Universitat Autònoma de Barcelona

Cataloguing the shape and strength of positive selection on 1000 Genomes Project data

Jesús Murga Moreno

Cataloguing the shape and strength of positive selection on 1000 Genomes Project data

Directors

Jesús Murga Moreno

Sònia Casillas Viladerrams

Antonio Barbadilla Prados

A Murga, a Pepillo el Habarero, a María la Cuarenta, al hambre que pasaron.

Kyrie, gloria, hosanna y eleison
son todas las palabras que recuerdo de misa,
para mis tíos en paro,
para los adictos al diazepam
que desayunan y trasnochan frente a la farmacia,
para mis primos, a los que no perdono,
al dolor que se esconde,
el que hubiera escondido,
el que hube escondido,
del que recuerdo el precio en pesetas.

Para los colmillos de mi abuela,
al nicho blanco que nadie encala en Noviembre,
las lentes bifocales, el rostro serio,
que mi memoria ya no acierta,
ni atisba, ni parece llorar.

Para el vino y los alcohólicos,
desde la lengua muerta de mi abuelo,
que ya no juzga,
que ya no llora,
que tampoco grita por las noches.

Para las que heredan viudas en vida sillas de mimbre,
para los huesos rotos de las cojas,
la mente exhausta de sus maridos, el rostro deshecho.

Kyrie,
kyrie, gloria, hosanna y eleison,
eleison para ellos.

Gracias a todos los que me abrieron la puerta de su casa estos diez últimos años. Gracias a todos los que con amor y paciencia habéis sorportado mis miedos y frustaciones estos últimos meses.

Publications

The following publications have being published during this thesis or are currently in preparation:

- Murga-Moreno, J., Coronado-Zamora, M., et al. PopHumanScan: the online catalog of human genome adaptation. *Nucleic Acids Research*, 47(D1):D1080–D1089, 2019a. ISSN 0305-1048. doi: 10.1093/nar/gky959.
- Murga-Moreno, J., Coronado-Zamora, et al. iMKT: the integrative McDonald and Kreitman test. *Nucleic Acids Research*, 47(W1):W283–W288, 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz372.
- Murga-Moreno, J., et al. Imputed McDonald and Kreitman test: a straightforward correction that increases significantly the power of gene-by-gene MKT (in preparation)
- Murga-Moreno, J., et al. Efficient inference of adaptation rate and strength (in preparation)

I contributed to the following papers but they do not form part of this thesis

- Di, C., Murga Moreno, J., et al. Decreased recent adaptation at human mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants. *eLife*, 10:e69026, 2021. ISSN 2050-084X. doi: 10.7554/eLife.69026. Publisher: eLife Sciences Publications, Ltd.
- Colomer-Vilaplana, A., Murga-Moreno, J., et al. PopHumanVar: an interactive application for the functional characterization and prioritization of adaptive genomic variants in humans. *Nucleic Acids Research*, page gkab925, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab925
- Kapun, M., Nunez, J.C.B., Bogaerts-Márquez, M., Murga-Moreno, J., Margot, P. et al. Drosophila Evolution over Space and Time (DEST): A New Population Genomics Resource. *Molecular Biology and Evolution*, (msab259), 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab259.

Contents

List of Figures

List of Tables

Acronyms

1000GP 1000 Genomes Project

MKT McDonald and Kreitman Test

impMKT imputed McDonald and Kreitman Test

SFS Site Frequency Spectrum

ABC Approximate Bayesian Computation

NGS Next Generation Sequencing

DGN Drosophila Genome Nexus

DGRP Drosophila Genetic Reference Panel

RAL North American Raleigh

SNP Single Nucleotide Polymorphism

LD Linkage Disequilibrium

iHS integrated Haplotype Score

XP-EHH Cross-Population Extended Haplotype Homozygosity

ag1000G *Anopheles gambiae* 1000 Genomes

SGDP Simon Genome Diversity Project

DPGP Drosophila Population Genomics Project

HGDP Human Genome Diversity Project

YRI Yoruba individuals from Nigeria

CEU European ancestry individuals from Utah

CHB Han Chinese individuals from Beijing

JPT Japanese individuals from Tokyo

AFR Africa

EAS East Asia

EUR Europe

SAS South Asia

AMR America

OoA Out-of-Africa

SMC Sequential Markov Chain

ARG Ancestral Recombination Graph

TMRCA Time to the common ancestor

MCMC Markov Chain Monte Carlo

HMM Hidden Markov Model

VCF Variant Calling Format

DFE Distribution of Fitness Effect

PRF Poisson Random Field

SDM Slightly Deleterious Mutations

BGS Background selection

HRi Hill-Robertson interference

LSBL Locus-Specific Branch Length

EHH Extended Haplotype Homozygosity

VIP Viral Interaction Protein

d Genetic distance between two orthologous sequences

d_N/d_S Rate of non-synonymous substitution relative to the rate of synonymous substitution

d_N Rate of non-synonymous substitutions per generation and site

d_S Rate of synonymous substitutions per generation and site

D_N Observed number of non-synonymous substitutions

D_S Observed number of non-synonymous substitutions

P_N Observed number of non-synonymous polymorphisms

P_S Observed number of synonymous polymorphisms

NI Neutrality Index

DoS Direction of Selection

ML Maximum Likelihood

fwwMKT Fay, Wycoff and Wu MKT

eMKT extended MKT

aMKT asymptotic MKT

uSFS unfolded Site Frequency Spectrum

Abstract

Since the split with chimpanzees, and especially since the migrations that led humans to colonize almost every place on Earth, our species has faced frequent environmental and social changes that have shaped the variation patterns of our genomes through the action of natural selection. These selection pressures left signatures in the landscape of genetic variation that can be identified in today's genomes. Numerous statistical methods have been proposed to analyze genomic data, allowing the detection and quantification of molecular adaptation at different temporal scales and providing essential insights into past and recent human evolutionary history. The availability of the most comprehensive worldwide nucleotide variation dataset so far, the 1000 Genomes Project, provides a resource to test population genetics hypotheses and eventually pinpoint targets of positive selection from the background evolutionary dynamics of genetic variation.

This thesis aims to trace the shape and strength of positive selection on 1000GP data, mainly focusing on population genetics methods that try to disentangle the adaptive selection contributing to between species and between populations diversification. For this purpose, the thesis develops statistical and bioinformatics approaches to solve issues of major importance in population genomics.

We performed a genome-wide scan of selection on the 1000GP data by surveying distinctive signatures of genomic variation left by selective events and created an online catalog of all candidates to facilitate their validation and thorough analysis. The outlier approach applied here detects sweeps at different historical ages and evidence of recurrent selection in the human lineage since the split between our species and chimpanzees. We provide new candidates and bring together studies that locate repeatedly the same target genes independently of data and methodologies. These results have been made available in a collaborative, online database, compiling and annotating adaptation events along with the human evolutionary history, which aims to be expanded in future studies.

In addition, we reviewed the McDonald and Kreitman test (MKT), one of the most powerful and robust methods to detect the action of recurrent natural selection at the DNA level, both at the gene and the genome level. First, although several modifications of the original MKT have been proposed to account for the potential biases underlying the MKT, most of these extensions mainly deal with the presence of slightly deleterious mutations (SDM). While more and more genome-wide analyses have been carried out, the simple G-test of the original MKT has become almost deprecated. For that reason, we present the imputed MKT (impMKT), an MKT extension that significantly improves gene-by-gene analyses maximizing the information to test the recurrent positive at the gene level. Second, in addition to SDM, demography, linked selection and weak adaptation have been repeatedly postulated as the possible cause of the much lower proportion of adaptive mutations measured by the MKT in humans and primates. Taking advantage of genome-wide information, we also develop an extension of the ABC-MK method. Our approach is a simpler and much more computationally efficient ABC-based inference procedure than the previous one, which accounts for the DFE of deleterious and beneficial alleles and incomplete recombination between selected genomic elements. We describe the inference procedure, assess its performance and robustness, and finally show that it is reasonably robust to non-equilibrium events or different configurations of adaptive selection. In addition, we present evidence for a substantial effect of RNA-viruses on human adaptation rates, providing new insight into the human drivers of adaptation.

Finally, in addition to our collaborative database and computationally efficient methods, we developed a web server that facilitates MKT analyses in the human lineage and custom analyses for humans and other species with population genomics data.

Resumen

Desde que los humanos y chimpancés se separaron evolutivamente, y posteriormente a través de las migraciones, nuestra especie se ha enfrentado a numerosos cambios ambientales y sociales. Estas presiones han moldeado los patrones de variación de nuestros genomas, dejando características huellas moleculares a lo largo del genoma que pueden identificarse mediante numerosos métodos estadísticos. Dichos métodos han permitido detectar y cuantificar la adaptación molecular a diferentes escalas temporales, proporcionando información esencial sobre la historia evolutiva pasada y reciente de nuestra especie. La disponibilidad del conjunto de datos de variación nucleotídica más completo hasta la fecha, el Proyecto 1000 Genomas, permite probar hipótesis de la genética de poblaciones en base a los patrones de variación y finalmente identificar caracteres sujetos a selección positiva.

Esta tesis tiene como objetivo inferir la forma y la fuerza de la selección positiva en datos los de 1000GP, centrándose principalmente en métodos estadísticos y bioinformáticos que pueden revelar la selección adaptativa que contribuye a la diversificación entre especies y entre poblaciones de nuestra especie.

Con este propósito, hemos realizado un escaneo de selección de todo el genoma en los datos de 1000GP mediante el análisis de improntas distintivas de variación genómica causados por diferentes sucesos selectivos y creado un catálogo de todas las regiones genómicas candidatas a estar sujetas a la acción de la selección natural, para así facilitar su validación y análisis exhaustivo. La aproximación presentada detecta barridos selectivos en diferentes momentos históricos y evidencias de selección recurrente en el linaje humano desde la división entre nuestra especie y los chimpancés. Proporcionamos nuevos candidatos y reunimos estudios que localizan repetidamente los mismos genes independientemente de los datos y las metodologías. Estos resultados se han puesto a disposición en una base de datos colaborativa, que recopila y anota eventos de adaptación junto con la historia evolutiva humana, la cual pretende ampliarse con estudios futuros.

Además, revisamos la prueba de McDonald y Kreitman (MKT), uno de los métodos más potentes y robustos para detectar la acción de la selección natural recurrente a nivel de ADN, tanto a nivel de gen como de genoma. En primer lugar, aunque se han propuesto varias modificaciones del MKT original para solventar sus posibles sesgos subyacentes, la mayoría de estas extensiones principalmente tratan la presencia de mutaciones levemente perjudiciales (SDM). Si bien se han cada vez se llevan a cabo más y más análisis a escala genómica, el simple G-test propuesto por el MKT original está desuso. Por esa razón, presentamos el imputed MKT (impMKT), una extensión de MKT que mejora significativamente los análisis gen por gen y maximiza la información para cuantificar la selección positiva a nivel génico. En segundo lugar, además de SDM, la demografía, la selección ligada y la adaptación débil se han postulado repetidamente como causantes de la menor proporción de mutaciones adaptativas en humanos y primates. Aprovechando la información genómica, hemos desarrollado una extensión del método ABC-MK. Nuestro enfoque es un procedimiento de inferencia basado en ABC más simple y eficiente que el anterior, modelando la DFE de alelos perjudiciales y beneficiosos y la recombinación incompleta entre elementos genómicos. Describimos el procedimiento de la inferencia, evaluamos su desempeño y robustez, y finalmente mostramos que es razonablemente robusto frente a eventos de no equilibrio o diferentes configuraciones de selección adaptativa. Además, presentamos evidencia de un efecto sustancial de los virus de ARN en las tasas de adaptación humana, proporcionando una nueva visión de los impulsores humanos de la adaptación.

Finalmente, además de nuestra base de datos colaborativa y métodos computacionalmente eficientes, creamos un servidor web que facilita los análisis MKT en el linaje humano y análisis personalizados.

Resum

Des que els humans i els ximpanzés es van separar evolutivament, i posteriorment a través de les migracions, la nostra espècie s'ha enfrontat a nombrosos canvis ambientals i socials. Aquestes pressions han modelat els patrons de variació dels nostres genomes, deixant característiques empremtes moleculars al llarg del genoma que es poden identificar mitjançant nombrosos mètodes estadístics. Aquests mètodes han permès detectar i quantificar l'adaptació molecular a diferents escales temporals, proporcionant informació essencial sobre la història evolutiva passada i recent de la nostra espècie. La disponibilitat del conjunt de dades de variació nucleotídica més complet fins ara, el Projecte 1000 Genomes, permet provar hipòtesis de la genètica de poblacions sobre la base dels patrons de variació i finalment identificar caràcters subjectes a selecció positiva.

Aquesta tesi té com a objectiu inferir la forma i la força de la selecció positiva en les dades de 1000GP. Per fer-ho ens centrem en mètodes estadístics i bioinformàtics que detecten la selecció adaptativa que contribueix a la diversificació entre espècies i entre poblacions.

Amb aquesta finalitat, hem realitzat un cribratge de selecció al llarg de tot el genoma, per totes les poblacions de 1000GP mitjançant l'anàlisi d'impromptus distintives de variació genòmica causades per diferents tipus d'esdeveniments selectius. El mètode emprat detecta arrossegaments selectius per diferents escales temporals i evidències de selecció recurrent al llinatge humà des de la separació evolutiva respecte als ximpanzés. A partir dels resultats, s'ha creat un catàleg que incorpora totes les regions genòmiques candidates a haver estat subjectes a l'acció de la selecció natural, per facilitar així, la seva validació i anàlisi en profunditat. Proporcionem nous candidats i reunim estudis que localitzen repetidament els mateixos gens independentment de les dades i les metodologies. Els resultats s'han posat a disposició en una base de dades col·laboratives en línia, amb l'objectiu de compilar i anotar esdeveniments d'adaptació d'estudis futurs.

Per altra banda, fem una revisió del test de McDonald i Kreitman (MKT), un dels mètodes històrics més potents i robusts per detectar l'acció de la selecció natural recurrent, tant en l'àmbit genètic com genòmic. En primer lloc, tot i la gran quantitat de modificacions proposades que corregeixen els potencials biaixos del test original, la majoria d'aquestes extensions principalment tracten la presència de mutacions lleument perjudicials (SDM). Si bé cada vegada tenim més i més anàlisis a escala genòmica, el simple G-test proposat pel MKT original està en desús. Per tot això, presentem imputed MKT (impMKT), una extensió del MKT que millora l'anàlisi i maximitza la informació que permet quantificar la selecció positiva a nivell genètic. En segon lloc, a més de la presencia de SDM, la recombinació, la demografia, la selecció positiva dèbil o la selecció lligada s'han postulats com a possible causa de la baixa proporció de mutacions adaptatives detectat en humans i primats. Aprofitant la informació de tot el genoma, desenvolupem una extensió del mètode ABC-MK. La nostra proposta és un procediment d'inferència basat en ABC més simple i eficient que l'anterior, modelant la DFE d'al·lels perjudicials i beneficiosos i la recombinació incompleta entre elements genòmics. Descrivim el procediment de la inferència, avaluem el seu rendiment i robustesa, demostrant que és raonablement robust per a esdeveniments demogràfics diversos i en diferents escenaris adaptatius. A més, presentem l'evidència d'un efecte substancial dels virus d'ARN en les taxes d'adaptació humana, proporcionant una nova perspectiva sobre la importància del virus d'ARN com a promotors d'adaptació molecular en humans.

Finalment, a part de la nostra base de dades col·laborativa i mètodes computacionalment eficients, implementem un servidor web que facilita l'anàlisi MKT al llinatge humà i anàlisis personalitzades.

Chapter 1

Introduction

Population genetics describe and interpret the changes in allelic frequencies or genetic structures of natural populations. The intra-population or species-specific changes, also called micro-evolution and macro-evolution, result from the time scale at which evolution is observed as micro-evolutionary changes lead to macro-evolutionary changes in the long term (Dobzhansky, 1982). In the first half of the 20th century, the population genetics principles and forces defining evolutionary changes were defined by the pioneering work of Fisher, Wright, and Haldane (Fisher, 1930; Wright, 1931; Haldane, 1932). These forces, namely natural selection, genetic drift, mutation, recombination, and gene flow, were defined and modeled, becoming the fundamental factors in the later theoretical and empirical developments. These studies widely explored the consequences of selection and randomness on allele frequency trajectories at a first attempt to model variation at natural populations while integrating principles of Mendelian inheritance, making population genetics the theoretical core of Darwin's theory of evolution. The interaction between population genetics and other disciplines such as the experimental evolution of populations, zoology or paleontology, resulted in the so-called Modern Synthesis of evolutionary biology, or Neo-Darwinism (Dobzhansky, 1982). The Modern Synthesis supposed the incorporation of Mendelian laws of inheritance and the gene concept in the original theory of evolution by natural selection of Darwin. Nonetheless, the near absence of actual genetic data constrained population genetics fundamentally to a mathematics development. Half a century ago, the technique of protein gel electrophoresis was applied to get the first estimates of genetic variation at the molecular level, providing nucleotide diversity measures for several loci and inaugurating the subfield of molecular population genetics. From then on, the fundamental forces defining evolution and genetic variation had been extensively explored. During the last decades, the tools, theoretical frameworks, and data that the field has offered to science have been essential for the birth and development of other disciplines. Thus, if not possible to conceive broad concepts such as personalized medicine, genome-wide association studies, or the migratory movements of the ancestral human populations without considering the bases of molecular population genetics. Molecular data in many levels has provided evidence to theorists, which interpreted and redefined concepts (Charlesworth and Charlesworth, 2017) while solving some classical questions and proposing others (Charlesworth, 2010). The explosion of genomic data has provided wide evidence of adaptation to trace phenotypes from genotypes, where the action of natural selection can lead to characteristic footprints on variation patterns. Nonetheless, genomic data have provided wide evidence of natural selection on genomes, notwithstanding the phenotype being the primary target of natural selection. Molecular population genetics has attempted to describe and infer the levels of genetic variation observed in natural populations, trying to model the relative importance of each evolutionary process aforementioned. The knowledge of the processes that intrinsically model these patterns and ultimately their population dynamics holds the answer to the other great question in population genetics: which traits or phenotypes are targets of natural selection. The genetic basis of any phenotype that has a reproductive advantage is modeled by natural selection. Therefore, natural selection contributes to shaping genetic variants in populations that lead to a phenotype. Over time, the process can result in traits that specialize in particular ecological niches and may eventually result in speciation. Nonetheless, resolving this question requires knowledge of the relationship between genotype and phenotype, the genotype-phenotype map, or the gene architecture of traits. Despite the massive information provided by next-generation sequencing and the thorough description of some essential adaptive traits in the human lineage (Kwiatkowski, 2005; Bersaglieri et al., 2004; Sabeti et al., 2007; Genovese et al., 2010; Beleza et al., 2013; Huerta-Sánchez et al., 2014; Schlebusch et al., 2015; Fumagalli et al., 2015; Minster et al., 2016; Mathieson and Mathieson, 2018), the role of each evolutionary force remains unsolved, and new genomic layers has to be incorporated to understand the mechanisms of variation better. The following sections will focus on the three main molecular population genetics milestones: the allozyme, the nucleotide sequence and the current population genomics eras. We will empathize how the long struggle for measurement of genetic variation has been evolving, solving old debates and bringing new ones. Besides, we will review the theoretical side, including the nearly neutral theory of molecular evolution, the role of recombination, and linked selection.

1.1 Molecular population genetics: from the allozyme era to SNPs genealogies

The first molecular dataset allowed the estimation of genetic variation at an average locus in natural populations. This section reviews the three main milestones fostered by the technological innovations to survey molecular genetic variation.

1.1.1 The allyzome era

Lewontin and Hubby (1966) and Harris (1966) provided the first estimates of variation at protein-level natural populations. Both works described electrophoretically detectable variation -or allozymes- which screens for protein migration on an electrophoretic gel. Proteins differing in electrophoretic mobility ultimately result from the existence of variation at the DNA level. Allozyme data were commonly used, measuring the levels of genetic diversity at populations, species, and taxa levels, showing that genetic diversity varies non-randomly and is much higher than expected from two main evolutionary selective scenarios of the time (Lewontin, 1974). The *classical* hypothesis predicted that natural selection mainly purges variation, and therefore most loci were thought to be homozygous. The *balance* hypothesis (Dobzhansky, 1955), which was the prevalence selective view, predicted that natural selection acts by maintaining genetic variation and thus, a large proportion of heterozygous loci. Nonetheless, the electrophoretic technique had significant limitations that made it difficult to reconcile the prevailing theories with the observed data. First, allozyme studies were doubly flawed. Not only was it limited to detecting non-synonymous changes at the DNA level, but such changes were only detectable if the amino acid change affected the mobility of the protein in the gel (Lewontin, 1991). Therefore, allozyme measures only consider a small part of possible mutations at the protein level, producing low-resolution data for understanding evolutionary forces and discriminating between classical and balancing hypotheses. In addition, Barbadilla et al. (1996) conclude that in highly polymorphic loci the commonly observed frequency pattern of electrophoretic variants is purely a consequence of statistical relations and conveys no information about the underlying evolutionary forces.

Despite its limitations, the first results showed that population size is a key parameter in population genetics. For example, the two measures provided by Lewontin and Hubby (1966) and Harris (1966) (H, the average proportion of loci that are heterozygous in an individual; and P, the average proportion of loci that are polymorphic in the population) exhibited higher values in Drosophila than humans (*P* 43% vs. 28% and *H* 12% vs. 7% for each species respectively) (Casillas and Barbadilla, 2017). Such measures were applied between species and taxa, showing that invertebrates tend to be highly polymorphic, whereas mammals and other animal taxa are only about half as variable on average (Nevo et al., 1984). A priori, large populations were expected to accumulate more variation. However, the narrow spectrum in genetic diversity levels was insufficient to explain considerable differences in population size, the so-called Lewontin's Paradox (Lewontin, 1974). Later studies showed population sizes exceeding several orders of magnitude, while genetic diversity levels only varied by a few orders (Buffalo, 2021). From the very beginning, explanations for diversity measured across species and taxa and explanations for Lewontin's Paradox in neutralist vs. selectionist terms were controversial.

1.1.2 The nucleotide sequence era

In the 90's, nucleotide sequencing replaced allozyme data completely. Nonetheless, in the 80's restriction enzyme techniques played an important role as the first mass genotyping technique (Charlesworth, 2010). Although limited to a few known sequences recognized by restriction enzymes, this technique represented a significant advance, because by the first time it allowed the survey of much larger genomic regions. Therefore, new statistics, such as the nucleotide diversity (π) proposed by (Nei and Li, 1979) were defined, and empirical published results regarding restriction enzyme techniques. are the basis of many hypotheses and studies nowadays (Casillas and Barbadilla, 2017). Some important examples are the positive correlation between nucleotide diversity and recombination (Begun and Aquadro, 1992) or the inference of the effective size of the human population through the mean nucleotide diversity per site (Robertson et al., 1983).

Kreitman (1983), for the first time, revealed the entire nucleotide sequence variation present in a gene region. Eleven *Adh* genes were sequenced from 11 chromosomes independently isolated from five natural populations of *D. melanogaster*. The study revealed 43 SNPs, only one being a non-synonymous polymorphism (responsible for the electrophoretic polymorphism LF), while the rest were silent polymorphisms found in the gene's coding and non-coding regions. The results led to the well-known conclusion that most non-synonymous mutations have deleterious effects on fitness that contribute little to within-population variation or divergence compared to silent changes in the DNA sequence (Charlesworth, 2010).

For a long time, the studies focused on specific genomic regions or genes. However, this crucial step facilitated the development of the first databases and statistics to initially characterize genetic diversity (Casillas et al., 2005; Casillas and Barbadilla, 2006). Despite the advance, genetic diversity studies based on a sample of genes or DNA regions may provide a biased view of genome-wide measurements. It was not until the advent of the Next Generation Sequencing (NGS) that studies have been expanded at the genomic level and were no longer limited to a set of model organisms.

1.1.3 The population genomics sequence era

In the last decade, the development of massive sequencing techniques has led to the current population genomic era. Thanks to NGS improvement and cheaper techniques, we have complete genomes for multiple species, which facilitated comparisons at the population and the phylogenetic levels. Furthermore, the errors associated with the sequencing of short-reads (assembly, SNP-calling, sequencing errors) fostered the creation of bioinformatics tools to overcome their limitations. As a result, we have the multiple full genome catalogs for humans and Drosophila, such as 1000 Genome Projects (1000GP) or Drosophila Genome Nexus (DGN), on which much of the development of this dissertation thesis is based.

Nevertheless, it was not until 2007, with the study by Begun et al. (2007), that the first study of population genetics was carried out. Until then, as mentioned above, the large catalogs of variation were based on non-random regions or samples of the genomes despite sequencing advances. This generated a partial, perhaps biased, view of the inferred processes that shape genetic variation in natural populations. The first population genomic study carried out by Begun et al. (2007) (followed by Macpherson et al. (2007) and Sackton et al. (2009)) challenged the population level predictions of the neutral theory. However, the data from these studies had severe limitations. First, in the study of Begun et al. (2007), the sequenced lines did not come from a common source, implying that this is not a population genomic study in the strictest sense, with the implications that population structure can have on natural variation. Secondly, the study was carried out through low coverage sequencing, which can significantly impact the detection of variants and in the estimation of allele frequencies in the population. The approach of Begun et al. (2007) once again highlights: i) that the study of Drosophila is, and has been, one of the centerpieces in population genetics, ii) an approach that followed the future of genomic variation analyses not only in Drosophila (Macpherson et al., 2007; Sackton et al., 2009), but in other species.

The limitations of these studies have led to what can now be considered the standard analysis of population genomics. The Drosophila Genetic Reference Panel (DGRP) (Mackay et al., 2012) project sequenced with high coverage a total of 205 lines derived from a North American Raleigh natural population (RAL). The study provided for the first time the opportunity to perform the most comprehensive population genetics study done so far in any species, corroborating preliminary hypotheses and results on smaller datasets. First, they demonstrated that the pattern of polymorphism and divergence by functional site class is consistent within and among chromosomes. Second, polymorphism levels between synonymous and nonsynonymous sites differ by order of magnitude. Third, the proportion of the genome that is subject to the action of purifying selection was estimated, being around $\approx 40\%$. Globally, Mackay et al. (2012) estimated that $\approx 25\%$ of substitutions are adaptive, and that the centromeric regions show little evidence of positive selection.

While genome-level sequencing has provided a new resolution to understand molecular population genetics, the advances in other technologies have allowed unraveling multiple layers of the Genotype-Phenotype map. Hence, new omics datasets (regulatory elements, gene expression, chromatin states, etc.) allow a deeper characterization of the targets of natural selection. Furthermore, the advent of this huge amount of data allows us to test and create new hypotheses, and for the first time, there are enough resources to confront seriously theory and empirical data.

Genome-wide catalogs of variation

Since the beginning of sequencing, the effort to generate nucleotide variation catalogs has been more significant in humans than in any other species. As early as 2005, the International HapMap Consortium began creating the first catalog of common human genetic variation in diverse populations. The first version of HapMap, put on the databases the information of 264 samples corresponding to 4 human populations. The final version of the project featured haplotype maps of 1.6 million single nucleotide polymorphisms (SNPs) in 1184 reference individuals from 11 global populations. The haplotypic information deposited in HapMap revealed a linkage disequilibrium and low haplotype diversity (Consortium, 2007), leading to substantial correlations of SNPs with many of their neighbor SNPs.

Thus, the HapMap project allows for the first time genome-wide variation description and the detection of positive natural selection through the human genome, as well as the development of new tests to infer natural selection. These methods were based on the relationship between SNPs and the extent of the surrounding linkage disequilibrium (LD), looking for recent adaptation on the human genome (Voight et al., 2006; Sabeti et al., 2007; Pickrell et al., 2009). In Section 1.2.6, we review the main examples, as the integrated Haplotype Score (iHS) (Voight et al., 2006) or the Cross-Population Extended Haplotype Homozygosity (XP-EHH) (Sabeti et al., 2007), which were developed to exploit HapMap data. On one hand, HapMap data tested how ubiquitous natural selection was in the human genome and what kinds of genes and biological processes determined human adaptation.

Nevertheless, since HapMap publication, sequenced individuals and projects have continued growing per year and species. Table 1.1 reviewed some of the most important projects in terms of quality, accessibility, and impact in population genomics studies from several species. Important genome catalogs in other species, such as 1001 Genomes (Alonso-Blanco et al., 2016), Ag1000G (Miles et al., 2017), Great apes (de Manuel et al., 2016) or Simon Genome Diversity Project (SGDP) (Mallick et al., 2016) as well as the increasing number of ancient genomes and the sampling of populations in space and time have opened the door to new studies on a temporal and longitudinal scale (Kapun et al., 2021; Machado et al., 2021; Speidel et al., 2021). In addition, sequencing techniques have provided a deeper resolution at the phylogenetic level. A clear example is Clark et al. (2007) or, more recently, Kim et al. (2020), where 101 lines encompassing 93 Drosophila species were assembled. Overall, these datasets constitute a landscape when natural variation is revealed at the DNA level.

The Drosophila Genome Nexus (DGN) and the 1000 Genome Project (1000GP) represent the most significant examples of nucleotide variation at the genomic level for Drosophila and humans to date, respectively. In this thesis we have explored the genome variation patterns led by natural selection through the population information deposited in the 1000GP data. In addition, we have explored the creation of statistics and resources for the detection of recurrent positive selection, which take advantage of all the information deposited in 1000GP and DGN.

DGN. The Drosophila Genome Nexus project (Lack et al., 2015, 2016) provides the genome sequences of 1,121 worldwide *D. melanogaster* individuals from 58 populations out of 23 countries spanning 5 continents.

Each population genomic sequence is assembled against a single common reference. This project aims to increase the comparability of different population genomic data sets (Lack et al., 2015). DGN re-aligned genome sequences from: DPGP1

Table 1.1: Catalogs of genome-wide variation

Year	**Dataset**	**Citation**
2005	HapMap phase I	Altshuler et al. (2005)
2005	HapMap phase I	Altshuler et al. (2005)
2005	Perlegen	Hinds et al. (2005)
2007	HapMap phase II	Consortium (2007)
2010	HapMap phase III	Altshuler et al. (2010)
2010	1000GP pilot	Consortium (2010)
2012	DPGP	Langley et al. (2012)
2012	DPGP2	Pool et al. (2012)
2012	DGRP	Mackay et al. (2012)
2014	Great apes	Prado-Martinez et al. (2013)
2015	1000GP phase III	Consortium (2012)
2015	DGN 1	Lack et al. (2015)
2016	SGDP	Mallick et al. (2016)
2016	DGN 2	Lack et al. (2016)
2016	Great apes	de Manuel et al. (2016)
2017	1000Ag phase I	Miles et al. (2017)
2019	*D. simulans*	Signor et al. (2018)
2020	1000Ag phase II	Consortium et al. (2020)
2020	HGDP	Bergström et al. (2020)
2020	GnomAD	Karczewski et al. (2020)
2021	DEST	Kapun et al. (2021)

(Langley et al., 2012), 27 genomes from Malawi; DPGP2 (Pool et al., 2012), 139 genomes from 22 populations, mainly from Africa; DPGP3 (Lack et al., 2015), 197 genomes from Zambia; DGRP (Mackay et al., 2012), 205 genomes from Raleigh, USA; the global diversity lines (Grenier et al., 2015): 85 genomes from Australia, China, the Netherlands, the USA and Zimbabwe; (Bergman and Haddrill, 2015): 50 genomes from France, Ghana and the USA; (Campo et al., 2013)): 35 genomes from California; (Kao et al., 2015), 23 genomes from 12 New World locations; and 306 new sequenced genomes from Ethiopia, South Africa, Egypt and France; resulting in a dataset of 1,067 complete sequence genomes which cover almost the complete geographical range of this species.

1000GP. The 1000GP project is the largest catalog of human nucleotide variation published to date. However, with decreasing costs and improvements in sequencing technologies, we have seen an increase in catalogs of a similar nature, such SDGP, or Human Genome Diversity Project (HGDP) (Mallick et al., 2016; Bergström et al., 2020). The main goal of the 1000GP project was to discover and describe the different forms of polymorphisms at the genomic level in multiple populations. Therefore, it was designed not only to genotype markers but to characterize at least 95% of all variants present in the genome.

The project was divided into three phases due to the initial cost of sequencing and the technologies available. The 1000G pilot phase included complete, low-coverage whole-genome sequencing of 179 individuals from four populations: Yoruba individuals from Nigeria (YRI), individuals of European ancestry from Utah (CEU), Han Chinese individuals from Beijing (CHB), and Japanese individuals from Tokyo (JPT) (Consortium, 2012). It concluded with the detection and cataloging of 14.4 million SNPs, 1.3 million short indels, and over 20,000 structural variants. The resulting dataset covers approximately 85% of the reference sequence and 93% of the coding sequence of the genome, with the vast majority (97%) of inaccessible sites being high copy number repeats or segmental duplications.

By 2012, 1000GP consortium reported completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations, including samples from Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and the Americas (AMR). The combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping showed for the first time that most variants in the human genome are rare. About 64 million SNPs have a frequency $< 0.5\%$, 12 million have a frequency between 0.5% and 5%, and only about 8 million have a frequency. Moreover, as predicted by the out-of-Africa (OoA) hypothesis, the individuals from African ancestry showed a more significant number of variant sites. At the same time, the variation at admixed populations (AMR) was proportional to the degree of recent African ancestry in their genomes. Thus, the broad spectrum of genetic variation increased to a total of 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants) all phased onto high-quality haplotypes. This resource includes 99% of SNP variants with a frequency of 0.1% for various ancestries, describing the distribution of genetic variation across the global sample. To date, the project continues growing and it encompasses more than 3,000 highly-coverage genomes (Byrska-Bishop et al., 2021). Nonetheless, project such as HGDP, have led to the discovery of hundreds of thousands of new variants that reflect substantial amounts of previously ignored common genetic variation which together with the geographic and antrophogical information shows the importance of further studies for understanding human diversity (Bergström et al., 2020).

1.1.4 From genomes to trees

The correlation between recombination and nucleotide variation along the genome has been one of the primary pieces of evidence describing nucleotide variation (Begun and Aquadro, 1992; Mackay et al., 2012). Begun and Aquadro (1992) showed for the first time the positive correlation between the exchange rate (recombination) of genetic material and nucleotide variation, results validated by Mackay et al. (2012) at deeper resolution. In addition, recombination is also crucial to promote adaptation since it modulates interference between deleterious and advantageous alleles as demonstrated by Castellano et al. (2016), reducing or increasing adaptation in genomes. Linked selection models, such as hitchhiking and BGS, along with recombination, can better predict patterns of variation and thus explain the observed variation (Kern and Hahn, 2018; Gillespie, 1994; Hahn, 2008; Johri et al., 2020).

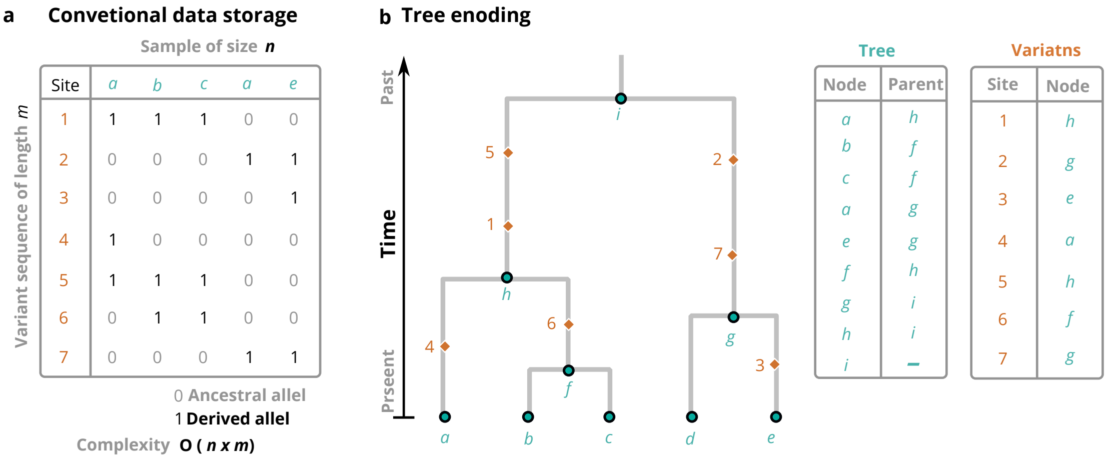
The increasing number of catalogs of nucleotide variation and NGS advances driven by new technologies, bioinformatic tools and statistics models have allowed the measure of local recombination rate, a crucial parameter in population genetics to understand patterns of genome variation. Comeron et al. (2012) integrated the power of classical genetics with NGS achieving the first integrated high-resolution description of the recombination patterns of both intragenomic and population variation. New methodologies try to overcome the limitation of theory (such as neutrally evolving sites, constant mutation and population size, demography, unphased data or pooled data) and computational requirements. Based on Sequential Markov Chain (SMC) models and machine-learning, we have methods that can infer recombination at a significantly higher resolution, scaling better to larger datasets (Spence and Song, 2019; Barroso et al., 2019; Adrion et al., 2020b).

Recombination generates different ancestries of a set of linked DNA sequences. Altogether, coalescence and recombination can be used to determine both the common sequence ancestor and the branching time. The historical process of recombination and coalescence that describes each site's evolutionary relationships and genealogy can be summarized in what is known as Ancestral Recombination Graphs (ARGs) (Dutheil, 2020). ARGs encode all the information by which a sample can be traced to a common ancestor. Possible recombination events cause the patterns to differ from one sample or site to another. ARGs summarize all the coalescence, recombination, and mutation information that produce the observed variation patterns. Consequently, inferring each site's genealogy in a genome sample comprises the information recorded at genomes describing different evolutionary processes. For example, variants under the effect of natural selection will show modified genealogies topology, since tree genealogies under natural selection result in shorter branches than expected in neutrally evolving populations (Harris, 2019). ARGs store information that standard summary statistics cannot assess, including introgression, time to the most common ancestors (TMRCA), recombination, and linkage disequilibrium (LD). Typically, ARGs are represented as marginal coalescence trees, including or not full information of the ARG, depending on the recombination time stored in such trees (Rasmussen et al., 2014; Brandt et al., 2021).

