while the derived, fixed amino acid is preferred (meaning it yields a positive ΔEP value). Of the total 168 loci where the derived substituted yielded a positive ΔEP value, most of these sites (134 loci) had an ancestral

amino acid and the derived amino acid with EP values higher than 0.2. Again, the analysis was restricted to these sites that moderate probability to ensure that there would be possible amino acids to switch between versus a single amino acid dominating the phylogeny. The expectation is that if these sites are less conserved then multiple amino acids would be permissible yielding similar probabilities. However, if the average EP value for a random mutation yields a low EP amino acid, then these would be candidates of positive selection.

As these are sites where there is an observed positive ΔEP value, primates would have to have an ancestral amino acid that is not very likely, and there must be at least one high EP amino acid at that site achievable by a single mutation. If most other amino acids are at low EP, then the average ΔEP should be negative or near zero. However, if there are several permissible amino acids then the average would skew above zero. The results that most loci yielded simulated ΔEP values below zero, skewing the distribution left to the observed one (Figure 15). This indicates that for most of these sites, there is just one likely amino acid which has become fixed in modern humans whereas all other non-human primates have the low EP amino acid. These sites are then possible candidates of being under positive selection as the more evolutionary likely amino acid is becoming fixed in the human lineage.

These are all comparisons of sites where modern humans and other non-human primates differ. However, comparing sites where amino acids are conserved across primate phylogeny allows for identification of the distribution of possible mutation effects where there are not known mutations in primates. The expectation here would be that most sites would have the highest probability amino acid with all other potential

mutations yielding less likely amino acids. Comparing the distributions observed for sites fixed in modern humans to the distribution of potential mutations at invariant sites across all primates, the majority of sites are fixed for the highest EP amino acid with very few even having multiple permissible amino acids (Figure 16).



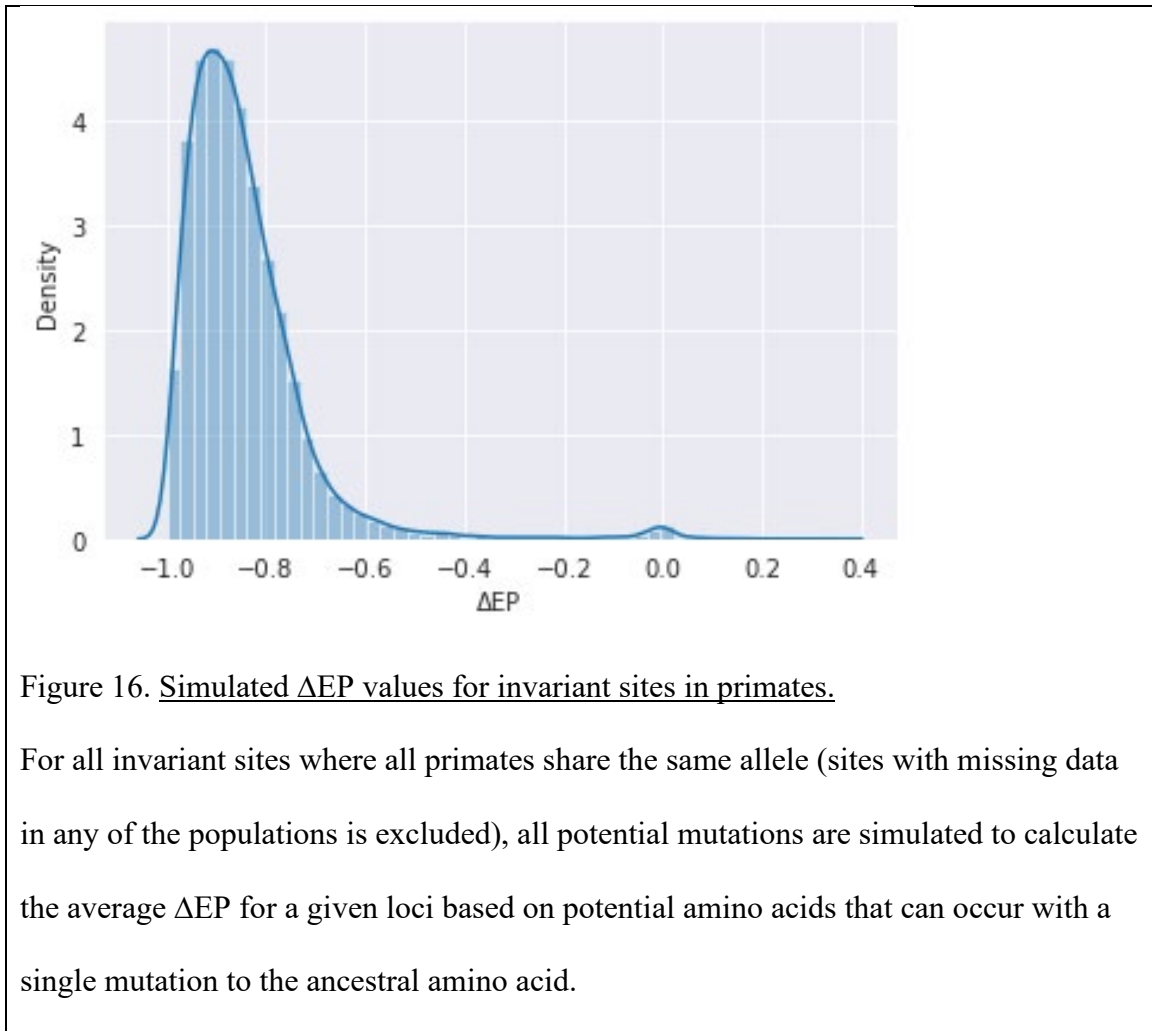


Figure 16. Simulated ΔEP values for invariant sites in primates.

For all invariant sites where all primates share the same allele (sites with missing data in any of the populations is excluded), all potential mutations are simulated to calculate the average ΔEP for a given loci based on potential amino acids that can occur with a single mutation to the ancestral amino acid.

Commonly Identified Phenotypic Traits of Adaptation in Humans do not Show any Significant Difference in ΔEP Values.

For each site, its corresponding gene was identified using REFseq annotation. Using the identified gene region for each site allowed for the comparison of the distribution of EP values for each of the three gene sets of interest: cognition, integumentary system, and musculoskeletal development associated with bipedalism. Each gene set was identified using Gene Ontology and literature searches for traits of interest. To identify whether the sites in these genes are under selection, the distribution

of probability values is compared to sites from a presumed neutral set of genes. In this case, the control set were genes identified as housekeeping genes which are predicted to be conserved across phylogeny. The housekeeping genes were identified based on Gini coefficient values (160). For each gene set, the distribution of ΔEP values was plotted with all three gene sets showing a similar pattern with a peak at -1 (Figure 17). This corresponds to sites where modern humans have the low EP derived allele and non-human primates have the ancestral, high EP allele. Additionally, the housekeeping genes had a similar distribution of ΔEP values with a peak around the extreme negative ΔEP values.

To test whether any of the three test gene sets had enrichment for either negative or positive ΔEP values, the number of nonsynonymous and synonymous substitutions were counted for each of the test sets and the control set of housekeeping genes. The number of synonymous substitutions acts as the control for rates of substitutions across each of the genes along with controlling for differences in gene lengths. If a test gene set would have significantly more negative ΔEP substitutions than expected given the distribution of the control set, then the adaptations are occurring via novel amino acids. Conversely, if there is significant difference in positive ΔEP counts, then adaptation is occurring by returning to a more stable amino acid. If neither, positive nor negative ΔEP counts are significantly different for a test gene set, then signals of adaptation were not found.

For negative ΔEP , for each of the comparisons the gene set was not significantly different from the housekeeping set using a Chi Square analysis (Cognition p-value: 0.75, Integumentary System p-value: 0.72, Bipedalism p-value: 0.51) (Table 4). For positive

ΔEP , each of these comparisons using a Chi-Square analysis was also not significant between the gene set of interest and the housekeeping genes (Cognition p-value: 0.99, Integumentary System p-value: 0.12, Biped p-value: 0.43) (Table 5). While there are many more negative ΔEP fixed sites than positive, there was no difference in the distribution for any of the three test sets and the control set indicating there is no evidence of selection on any of these three test gene sets.

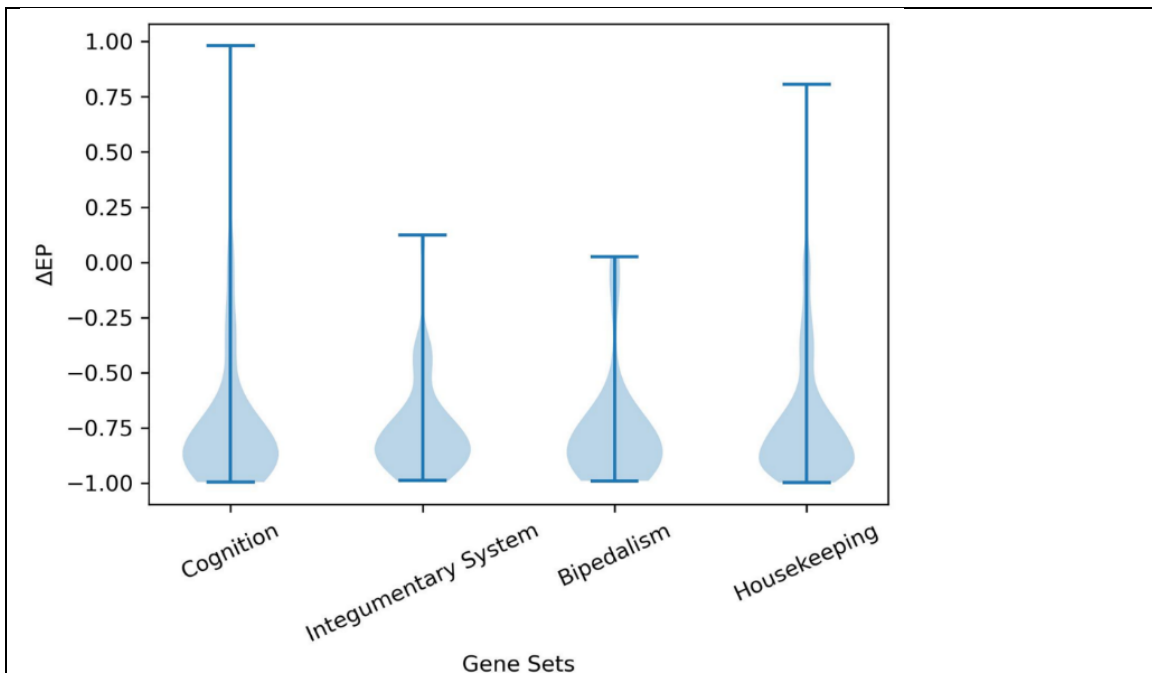


Figure 17. Distribution of ΔEP values for each of the three gene sets and the control housekeeping gene set.

Each nonsynonymous substituted site was annotated with the gene in which it fell. Using GO terms for cognition, skin development, and skeletal system development, lists of genes for the three categories of interest were assembled. For housekeeping genes as a control, genes with a low Gini coefficient that are conserved across species were assembled. Distributions of ΔEP for sites that fell within each of the four gene set categories are plotted in a violin plot.

Table 4. Counts of fixed, negative ΔEP alleles between human and non-human primates.

	Cognition	Integumentary System	Bipedalism	Housekeeping
Nonsynonymous	91	73	42	951
Synonymous	241	119	79	1886

Counts of nonsynonymous and synonymous fixed sites that yield negative ΔEP values for each of the test gene sets in addition to the control housekeeping gene set.

Table 5. Counts of fixed, positive ΔEP alleles between human and non-human primates.

	Cognition	Integumentary System	Bipedalism	Housekeeping
Nonsynonymous	4	2	2	40
Synonymous	241	119	79	1886

Counts of nonsynonymous and synonymous fixed sites that yield positive ΔEP values for each of the test gene sets in addition to the control housekeeping gene set.

No Evidence of Enrichment of Substituted Sites for Genes Identified to be Under Positive Selection.

Genes where the hominin branch was identified to be under positive selection were identified using aBSREL. Of the 4438 genes for which there were primate alignments available after filtering, 31 genes were identified to be under positive selection on the hominin branch (Appendix C Table 14). Of these 31 genes, there is an enrichment for gene sets associated with ammonium including ammonium homeostasis

and ammonium transmembrane transport along with the Rh blood group pathway (Appendix C Table 15). Only 11 nonsynonymous derived substitutions on the hominin branch are found in the 31 positively selected genes (Appendix C Table 16). The 11 fixed differences on the hominin branch are found across 6 of the positively selected genes: RAD54L, BBS12, MSH4, PIGR, FAM161A, and APLF. The distribution of ΔEP values for the substitutions found in the positively selected genes in comparison the substitutions of the genes not found to be under positive selection are not significantly different (Appendix C Figure 28).

Discussion

Thousands of sites are found where there is a substitution on the hominin branch in comparison with the non-human primates in the phylogeny which aligns with previous results (57,154). These sites represent mutations that change the protein coding sequence, and thus can create a functional impact potentially affecting the phenotype. Here, we analyze the evolutionary probability of the sites to determine whether the mechanism of adaptation can be elucidated: adaptation arising by novelty or stability. We find that a small number of sites are instances where humans have a derived mutation that creates a low EP amino acid which transitioned from an equally unlikely ancestral amino acid found in the other non-human primates. This could be indicative regions of relaxed selection with lower conservation on these amino acids. This would mean the sites are better able to handle novel amino acids (171,172) with some previous literature supporting relaxed selection as a mechanism for phenotypic adaptation in humans (173,174). However, when random mutations are simulated at these sites, we found that a

random mutation is more likely to yield a less permissible amino acid whereas we observe a more likely permissible amino acid in modern humans. Both non-human primates and humans have two different low EP amino acids meaning either these sites are pushed to fixation due to genetic drift by some random chance or there is a potential for these sites to have been fixed due to selection.

While these sites made up a small fraction of the discovered mutations, the largest number of substitutions are instances where the derived amino acid in modern humans has low EP whereas the ancestral state found in all other non-human primates has a high EP amino acid. This is perhaps expected as the ancestral allele shared across non-human primates would be much older and more likely to be shared widely across the vertebrate phylogeny (14). This means for modern humans that at some point since the common ancestor to chimpanzees there was a mutation to a novel amino acid which is now carried by all modern humans. This indicates that most substitutions in the coding region is occurring due by way of novel amino acids arising in the population which then become fixed in the lineage. Of the sites that become fixed in modern humans, a subset of these are neutral or weakly deleterious mutations being fixed due to genetic drift. As the majority of de novo mutations are weakly deleterious, this is likely a large subset especially if the mutation is only very weakly deleterious (8,175,176). Of the substitutions identified in modern humans, most of the mutations are also found in all four genomes of archaic humans indicating the mutation arose prior the speciation between archaic and modern humans estimated around 0.9 – 1.4 million years ago (94,177,178). For thirty of the derived amino acid substitutions in modern humans do all four archaic genomes have the ancestral amino acid shared with all other non-human

primates which indicates the derived mutation arose more recently (Appendix C Figure 29). Another forty sites are found polymorphic in the four archaic human populations. This is indicative of more complex demographic history such as these sites being polymorphic in the common ancestor of archaic and modern humans or arising due to gene flow between the populations.

Having identified several thousand fixed, derived alleles on the hominin branch, we looked at the probability of the corresponding amino acid for these mutations. As the question was to identify the mechanism of adaptation, the sites that fell into three distinct gene tests were compared to a control set of housekeeping genes. The test sets were genes associated with phenotypes of increased cognition, integumentary development, and changes in the musculoskeletal system for manipulation and bipedalism (155–158). We found that there appears to be no significant difference between the distribution of each of the test gene sets with the control set of housekeeping genes. The housekeeping genes are predicted to be well conserved across the phylogeny and crucial to function (160) yet across all four gene sets have approximately the same distribution of nonsynonymous to synonymous substitutions. While most substitutions on the hominin lineage are arising due to novel amino acids as compared to the vertebrate tree, these substitutions do not appear to correspond directly with gene sets associated with common evolutionary phenotypes in modern humans.

To disentangle any biases introduced through selection of the test gene sets, genes under selection were also identified using aBSREL (166). Since the majority of substitutions were found to yield novel amino acids on the hominin branch, we can test whether a drop in evolutionary probability is a property of in the rapid evolution of genes.

To do this, we used the measure of dN/dS to identify genes where positive selection was detected specifically on the hominin branch. However, of the 31 genes identified to be under positive selection, only 11 substitutions appeared in 6 of those 31 genes. Thus, there was no evidence of an enrichment of substituted sites in genes identified to be under selection on the hominin branch.

While there was no evidence of clear candidates of adaptation in the pool of substitutions identified in modern humans, there were several thousand substitutions identified that were found in both modern and archaic humans that arose. A further direction would be to further analyze these sites using other methods of variant prediction to better understand the effects of these mutations on the gene function (116,179). In addition, regulatory variation has been identified as being another mechanism of adaptation (154) and further work needs to be done to identify substitutions in the regulatory regions which can then be better characterized (180,181). The issue of identifying the causative mutations of adaptation in modern humans and attempting to better understand the mechanism by which these mutations are arising clearly needs to be approached from both a protein coding and regulatory perspective. While the phenotypic changes in modern humans are clear, the mechanism by which these changes have occurred are still not understood and further work characterizing the substitutions between modern humans and non-human primates may elucidate the issue.

REFERENCES

1. Maruyama T. The age of an allele in a finite population*. *Genet Res.* 1974 Apr;23(2):137–43.
2. Kimura M, Ohta T. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics.* 1969 Mar;61(3):763–71.
3. Fay JC, Wyckoff GJ, Wu CI. Positive and Negative Selection on the Human Genome. *Genetics.* 2001 Jul 1;158(3):1227–34.
4. Pálsson S, Pamilo P. The Effects of Deleterious Mutations on Linked, Neutral Variation in Small Populations. *Genetics.* 1999 Sep 1;153(1):475–83.
5. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. *Science.* 2011 Feb 18;331(6019):920–4.
6. Schrider DR, Kern AD. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Mol Biol Evol.* 2017 Aug 1;34(8):1863–77.
7. Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol Evol.* 2017;8(6):700–16.
8. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLoS Genet* [Internet]. 2008 May 30 [cited 2020 May 11];4(5). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2377339/>
9. Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, Boomsma DI, et al. Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency. *PLoS Genet* [Internet]. 2013 Feb 28 [cited 2020 Jun 8];9(2). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3585140/>
10. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001 Feb;409(6822):860–921.
11. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. *Science.* 2001 Feb 16;291(5507):1304–51.
12. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. *Nature.* 2015 Oct;526(7571):68–74.
13. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. *Nature.* 2015 Oct;526(7571):82–90.

14. Liu L, Tamura K, Sanderford M, Gray VE, Kumar S. A Molecular Evolutionary Reference for the Human Variome. *Mol Biol Evol.* 2016 Jan 1;33(1):245–54.
15. Patel R, Scheinfeldt LB, Sanderford MD, Lanham TR, Tamura K, Platt A, et al. Adaptive Landscape of Protein Variation in Human Exomes. *Mol Biol Evol.* 2018 Aug 1;35(8):2015–25.
16. Patel R, Kumar S. On estimating evolutionary probabilities of population variants. *BMC Evol Biol.* 2019 Jun 25;19(1):133.
17. Ragoussis J. Genotyping Technologies for Genetic Research. *Annu Rev Genomics Hum Genet.* 2009;10(1):117–33.
18. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinform.* 2009 Jul 1;10(4):354–66.
19. Sohn J il, Nam JW. The present and future of de novo whole-genome assembly. *Brief Bioinform.* 2018 Jan 1;19(1):23–40.
20. Eklblom R, Wolf JBW. A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl.* 2014;7(9):1026–42.
21. Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglu S. Whole-Genome Sequencing and Assembly with High-Throughput, Short-Read Technologies. *PLOS ONE.* 2007 May 30;2(5):e484.
22. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet.* 2020 Oct;21(10):597–614.
23. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet.* 2011 Nov;12(11):745–55.
24. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature.* 2021 Nov;599(7886):628–34.
25. Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, et al. The sequences of 150,119 genomes in the UK Biobank. *Nature.* 2022 Jul;607(7920):732–40.
26. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet.* 2012 Oct 15;21(R1):R1–9.
27. Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. Ultrarare variants drive substantial cis heritability of human gene expression. *Nat Genet.* 2019 Sep;51(9):1349–55.

28. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*. 2021 Sep;597(7877):527–32.
29. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 2007 Aug;8(8):610–8.
30. Saint Pierre A, Génin E. How important are rare variants in common disease? *Brief Funct Genomics*. 2014 Sep 1;13(5):353–61.
31. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*. 2010 Jun;11(6):415–25.
32. Pizzo L, Jensen M, Polyak A, Rosenfeld JA, Mannik K, Krishnan A, et al. Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genet Med*. 2019 Apr;21(4):816–25.
33. Zhu N, Swietlik EM, Welch CL, Pauciulo MW, Hagen JJ, Zhou X, et al. Rare variant analysis of 4241 pulmonary arterial hypertension cases from an international consortium implicates FBLN2, PDGFD, and rare de novo variants in PAH. *Genome Med*. 2021 May 10;13(1):80.
34. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. *J Hum Genet*. 2021 Jan;66(1):11–23.
35. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics*. 1993 Aug 1;134(4):1289–303.
36. Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, et al. Population Genetic Inference from Personal Genome Data: Impact of Ancestry and Admixture on Human Genomic Variation. *Am J Hum Genet*. 2012 Oct 5;91(4):660–71.
37. Mathieson I, McVean G. Demography and the Age of Rare Variants. *PLoS Genet* [Internet]. 2014 Aug 7 [cited 2020 May 12];10(8). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4125085/>
38. Watterson GA. Some Theoretical Aspects of Diffusion Theory in Population Genetics. *Ann Math Stat*. 1962 Sep;33(3):939–57.
39. Kimura M, Ohta T. The Age of a Neutral Mutant Persisting in a Finite Population. *Genetics*. 1973 Sep 1;75(1):199–212.
40. Li WH. The first arrival time and mean age of a deleterious mutant gene in a finite population. *Am J Hum Genet*. 1975 May;27(3):274–86.

41. Griffiths RC, Tavaré S. The age of a mutation in a general coalescent tree. *Commun Stat Stoch Models*. 1998 Jan 1;14(1–2):273–95.
42. Griffiths RC, Tavaré S. Sampling Theory for Neutral Alleles in a Varying Environment. *Philos Trans Biol Sci*. 1994;344(1310):403–10.
43. Slatkin M, Rannala B. Estimating Allele Age. *Annu Rev Genomics Hum Genet*. 2000;1(1):225–49.
44. Rannala B, Slatkin M. Likelihood Analysis of Disequilibrium Mapping, and Related Problems. *Am J Hum Genet*. 1998 Feb 1;62(2):459–73.
45. Slatkin M, Rannala B. Estimating the age of alleles by use of intraallelic variability. *Am J Hum Genet*. 1997 Feb;60(2):447–58.
46. Mahmoudi A, Koskela J, Kelleher J, Chan Y ban, Balding D. Bayesian inference of ancestral recombination graphs. *PLOS Comput Biol*. 2022 Mar 9;18(3):e1009960.
47. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genet*. 2014 May 15;10(5):e1004342.
48. Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet*. 2019 Sep;51(9):1321–9.
49. Albers PK, McVean G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biol*. 2020 Jan 17;18(1):e3000586.
50. Platt A, Pivrotto A, Knoblauch J, Hey J. An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations. *PLOS Genet*. 2019 Aug 19;15(8):e1008340.
51. Li N, Stephens M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*. 2003 Dec 1;165(4):2213–33.
52. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. *Genet Res*. 1974 Feb;23(1):23–35.
53. Uricchio LH, Petrov DA, Enard D. Exploiting selection at linked sites to infer the rate and strength of adaptation. *Nat Ecol Evol*. 2019 Jun;3(6):977–84.
54. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. *Genome Res*. 2014 Jun 1;24(6):885–95.
55. Galtier N. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLOS Genet*. 2016 Jan 11;12(1):e1005774.

56. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The Role of Geography in Human Adaptation. *PLOS Genet.* 2009 Jun 5;5(6):e1000500.
57. Waterson RH, Lander ES, Wilson RK, The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature.* 2005 Sep;437(7055):69–87.
58. Zhen Y, Huber CD, Davies RW, Lohmueller KE. Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and *Drosophila melanogaster*. *Genome Res.* 2021 Jan 1;31(1):110–20.
59. Huber CD, Kim BY, Marsden CD, Lohmueller KE. Determining the factors driving selective effects of new nonsynonymous mutations. *Proc Natl Acad Sci.* 2017 Apr 25;114(17):4465–70.
60. Garud NR, Messer PW, Petrov DA. Detection of hard and soft selective sweeps from *Drosophila melanogaster* population genomic data. *PLOS Genet.* 2021 Feb 26;17(2):e1009373.
61. Harris RB, Sackman A, Jensen JD. On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *PLOS Genet.* 2018 Dec 28;14(12):e1007859.
62. Charlesworth B, Jensen JD. Effects of Selection at Linked Sites on Patterns of Genetic Variability. *Annu Rev Ecol Evol Syst.* 2021;52(1):177–97.
63. Souilmi Y, Tobler R, Johar A, Williams M, Grey ST, Schmidt J, et al. Admixture has obscured signals of historical hard sweeps in humans. *Nat Ecol Evol.* 2022 Dec;6(12):2003–15.
64. Novembre J, Galvani AP, Slatkin M. The Geographic Spread of the CCR5 Δ 32 HIV-Resistance Allele. *PLOS Biol.* 2005 Oct 18;3(11):e339.
65. Muktopavela RA, Petr M, Ségurel L, Korneliussen T, Novembre J, Racimo F. Modeling the spatiotemporal spread of beneficial alleles using ancient genomes. Andrés AM, Perry GH, Alves I, Eriksson A, editors. *eLife.* 2022 Dec 20;11:e73767.
66. Pyott SJ, van Tuinen M, Screven LA, Schrode KM, Bai JP, Barone CM, et al. Functional, Morphological, and Evolutionary Characterization of Hearing in Subterranean, Eusocial African Mole-Rats. *Curr Biol.* 2020 Nov 16;30(22):4329-4341.e4.
67. Dolatyabi S, Peighambari SM, Razmyar J. Molecular detection and analysis of beak and feather disease viruses in Iran. *Front Vet Sci [Internet].* 2022 [cited 2023 Aug 29];9. Available from: <https://www.frontiersin.org/articles/10.3389/fvets.2022.1053886>

68. Xu K, Kosoy R, Shameer K, Kumar S, Liu L, Readhead B, et al. Genome-wide analysis indicates association between heterozygote advantage and healthy aging in humans. *BMC Genet.* 2019 Jul 2;20(1):52.
69. Tian R, Pan Y, Etheridge THA, Deshmukh H, Gulick D, Gibson G, et al. Pitfalls in Single Clone CRISPR-Cas9 Mutagenesis to Fine-Map Regulatory Intervals. *Genes.* 2020 May;11(5):504.
70. Ose NJ, Campitelli P, Patel R, Kumar S, Ozkan SB. Protein dynamics provide mechanistic insights about epistasis among common missense polymorphisms. *Biophys J.* 2023 Jul 25;122(14):2938–47.
71. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl comparative genomics resources. *Database.* 2016 Jan 1;2016:bav096.
72. Wright S. The Distribution of Gene Frequencies in Populations. *Proc Natl Acad Sci.* 1937 Jun;23(6):307–20.
73. Wright S. The Distribution of Gene Frequencies Under Irreversible Mutation. *Proc Natl Acad Sci.* 1938 Jul;24(7):253–9.
74. Kimura M. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet Res.* 1968 Jun;11(3):247–70.
75. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 2004 Jan 1;32(suppl_1):D493–6.
76. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2005 Jan 1;33(suppl_1):D501–4.
77. The International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007 Oct;449(7164):851–61.
78. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D884–91.
79. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics.* 2019 Nov 1;35(21):4453–5.
80. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. *Nucleic Acids Res.* 2022 Jan 7;50(D1):D988–95.

81. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*. 2012 Apr 1;6(2):80–92.
82. The ENCODE Project Consortium, Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*. 2020 Jul 30;583(7818):699–710.
83. Farrell CM, Goldfarb T, Rangwala SH, Astashyn A, Ermolaeva OD, Hem V, et al. RefSeq Functional Elements as experimentally assayed nongenic reference standards and functional interactions in human and mouse. *Genome Res*. 2022 Jan 1;32(1):175–88.
84. Lesurf R, Cotto KC, Wang G, Griffith M, Kasaian K, Jones SJM, et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res*. 2016 Jan 4;44(D1):D126–32.
85. Box GEP, Cox DR. An Analysis of Transformations. *J R Stat Soc Ser B Methodol*. 1964 Jul 1;26(2):211–43.
86. Spence JP, Song YS. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci Adv*. 2019 Oct 23;5(10):eaaw9206.
87. Wright S. Coefficients of Inbreeding and Relationship. *Am Nat*. 1922 Jul;56(645):330–8.
88. Kelleher J, Thornton KR, Ashander J, Ralph PL. Efficient pedigree recording for fast population genetics simulation. *PLOS Comput Biol*. 2018 Nov 1;14(11):e1006581.
89. Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, et al. A community-maintained standard library of population genetic models. Coop G, Wittkopp PJ, Novembre J, Sethuraman A, Mathieson S, editors. *eLife*. 2020 Jun 23;9:e54967.
90. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science [Internet]*. 2012 Jul 6 [cited 2023 Jan 20]; Available from: <https://www.science.org/doi/10.1126/science.1219240>
91. Mi H, Thomas P. PANTHER Pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol Clifton NJ*. 2009;563:123–40.
92. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci*. 2022;31(1):8–22.

93. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc.* 2019 Mar;14(3):703–21.
94. Mafessoni F, Grote S, de Filippo C, Slon V, Kolobova KA, Viola B, et al. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc Natl Acad Sci.* 2020 Jun 30;117(26):15132–6.
95. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature.* 2014 Jan;505(7481):43–9.
96. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science.* 2017 Nov 3;358(6363):655–8.
97. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science.* 2012 Oct 12;338(6104):222–6.
98. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell.* 2018 Mar 22;173(1):53-61.e9.
99. Siewert KM, Voight BF. BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data. *Genome Biol Evol.* 2020 Feb 3;12(2):3873–7.
100. Fisher RA. *The genetical theory of natural selection.* Рипол Классик; 1930. 289 p.
101. Dobzhansky T. Mendelism, Darwinism, and Evolutionism. *Proc Am Philos Soc.* 1965;109(4):205–15.
102. De Sanctis B, Krukov I, de Koning APJ. Allele Age Under Non-Classical Assumptions is Clarified by an Exact Computational Markov Chain Approach. *Sci Rep.* 2017 Sep 19;7(1):11869.
103. Dobzhansky T. *Genetics of the Evolutionary Process.* Columbia University Press; 1970. 524 p.
104. Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, et al. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science.* 2013 Mar 29;339(6127):1578–82.
105. Bitarello BD, de Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D, et al. Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biol Evol.* 2018 Mar 1;10(3):939–55.

106. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol*. 2005;128(2):415–23.
107. Soni V, Vos M, Eyre-Walker A. A new test suggests hundreds of amino acid polymorphisms in humans are subject to balancing selection. *PLOS Biol*. 2022 Jun 2;20(6):e3001645.
108. Aqil A, Speidel L, Pavlidis P, Gokcumen O. Balancing selection on genomic deletion polymorphisms in humans. Messer PW, Weigel D, editors. *eLife*. 2023 Jan 10;12:e79111.
109. Assaf ZJ, Petrov DA, Blundell JR. Obstruction of adaptation in diploids by recessive, strongly deleterious alleles. *Proc Natl Acad Sci*. 2015 May 19;112(20):E2658–66.
110. Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. Estimating the mutation load in human genomes. *Nat Rev Genet*. 2015 Jun;16(6):333–43.
111. Sellis D, Callahan BJ, Petrov DA, Messer PW. Heterozygote advantage as a natural consequence of adaptation in diploids. *Proc Natl Acad Sci*. 2011 Dec 20;108(51):20666–71.
112. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. *G3 GenesGenomesGenetics*. 2015 Aug 1;5(8):1543–50.
113. Schwarze K, Buchanan J, Fermont JM, Dreau H, Tilley MW, Taylor JM, et al. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genet Med*. 2020 Jan 1;22(1):85–94.
114. Almogly G, Pratt M, Oberstrass F, Lee L, Mazur D, Beckett N, et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform [Internet]. *bioRxiv*; 2022 [cited 2024 Jan 19]. p. 2022.05.29.493900. Available from: <https://www.biorxiv.org/content/10.1101/2022.05.29.493900v4>
115. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, M. Mastrogiannis G, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008 Oct;455(7216):1061–8.
116. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018 Jan 4;46(D1):D1062–7.

117. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020 May;581(7809):434–43.
118. Piñero J, Ramírez-Anguaita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020 Jan 8;48(D1):D845–55.
119. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet* [Internet]. 2020 [cited 2023 Nov 3];11. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00424>
120. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N Engl J Med*. 2013 Oct 17;369(16):1502–11.
121. Goh G, Choi M. Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases. *Genomics Inform*. 2012 Dec;10(4):214–9.
122. Salfati EL, Spencer EG, Topol SE, Muse ED, Rueda M, Lucas JR, et al. Re-analysis of whole-exome sequencing data uncovers novel diagnostic variants and improves molecular diagnostic yields for sudden death and idiopathic diseases. *Genome Med*. 2019 Dec 17;11(1):83.
123. Ma Y, Jun GR, Zhang X, Chung J, Naj AC, Chen Y, et al. Analysis of Whole-Exome Sequencing Data for Alzheimer Disease Stratified by APOE Genotype. *JAMA Neurol*. 2019 Sep 1;76(9):1099–108.
124. Grzymski JJ, Elhanan G, Morales Rosado JA, Smith E, Schlauch KA, Read R, et al. Population genetic screening efficiently identifies carriers of autosomal dominant diseases. *Nat Med*. 2020 Aug;26(8):1235–9.
125. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. *Genet Med*. 2016 Sep;18(9):906–13.
126. Grzymski JJ, Elhanan G, Smith E, Rowan C, Slotnick N, Dabe S, et al. Population Health Genetic Screening for Tier 1 Inherited Diseases in Northern Nevada: 90% of At-Risk Carriers are Missed [Internet]. 2019 May [cited 2022 Jan 7] p. 650549. Available from: <https://www.biorxiv.org/content/10.1101/650549v1>
127. The “All of Us” Research Program. *N Engl J Med*. 2019 Aug 15;381(7):668–76.

128. Slatkin M. Allele age and a test for selection on rare alleles. Charlesworth B, Harvey PH, editors. *Philos Trans R Soc Lond B Biol Sci*. 2000 Nov 29;355(1403):1663–8.
129. Hateley S, Lopez-Izquierdo A, Jou CJ, Cho S, Schraiber JG, Song S, et al. The history and geographic distribution of a KCNQ1 atrial fibrillation risk allele. *Nat Commun*. 2021 Nov 8;12(1):6442.
130. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 2013 Jan;493(7431):216–20.
131. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genet*. 2009 Oct 23;5(10):e1000695.
132. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Comput Biol*. 2016 May 4;12(5):e1004842.
133. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021 Feb 16;10(2):giab008.
134. Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Mol Ecol Resour*. 2019;19(2):552–66.
135. Kim BY, Huber CD, Lohmueller KE. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics*. 2017 May 1;206(1):345–61.
136. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet*. 2017 Oct;18(10):599–612.
137. Biddanda A, Rice DP, Novembre J. A variant-centric perspective on geographic patterns of human allele frequency variation. Goldberg A, Perry GH, editors. *eLife*. 2020 Dec 22;9:e60107.
138. Y. C. Brandt D, Wei X, Deng Y, Vaughn AH, Nielsen R. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*. 2022 May 1;221(1):iyac044.
139. Ragsdale AP, Thornton KR. Multiple Sources of Uncertainty Confound Inference of Historical Human Generation Times. *Mol Biol Evol*. 2023 Aug 1;40(8):msad160.
140. Karlin S, McGregor J. The number of mutant forms maintained in a population. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967.

141. Kimura M. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. *Genetics*. 1969 Apr;61(4):893–903.
142. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016 Aug;536(7616):285–91.
143. McVean G, Awadalla P, Fearnhead P. A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. *Genetics*. 2002 Mar 1;160(3):1231–41.
144. Blumenfeld OO, Patnaik SK. Allelic genes of blood group antigens: A source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. *Hum Mutat*. 2004;23(1):8–16.
145. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*. 2007 Jan;39(1):31–40.
146. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007 Jun;447(7145):661–78.
147. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. *Nature*. 2020 Jan;577(7789):179–89.
148. Harris EE, Meyer D. The molecular signature of selection underlying human adaptations. *Am J Phys Anthropol*. 2006;131(S43):89–130.
149. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. *Science*. 2016 Nov 11;354(6313):760–4.
150. Jovanovic VM, Sarfert M, Reyna-Blanco CS, Indrischek H, Valdivia DI, Shelest E, et al. Positive Selection in Gene Regulatory Factors Suggests Adaptive Pleiotropic Changes During Human Evolution. *Front Genet* [Internet]. 2021 [cited 2023 Nov 16];12. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2021.662239>
151. Anderson JA, Vilgalys TP, Tung J. Broadening primate genomics: new insights into the ecology and evolution of primate gene regulation. *Curr Opin Genet Dev*. 2020 Jun 1;62:16–22.
152. Bakewell MA, Shi P, Zhang J. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci*. 2007 May;104(18):7489–94.
153. Pivirotto AM, Platt A, Patel R, Kumar S, Hey J. Analyses of allele age and fitness impact reveal human beneficial alleles to be older than neutral controls [Internet].

bioRxiv; 2023 [cited 2023 Nov 16]. p. 2023.10.09.561569. Available from: <https://www.biorxiv.org/content/10.1101/2023.10.09.561569v2>

154. Suntsova MV, Buzdin AA. Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species. *BMC Genomics*. 2020 Sep 10;21(7):535.
155. Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, et al. Accelerated Evolution of Nervous System Genes in the Origin of *Homo sapiens*. *Cell*. 2004 Dec 29;119(7):1027–40.
156. Dumas G, Malesys S, Bourgeron T. Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition. *Genome Res*. 2021 Mar 1;31(3):484–96.
157. Prang TC, Ramirez K, Grabowski M, Williams SA. *Ardipithecus* hand provides evidence that humans and chimpanzees evolved from an ancestor with suspensory adaptations. *Sci Adv*. 2021 Feb 24;7(9):eabf2474.
158. Zhang J, Webb DM, Podlaha O. Accelerated Protein Evolution and Origins of Human-Specific Features: *FOXP2* as an Example. *Genetics*. 2002 Dec 1;162(4):1825–35.
159. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. *Nature*. 2013 Jul;499(7459):471–5.
160. Joshi CJ, Ke W, Drangowska-Way A, O'Rourke EJ, Lewis NE. What are housekeeping genes? *PLoS Comput Biol*. 2022 Jul 13;18(7):e1010295.
161. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000 May;25(1):25–9.
162. The Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023 May 2;224(1):iyad031.
163. Kypriotou M, Huber M, Hohl D. The human epidermal differentiation complex: cornified envelope precursors, S100 proteins and the 'fused genes' family. *Exp Dermatol*. 2012;21(9):643–9.
164. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. *Nucleic Acids Res*. 2023 Jan 6;51(D1):D933–41.
165. Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, et al. Ensembl 2024. *Nucleic Acids Res*. 2024 Jan 5;52(D1):D891–9.

166. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Mol Biol Evol.* 2015 May 1;32(5):1342–53.
167. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics.* 2005 Mar 1;21(5):676–9.
168. Wisotsky SR, Kosakovsky Pond SL, Shank SD, Muse SV. Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril. *Mol Biol Evol.* 2020 Aug;37(8):2430–9.
169. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci.* 2005 Oct 25;102(43):15545–50.
170. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.* 2003 Jul;34(3):267–73.
171. Lahti DC, Johnson NA, Ajie BC, Otto SP, Hendry AP, Blumstein DT, et al. Relaxed selection in the wild. *Trends Ecol Evol.* 2009 Sep 1;24(9):487–96.
172. Snell-Rood EC, Van Dyken JD, Cruickshank T, Wade MJ, Moczek AP. Toward a population genetic framework of developmental evolution: the costs, limits, and consequences of phenotypic plasticity. *BioEssays.* 2010;32(1):71–81.
173. Deacon TW. A role for relaxed selection in the evolution of the language capacity. *Proc Natl Acad Sci.* 2010 May 11;107(supplement_2):9000–6.
174. Tiwary BK. The cognitive and speech genes are jointly shaped by both positive and relaxed selection in the human lineage. *Genomics.* 2020 Sep 1;112(5):2922–7.
175. Charlesworth B, Charlesworth D. Rapid fixation of deleterious alleles can be caused by Muller’s ratchet. *Genet Res.* 1997 Aug;70(1):63–73.
176. Santiago E, Caballero A. Joint Prediction of the Effective Population Size and the Rate of Fixation of Deleterious Mutations. *Genetics.* 2016 Nov 1;204(3):1267–79.
177. Bergström A, Stringer C, Hajdinjak M, Scerri EML, Skoglund P. Origins of modern human ancestry. *Nature.* 2021 Feb;590(7845):229–37.
178. Rogers AR, Harris NS, Achenbach AA. Neanderthal-Denisovan ancestors interbred with a distantly related hominin. *Sci Adv.* 2020 Feb 21;6(8):eaay5483.

179. Du J, Sudarsanam M, Pérez-Palma E, Ganna A, Francioli L, Iqbal S, et al. Variant Score Ranker—a web application for intuitive missense variant prioritization. *Bioinformatics*. 2019 Nov 1;35(21):4478–9.
180. Liu L, Sanderford MD, Patel R, Chandrashekar P, Gibson G, Kumar S. Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat Commun*. 2019 Dec;10(1):330.
181. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet*. 2017 Apr;49(4):618–24.
182. Tian X, Browning BL, Browning SR. Estimating the Genome-wide Mutation Rate with Three-Way Identity by Descent. *Am J Hum Genet*. 2019 Nov 7;105(5):883–93.
183. Tremblay M, Vézina H. New Estimates of Intergenerational Time Intervals for the Calculation of Age and Origins of Mutations. *Am J Hum Genet*. 2000 Feb 1;66(2):651–8.
184. Takahata N. Allelic genealogy and human evolution. *Mol Biol Evol*. 1993 Jan 1;10(1):2–22.

APPENDIX A

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

Table 7. Δ EP measures for fixed and polymorphic alleles.

Measure	Negative Δ EP		Positive Δ EP	
	Fixed (%)	Polymorphic (%)	Fixed (%)	Polymorphic (%)
Count	29348 (13.0%)	195747 (87.0%)	6941 (60.2%)	4585 (39.8%)
Mean frequency	1	0.025	1	0.096
Mean Ancestral EP	0.714	0.848	0.122	0.249
Mean Derived EP	0.052	0.017	0.602	0.515
Mean Δ EP	-0.663	-0.832	0.480	0.266

Based on maximum-likelihood rooting estimates of ancestral alleles (see Figure 1 for values based on Ensembl rooting). Simulated mean Δ EP was calculated for each SNP by considering all possible non-synonymous mutations and the corresponding EP value for the resulting amino acid in proportion to their mutation probabilities based on empirical estimates. 95% confidence intervals on the mean, determined by bias-corrected bootstrap, are given in parentheses.

Table 8. Results of simulation-based tests of dispersion of positive ΔEP SNPs.

Chr	# Genes	Observed Mean SNP Density / bp	Mean simulated density	Observed Variance in Density	Mean simulated variance in density	Estimated probability of simulating variance higher than observed
1	1849	0.00035	0.00034	1.21E-06	3.50E-07	0
2	1134	0.0003	0.0003	1.45E-06	2.72E-07	0
3	1000	0.00027	0.00027	8.35E-07	2.39E-07	0
4	695	0.00039	0.00039	1.70E-06	3.75E-07	0
5	815	0.00043	0.00043	2.49E-06	4.10E-07	0
6	950	0.00026	0.00025	6.85E-07	2.69E-07	0
7	807	0.00038	0.00038	1.75E-06	3.65E-07	0
8	603	0.00038	0.00039	1.93E-06	3.70E-07	0
9	701	0.0004	0.0004	1.67E-06	3.81E-07	0
10	667	0.00035	0.00036	1.67E-06	3.25E-07	0
11	1207	0.00055	0.00055	2.87E-06	5.76E-07	0
12	965	0.00025	0.00026	6.31E-07	2.47E-07	0
13	299	0.00023	0.00023	5.37E-07	1.95E-07	0
14	551	0.00023	0.00024	4.00E-07	2.27E-07	0
15	517	0.00021	0.00022	3.47E-07	1.83E-07	0
16	744	0.00057	0.00057	4.37E-06	5.82E-07	0
17	1052	0.00043	0.00043	1.80E-06	4.31E-07	0
18	251	0.00023	0.00023	3.11E-07	1.99E-07	0.055
19	1296	0.00076	0.00076	4.32E-06	7.38E-07	0
20	506	0.00035	0.00035	9.00E-07	3.99E-07	0
21	205	0.00064	0.00065	5.66E-06	9.64E-07	0
22	395	0.00049	0.00049	1.74E-06	4.78E-07	0

Table 9. Gene ontology results

Panther Pathways	Observed	Expected	Fold Enrichment	Raw P Value	FDR
Plasminogen activating cascade	4	0.06	64.59	9.14E-07	7.31E-05
Blood coagulation	6	0.17	35.81	3.33E-08	5.32E-06

GO Biological Process	Observed	Expected	Fold Enrichment	Raw P Value	FDR
plasminogen activation	3	0.04	74.87	1.62E-05	1.82E-02
blood coagulation, fibrin clot formation	5	0.09	57.19	6.10E-08	3.19E-04
protein activation cascade	5	0.09	52.79	8.68E-08	3.40E-04
blood coagulation	7	0.63	11.11	4.08E-06	7.11E-03
coagulation	8	0.64	12.55	3.30E-07	1.04E-03
hemostasis	7	0.65	10.80	4.89E-06	6.97E-03
positive regulation of heterotypic cell-cell adhesion	3	0.05	54.90	3.60E-05	3.32E-02
acute-phase response	6	0.15	39.22	2.03E-08	3.18E-04
acute inflammatory response	7	0.3	23.43	3.39E-08	2.65E-04
platelet aggregation	4	0.16	25.54	2.54E-05	2.65E-02
cell-cell adhesion	10	1.98	5.06	3.02E-05	2.96E-02
negative regulation of endopeptidase activity	8	0.9	8.86	4.10E-06	6.43E-03
regulation of endopeptidase activity	10	1.54	6.51	3.54E-06	6.95E-03
regulation of peptidase activity	11	1.65	6.68	8.55E-07	1.91E-03

Table 9. (continued)

GO Biological Process	Observed	Expected	Fold Enrichment	Raw P Value	FDR
regulation of proteolysis	14	2.71	5.17	4.91E-07	1.28E-03
negative regulation of peptidase activity	8	0.94	8.55	5.29E-06	6.91E-03
negative regulation of proteolysis	9	1.26	7.16	5.33E-06	6.43E-03

GO Molecular Function	Observed	Expected	Fold Enrichment	Raw P Value	FDR
serine-type endopeptidase inhibitor activity	6	0.37	16.31	2.52E-06	2.51E-03
endopeptidase inhibitor activity	8	0.67	12.00	4.57E-07	1.14E-03
endopeptidase regulator activity	8	0.72	11.15	7.81E-07	9.72E-04
peptidase regulator activity	9	0.86	10.51	2.46E-07	1.22E-03
peptidase inhibitor activity	8	0.69	11.56	6.01E-07	9.96E-04
protease binding	6	0.51	11.68	1.57E-05	1.30E-02

GO Cellular Component	Observed	Expected	Fold Enrichment	Raw P Value	FDR
fibrinogen complex	4	0.03	> 100	7.75E-08	1.98E-05
extracellular space	26	12.49	2.08	1.41E-04	1.92E-02
extracellular region	30	16.01	1.87	3.16E-04	3.23E-02
endocytic vesicle lumen	3	0.08	37.43	9.97E-05	1.57E-02

Table 9. (continued)

GO Cellular Component	Observed	Expected	Fold Enrichment	Raw P Value	FDR
platelet alpha granule lumen	7	0.24	28.68	9.23E-09	3.77E-06
platelet alpha granule	7	0.33	21.12	6.62E-08	1.93E-05
secretory granule	13	3.2	4.06	1.75E-05	2.98E-03
secretory vesicle	13	3.83	3.40	1.07E-04	1.57E-02
secretory granule lumen	12	1.17	10.29	2.56E-09	1.74E-06
cytoplasmic vesicle lumen	12	1.18	10.20	2.83E-09	1.45E-06
vesicle lumen	13	1.18	10.98	2.43E-10	2.48E-07
blood microparticle	13	0.52	24.78	1.43E-14	2.93E-11
endoplasmic reticulum lumen	11	1.14	9.62	2.46E-08	8.38E-06
collagen-containing extracellular matrix	11	1.58	6.96	5.79E-07	1.31E-04
extracellular matrix	11	2.09	5.25	8.22E-06	1.68E-03
external encapsulating structure	11	2.1	5.24	8.35E-06	1.55E-03
extracellular exosome	19	7.65	2.48	1.52E-04	1.94E-02
extracellular vesicle	19	7.73	2.46	1.75E-04	2.10E-02
extracellular membrane-bounded organelle	19	7.74	2.46	1.76E-04	1.89E-02
extracellular organelle	19	7.74	2.46	1.76E-04	2.00E-02

Panther Protein Class	Observed	Expected	Fold Enrichment	Raw P Value	FDR
protease inhibitor	6	0.48	12.57	1.05E-05	2.06E-03

Table 9. (continued)

Reactome Pathways	Observed	Expected	Fold Enrichment	Raw P Value	FDR
LDL remodeling	2	0.02	91.51	3.59E-04	2.98E-02
Plasma lipoprotein remodeling	3	0.12	25.74	2.76E-04	2.38E-02
Plasma lipoprotein assembly, remodeling, and clearance	4	0.24	16.39	1.29E-04	1.29E-02
GRB2: SOS provides linkage to MAPK signaling for Integrins	4	0.05	78.43	4.71E-07	2.35E-04
Integrin signaling	4	0.09	42.23	4.08E-06	1.02E-03
Platelet Aggregation (Plug Formation)	4	0.14	28.90	1.61E-05	3.35E-03
Platelet activation, signaling and aggregation	9	0.95	9.50	5.57E-07	1.98E-04
p130Cas linkage to MAPK signaling for integrins	4	0.05	73.21	5.95E-07	1.85E-04
Regulation of TLR by endogenous ligand	4	0.08	52.29	1.91E-06	5.29E-04
MyD88 deficiency (TLR2/4)	3	0.06	48.44	5.01E-05	5.94E-03
Diseases associated with the TLR signaling cascade	3	0.11	26.57	2.53E-04	2.34E-02
Diseases of Immune System	3	0.11	26.57	2.53E-04	2.43E-02
IRAK4 deficiency (TLR2/4)	3	0.07	45.75	5.83E-05	6.60E-03
Common Pathway of Fibrin Clot Formation	3	0.08	37.43	9.97E-05	1.08E-02
Formation of Fibrin Clot (Clotting Cascade)	5	0.14	35.19	5.35E-07	2.22E-04

Table 9. (continued)

Reactome Pathways	Observed	Expected	Fold Enrichment	Raw P Value	FDR
Signaling by high-kinase activity BRAF mutants	4	0.13	31.37	1.20E-05	2.71E-03
Oncogenic MAPK signaling	4	0.3	13.39	2.71E-04	2.42E-02
MAP2K and MAPK activation	4	0.14	28.16	1.77E-05	3.40E-03
Signaling by RAF1 mutants	4	0.15	27.45	1.95E-05	3.23E-03
Post-translational protein phosphorylation	10	0.39	25.66	1.39E-11	3.47E-08
Signaling downstream of RAS mutants	4	0.16	24.4	3.00E-05	4.67E-03
Signaling by RAS mutants	4	0.16	24.4	3.00E-05	4.15E-03
Paradoxical activation of RAF signaling by kinase inactive BRAF	4	0.16	24.4	3.00E-05	4.39E-03
Signaling by moderate kinase activity BRAF mutants	4	0.16	24.4	3.00E-05	3.93E-03
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	10	0.45	22.14	5.43E-11	6.76E-08
Platelet degranulation	9	0.46	19.45	1.54E-09	1.28E-06
Response to elevated platelet cytosolic Ca ²⁺	9	0.48	18.72	2.12E-09	1.32E-06
Signaling by BRAF and RAF1 fusions	4	0.23	17.16	1.09E-04	1.14E-02