So far, inference from ARGs has been limited due to computationally requirements because mutation and recombination events finally result in an intractable probabilistic space (Dutheil, 2020). Based on SMC, introduced by McVean and Cardin (2005), `ARGweaver` software pioneered ARG inference in accuracy and computational performance (Rasmussen et al., 2014). Because `ARGweaver` is able to perform full ARG inference and quantifying inference uncertainty by sampling from the posterior distribution, it was restricted to a limited number of individuals and scaled poorly with the sample. Overall, the number of studies that benefit from ARGs' inference was low, but, in any case, the approach was prepared for genome-wide inference.

For a while, the computational cost of the inference and the many genomes available implied that conventional population genetic summary statistics overcomes ARGs (Hejase et al., 2020). Nonetheless, the state of ARGs inference has recently drastically changed. Simultaneously, Speidel et al. (2019) (`Relate`) and Kelleher et al. (2019) (tree sequence framework) developed new methodologies to approximate ARG while producing genealogies SNPs to SNPs at genome-wide level. Unlike `ARGweaver`, both methodologies cannot work with full ARG, but only encode topology change recombination events (Kelleher et al., 2019) or allow more than one recombination event between trees but finally encoding average of multiple coalescence trees (Brandt et al., 2021). Such methods approximate the coalescence with recombination using a modification of the Li and Stephens (2003) model to infer local tree topologies. As further explained in Brandt et al. (2021), `Relate`, use the Li and Stephens (2003) model to infer the topologies and use Markov Chain Monte Carlo (MCMC) under a coalescent prior to infer coalescence times. On the other hand, the tree sequence framework recreates ancestral haplotypes based on allele sharing between samples and applies a Hidden Markov Model (HMM) and the Li and Stephens (2003) model to generate the tree topology through using ancestral and the sampled haplotypes (Brandt et al., 2021). Both methods were tested in the most comprehensive human variation catalogs (1000GP, SGDP), analyzing selection, introgression, and population structure, among other patterns. Through extensive analysis using inference ARGs, Speidel et al. (2019)

dated *TCC* to *TTC* enrichment mutational signal (Harris and Pritchard, 2017) around 10,000 to 20,000 years ago, and introgression between Neanderthals and modern humans in Eurasia and between modern East and South Asians and Denisovans, alongside to other signals specific to African groups. Moreover, they tested for selection signals on complex traits while finding widespread directional polygenic adaptation at enriched SNPs in GWAS analysis. Remarkably, Kelleher et al. (2019) not only revealed subtle genetic distinctions among the populations of London, Edinburgh, and their rural outskirts or characterized ancestral relationships in 1000GP and SGDP datasets but also achieved store the information into a new data highly efficient tree sequence-structure called *tree succinct*.



Site	*a*	*b*	*c*	*a*	*e*
1	1	1	1	0	0
2	0	0	0	1	1
3	0	0	0	0	1
4	1	0	0	0	0
5	1	1	1	0	0
6	0	1	1	0	0
7	0	0	0	1	1

Node	Parent
a	*h*
b	*f*
c	*f*
a	*g*
e	*g*
f	*h*
g	*i*
h	*i*
i	–

Site	Node
1	*h*
2	*g*
3	*e*
4	*a*
5	*h*
6	*f*
7	*g*

Figure 1.1: *Tree succint* representation. A. Conventional matrix describing Variant Calling Format (VCF) storage. B. Genealogy encoding the data and constant variant storage. Figure taken from Kelleher et al. (2019).

While Variant Calling Format (VCF) data was the standard format to encode catalogs of genome-wide variation, the *tree succinct* seems to be the natural successor. *Tree succinct* benefits of the ancestral information provided by ARGs. As detailed in Kelleher et al. (2019), *tree succinct* is the result of recording mutations using the genealogy at particular sites, recording variation in the ancestry where these mutations arose. It allows reducing the classic matrix of n samples and m sites (see Figure 1.1), commonly used, to a file where each variant maintains a constant size and format (Kelleher et al., 2019). Therefore each variant is recorded where the mutation arose in the ancestry through edges and nodes. For example, at the largest simulation provided at Kelleher et al. (2019), representing the ancestry of 10^7 chromosomes, each of which is 100Mb long, *tree succinct* showed improvements in several orders of magnitude not only in storage (from TB to GB) but also in accessing data and computing statistics (from hours to seconds) concerning a compressed and more efficient version of VCF (Kelleher et al., 2019). *Tree succinct* format shows the future of massive sequenced data, such as the UK Biobank, which provides information about several thousand of individuals, overcoming the future limitations of the VCF in terms of performance, storage and scalability.

Speidel et al. (2019) and Kelleher et al. (2019) achieved a new milestone into molecular population genetics. The software and data structure will be key to explore evolutionary questions. New standard statistics will provide new accuracy level testing for introgression, selection, and demography (Harris, 2019).

1.2 Primary evolutionary forces: the theory

Population genetics moved from fundamentally theoretical science to a landscape where massive genomic data at different levels became available. Thus, the information encoded at the genetic level and its interpretation led to redefining population dynamics. The first population genetics model was proposed by G.H. Hardy and W. R. Weinberg in 1908 (the Hardy-Weinberg principle, (Hardy, 1908)). It can be defined as the zero-force state model and serves as a null model to explain the maintenance of genetic variation in populations. The principle states that allele frequencies would remain unchanged generation after generation once they reach the equilibrium state in an ideal population and the absence of any other evolutionary forces. A population in Hardy-Weinberg equilibrium underlies the following assumptions: i) diploid organism, ii) infinite population size, iii) sexual reproduction, iv) allele frequencies do not differ between sexes, v) absence of external forces affecting mutation dynamics, such as gene flux or selection, vi) random mating, vii) non-overlapping generations. From this perspective, the theoretical Wright, Fisher and Haldane modeled the fundamental evolutionary forces, conceiving population genetics as a theory on which diverse forces can affect the allele frequencies in a population: mutation, migration, natural selection, recombination, and genetic drift.

This section describes the main evolutionary forces through the Wright-Fisher model, the extension and predictions proposed by Kimura and Otha in the nearly neutral theory of molecular evolution, and the importance of these models in the current coalescence and forward-in-time simulations.

1.2.1 Wright-fisher model: genetic drift and probability of fixation

The assumptions of random mating and infinite population size in the Hardy-Weinberg law could hardly be associated with natural populations. Under the assumption of panmixia, for some populations with large individual sizes, it could serve as a first approximation to explain natural variation. However, for most species, the population size is not big enough to overcome the effect of genetic drift. Thus genetic drift occurs because alleles are transmitted to the next generation by chance. In a finite, diploid, sexual population, following Mendel's principles, only one of the two alleles in an individual, chosen at random at a locus, is transmitted to the offspring. These random samplings can significantly affect the evolutionary process (Gillespie, 1994). Wright and Fisher were the first to explore theoretically the impact of genetic drift in depth. Nowadays, we describe the Wright-Fisher model for the algorithm that defines a simplified biological model considering:

- We consider a constant diploid population
- Two segregating alleles, A1 and A2
- Discrete generations

In this model, each parental allele has the same probability of contributing to the next generation, so each new copy of a gene in the new generation depends only on that gene frequency in the previous generation. For many years, the genetic drift algorithm has been computationally simplified as sampling a bag of marbles with two colors (alleles A1 and A2). Analogous to a discrete generation, sampling $2N$ marbles randomly and replacing them with another bag can cause a fluctuation in the frequency of colors. By chance, type A1 (or A2) individuals may leave more or less offspring in the next generation. In genetic terms, to develop the current population, we randomly sample the parental population with replacement where, at the individual level, any parent allele has the same probability of appearing in the gamete, and at the population level, different individuals in the population can contribute unequally to the offspring, being the primary sources of randomness in the process (Gillespie, 2004). The random sampling process is mathematically described with a Binomial random variable. Following the examples at Masel (2011) and Gillespie (2004), Figure 1.2 shows genetic drift following a simplified version of the Wright-Fisher model. The simulation showed five independent populations considering 100 and 1000 individuals, where an allele A1 is at frequency 0.5 the first generation.

Overall, the changes produced uniquely by the binomial sampling is neutral, and the direction of the random changes cancels out in the long term. However, any trajectory is a random walk and behaves differently of any other, showing the stochastic nature of evolution in a future generation. Nonetheless, some alleles are fixed or lost to the population, removing genetic variability. As shown in Figure 1.2, this effect may be especially relevant in small populations. As the Wright-Fisher model shows, every segregating neutral mutation in a population is eventually fixed or lost by genetic drift. Furthermore, as shown in Figure 1.3, the initial allele frequencies influence the fixation or loss probabilities. Lower initial allele frequency is associated with higher frequent loss, just as higher initial frequencies with frequent fixations. Finally, binomial sampling allows us to estimate the final fixation probability of a neutral allele as a function of its initial frequency (i/N), and its extinction probability ($1 - i/N$), a classic result in population genetics which can be derived from diffusion theory too (Kimura, 1955).

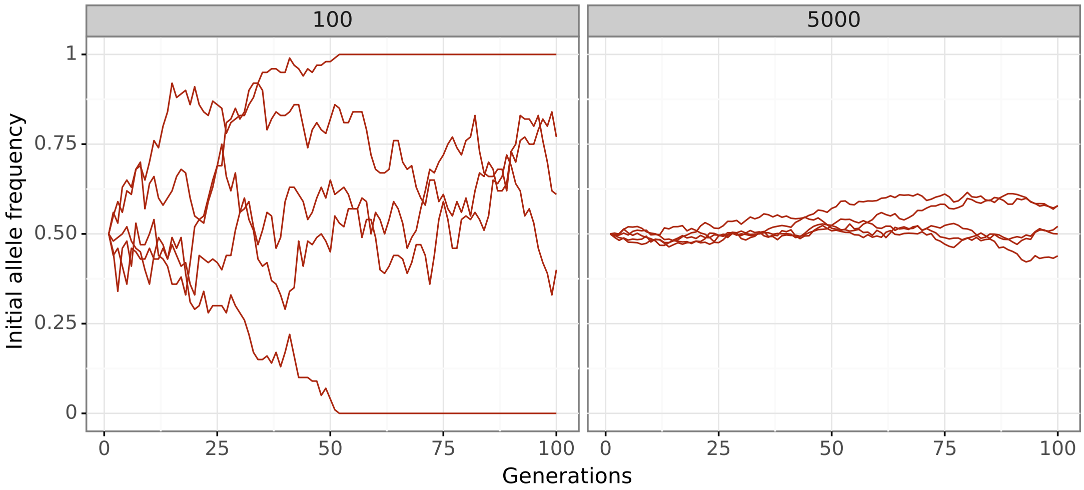


Figure 1.2: Simplified Wright-Fisher algorithm representing the stochastic nature of genetic drift. The frequency of the allele A1 is plotted over time and the initial frequency $p(0) = 0.5$. Each generation the initial frequency of the A1 allele (0.5) varies due to sampling with replacement. Panel A and B represent 5 independent replicates for an A1 allele with constant populations of 100 and 5000 individuals respectively.

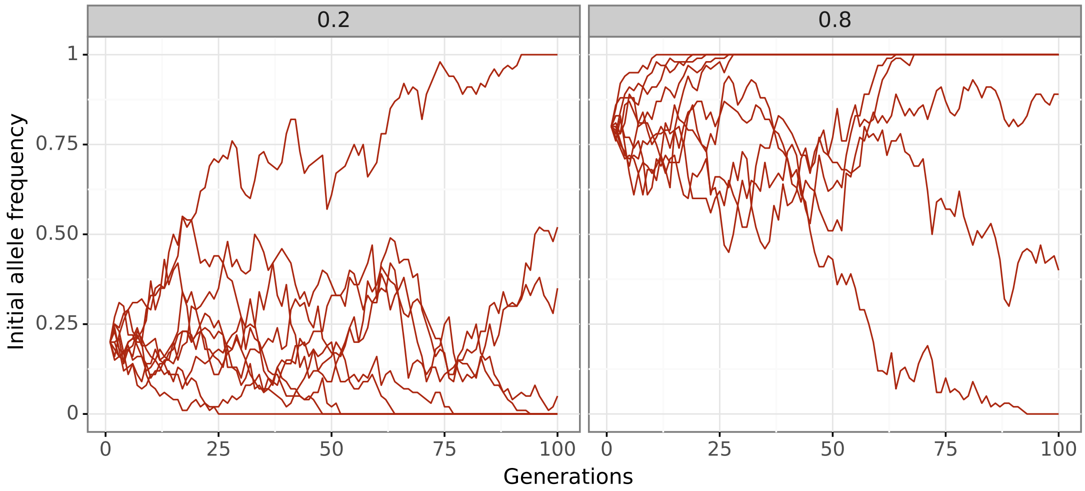


Figure 1.3: Genetic drift depending on initial allele frequency for a constant population of 100 individuals. Because fixation probability only depends on the frequency at $t = 0$, panel B, starting at frequency $p(0) = 0.8$ shows a higher number of fixed alleles, while neither replicas of panel A reach fixation.

The above assumptions define the simplest dynamic of the Wright-Fisher model. Nevertheless, considering that the state of A1 at time $t + 1$ only depends on its state at time t, it is possible to predict how genetic drift causes fixation or loss, as well the chances to find a population at a specific state (see Hartl (2020)). The likelihood to go from one state to another can be defined by extending the binomial formula.

$$P(i \to j) = \left(\frac{2N!}{j!(2N-j)!}\right) p^i q^{2N-j} \tag{1.1}$$

where j will be the number of alleles obtained from the initial number of i alleles after the sampling process. For i and j equal to $1, ..., 2N$ the values of $P(i \to j)$ define a square transition matrix, which show the probability of state i changing to state j in a single generation (Hartl, 2020), and commonly known as the transition probability matrix. The simplest case, a diploid locus in a population, can be summarized considering 0, 1 or 2 copies $P(0)$, $P(1)$, and $P(2)$. Considering the extreme case of obtaining two A1 alleles in the next generation, $P(2)$ can be defined as the sum of the transition probabilities given the combination and sampling error.

$$P_{t=1}(2) = (P_{2\to2})P_{t=0}(2) + (P_{1\to2})P_{t=0}(1) + (P_{0\to2})P_{t=0}(0) \tag{1.2}$$

Transition probabilities are calculated with the binomial formula, whereas probabilities at time $t = 0$ represent the frequencies of populations with a given allelic state. The Wright-Fisher model of genetic drift is a special Markov chain process with two absorbing states corresponding to the allele frequencies $p = 0$ and $p = 1$. Each fixation state is called absorbing because, once a subpopulation is fixed, it remains fixed. Eventually, each subpopulation is absorbed at either $p = 0$ or $p = 1$.

Equation (1.1) can model discrete generations for the Wright-Fisher model, where time and allele states move forward from the initial condition to another at time $t + 1$. The discrete modeling shows the role of genetic drift in actual biological populations over generations. Other models have been proposed to deal with the main assumptions of the Wright-Fisher (Moran, 1962; Cannings, 1975). While equation (1.1) fits the Wright-Fisher model, this discrete process can be approximated using particle diffusion theory and partial differential equations, where time and allele frequency are continuous variables. This approach is known as the diffusion approximation for genetic drift. Although Wright used diffusion approximation to model some aspect of genetic drift (Wright, 1938), the complete approximated solutions were solved by Motoo Kimura (Kimura, 1955).

Like Markov chains, diffusion theory can predict the probability distribution of frequency alleles over time. Thus, both predictions become similar to the outcome of neutrality and genetic drift. Nonetheless, diffusion theory assumes that populations are large enough, which turns the probability distribution into a continuous and smooth function compared to discrete prediction over generations output from the Markov chain estimations, predicting the distribution probability of an allele given a population. The diffusion approximation has been used to precisely describe genetic drift while being flexible to include other evolutionary factors, such as migration or selection, including non-equilibrium populations (Evans et al., 2007), as well as to measure the average time to fixation and loss for alleles the population. As previously mentioned, diffusion approximation converges to the classical probability of fixation of neutral alleles $i/2N$. Moreover, using the diffusion approximation, Kimura was able to define the probability of fixation of a selective allele (Kimura, 1964, 1968), being one the essential expressions in our field. Figure 1.4 shows the probability of fixation depending on the population size and coefficient product and the probability of fixation of a neutral allele.

$$P_{fixation} = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}} \tag{1.3}$$

where p is the initial frequency of the mutation, $s$is the selection coefficient, and N is the population size.

From equation (1.3) and by extension Figure 1.4 it is important to emphasize two key points. First, as described by Ohta (1973), the probability of fixation of a mutation depends not only on the effect it has on the biological fitness of the individual carrier, but also on the population size. The product of both parameters, Ns, the population scaled selection coefficient, predicts the relationship of forces between drift and selection. Second, that in the interval $-1 < Ns < 1$ the probability of fixation approaches that of neutrality. These nearly neutral or effectively neutral mutations are essential in explaining deviation of the neutral patterns of polymorphism and divergence that we observe in natural populations, culminating in the nearly neutral theory of evolution. For $|Ns| > 1$ selection significantly impacts the probability and time of fixation. These mutations are commonly referred to as slightly selected. At mutations with population scaled selection coefficient $|Ns| > 10$ natural selection exerts dramatic power over the population dynamics and genetic drift. These mutations are commonly referred to as strongly selected. Therefore, the selection coefficients scaled to population size are of evolutionary relevance. In Section 1.2.3 we review these concepts deeply, as well as the Distribution of the Effect (DFE) from which these values can be determined.

From diffusion theory and the Wright-Fisher model it can be derived another important equation commonly employed to define the population stationary frequency distribution. The stationary frequency distribution describes the density probability of a mutation i at frequency $x + dx$ (Wright, 1938), allowing the calculation of the frequency spectrum at different distributions of selection coefficients. It has been largely explored since Wright considered it for the first time (Wright, 1938), including non-equilibrium populations (Evans et al., 2007). The irreversible equation for the stationary distribution is commonly be stated as:

$$\phi(x) = \frac{1}{1-x} \frac{e^{4Ns}1 - e^{4Ns(1-x)}}{e^{4Ns} - 1} \tag{1.4}$$

Wright-Fisher model and diffusion theory for genetic drift have been extensively used in Chapters 4 and 5 of this dissertation. First, we account for forward simulations assuming Wright-Fisher populations, including selection and complex demography. Second, the Wright-Fisher model and the stationary distribution of alleles were used in Chapter 4 through a Maximum Likelihood approach based on the Poisson Random Field (PRF) framework to estimate the proportion of adaptive mutations. Finally, Chapter 5 uses diffusion theory to construct a population genetic model when solving the proportion of adaptive mutations using an Approximate Bayesian Computation.

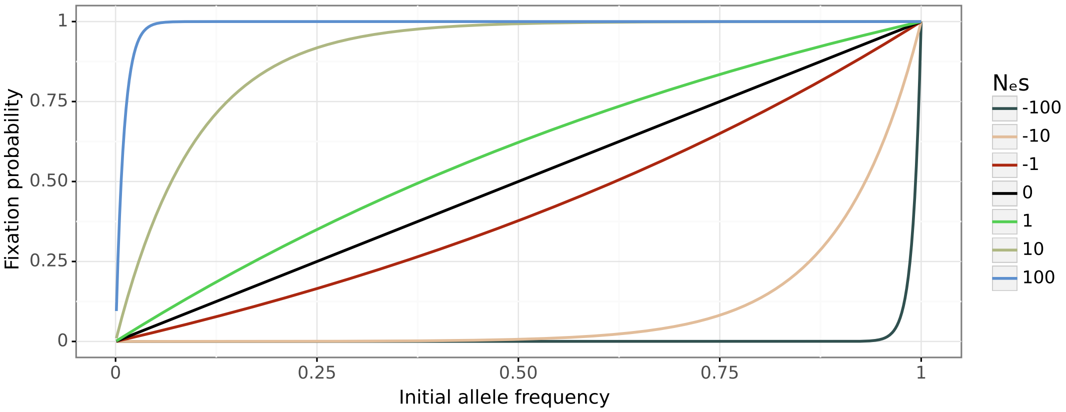


Figure 1.4: Fixation probability according to diffusion equations approximation with selection. Selection coefficients are shown in $4N_es$ units. The fixation probability of a neutral allele is equal to its initial frequency.

1.2.2 N_e vs N_c

Focusing on the mathematical definition of genetic drift, selection, and the extensions of the diffusion theory, we can observe that the dynamics of genetic variation depends highly on population size (N). In the Wright-Fisher population's context, the concept of effective population size (N_e) is the size of the idealized Wright–Fisher population that would show the same amount of genetic diversity or other parameters of interest as the actual population. Therefore the population size definitions depend on how genetic variation changes over time and rely on the dynamics of genetic variation in the population. Considering the Wright-Fisher model, A1 copies into the next generation are defined by a binomial random variable, implying a binomial variance. Hence, for big N, the variance in the frequency of allele A will be minor than under lower N. N_e, the effective population size, should be understood as the number of individuals that satisfies the expected variance in allele frequencies under the binomial distribution, not as the number of individuals we count in a population, the so-called population census (N_c). Thus, for example, a population showing frequency changes slowly over time is expected to be associated with a relatively large population size, as shown in previous simulations when considering the exclusive influence of genetic drift. In that sense, the effective population size allows us to compare the idealized drift expected in different frameworks, such as different populations or species. Wright-

Fisher model assumptions are commonly violated in the natural population. Several forces affect the offspring variance. Such is the case where populations size fluctuates over generations, where females and males differ, due to variation in the number of offspring per individual, or by continuous, overlapping, generations. Hence, the expected population size in a Wright-Fisher model, the effective population size, is usually smaller than the current census population size. In an equilibrium population, Wright-Fisher population size and N_e will converge. In populations with nonequilibrium histories, effective population size will differ from Wright-Fisher (Hahn, 2018).

In addition, it is essential to note that directional selection, which would increase the variance in the number of offspring per parent, also acts by decreasing N_e concerning N_c. Understanding the N_e dynamics is fundamental to explaining the evolutionary role of genetic drift and the interactions with other forces such as mutation, migration, recombination, and selection. Then, N_e values significantly affect genetic variability and the rate of evolution. For example, as extensively discussed at Charlesworth (2009), mutation rate and N_e will determine the equilibrium level of neutral or weakly selected variability in a population and the dynamic (fixation or loss) and effectiveness of selected mutation, since N_e can vary the intensity of selection (as explained in the previous section). All in all, N_e estimations allows us to understand the role of genetic drift and the other forces modeling genetic variation, capturing long-term population dynamics. Although N_e rather than population census size, N_c, is the parameter summarizing the actual number of individuals contributing to the offspring (Charlesworth, 2009), N_c determines, however, the number of new mutations entering in the population each generation. Karasov et al. (2010) presented a plaing example in the evolution of the *Ace* gene (a pesticide resistance). For accounting for the quick, repeated convergent mutations in this gene, they assume a $N_c \approx 10^9$, a value 100-fold the estimated effective population size of *D. melanogaster*. Therefore, despite N_e would finally determine the fate of beneficial alleles, N_c has to be accounted for to infer the adaptive potential of a gene or the whole genome.

1.2.3 Neutral, nearly neutral theory and Distribution of Fitness Effect

Hundred studies measured protein genetic variability within and between species with the advent of electrophoretic data (Nevo et al., 1984). Because of the amount of genetic variation exposed in these studies, Kimura proposed the neutral theory of evolution since the segregating load (genetic load) was incompatible with the classical thought that mutations should have selective effects. Furthermore, the underlying process driving the amino acid substitutions and the associated variation patterns was inexplicable for the balance hypothesis. In addition, previously, Zuckerkandl and Pauling (1965) estimated that mammal hemoglobins evolve at a roughly constant rate of amino acid substitutions per year (the so-called molecular clock hypothesis) whose substitution load was unaffordable for the survival of species .

Based on both the segregation and the substitution loads, Kimura suggests a radical alternative to explain genetic variation: the vast majority of segregating mutations should have little or no fitness advantage or disadvantage and therefore be selectively neutral. In this frame, most polymorphic and fixation patterns can be explained by genetic drift and mutation rate, the main evolutionary force that dictates the trajectory of mutations present in a population. Under the neutral theory, the frequency dynamics in the population are determined by the rate of mutation and random genetic drift. Considering neutral variants, the bulk of existing polymorphisms and fixed differences between species are selectively neutral and functionally equivalent. As reviewed by Casillas and Barbadilla (2017), the following statements summarize the neutral theory (also called the mutation-drift balance hypothesis):

- Deleterious mutations are rapidly eliminated from the population, and adaptive mutations are rapidly fixed. Therefore, within-species variation must be selectively neutral (i.e., the derived allele has the same biological fitness as the ancestral allele).
- Polymorphism is a transient phase of molecular evolution between extinction and fixation, rather than balanced selection.
- The levels of neutral polymorphism (θ) are the product of the neutral mutation rate and the effective size, N_e. Large populations will have more polymorphism than small populations.
- Neutral mutations are fixed at a constant rate (K) that is equal to the product of the mutation rate per generation (μ_0) and the proportion of new neutral mutations (f). $K = f\mu_0$.

Then, considering polymorphism as the transitional state to fixation (which ultimately contributes to divergence) or lost, neutral theory becomes the basic framework in which natural selection can easily be compared to neutral genetic drift in terms of allele trajectories and fixation patterns. Under neutrality, directional selected mutations will be lost or fixed faster than by genetic drift, since natural selection will deterministically reduce or increase their frequencies (see Figure 1.5). On the other hand, neutral mutations give rise to a random walk in frequencies that allele frequency dynamics and genetic drift can predict (see Figure 1.5). Many tests for natural selection underlie these neutral assertions, such as the McDonald and Kreitman test (MKT) (see Section 1.3).

A consequence of the neutral hypothesis is the minimal equation $K = \mu$ (Kimura, 1968). Under neutrality, the rate at which allelic changes are fixed in a given species (K) equals the mutation rate (μ). Also, this linear accumulation of substitutions over generations predicted by the neutral theory is the theoretical frame for the molecular clock hypothesis. We can define K as the rate at which mutations are fixed in each generation in a species. Hence, K informs about the rate at which species diverge over their evolutionary time. The fate of new variation also depends on the probability of fixation of each new mutation. This probability depends on two factors: the strength of selection (s) and the population size, assuming the simplification that the effective population size N_e equals N_c. Specifically, mutations enter the population at a rate of $2N\mu$ (the mutation rate is per site per generation, and in a diploid population, there are $2N$ potential chromosomes to mutate), so the overall molecular evolutionary rate taking into account all mutations is determined by the general expression.

$$K = 2N\mu_0 \int_{-\infty}^{\infty} u(N,s)f(s)ds \tag{1.5}$$

Considering neutrality, most of the mutation will be neutral or strongly deleterious rather than beneficial, hence s will be $s = 0$ and $s << 0$ and K is defined as

$$K = 2N[\mu_0 u(N, s = 0) + (\mu - \mu_0)u(N, s << 0)] \tag{1.6}$$

Note that the probability of fixation of a neutral mutation equals its initial frequency in the population $u(N, s) = \frac{1}{2N}$ and fixation probabilities of strongly deleterious mutations are null, then K turns to the minimal Kimura's expression.

$$K = 2N\mu_0 \frac{1}{2N} = \mu_0 \tag{1.7}$$

Several studies in the mid-1970s showed that different proteins had different molecular clocks. For example, Dickerson showed that the values ranged from $9 \cdot 10^-9$ substitutions per site per year, like fibrinopeptide to $10 \cdot 10^-11$ per site per year, like histone IV. The neutral theory was challenged because the rates of protein evolution were proportional to absolute time (in years), not to generation time; the expected time unit if mutation rate depends on generation time. In addition to Lewontin's paradox, this assertion represented the most controversial points of the neutral theory, despite the multiple proposal attempting to explain it (Ohta, 1972; Lynch, 2006; Corbett-Detig et al., 2015; Buffalo, 2021)

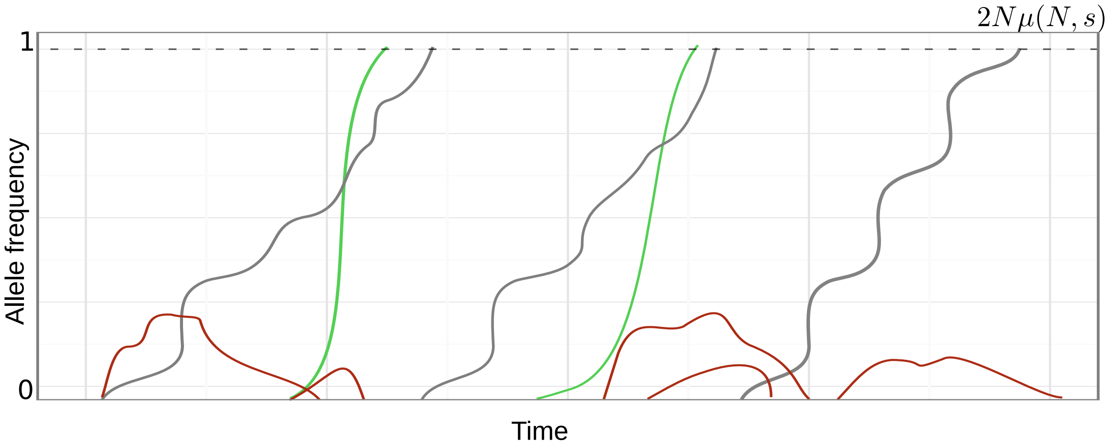


Figure 1.5: Frequency trajectories of neutral (blue), adaptive (green) and deleterious alleles (red). Neutral alleles are eventually fixed following a random walk that can be predicted by genetic drift. Selected alleles are rapidly fixed or lost depending on their selection coefficient and the action of natural selection.

Tomoko Ohta redefined neutral theory by introducing a new class of mutation, nearly neutral mutation (Ohta, 1973). To understand Ohta's proposal, known as the nearly neutral theory (Ohta, 1992; Ohta and Gillespie, 1996), it is needed to consider the Distribution of Fitness effect (DFE), another core concept in molecular population genetics. The DFE reflects the distribution of selection coefficients of new mutations. As we have just mentioned, the neutral theory considers only three types of effects: neutral, deleterious, or beneficial, of which neutral mutations account for the vast majority of polymorphism, since deleterious mutations are rapidly eliminated from the population by the action of natural selection and do not contribute to intra or interspecific variation.

Otha's proposal finally reflected that the DFE should be continuous and proposed to incorporate two new mutations concerning the neutral theory. On the one hand, effectively neutral mutations, mutations either slightly beneficial or deleterious that behave as neutral. These mutations have a fitness effect coefficient much smaller in magnitude than $1/N_e$, spanning the range $-1 < N_e s < 1$. Such mutations act as effectively neutral because their fate is basically controlled by genetic drift. On the other hand, nearly neutral mutations, which are mutations that have fitness effects on the order of $1/N_e$. These kinds of mutations range is $|10 < N_e s < 1|$. Note that s

can be positive or negative too. Summarizing, nearly neutral mutation can be slightly deleterious or advantageous, and their fate depends on natural selection and genetic drift.

Since more mutations will fall in the range $-1 < N_e s < 1$ in populations with small effective population sizes, they will have a larger number of effectively neutral mutations. Thus, in small populations, genetic drift surpasses natural selection often. Conversely, as N_e increases, less mutations fall within a coefficient $|N_e s| < 1$, purging deleterious mutations or favoring the fixation of beneficial mutations.

Thus, N_e dramatically influences the strength of purifying selection when purging slightly deleterious mutations (SDM). Because larger populations usually have shorter generation times, you also expect higher mutation rates. Since the proportion of effectively neutral mutations (f_0) and neutral mutation rate counteracts ($K = f_0 \mu_0$), this can explain why neutral substitution rate (K) is constant per year between species, and protein evolution is relatively insensitive to generation time contrary to strict Kimura neutral theory. Considering the DFE and the effective population size, Kimura and Ohta (1971) redefined the molecular clock by assuming that less important proteins, or parts that are less important for their function, evolve more rapidly, while other less critical parts or proteins are constrained because of their fitness. Thus, assuming nucleotide polymorphism as a phase of molecular evolution (Kimura and Ohta, 1971).

The shape of DFE is a fundamental information in population genetics and other research fields. We need to know the shape to predict the polymorphism and divergence of any given species, to explain the maintenance of quantitative and phenotypic genetic variation in quantitative genetics, to understand the relationship between evolution of sex and recombination, or predict the rate of genomic degradation due to Muller's ratchet. It has been the subject of debate for 30 years since Otha's proposal, in which, for the first time, underlyingly and due to new types of mutations, it is stated that the DFE must be continuous. Today we know that this is the case thanks to a range of studies ranging from mutation accumulation experiments to statistical inference using synonymous and nonsynonymous polymorphism. That continuum would encompass strongly deleterious, slightly deleterious, effectively neutral, and slightly beneficial and highly beneficial mutations, as opposed to Kimura's original neutral theory.

The fundamental question, therefore, is what shape such a DFE takes. Over the last decade, many models have been developed to infer the DFE. Inferring the DFE can help us resolve which traits or phenotypes are subject to natural selection since we might know what proportion of mutations are selective. Statistical methods typically compare the levels of synonymous and nonsynonymous polymorphism to find out. Considering that synonymous variability evolves neutrally, the levels of nonsynonymous variability that differ may reflect the nature of the DFE in nonsynonymous variants, which, a priori, are more likely to affect protein functions. Such approximations have determined that in *D. melanogaster* and humans around 20-30%, and 6% of non-synonymous mutations are effectively neutral ($-1 < N_e s < 1$), while between 10-20% are slightly deleterious ($-10 < N_e s < -1$) (Eyre-Walker et al., 2006; Boyko et al., 2008; Eyre-Walker and Keightley, 2009). More importantly, there is no single DFE; each nucleotide depending on the functional class to which it belongs, or mutation type, such as insertions or inversions, has its own DFE and may vary across the genome. Over the last decade, several mathematical models have been proposed to infer the DFE, although it is unclear what type of distribution may best fit the data, whether multiple species share a similar form of DFE (Galtier and Rousselle, 2020) or even if it follows a continuous distribution, as extensively reviewed by Kousathanas and Keightley (2013). Some representative DFE models are discussed in depth in Section 1.3. and used in Chapter 4 and 5

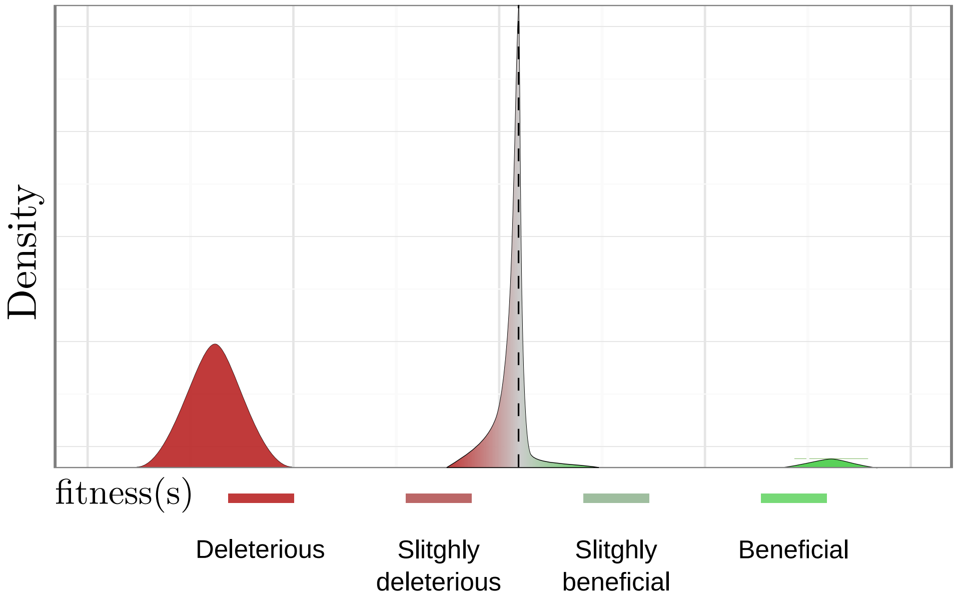


Figure 1.6: Continuous DFE according to the nearly neutral theory of molecular evolution. Mutation mutations are colored from red to green according to their fitness effect. Effectively neutral mutation are shown in gray.

1.2.4 Lewontin paradox, background selection and hitchhiking

Lewontin's paradox remains one of the Achilles' heels of population genetics. This paradox, described by Lewontin (1974), arose with the data provided by the allozyme era, and despite the limitations of this technique, it remains unsolved by DNA sequencing data.

Neutral and nearly neutral theories assume that nucleotide variation is mainly determined by the balance of new mutations and drift. Hence, considering nearly neutral theory, mutations introduce genetic variation at $2N$ while genetic drift removes it depending on the population size at a rate $1/2N$. As we explained in the previous section, in small populations genetic drift removes variation faster than mutation adds. It is therefore intuitive that there must be a relationship between variation and effective population size. Under neutrality, this relationship can be summarized through the expected neutral heterozygosity (θ). θ is defined as $\theta = 4N\mu$, from which it is easy to predict from his linear relationship that small populations are expected to show lower variation levels. Unfortunately, this prediction was precisely challenged with the advent of the allozyme-era data and remains unresolved to date. The failure of this prediction is known as Lewontin's paradox. Lewontin (1974) and subsequent studies (Buffalo, 2021; Leffler et al., 2012) showed that heterozygosity levels vary only a few orders of magnitude between taxa, while species' population sizes vary up to several (Buffalo, 2021).