Table 9. (continued)

Reactome Pathways	Observed	Expected	Fold Enrichment	Raw P Value	FDR
Integrin cell surface interactions	5	0.31	16.34	1.80E-05	3.21E-03
Binding and Uptake of Ligands by Scavenger Receptors	4	0.37	10.77	6.01E-04	4.83E-02

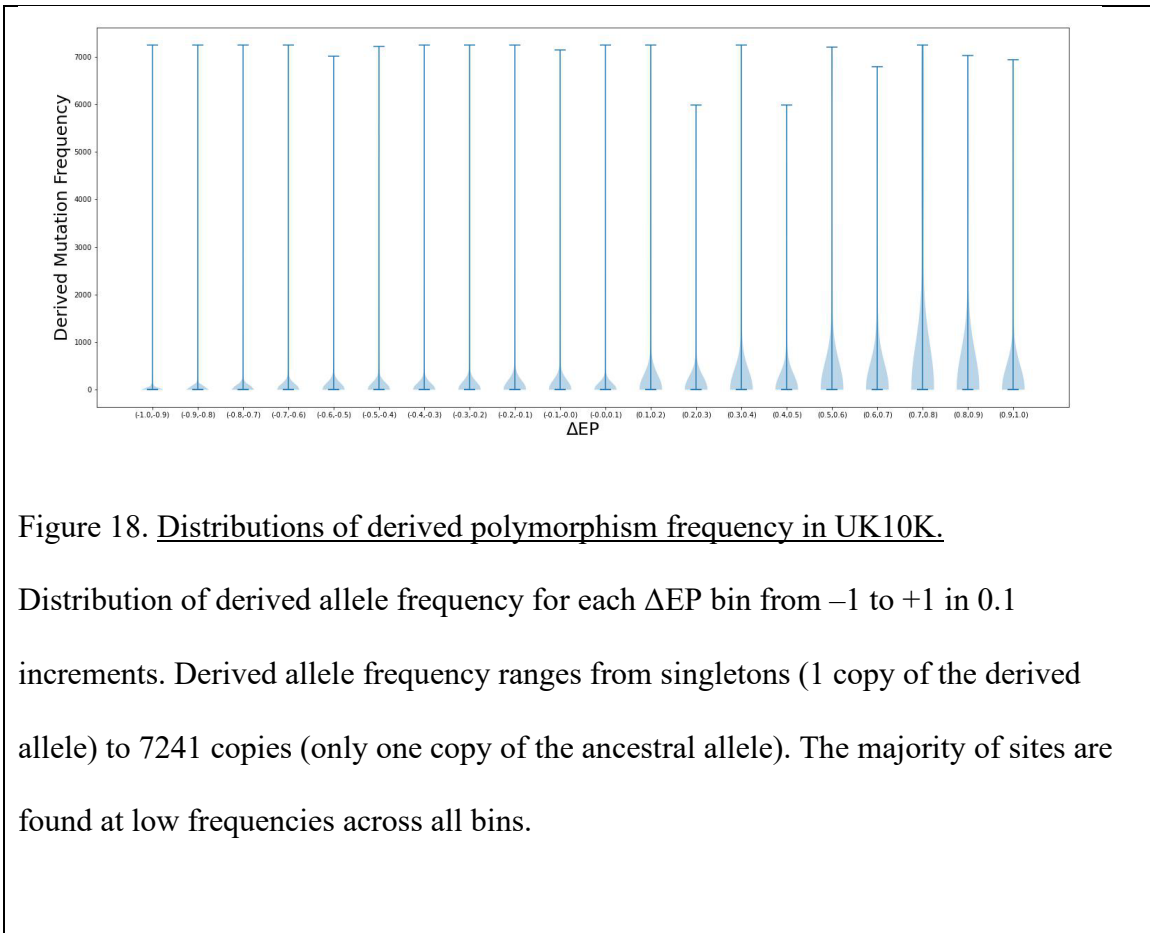
Table 9. F_{ST} Values across ΔEP spectrum of values.

ΔEP	Africa vs Eurasia	Europe vs Asia	Great Britain vs Italy
-0.95	0.5615	0.5254	0.4913
-0.85	0.5416	0.5249	0.5097
-0.75	0.5388	0.5219	0.5112
-0.65	0.521	0.5344	0.5106
-0.55	0.5209	0.5211	0.5003
-0.45	0.514	0.511	0.4985
-0.35	0.4986	0.5182	0.5173
-0.25	0.5178	0.4934	0.4936
-0.15	0.5085	0.5328	0.4928
-0.05	0.511	0.5223	0.4485
0.05	0.5076	0.5423	0.5211
0.15	0.5052	0.5544	0.4942
0.25	0.479	0.5825	0.5861
0.35	0.4844	0.532	0.5286
0.45	0.4902	0.547	0.4228
0.55	0.5889	0.5955	0.4614
0.65	0.5295	0.4961	0.4488
0.75	0.5159	0.6301	0.5388
0.85	0.4205	0.4204	0.4972
0.95	0.5693	0.4416	0.5675

Mean F_{ST} rank value for UK10K SNPs in ΔEP bins for three population contrasts. Values are for SNPs that are in the UK10K sample and occur with at least 10 derived alleles in the pooled populations of the contrast. For each ΔEP SNP the observed F_{ST} was ranked against that for control alleles of the same derived allele frequency.

Table 10. Statistical power for detecting excess heterozygosity.

Selection Coefficient	Probability of rejecting null hypothesis at false positive rate of 0.05
0.0	0.0485
0.0001	0.0545
0.0002	0.0560
0.0005	0.0615
0.001	0.0830
0.002	0.136
0.005	0.417
0.01	0.870
0.02	1.0
0.05	1.0
0.1	1.0



APPENDIX B

SUPPLEMENTAL MATERIAL FOR CHAPTER 3

Supplemental Methodology. Time of Coalescence Estimator Modification

Time of coalescence, t_c , is calculated as a maximum likelihood based on the rates of recombination and mutation in combination with the maximum shared haplotype, msh , as determined by the shared haplotypes tract surrounding a focal allele in a population of genomes. This is described in Platt et al. 2019. However, in the case of missing data due to only sequencing the coding regions, the msh may start or end due to mutation or recombinational events in the intronic or intergenic regions. In the case of exome sequence data, the dataset is missing mutations between the exons. Since this data is missing, the end of the shared haplotype would not be identified until the next closest exon. To attempt to mitigate this issue, we have implemented a modified version of the time of coalescence estimator. In this modified version, since the haplotype will always end due in an exon due to the sequencing data type, instead we estimate that this shared haplotype ended somewhere in the non-sequenced region. To do this, the maximum shared haplotype is measured as a random distance between the exon where the haplotype was found to end and the next closest exon to the focal mutation. With modified msh value assigned as this random distance, the maximum likelihood estimator of t_c is calculated as normal (50).

Table 11. Constant population simulation parameters

Parameter	Neutral Model	Selection Model
Engine	msprime Version: 1.2.0 (132)	SLiM Version 4.1 (134)
Model	PiecewiseConstant Piecewise constant size population model over multiple epochs.	
Seed	14	
Sample Size (genomes)	Varies depending on analysis: 50, 500, 3621, 5000	5000
Contig Ploidy	2	
Sample Time (generation)	0	
Contig Length (bp)	50818468.0 (10)	
Mean Recombination Rate (per bp per gen)	2.1057233894035443e-08 (77)	
Mean Mutation Rate (per bp per gen)	1.29e-08 (182)	
Genetic Map	HapMapII_GRCh38 (77)	
Generation Time (years)	30 (183)	
Ancestral Population Size (individuals)	10000 (184)	
Distribution of fitness effects (DFE)		Gamma_K17 (135)

The parameters for a simulation of a constant population model with either neutral mutations or negatively selected mutations. Citations listed for parameters.

Documentation can be found at: <https://popsim-consortium.github.io/stdpopsim-docs/stable/api.html#stdpopsim.PiecewiseConstantSize>

Table 12. Complex population simulation parameters

Parameter	Neutral Model	Selection Model
Engine	msprime Version: 1.2.0 (132)	SLiM Version 4.1 (134)
Model	OutOfAfrica_3G09 Three population out-of-Africa. (131)	
Seed	14	
Sample Size (genomes)	Varies depending on analysis: YRI:0; CEU: {50, 500, 3621, 5000}; CHB:0	5000
Contig Ploidy	2	
Sample Time (generation)	0	
Contig Length (bp)	50818468.0 (10)	
Mean Recombination Rate (per bp per gen)	2.1057233894035443e-08 (77)	
Mean Mutation Rate (per bp per gen)	2.35e-08	
Genetic Map	HapMapII_GRCh38 (77)	
Generation Time (years)	25	
Ancestral Population Size (individuals)	7300	
Distribution of fitness effects (DFE)		Gamma_K17 (135)

The parameters for simulation of an out-of-Africa population model based on Gutenkunst et al. 2009. Parameters indicated for either neutral mutations or negatively selected mutations. Documentation for this simulation can be found at:

https://popsim-consortium.github.io/stdpopsim-docs/stable/catalog.html#sec_catalog_homsap_models_outofafrica_3g09

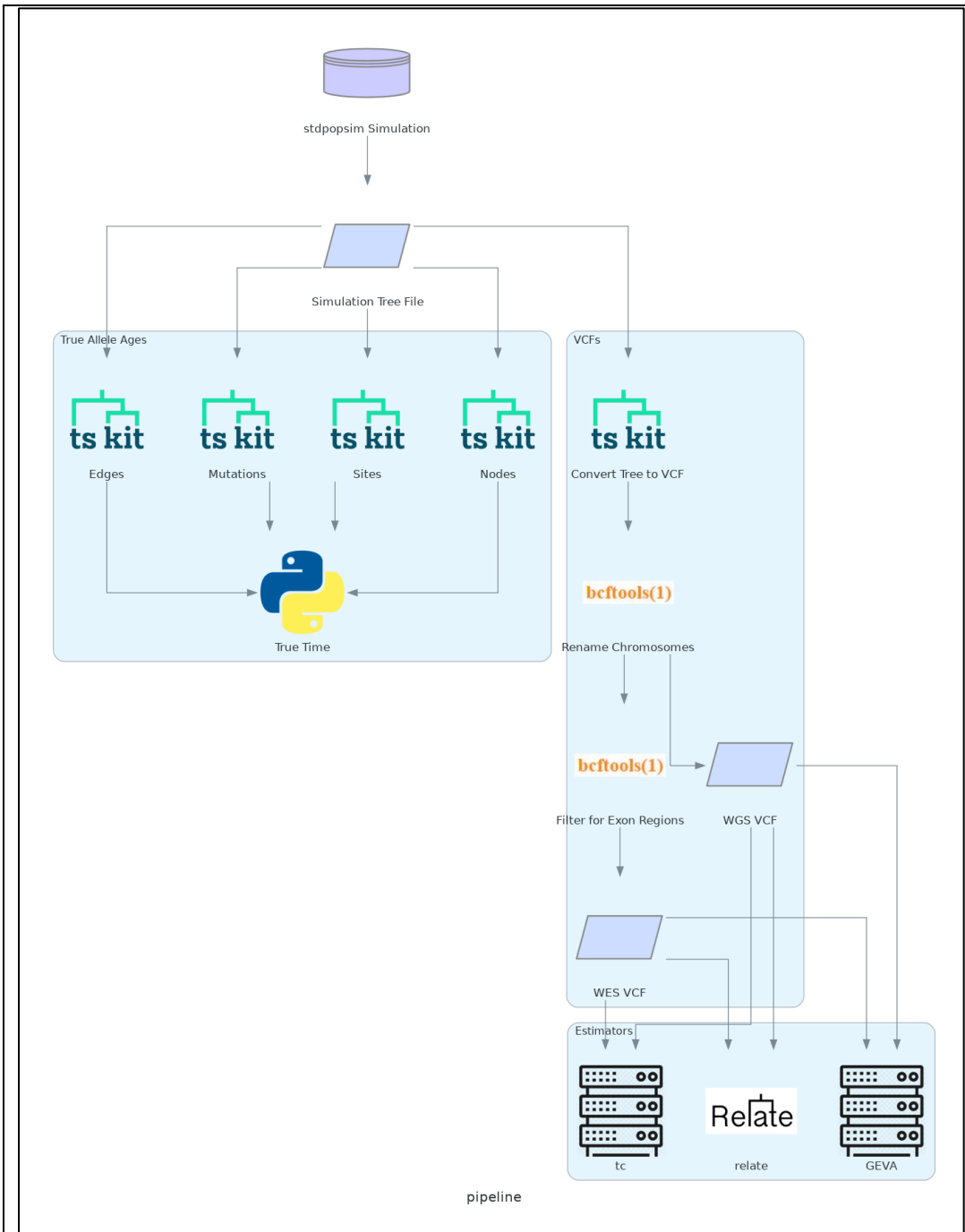


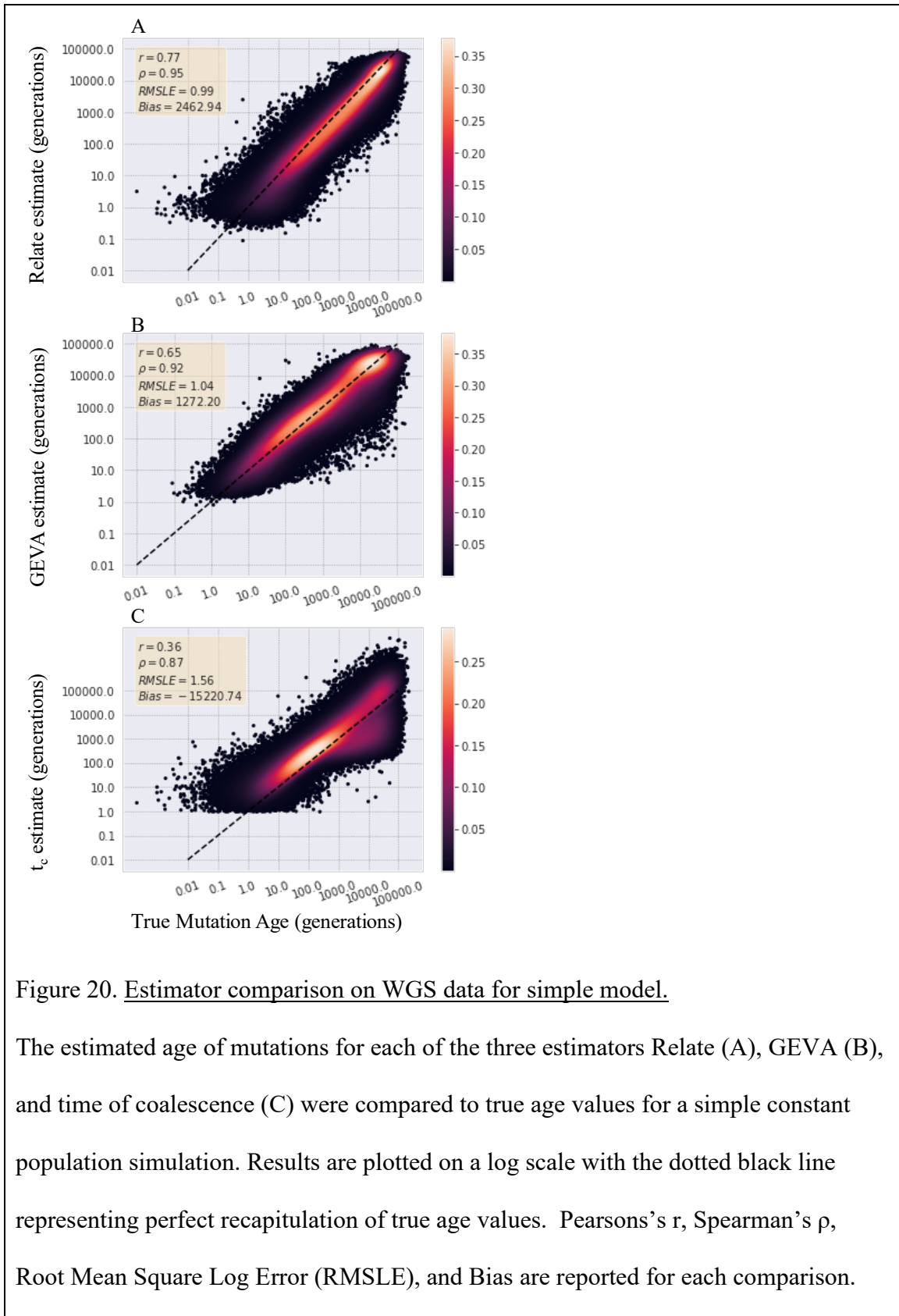
Figure 19. Depiction of simulation pipeline.

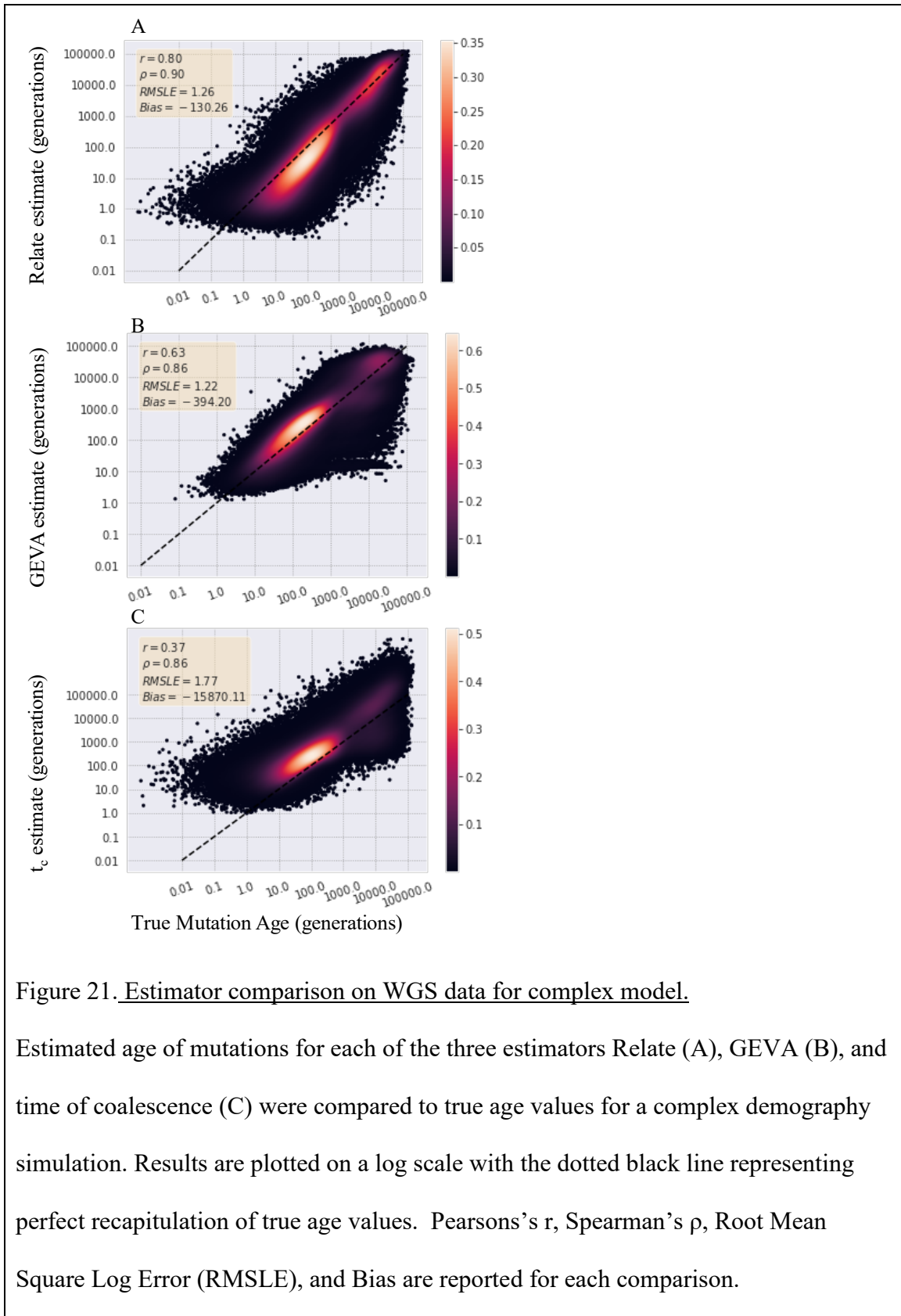
Pipeline for process of simulating dataset of mutations, extracting true age of mutations, converting simulations to VCF format and filtering for sites in the exon regions, and estimating age of sites using three estimators: GEVA, relate, tc.

Table 13. Summary statistics from three estimators for simple and complex model.

Estimators	Simple Model				Complex Model			
	WES		WGS		WES		WGS	
	Time (hrs)	Num SNPs	Time (hrs)	Num SNPs	Time (hrs)	Num SNPs	Time (hrs)	Num SNPs
Relate (12 threads)	8.448 (0.704)	3981	140.616 (11.718)	171107	10.62 (0.885)	9400	162.204 (13.517)	407286
GEVA	18.62	2672	2110.3	120499	22.3	3254	1530.8	169149
t_c	65.45	3965	3023.5	171876	82.52	9314	4474.8	406388

Each of the three allele age estimators (Relate, GEVA, t_c) were run on neutral simulations of the simple constant population size model and the out-of-Africa expansion population size model for both the whole exome sequences (WES) and the whole genome sequences (WGS). The time each estimator took to return estimates was recorded in hours and the number of single nucleotide polymorphisms (SNPs) estimated for each dataset and method were also recorded. For Relate the actual time taken to run on 12 threads is recorded in the parentheses for each simulation model. Each estimator filtered some subset of the sites due to filtering mechanisms of the method as described in the methods.