Lewontin's paradox implies the lack of a model that can explain the levels of nucleotide variation between species. As mentioned in the previous section, together with the on year substitution rate of the molecular clock, it was one of the most criticized points of the neutral theory. Moreover, Tomoko Ohta's revision through the nearly neutral theory has not provided a reliable explanation either. This situation continued the neutralist-selectionist debate for years, despite the incorporation of two significant models broadly consistent with current data that can redefine the null hypothesis (see Figure 1.7). Both models can usually be interpreted as extensions of the nearly neutral theory once it is considered that sites do not segregate independently and the effect of selection on neutral linked variants.

On the one hand, Smith and Haigh (1974) introduced genetic hitchhiking as an molecular evolutionary force. The hitchhiking model proposes that when a beneficial allele reaches fixation, the diversity around the allele is reduced because neutral variants are swept along the selective fixation, a process which was named later selective sweep. Hence, the reduction of genetic diversity is directly related to the rate of fixation of neutral-linked variants, and levels of genetic variation are determined over time by mutation, genetic drift and recombination. John Gillespie developed a stochastic model that considers both the effects of genetic drift and recurrent hitchhiking, called genetic draft (Gillespie, 1994). Like genetic drift, genetic draft eliminates genetic variation and depends on population size. As we have seen in the previous section, genetic drift is less effective in eliminating alleles on large populations, while natural selection is more effective. If we also consider that more adaptive mutations will occur by chance in these populations, genetic variation should be significantly reduced due to a more significant number of hitchhiking events. With this model, population size and genetic diversity can be decoupled, potentially resolving Lewontin's paradox. In addition, the genetic draft model is consistent with high adaptation rates in large species (Hahn, 2008, 2018)

On the other hand, Charlesworth et al. (1993) proposed the background selection model (BGS) . This model can be considered similar to the hitchhiking model. However, it involves deleterious mutations rather than beneficial. Thus, in the BGS model, linked neutral variation is removed from the population due to linkage with deleterious alleles, achieving not fixation but the loss of alleles by the action of purifying selection. BGS is expected to reduce genetic variation like the hitchhiking effect, although it can hardly mimic the pattern of genetic variation (Schrider, 2020). The effects of BGS on fixations and frequency spectrum have been the subject of much theoretical work (Charlesworth et al., 1993; Charlesworth, 1994; Hudson and Kaplan, 1995; Barton, 1995; Nordborg et al., 1996). In Chapter 3, we explored the role of BGS in the MKT. We use the classical BGS model described in Charlesworth et al. (1993), Charlesworth (1994), Hudson and Kaplan (1995) and Nordborg et al. (1996) to model the deleterious alleles considering analytical estimation of positively selected allele fixation and BGS model, while accounting for linked neutral diversity reduction and the reduction of fixation probability of a positively selected allele under BGS (Charlesworth, 1994).

Both models, selective sweeps and background selection, have been extensively applied to numerous species and genome regions. However, although the effect appears ubiquitous, Hahn (2018) remarks that *what remains to be determined is the relative importance of each process across regions of the genome and species.* For that reason, neither genetic draft nor BGS can constitute a null model, whereas the nearly neutral theory can fit any species or part of the genome in a generalized way, allowing us to test the role of natural selection in cases where we obtain unexpected patterns.

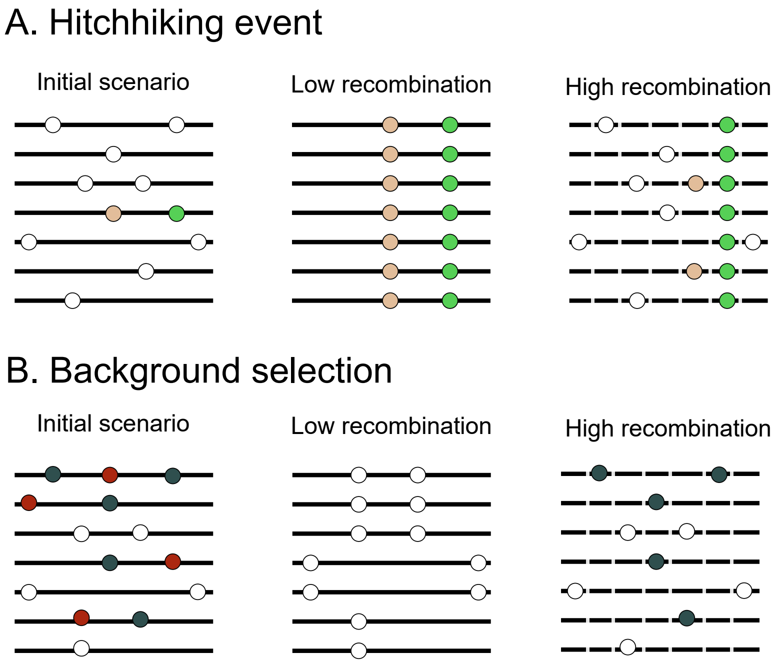


Figure 1.7: Hitchhiking and BGS effect on neutral linked sites considering low and high recombination scenarios. A. Neutral mutations (gold dots) are fixed along with beneficial mutations (green dots) because of linkage, resulting in a reduction of genetic diversity. If recombination is sufficiently high, the neutral levels (white dots) of polymorphisms can be recovered. B. Neutral mutations (blue dots) are purged along with deleterious mutations (red dots) because of linkage, resulting in reduced genetic diversity. However, if recombination is sufficiently high will break the haplotype. Hence linked neutral alleles (blue dot) remain in the population.

1.2.5 Hill-Robertson interference

Although Lewontin's paradox remains unsolved and requires complementary models to understand the relationship between diversity and population size, the hitchhiking and BGS models of linked selection place the importance of recombination in nucleotide diversity levels. The fate of a new mutation is conditioned not only by the selective advantage or disadvantage it confers, but also by the genomic context in which it appears. A classic paper of population genetics, presented by Hill and Robertson (1966), predicts that when selection is common, an increased linkage between sites will limit the effectiveness of both positive and purifying selection since selection at one site interferes with the selection at other linked sites. Therefore, if one or more selected mutations surround a newly selected mutation, they will interfere with each other since they do not segregate independently. Hence, recombination can determine not only the fate of the selected mutation but also the fate of surrounding mutations.

The reduction in selection efficiency due to interaction between linked sites is known as Hill-Robertson interference (HRi (Hill and Robertson, 1966), see Figure 1.8). HRi can occur in two different ways, involving either beneficial alleles or deleterious alleles, and both types can compromise the adaptation of genomes. HRi involving beneficial alleles occurs when beneficial mutations segregate simultaneously in different haplotypes and compete for fixation. This type of HRi is known as clonal interference. The second type (a Ruby in the Rubbish (Peck, 1994)), involving deleterious alleles, occurs when a beneficial mutation appears in a genetic background loaded with segregating deleterious mutations.

The context of genomic recombination determines both types of HRi. In a low recombination context, the beneficial alleles will compete until one of them will become fixed together with the carrier haplotype, while the other beneficial alleles are lost. In these cases, deleterious mutations linked to the beneficial mutation could be carried over to fixation due to clonal interference. In a Ruby in the Rubbish scenario, beneficial mutations in linkage with deleterious ones are lost. In contrast, when recombination is high enough, haplotypes carrying beneficial mutations can swap alleles generating a new haplotype carrying both adaptive mutations and fix clonal interference. Similarly, deleterious alleles can be eliminated, and adaptive alleles can be fixed without interfering with each other. This interference can be the result of the BGS process. Both types of interference limit the rate of adaptation in genomes, mainly affecting the fixation of slightly beneficial alleles (Uricchio et al., 2019). These concepts are discussed in-depth in Chapter 5. As mentioned above, the chapter shows the incorporation of the BGS model into the MKT, which indirectly measures interference by correcting α when selection is weak

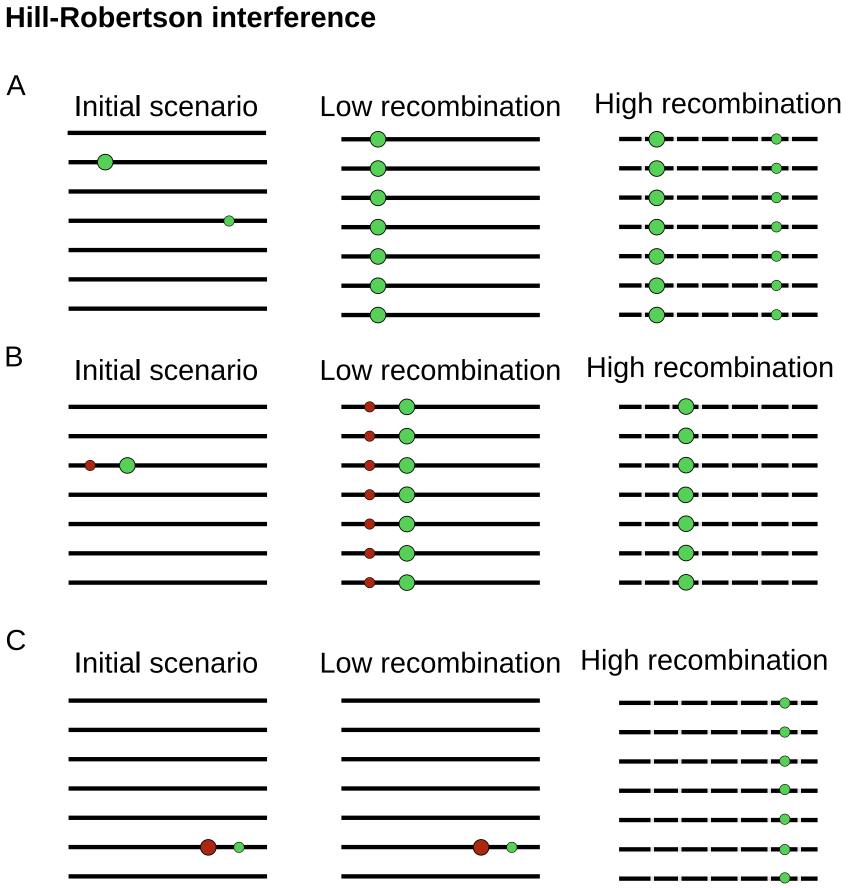


Figure 1.8: Hill-Robertson interference considering adaptive and deleterious mutations. Dot color represents beneficial (green) or deleterious (red) mutations and size the strength of selection. A. Two adaptive mutations occur in two different haplotypes and compete for fixation. If recombination is sufficiently high, it will generate a new haplotype carrying both adaptive mutations and can both be fixed. B-C. Deleterious alleles are dragged to fixation due to linkage to beneficial one or beneficial alleles are purged depending on the selection coefficients. If recombination is sufficiently high, deleterious alleles can be purged and beneficial alleles are fixed without interference.

1.2.6 Signatures and tests of positive selection

Much has been debated about the ubiquitous effect of BGS, positive, linked selection and neutrality (Hahn, 2008; Kern and Hahn, 2018; Jensen et al., 2019; Johri et al., 2020). Although the effect of linked selection is one of the most plausible explanations to unexpected neutral patterns in the genomes and Lewontin's Paradox, the most recent studies show we still need a selection model that fully describes the patterns of genetic variation among species (Corbett-Detig et al., 2015; Buffalo, 2021). Despite the controversy, there is ample evidence that positive selection is ubiquitous in the genome, and during the last decade, neutral theory has provided the necessary null model against which to evaluate non-neutral hypotheses. These patterns can be tremendously valuable for inferring the presence and effects of mutations, including mutations that we can consider advantageous (direct selection) and the effects on linked mutations (linked selection).

Numerous studies verified that, more than positive selection, is BGS the pervasive form of selection across the genome (McVicker et al., 2009; Murphy et al., 2021). McVicker et al. (2009) summarize the effect of background selection through the B statistic. The author stated that a sizable genomic fraction (19-26%) has a reduced neutral diversity due to BGS. BGS has been commonly explored at the genome-wide level (Lohmueller and Nielsen, 2021), showing that most mutations tend to be deleterious, as proposed by the neutral and nearly neutral theory.

However, positive selection can leave much stronger signals in the variation pattern. Nowadays, a great debate has arisen regarding the relative importance of demography, linked selection, and genome-wide BGS. While some argue that much of the nucleotide patterns can be explained by different modes of selection (such as hard and soft sweeps), others argue for incorporating a new null model, under which demographic factors and BGS can alternatively explain the patterns attributed to positive linked selection (Kern and Hahn, 2018; Johri et al., 2020). However, a detailed exploration of this problem shows that BGS patterns can hardly resemble the diversity patterns produced by a selective sweep (Schrider, 2020). Regardless of the debate, it is clear that natural selection results in patterns of diversity and linkage disequilibrium that cannot be explained under the null neutral model.

During the last decade, neutral theory has provided the necessary null model to evaluate non-neutral hypotheses such as hitchhiking or BGS. The most direct consequence of genetic hitchhiking is the reduction of neutral diversity around the selected mutation. There are other effects besides the reduction of genetic diversity, such as the deviation of the neutral SFS pattern and the increased homozygosity of haplotypes (Lohmueller and Nielsen, 2021). The deviations of the neutral diversity patterns depend on the sweep's time or strength and have been widely used to detect and measure the effect of positive directional selection. The following section describes the main summary statistics used to describe the patterns of genetic variation and to pinpoint genomic regions subjected to positive selection. They are classified into: deviations from neutral SFS, population differentiation, and LD (see Figure 1.9). In addition, because this thesis is mainly focused on the MKT, we also dedicate a section to describe the signatures of recurrent positive selection.

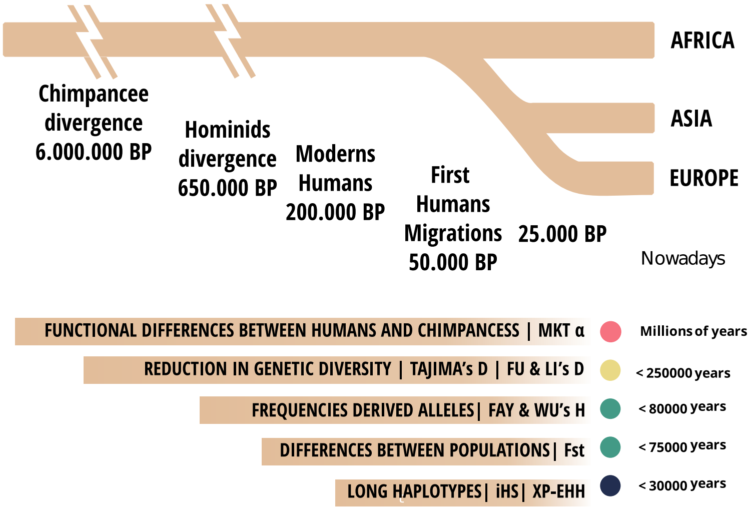

AFRICA
ASIA
EUROPE
Chimpancee divergence 6.000.000 BP
Hominids divergence 650.000 BP
Moderns Humans 200.000 BP
First Humans Migrations 50.000 BP
25.000 BP
Nowadays
FUNCTIONAL DIFFERENCES BETWEEN HUMANS AND CHIMPANCESS | MKT α
Millions of years
REDUCTION IN GENETIC DIVERSITY | TAJIMA's D | FU & LI's D
< 250000 years
FREQUENCIES DERIVED ALLELES| FAY & WU's H
< 80000 years
DIFFERENCES BETWEEN POPULATIONS| Fst
< 75000 years
LONG HAPLOTYPES| iHS| XP-EHH
< 30000 years

Figure 1.9: Statistics and signatures to infer selection at DNA level regarding the timing of the selective events. Adapted from (Sabeti et al., 2007)

Deviations from neutral SFS

The reduction in neutral diversity produced by selective sweeps can result in unexpected patterns at the SFS depending on the sweep fixation past time. A significant proportion of intermediate neutral alleles drifting in the population are purged by the hitchhiking process (Lohmueller and Nielsen, 2021). Once the advantageous mutations and linked variants get fixed, new mutations will restore neutral diversity levels. However, new mutations will appear slowly, and all the new variants in the population will be at low frequency. Hence, fixed sweeps can be detected not only by the reduction in genetic diversity, but also by an excess of rare variants at SFS (Sabeti et al., 2007). The strength of selection finally determines the reduction in genetic diversity and the size of the affected regions. Stronger adaptive events result in rapid fixations and, finally, higher excess of rare alleles, since recombination cannot rescue unlinked neutral variants at intermediate frequencies (Lohmueller and Nielsen, 2021). This kind of measure is interesting because the reduction of genetic diversity persists more extensively than other signatures ($< 250,000$kyrs, see Figure 1.9).

Tajima's D (Tajima, 1989), Fay and Wu's H (Fay and Wu, 2000), and Fu and Li's D and F (Fu and Li, 1993) are among the summary statistics most widely used to detect this kind of selection signature. Tajima's D measures the difference between two estimators of the population variability (θ_w and π). Under neutrality, the means of θ_w (Watterson, 1975) and π should approximately equal one another. Therefore, the expected value of Tajima's D for a population conforming to a standard neutral model is zero. Significant deviations from zero indicate a skew in the allele frequency distribution relative to neutral expectations. Positive values of Tajima's D arise from an excess of intermediate frequency alleles and can result from population bottlenecks, structure, or balancing selection. Negative values of Tajima's D indicate an excess of low-frequency alleles and result from population expansions or positive selection. Fu and Li's D and F are based on the number of old and recent mutations found and expected under neutrality. It is computed by comparing the number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants or the mean pairwise difference between sequences. Thus, the expected value considered under neutrality is zero, and significant deviations from zero are informative about distinct demographic and/or selective events.

Before sweep fixation, neutral linked alleles will be found to be at high frequencies since they had hitchhiked along with the selected alleles. Eventually, once the selected alleles get fixed, neutral linked alleles become fixed too, and the excess of high-frequency alleles will disappear (Zeng et al., 2006). However, positive selection can create an excess of high-frequency derived alleles due to incomplete sweeps or due to recombination of the selected variants during hitchhiking (Lohmueller and Nielsen, 2021), opposite to the excess of rare alleles where positive selection signals are reflected because of the associated reduction of diversity and the presence of new alleles. Fay, and Wu's H detects the presence of an excess of high frequency derived alleles in a sample by comparing pairwise differences in the sample to the total number of the homozygous allele for the derived allele (Vitti et al., 2013). The derived allele is defined from the ancestral alleles of a closely related species, assuming each mutation occurred only once since the species split. More sophisticated methods to infer the ancestral state include multiple species comparison (Keightley and Jackson, 2018). The excess of rare and high-frequency derived alleles has been widely used as signals of positive selection in the genome. Nonetheless, several simple demographic processes can mimic both. For example, population expansion can result in higher proportions of new rare alleles, whereas population splits can result in high-frequency differentiated alleles. In addition, these patterns primarily reflect hitchhiking events from de novo mutations. When considering positive selection on standing variation, the selected alleles probably recombine several times before the action of natural selection. Thus, the hitchhiking process will occur on different genetic backgrounds affecting both the reduction of the genetic variation and the expected SFS patterns.

Population differentiation

Different populations can be subjected to different selective pressures, resulting in high levels of genetic differentiation between them. The measure of genetic differentiation between populations can disentangle the role of positive selection on both. Hence, if natural selection is acting on a specific trait in a population, then the adaptive allele and neutral linked sites can differ significantly regarding frequencies. The population undergoing selection is expected to have high frequencies at selected and neutrally linked alleles or even be a private mutation in the selected population. F_{ST} relates the amount of genetic variation among populations to the total genetic variation of overall populations. Genetic drift, migration, and admixture define genetic diversity between populations considering the nearly-neutral theory. Nonetheless, local adaptation will contribute to the level of population differentiation at a particular locus, creating unexpected patterns of diversity resulting in large F_{ST} values. The largest values of F_{ST} at a locus indicate differentiation between populations. First proposed metrics of population differentiation have been widely extended to improve power. For example, the locus-specific branch length metric (LSBL) compares pairwise F_{ST} measures between three or more populations (Shriver et al., 2004). LBSL have benefited from different genetic contexts to isolate population-specific differentiation (Vitti et al., 2013). Population differentiation can only arise when populations are partially isolated reproductively and both are subject to different selective pressures (Sabeti et al., 2006). Natural selection may change an allele frequency in one population but not in another, or act in the opposite direction. The extreme differentiation patterns detected will finally depend on the nature of the selection and the direction of selection in each population. Unlike other methods, population differentiation–based approaches can detect many types of selection, including classic sweeps, sweeps on standing variants, negative, and balancing selection (Lohmueller and Nielsen, 2021).

Linkage Disequilibrium (LD)

With the fast rise in frequency of a selected allele and the associated hitchhiking process, there is not enough time for recombination to break down the association with the neighboring loci on the ancestral chromosome. Such a collection of alleles in a chromosomal region that occurs together in individuals is termed haplotypes. Long haplotype methods can be beneficial in detecting ongoing sweeps. However, the unexpected patterns of LD would persist shortly in time because of the action of recombination. After 30,000 years, a typical chromosome will have undergone more than one crossover per 100 kb, leaving fragments that are too short to detect. However, it will depend on selection strength and local recombination rate too.

The Extended Haplotype Homozygosity (EHH) test was one of the first statistical methods to explore these kinds of signals. As defined by Sabeti et al. (2002), EHH is *the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent for the entire interval from the core region to the point x.* It captures the decay of identity and the distance of a haplotype carrying a specific allele at a position of interest, correcting for local recombination rates. EHH values decrease from 1 to 0 with increasing distance from the core-site. In the case of strong positive selection, due to the rapid increment in frequency, haplotype homozygosity will tend to extend much further than expected under neutrality (Voight et al., 2006). EHH is one of the most sophisticated methods to measure the effect of positive selection on haplotype structure. Major limitations related to EHH include non-uniform recombination levels and arbitrary physical distances to define perfect homozygosity to the core haplotype. Voight et al. (2006) proposed the integrated haplotype score (iHS) to solve EHH's major limitation. First, iHS measures the EHH decay as a function of the distance to the core haplotype. The integration of the function represents a less arbitrary statistic than EHH measured for a fixed distance to the core haplotype. Second, iHS used genetic distance rather than physical, capturing the recombination between the tested allele and the tested position. It has good power to detect ongoing sweeps where the selected allele has a frequency between 50% and 80% (Pickrell et al., 2009). $\mathrm{nS_L}$ is conceptual and mathematically similar to iHS. However, it measures the length of homozygosity between a pair of haplotypes in the number of mutations and the remaining haplotype at the region (Ferrer-Admetlla et al., 2014). Other important statistical methods have been developed, benefiting from the ideas explored at EHH and iHS. The cross-population extended haplotype homozygosity (XP-EHH) uses IHH measures around the same allele in two different populations. XP-EHH is similar to $\mathrm{F_{ST}}$, but using haplotype structure rather than changes in allele frequencies. It has more power to detect selective sweeps with haplotypes at above 80% frequency.

LD-based approaches have been widely used to detect recent adaptation on standing variation during the last decade too. Garud et al. (2015) developed H12, which measures the increased levels of haplotype homozygosity due to the increase in the frequency of adaptive alleles. Nonetheless, assuming that the adaptive alleles are drifting in the population, different haplotype backgrounds are expected to be at high frequency. H12 evaluates the frequency and presence of the most common background haplotypes making inferences of soft-sweeps in addition to classic hard sweeps (Garud et al., 2015, 2021). In addition to H12, other methods considering selection on standing variations have been developed and tested over different datasets and species, such as diplo-SHIC (Kern and Schrider, 2018), LASSI (Harris and DeGiorgio, 2020b), saltiLASSI, (DeGiorgio and Szpiech, 2021), SS-H12 (Harris and DeGiorgio, 2020a), or G123 (Harris et al., 2018).

It is also essential to consider that different tests can measure different types of signals, which on the one hand, can be helpful since knowing each statistic can be accurate in solving the proposed hypothesis. However, on the other hand, it can be challenging to establish which combination of methods is most beneficial to the whole-genome positive selection detection. To address this problem, composite of multiple signals, machine learning approaches and Approximate Bayesian Computation (ABC) can provide powerful ways to disentangle the patterns of natural selection considering any possible demography or selection background.

1.2.7 Simulations

Considering the Wright-Fisher model and the neutral theory, nucleotide variation data can be evaluated against a null hypothesis. However, testing such hypotheses requires a comparison between the observed data and data that can reproduce neutrality. Therefore, modern population genetics usually use stochastic data in silico, following a simulation process of the sequence data. The simulation process can be a valuable resource to evaluate evolutionary hypotheses and new methodologies, as it is an available scenario that can assess the expected performance. This section describes the main types of population genomic simulation processes. These methodologies correspond to two main approaches: coalescence and forward-in-time simulations.

Keeping in mind the description of genetic drift in the previous sections, we can intuit that the mathematical modeling developed by Haldane, Wright and Fisher, as well as Kimura and Otha, was done by imagining a stochastic process that progresses in time. From the binomial sampling in the simplest Wright-Fisher model (like the marble bag examples) to the exploration of diffusion equation that led to the neutral theory, genetic drift is modeled either at a time point t or on a scale $t + i$ (where i is the number of generations). Nonetheless, once DNA alignment became available, the paradigm changed to understand what evolutionary process led to such alignment and what forces played a significant role in it.

In the early 1980s, Kingman (1982) developed coalescence theory, a stochastic retrospective theory that operates at the level of samples, not populations, and from the present backward in time. Coalescence theory describes this backward process regarding the probabilities of lineage pairs coming together (finding common ancestry) in a randomly admixing population (Hejase et al., 2020). Coalescence theory allows us to estimate the expected levels and patterns of genetic variation in a sample of size n, given a stochastic evolutionary model. Coalescence can mathematically describe the most common ancestor of a sample of sequences in a population. Since lineages in the sample that are extinct or not sampled are ignored during the data generation process, coalescent simulations are much more efficient in both time and resources (Hejase et al., 2020). Usually, to generate the observed polymorphism, the mutations are placed along the sampled lineage using Poisson modeling and a constant mutation rate parameter (Kim and Wiehe, 2009). Many extensions have been proposed to overcome the limitations of Kingman's model that initially accounted for Wright-Fisher simplifications (Kim and Wiehe, 2009). Such extensions include general models of recombination, mutation, and demography. Probably the most important and widely used coalescent simulator was `ms`, proposed by Hudson in 2002 (Hudson, 2002). Its successor, `msprime`, dramatically improves computational and storage efficiency by using the tree sequence storage, aforementioned in Section 1.1.4. Other coalescent simulators, such as `discoal` (Kern and Schrider, 2016), `msms` (Ewing and Hermisson, 2010) or `cosi2` (Shlyakhter et al., 2014), can handle selection by conditioning on the trajectory of beneficial alleles (Kim and Wiehe, 2009; Hejase et al., 2020).

Despite the limited options of coalescence simulations regarding selection regimes, they remain helpful because of time and resource efficiency and provide null distributions to predict deviations of neutral theory, which should be caused by natural selection.

While coalescence simulations are an elegant and efficient mathematical framework, forward-in-time simulations are a brute-force attack. Coalescence can reproduce the conditions that lead to specific scenarios. On the contrary, forward-in-time simulations start from an initial model that progresses from generation to generation and finally leads to the model features. Therefore, there is no prior assumption, such as coalescence, but rather the simulation starts with an ancestral population and tracks it forward in time (Hejase et al., 2020). Furthermore, forward-in-time simulations can consider any virtually evolutionary condition, which allows for any combination of selected alleles. This flexibility makes these simulators especially interesting for studying linkage effects, such as recurrent hitchhiking or background selection scenarios.

However, flexibility comes at a cost. Forward-in-time simulations need more resources and time-consuming than any coalescent simulation. Assuming that we need to establish an ancestral population that will evolve, it is not difficult to imagine that population size is one of the limiting factors of these simulations. Moreover, many evolutionary parameters are products of the population size, and complex demography scenarios can be prohibitive, especially when considering complex migrations or exponential growth patterns. A common strategy to solve performance problems has been to rescale the ancestral population size. However, it is not straightforward to determine the stochastic effect of evolution on small populations because rescaling may fail to represent the original population genetics accurately when selection in strong (Uricchio and Hernandez, 2014). Thus, although rescaling is a widely used measure, it should be considered with care and validated against tractable, non scaled scenarios (Hejase et al., 2020; Uricchio and Hernandez, 2014).

Most of the results in Chapters 4 and 5 are based mainly on simulated forward-in-time data, as we tested several conditions based on recurrent positive selection and background selection. Today, several forward-in-time simulators deal pretty well in terms of performance and flexibility. However, each one has its peculiarities, and the choice depends on the user's requirements (Thornton, 2014; Hernandez, 2008; Haller and Messer, 2019). For this thesis, all simulations have been carried out using `SLiM` (Haller and Messer, 2019). To date, `SLiM` is probably the most popular forward-in-time simulator. Not only does it support complex demographic and selection scenarios under different mating and breeding strategies, but it has also been extended to non Wright-Fisher models, allowing the breeding of spatially structured populations. More interestingly, `SLiM` has recently incorporated two crucial capabilities that can significantly increase performance. First, it allows simulations using the aforementioned tree sequences, which improves performance, especially when performing genome-wide simulations. Second, it allows combinations of forward and backward simulations, where `msprime` can be used to generate coalescent histories for a selected prior population forward in time, efficiently generating neutral variation (Hejase et al., 2020).

In the last decade, and especially in the last few years, simulations have played an essential role in developing new methodologies, and without them, two of the most robust and promising methodologies could not have been developed. Both ABC and machine-learning methodologies can bypass direct computation of the likelihood function, which can be challenging to disentangle when considering an accurate number of parameters affecting an evolutionary model. Such methodologies can be exciting in detecting selection across different evolutionary timescales and inferring selection strength, timing, or even recombination since the inference is, *a priori*, unlimited to any parameter combinations producing the simulations. However, the main limitation of both lies in the number of simulations required and the selection of summary statistics that best capture the model. Training machine-learning methods require more realistic models to perform inference and ABC approaches a good understanding of the model to generate the prior distribution properly. Therefore, although unlimited theoretically, both can be highly computationally expensive and, in most, intractable unless using High-Parallel Computing platforms.

1.3 Recurrent positive selection: the McDonald and Kreitman Test

In humans, migrations since the OoA led humans to colonize almost everywhere on Earth, often facing new selective pressures, leading to potential new targets of positive selection. Most statistical measures described in the Section 1.2.6 detect selective sweeps that have not been fixed or have taken relatively little time using genetic diversity data. However, natural selection can act on a larger temporal scale, finally contributing to species differentiation.

Measuring the effect of natural selection at a larger temporal scale can allow us to detect how much positive selection occurred genome-wide since divergence with outgroup species or significant functional differences caused by positive selection events. Similar to hitchhiking, the effect of recurrent positive selection over time can also be captured by unexpected patterns, usually reflected in increased fixation rates. Since most mutations are usually neutral or deleterious considering nearly neutral theory, higher fixation rates will increase the proportion of beneficial substitutions. These signatures can detect natural selection on larger evolutionary time scales, comparing genetic data across lineages. When searching for higher fixation rates, we focus only on the beneficial alleles themselves, testing for the aforementioned directional selection (Sabeti et al., 2006). Nonetheless, multiple recurrent adaptive fixations are required to detect higher substitution rates than the background neutral mutation rate (Hahn, 2018), limiting the power of the analysis. Some clear examples of recurrent adaptive selection are genes involved in gametogenesis or viral interaction proteins (VIPs), which have a high proportion of nonsynonymous substitutions (Nielsen et al., 2005; Bustamante et al., 2005; Enard et al., 2016; Enard and Petrov, 2020).

One of the most straightforward ways to measure the effect of direct selection is through phylogenetic methods. Due to the lack of population data, the divergence accumulation between two orthologs sequences was initially used to infer the rate of adaptation. Because substitutions between species are a long-term consequence of polymorphism within species (Ohta, 1973), these methods are the base defining more sophisticated statistics which use divergence and polymorphism data (Hahn, 2018). As a result, the expected d, the distance between two orthologs sequences is defined as:

$$\mathbf{E}[d] = 2tK \tag{1.8}$$

K represents the average substitution rate of new alleles across sites, and t is the divergence time between species. Note that the genetic distance is measured using $2t$ since the substitutions can occur on both branches. Therefore, species are differentiated depending on the number of alleles that appear and are fixed and the time of separation between species. As previously explained (Section 1.2.1), considering neutral mutations, the fixation probability depends only on the initial frequency ($n/2N$), and the rate at which neutral mutation arises will only depend on the mutation rate, following $K = \mu$. Then, the substitution rate is equal to the mutation rate, regardless of population size. K represents the average substitution rate of new alleles across sites, and t is the divergence time between species.

Genetic distance is measured empirically by aligning two orthologous sequences and counting the number of differences between them. However, two factors can affect the measurement of genetic distance. First, the presence of derived alleles which may be polymorphic in the ancestral population is usually ignored. Such presence can inflate the number of sites that we consider divergent, especially when considering a closely related outgroup. Thus, closely related outgroups can potentially bias estimates of the rate of adaptive substitutions due to shared polymorphisms. On the other hand, if we consider that a site may be subject to more than one change from divergence, we might underestimate the actual number of differences between the two sequences. Finally, supposing that the divergence between the outgroup and focal species is too high, we may suffer the same bias as phylogenetic methods toward the most conserved genes, as rapidly evolving genes will not produce reliable sequence alignments. Keightley and Jackson (2018) show that these limitations can be overcome by using the multiple outgroup species, spanning multiple levels of divergence, and extracting local substitution rate information (Moutinho et al., 2019a). In the most extreme case, when each site has changed several times, we would still have 25% of the bases matching at random and $d = 0.75$. d has been widely explored to account for multiple substitutions. Among the most widely used models which parameterize the substitution process, we find those proposed by Jukes and Cantor (1969), Kimura (1980), Felsenstein (1981) or

Tamura and Nei (1993). The first correction was proposed by Jukes and Cantor (1969), where d becomes:

$$d = -\frac{3}{4} ln(1 - \frac{4}{3}a) \tag{1.9}$$

and a is the count of divergent sites. There are several methods to estimate the strength and direction of selection using between-species sequence data. The genetic distance can be used to perform the d_N/d_S ratio estimation (also referred to as Ka/Ks ratio, or simply ω), one of the classical approaches to test the direction of selection. d_N/d_S approach is defined by the number of replacement substitutions (D_N) per non-synonymous site (d_N), and the number of silent substitutions (D_S) per synonymous site (d_S). Since the mutation rate varies throughout a genome, Kimura (1977) suggested correcting D_N with D_S, equivalent to controlling the differences in neutral mutation rates. d_N/d_S ratio can indicate the general impact of natural selection on a sequence, but include the combined effect of neutral, advantageous, and deleterious mutations. For phylogenetic and population genetic analyses, divergence is one key parameter and must be estimated as accurately as possible. Therefore, erroneous divergence measures would affect subsequent estimates of adaptive substitutions using phylogenetic or population genetic methods.

dN/dS is expected to be 1 when changes have been selectively neutral during the evolution of the sequence. A d_N/d_S ratio higher than 1 is indicative of recurrent positive selection. However, a ratio above 1 can only be obtained if a considerable fraction of the mutations were advantageous, and only a few genes will ever reach a d_N/d_S higher than 1. Thus, since most mutations are deleterious following neutral and nearly theory, the ratio usually is $d_N/d_S < 1$. ω is a first approximation to pin-point putative genes or positions under the action of natural selection, measured through a pattern statistically incompatible with neutral theory.

1.3.1 McDonald and Kreitman Test

The McDonald and Kreitman Test (MKT) is an alternative to the ω estimate for the detection of positive selection, which takes advantage of phylogenetic and population information. It can detect the action of recurrent positive selection by analyzing polymorphism and divergence data altogether. MKT covers the evolutionary period spanning from the divergence of the outgroup species to the present.

The original MKT (McDonald and Kreitman, 1991) is one of the most powerful and robust methods we have to detect the action of natural selection at the DNA level. Polymorphic data correct for purifying selection on divergent non-synonymous sites, significantly increasing the detection power of recurrent positive selection. Four different counts are needed to conduct the MKT: the count of polymorphisms at synonymous (P_S) and non-synonymous sites (P_N), as well as the count of substitutions at synonymous (D_S) and non-synonymous sites (D_N). The four counts are placed in a 2×2 contingency table to test the null hypothesis of neutrality. Under neutrality, all non-synonymous mutations are expected to be neutral, and the D_N/D_S ratio will be roughly equal to the P_N/P_S ratio.

Because infrequent adaptive mutations fix fast relatively to common neutral mutations, they contribute almost exclusively to divergence and not to polymorphism; therefore, an excess of the divergence ratio relative to polymorphism can be interpreted as positive selection signals. Considering neutrality, the resulting 2×2 will show no deviation of the neutral expected ratio. The significance is commonly assessed through a Fisher Exact test or a chi-square test.

Several parameters have been derived to quantify the amount of selection using the MKT, such as the Neutrality Index (NI) (Rand and Kann, 1996) or the Direction of Selection (DoS) index (Stoletzki and Eyre-Walker, 2011). NI indicates to what extent the polymorphism to divergence ratio in the testing region departs from the expected under the neutral model. Under neutrality, P_N/P_S equals D_N/D_S, and thus NI equals 1. NI below 1 can be interpreted as an excess of divergence between species due to adaptive selection. NI above 1 is interpreted as an excess of polymorphic variation compared to neutral regions, which can be interpreted as evidence of purifying selection.

$$NI = \frac{P_N/P_S}{D_N/D_S} \tag{1.10}$$

The DoS statistic, proposed by Stoletzki and Eyre-Walker (2011), is an unbiased metric calculated as

$$DoS = \frac{D_N}{D_N + D_S} - \frac{P_N}{P_N + P_S} \tag{1.11}$$

Positive values of DoS show evidence of adaptive evolution at nonsynonymous sites, whereas negative values indicate negative selection. Because NI statistic is estimated as a ratio of two ratios, they tend to be biased and to have a large variance towards negative values, specially when data is sparse (any count < 5) (Stoletzki and Eyre-Walker, 2011).