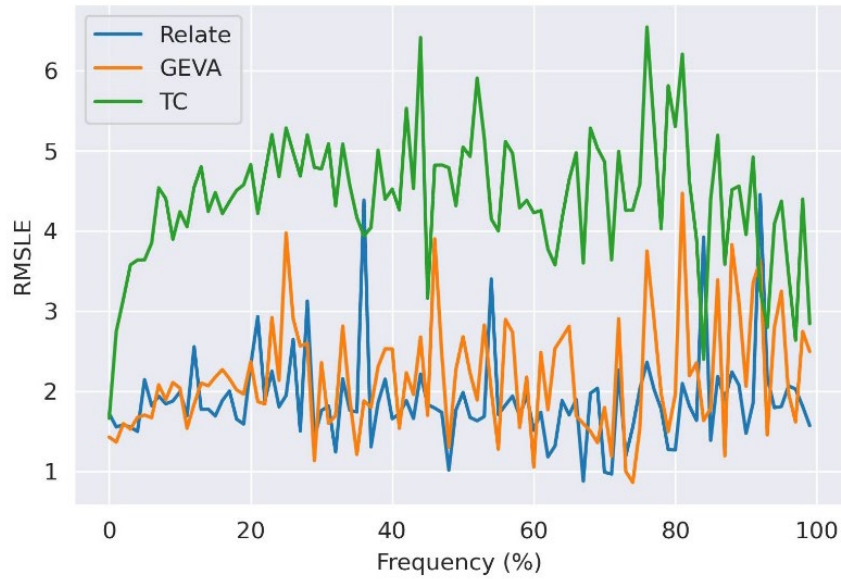
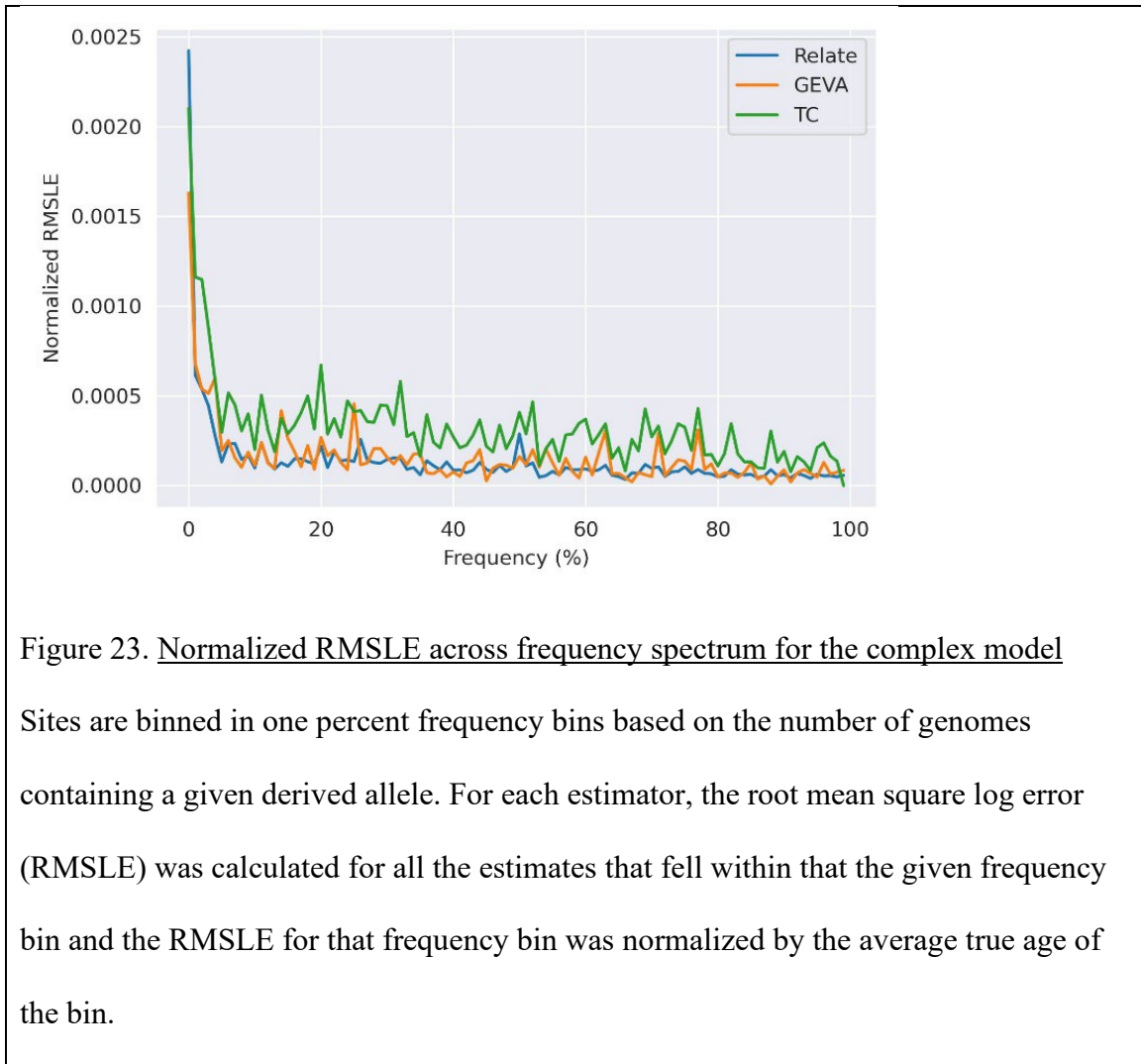
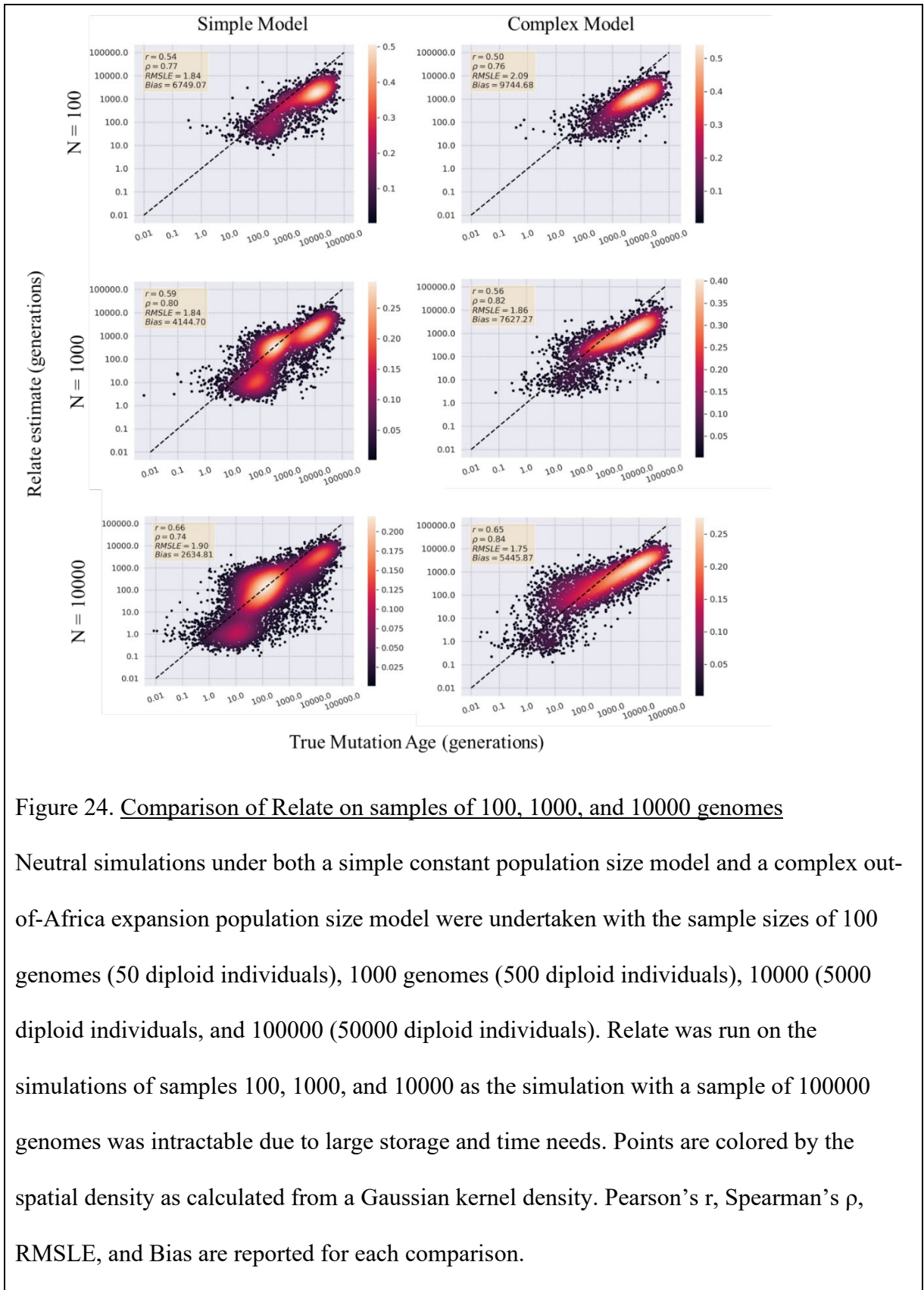


Figure 22. RMSLE across entire frequency spectrum of mutations.

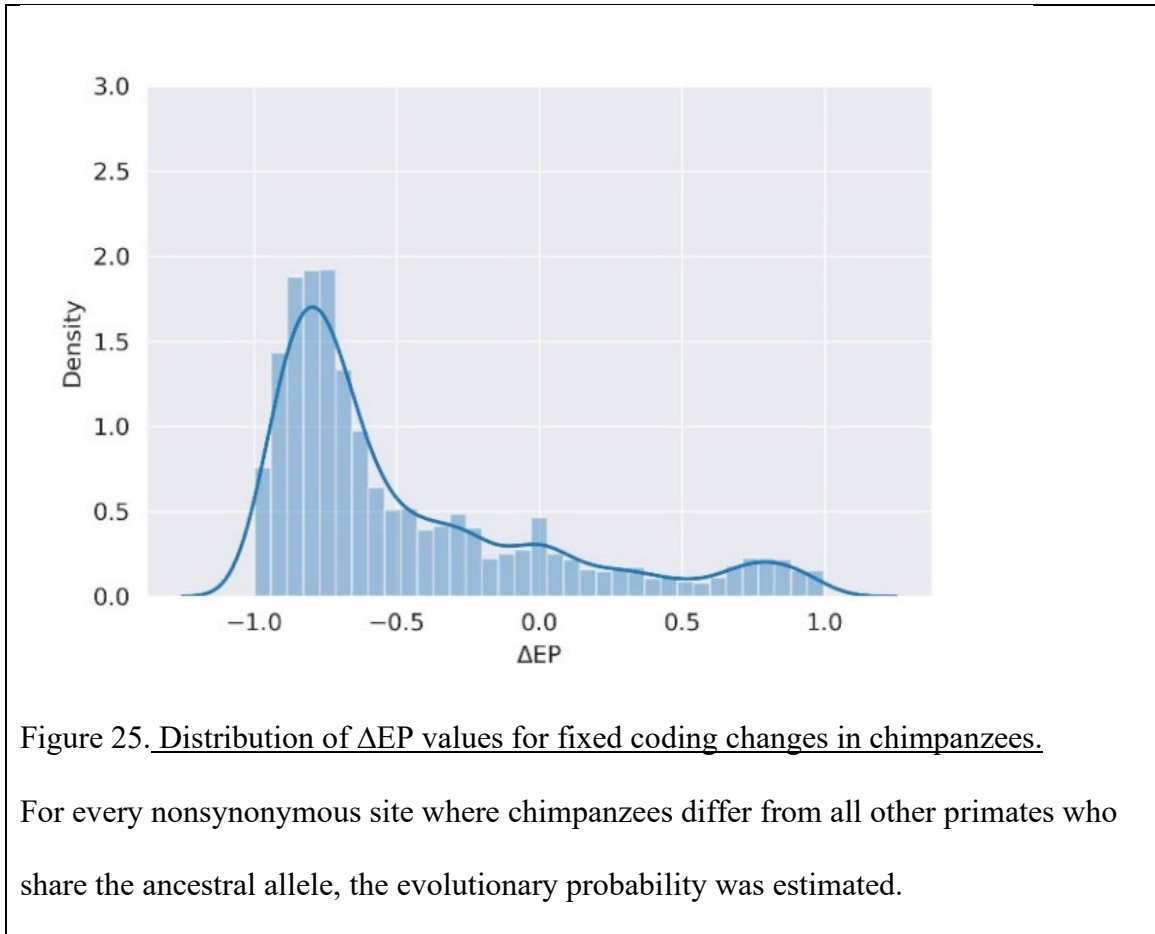
Sites are binned in one percent frequency bins based on the number of genomes containing a given derived allele. For each estimator, the root mean square log error (RMSLE) was calculated for all the estimates that fell within that the given frequency bin.





APPENDIX C

SUPPLEMENTAL MATERIAL FOR CHAPTER 4



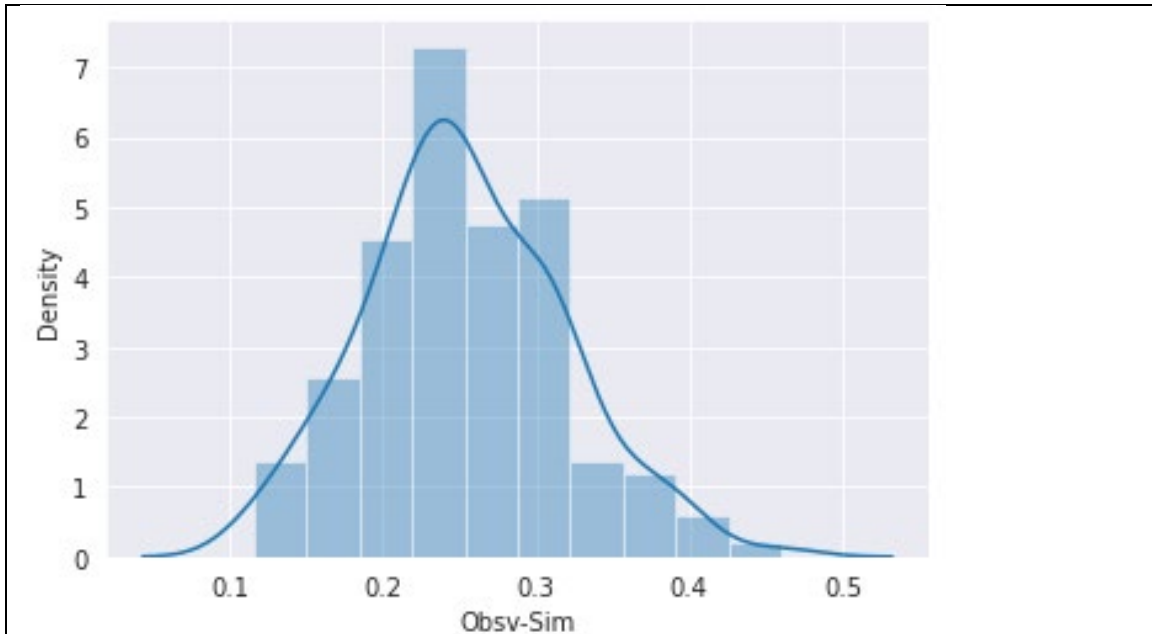


Figure 26. Distribution of difference in observed and simulated ΔEP values for loci where the ancestral and derived amino acids have a similar EP.

All loci where the ancestral amino acid and substituted derived amino acid have a similar magnitude EP value above 0.2, a random mutation in the ancestral amino acid codon is simulated and the EP values from every possible single mutation is calculated. The difference between the observed ΔEP value for a site and the average simulated value for that locus based on the ancestral amino acid is calculated.