The most popular summary statistic derived from MKT is the proportion of substitutions that have been fixed by adaptive evolution: α (Charlesworth, 1994; Smith and Eyre-Walker, 2002).

$$\alpha = 1 - \frac{D_S}{D_N}\frac{P_N}{P_S} \tag{1.12}$$

α have been widely used during the last decade to test regions and genes where natural selection would hypothetically act, as well as the frequency of adaptive mutations along the genome. Most existing approaches to computationally estimate the fraction of non-synonymous substitutions (α) derive from MKT and Poisson Random Field (PRF) frameworks (Sawyer and Hartl, 1992), both of which use divergence and polymorphism data to infer the adaptation rate (see Figure 1.10). PRF derive P_N, P_S, D_N and D_S modeling the mutation process, selection and genetic drift at evolving sampled independent sites, in addition to the associated population-scaled selection coefficient ($\gamma = 2N_e s$) and population size (N_e) (Moutinho et al., 2019a). PRF approach is the base model for multiple statistical approaches trying to estimate the proportion of adaptive substitutions, including Maximum Likelihood (ML) estimations of the DFE (see Section 1.3.4), but also Bayesian models to infer the population-scaled selection coefficients (Bustamante et al., 2002b; Sawyer et al., 2003)

Nonetheless, MKT and PRF-derived approaches have several drawbacks that could finally bias the estimation. First of all, it assumes the strict neutrality of segregating sites. However, several studies in multiple species have shown that selected mutations could be drawn following different forms of the DFE, resulting in unexpected patterns in the polymorphism ratio. A clear example could be the genomes where weak negative selection abounds (Casillas and Barbadilla, 2017). In these cases, where natural selection is not efficient purguing deleterious mutations, the SFS tends to accumulate SDM at low frequencies. Another of the most unrealistic assumptions of the MKT is that the neutral mutation rate is constant over time, and so is the selective constraint. However, the neutral mutation rate is heavily affected by changes in the effective population size (Balloux and Lehmann, 2012; Lanfear et al., 2014; Galtier and Rousselle, 2020; Rousselle et al., 2020). For example, suppose a population that has been expanding. In that case, slightly deleterious substitutions can lead to an overestimation of α as they could have been fixed in the past (thus, contributing to divergence) due to the larger impact of genetic drift in small populations (Eyre-Walker and Keightley, 2009). Another illustrative example of why neither the neutral mutation rate nor the selective constraint is constant over time is the trajectory of newly duplicated genes. For a newly duplicated gene, the strength of selection is initially relaxed, and then it might become under selective constraint if a new function is acquired. The strength of selection may also fluctuate over time in single-copy genes. In these cases, the MKT results can be misleading. However, this effect is expected in general to be negligible in the MKT for single genes because the fluctuations in the selection strength over time should not have directionality (Fay et al., 2001).

Besides, there is evidence that weakly advantageous mutations are segregating at the SFS (Galtier, 2016; Tataru et al., 2017; Uricchio et al., 2019). The presence of this non-neutral polymorphism can mask the effect of adaptive selection, as it acts in the opposite directions in the MKT. Lastly, as recently described at (Uricchio et al., 2019), the patterns of nucleotide diversity can be affected by hitchhiking or background selection, leading to patterns of fixation not assumed by strict neutrality due to linkage with slightly deleterious segregating alleles.

Over the last decades, several modifications of the original MKT have been proposed to account for the potential biases in estimating the proportion of adaptive substitutions (α). However, as aforementioned, other forces are affecting the SFS (such as recombination, demography, ancestral population sizes or weak adaptation) and several studies have shown the importance of dealing with slightly deleterious alleles (Fay et al., 2001; Mackay et al., 2012; Eyre-Walker and Keightley, 2009; Messer and Petrov, 2013b). Therefore, the presence of slightly deleterious mutation segregating and the subsequent distortion of the SFS have been repeatedly shown to be one of the essential factors biasing downwards biasing α.

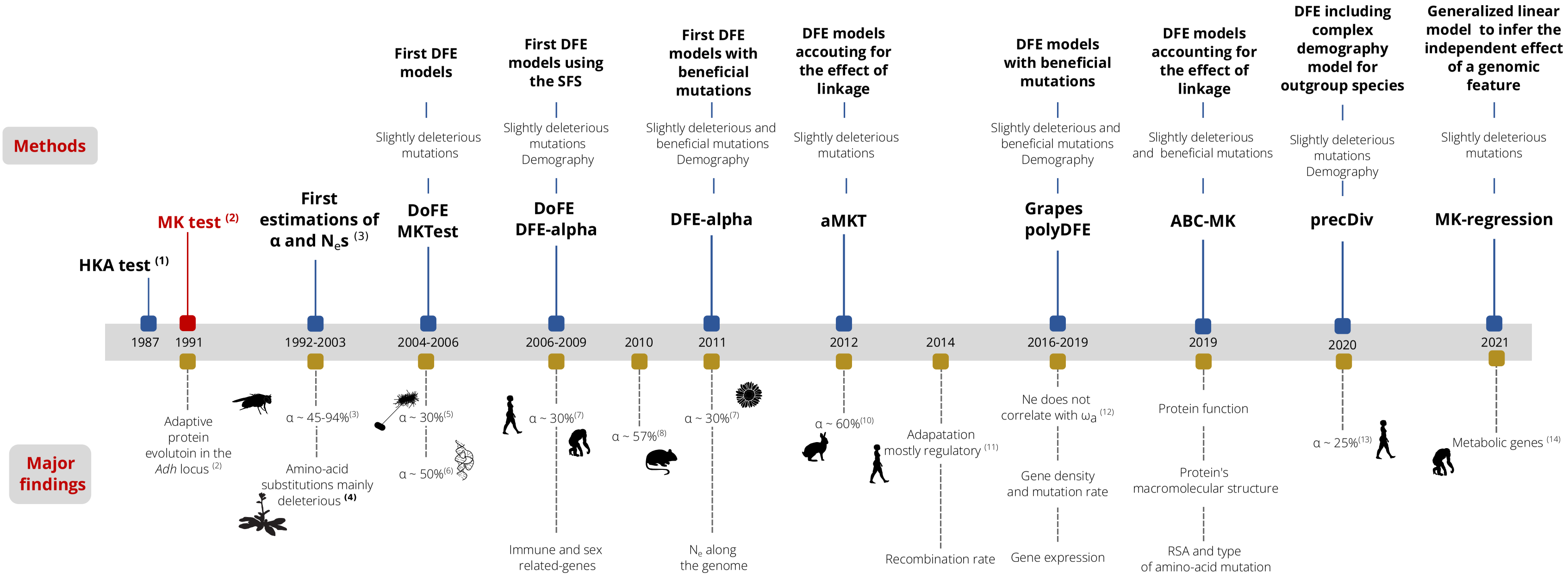

Methods
Major findings
HKA test (1)
MK test (2)
First estimations of α and N_es (3)
First DFE models
Slightly deleterious mutations
DoFE MKTest
First DFE models using the SFS
Slightly deleterious mutations Demography
DoFE DFE-alpha
First DFE models with beneficial mutations
Slightly deleterious and beneficial mutations Demography
DFE-alpha
DFE models accouting for the effect of linkage
Slightly deleterious mutations
aMKT
DFE models with beneficial mutations
Slightly deleterious and beneficial mutations Demography
Grapes polyDFE
DFE models accounting for the effect of linkage
Slightly deleterious and beneficial mutations
ABC-MK
DFE including complex demography model for outgroup species
Slightly deleterious mutations Demography
precDiv
Generalized linear model to infer the independent effect of a genomic feature
Slightly deleterious mutations
MK-regression
1987
1991
1992-2003
2004-2006
2006-2009
2010
2011
2012
2014
2016-2019
2019
2020
2021
Adaptive protein evolutoin in the Adh locus (2)
α ~ 45-94% (3)
Amino-acid substitutions mainly deleterious (4)
α ~ 30% (5)
α ~ 50% (6)
α ~ 30% (7)
Immune and sex related-genes
α ~ 57% (8)
α ~ 30% (7)
N_e along the genome
α ~ 60% (10)
Adapatation mostly regulatory (11)
Recombination rate
Ne does not correlate with ωa (12)
Gene density and mutation rate
Gene expression
Protein function
Protein's macromolecular structure
RSA and type of amino-acid mutation
α ~ 25% (13)
Metabolic genes (14)

Figure 1.10: Timeline of MKT extensions to infer α and the major findings on the factors impacting the variation of the molecular adaptive rate. α: proportion of adaptive amino-acid substitutions; N_e: effective population size; s: selection coefficient; ω_a: rate of adaptive non-synonymous substitutions; RSA: relative solvent accessibility. (1) Hudson et al. (1987); (2) McDonald and Kreitman (1991); (3) Sawyer and Hartl (1992); Charlesworth (1994); Fay et al. (2001); Smith and Eyre-Walker (2002); Bustamante et al. (2002b); Sawyer et al. (2003); (4) Bustamante et al. (2002b); (5) (Charlesworth and Eyre-Walker, 2006); (6) Williamson (2003); (7) Hvilsom et al. (2012); Eyre-Walker and Keightley (2009); (8) Halligan et al. (2010); (9) Gossmann et al. (2010); Strasburg et al. (2011); (10) Carneiro et al. (2012); (11) Enard et al. (2014); (12) Galtier (2016); (13) Zhen et al. (2021); (14) Huang (2021). Figure adapted from Moutinho et al. (2019a)

1.3.2 Heuristic extensions

The very first correction to the MKT approach was suggested by Templeton (1996). He suggested dealing with the presence of SDM, extending the contingency table to a 3×2 contingency table. The resulting table divided the polymorphic counts into singletons and multitons. Singleton categories are expected to be overrepresented due to the presence of SDM. While there are neutral nonsynonymous polymorphisms at low frequency, under neutrality there should also be a proportional number of synonymous polymorphisms at the same frequency. Later, Akashi (1999) proposed a more robust method that considers the complete distribution of allele frequencies rather than only multitons and singletons. Although both extensions are more powerful than the original MKT approach, their results are challenging to interpret when the ratio of non-synonymous to synonymous differences varies among allele frequency classes (Hahn, 2018).

fwwMKT. Fay et al. (2001) developed a straightforward methodology that removes all polymorphisms segregating at a frequency below a given threshold (normally 5%–15%). Although there is no consensus about the exact value of the this threshold, (Charlesworth and Eyre-Walker, 2008) explored the MKT and concluded that α estimates are robust using a frequency threshold of 15%, below which most slightly deleterious polymorphisms are found and removed. These estimates are reasonably accurate only when the rate of adaptive evolution is high, and the Distribution of Fitness Effects (DFE) of deleterious mutations is leptokurtic (Charlesworth and Eyre-Walker, 2008). α is estimated using the standard MKT equation, but considering only those polymorphic sites (for both neutral and selected classes) whose counts are above the established frequency j.

$$\alpha_{FWW} = 1 - \left(\frac{P_{N(j>15\%)}}{P_{S(j>15\%)}} \cdot \frac{D_S}{D_N} \right) \tag{1.13}$$

eMKT. Mackay et al. (2012) proposed the extended MKT (eMKT). Instead of simply removing low-frequency polymorphism below a given threshold, the count of segregating sites in non-synonymous sites is partitioned in the number of neutral variants (using neutral sites as a proxy) and the number of weakly deleterious variants. This increases the power of detecting adaptive selection (as it does not remove as much data as the fwwMKT) and allows the independent estimation of both adaptive and weakly deleterious substitutions. P_N, the count of segregating sites in the non-synonymous class, is decomposed into the number of neutral variants and the number of weakly deleterious variants, $P_N = P_{N_{neutral}} + P_N$. The estimation of both numbers allows estimating positive (adaptive) and negative selection independently. α is estimated from the standard MKT table discounting weakly deleterious variants: P_N is substituted by the expected number of neutral segregating sites, $P_{N_{neutral}}$. The corrected estimate of α is then

$$\alpha_{extended} = 1 - \left(\frac{P_{Nneutral}}{P_S} \cdot \frac{D_S}{D_N} \right) \tag{1.14}$$

The fraction of sites segregating neutrally ($P_{N_{neutral}}$) is estimated through the neutral polymorphic ratio given the frequency threshold (15%) over the SFS frequencies (j)

$$\hat{f}_{neutral\ j<15\%} = \frac{P_{S(j<15\%)}}{P_S} \tag{1.15}$$

Therefore, the expected number of segregating sites in the non-synonymous class neutral evolving given the threshold is

$$P_{N_{neutral}<15\%} = P_N \cdot \hat{f}_{neutral\ j<15\%} \tag{1.16}$$

Despite being inspired in the fwwMKT, since $P_{N_{neutral}}$ is considered a fixed proportion of nearly neutral variants given the synonymous count, this eMKT assumption result in associated biased estimations, and the amount of bias directly depends on the selected frequency cutoff (see Chapter 4).

aMKT. Messer and Petrov (2013b) proposed the asymptotic MKT. This MKT extension is robust to the presence of selective sweeps and to the segregation of slightly deleterious polymorphism. In this approach, the authors defined α as a function that depends on the SFS of alleles. Therefore, α is estimated in different frequency intervals. Given the frequency spectrum distribution in the frequency interval $[0, 1]$, the estimate of α_x results in an exponential function of the form. $\alpha_{(x)} = a + b \cdot e^{-cx}$. The best fit of the exponential at $x = 1$, eliminates the effect of a slightly deleterious allele. The exponential fit is suitable as the non-synonymous allele frequency is expected to decay exponentially over the respective levels of synonymous polymorphisms, since the fixation probability differs highly from the neutral alleles (Messer and Petrov, 2013b).

$$\alpha_{(x)} = 1 - \frac{D_S}{D_N} \cdot \frac{P_{N(x)}}{P_{S(x)}} \tag{1.17}$$

aMKT does not assume that sites evolve independently and do not require to invoke demography . Although they showed the approach is robust to both the underlying DFE and recent demographic events when selection is strong, the method misses main scenarios: i) BGS; ii) it does not account for the presence of weakly beneficial alleles (Tataru et al., 2017).

1.3.3 ABC-MK

Uricchio et al. (2019) presented the ABC-MK approach to overcome the main limitation presented at the aMKT. They hypothesized that a method considering the effect of weakly beneficial alleles could be developed by exploiting the impact of BGS on the fixation rate. ABC-MK interrogates α as a function of BGS to infer the rate of adaptation and strength of beneficial alleles jointly. To test pros and cons of aMKT, the authors explored analytical theory to investigate α when adaptation is strong and weak while accounting for BGS. They contemplate weakly selected polymorphism segregating at the SFS, slightly deleterious alleles, BGS strength, and weakly beneficial fixations altogether with strong beneficial alleles. The authors demonstrate that aMKT is strongly biased when weakly beneficial alleles contribute substantially to segregating polymorphism across the genome as a function of the strength of BGS (Uricchio et al., 2019). Nonetheless, since demography impacts the SFS, it is not straightforward to calculate the SFS through analytical theory, including generalized selection, demography, and linkage models. To exploit the explored co-variation between α and BGS from actual data, the authors performed an Approximate Bayesian Computation (ABC) method to relax most of the analytical assumptions and separately infer the rate and strength of adaptation (Uricchio et al., 2019). The authors followed a generic ABC algorithm in which they i) run forward simulations with a fixed DFE over non-synonymous mutations accounting for empirical BGS values and known demography to simulate the model; ii) estimate informative summary statistics from forward simulations following a biased-resampling strategy that avoids simulating the full model for different parameter combinations; iii) supply thousands sets of summary statistics corresponding to parameters sampled from prior distributions into a published ABC framework (Thornton, 2009) to infer parameters.

1.3.4 ML models of the DFE

In addition to the heuristic MKT extensions and ABC-MK, ML models of the DFE that assume PRF framework can estimate the expected proportion of adaptive fixations given the inferred DFE from the MKT data. In such approaches, the expected levels of fixations and polymorphism are used to perform likelihood estimates, while considering different evolutionary models. ML estimations of the DFE have varied from the first models inferring constant selection parameters across all loci to including models with continuous distributions of both positive and negative selection coefficients (Bierne and Eyre-Walker, 2004; Eyre-Walker et al., 2006; Boyko et al., 2008; Eyre-Walker and Keightley, 2009; Galtier, 2016; Racimo and Schraiber, 2014; Galtier, 2016; Tataru et al., 2017; Zhen et al., 2021), correcting the aforementioned assumptions to calculate how many non-adaptive substitutions are expected to become fixed given the empirical DFE.

Notwithstanding, the newest and more sophisticated ML implementations (Galtier, 2016; Tataru et al., 2017) take advantage of that proposed by Eyre-Walker et al. (2006) and Eyre-Walker and Keightley (2009), which assumes that non-neutral deleterious mutations arise from a DFE in the form of a Gamma distribution. Nonetheless, unlike Eyre-Walker et al. (2006), these methods also model the effect of weakly advantageous alleles through an exponentially distributed function, where DFE is a mixture distribution between the gamma and exponential distributions. Therefore, the state-of-the-art methods usually follow a standard population genetic model based on PRF presented in Galtier (2016) and Tataru et al. (2017) to later perform ML of the DFE.

To estimate the expected counts of polymorphism and divergence, the model considers a Wright-Fisher panmictic population of size N_e, which diverged in a time t where mutation occurs at a mutation rate μ, per site per generation (Galtier, 2016). Figure 1.11 illustrate the equations defining the expected polymorphic and divergence counts. Let consider synonymous mutation as neutral (P_S) and non-synonymous mutation as selected (P_N), from PRF the expected counts given a frequency i is estimated as

$$\mathbf{E}[P_{S[i]}] = \frac{4N_e\mu L_S}{i} \tag{1.18}$$

$$\mathbf{E}[P_{N[i]}] = 4N_e\mu L_N \int_0^1 B(i,n,x)H(s,x)dx \tag{1.19}$$

where L_N and L_S are the total number of synonymous and nonsynonymous sampled alleles, $B(i,n,x)$ is the probability of observing a mutation at frequency i in n sequences when the true allele frequency is x

$$B(i,n,x) = \binom{n}{i} x^i(1-x)^{n-i} \tag{1.20}$$

and $H(s,x)$ is the time that a new semi dominant mutation with selection coefficient s (in the heterozygous) spends between the frequency x and the frequency $x+dx$ from diffusion theory (Wright, 1938)

$$H(s,x) = \frac{1-e^{-s(1-x)}}{x(1-x)(1-e^{-s})} \tag{1.21}$$

Note that to obtain the expected polymorphic count given the underlying DFE of new mutations equation (1.19) should be integrated over the full DFE. Following (Eyre-Walker and Keightley, 2009), the underlying DFE for new deleterious mutations is defined by expression ϕ

$$\phi(s;a,b) = a^b s^{b-1}\frac{e^{-as}}{\Gamma(b)} \tag{1.22}$$

where a and b are scale and shape parameters from the Gamma distribution. Therefore, the expected polymorphic count given a particular DFE is defined as:

$$\mathbf{E}[P_{N[i]}] = 2N_e\mu L_N \int\limits_{-\infty}^{\infty}\int\limits_{0}^{1} B(i,n,x)H(s,x)\phi(s;\alpha,\beta)dxds \tag{1.23}$$

In addition, the method proposed by Galtier (2016) makes the DFE more flexible by considering that the most appropriate way to model the full DFE may not be the classical Gamma distribution over negative alleles. Therefore, Galtier's modeling included two different versions of the Fisher's geometric model, and a model assuming a Beta-shaped distribution of weak effect mutations, instead of a Gamma distribution. However, as explored further in Galtier (2016), Galtier and Rousselle (2020) and Rousselle et al. (2020), the shifted negative modeling and the two DFE models based on the Fisher geometric model generally do not perform well in the analysis.

Following Galtier (2016) and Tataru et al. (2017) procedure, the expected number of synonymous (D_S) and non-synonymous (D_N) substitutions can be estimated as:

$$D_S = L_S\mu t \tag{1.24}$$

$$\mathbf{E}[D_N] = 2N_e\mu L_N \int\limits_{-\infty}^{\infty} \frac{2s}{1-e^{(-4N_e s)}}\phi(s)ds \tag{1.25}$$

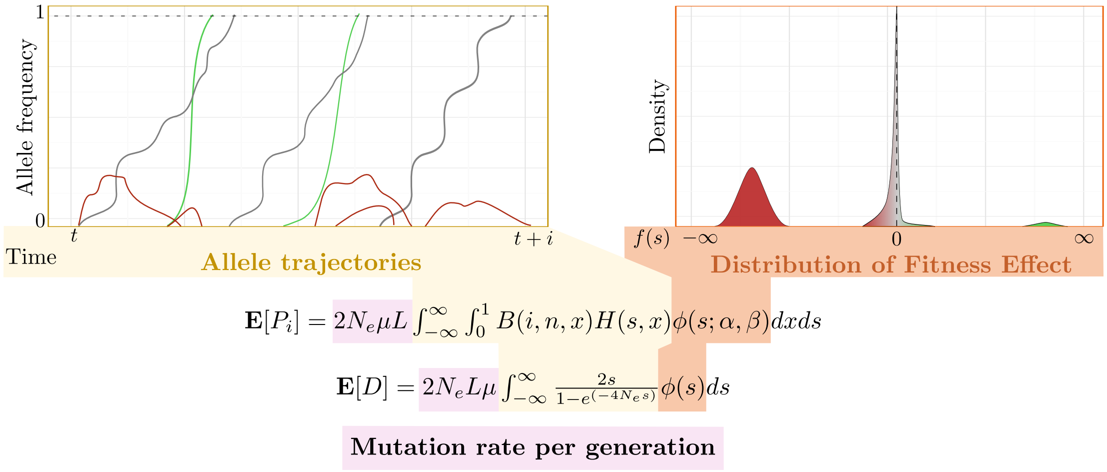
The proportion of adaptive mutations (α) is estimated using the ML inference of DFE parameters and the expected counts of non-adaptive mutations. α therefore is estimated subtracting the non-adaptive substitutions and neutral substitutions from the total observed divergence counts at selected sites.

$$\alpha = (d_N - d_N^{na})/d_N) \tag{1.26}$$

note that d_N^{na} is defined following Galtier (2016). The equation decomposition is similar to the equations shown at (Tataru et al., 2017) and (Eyre-Walker and Keightley, 2009)

$$d_N^{na} = \frac{2L_N N_e \mu \int_{-\infty}^{s_{adv}} \frac{2s}{1-e^{(-4Nes)}} \phi(s) ds}{L_N} \tag{1.27}$$

To estimate the non-adaptive substitutions, the approach proposed by Eyre-Walker and Keightley (2009) and Tataru et al. (2017) integrate the DFE from $-\infty \to 0$ ($s_{adv} = 0$), taking into account the deleterious fixations and the negative nearly neutral fixations given the DFE. Galtier (2016) considers positive s_{adv} values to subtract nearly neutral positive fixations given the s_{adv} threshold too.



$$\mathbf{E}[P_i] = 2N_e\mu L \int_{-\infty}^{\infty}\int_0^1 B(i,n,x)H(s,x)\phi(s;\alpha,\beta)dxds$$

$$\mathbf{E}[D] = 2N_eL\mu \int_{-\infty}^{\infty} \frac{2s}{1-e^{(-4N_es)}}\phi(s)ds$$

Figure 1.11: Expected polymorphic and divergence sites based on PRF approach. Note that for $s = 0$ the equations became similar to equations (1.18) and (1.24). Adapted from Casillas and Barbadilla (2017).

It is important to emphasize that the shape of the SFS may be biased and differ from the previously expected values. The distortion may mainly be due to the effect of demographic events. To account for such distortions, the different methods usually incorporate nuisance parameters following (Eyre-Walker et al., 2006). The nuisance parameter r modifies each frequency of the SFS individually. Thus, r modifies the effective mutation rate at each frequency i, considering the relative mutation rate at i with respect to the mutation rate at singletons Eyre-Walker et al. (2006). The perturbing parameter r_i considers the same amount of distortion between nonsynonymous and synonymous SFS. Although the assumption is unrealistic, simulations showed that the original correction proposed by Eyre-Walker et al. (2006) is robust enough to take into account demographic effects. However, other models correct for demographic effects by considering explicit changes in the effective population during the modeling process (Eyre-Walker and Keightley, 2009; Zhen et al., 2021).

Despite the efforts for modeling the DFE using polymorphic data, Booker (2020) suggested that inferred parameters from ML approaches must be reviewed. Underlying assumptions regarding the DFE shape, selection coefficient strengths, or population sizes can affect the estimations and capture distinct aspects of the DFE (Booker, 2020; Zhen et al., 2021). Such a situation is plausible in several *D. melanogaster* studies. For example, Campos et al. (2017) and Keightley et al. (2016) found differences between selection coefficients and probabilities of beneficial alleles. As extensively discussed in Booker (2020), it is plausible due to different methodological assumptions or if the DFE for advantageous mutations is bimodal. This latter case should be especially considered for strongly beneficial alleles, which would be undetectable by analyzing the unfolded SFS (uSFS) (Booker, 2020).

Messer and Petrov (2013b) compared the performance of their method -the aMKT- and the `DFE-alpha` method through simulations. They claimed that DFE-alpha correctly estimated when the model allowed for population size change, but the demography inferred was found to be biased, mainly due to background selection acting at linked sites. Genetic draft leaves signatures in the SFS similar to those observed under a recent population size expansion. The DFE-alpha method inferred systematically a population expansion even though no expansion was set in the simulation . Another limitation of PRF approaches is that it becomes computationally intensive, especially when a change in demographic model is applied. Since DFE-alpha can only consider two population-size changes, it becomes insufficient for capturing the excess of rare variants due to the complex demographic history of some populations, like the human history (Zhen et al., 2021).

1.4 Surveys of positives selection

1.4.1 Surveys on candidate genes

Before the advent of genome-wide catalogs of variation, we had only a bunch of clear examples of the action of natural selection in DNA sequences, and the inference required an *a priori* hypotheses. Moreover, the absence of functional information regarding the non-coding part of the genome limited such hypotheses to a few candidate genes, where the action of natural selection was plausible usually due to the presence of some phenotypic evidence. Considering that the human species has been able to colonize the globe and live in absolutely different environments, it could be assumed that the number of selective hypotheses to test would be directly related to each population's differences and environmental peculiarities. However, while such a statement may be true, it is not easy to focus on the right questions, and solving them poses the challenge of linking a phenotypic trait to a single causal gene. Hence, the number of candidate genes under natural selection has permanently been reduced. Despite the lack of information regarding potential candidate genes, the appearance of molecular data allows inferring the action of natural selection and to carry out thorough studies. The data allows linking genetic information and phenotypic evidence and includes the historical moment in which natural selection took place. Therefore, genetic evidence has provided new insight into human history in conjunction with archaeological and historical data. Classic examples include the gene associated with lactose tolerance in adulthood and genes that reduce susceptibility to malaria infection (Tishkoff et al., 2001; Bersaglieri et al., 2004). Both examples neatly illustrate how molecular data allowed us to search and find evidence for the action of natural selection.

In the case of malaria, numerous genes have been described over the last decade that would likely confer resistance to infection. In 1954, the study performed by Allison (1954) showed the correlation between the geographical distribution of sickle cell disease and malaria endemicity in Africa (Tishkoff et al., 2001). The sickle cell disease is produced by an amino acid change at β-globin gene (HBB). Although the mutation is strongly deleterious in homozygous, it is prevalent in regions where malaria is endemic. That prevalence suggests that heterozygous carriers are protected against *Plasmodium falciparum* malaria, and the mutation is maintained in the population by the action of natural selection. The study at the population-genetic level allowed inferring the action of selection on the malaria resistance trait. Currat et al. (2002) and Ohashi et al. (2004) showed the first evidence of natural selection at the HBB locus. Polymorphic data obtained from marker genotyping and chromosomal sequencing allow both studies to prove the origins of the haplotypes carrying the sickle cell mutation in populations of different ancestry but show unexpected patterns of linkage disequilibrium and longer haplotypes. The results showed that mutations in the HBB groups had recently occurred, hence the rapid increase in allele frequency, dating back 2000 years.

The first population-genomic results provided on malaria resistance candidate genes provided a holistic perspective of the adaptation of a human trait considering the origin of the sickle cell mutation in response to malaria (Sabeti et al., 2007). Moreover, the estimates of the mutation age are fully compatible with theories of the recent expansion in the human population, whereby the advent of agriculture led to the population densities necessary for the efficient spread of malaria (Fan et al., 2016).

Nonetheless, the survey of HBB genes is not the only example of malaria candidate genes under natural selection. Tishkoff et al. (2001) reconstructed the evolutionary history of the G6PD genes. Like in HBB clusters, G6PD presents highly deleterious variants resulting in enzyme deficiency associated with specific geographic regions highly correlated to the distribution of malaria endemicity (Tishkoff et al., 2001). The linkage disequilibrium patterns allow the reconstruction of haplotype diversity and date the appearance of the enzyme deficiency associated variants about 11000 and 6000 years ago (depending on the studied haplotype).

In the case of lactase persistence, the ability to digest lactose in adulthood varies among populations and is also genetically determined (Bersaglieri et al., 2004). Moreover, the geographic distribution of tolerance ratios coincides with the historical occurrence of livestock domestication across the globe, including today's major production regions (Beja-Pereira et al., 2003). These signals prompted early studies and hypotheses proposing that the persistence of lactose tolerance, especially in northern European populations, could be explained by the action of positive selection. Bersaglieri et al. (2004) performed the first population-genetics approach to assess this question. They surveyed the role of natural selection over 100 genotyped SNPs in multiple populations, using an F_{ST} and an F_{ST} extension measure. The approach showed that: i) SNPs near LCT show significant differences in allele frequencies among populations; and ii) the putative haplotype associated to persistence is longer than would be expected in the absence of selection (Bersaglieri et al., 2004).

These first studies surveying targets of selection using molecular data showed the power of these data to solve evolutionary questions. High-resolution LD maps allow detecting recent adaptation events along the human genome (Akey et al., 2002). Such examples of genes continue to be the starting point for new studies and methodologies. Over the years, we have succeeded in inferring the strength, direction, and causes and effects of natural selection. Nonetheless, as extensively discussed in Akey (2009), the study of natural selection over candidate genes has two significant issues. In the first place, the pinpointing of candidate genes requires an *a priori* hypothesis. Natural selection acts on phenotypes, which ultimately shapes the genetic variants in populations. However, even in these cases where the hypothesis is elaborated directly on scientific evidence on phenotypic changes, it is necessary to understand genotype-phenotype relationships to pinpoint specific genes. In practice, only Mendelian architecture traits are susceptible to directly linking genotype and phenotype. Such is the case for the lactose tolerance trait. Moreover, Mendelian architecture traits were commonly associated with functional protein-coding genes ignoring important traits at regulatory levels.

Second, variation in DNA may not only occur through the presence of positive selection. There are a variety of factors that can lead to misinterpretation of putative selection signals (Akey, 2009; Booker et al., 2020; Schrider, 2020). The interplay of forces shaping genetic variation and natural selection can result in patterns similar to those that *a priori* could be detected as natural selection. Of particular relevance is the role of genetic drift and the demographic history of populations (Schraiber and Akey, 2015), since diversity patterns are affected by changes in population size. Nonetheless, other factors such as nonrandom mating and admixture should be considered too when scanning patterns of adaptation (Akey, 2009). Therefore, the genetic variation at candidate genes has to be reviewed in their genomic context to make robust inferences of positive selection while understanding realistic models of population history, recombination, and other selective regimes.

1.4.2 Genome-wide catalogs of positive selection

Starting in the last decade, genomic data is allowing us to unravel the interplay between the different forces that modulate genetic variation. Its high resolution is especially helpful in distinguishing between demographic history and natural selection to develop a coherent narrative of human evolutionary history. On the one hand, it has allowed to have a clear view of the demographic history of anatomically modern humans: the time of emergence as species ($\approx 200,000$ years ago, probably in East Africa), the routes and migratory events along our colonization of the globe (Schraiber and Akey, 2015). On the other hand, genomic data has allowed us to elucidate the action of natural selection on classic examples of candidate genes at different levels, unraveling when and how natural selection action took place. More importantly, genomic data allows us for the first time to survey the genome without any a priori hypothesis searching for positive selection patterns. This way, genome-wide scans of selection have resulted in an extensive catalog of putatively adaptive regions.

Considering the history of anatomically modern humans, demography is especially relevant for several reasons: i) population expansions from OoA dispersal could be one reason for the accelerated rate of adaptation, facing humans to new selective pressures; ii) migration and isolation can determine local events of adaptation and population differentiation; iii) population contraction can be directly associated with selective pressure, making the two forces that shape natural variation indistinguishable (Lohmueller and Nielsen, 2021). In addition, it is well-documented that complex demography history can mimic variation patterns produced by hitchhiking and background selection (Schraiber and Akey, 2015).

From early analyses accounting for very sparse maps of genetic variation (Akey et al., 2002; Payseur et al., 2002), genome-wide scan analyses of selection exploded in the mid-2000s. The advent of Perelegen and HapMap data (Altshuler et al., 2005; Hinds et al., 2005) provided the opportunity for studies at different scales, from which a large number of genome-wide catalogs of recent and ongoing selection emerged. As a result, we now have numerous genome-wide selection scans in the human genome, including different populations and signals, and more comprehensive datasets than Perlegen or HapMap (Table 1.1). The identification of candidate regions has demonstrated the ubiquity of selection along the genome. Some of the selection events detected in humans, as in other species, are related to clear examples of selective pressure. The genomic survey has repeatedly shown adaptive signals related to pathogens, high altitude, toxic environments, change in diets, ultraviolet exposure, or even noninfectious genetic diseases (Fan et al., 2016). This relationship between putatively selected variants and identified forces of selective pressure represents an essential point for understanding species differentiation and history because phenotypes are the primary target of selection that alleles putatively selected are likely to have functional relevance (Akey, 2009).

The outlier approach is a primary methodology to carry out genome-wide analysis detecting the potential evidence of natural selection. In this type of analysis, many loci, or the whole genome, are sampled to calculate summary statistics. Since natural selection will act over specific regions or genes, a putative candidate will appear at the extreme tail when constructing an empirical distribution. Thus, outlier approaches can be a straightforward methodology to address this unsolved problem considering candidate genes analysis because natural selection will operate under specific functional loci, while confounding factors, such as demography, would affect variation along the genome similarly (Akey, 2009).

Figure 1.12 represents graphically the application of the outlier approach to a general summary statistic. A standard population genomic survey will sample i loci to estimate a summary statistic T. Under neutrality, genetic drift is the leading force affecting the loci i. Although other genetic forces can affect the expected patterns, such as background selection or local recombination rates, the effect is ubiquitous along the genome. Therefore, the T_i estimation will equally capture the forces affecting the surveyed loci. Nonetheless, variants under positive selection can show biased T_i estimations, mainly enriching in the tail of the distribution. For example, in Figure 1.12, the estimation T_i is represented by the loci genealogies of three individuals. Putative neutral site sites would share similar topologies (gray boxes), and putatively selected sites result in unexpected topologies (red boxes) considering whole genome distributions. In these examples, differences in loci genealogies will be summarized in T_i estimations, where T_i at the putatively selected alleles enrich the most extreme values (outliers) of the empirical distributions of T.

Hence, following the Figure 1.12, the approach proceeds: i) surveying any i loci to estimate the summary statistic T; ii) estimating the empirical distribution of T. The last step is to consider a cutoff in the empirical distribution under which outliers are considered. This cutoff is usually arbitrary, and data in the 99th or 95th percentile has been primarily used in many studies. Because the chosen percentile only represents the distribution tails of summarized values, unexpected patterns of variation due to confounding factors (such as demography or recombination) can mimic positive selection patterns and the selected cutoff itself can increase the number of false positives and false negatives obtained by this methodology.

Neutral simulations that incorporate realistic demographic models and selection are essential for choosing a cutoff. Ultimately, outliers in the empirical distribution only indicate an extreme variation pattern, which may be caused by other cofactors and not by natural selection. Thus, a cutoff based on simulations can be set based on the expected patterns under neutrality and any other cofactors resulting in robust detection (Akey, 2009). While demographic and local recombination inferences have been refined over the years, simulation studies are far from perfect. Most genome-wide positive selection scans are accompanied by coalescent simulations using programs such as `discoal` (Kern and Schrider, 2016), `msms` (Ewing and Hermisson, 2010) or `cosi2` (Shlyakhter et al., 2014). These simulations usually account for selection at a single locus. However, other factors, such as codominance or background selection, are not considered because of the expensive computational cost and the assumptions of coalescence theory. Even though, in some specific cases, when they can fail to reproduce even a classic hard sweep, simulations have repeatedly shown to be quite valuable to establish a cutoff considering the neutrality, demography, and recombination.

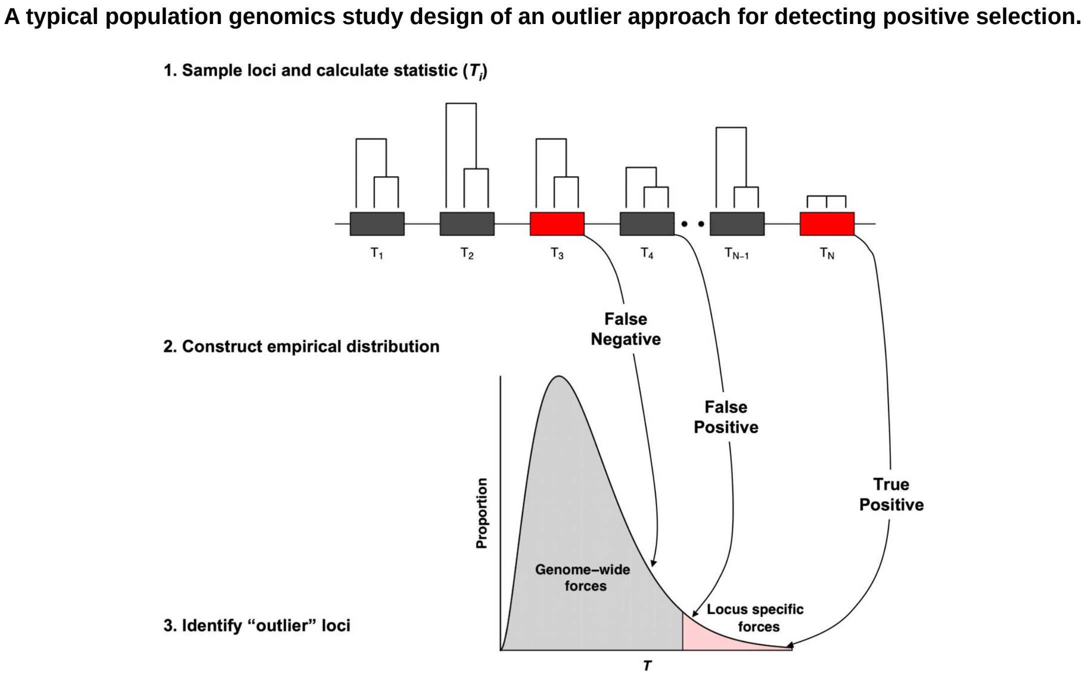

A typical population genomics study design of an outlier approach for detecting positive selection.
1. Sample loci and calculate statistic (T_i)
T_1
T_2
T_3
T_4
T_{N-1}
T_N
2. Construct empirical distribution
False Negative
False Positive
True Positive
Proportion
Genome–wide forces
Locus specific forces
3. Identify "outlier" loci
T

Figure 1.12: Schematic view of an outlier approach. An estimator T is applied on the sample (usually SNPs along the genome). Assuming that the entire genome will be affected by the same confounding factors, such as demographic events, the T estimate at putatively selected loci should be reflected as an extreme value in the empirical distribution of T. Taken from Akey (2009).

Kelley et al. (2006) and Teshima et al. (2006) explored the performance of the outlier approach by performing coalescent simulations. The simulations performed by Kelley et al. (2006) anticipated several essential points. First, they found that the number of false positives can be high even when considering ascertained data. It will ultimately depend on the parameters that modulate selection and the fraction of loci targeted by positive selection. However, although the authors detailed the problem dealing with the cutoffs, Kelley et al. (2006) claimed: *if the goal of a study is to identify a restricted set of candidate selection genes to study in more detail, then our data suggest that an outlier approach is a reasonable study design as long as one accepts that a substantial proportion of candidates may be false positives.* Nonetheless, once the simulations incorporate accurate models in mutation rates, recombination, and selection coefficients, the increase in variance in summary statistics due to perturbations will be similar to that of natural populations. Therefore, the identification of outliers based on more realistic simulations should result in more accurate inferences. The simulations by Kelley et al. (2006) and Teshima et al. (2006) manifest that extreme outlier values arise under neutrality, considering the stochastic nature of evolution. However, both studies used classical neutrality tests to measure the effect of positive selection. While such tests have high power to identify completed or near fixed sweeps from different periods and stages, they generally have low statistical power to detect ongoing or partial sweeps. Although it is necessary to expand the work of Kelley et al. (2006) and Teshima et al. (2006) regarding the outcomes of applying other methods on the outlier approach, Enard et al. (2014) and, more recently, Booker et al. (2020) also carried out similar analyses. The analysis performed by Enard et al. (2014) and Booker et al. (2020) reviewed complex models of selection and recombination further than the selected cutoff and coalescence simulations in the outlier approach, pinpointing out limitations of the future genome-wide analysis.

As shown in table 1.2, the number of genome-wide scans of positive selection accounts for different datasets and methodologies, testing natural selection at different levels and populations (Haasl and Payseur, 2016; Lohmueller and Nielsen, 2021). Overall, these studies show that up to 10% of the genome could be subject to positive selection (Akey, 2009), being ubiquitous throughout the genome. These values may be up to 20% higher in the case of Drosophila (Mackay et al., 2012). Nevertheless, genome-wide approaches clearly show a pervasive effect in the studied species (Haasl and Payseur, 2016). Interestingly, as discussed in Lohmueller and Nielsen (2021) and Akey (2009), these studies lead to the creation of catalogs of positive selection and show that a considerable number of human genes are subject to the action of natural selection. Moreover, functional studies of these genes suggest that they are consistent with previous hypothesis and studies at the gene level, showing enrichment of signals in genes related to immunity, olfactory receptors, pigmentation, metabolic pathways, and other cell cycle signals. However, the vast majority of studies show more significant signal enrichment in non-genic regions. Although many of these regions can be attributed to regulatory regions close to genes, other intergenic signals can be attributed to distal enhancers, non-coding RNAs, or even elements related to genome organization. For example, Enard et al. (2014) extensively observed that positive selection signals correlate better with regulatory sequences than classical amino acid substitutions.

Through these studies, clear examples of selection have been elucidated over the last decades. Overlapping results is one of the simplest ways to assess confidence in the results of the genome-wide selection. Considering the various statistics used to measure selection strength and timing and the wide variety of data and populations, one cannot expect concordance between false positives. However, the agreement between results is also far from perfect. For example, the overlap between adaptive events results in only 14% of putatively adaptive regions (Akey, 2009). Indeed, the figure may be worrisome, but considering there is no consensus to use a cutoff or, even more importantly, standardization in neutral simulations, the result should not be surprising either. Other genome-wide approaches have been performed using new methodologies, and more accurate data. Nevertheless, even considering the results described by CMS Grossman et al. (2013) or Johnson and Voight (2018), the agreement remains low.

Overall, it may be easy to foresee the criticisms in this type of analysis. Among this low concordance, the studies can locate multiple constantly repeated genes independently of data and methodologies. Among these examples, we have found new candidates and old and clear examples, such as LCT or genes associated with resistance to malaria. These results have given rise to the complete maps of positive selection in the human genome to date, combining genome-wide scans with the first studies that had an a priori hypothesis.

Table 1.2: Genome-wide scan of positive selection on human populations over the last decade. Adapted from Haasl and Payseur (2016), and Lohmueller and Nielsen (2021)

Year	Study	Data	Statistic	Data source
1999	Huttley et al. (1999)	STR	Extended LD	European Utah and Amish CEPH
2002	Akey et al. (2002)	26.5K SNPs	Fst	European-American African, American Chinese-American, (The SNP Consortium)
2002	Payseur et al. (2002)	STR	SFS	European
2003	Kayser et al. (2003)	STR	lnRV, RST	Africans and Europeans
2003	Clark et al. (2003)	WES	dN/dS	
2004	Storz et al. (2004)	STR	FST	Africans, Europeans and Asians
2004	Shriver et al. (2004)	SNP	FST	European-Americans, 20 African-Americans, 10 Chinese and 10 Japanese
2005	Altshuler et al. (2005)	1M SNPs	rEHH	HapMap phase I
2005	Bustamante et al. (2005)	WES	d_N/d_S	
2005	Carlson et al. (2005)	SNP	SFS	24 African-American individuals, 24 CEPH individuals, and 24 Chinese-Americans
2005	Weir et al. (2005)	SNP	FST	Perlgen, HapMap phase I
2005	Nielsen et al. (2005)	HapMap II and Seattle SNPs	Site frequency Spectrum (CLR, MWu)	Europeans and African-American
2006	Kelley et al. (2006)	1.6M SNPs assigned to 14,589 gene regions	Tajima's D	Perlegen
2006	Voight et al. (2006)	1M SNPs	iHS	HapMap phase I
2006	Wang et al. (2006)	1.6M SNPs	ALnLH	HapMap phase I, Perlegen
2006	Zhang et al. (2006)	100K SNPs	WGLRH	Perlegen, HapMap phase I
2006	Mattiangeli et al. (2006)	STR	Ewens Watterson test	72 unrelated Irish individual
2006	Bubb et al. (2006)	SNP Consortium	High SNP density	African-American
2007	Consortium (2007)	3.1 M SNPs	LRH and iHS	HapMap phase II
2007	Kimura et al. (2007)	1M SNPs	rMHH	HapMap I
2007	Sabeti et al. (2007)	3.1M SNPs	rEHH, iHS, XP-EHH	HapMap phase II
2007	Tang et al. (2007)	1.5M Perlegen SNPs 3.6M HapMap SNPs	lnRsb	Perlgen, HapMap phase III
2007	Williamson et al. (2007)	100 K SNPs	CLR	Perlegen
2007	Haygood et al. (2007)	SNPS		

2008	Johansson and Gyllensten (2008)	1.6 M SNPs	Haplotype block length and Fst	Perlegen
2008	Kimura et al. (2008)	500 K SNPs	LD (AREHH, rHH)	Melanesian, Polynesian
2008	O'Reilly et al. (2008)	1.6M or 1 M SNPs	LD (Ped/pop)	HapMap, Perlegen
2008	Oleksyk et al. (2008)	200K SNPs	Fst, SFS	European-American, African- American
2008	Hancock and Di Rienzo (2008)	SNP	GLMM (climatic variables)	
2008	Myles et al. (2008)	SNP	FST	HapMap phase I
2009	Amato et al. (2009)	4M SNPs (HapMap III)	Population differentiation (FST)	HapMap phase II
2009	Herráez et al. (2009)	900K SNPs	LD lnRsb	HGDP
2009	Nielsen et al. (2009)	13,400 genes (sequence)	SFS, Fst	European-Americans African-Americans
2009	Pickrell et al. (2009)	650K SNPs	iHS, XP-EHH, Fst	HGDP
2009	Chen et al. (2009)	HapMap	Modified MKT	HapMap phase II
2009	Andrés et al. (2009)	WES	CLRT	19 African-Americans and 20 European-Americans
2010	Chen et al. (2010)	3.6M SNPs	XP-CLR	HapMap phase II
2010	Grossman et al. (2010)	3.1M SNPs	CMS	3 HapMap II
2010	Hancock et al. (2010)	0.64M SNPs	Correlation of SNP frequency with environmental variable	61 pops: HGDP, Luhya, Maasai, Tuscan, Gujarati from HapMap III,!Kung, Amhara, Yup'ik, Chukchee, and Aborigine
2010	Altshuler et al. (2010)	3.1M SNPs	CMS	3 non-admixed HapMap phase III (TSI, LWH, MKK)
2010	Lappalainen et al. (2010)	500K SNPs	iHS, rEHH and Fst	Finnish, Swedish, German, British— 100–350 each, HapMap
2010	Tennessen et al. (2010)	25,769 kb exome sequence (56,000 SNPs)		4 African, 6 European
2010	Consortium (2010)	Low coverage whole genomes	Fst	HapMap phase II
2010	Yi et al. (2010)	WES	PBS	
2010	Albrechtsen et al. (2010)	SNP	Excessive Identity by descent	
2010	Bigham et al. (2010)	SNP, CNV	lnRH, WGRLH, SFS	
2010	Simonson et al. (2010)	SNP	iHS, XP-EHH	
2010	Beall et al. (2010)	SNP	Allele frequency differences	
2010	Lappalainen et al. (2010)	SNP	iHS, LRH, EHH, FST	
2010	Xu et al. (2011)	SNP	iHS, XP-EHH, XP-CLR, FST	

2010	Enard et al. (2010)	SNP	Novel version of HKA test	
2011	Bhatia et al. (2011)	900K SNPs	Fst	Gambian African-American Nigerian
2011	Cai et al. (2011)	4M SNPs (HapMap II)	Shared genomic segment (SGS)	3 populations HapMap II
2011	Tennessen and Akey (2011)	100K SNPs	Fst	19 of HGDP pops
2011	Fan et al. (2011)	3.1M SNPs	XP-EHH	3 HapMap II
2011	Metspalu et al. (2011)	SNP	XP-EHH, iHS	
2011	Fumagalli et al. (2011)	SNP	GLMM	
2011	Hancock et al. (2011)	SNP	GLMM	
2012	Suo et al. (2012)	SNP	iHS, XP-EHH	HapMap, HGDP, SGVP
2012	Granka et al. (2012)	SNP	iHS, XP-EHH	
2012	Piras et al. (2012)	SNP	EHH, XP-EHH	
2012	Vernot et al. (2012)	WGS	Fst	
2012	Zhang et al. (2012)	SNP, CNV	CNV frequency differentiation	
2012	Jarvis et al. (2012)	SNP	Fst, XP-EHH, iHS	
2012	Andersen et al. (2012)	SNP, WGS	CMS	
2012	Scheinfeldt et al. (2012)	SNP	LSBL	
2013	Grossman et al. (2013)	Low coverage WGS	CMS	1000GP pilot phase
2013	Migliano et al. (2013)	SNP	iHS, XP-EHH	
2013	Somel et al. (2013)	SNP	d_N/d_S	
2013	Hider et al. (2013)	WGS	SFS, Rsb, PBS	
2013	Frichot et al. (2013)	SNP	Latent factor mixed models	
2013	Raj et al. (2013)	SNP	iHS, FST	
2013	Liu et al. (2013)	SNP	LHR	
2013	Leffler et al. (2012)	WGS	Haplotype sharing between species	
2014	Colonna et al. (2014)	SNP, indel	iHS, XP-EHH, FST	
2014	Eichstaedt et al. (2014)	SNP	iHS, XP-EHH, FST	
2014	Clemente et al. (2014)	SNP	iHS, SFS	
2014	Haasl et al. (2014)	WGS	Novel ksk2 test	
2014	Ali et al. (2014)	SNP	iHS, XP-EHH	
2014	Wuren et al. (2014)	SNP	iHS, XP-EHH	
2014	Fagny et al. (2014)	WGS	iHS and DIND	

2014	Enard et al. (2014)	WGS	iHS, XP-EHH	
2014	(Sjöstrand et al., 2014)	SNP	Novel Maximum Frequency of Private Haplotypes test	
2014	Pybus et al. (2014)	Low coverage WGS	Hierachical boosting	1000GP pilot phase
2016	Field et al. (2016)	SNP	SDS	UkBiobank
2017	Schrider and Kern (2017)	Low coverage WGS	S/HIC	1000GP phase III
2018	Sugden et al. (2018)	SNP, WES, WGS	SWIFT(r)	‡Khomani San, 1000GP pilot phase
2018	Johnson and Voight (2018)	Low coverage WGS	iHS	1000GP phase

STR: microsatellite (short tandem repeat); CNV: copy number variant; WGS: whole-genome sequence; WES: whole-exome sequence; ESS: exonic scan for selection; GLMM: use of generalized linear mixed model methodology; CLRT: composite likelihood ratio test; iHS: integrated haplotype statistics; EHH: extended haplotype homozygosity; XP-EHH: cross-population EHH; LRH: long-range haplotype test; WGRLH: whole-genome LRH; SFS, site frequency spectrum statistic(s); PBS: population branch statistic; XP-CLR: cross-population composite likelihood ratio; HKA: Hudson–Kreitman–Aguade test.

Chapter 2

Objectives

The ultimate goal of this thesis is to characterize adaptation that has occurred at different time scales in the human lineage through the analysis of the 1000 Genomes Project (1000GP) dataset. The 1000GP represents, to date, the largest and highest quality genome-wide dataset of human nucleotide variation, and contains a treasure trove of information about human evolution. In particular, this thesis places special emphasis on methods for the detection of recurrent positive selection events.

In order to achieve these objectives, the following specific objectives are proposed.

1. Perform an exhaustive genome-wide scan combining different population genetics statistics to detect candidate regions under selection in 22 non-admixed human populations from 1000GP. The combination of different statistics and the pinpointing of those regions that stand out from the background genomic variability, including abnormally long haplotypes, shifts in the Site Frequency Spectrum (SFS), or an excess of non-synonymous substitutions, should result in an exhaustive understanding of human genetic adaptation. In addition, because of the lack of consistency between previous genome-wide scans of positive selection, we aimed to construct a robust methodology, not only to detect new adaptive events, but also to replicate genome-wide scans of positive selection to date.

2. To construct an online, user-friendly database to facilitate the thorough analysis of candidate regions under selection in the human genome by putting together the evidence of selection with structural and functional annotations of the regions and cross-references to previously published articles.

3. To develop a new, more efficient McDonald and Kreitman (MK) approach specifically designed to improve gene-by-gene analyses. We aimed to construct a metric that does not exclude all variants below a frequency threshold, avoiding dropping out a large fraction of the data.

4. Reformulate the ABC-MK, an MK-based approach that incorporates linked selection to the MK test. We aimed to extend ABC-MK theoretical and bioinformatically to avoid the prohibitively expensive computational cost and High Parallel Computing.

5. To develop bioinformatics tools to facilitate the analysis and integration of other datasets or species regarding genome-wide scans of positive selection and MK test approaches.

Chapter 3

PopHumanScan: the online catalog of human genome adaptation

Abstract

Since the migrations that led humans to colonize Earth, our species has faced frequent adaptive challenges that have left signatures in the landscape of genetic variation and that we can identify in our today's genomes. Here we (i) perform an outlier approach on eight different population genetic statistics for 22 non-admixed human populations of the Phase III of the 1000 genomes project to detect selective sweeps at different historical ages, as well as events of recurrent positive selection in the human lineage; and (ii) create PopHumanScan, an online catalog that compiles and annotates all candidate regions under selection to facilitate their validation and thoroughly analysis. Well-known examples of human genetic adaptation published elsewhere are included in the catalog, as well as hundreds of other attractive candidates that will require further investigation. Designed as a collaborative database, PopHumanScan aims to become a central repository to share information, guide future studies and help advance our understanding of how selection has modelled our genomes as a response to changes in the environment or lifestyle of human populations. PopHumanScan is open and freely available at https://pophumanscan.uab.cat

Introduction

Since the split with chimpanzees, and especially since the migrations that led humans to colonize almost every single place on Earth, our species has faced frequent environmental and social changes that have shaped the variation patterns of our genomes through the action of natural selection (Nielsen et al., 2017). These environmental challenges include, for example, extreme cold temperatures in much of the Americas and Eurasia during the last ice age, limiting exposure to sunlight as we moved to higher latitudes, or contact with new pathogens. Part of the incorporated genetic innovations may have been introgressed from archaic hominins that left Africa before us, including Neanderthals and Denisovans, with whom we encountered and interbred before they got extinct. Around 1 to 6% of any modern non-African human genome can be traced back to the genomes of these archaic populations (Racimo et al., 2015). Another dramatic change occurred within the past 10,000 years coinciding with the transition from a hunting-gathering lifestyle to farming. Selection pressures for adapting to large settlements and new diets favored genetic variants associated with innate immune response, fatty acid metabolic efficiency, and lactose tolerance, among others (Fan et al., 2016).

These selection pressures left signatures in the landscape of genetic variation that can be identified in our today's genomes (Sabeti et al., 2006). Starting from single-locus studies to the first large-scale catalogs of genetic variation (Hinds et al., 2005; Altshuler et al., 2005; Consortium, 2007; Altshuler et al., 2010), dozens of targets of positive selection have been identified, providing important insights into recent human evolutionary history (Sabeti et al., 2007; Akey, 2009; Fan et al., 2016). Even though genome-wide HapMap genotyping data is able to disentangle the effects of demography and selection better than single-locus approaches, it still has the problem of ascertainment bias, which may alter the site frequency spectrum of analyzed single nucleotide polymorphisms (SNPs) (Kelley et al., 2006). The availability of the most comprehensive worldwide nucleotide variation data set so far from the 1000 Genomes Project (1000GP) (Consortium, 2012; Auton et al., 2015), based on whole-genome re-sequencing, provides the human lineage with an abundant, ascertained variation dataset on which to test molecular population genetics hypotheses and eventually pinpoint targets of positive selection in one or more human populations that escape from the background evolutionary dynamics of genetic variation (Johnson and Voight, 2018).

To gain deeper understanding of how environmental and social challenges have shaped our genomes through the action of natural selection, here we (i) perform a genome-wide scan of selection on the latest version of the 1000GP data by surveying distinctive signatures of genomic variation left by different selective events, and (ii) create an online catalog of all candidate genomic regions under selection to facilitate their validation and thorough analysis. As far as we are concerned, dbPSHP (Li et al., 2014) and the 1000 Genomes Selection Browser 1.0 (Pybus et al., 2014) are the only previous online databases that compiles putative positively selected loci in human evolution. In the dbPSHP database, regions were extracted from curated publications based on genotyping data –instead of whole-genome re-sequencing data– of the HapMap III (Altshuler et al., 2010) and the 1000GP Pilot 1 (Consortium, 2012), and the last update is reported as far as May 2014. In the 1000 Genomes Selection Browser, they use data of three populations from the 1000GP Pilot 1 (Consortium, 2012), and identify regions under selection by means of a machine-learning algorithm that combines the results of multiple neutrality tests (Pybus et al., 2015). Here we perform an outlier approach on the greatest number of population genetic statistics and sampled populations available so far. This genome-wide scan of selection is able to detect sweeps at different historical ages, as well as evidence of recurrent selection in the human lineage since the split between our species and chimpanzees. Results have been made available in a collaborative, online database, PopHumanScan, which is aimed at compiling and annotating adaptation events along the human evolutionary history.

Well-known examples of human genetic adaptation published elsewhere are included in the catalog, as well as hundreds of other attractive candidates that will require a more thoroughly analysis. PopHumanScan graphically represents each signature of selection within the empirical distributions of the corresponding DNA diversity statistic across populations. It also provides structural and functional annotations of the region, links to external databases, and cross-references to 268 publications.

PophumanScan analysis pipeline

We have designed and implemented a custom pipeline (Figure 3.1) to perform a genome-wide scan of selection. Specifically, the pipeline processes eight different neutrality tests calculated either in sliding windows along the genome or for each protein-coding gene, for 22 non-admixed human populations. The genomic regions identified should show signatures that are compatible with natural selection having driven the evolution of the region at one or different timescales, from recent selective sweeps to recurrent selection since the split between our species and chimpanzees. These candidate regions under selection are further characterized with structural and functional annotations of that particular region. Furthermore, 268 articles reporting evidences of natural selection in genomic regions and genes using different statistical methods have been manually curated and cross-referenced to the candidate regions detected with our pipeline.

Pre-processing of the PopHuman data

Population genomic data was retrieved from PopHuman (Casillas et al., 2018) for 22 non-admixed populations of the Phase III of the 1000GP (Auton et al., 2015) (see Table A.1), mapped to GRCh37/hg19. Specifically, values for seven different neutrality tests have been obtained for each population for 186,549 10-kb non-overlapping sliding windows along the autosomes and the X chromosome (Figure 3.1). In addition, the McDonald and Kreitman test (MKT) (McDonald and Kreitman, 1991), as well as the proportion of substitutions that are adaptive (α) (Charlesworth, 1994; Smith and Eyre-Walker, 2002), were calculated on the protein-coding genes overlapping the candidate regions under selection identified with the other seven statistics and that showed some variability in both polymorphism and divergence, according to gene annotations from GENCODE release 27 (Harrow et al., 2012) and PopHuman polymorphism and divergence data (Figure 3.1). MKT-derived calculations were performed using the R package iMKT (https://github.com/BGD-UAB/iMKT; last accessed: February 2018). In total, eight different neutrality tests were performed, each spanning different timescales ranging from several million years ago to the present (Sabeti et al., 2006; Casillas and Barbadilla, 2017).

Linkage Disequilibrium (LD) signature ($<$ 30 kya). LD signatures were detected using two complementary measures based on linkage disequilibrium: iHS (Voight et al., 2006) and XP-EHH (Sabeti et al., 2007). iHS has good power to detect selective sweeps with haplotypes at moderate frequency (50%–80%), while XP-EHH is more powerful for detecting selective sweeps when the selected haplotype has a frequency $>$ 80%. In the case of XP-EHH, which analyzes pairs of populations, only pairs CEU-YRI, CEU-CHB and YRI-CHB were considered, and the population showing the evidence of selection was identified with the locus-specific branch length method (LSBL) (Shriver et al., 2004).

Site Frequency Spectrum (SFS) signature($<$ 80 kya). Five statistics have been considered.: One is based on population differentiation -F_{st} (Wright, 1950; Weir and Cockerham, 1984)-, and the other four are based on both the allele frequency spectrum and the levels of variability -Fay and Wu's H (Fay and Wu, 2000), Fu and Li's D and F (Fu and Li, 1993), and Tajima's D (Tajima, 1989)-. Fay and Wu's H detects an excess of high-frequency derived SNPs, compatible with an incomplete sweep, or recombination breaking swept linked SNPs. F_{st} detects population-specific selective events that changed the genetic composition of the affected population. It analyzes pairs of populations. The pairs CEU-YRI, CEU-CHB and YRI-CHB were considered, and the population showing the evidence of selection was identified with the locus-specific branch length method (LSBL) (Shriver et al., 2004). Fay and Wu's H detects an excess of high-frequency derived SNPs, compatible with an incomplete sweep, or recombination breaking swept linked SNPs.

Protein changes signatures (many millions of years). Recurrent selection since the split between our species and chimpanzees (¡6 mya) is detected using a test based on comparisons of polymorphism and divergence -MKT (McDonald and Kreitman, 1991)-, and the result of the test is summarized with the estimator α (Charlesworth, 1994; Smith and Eyre-Walker, 2002). For this calculation, we used a MKT-based methodology, which corrects for the presence of nonsynonymous slightly deleterious segregating sites in order to avoid underestimating α for methodological details of this method, see (Mackay et al., 2012).

Genome-wide scan of selection

For the parameter α of MKT, evidence of positive selection for protein-coding genes was inferred when $\alpha > 0$ and the Fisher's Exact Test for the 22 MKT contingency table was significant (P-value < 0.05) (Figure 3.1). Because the other seven selection statistics have not been associated with a simple parametric distribution, candidate windows under selection were identified as the most extreme values (within the 0.05% tail) in the corresponding empirical distribution. These empirical distributions were performed independently for each of the 22 populations (or three population pairs in the case of XP-EHH and F_{ST}), and independently for the autosomes and the X chromosome (to account for different demographic histories and the different effective population size of the autosomes compared to the X chromosome; chromosome Y was not analyzed). In total, 116 empirical distributions were obtained for autosomal regions, and 91 for the X chromosome (data of iHS and XP-EHH was not available for the X chromosome in PopHuman) (Figure 3.1). From the initial 186,549 10-kb non-overlapping windows from PopHuman for each population and statistic, those containing < 5 segregating sites were discarded ($< 0.2\%$, Figure A.1). Then, an empirical P-value was assigned to each of the remaining windows for each of the 116 combinations of population (or population pair) and statistic, separately for the autosomes as a whole and the X chromosome. Specifically, for each window i in a population (or population pair), p is the quantile of that window for statistic j, that is, its empirical P-value. In the case of Tajima's D, Fu & Li's D and F, and Fay and Wu's H, two-tailed P-values were calculated. Once the significance for each individual 10-kb window in the genome was assessed, a candidate region under selection was defined as being a contiguous genomic region containing at least one 10-kb significant window (P-value < 0.0005, Figure 3.2) and spanning adjacent windows with P-values < 0.005 (Figure A.2). In addition, this region may span stretches $< 20kb$ of contiguous nucleotides not analyzed in PopHuman (i.e., because they contain non-accessible bases according to the Pilot-style Accessibility Mask of the 1000GP (Auton et al., 2015; Casillas et al., 2018)). This outlier approach was designed to face the unique features and limitations of our PopHuman source data and to be highly conservative defining candidate regions under selection. We expect that it likely results in an enriched set of genomic regions that have been targets of natural selection along the human evolutionary history (Kelley et al., 2006), and refer to the outlier regions as candidate regions showing signatures of selection. Once candidate selected regions (or genes) were assigned for the 22 populations (or three population pairs) and eight statistics, they were collapsed according to their coordinates into a joint set of 2,879 candidate regions under selection genome-wide. Of these, 20 regions were removed because they were completely located in DAC Blacklisted regions (i.e., regions of the reference genome which are troublesome for high throughput sequencing aligners) or partially overlapped genomic gaps, as obtained from the UCSC (Casper et al., 2018). Therefore, a total of 2,859 regions were finally considered.

Structural and functional annotations

The final 2,859 candidate regions under selection were structurally and functionally characterized according to 15 different annotations categorized into 5 groups, extracted from the UCSC (Casper et al., 2018) and two publicly available databases (Vernot et al., 2016; Martínez-Fundichely et al., 2014) (Figure 3.1).

Sequencing. (i) Mappability was assessed as the percentage of bases in the region that do not present any troublesome to high-throughput sequencing aligners according to the DAC Blacklisted regions of the UCSC Casper et al. (2018). (ii) *Distance to closest GAP* was computed as the distance (in Mb) to the closest gap (Casper et al., 2018).

Regulation. (iii) CpG Islands (Gardiner-Garden and Frommer, 1987) and (iv) Vista Enhancers (Pennacchio et al., 2006) were computed as the percentage of bases in the region that overlap these genomic elements. (v) Transcription Factor Binding Sites (TFBSs) (Casper et al., 2018) and (vi) ORegAnno Regulatory Elements (Lesurf et al., 2016) were calculated as the total number of elements contained in the region.

Comparative genomics. Evolutionary conservation of the regions was assessed by considering the results of three different algorithms -phastCons, PhyloP and GERP- on the multiple alignments of the genomes of 100 vertebrate species (Pollard et al., 2010). (vii) PhyloP Evolutionary Conservation and (viii) GERP Constrained Elements were assessed as the percentage of bases in the region that have a score > 2 for the given statistic (i.e., constrained sites) (Casper et al., 2018). (ix) phastCons Evolutionary Conservation was calculated as the percentage of bases that overlap phastCons conserved elements (Casper et al., 2018).

Structural variation. (x) InvFEST Inversions (Martínez-Fundichely et al., 2014), (xi) DGV Structural Variants (MacDonald et al., 2014), (xii) RepeatMasker (Bao et al., 2015), (xiii) Segmental Duplications (Bailey et al., 2002) and (xiv) TRF Simple Tandem Repeats (Benson, 1999) were assessed as the percentage of bases in the region that overlap these genomic elements.

Archaic introgression. (xv) Archaic introgression was assessed as the percentage of bases in the region that overlap either Neanderthal or Denisova introgressed haplotypes (Vernot et al., 2016).

Published references. A total of 268 publications from 1954 to 2018 reporting either specific loci or multiple regions from a genome-wide scan of selection in the human genome were cross-referenced with our final 2,859 candidate regions under selection (Figure 3.1, Table A-2 online https://doi.org/10.1093/nar/gky959). Of these, 132 publications were directly extracted from the dbPSHP database (Li et al., 2014), while the other 136 were manually curated here. Exhaustive information from the main text and/or supplementary figures and tables was extracted for each reported loci, including the genomic coordinates, affected population(s), statistic(s), type of selection and PubMed ID. Genomic coordinates were lifted over to GRCh37/hg19 using the LiftOver tool of the UCSC (Casper et al., 2018), or deduced from protein-coding gene location, if necessary.

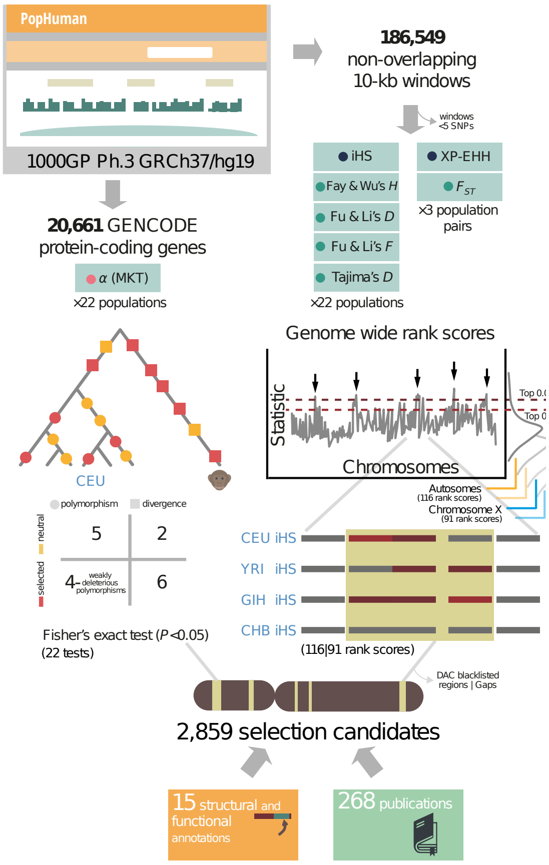


Figure 3.1: PopHumanScan pipeline. Starting from population genomic data retrieved from PopHuman, 8 different neutrality tests are analyzed in 22 non-admixed human populations (or 3 population pairs). Tests are color-coded depending on the type of signature they are able to detect: Linkage Disequilibrium (LD), Site Frequency Spectrum (SFS) and Protein Changes. The significance of each test is assessed either with a Fisher's exact test or a rank score, for each of the 22 populations (or 3 population pairs) independently, and independently for autosomes and the X chromosome. Finally, candidate regions under selection are structurally and functionally annotated, and cross-referenced with 268 publications

Overview of the PopHumanScan online catalog

In addition to the exhaustive genome-wide selection scan that has been performed, we have also created PopHumanScan, a collaborative, online database that is aimed at compiling and annotating adaptation events along the human evolutionary history (Figure 3.2). PopHumanScan reports each evidence of selection with the empirical distributions of the corresponding DNA diversity statistic across the human genome and among populations, structural and functional annotations of the region, links to external databases, as well as cross-references to 268 publications.

Implementation

PopHumanScan is currently running under Apache on a CentOS 7.2 Linux x64 server with 16 Intel Xeon 2.4 GHz processors and 32 GB RAM. It is mainly built on PHP as backend framework. It also includes AJAX for specific file requests and MySQL for data storage. The client-side is build on JavaScript and uses several JavaScript libraries, including jQuery, the jQuery plugin DataTables, and Plotly.js, as well as a custom Bootstrap 4 framework.

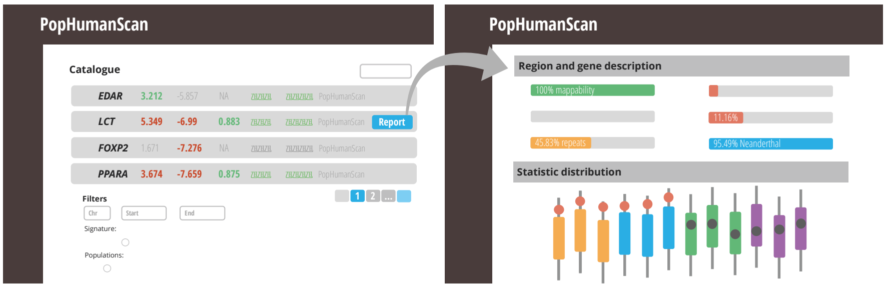
The PopHumanScan catalog

Main table. All 2,859 candidate regions under selection are displayed as rows in an interactive table (Figure 3.2, left). The information displayed in each row includes: (i) the genomic coordinates of the candidate locus, (ii) genes contained in or partially overlapping the region (if any), (iii) the most extreme value for each of the eight statistics considered (i.e., most extreme value in any 10-kb window included in the region, for any population or population pair; green (positive) and red (negative) values are outliers (P-value < 0.0005) in the corresponding empirical distribution), (iv) color-coded dots depicting the type of the selection signatures (i.e., ● Linkage Disequilibrium, ● Site Frequency Spectrum, and/or ● Protein Changes), (v) color-coded dots depicting the meta-population(s) that show signatures of selection (i.e., ● European, ● African, ● South-Asian, and/or ● East-Asian), and (vi) the source that contributed the candidate region under selection. At the time of writing, all 2,859 regions came uniquely from our genome-wide selection scan (i.e., source labelled as PopHumanScan), but additional data sources by contributors from the scientific community are expected once PopHumanScan is published (see next section). By clicking the + icon at the beginning of each row, detailed information of the particular candidate region under selection is displayed, including the values for all significant statistics in all target populations (or population pairs) and an overview of the main structural and functional annotations and cross-referenced publications (i.e., non-gray buttons represent overlapping annotations or cross-referenced publications), as well as access to the complete report for the corresponding candidate region under selection. Finally, several filters are available at the bottom of the page to narrow the search.

Complete report. A complete report for each candidate region can be accessed from the main table (Figure 3.2, right). The first section of the report displays all the structural and functional annotations of the region, together with links to external databases: (i) PopHuman (Casillas et al., 2018), which complements the population genomics information; (ii) HaploReg (Ward and Kellis, 2016), which allows the exploration of evolutionary conservation, expression eQTSs, epigenomic data, and regulatory annotations; and (iii) Ensembl (Zerbino et al., 2018), which allows the exploration of the linkage disequilibrium of the region, among others. The second section lists all the genes contained in or partially overlapping the region (if any). For each encoded gene, a short description of the gene and associated Gene Ontology terms for the Biological Process classification (The Gene Ontology Consortium, 2017) are provided, along with links to external databases: Ensembl (Zerbino et al., 2018), NCBI (Brown et al., 2015), Uniprot (The UniProt Consortium, 2017), UCSC (Casper et al., 2018), Expression Atlas (Papatheodorou et al., 2018), OMIM (Amberger et al., 2015), Open Targets (Koscielny et al., 2017), and HumanMine (Lyne et al., 2015). The third section contains cross-referenced publications that support the selection evidence found in the region. The fourth section contains an interactive graph showing recombination rate values in cM/Mb along the chromosome in which the region is located location, calculated from the recombination map by Bhérer et al. (2017)and extracted from PopHuman (Casillas et al., 2018). The specific location of the candidate region under selection is indicated with dashed vertical lines, and the solid horizontal line represents the average recombination rate value in the candidate region. Finally, in the fifth section boxplots show the distribution of each significant statistic in all the populations (or population pairs). Highlighted values correspond to those in the candidate region, and those in red are outliers of the empirical distribution (P-value < 0.0005).