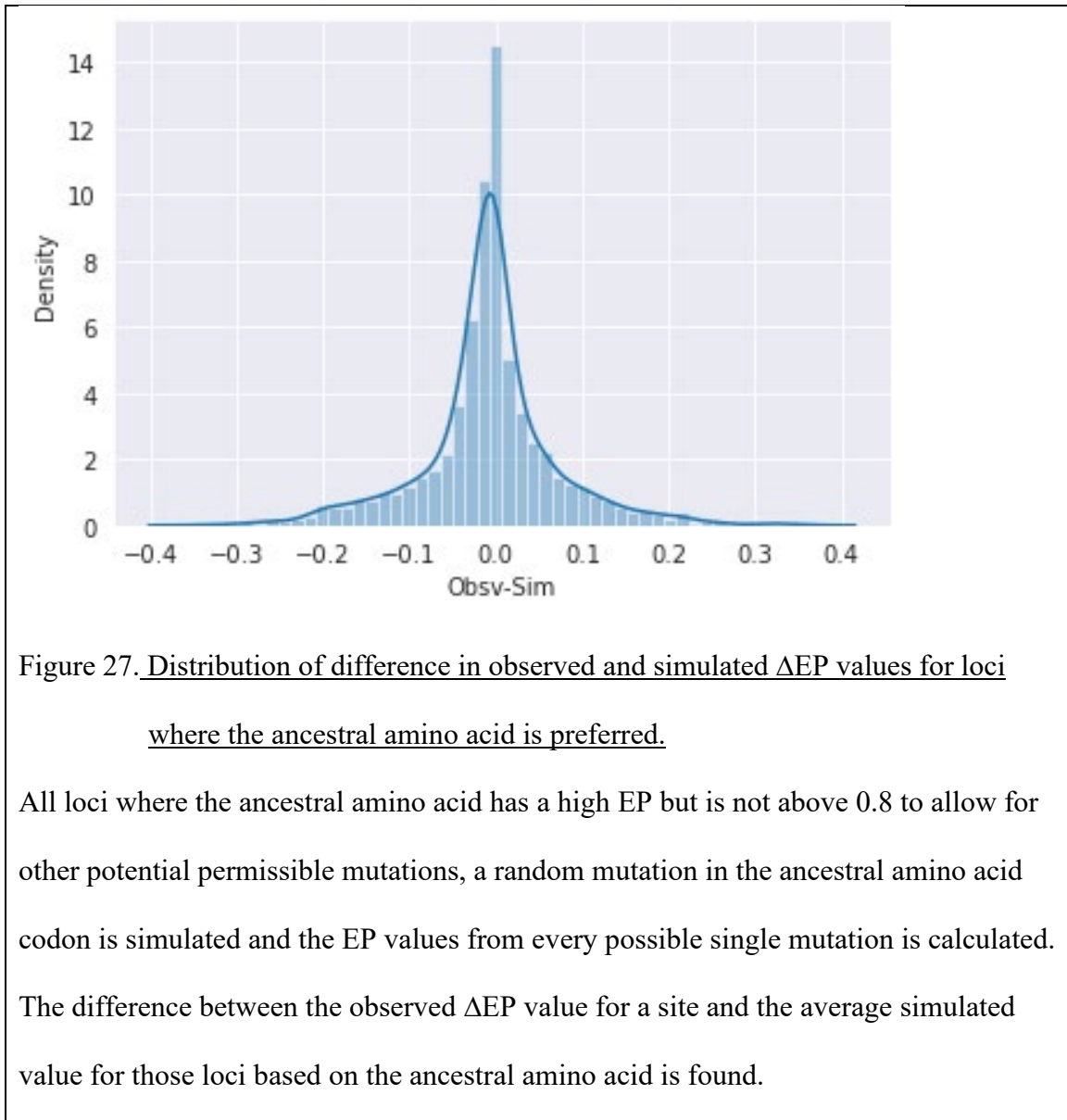


Table 14. Selected genes identified by aBSREL.

REFseq ID	Sites	# of Seq	BUSTE D-E filtered sites	Fraction of sites filtered	Error p-value	Chr	Gene Name
NM_001321103	1259	23	46	0.001589	0.5	3	SLC4A7
NM_002440	937	22	94	0.00456	0.5	1	MSH4
NM_001352754	818	23	83	0.004412	0.320688	2	ARMC9
NM_002644	765	23	0	0	0.5	1	PIGR
NM_001142548	747	23	47	0.002736	0.325504	1	RAD54L
NM_001038705	731	23	101	0.006007	0.076402	3	GPR149
NM_001178007	710	23	78	0.004776	0.5	4	BBS12
NM_144698	1001	14	39	0.002783	0.072364	1	ANKRD35
NM_032180	660	19	44	0.003509	0.5	2	FAM161A
NM_173545	512	23	40	0.003397	0.478966	2	APLF
NM_001319051	618	19	20	0.001703	0.055296	3	GALNT15
NM_001297608	506	23	60	0.005156	0.178646	4	SPATA18
NM_032009	823	13	0	0	0.493819	5	PCDHGA2
NM_020407	458	23	47	0.004462	0.209786	1	RHBG
NM_001199797	434	22	25	0.002618	0.169249	1	PTPN7
NM_001496	411	22	18	0.001991	0.102322	5	GFRA3
NM_198406	344	24	62	0.00751	0.111432	1	PAQR6
NM_052931	331	23	30	0.003941	0.256328	1	SLAMF6
NM_001040260	766	10	31	0.004047	0.491673	4	DCLK2
NM_001166664	273	22	0	0	0.390138	1	CD244
NM_020485	417	14	25	0.004282	0.5	1	RHCE
NM_001207039	262	19	21	0.004219	0.175376	6	ETV7
NM_001329564	248	20	10	0.002016	0.5	5	ZBED3

Table 14. (continued)

REFseq ID	Sites	# of Seq	BUSTED-E filtered sites	Fraction of sites filtered	Error p-value	Chr	Gene Name
NM_001039211	413	12	17	0.00343	0.5	1	ATAD3C
NM_001105519	201	23	0	0	0.5	2	FAM166C
NM_001282870	463	10	0	0	0.5	1	RHD
NM_018659	136	22	5	0.001671	0.221835	4	CYTL1
NM_001330180	253	11	0	0	0.5	2	MTERF4
NM_002620	105	23	0	0	0.5	4	PF4V1
NM_001243328	213	10	0	0	0.5	6	RAET1E
NM_001320010	113	14	15	0.009482	0.482804	1	PYDC5

Listed genes were identified as being positively selected on the hominin branch utilizing aBSREL and BUSTED-E filtering pipeline. Sites are identified as being positively selected based on ω calculated as dN/dS.

Table 15. GSEA results for positively selected genes from aBSREL.

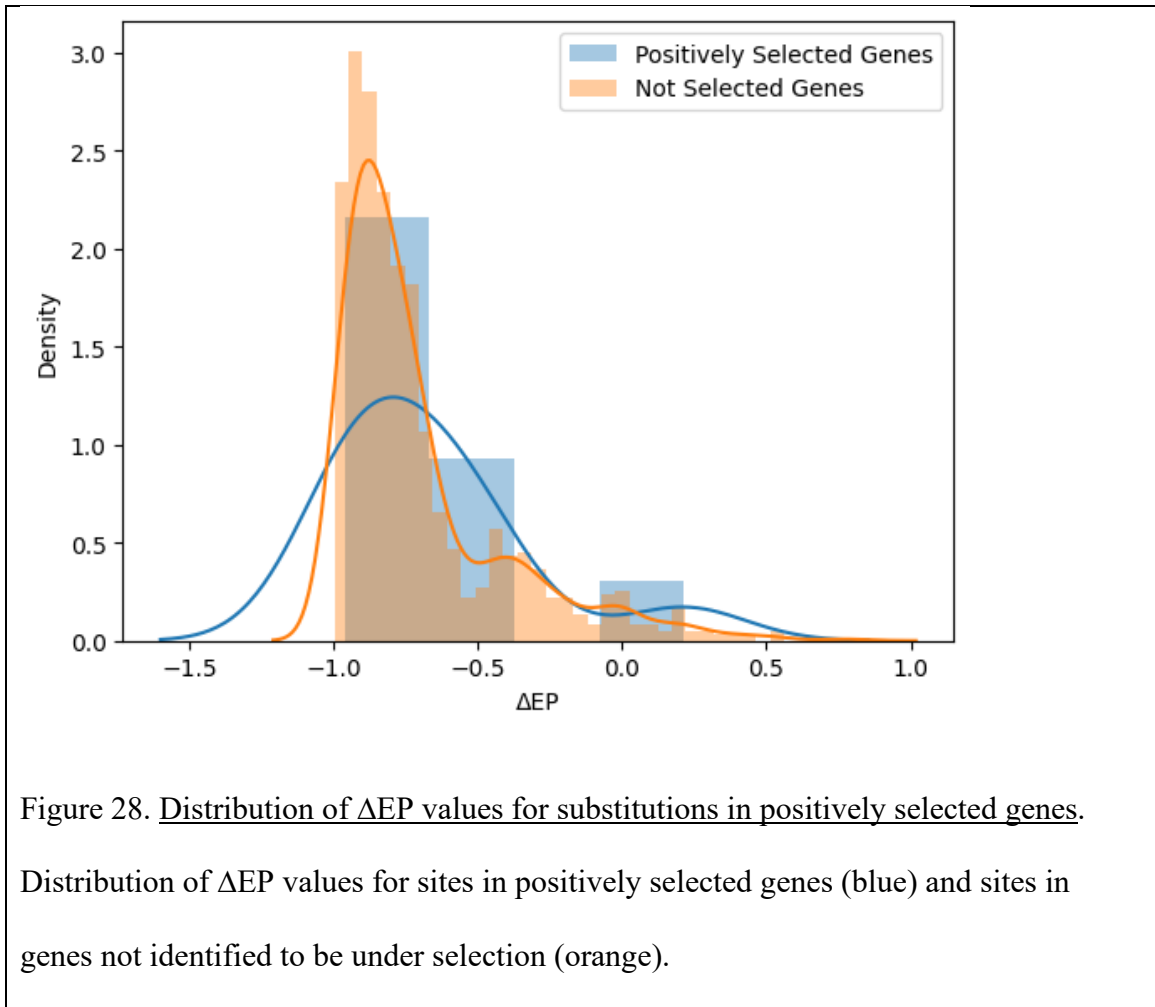
Gene Set Name	# Genes in Gene Set (K)	# Genes in Overlap (k)	k/K	p-value	FDR q-value
Ammonium Homeostasis (GOBP)	6	3	0.5	6.91E-09	2.39E-04
Ammonium Transmembrane Transport (GOBP)	12	3	0.25	7.58E-08	8.73E-04
Ammonium Transmembrane Transporter Activity (GOMF)	12	3	0.25	7.58E-08	8.73E-04

Genes identified as positively selected from aBSREL were analyzed for gene set enrichment for the hallmark set of genes. Three gene sets were identified with significant overlap between the sets and the 31 positively selected genes. False discovery rate and p-value for the enrichment analysis are included.

Table 16. Human substituted sites identified in positively selected genes.

Position	MH Codon	MH AA	MH EP	P Codon	P AA	P EP	Chr	Gene Name	ΔEP
46739041	TAT	Y	0.5537	CAT	H	0.3384	1	RAD54L	0.2153
123663637	TAT	Y	0.0052	TCT	S	0.4868	4	BBS12	-0.4816
123663677	AAC	N	0.0087	AAG	K	0.7706	4	BBS12	-0.7619
123664588	ATC	I	0.0073	ACC	T	0.6954	4	BBS12	-0.6881
123664700	CAT	H	0.0087	CAA	Q	0.8773	4	BBS12	-0.8687
76349404	GTT	V	0.0073	ATT	I	0.9691	1	MSH4	-0.9618
207106446	AAA	F	0.0052	AAC	V	0.8346	1	PIGR	-0.8294
62067198	GCT	S	0.1732	GTT	N	0.6839	2	FAM161A	-0.5107
62067465	TCT	R	0.0116	TGT	T	0.5507	2	FAM161A	-0.5391
68740752	ACT	T	0.0021	GCT	A	0.9595	2	APLF	-0.9574
68753330	GAG	E	0.0021	AAG	K	0.9494	2	APLF	-0.9473

Substitutions on the hominin branch found in any of the 31 positively selected genes are identified and reported here with the position of the derived mutation, the evolutionary probability values, the gene name, and the human and primate alleles, codons, and amino acids. MH refers to modern human for allele (BP), codon, and amino acid (AA). P refers to the primate allele (BP), codon, and amino acid (AA).



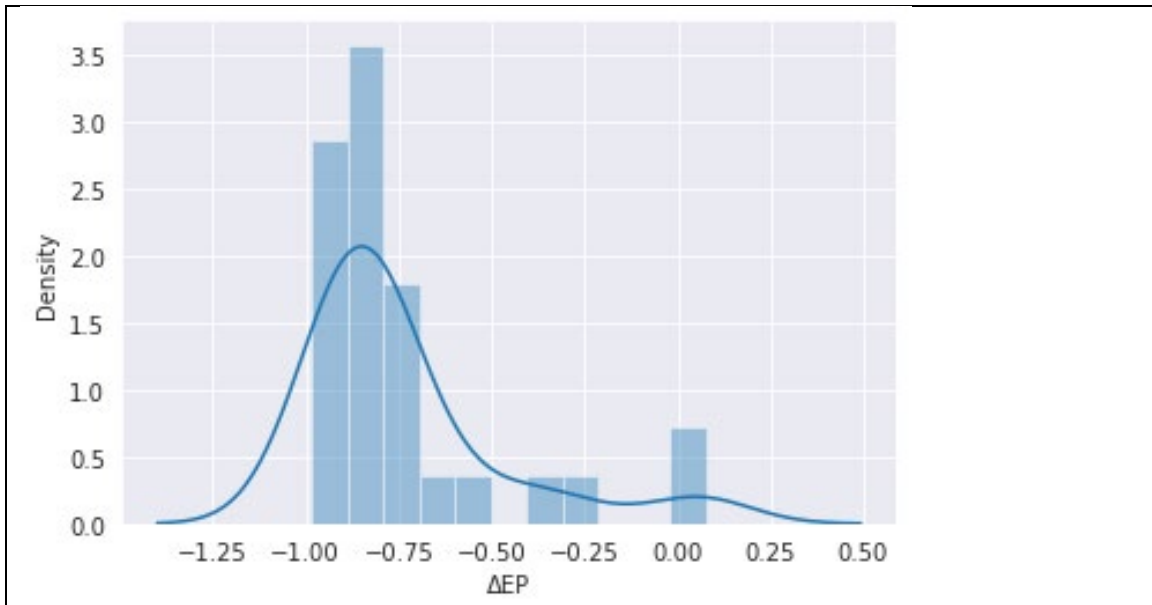


Figure 29. Distribution of ΔEP values for fixed coding changes in archaic humans. Sites where archaic humans have the ancestral amino acid (shared with non-human primates) at the loci whereas modern humans have the fixed, derived amino acid.