Utilities and support resources

Contributing to PopHumanScan. PopHumanScan has been devised as a collaborative database. In order to incorporate information contributed by the scientific community, two password-protected tools have been implemented. The first one allows users to add additional candidate regions under selection in the catalog. All contributed regions will be subjected to manual curation and clearly labelled with a data source tag. The second tool allows manually cross-referencing candidate regions already present in the database.



EDAR	3.212	-5.857	NA	PopHumanScan
LCT	5.349	-6.99	0.883	PopHumanScan
FOXP2	1.671	-7.276	NA	PopHumanScan
PPARA	3.674	-7.659	0.875	PopHumanScan

Figure 3.2: Simplified representation of the PopHumanScan interface. The main PopHumanScan table is displayed to the left, while the complete report for a particular candidate region under selection is displayed to the right

Help and Tutorial.

This section documents the data used and the procedures implemented in PopHumanScan, as well as instructions on how to contribute to it. Interestingly, it also contains a complete tutorial introducing to the usage of the database through a step-by-step worked example.

Contents of PopHumanScan

At the time of writing, the PopHumanScan database contains 2,859 candidate regions under selection derived from the genome-wide selection scan pipeline presented here. Regions are distributed homogeneously along the autosomes and the X chromosome (Table 3.1, Figure A.2). Of these, 1,453 regions (50.8%) overlap GENCODE protein-coding genes, and 1,986 regions (69.5%) are cross-referenced with at least one publication (Table 3.1 Figure 3.3).

Table 3.1: Summary of the candidate regions under selection included in PopHumanScan

		Regions with selection signatures in meta-populations				Regions with different types of signatures				
Chr	**Number of candidate regions**	● European (EUR)	● African (AFR)	● South-Asian (SAS)	● East-Asian (EAS)	● Linkage disequilibrium (LD)	● Site Frequency Spectrum (SFS)	● Protein Changes	**Regions overlapping protein-coding genes**	**Regions cross-referenced with publications**
1	214	57	84	48	66	46	176	2	123	152
2	253	77	95	60	81	56	203	2	131	191
3	201	61	81	49	54	34	173	0	116	153
4	241	62	110	68	59	42	207	1	111	173
5	166	46	62	41	52	32	140	1	85	119
6	171	42	83	42	49	29	146	1	81	118
7	144	53	58	49	49	25	125	0	74	115
8	164	41	57	51	48	32	135	0	68	108
9	96	28	33	22	34	12	85	1	51	59
10	141	45	47	45	48	21	126	1	70	103
11	120	35	37	34	44	22	99	2	62	86
12	142	44	52	35	42	24	123	1	87	105
13	82	17	31	27	21	9	75	0	23	51
14	74	19	29	25	22	15	61	0	39	57
15	79	34	29	17	25	12	70	1	47	57
16	88	20	20	41	30	21	68	1	49	69
17	86	30	32	34	24	18	71	1	64	69
18	56	15	23	14	18	7	49	0	24	43
19	67	21	29	21	27	5	62	4	42	38
20	59	19	22	14	23	10	52	1	29	49
21	34	13	11	10	16	7	28	0	11	26
22	39	8	16	8	15	8	35	1	27	32
X	142	44	49	36	37	ND	142	0	39	13
Total	2859	831	1090	791	884	487	2451	21	1453	1986
		29.1%	38.1%	27.7%	30.9%	17.0%	85.7%	0.7%	50.8%	69.4%

ND = Not Determined

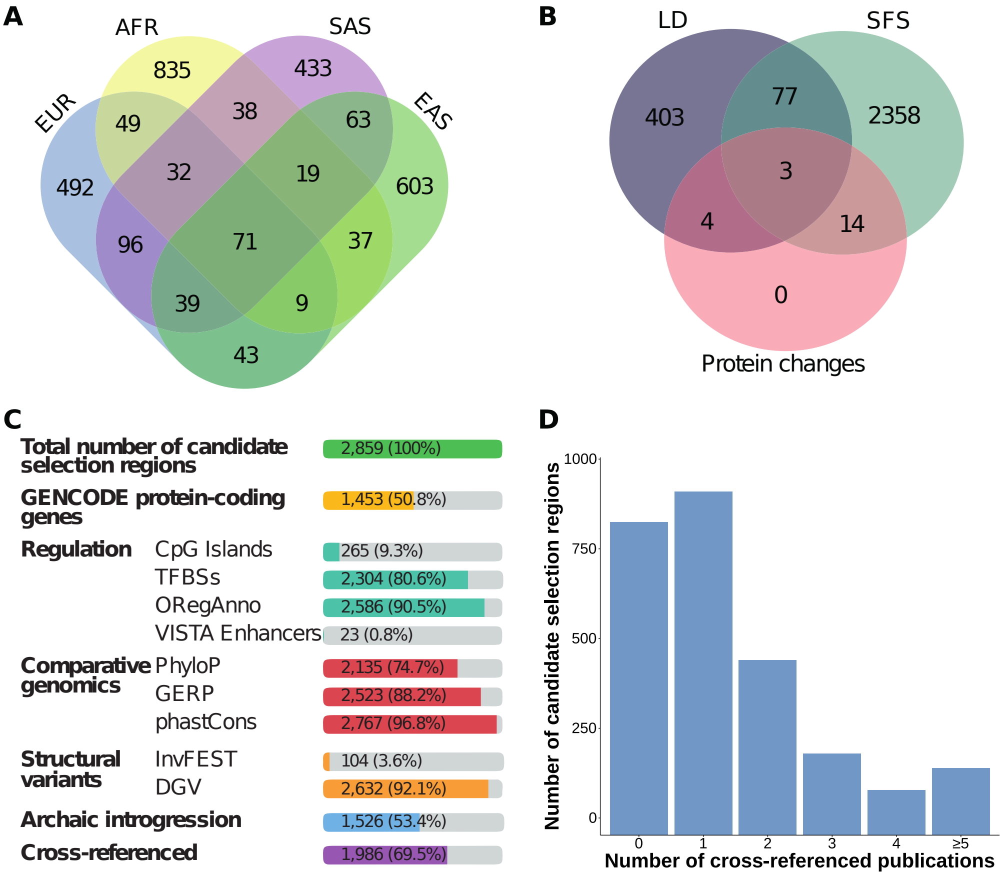


Figure 3.3: Summary of the contents of PopHumanScan. (A) Number of candidate regions under selection unique and shared among the four meta-populations: EUR, AFR, SAS EAS. (B) Number of candidate regions under selection unique and shared among the three different signature types: Linkage Disequilibrium (LD), Frequency Spectrum (SFS) and Changes. (C) Number of candidate regions under selection overlapping different structural and functional annotations. (D) Number of candidate regions under selection cross-referenced with $0, 1, 2, 3, 4$ or ≥ 5 published papers

Selection signatures in meta-populations

The total number of candidate regions showing signatures of selection in the four meta-populations is: 831 (29.1%) in Europe, of which 413 (49.7%) overlap protein-coding genes; 1,090 (38.1%) in Africa, of which 580 (53.2%) overlap protein-coding genes; 791 (27.7%) in South-Asia, of which 401 (50.7%) overlap protein-coding genes; and 884 (30.9%) in East-Asia, of which 424 (48.0%) overlap protein-coding genes (Table 1). Most of the regions (82.5%) show signatures that are unique to one single meta-population (Figure A.3-A): 492 (17.2%) show signatures that are unique in Europe, 835 (29.2%) are unique in Africa, 433 (15.1%) are unique in South-Asia, and 603 (21.1%) are unique in East-Asia. Of the 1,090 regions showing signatures in Africa, 76.6% are unique to Africans; while a lesser percentage -59.2%, 54.7% and 68.2%- of the regions showing signatures in Europe, South-Asia and East-Asia, respectively, are unique to their meta-population. About one third (29.0%) of the candidate regions under selection are shared across populations within the same meta-population. This percentage is higher for candidate regions showing both LD and SFS signatures (52.7%), it is 33.6% for candidate regions showing LD signatures only, and 27.1% for candidate regions showing SFS signatures only.

Types of selection signature

The total number of candidate regions showing distinct types of signatures of selection is: 487 (17.0%) for Linkage Disequilibrium (LD); 2,451 (85.7%) for Site Frequency Spectrum (SFS); and 21 (0.7%) for Protein Changes (i.e., recurrent selection since the split between humans and chimpanzees) (Table 1). Most of the regions (96.6%) show one single signature of selection (Figure A.3-B): 403 (14.1%) show LD signatures only; and 2,358 (82.5%) show SFS signatures only. All genes showing evidence of recurrent selection also show signatures in either LD and/or SFS, as only genes overlapping candidate regions under selection detected by LD and/or SFS were tested for α (MKT). These results would indicate that the statistics we used in our genome-wide scan of selection look at different characteristics of the genetic variability of the region, and that they are largely complementary.

Structural description of the regions

Region length. Most of the candidate regions under selection (63.6%) span one single 10-kb window, and the variable lengths of candidate regions follows a reversed J-shaped distribution (Figure A.4-A).

Distance between consecutive regions. The average distance between consecutive candidate regions is $\approx 1Mb$, and the distribution of distances is also reversed J-shaped (Figure A.4-B).

Recombination. The average recombination rate of the candidate regions is 0.71 cM/Mb, and the distribution of recombination rates is again reversed J-shaped (Figure A.4-C). There is a strong, negative, nonlinear association between recombination rate and both region length (Figure A.5-A) and distance between consecutive candidate regions (Figure A.5-B).

Functional description of the regions

Regulation. Most of the candidate regions (90.5%) contain at least one regulatory element annotated in the ORegAnno database, and 80.6% contain TFBSs (Figure 3.3-C). On the contrary, VISTA enhancers are much less abundant in the genome and they are only found in 23 of the 2,859 candidate regions (0.8%). CpG Islands are also in shortage and they are present in 9.3% of the regions.

Comparative genomics. Nearly all (96.8%) candidate regions overlap phastCons conserved elements. In the case of GERP and PhyloP, 88.2% and 74.7% of the regions, respectively, overlap constrained bases with score > 2. Structural variation. The Database of Genomic Variants (DGV) (MacDonald et al., 2014) is a very exhaustive database of structural variants annotated in the human genome. One or more elements annotated in this database are present in 92.1% of the candidate regions under selection. On the contrary, only 104 regions (3.6%) overlap validated polymorphic inversions from the manually curated InvFEST database (Martínez-Fundichely et al., 2014).

Archaic introgression. A total of 1,526 of the candidate regions (53.4%) overlap haplotypes introgressed from either neanderthals or denisovans. This percentage is expected, as introgressed haplotypes persisting in different present-day human individuals cover 46.7% of the reference genome (Vernot et al., 2016).

Cross-references with publications. A percentage of 69.5% of the candidate regions are cross-referenced with at least one publication, and 36.0% are cross-referenced more than once (Figure 3.3)

Gene Ontology analysis

Our candidate regions overlap a total of 1,447 unique GENCODE protein-coding genes. These were functionally classified into Gene Ontology (GO) terms (The Gene Ontology Consortium, 2017) according to the PANTHER GO-Slim annotation dataset using the PANTHER Classification System (Mi et al., 2017) (Figure A.6). In addition, statistically over- and under-represented functions were analyzed using the complete GO annotation dataset (The Gene Ontology Consortium, 2017) using the same tool (Tables S3, S4, and S5). Interestingly, among all Biological Process categories, regulation of neuron projection development is over-represented (fold enrichment 1.88, FDR 1.23E-02), in addition to cellular component organization (fold enrichment 1.24, FDR 1.59E-03) (Table S4). Finally, several Cellular Component categories are statistically over-represented, including presynaptic membrane (fold enrichment 2.72, FDR 4.32E-02) (Table S5). In spite of finding some statistically over-represented GO categories in our genes list, selection signatures seem to be heterogeneous and a detailed analysis of each candidate region is required to understand the real story under each selective event.

Pophumanscan with an example: selection at the lactase locus

The introduction of agriculture and cattle domestication in the Middle East and North Africa $\approx 10,000$ years ago lead to strong selection pressure for the ability to digest milk as adults. This is accomplished if the enzyme lactase that metabolizes lactose, encoded by the LCT gene, maintains high levels into adulthood, a characteristic that is called lactase persistence. Several variants near the LCT locus show some of the strongest signals of selection in the human genome for those populations that have traditionally practiced dairying, including a genetic variant in an intron of the gene MCM6, upstream of LCT (Fan et al., 2016). The LCT locus is found inside the longest candidate region under selection reported in PopHumanScan ($\approx 1Mb$). The region is located in the long arm of chromosome 2 and contains 8 GENCODE protein-coding genes, including LCT and MCM6 (Figure 3.4). Our genome-wide scan of selection has detected signatures at four different statistics that span the three types of signatures: LD (iHS and XP-EHH), SFS (Fu and Li's D), and Protein Changes (α). LD signatures involve basically European and African populations, while the signature of recurrent selection is more general to the four meta-populations. The region contains thousands of TFBSs and hundreds of ORegAnno regulatory elements, it overlaps evolutionary constrained elements, and $> 95\%$ of the region overlaps haplotypes introgressed from Neanderthals. It has been reported in 24 published articles (of the set of 268 that we considered).

Conclusions

In summary, our exhaustive approach combining eight different statistics to detect candidate regions under selection in 22 non-admixed human populations has been able to locate distinct signatures in 2859 regions that stand out from the background genomic variability, including abnormally long haplotypes, shifts in the SFS or excess of non-synonymous substitutions between our species and chimpanzees. Many of these regions probably manifest the footprints of selective sweeps that occurred at different historical ages, or recurrent selection that has been taking place during the last millions of

years. The PopHumanScan online database is going to facilitate the thorough analysis of candidate regions under selection in the human genome by putting together all these evidences of selection with structural and functional annotations of the regions and cross-references to previously published articles. Furthermore, the database can incorporate new data from the scientific community through specific build-in utilities. All in all, PopHumanScan aims to become a central repository to share information, guide future studies and contribute to the research on human genome adaptation.

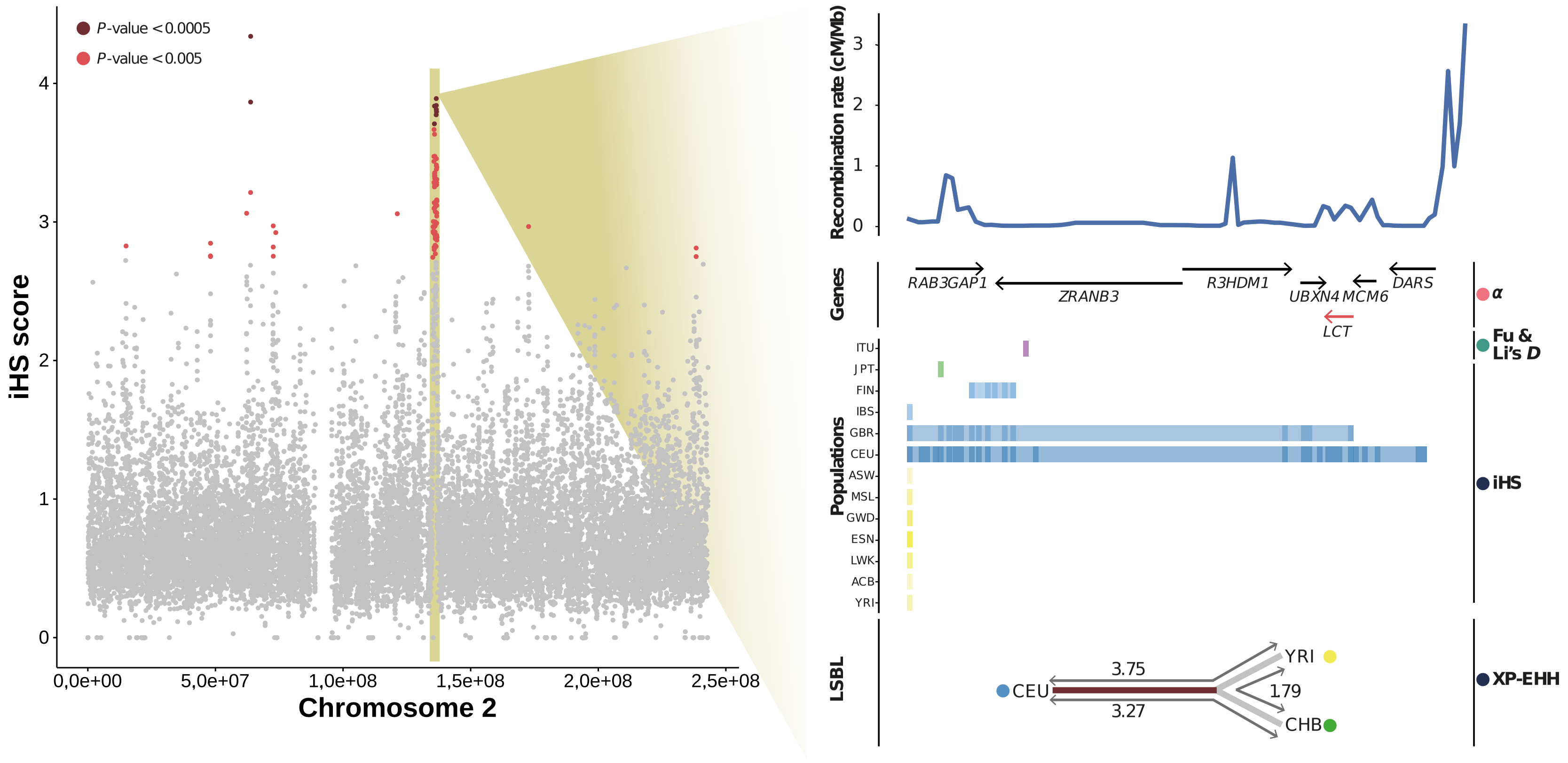


Figure 3.4: Signatures of selection detected at the lactase locus. The distribution of iHS values for the CEU population in 10-kb windows along chromosome 2 are displayed to the left; windows with a P-value < 0.0005 or P-value < 0.005 in the empirical distribution are highlighted. The candidate region under selection including the LCT gene is zoomed-in to the right, where all significant signatures at four different statistics spanning three different signature types are represented: Protein Changes, Site Frequency Spectrum (SFS) and Linkage Disequilibrium (LD). Signatures in each population are colored according to its meta-population: EUR, AFR, SAS and EAS.

Data availability

Scripts for the PopHumanScan analysis pipeline are available as Jupyter Notebooks at https://github.com/BGD-UAB/PopHumanScan. All data, tools and support resources provided by the PopHumanScan database are freely available at https://pophumanscan.uab.cat. Log-in information to contribute data to PopHumanScan is available upon request.

Acknowledgements

We thank Carla Giner for helpful comments on the PopHumanScan data and implementation, and Esteve Sanz for help with the informatics infrastructure in which PopHumanScan is implemented. We also thank two anonymous referees for very helpful comments on the PopHumanScan implementation and manuscript.

Funding

Ministerio de Economía y Competitividad (Spain) [CGL2017-89160P to Mauro Santos, A.B.]; AGAUR (Generalitat de Catalunya) [2017SGR-1379 to Alfredo Ruiz]; Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement (Generalitat de Catalunya) [FI-DGR2015 to M.C-.Z.]; Departament de Genètica i de Microbiologia (UAB) [PIF to J.M-.M.]; Servei de Genòmica i Bioinformàtica de la UAB. Funding for open access charge: Ministerio de Economía y Competitividad (Spain).

Conflict of interest statement. None declared.

Chapter 4

Imputed McDonald and Kreitman test: a straightforward correction that increases significantly the power of gene-by-gene MKT

Abstract

The McDonald and Kreitman test (MKT) is one of the most powerful and widely used methods to detect and quantify recurrent natural selection in DNA sequence data. One of its main limitations is the underestimation of positive selection due to the presence of slightly-deleterious polymorphisms segregating at low frequencies. Although several approaches have been developed to overcome this limitation, most of them work on gene pooled analyses. Here we present the impMKT, a new straightforward approach for the detection of positive selection and other selection components of the Distribution of Fitness Effect (DFE) at the gene level. We compare impMKT with other widely-used MKT approaches considering both simulated and empirical data. By applying impMKT to human and Drosophila data at the gene level, we substantially increase the statistical evidence of positive selection with respect to previous approaches (e.g. 50% and 157% compared with the MKT in Drosophila and human, respectively). We review the minimum number of genes required to obtain a reliable estimation of the proportion of adaptive substitution (α) in gene pooled analyses comparing impMKT and other MKT implementations. Because of its simplicity and increased statistical power to test recurrent positive selection on genes, we propose impMKT as a first straightforward approach for testing specific evolutionary hypotheses at the gene level. The software implementation and population genomics data is available at the web-server imkt.uab.cat.

Introduction

Natural selection leaves characteristic footprints at the patterns of genetic variation. Since the advent of next-generation sequencing, numerous statistical methods have been proposed to analyze genomic data (Casillas and Barbadilla, 2017), allowing the detection and quantification of molecular adaptation at different temporal scales. The McDonald and Kreitman test (MKT) (McDonald and Kreitman, 1991) is one of the most powerful and robust methods to detect the action of recurrent natural selection at the DNA level. Unlike the ω ratio (Kimura, 1977), which compares the number of synonymous (D_S) and non-synonymous (D_N) divergent sites, the MKT combines both divergence (D_S, D_N) and polymorphism (P_S, P_N) data. Polymorphic data allows taking into account purifying selection on divergent non-synonymous sites, significantly increasing the power of detecting recurrent positive selection.

The null model of the original MKT approach is the neutral theory (Kimura, 1968, 1977; Ohta, 1973). Neutral theory assumes that positively selected (adaptive) mutations get fixed relatively fast compared to neutral mutations, contributing almost exclusively to divergence and not to polymorphism. Therefore, an excess of the divergence ratio relative to the polymorphism is the signal of positive selection acting on non-synonymous sites (D_N/D_S¿ P_N/P_S). Temporally, the MKT covers the evolutionary period spanning from the present to the time back to divergence between the target and the outgroup species, and it allows the estimation of the fraction of adaptive non-synonymous substitutions (α) (Charlesworth, 1994; Smith and Eyre-Walker, 2002). Nonetheless, the MKT, as originally formulated, has multiple drawbacks that could bias the estimation of α. First, the MKT assumes strict neutrality on segregating (polymorphic) sites. However, several studies in multiple species have shown an excess of low-frequency variants (Smith and Eyre-Walker, 2002; Messer and Petrov, 2013a; Galtier, 2016). These variants are attributed to slightly deleterious mutations (SDM), which will not usually reach fixation, contributing more to polymorphism than divergence. SDM reduce the MKT statistical power and underestimate α (Eyre-Walker and Keightley, 2009). Second, MKT assumes that the neutral mutation rate is constant over time, and so is the selective constraint. However, the nearly-neutral mutation rate depends on the effective population size (N_e) (Balloux and Lehmann, 2012; Galtier and Rousselle, 2020; Lanfear et al., 2014; Rousselle et al., 2019) and, therefore, changes in population size can affect the MKT considerably. SDM get fixed at higher rates in populations with past smaller sizes, contributing to divergence and leading to an overestimation of α (Eyre-Walker and Keightley, 2009). Besides, recent evidence shows that weakly advantageous mutations can also be segregating within populations (Galtier, 2016; Tataru et al., 2017; Uricchio et al., 2019). The presence of this positively selected polymorphism, like SDM, can mask the effect of adaptive selection, since it counteracts the excess of the divergence ratio relative to the polymorphism tested by the MKT.

Over the last decades, several modifications of the original MKT have been proposed to account for the potential biases in the estimation of α (Templeton, 1996; Fay et al., 2001; Eyre-Walker and Keightley, 2009; Mackay et al., 2012; Messer and Petrov, 2013a; Galtier, 2016). Most of these extensions deal with the presence of SDM. Although other forces affect the Site Frequency Spectrum (SFS) of segregating variants, such as recombination, demography, ancestral population sizes or weak positive selection, several studies have pointed out the relevance of SDM (Eyre-Walker et al., 2006; Eyre-Walker and Keightley, 2009). SDM distort the non-synonymous SFS, and have been repeatedly shown to be a main factor biasing α downwards (Charlesworth and Eyre-Walker, 2008; Eyre-Walker and Keightley, 2009; Fay et al., 2001; Galtier, 2016).

New model-based approaches for the estimation of α have benefited from the increasing number of genomics data sets available, which allow dealing, implicitly or explicitly, with the underlying Distribution of Fitness Effects (DFE) of new mutations, including the presence of SDM or controlling for correlated genomic features (Eyre-Walker and Keightley, 2009; Galtier, 2016; Messer and Petrov, 2013a; Tataru et al., 2017; Huang, 2021; Uricchio et al., 2019). However, these advanced methodologies need extensive data sets to fit complex parametric evolutionary models by applying maximum likelihood (ML) inference, exponential fitting or generalized linear models and they work properly for genome-wide analyses or on large pools of genes. In contrast, these methodologies are rarely applicable over specific genes to test particular evolutionary hypotheses, as the original MKT does (McDonald and Kreitman, 1991).

While more and more genome-wide analyses of evolution of protein coding genes have been carried out through these MKT extensions, the simple G-test or the independence chi-square test of the original MKT (McDonald and Kreitman, 1991) is currently almost deprecated. Most MKT heuristic alternatives exclude all variants below a frequency threshold for the minor frequency allele (MAF) (Templeton, 1996; Akashi, 1999; Fay et al., 2001). Since the MAF distribution resembles an exponential one, dropping this data inevitably leads to the loss of most of the polymorphic information, consequently performing very poorly on gene-by-gene testing.

Here, we present the imputed MKT (impMKT), a modification of the Fay, Waycoff, and Wu MKT approach (fwwMKT) (Fay et al., 2001) to improve gene-by-gene analyses. We propose a methodology that imputes the proportion of SDM at the SFS rather than removing all variants below a frequency threshold. The impMKT maximizes the information to test the excess of divergence ratio relative to polymorphism at the gene level. We compare our imputation method to previous and recent MKT approaches, using simulated data to test its accuracy and efficiency. Moreover, we test the impMKT on the human African lineage samples of the 1000 Genome Project (1000GP) (Auton et al., 2015) and the Zambian population of the Drosophila Genome Nexus (DGN) (Lack et al., 2016). impMKT considerably increases the number of statistically significant genes under positive selection in Drosophila and humans, respectively, compared to other MKT approaches. Despite the limitations of heuristic MKT and MKT-derived methods, the impMKT has the advantages of simplicity, intuitiveness, ease of use, and increased statistical power to test recurrent positive selection on genes, thus it can be used as a first straightforward approach for testing specific evolutionary hypotheses at the gene level.

Materials and methods

Simulated data

We used `SLiM 3` (Haller and Messer, 2019) to test the accuracy and performance of the impMKT compared to other MKT approaches on simulated data. We tested 15 different genetic scenarios following the procedure proposed by Campos and Charlesworth (2019) and Booker (2020).

We simulated the evolution of a population of 10,000 diploid individuals for 220,000 generations while setting a uniform population-scaled mutation and recombination rates of $4N_e r = 4N_e \mu = 0.001$. To improve performance we re-scaled by a factor of 10 and substitutions were recorded $14N_e$ generations after burn-in following Booker (2020). Each simulation contained seven genes spaced by 8100 pb neutral intergenic regions. For each gene, we simulated five exons of 300 pb separated by 100 pb neutrally-evolving introns. We assumed a proportion of 0.25 and 0.75 for synonymous and non-synonymous alleles, respectively. Deleterious alleles were modeled following a Gamma distribution, whereas beneficial alleles were modeled following a point-mass distribution. We assumed that the Gamma distribution of deleterious alleles followed a shape (β) parameter of 0.3, and population-scaled selection coefficients of $2N_e s_- = 2000$. For beneficial alleles, we assumed a population-scaled selection coefficients $2N_e s_+ = 250$. We solved the analytical approach described in Uricchio et al. (2019) to obtain the fixation probabilities of beneficial alleles considering an adaptation rate value of 0.4. Finally, we used the estimated fixation probabilities to define the relative proportion of beneficial and deleterious alleles as p_a and $0.75 - p_a$ in our model.

We performed 2000 replicas, totalizing 14,000 simulated genes (2000 replicas $\times$ 7 genes), sampling 20 individuals. Besides, seven parameters were modified to test for multiple scenarios (see Table 4.1). Each scenario independently replaces a genetic feature to identify limitations and advantages of the method regarding the underlying DFE, the global adaptation rate, or the number of polymorphic sites.

D. melanogaster and human data

We followed the pipeline described at Murga-Moreno et al. (2019b) to retrieve polymorphic and divergence genome data from *D. melanogaster* and the human lineage.

Table 4.1: `SLiM` simulated parameters.

Simulations	N_e	n	$2N_es_-$	$2N_es_+$	β	p_a	ρ	θ	**Genes**
Base	1000	20	-2000	250	0.3	0.00021	0.001	0.001	14000
$2N_es+ = 500$	1000	20	-2000	500	0.3	0.00012	0.001	0.001	14000
$2N_es+ = 100$	1000	20	-2000	100	0.3	0.00048	0.001	0.001	14000
$2Nes- = 1000$	1000	20	-1000	250	0.3	0.00021	0.001	0.001	14000
$2Nes- = 500$	1000	20	-500	250	0.3	0.00021	0.001	0.001	14000
$\beta = 0.1$	1000	20	-2000	250	0.1	0.00115	0.001	0.001	14000
$\beta = 0.2$	1000	20	-2000	250	0.2	0.00048	0.001	0.001	14000
28000 genes	1000	20	-2000	250	0.3	0.00021	0.001	0.001	28000
1000 genes	1000	20	-2000	250	0.3	0.00021	0.001	0.001	1000
$\rho = 0.01$	1000	20	-2000	250	0.3	0.00021	0.01	0.001	14000
$\rho = 0.0001$	1000	20	-2000	250	0.3	0.00021	0.0001	0.001	14000
$\theta = 0.01$	1000	20	-2000	250	0.3	0.00021	0.001	0.01	14000
$\theta = 0.0001$	1000	20	-2000	250	0.3	0.00021	0.001	0.0001	14000
$\alpha = 0.1$	1000	20	-2000	250	0.3	0.000036	0.001	0.001	14000
$\alpha = 0.1$	1000	20	-2000	250	0.3	0.00075	0.001	0.001	14000

N_e: Effective population size; n: sample size; $2N_es$: population-scaled selection coefficient; Shape parameter of the Gamma distribution; p_a: Relative proportion of advantageous mutations; ρ: population-scaled recombination rate; θ: population-scaled mutation rate; α: proportion of adaptive mutation

In brief, for *D. melanogaster* we retrieved polymorphic and divergence data from the DGN data, using the genome sequence of *D. simulans* as outgroup (release 2) (Lack et al., 2016). Especifically, we subset data from 13,753 protein-coding genes from the Zambian population (197 individuals). We binned the output SFS considering a sample of 20 individuals. The ancestral state of each segregating site was inferred from the sequence comparison with the outgroup species *D. simulans*. The *D. melanogaster* genome reference sequence and annotations correspond to the 5.57 FlyBase release. Gene-associated recombination rate estimates at 100 kb non-overlapping windows were retrieved from Comeron et al. (2012).

For the human lineage, we retrieved polymorphic data and ancestral states for all African populations of the 1000GP Phase III (Auton et al., 2015). We used chimpanzee (*Pan troglodytes*) as the outgroup species to compute human divergence metrics. We downloaded hg19-panTro4 alignment from PopHuman (Casillas et al., 2018). Annotations retrieved from GENCODE (release 27) (Derrien et al., 2012) were used to assess the functional class of each genomic position. Recombination rate values associated with each protein-coding gene were obtained from Bhérer et al. (2017) and correspond to the sex-average estimates. We retrieved polymorphic and divergence data from 20,643 protein-coding genes. We binned the output SFS considering a sample of 20 individuals.

MKT approaches

To test the performance and accuracy of the impMKT, we compared it against four already published heuristic MKT methods: (i) the original MKT (McDonald and Kreitman, 1991); (ii) the Fay, Wickoff and Wu correction (fwwMKT) (Fay et al., 2001); (iii) the extended MKT (eMKT) (Mackay et al., 2012); and (iv) the asymptotic MKT (aMKT) (Messer and Petrov, 2013a) following Haller and Messer (2017) cutoffs. In addition, we included the Grapes software (Galtier, 2016), a Maximum-Likelihood (ML) method fitting the DFE. We ran Grapes using the Gamma-Zero and Gamma-Exponential DFE distributions and estimated α for 100 bootstrap datasets. We measured α confidence interval (CI) through the boundaries for α estimation in Grapes using *α_down* and *α_up* parameters independently for each bootstrapped dataset (Galtier, 2016).

aMKT and Grapes (as well other DFE related methods) are commonly used to estimate α using a large pool of genes or genome-wide data (Messer and Petrov, 2013a; Rousselle et al., 2019). Both methodologies have been previously shown to perform the most accurate estimations in the presence of SDM and demography events (Eyre-Walker and Keightley, 2009; Messer and Petrov, 2013a). Since impMKT is specially designed to perform gene-by-gene analyses, we tried to determine in which cases the amount of data was large enough to perform estimations using aMKT and Grapes compared to impMKT.

Results and discussion

imputed MKT (impMKT)

Our main goal is to devise a derived MKT approach that enhances the power to detect selection at gene-level. To do this, we modified the approach proposed by Fay et al. (2001) (fwwMKT), which removes all non-synonymous (P_N) and synonymous (P_S) polymorphic sites below a derived allele frequency cutoff j, assuming that SDM segregate at low frequencies. Removing variants below a cutoff, typically 5% or 15% (Fay et al., 2001; Mackay et al., 2012), implies losing a considerable amount of data. Consider the example of Nielsen and Slatkin (2013) for the standard neutral coalescence model, a 15% cutoff implies up to 44% of excluded variants of the expected SFS for a sample of $n = 10$ haploid individuals. We observed the same trend considering the *D. melanogaster* and human data, for which considering a virtual gene containing the mean polymorphic level a 15% cutoff implies up to 80% of excluded variants of the expected SFS for *D. melanogaster* and up to 90% in humans. This amount of data exclusion may make the computation of the MKT impracticable, especially in species with low levels of polymorphism.

Here we propose a new MKT approach that modifies the fwwMKT to impute the actual number of SDM (P_{wd}) segregating within P_N. The resulting approach, impMKT, just removes the imputed number of SDM instead of all polymorphism segregating below a given threshold as fwwMKT does, thus retaining a larger fraction of the data and increasing the power to detect positive selection.

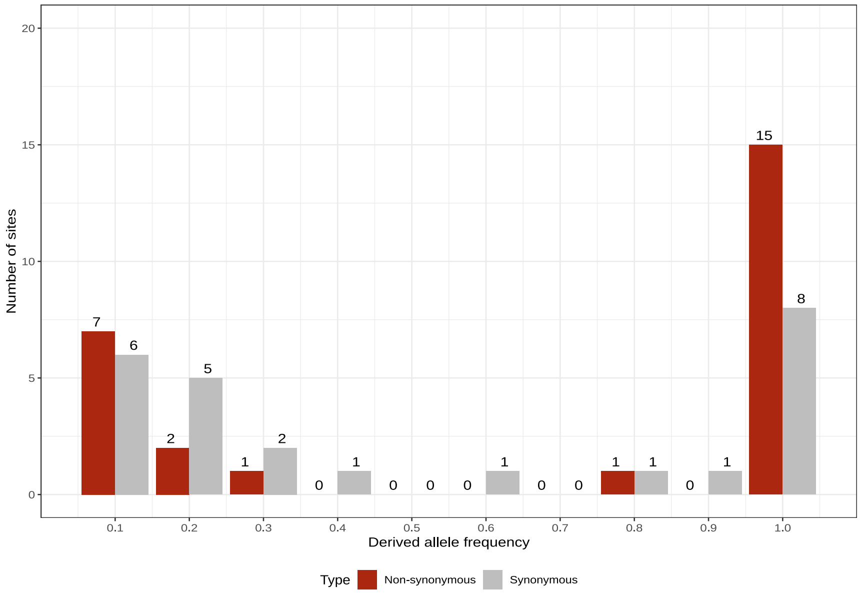


Figure 4.1: Hypothetical SFS and fixed differences from Hahn (2018)

Consider the SFS and fixed differences of a hypothetical gene as illustrated by Hahn (2018) (Figure 4.1). Table 4.2 shows the 2 × 2 contingency tables to perform the original MKT, fwwMKT and impMKT. Charlesworth and Eyre-Walker (2008) investigated how the removal of low-frequency polymorphism affects the estimation of at MKT approaches depending on the continuous function defining the DFE for different non-arbitrary cutoffs. To develop the impMKT, we followed Charlesworth and Eyre-Walker (2008) results, which show that any derived allele frequency cutoff $j > 15\%$ is a near-optimal solution to the problem of SDM segregating at the SFS (Charlesworth and Eyre-Walker, 2008).

Consequently, considering that SDM are the main force biasing downward α and assuming that SDM do not segregate at frequencies above 15% ($P_{wd} \rightarrow 0$), we impute the actual proportion of SDM (P_{wd}) segregating below the frequency cutoff by considering that the expected neutral polymorphism non-synonymous/synonymous ratio is $P_{N_{(j>15\%)}}/P_{S_{(j>15\%)}}$. This ratio can be used to infer the number of SDM in our data set (P_{wd}). If $P_{wd} \neq 0$ bellow $j < 15\%$, then $P_{N_{(j<15\%)}}/P_{S_{(j<15\%)}}$ exceeds the expected polymorphic ratio because $P_{N_{(j<15\%)}}$ includes P_{wd}. That is, $P_{N_{(j<15\%)}} = P_{neut_{(j<15\%)}} + P_{wd_{(j<15\%)}}$, where $P_{neut_{(j<15\%)}}$ refers to the number of non-synonymous segregating sites that are effectively neutral. Accordingly, we can estimate (impute) P_{wd} from expression

$$\frac{P_{N(j<15\%)} - P_{wd(j<15\%)}}{P_{S(j<15\%)}} = \frac{P_{N(j>15\%)}}{P_{S(j>15\%)}} \tag{4.1}$$

rearranging we have

$$P_{wd} \approx P_{wd(j<15\%)} = P_{N(j<15\%)} - \frac{P_{N(j>15\%)} \cdot P_{S(j<15\%)}}{P_{S(j>15\%)}} \tag{4.2}$$

Considering our example in Table 4.2, P_{wd} is

$$P_{wd} = 7 - \frac{4 \cdot 6}{11} \approx 5 \tag{4.3}$$

and thus 5 is the number of sites removed from the non-synonymous polymorphism counts (see Table 4.2-D).

As can be seen in Table 4.2-C, the approach proposed by Fay et al. (2001) shows that removing all low-frequency polymorphisms below a given threshold j significantly increases the power of detection of positive selection by conducting a 2 × 2 test. Thus testing for the ratio of replacement on fwwMKT 2 × 2 contingency table through a Fisher exact test decreases the P-value significance from 0.093 to 0.045 in our example. Nonetheless, it implies a reduction of 46% of the analyzed data, reducing P_N from 11 to 4 and P_S from 17 to 11, respectively. In comparison, by simply removing the expected number of SDM (P_{wd}), we reduced the data loss to only 15%, while decreasing the P-value from 0.093 to 0.017 (see Table 4.2-D). Therefore, the impMKT allows maximizing gene-by-gene analyses where information can be limited to a small number of polymorphic sites.

In addition, we can correct α, the proportion of adaptive substitutions, by removing the expected proportion of SDM (P_{wd}) with the expression

$$\alpha_{imputed} = 1 - \left(\frac{P_N - P_{wd}}{P_S} \cdot \frac{D_N}{D_S} \right) \tag{4.4}$$

Table 4.2: Original and impMKT contigency table

A. Definition of the MKT 2×2 contingency table.

	Polymorphism	**Divergence**
Non-synonymous	$P_{Neutral} = P_N - P_{wd}$	P_S
Synonymous	D_N	D_S

B. Example of MKT $2x2$ contingency table. Including all polymorphic sites.

	Polymorphism	**Divergence**
Non-synonymous	11	15
Synonymous	17	8

$2x2$ Fisher exact test; P-value = 0.093

C. Example of fwwMKT $2x2$ contingency table. Removing all polymorphic sites with a derived allele frequency below 15%

	Polymorphism	**Divergence**
Non-synonymous	***11 - 7 = 4***	15
Synonymous	***17 - 6 = 11***	8

$2x2$ Fisher exact test; P-value = 0.045

D. Example of impMKT $2x2$ contingency table. Removing the expected SDM with a derived allele frequency below 15% (see equation (4.2))

	Polymorphism	Divergence
Non-synonymous	***11 - 5 = 6***	15
Synonymous	17	8

$2x2$ Fisher exact test; P-value = 0.017

Other selection regimes. The SDM imputation can be used to estimate other selective components shaping the DFE. Let consider the model proposed by Eyre-Walker and Keightley (2009) and nearly-neutral theory (Ohta, 1973), where selected segregating alleles are drawn from a continuous Gamma distribution and categorized as strongly deleterious, slightly deleterious and effectively neutral mutations. Analogous to Mackay et al. (2012), we define the statistics d, d_w and d_0, which measure the different types of purifying selection, both at genome-wide and gene levels.

Let d be the proportion of strongly deleterious mutations. We estimated d

following Mackay et al. (2012) as the missing fraction of segregating non-synonymous sites

$$\hat{d} = 1 - \frac{P_N}{P_S} \cdot \frac{m_S}{m_N} \tag{4.5}$$

where m_S and m_N are the total number of synonymous and non-synonymous sites, respectively.

Let d_w be the fraction of slightly deleterious mutations at non-synonymous sites

$$\hat{d_w} = \frac{P_{wd}}{P_S} \cdot \frac{m_S}{m_N} \tag{4.6}$$

Lastly, the fraction of effectively neutral mutations d_0 can be estimated as the remaining fraction

$$\hat{d_0} = 1 - d - d_w \tag{4.7}$$

Properties of the impMKT estimator

We tested the accuracy and performance of the impMKT compared to other MKT approaches at estimating the fraction of substitutions fixed by positive selection (α) under different scenarios that were simulated using SLiM 3 (Haller and Messer, 2019). The different scenarios considered the combined effects of different genetic features: the level of polymorphism in terms of segregating sites (θ), the number of simulated genes, the proportion of adaptive mutations (p_a), the proportion of SDM (β), the recombination rate (ρ) and the selection strength ($2N_es$) (Table 4.1). In addition to $j > 15\%$, we explored derived allele frequency cutoffs larger than 15% ($j > 25\%$ and $j > 35\%$). We also tested 5% ($j > 5\%$) frequency cutoff as in Mackay et al. (2012).

In all simulations, the original MKT underestimates considerably the α values (Figure 4.2) due to the presence of SDM segregating at low frequencies, excluding simulations where the contribution of SDM is negligible. Overall, the aMKT and Grapes performed better under the presence of SDM and achieved the best results when considering both unbiasedness and efficiency of the estimator (minimum variance)

(Figure 4.2, Table 4.3). While heuristic MKT approaches tend to underestimate α, Grapes tends to slightly overestimate α in most of the scenarios, while aMKT tends to provide slight underestimations (Figure 4.2, Table 4.3, Table B.1).

As previously shown in Charlesworth and Eyre-Walker (2008), α estimates converge to the actual value depending mainly on the shape of the DFE (β) and the amount of adaptive evolution (α). We considered three different values of β (0.3 (baseline), 0.2, and 0.1) to test such effect. We observed the same trend for all MKT-derived approaches: the underestimation for the different MKTs is smaller the more leptokurtic DFE is, which in turn implies less SDM. The same effect was found when increasing the rate of adaptive evolution (from $\alpha = 0.1$ to $\alpha = 0.7$, Table 4.3, Table B.1, Figure B.1).

For all the simulated scenarios, the fwwMKT and the impMKT behave similarly to the MKT, mainly depending on the frequency cutoff. As expected, lower cutoffs (i.e., 5%) resulted in minor accuracy improvements in the estimation of α compared to the original MKT approach, except when SDM contributed little to the SFS (smaller β and larger α values). Conversely, larger cutoffs (i.e., 15%, 25% or 35%) resulted in better estimates of α. Specifically, a 35% cutoff is large enough to deal with SDM and to perform estimations similar to the aMKT and Grapes in all simulations. Both the impMKT and the fwwMKT performed very similarly due to the large amount of data considered from the simulations. Contrarily, the eMKT was not able to deal with the presence of SDM and higher frequency cutoffs did not improve the estimations of α (Figure 4.2, Figure B.1, Figure B.2, Table 4.3, Table B.1; see Discussion).

In scenarios simulating low levels of polymorphism in terms of segregating sites (i.e., reduced number of simulated genes, or reduced mutation rate θ), the accuracy and efficiency of the aMKT and Grapes diminishes (Figure B.1, Figure B.4 and Table B.2). Under these circumstances, the aMKT could be applied to approximately 70% of the cases only, and provided worse estimations of α than the impMKT. We observe the same trend when measuring the standard deviation of the estimators. impMKT provided better results in comparison to aMKT while showing similar accuracy to Grapes (Table B.1, Figure B.1). Similarly, the confidence intervals (CI) estimated by Grapes increased by one order of magnitude, from range [0.01,0.06] (considering the other scenarios) to 0.16 (for the scenario with 2000 simulated genes) and 0.19 (for the scenario with $\theta = 0.0001$) (Figure 4.2, Figure B.1, Figure B.3, Figure B.4).

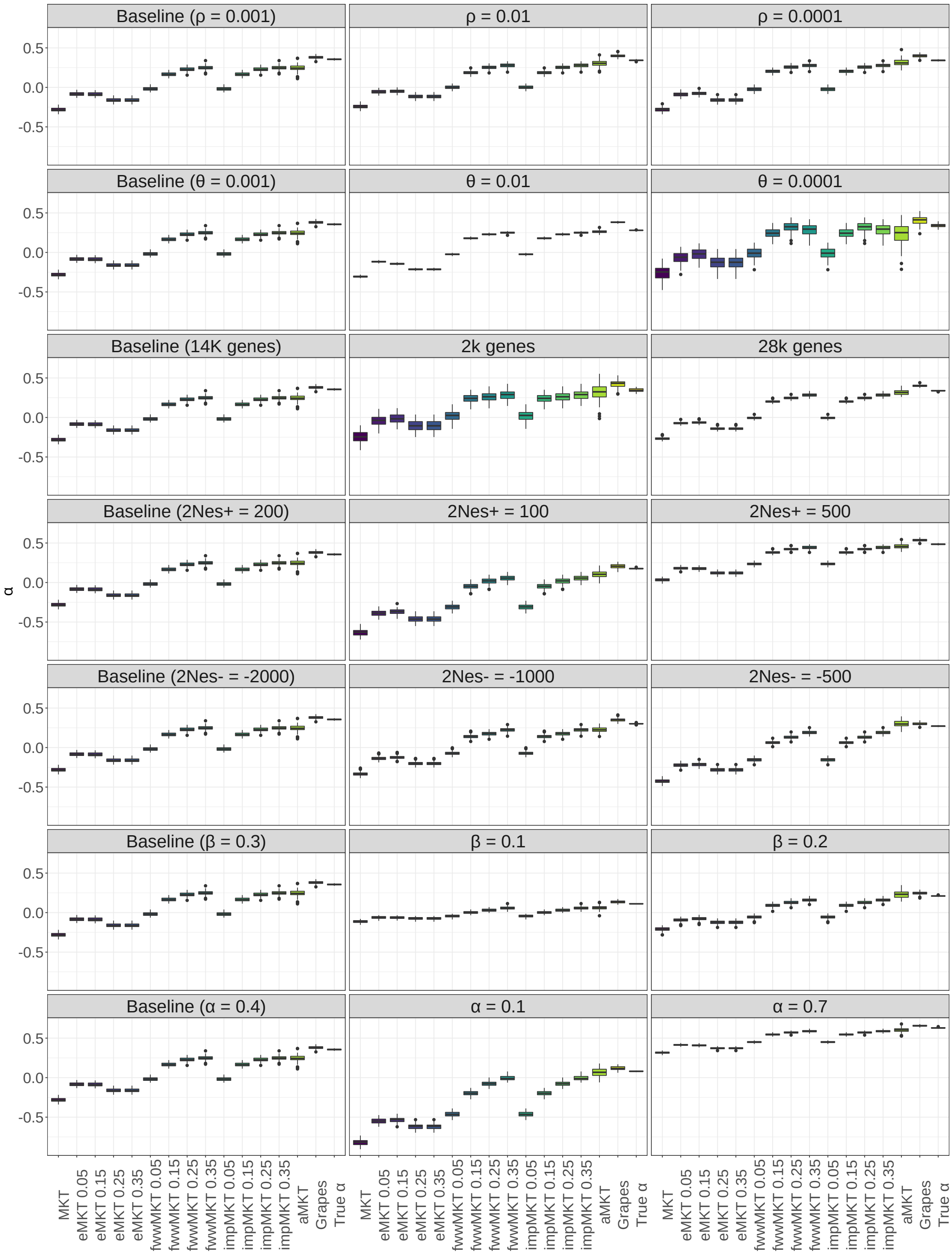


Figure 4.2: α MKT estimations on different SLiM simulated scenarios. ρ and θ are the population-scaled recombination and mutation rates ($\rho = 4N_er$, $\theta = 4N_e\mu$). $2N_es$ is the scaled-population selection coefficient for beneficial and deleterious alleles. β is the shape parameter of the gamma DFE.

Table 4.3: Mean error bias for each scenario and MKT approach. Error bias were measured through the difference of mean values of α for each method and the true value of α.

Simulations	MKT	eMKT 0.05	eMKT 0.15	eMKT 0.25	eMKT 0.35	fwwMKT 0.05	fwwMKT 0.15	fwwMKT 0.25	fwwMKT 0.35	impMKT 0.05	impMKT 0.15	impMKT 0.25	impMKT 0.35	aMKT	Grapes
Baseline	0.638	0.441	0.443	0.517	0.517	0.374	0.189	0.127	0.106	0.374	0.189	0.127	0.106	0.111	0.025
$2N_es_- = -1000$	0.634	0.438	0.425	0.502	0.502	0.373	0.161	0.125	0.079	0.373	0.161	0.125	0.079	0.074	0.049
$2N_es_- = -500$	0.698	0.493	0.485	0.553	0.553	0.425	0.21	0.141	0.081	0.425	0.21	0.141	0.081	0.03	0.028
$2N_es_+ = 100$	0.812	0.565	0.546	0.639	0.639	0.487	0.226	0.159	0.122	0.487	0.226	0.159	0.122	0.073	0.029
$2N_es_+ = 500$	0.451	0.304	0.308	0.363	0.363	0.249	0.102	0.062	0.042	0.249	0.102	0.062	0.042	0.027	0.053
$\rho = 0.0001$	0.624	0.432	0.42	0.503	0.503	0.365	0.138	0.085	0.063	0.365	0.138	0.085	0.063	0.027	0.058
$\rho = 0.01$	0.585	0.4	0.391	0.458	0.458	0.343	0.156	0.09	0.066	0.343	0.156	0.09	0.066	0.037	0.055
$\beta = 0.1$	0.224	0.173	0.175	0.185	0.185	0.157	0.11	0.082	0.055	0.157	0.11	0.082	0.055	0.051	0.023
$\beta = 0.2$	0.42	0.306	0.288	0.334	0.334	0.268	0.12	0.083	0.051	0.268	0.12	0.083	0.051	0.022	0.037
$\theta = 0.0001$	0.6	0.408	0.364	0.467	0.467	0.35	0.098	0.018	0.057	0.35	0.098	0.018	0.057	0.113	0.066
$\theta = 0.01$	0.586	0.397	0.425	0.494	0.494	0.304	0.101	0.051	0.031	0.304	0.101	0.051	0.031	0.017	0.102
$\alpha = 0.1$	0.901	0.629	0.616	0.699	0.699	0.543	0.278	0.158	0.087	0.543	0.278	0.158	0.087	0.014	0.038
$\alpha = 0.7$	0.313	0.216	0.221	0.257	0.257	0.179	0.083	0.059	0.041	0.179	0.083	0.059	0.041	0.027	0.028

Statistical power of the impMKT at gene-by-gene analysis

We estimated α at the gene level on *D. melanogaster* (Zambia, ZI; 197 individuals) and human (Africa, AFR; 661 individuals) population data. Table 4.4 shows the mean values and the number of analyses performed considering the MKT approaches. We removed from the analysis those genes with zero divergence or zero polymorphism, either for synonymous or nonsynonymous sites.

Due to the amount of raw data, the original MKT was the approach that allowed us to estimate α on the largest number of protein-coding genes: 12,024 (87%) genes in the *D. melanogaster* Zambian population. The statistical significance for both positively and negatively selected genes was determined using the Fisher's exact test; 1,495 and 1,331 were detected under positive and negative selection, respectively. The number of analyzable genes decreased 14% when applying the eMKT correction, from 12,024 to 10,340, but slightly increasing the number of genes under positive selection, from 1,493 to 1,571. We found a decreased of 37% when applying the fwwMKT correction, from 12,024 to 7,574 genes, as well as in the number of genes under positive selection, from 1,495 to 929. More importantly, for both approaches we found a drop in the number of genes under negative selection, from 1,131 to 700 and 38 genes for eMKT and fwwMKT respectively.

The impMKT was able to analyze the exact same number of genes as the fwwMKT approach (7,588 genes), since impMKT needs data to compute the P_N/P_S ratio above the threshold, as the fwwMKT. However, the number of positively selected genes increased from 1495 in the original MKT approach or 929 in the fwwMKT to 2,244 (Figure 4.3-A). Therefore, the impMKT increased the detection of positive selection by 50% in the *D. melanogaster* Zambian population compared to the original MKT (1,495 vs 2,244 genes), by 141% compared to the fwwMKT (929 vs 2,244) and by 42% regarding eMKT (from 1,571 to 2,242). In addition, the impMKT also detected 792% more genes under negative selection than the fwwMKT correction (from 38 in the fwwMKT to 339 genes in the impMKT). We noted a significant drop in the number of genes under negative selection regarding the MKT and eMKT. Nonetheless, since neither MKT nor eMKT are able to deal properly with SDM, as shown in simulations, such trend was not unexpected.

We found similar patterns for the human dataset regarding the MKT and fwwMKT. MKT was the methodology that estimated α on the largest number of genes (13,078, 68%), as expected, while fwwMKT and impMKT only analyzed 3,145 genes. Nonetheless, the increase in the number of genes under positive selection detected by the impMKT is especially significant, rising by 159% (from 79 positively selected genes in the MKT to 203 in the impMKT) (Figure 4.3-B), and the fwwMKT only detected 18 genes under positive selection. Interestingly, contrary to *D. melanogaster* data eMKT detected less genes than MKT under positive selection. Considering eMKT results from simulations regarding SDM and the associated protein-coding DFE in humans (Booker, 2020), we determined that eMKT very sensitive to the underlying DFE.

Overall, in populations with low levels of polymorphism, the impMKT allowed detecting genes under positive selection more efficiently than the other methodologies because it does not remove all the data below a threshold, as the fwwMKT does. By just removing the imputed fraction of SDM, the impMKT can maintain a reasonably good statistical power and, contrary to the fwwMKT, is able to analyze data from datasets with low levels of polymorphism, such as human data. We do not tested aMKT nor Grapes since both methods are not performant or are inaccurate on single-gene sequence data and preferably used in large pools of genes or genome-wide levels.

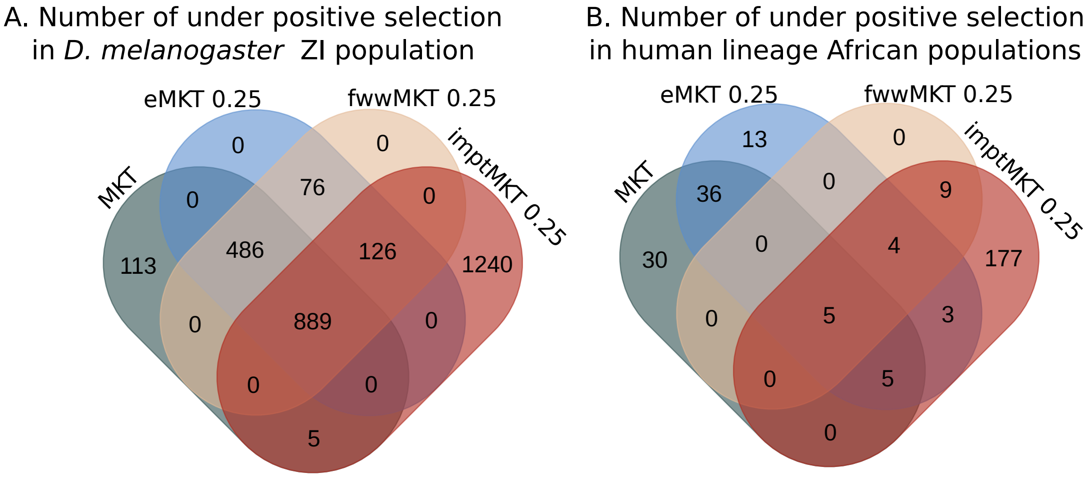


Figure 4.3: A. Positively selected genes in the in the Drosophila Zambian population as detected by each MKT approach. B. Positively selected genes human African population as detected by each MKT approach.

Table 4.4: Gene-by-gene analysis. Total number of analyzable, positively and negatively selected genes by MKT approach

		MKT		eMKT 0.25		fwwMKT 0.25		impMKT 0.35	
Population	Set	α	N	α	N	α	N	α	N
ZI	Analyzable	-0.721 ± (2.823)	12069	-0.376 ± (1.54)	7588	-0.032 ± (1.664)	7588	-0.032 ± (1.664)	7588
ZI	Negative	-4.907 ± (7.044)	1131	-3.526 ± (3.12)	586	-10.558 ± (10.472)	38	-4.698 ± (4.888)	339
ZI	Positive	0.762 ± (0.135)	1495	0.728 ± (0.129)	1136	0.844 ± (0.095)	929	0.775 ± (0.121)	2244
AFR	Analyzable	-1.688 ± (3.188)	13078	-1.17 ± (2.304)	3230	-0.679 ± (2.21)	3230	-0.679 ± (2.21)	3230
AFR	Negative	-7.408 ± (6.323)	1037	-4.864 ± (4.402)	338	-12.695 ± (5.783)	11	-5.375 ± (4.676)	244
AFR	Positive	0.816 ± (0.116)	79	0.753 ± (0.173)	21	0.893 ± (0.093)	18	0.759 ± (0.121)	205

Statistical power of the impMKT in gene pooling

Next, we explored the performance of the impMKT, compared to the aMKT, Grapes and the original MKT approach, on pooled gene data. By adding up polymorphism and divergence data from multiple genes, this type of analysis increases the number of polymorphic sites to estimate the SFS, which provides the statistical power necessary to implement both the aMKT and ML approaches. We created gene pools to obtain a reliable measure of the average α. Specifically, we first selected 3,500 random protein-coding genes from both the Drosophila and the human datasets. Then, we resampled the genes 1,000 times with replacement to create pools of 1, 2, 5, 10, 25, 50, 75, 100, 250, 750, and 1,000 genes on which we computed the SFS and estimated α (Table 4.5).

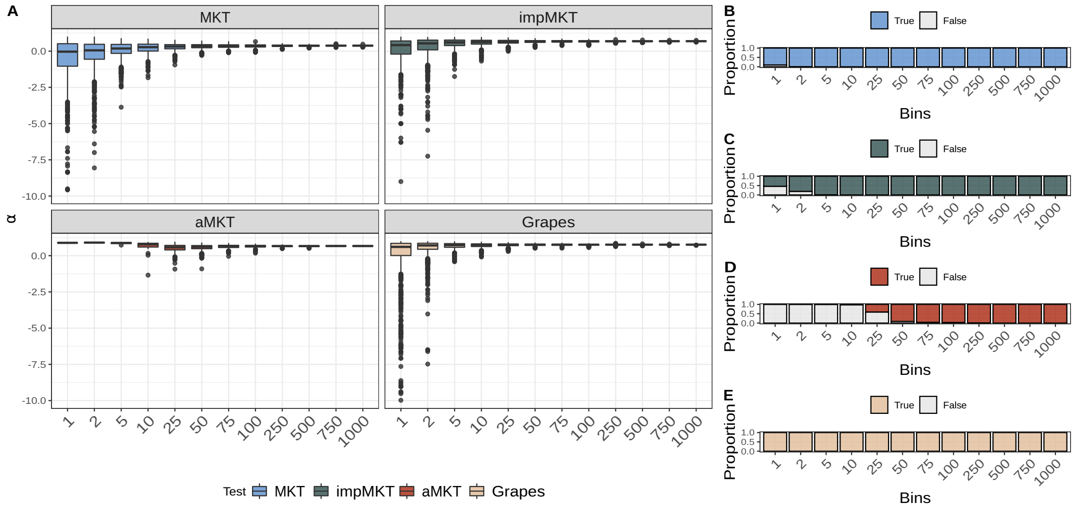


Figure 4.4: Gene pooled analysis. 3500 random protein-coding genes were picked from the ZI dataset. We pooled the genes to obtain SFS of 1, 2, 5, 10, 25, 50, 75, 100, 250, 750, and 1,000 genes by resampling them 1000 times with replacement. A. estimates by MKT correction. B. Proportion of analysis performed by impMKT. C. Proportion of analysis performed by aMKT. D. Proportion of analysis performed by Grapes

***D. melanogaster* ZI population.** Resampling analysis results in the *D. melanogaster* ZI population showed that estimated α converges to an average value as more and more genes are pooled (Figure 4.4-A). First, in the case of impMKT, pools of 5 genes or more already allowed estimating α in 90% of the cases, reaching 100% in pools of 10 or more genes (Figure 4.4-B). Second, aMKT required larger pools to analyze the data; pools of 50 genes or more allowed estimating in 90% of the cases, while 500 or more genes were required to estimate α in all of the replicates (Figure 4.4-C).

Third, MKT and Grapes could analyze the vast majority of replicates (except for a few replicates in bins with only 1 or 2 pooled genes). Nonetheless, we noted 1.9-fold (from 1.2 to 2.3) and 8-fold increase (from 1.2 to 9.7) in α variance regarding MKT and Grapes compared to impMKT respectively at the first pool, showing the lack of power on small dataset. As the number of genes grow, the mean converging value of α was very similar for the impMKT, the aMKT and MKT, and higher for Grapes (Figure 4.4-A), an expected result considering previous results with simulated data (see previous section). In addition, impMKT showed similar (or higher) α values than aMKT and was applicable to the smallest gene pools.

Human protein-coding genes. Due to the low polymorphism levels in human protein-coding genes compared to *D. melanogaster*, the minimum number of genes pooled to estimate accurate measures of α was larger, especially for aMKT (Figure 4.5). Specifically, aMKT required pools of 500 genes or more to estimate α in 90% of the of the replicas, while more than 1000 were required to estimate α in all of the replicates (Figure 4.5-C). In the case of Grapes, we found most of the analysis can be performed but showing the 1.7-fold (from 3.4 to 5.8) increase in α variance regarding impMKT estimations. impMKT could estimate most replicates with 5 or more genes pooled, and all of the replicates with 25 or more genes pooled (Figure 4.5-B), showing similar or higher α values than aMKT.

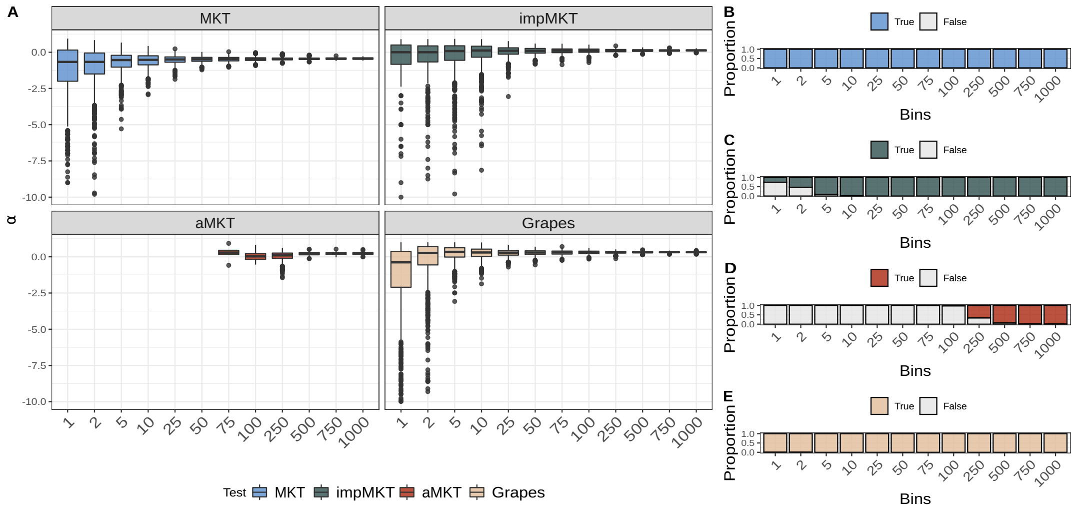


Figure 4.5: Gene pooled analysis. 3500 random protein-coding genes were picked from the human dataset. We pooled the genes to obtain SFS of 1, 2, 5, 10, 25, 50, 75, 100, 250, 750, and 1,000 genes by resampling them 1000 times with replacement. A. α estimates by MKT correction. B. Proportion of analysis performed by impMKT. C. Proportion of analysis performed by aMKT. D. Proportion of analysis performed by Grapes

Table 4.5: α estimates by pooled genes. Each bin number corresponds to the number of pooled genes. Mean estimates and 95 percentiles estimates are shown by MKT approach and bin.

Bins		**1**	**2**	**5**	**10**	**25**	**50**	**75**	**100**	**250**	**500**	**750**	**1000**
Population	**Test**												
ZI	**impMKT**	-0.015 (-2.332 - 0.915)	0.245 (-1.335 - 0.915)	0.528 (-0.133 - 0.898)	0.596 (0.176 - 0.874)	0.641 (0.403 - 0.828)	0.662 (0.486 - 0.799)	0.67 (0.53 - 0.783)	0.674 (0.557 - 0.774)	0.682 (0.611 - 0.75)	0.684 (0.633 - 0.729)	0.686 (0.647 - 0.722)	0.686 (0.652 - 0.717)
ZI	**aMKT**	0.888 (0.888 - 0.888)	0.876 (0.865 - 0.894)	0.849 (0.817 - 0.877)	0.661 (0.25 - 0.864)	0.56 (0.17 - 0.821)	0.6 (0.352 - 0.775)	0.628 (0.448 - 0.767)	0.634 (0.482 - 0.762)	0.645 (0.544 - 0.729)	0.65 (0.583 - 0.709)	0.654 (0.595 - 0.703)	0.656 (0.607 - 0.696)
ZI	**Grapes**	-0.569 (-5.926 - 1.0)	0.4 (-0.843 - 0.992)	0.658 (0.177 - 0.927)	0.704 (0.394 - 0.916)	0.738 (0.551 - 0.876)	0.752 (0.623 - 0.86)	0.758 (0.655 - 0.848)	0.762 (0.672 - 0.839)	0.768 (0.714 - 0.82)	0.769 (0.73 - 0.804)	0.771 (0.74 - 0.799)	0.771 (0.745 - 0.795)
AFR	**impMKT**	-0.767 (-5.928 - 0.86)	-0.746 (-4.25 - 0.78)	-0.423 (-3.471 - 0.76)	-0.191 (-2.008 - 0.74)	0.023 (-0.862 - 0.6)	0.077 (-0.43 - 0.491)	0.081 (-0.361 - 0.429)	0.093 (-0.269 - 0.39)	0.098 (-0.142 - 0.306)	0.101 (-0.049 - 0.236)	0.101 (-0.015 - 0.213)	0.099 (-0.002 - 0.199)
AFR	**aMKT**	nan (nan - nan)	nan (nan - nan)	nan (nan - nan)	nan (nan - nan)	0.487 (0.2 - 0.665)	0.25 (-0.245 - 0.744)	0.086 (-0.555 - 0.573)	0.082 (-0.586 - 0.536)	0.2 (-0.097 - 0.456)	0.189 (0.031 - 0.362)	0.176 (0.051 - 0.327)	0.166 (0.057 - 0.286)
AFR	**Grapes**	-1.66 (-8.995 - 1.0)	-0.922 (-6.572 - 1.0)	0.119 (-1.155 - 0.954)	0.19 (-0.605 - 0.76)	0.221 (-0.273 - 0.615)	0.237 (-0.121 - 0.537)	0.245 (-0.059 - 0.502)	0.245 (-0.005 - 0.463)	0.257 (0.101 - 0.396)	0.26 (0.152 - 0.353)	0.264 (0.175 - 0.342)	0.265 (0.194 - 0.335)

Discussion

Effect of slightly deleterious mutations (SDM) on α estimation

SDM segregating at low frequencies impact the power of MKT and the estimation of α (Akashi, 1999; Fay et al., 2002; Galtier, 2016; Messer and Petrov, 2013a; Fay et al., 2001; Templeton, 1996; Bustamante et al., 2002a, 2005; Bierne and Eyre-Walker, 2004). As Bierne and Eyre-Walker (2004) pointed out, unless the methodology considers the presence of SDM, estimations using *D. melanogaster* data are likely underestimating α. We verify such statements by thoroughly exploring the MKT-derived approaches using both in silico and empirical data, assessing the benefits and drawbacks of each methodology, considering the nature of the data and the study design. Simulations with SLiM 3 have been carried out to benchmark the performance of the four MKT methodologies and the impMKT under different evolutionary scenarios. Predefined α values were used to assess the closest estimation. aMKT and Grapes are the best methods with respect to unbiasedness and efficiency of estimated values of α. However, their performance decreases in scenarios with a small number of polymorphic variants (shorter genomic regions or lower mutation rate) or could not even be applied due to low variant counts. Our results are consistent with previous explorations of MKT-derived approaches (Charlesworth and Eyre-Walker, 2008; Messer and Petrov, 2013a). Hence, we found similar results exploring aMKT and Grapes in Drosophila and human genome sequence data and showed similar accuracy in simulations. Overall, both approaches allow efficient removal of SDM in all frequencies and not only below a threshold as in fwwMKT or impMKT methods.

Strikingly, both procedures lack power when applied to individual genes or small pooled datasets. Despite the high polymorphic and divergence levels in *D. melanogaster*, it is not enough for the aMKT to fit the exponential curve and calculate α for single genes, and the number of analyzable genes is dramatically reduced (see Table 4.4). We showed that pooled sets of genes allow overcoming data limitations to estimate an overall α value (Boyko et al., 2008; Eyre-Walker and Keightley, 2009). Thus we explored the minimum number of pooled genes to perform aMKT regarding *D. melanogaster* and human populations. For aMKT we found that a minimum of 500 genes is required to perform 1000 replicas when bootstrapping a set of 3500 random genes (Figure 4.4). Such a number increased to more than 1000 when using the human dataset (Figure 4.5). We found that Grapes can perform the estimation most of the time (only a few negligible analyses were not performed, see Figure B.4, Figure B.5), considering gene-by-gene analysis or pooled analysis. Nonetheless, we found extremely high variance in

α estimates and we noted that the associated CI to α estimation for each bootstrapped datasets is only acceptable once the analysis accounts for a minimum number of 50 genes in Drosophila and humans (see Figure B.8). The same trend is observed in those simulated scenarios producing less polymorphism regarding the percentage of aMKT analysis and Grapes CIs (Figure 4.2, Figures B.3). The results for both populations can be considered as a generalized proxy, given the high levels of polymorphism in *D. melanogaster* compared to humans.

Such findings show the limitation of aMKT and Grapes (and other ML methods) when performing MKT at the gene-by-gene level or using small pooled datasets (Eyre-Walker and Keightley, 2009; Racimo and Schraiber, 2014; Tataru et al., 2017). Among non-ML approaches, fwwMKT and impMKT produced quite similar results. However, only when using higher frequency cutoffs than the commonly-used 15% they showed results close to those by aMKT and Grapes (see Table 4.2, Table 4.3) although such statement will depends on the underlying DFE. Such cutoffs can be astringent considering empirical data, especially in the case of fwwMKT. Instead of removing all polymorphism at low frequencies at both synonymous and non-synonymous sites, as fwwMKT does, the new impMKT separates P_N into the number of effectively neutral variants and the number of SDM, and only removes the latter. In this way, impMKT allows increasing the frequency cutoff without compromising the amount of data that much. As a result, impMKT is the most powerful method to detect selection at the gene-by-gene level, substantially increasing the number of statistically-significant genes under positive selection compared to other methodologies (see Figure 4.3 and Table 4.4). In the case of pooled analyses, impMKT reduced dramatically the minimum number of genes required to perform the analysis in both Drosophila and human datasets (5 and 10 respectively, B.3, Figure B.4).

Even though strongly deleterious (d), slightly deleterious (d_w) and effectively neutral (d_0) mutations are commonly defined given DFE ranges $-10 < N_e s$, $-10 < N_e s < -1$ and $N_e s > -1$, respectively, we observed mutations segregating in the range $-10 < N_e s$. Hence, if d is the proportion of no segregating mutations because of strong purifying selection, as stated above, we estimated d_w including any segregating mutation below the threshold $N_e s < -1$. Table B.4 and Figure B.5 show impMKT unbiased estimations of d, d_w and d_0 using 5% and 35% cutoffs. Similarly to α estimation, the estimator require larger cutoff than 5-15% (Charlesworth and Eyre-Walker, 2008; Mackay et al., 2012) to properly impute SDM and estimate d_w and d_0 accurately. Hence, the new impMKT provides easier and faster estimations of d, d_w and d_0 than ML approaches, representing the actual mutation proportions subject to different selection regimes and quantitative measures of the DFE along the genome or at the gene level.

The effect of pooling data

We showed that most MKT approaches could provide an accurate estimate of the average α if data from a large number of genes are collected (Hahn, 2018). Therefore, the process of pooling genes to create single evolutionary entities is a proper strategy to overcome the problem of lacking enough polymorphism data to conduct an MKT. In the majority of the performed analyses, this process does not seem to affect the results. However, some caveats must be taken into account when interpreting results obtained by this procedure.

First, pooled genes do not necessarily share the same recombination context, GC-content, or gene density rate, which also affect the adaptive potential of genes. Although pooling genes by one or more features at a time have been widely used to disentangle the potential drivers of adaptation (Castellano et al., 2016; Moutinho et al., 2019b; Soni et al., 2021; Uricchio et al., 2019), such approaches can report a spurious association between adaptation signals and other features if they are strongly correlated (Huang, 2021). Huang (2021) developed the so-called MK-regression to overcome biases of pooling analyses applied to one genomic feature at a time, by jointly evaluating the effects of correlated genomic features on α estimation. Nonetheless, MK-regression is designed to measure the adaptation rate at the genomic level, and consequently not the preferred approach to pinpoint individual genes neither (Huang, 2021). Interestingly, we have noticed that MK-regression followed the strategy proposed by Fay et al. (2001) to deal with SDM. We propose to apply our impMKT approach instead, to preserve data and extending the implementation at the gene-by-gene level.

Second, by pooling hundreds of genes, it is more difficult to detect a signal of positive selection if it is due to a few genes of the pool. In other words, all the evolutionary forces acting differently on different genes contribute to the dilution of potential biological signals.

Third, although this data pooling increases the power of detecting selection, it could lead to the Simpson's paradox (Simpson, 1951) if a significant trend in the 2×2 contingency tables disappears or reverses when the data is combined into a single table (Hahn, 2018; Stoletzki and Eyre-Walker, 2011). Regarding MKT data, this can happen when large differences in the number of non-synonymous fixations (D_N) between genes lead to incorrect inferences about the selection operating in different regions (Stoletzki and Eyre-Walker, 2011).

α estimation on the presence of recent positive selection

Several studies have showed the contribution of slightly beneficial mutations (SBM) to the SFS at medium/high-frequency over the last years, representing a new distortion source in the MKT approaches. These alleles can segregate in the frequency spectrum and eventually fix in the population depending on the selective strength. Multiple methods have been proposed to overcome this limitation (Galtier, 2016; Tataru et al., 2017). Nonetheless, many natural patterns remain unanswered, and they can be attributed to the effect of linked selection, since methods that incorporate weak selection assume that sites evolve independently. Uricchio et al. (2019) proposed a new MKT approach that incorporates background selection (BGS), estimates the fraction of weak selection and discerns the role of linkage in α estimations.

We tested such effect following Uricchio et al. (2019) simulations to evaluate SBM as well as BGS. We simulated the exact global adaptation rate as in the baseline simulation and 50% of α corresponded to the contribution of weakly advantageous alleles following a point-mass distribution with selection coefficient $2N_es = 5$.

In addition to the contribution of SBM to the fixation process, one expects a higher concentration of SBM at high frequencies, since the Hill-Robertson effect prevents them to reach fixation due to linkage to other SBM or SDM whether BGS is acting. Under these assumption, we modify the impMKT approach to account for such an excess of non-neutral alleles at high frequencies. The proposed modification would follow the main assumptions described for SDM, in this case exploring a new frequency cutoff at high frequencies to remove SBM, while incorporating the estimated excess to fixations. In addition, we executed Grapes using the Gamma-Exponential model and considered adaptive mutations using $2N_es > 5$ threshold.

Despite the possible excess at high frequencies, SBM may be seen to segregate across the spectrum, depending not only on the Hill-Roberston effect but also on selective strength, linkage disequilibrium patterns and fixation times (Figure B.6). Assuming that SBM can segregate at any frequency, impMKT cannot deal with weak adaptation, even imputing nearly fixed variants. Therefore our heuristic approach, extending aMKT results from Uricchio et al. (2019), can also be affected by the presence of SBM and BGS but also Grapes especially when BGS is acting (Figure B.7, Table B.4). All in all, the effect of linkage and the contribution of weak selection at the gene level remain unexplored. Thus, new approaches are needed to pinpoint genes under weak positive selection.

Software and availability

Human and *D. melanogaster* processed data and the new impMKT software implementation are available at imkt.uab.cat (Murga-Moreno et al., 2019b). The supporting figures as well as notebooks and code used to perform the analyses can be found at https://github.com/jmurga/mkt_comparison.

Chapter 5

An efficient and robust ABC approach to infer the rate and strength of adaptation

Abstract

More than a decade after genomes became available in several model species, the question of how much genomic evolution is driven by natural selection or neutral forces still remains to be solved in population genetics. In particular, quantifying positive selection is a main challenge because of the multiple confounder variables. It has become clear that approaches to quantify positive selection need to account for the diverse shapes that positive selection can take, while also being robust to a diversity of demographic events and other neutral and non-neutral processes. The growing availability of population genomics data in non-model species where characterizing past adaptation is of evolutionary interest, but with poorly characterized demographic history (or mutational processes, or recombination patterns, etc.), makes the need for robust approaches even more pressing. Here, we introduce an efficient Approximate Bayesian Computation version of the McDonald-Kreitman test, called ABC-MK, to quantify long-term protein adaptation in specific lineages of interest. Compared to the previous implementation, the new ABC-MK runs in a few hours for the first run, and seconds for subsequent runs on an entire proteome, instead of days required by the previous method. This new version of ABC-MK is robust to a wide range of past demographic perturbations and to a broad range of positive selection configurations and strength that make it particularly useful in the context of ecological genomics analyses of non-model species. Using ABC-MK on the human proteome, we find that RNA viruses have driven more long-term strong adaptation than DNA-viruses.

Introduction

Genomes contain a record of the evolutionary processes that shape diversity within and across species, and software tools that use genomic sequences to infer aspects of the evolutionary past are now an integral part of population genetics research. Of particular interest to evolutionary biologists are methods that can disentangle various processes that may contribute to diversification between species, such as adaptation and genetic drift. Such methods have the potential to resolve fundamental questions about the evolutionary (e.g. Corbett-Detig et al. (2015); Galtier (2016); Galtier and Rousselle (2020)) and biological (e.g. Enard et al. (2016); James et al. (2016) drivers of diversification at the genomic level. Though numerous methods have been proposed to this end, it remains challenging to generate accurate and unbiased methods. Studies addressing the potential biases of the available approaches unaccounted-for evolutionary processes and assessing evidence for genome adaptation is still an intense area of research in molecular population genetics (McDonald and Kreitman, 1991; Gillespie, 1994; Smith and Eyre-Walker, 2002; Hahn, 2008; Fay, 2011; Tataru et al., 2017; Kern and Hahn, 2018; Jensen et al., 2019; Johri et al., 2020).

The development of methods that are both computationally efficient and reasonably robust to model misspecification remains a major challenge. Most computational approaches that infer the rate of long-term adaptation at the DNA level derive from the McDonald and Kreitman (MKT) framework (McDonald and Kreitman, 1991) or the related Poisson Random Field (PRF) framework (Sawyer and Hartl, 1992). Both methods use divergence and polymorphism data to estimate the proportion of non-synonymous substitutions fixed by positive selection in coding sequences, comparing alleles that are likely to have fitness effects (putatively selected) to those less likely to be under selection (putatively neutral). A significant excess of fixed differences among the putatively functional set relative to the putatively neutral set is taken as a signal of positive selection. The rate of adaptation is often summarized by the quantity α, which is defined as the proportion of non-synonymous (or putatively functional) fixed differences that were under positive selection along a particular evolutionary branch. When α is close to 1, then positive selection is the predominant determinant of molecular divergence. If α is close to 0, then drift dominates sequence divergence. Smith and Eyre-Walker (2002) applied a simple theoretical model of directional selection relating polymorphism and divergence with adaptation rate, and showed that the rate of adaptation α could be inferred with the quantity

$$\alpha = 1 - \left(\frac{D_S}{D_N} \frac{P_N}{P_S} \right) \tag{5.1}$$

where D_S is the number of synonymous fixed differences in a sequencing sample, D_N represents nonsynonymous fixed differences, P_N is the number of nonsynonymous polymorphic sites, and P_S represents polymorphic synonymous sites. This convenient formula has been widely applied to estimate molecular adaptation, in part because of its simplicity. Indeed, the quantities on the right hand side of equation (5.1) are commonly inferred by comparing a population sample of sequenced individuals (sequencing sample) to a closely related outgroup species. Although widely used, it should be noted that MKT and PRF-based approaches have multiple drawbacks that could bias the estimation of α. For instance, equation (5.1) relies on a null model derived from nearly-neutral theory (Kimura, 1968; Ohta, 1974; Kimura, 1977), and assumes that selected polymorphism, either negative or positive, is rarely observed. Subsequent modeling and empirical studies argued that weakly selected alleles can attain high frequencies and may cause substantial biases in inferences that use equation (5.1) (Balloux and Lehmann, 2012; Lanfear et al., 2014; Booker and Keightley, 2018; Galtier and Rousselle, 2020; Rousselle et al., 2020). Though weakly deleterious alleles are less likely to reach fixation, if they contribute to the class P_N, then the neutral mutation rate in the putatively functional class may be overestimated, which makes the estimation of α downwardly biased (Charlesworth and Eyre-Walker, 2008). Note that the fixation of weakly deleterious alleles could also cause overestimation of α (see Eyre-Walker and Keightley (2009) and Section 5). Altogether, weakly selected polymorphism could drive substantial biases in the inference of adaptation rates and strength.

The presence of slightly deleterious mutations has been addressed by MKT- and PRF-based methods by explicitly modeling the Distribution of Fitness Effects (DFE) for negatively selected variants (Boyko et al., 2008; Eyre-Walker and Keightley, 2009; Messer and Petrov, 2013a; Racimo and Schraiber, 2014; Galtier, 2016). Beneficial alleles also can be found at intermediate frequency or high frequency (Tataru et al., 2017; Uricchio et al., 2019), especially when the rate of strongly beneficial mutations is high (as might be expected in a large population) or weakly beneficial alleles contribute substantially to polymorphism (as might be expected under some polygenic selection models). Despite the development of several methods that account for weakly selected polymorphism, some empirical observations remain challenging to explain under existing models, such as the apparent low rate of adaptation in primates, the constrained range of genetic diversity across species, and differences in the rate of adaptation among taxa (Galtier, 2016; Castellano et al., 2018, 2019a). Generating a deeper biological and evolutionary understanding of the drivers of differentiation across species may require new methods and models that can efficiently estimate the DFE while simultaneously accounting for many (potentially confounding) evolutionary processes.

Demographic processes (such as population contractions, expansions, and migrations) are major potential sources of bias in the inference of selection (Jensen et al., 2019; Johri et al., 2020), just as selection is a major potential confounder in the inference of demography (Schrider et al., 2016; Torres et al., 2018). The developers of robust inference methods have typically sought to account for both selection and demographic processes simultaneously. The cost of incorporating both demography and selection is accrued in terms of model complexity and loss of efficiency, as it is much more challenging to compute likelihoods or summary statistics under joint demography/selection models. There is some hope however that methods based on the asymptotic-MKT (Messer and Petrov, 2013a) may have some inherent robustness, since these approaches rely on summary statistics that involve ratios of functional and (putatively) non-functional alleles. Hence, some of the effects of demography should be absorbed into the ratio, as both categories of alleles will be affected. Indeed, Uricchio et al. (2019) reported that moderate levels of demographic model-misspecification resulted in tolerable inaccuracies for parameter inference.

Here we develop an extension of the Approximate Bayesian Computation ABC-MK method presented in Uricchio et al. (2019) that greatly improves the efficiency of the population genetics inferences. In Uricchio et al. (2019), analytical calculations were used to explore the effect of background selection and selective interference on weakly beneficial alleles, but the estimation procedure employed was based on computationally intensive forward simulations and took days even on a High Performance Computing cluster. We developed a simpler and much more computationally efficient ABC-based inference procedure that accounts for the DFE of deleterious and beneficial alleles and partial recombination between selected genomic elements. We describe the inference procedure, assess its performance and robustness to non-equilibrium demographic scenarios and different intensities of adaptation, and apply it to human genomic data. We show that the method is reasonably robust to non-equilibrium events or different fitness values of adaptation, and provide additional evidence for a substantial effect of RNA-viruses on human adaptation rates, and discuss caveats and potential extensions of our work.

The robustness of ABC-MK to a variety of evolutionary scenarios makes it particularly useful in the context of genomic datasets with poorly characterized past evolution, both at the level of demography or at the level of the nature of adaptation.

Materials and Methods

Our first goal is to calculate the expected rate of fixation and the expected site frequency spectrum (SFS) of neutral and selected polymorphism sites under a model of directional selection with partial recombination. To do so, we follow the results of Uricchio et al. (2019), which in turn extended the results of several earlier studies (e.g., Eyre-Walker and Keightley (2009); Messer and Petrov (2013b)).

Our ultimate goal is to estimate α, the proportion of nonsynonymous substitutions fixed by positive selection (Smith and Eyre-Walker, 2002), as well as the DFE for de novo nonsynonymous mutations. We suppose that selection is directional, with both positively selected and negatively selected mutations. We first consider the case where each selected locus evolves independently, and in subsequent sections we consider cases with background selection and selective interference. As in Uricchio et al. (2019), the DFE over beneficial alleles consists of two point masses, one representing strongly beneficial alleles and the other representing weakly beneficial alleles.

Finally, we extend the calculations by developing a random sampling scheme that accounts for the Poisson variance in mutation and fixation rates, and allows us to develop a simple inference pipeline avoiding forward simulations. We briefly review the core aspects of the theoretical framework, while a more detailed summary can be found in the Supplemental Materials of Uricchio et al. (2019).

Theoretical approximation to α

The rate of adaptation can be decomposed into weakly and strongly beneficial components, $\alpha = \alpha_W + \alpha_S$. The substitution rate for nonsynonymous alleles is denoted as d_N, with d_{N_+} , d_{N_-}, and d_{N_0} representing the rates for positively selected, negatively selected, and neutral alleles respectively (note that $d_N = d_{N_+} + d_{N_-} + d_{N_0}$). In the same way, we denote as d_S the substitution rate of synonymous mutations, which are assumed to be neutral. We can write α as

$$\mathbf{E}[\alpha] = \frac{d_{N_+}}{d_N} = \frac{d_N - (d_{N_-} + d_{N_0})}{d_N} = 1 - \frac{(d_{N_-} + d_{N}0)d_S}{d_S d_N} \tag{5.2}$$

Note that we define α as the actual proportion of positively selected substitutions along the branch, and hence equation (5.2) is an expression for the expectation of α. As noted by Messer and Petrov (2013a), d_S can be estimated from sequence alignments with the ratio D_N/D_S under the assumption that the observed number of substitutions along a branch should be proportional to the rate. The ratio $(d_{N_-} + d_{N_0})/d_S$ is more complex to estimate, because it relies on partitioning substitutions by their fitness effects. Under the assumption that polymorphic alleles are rarely selected (because deleterious sites are removed from the population quickly and beneficial sites go to fixation rapidly), previous work (Smith and Eyre-Walker, 2002) showed that this ratio can be approximated by substituting $\frac{P_N}{P_S}$ into equation (5.2), and a point estimate of α as

$$\alpha \approx 1 - \left(\frac{P_N}{P_S}\frac{D_S}{D_N}\right) \tag{5.3}$$

However, if selected polymorphisms segregate in the sample, then P_N in equation

(5.3) will be inflated relative to the true rate of mutation for neutral nonsynonymous alleles, which results in underestimation of α. A potential solution is to exclude alleles with derived allele frequencies lower than some threshold from the quantities P_N and P_S, since most (negatively) selected alleles should be constrained to lower frequency (Fay et al., 2001). While this solution works well for some DFEs (for example, when all deleterious alleles are strongly selected), weakly deleterious alleles can reach appreciable frequencies and bias inference regardless of the selected frequency threshold. Messer and Petrov (2013a) extended this idea by developing a very simple estimator of α that uses all frequencies simultaneously by rewriting the estimator of Smith and Eyre-Walker (2002) as

$$\alpha \approx 1 - \left(\frac{P_{N(x)}}{P_{S(x)}} \frac{D_S}{D_N} \right) \tag{5.4}$$

where $P_{N(x)}$ and $P_{S(x)}$ are the number of non/synonymous alleles at frequency x in a sequencing sample. They fit a simple exponential curve to the $\alpha_{(x)}$ data points, and the asymptote of this curve is taken as an estimate of α. This method improves the quality of α estimates by using all of the frequency data simultaneously and providing confidence intervals for α, but does not provide an estimate of the DFE and assumes that beneficial alleles do not contribute to P_N (see the Supplemental Material of Uricchio et al. (2019) for more details).

Generic model to the expected fixation rates and frequency spectra

A complementary approach to that of Messer and Petrov (2013a) is to directly model the effects of the DFE for beneficial and deleterious alleles on the shape of the $\alpha_{(x)}$ curve, and to infer the best fitting model parameters. Nonetheless, note that while $\mathbf{E}[\alpha_{(x)}] = 1 - \mathbf{E}[\frac{D_S P_{N(x)}}{D_N P_{S(x)}}]$ is not straightforward to calculate because it depends on the ratio of several random variables, the expectation of each component in equation 5.4 ($P_{S(x)}$, $P_{N(x)}$, D_S, D_N) is easily calculated in a directional selection model from first principles using diffusion theory (Evans et al., 2007). Therefore, we make a first-order approximation

$$\mathbf{E}[\alpha_{(x)}] \approx 1 - \frac{\mathbf{E}[D_S]\mathbf{E}[P_{N(x)}]}{\mathbf{E}[D_N]\mathbf{E}[P_{S(x)}]} \tag{5.5}$$

In this manuscript we assume that positively selected mutations have fitness effects drawn from a point mass distribution (although such assumption is relaxed in the Approximate Bayesian Computation), while negatively selected mutations drawn from a gamma distribution. In general, our approach can be applied to any distribution for which we can analytically solve the fixation rates and expected frequency spectra. Brief descriptions of these calculations follow this section, while detailed descriptions of these calculations please refer to the Supplemental Materials of Uricchio et al. 2019 or the online documentation for our software at web address https://jmurga.github.io/Analytical.jl/dev/.

Expected number of fixations. Considering the distribution of selection coefficients over new mutations μs (selection coefficient underlying the mutation rate) and the fixation probability π_s, we calculate the expected number of substitutions along a branch of time T in a locus of length L as

$$\mathbf{E}[D] = LTd = LT \int\limits_s 2N\mu_s\pi_s ds \tag{5.6}$$

For positively selected mutations with large selection coefficients ($s > 0.01$), we follow the procedure described in Uricchio and Hernandez (2014) for determining the probability of fixation, which treats the initial trajectory of the mutation as a Galton-Watson process.

Expected frequency spectrum. The expected number of alleles at frequency x is estimated from the standard diffusion theory for the site frequency spectrum in an equilibrium population (e.g., see equation 31 of Evans et al. (2007)).

$$\Psi(x) = \int\limits_s \theta_s \frac{1}{x(1-x)} \frac{e^{4Ns}(1 - e^{-4Ns(1-x)}}{e^{4Ns} - 1} ds \tag{5.7}$$

where $\Psi(x)$ is the number of alleles at frequency x in a population of size N and $\theta_s = 4N\mu s$ is the population-scaled mutation rate for mutations with selection coefficient s. To obtain the downsampled frequency spectrum in a finite sample of $2n$ chromosomes, we convoluted equation (5.7) with the binomial distribution.

Background selection and adaptive divergence. Background selection (BGS) (Charlesworth et al., 1993; Hudson and Kaplan, 1995; Nordborg et al., 1996) and selective interference (e.g., Hill-Robertson interference, (Hill and Robertson, 1966)) could affect the rate of fixation of weakly deleterious or beneficial alleles. Up to this point, we have considered only selected loci that evolve independently of all other selected loci. In this section we will relax this assumption by exploring the effects of selective interference on fixation rates and the frequency spectrum. Note that this will not be a full treatment of these topics, which are active areas of research. Rather we will follow approximations that will apply in some circumstances (in particular, when BGS is the predominant driver of selective interference), but may fail when strongly beneficial alleles interfere.

To explore $\alpha(x)$ accounting for recombination and BGS impact, we focused on a model in which the coding locus is flanked on each side by loci of length L, which contain deleterious alleles (see Figure 5.1). We modeled deleterious alleles with a population-scaled selection coefficient $-2Nt$ undergoing persistent deleterious mutation at rate $4N\mu_{-}$. The whole flanking loci recombined at a rate r per-base, per-generation. Previous work has shown that diversity at the coding locus (π) is decreased relative to its neutral expectation (π_0), and closed form expressions for the expected reduction in diversity are available (Hudson and Kaplan, 1995; Nordborg et al., 1996). The effects of BGS on fixations and frequency spectra have been subject of much theoretical work (Charlesworth et al., 1993; Charlesworth, 1994; Hudson and Kaplan, 1995; Barton, 1995; Nordborg et al., 1996). While patterns of sequence variation induced by BGS can be quite complex (Nicolaisen and Desai, 2013; Good et al., 2014; Torres et al., 2018, 2020), to a first approximation the effect of BGS can be thought of as a reduction in the effective population size N_e, with $N_e = N\frac{\pi}{\pi_0}$ (McVicker et al., 2009). To account for the role of BGS on the fixation rate of deleterious alleles, we replace N in the prior equations with N_e after accounting for BGS. We also replace N with N_e in formulae for the frequency spectra.

For beneficial alleles, the effects of selective interference are slightly more complex. Strongly beneficial alleles are essentially unaffected by BGS, in that their fixation probabilities almost do not depend on the reduction in neutral diversity. Weakly beneficial alleles can have their fixation probabilities substantially reduced by BGS. We followed Barton (1995) to derive formulae for the reduction in fixation rate of weakly and strongly beneficial alleles after accounting for BGS, as described in the Supplemental Materials of Uricchio et al. (2019) (see *Background selection and adaptive divergence* section). The reduction in fixation probability for a weakly beneficial allele with selection coefficient s under interference with deleterious alleles with selection coefficient t is given by

$$\phi(t,s) = e^{[\frac{-2\mu}{t(1+\frac{rL}{t}+\frac{2s}{t})}]} \tag{5.8}$$

where l is the distance in base pairs from the region of interest, $1 \leq l \leq L$ (see equation 17d of Barton (1995)). Multiplying across all deleterious linked sites and factoring in flanking sequences to both the left and right of the focal site (which requires us to square the product below), we find that the reduction in the probability of fixation relative to the case with no linkage (Φ) is

$$\Phi = \prod_{1}^{L} \phi(t,s) = e^{\frac{-2t\mu(\psi[1,\frac{r+2s+L}{r}]-\psi[1,\frac{r(L+1)+2s+t}{r}])}{r^2}} \tag{5.9}$$

where ψ is the polygamma function. Evidence for the adequacy of these approximations is provided in Uricchio et al. (2019) by comparing the results to forward simulations. However, we note that these expressions are not expected to hold under very high rates of mutation for beneficial alleles, and will be less accurate for strongly beneficial alleles than weakly beneficial alleles. Given these expressions, we can replace the fixation rates for beneficial alleles in our prior formulae with the fixation rates after accounting for selective interference.

Poisson-sampling process. The previous sections described the expectation of fixation rates and frequency spectra under a model of directional selection and selective interference. We now develop a simple random sampling scheme around these expectations that accounts for sampling and process variance, linking analytical estimation and ABC procedure for finally avoiding forward simulations. We note that the model we explore is quite similar to the BGS model in DeGiorgio et al. (2016), though while we are interested in the long-term accumulation of fixations, DeGiorgio et al. (2016) is primarily interested in the non-equilibrium signature of a recent or ongoing selective sweep. Following the Poisson Random Field model (Sawyer and Hartl, 1992), we supposed that the number of fixed differences and polymorphic sites were Poisson random distributed variables with mean values given by the expectations in the previous sections.

To avoid performing branch length estimations in our computation, we assumed

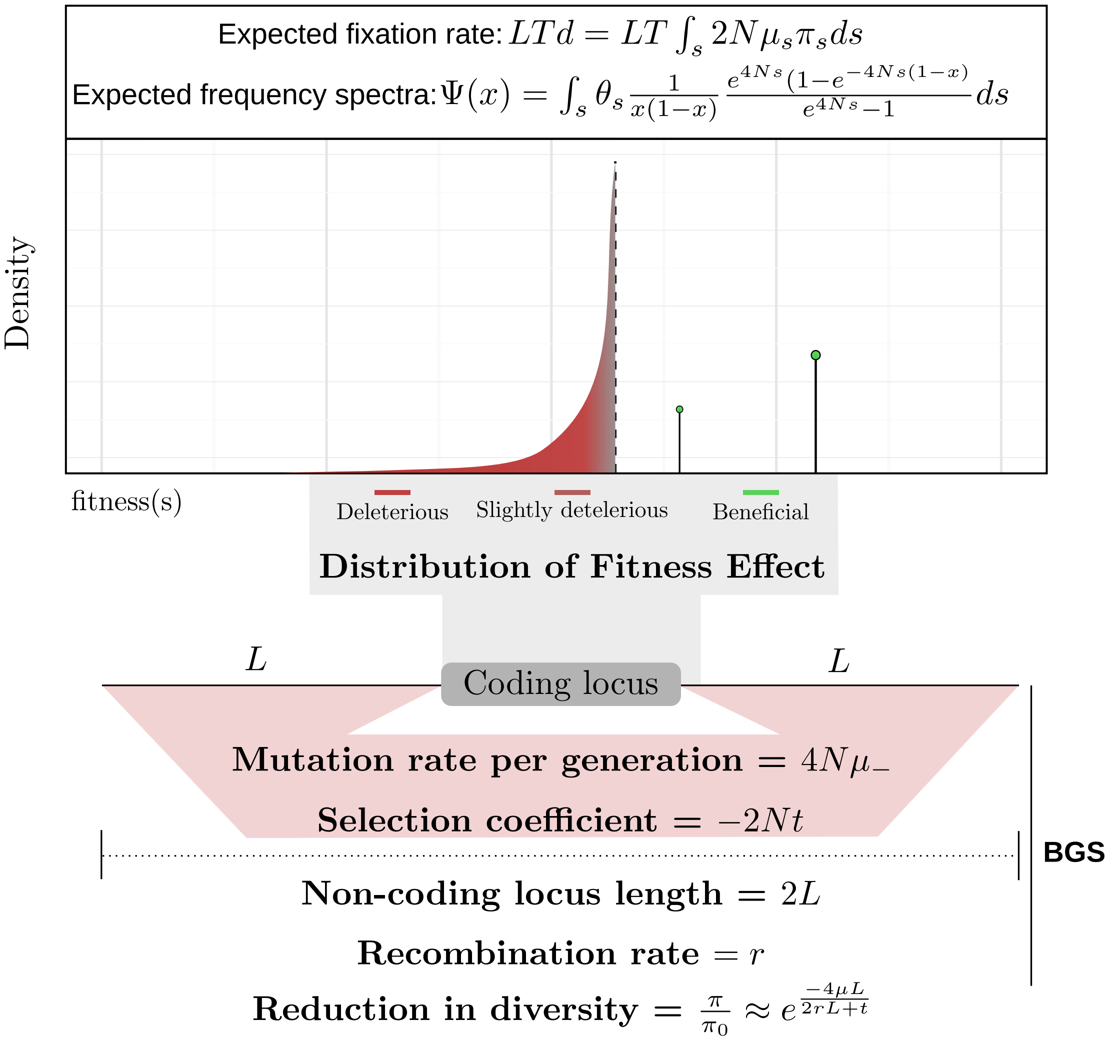

Expected fixation rate: $LTd = LT\int_s 2N\mu_s\pi_s ds$
Expected frequency spectra: $\Psi(x) = \int_s \theta_s \frac{1}{x(1-x)} \frac{e^{4Ns}(1-e^{-4Ns(1-x)}}{e^{4Ns}-1} ds$
Density
fitness(s)
Deleterious
Slightly detelerious
Beneficial
Distribution of Fitness Effect
L
Coding locus
L
Mutation rate per generation $= 4N\mu_-$
Selection coefficient $= -2Nt$
BGS
Non-coding locus length $= 2L$
Recombination rate $= r$
Reduction in diversity $= \frac{\pi}{\pi_0} \approx e^{\frac{-4\mu L}{2rL+t}}$

Figure 5.1: Graphical representation of the model used to estimate the fixation ratio and frequency spectra under BGS. The coding locus follows a combination of Gamma and discrete distributions while the non-coding locus undergoing persistent deleterious mutation at rate $4N\mu_-$ accounting for population-scaled selection coefficient $-2Nt$. The expected reduction in diversity is used to estimate the associated reduced fixation probabilites given the BGS model described.

that the empirically observed number of fixations should be proportional to the length of the evolutionary branch of interest, T, the locus length L and mutation ration μ. We take the observed number of fixations as a proxy for the expected number, and then sample weakly deleterious, neutral, and beneficial substitutions proportional to their relative rates for a fixed set of model parameters. The expected number of substitutions for positively selected substitutions is then

$$\lambda_{D_{N+}} = D\frac{\mathbf{E}[d_{N+}]}{\mathbf{E}[d_{N+}] + \mathbf{E}[d_{N-}] + \mathbf{E}[d_S]} \tag{5.10}$$

where D is the observed number of substitutions in a dataset of interest.

It should be noted that both sampling variance and process variance affect the number of variable alleles at any particular allele frequency in a sequencing sample. The process variance arises the random mutation-fixation process along the branch, while the sampling variance arises from the random subset of chromosomes that are included in the sequencing data. We sampled a Poisson distributed number of polymorphic alleles at frequency x relative to their rate given the expected frequency spectra. The expected frequency spectra were downsampled using a binomial (with probability of success given by the frequency $\begin{pmatrix} x \\ 2n \end{pmatrix}$ in a sample of $2n$ chromosomes) to account for the sampling variance. In a manner exactly analogous to fixed variants as described above, to account for the process variance

$$\lambda[P_N] = \sum_{x=0}^{1} P_{(x)} \frac{\mathbf{E}[p_{N+(x)}] + \mathbf{E}[p_{N-(x)}]}{\mathbf{E}[p_{N+(x)}] + \mathbf{E}[p_{N-(x)}] + \mathbf{E}[p_{S(x)}]} \tag{5.11}$$

To account for BGS, we solved equation (5.9) using any expected B values from McVicker et al. (2009) at each polymorphic or fixed site . We note that inferred background selection strength is not available for most species and our software can run without this information, but when it is available, the inference can be limited to such values. We discount the fixation probability of deleterious alleles by the predicted value of B at each site, and we use the predicted reduction in fixation probability given by equation (5.9) for weakly beneficial alleles. In practice, we bin sites into B-value ranges, such that all sites within (for example) a 2.5% B-value range of 0.675 to 0.7 experience the same N_e and the same strength of selective interference (for example $N_e = 0.675N$, which is the midpoint of this B-value window).

Computational workflow. Our ultimate goal is to infer α, α_W, and α_S given a set of observed α values from a sequencing dataset. Since α in our framework is a model output and not a parameter per se (i.e., α depends on the random number of fixations along an evolutionary branch, which in turn depend on the parameters of the evolutionary model), we cannot immediately obtain the corresponding fixation rates and frequency spectra values for a given set of expected α values without first solving for the mutation rates and fixation probabilities considering the input model (see Table 5.1). Given the α and α_W (which together uniquely determine α_S), a DFE over negatively selected alleles, a known B-value, a selection coefficient for beneficial alleles, a selection coefficient for flanking deleterious alleles, a recombination (ρ) and mutation rate on the coding locus (θ), we numerically solve for the probability of fixation of beneficial alleles and the mutation rate on the flaking locus that correspond to the desired α values given the BGS strength. This allows us to calculate rapidly the expected frequency spectra and fixation rates that will correspond to the desired α values and generate a sample of $\alpha_{(x)}$ values following the Poisson-sampling process under the corresponding evolutionary model.

We used a generic Approximate Bayesian Computation (ABC) algorithm to infer the rate and strength of adaptation. ABC procedure first samples the parameter values from prior distributions; second simulates random model calculating informative summary statistics; and third compares the simulated summary statistics to observed empirical data. The summary statistics producing best match to the observed empirical data form an approximate parameter posterior distribution. Our approach used empirical data to both perform computational workflow and Poisson-sampling scheme described above to sample $\alpha_{(x)}$ generating summary statistics corresponding to different evolutionary scenarios and to finally compare such summary statistics to empirical $\alpha_{(x)}$ estimations. Since we do not know a priori the values of any model parameter, to estimate summary statistics, we sample 10^5 sets of parameters randomly from a prior uniform distribution, which allows for flexibility in the DFE of deleterious and beneficial alleles. We supplied the summary statistics and empirical $\alpha_{(x)}$ into ABCreg (Thornton, 2009) to estimate the empirical values of α_W, α_S and α while accounting for BGS in bins of 2.5% from $\frac{\pi}{\pi_0} = 0.1$ to $\frac{\pi}{\pi_0} = 1$.

For each analysis, we used 100 bootstrapped datasets to generate summary statistics and inputs to the ABC inference. As summary statistics, we used the value of $\alpha_{(x)}$ for $x \in (2, 5, 20, 50, 200, 661, 925)$ and $x \in (2, 5, 20, 50, 200, 500, 700)$ regarding the input data. These values are similar to the ones used in Uricchio et al. (2019), though we excluded singletons because very low frequency alleles are particularly sensitive to sequencing errors and distortions due to demographic processes or other model misspecifications. We inferred posterior distributions for each bootstrapped dataset, each one using the same 10^5 summary statistics as a prior. We set the tolerance threshold in ABCreg to 0.01 such that 10^3 values were accepted from posterior distributions.

Table 5.1: Model parameters

Model parameters	γ	Population-scaled selected coefficient of deleterious alleles
	s_w	Population-scaled selected coefficient of weakly beneficial alleles
	s_s	Population-scaled selection coefficient of strong beneficial alleles
	α_w	Proportion of weakly adaptive substitutions
	α	Proportion of adaptive substitutions
	θ_{coding}	Population-scaled mutation rate at coding locus
	θ	Scale parameter
	β	Shape parameter
	B	B value from McVicker et al. (2009)
	N	Effective population size
	n	Sample size
	L	Non-coding locus length
	ρ	Population-scaled recombination rate
Derived parameters	l_W	Fixation probability of weakly beneficial alleles
	l_S	Fixation probability of strong beneficial alleles
	$\theta_{non-coding}$	Population-scaled mutation rate at non-coding locus

Forward-in-time simulations

We used `SLiM 3` (Haller and Messer, 2019) to generate simulated sequence variation data under our model and test the predictions of our approach. We performed three different sets of simulations accounting for demography or not, in which we modeled the same rates of adaptation and BGS. For each set of simulations, we considered branch length that mimics the human split from chimpanzee (estimated to be $\approx$ 5.5M years ago). In simulations with *realistic* non-equilibrium human demography, we added demographic events following Tennessen et al. (2012) to model the variation in the 661 African individuals whose genomes are included in the 1000 Genome Project (Auton et al., 2015). Each simulation represents a coding locus of $2 \cdot 10^3$ bp flanked on each side by a 10^5 bp non-coding locus. A total of $5 \cdot 10^4$ genes were simulated accounting for a total of $10^8 bp$ of coding sequence. We performed the simulations following previously estimated values of negative selection of human proteins (Boyko et al., 2008), where the distribution of deleterious alleles follows a gamma-distribution with scale and shape parameters of 0.184 and 0.000402 respectively, which implies a mean fitness of $2Ns = -457$ for negatively selected nonsynonymous alleles. Strongly and weakly beneficial alleles followed a point-mass distribution given the population-scaled selections coefficients of $2Ns = 10$ and $2Ns = 500$ respectively. Our simulations supposed that 25% of mutations in each coding locus are synonymous while and 75% are nonsynonymous. We used a mutation within each coding locus of $\theta = 4N\mu = 0.001$ and a mean human recombination rate in the flanking sequence of $\rho = 4Nr = 0.001$. See Table 5.2 for parameter values of forward simulations.

Table 5.2: Prior values to SLiM simulations

Scenarios	N_{anc}	μ_{coding}	r	$\mu_{non-coding}$	s_W	s_S	l_W	l_S	α_W	α	B
Non-demography	500	1.32e-06	5e-07	5e-07	10	500	0.0038	0.00039	0.1	0.4	0.2
	500	1.32e-06	5e-07	5e-07	10	500	0.0077	0.00026	0.2	0.4	0.2
	500	1.32e-06	5e-07	5e-07	10	500	0.0116	0.00013	0.3	0.4	0.2
	500	7.52e-07	5e-07	5e-07	10	500	0.0038	0.00039	0.1	0.4	0.4
	500	7.52e-07	5e-07	5e-07	10	500	0.0077	0.00026	0.2	0.4	0.4
	500	7.52e-07	5e-07	5e-07	10	500	0.0116	0.00013	0.3	0.4	0.4
	500	1.83e-07	5e-07	5e-07	10	500	0.0038	0.00039	0.1	0.4	0.8
	500	1.83e-07	5e-07	5e-07	10	500	0.0077	0.00026	0.2	0.4	0.8
	500	1.83e-07	5e-07	5e-07	10	500	0.0116	0.00013	0.3	0.4	0.8
	500	8.22e-10	5e-07	5e-07	10	500	0.0038	0.00039	0.1	0.4	0.999
	500	8.22e-10	5e-07	5e-07	10	500	0.0077	0.00026	0.2	0.4	0.999
	500	8.22e-10	5e-07	5e-07	10	500	0.0116	0.00013	0.3	0.4	0.999
Tennesen model	7310	1.32e-06	3.42e-08	3.42e-08	10	500	0.0038	0.00039	0.1	0.4	0.2
	7310	1.32e-06	3.42e-08	3.42e-08	10	500	0.0077	0.00026	0.2	0.4	0.2
	7310	1.32e-06	3.42e-08	3.42e-08	10	500	0.0116	0.00013	0.3	0.4	0.2
	7310	7.52e-07	3.42e-08	3.42e-08	10	500	0.0038	0.00039	0.1	0.4	0.4
	7310	7.52e-07	3.42e-08	3.42e-08	10	500	0.0077	0.00026	0.2	0.4	0.4
	7310	7.52e-07	3.42e-08	3.42e-08	10	500	0.0116	0.00013	0.3	0.4	0.4
	7310	1.83e-07	3.42e-08	3.42e-08	10	500	0.0038	0.00039	0.1	0.4	0.8
	7310	1.83e-07	3.42e-08	3.42e-08	10	500	0.0077	0.00026	0.2	0.4	0.8
	7310	1.83e-07	3.42e-08	3.42e-08	10	500	0.0116	0.00013	0.3	0.4	0.8
	7310	8.22e-10	3.42e-08	3.42e-08	10	500	0.0038	0.00039	0.1	0.4	0.999
	7310	8.22e-10	3.42e-08	3.42e-08	10	500	0.0077	0.00026	0.2	0.4	0.999
	7310	8.22e-10	3.42e-08	3.42e-08	10	500	0.0116	0.00013	0.3	0.4	0.999

Software and data availability

We developed user-friendly software in order to execute our model. ABC-MK is freely available at https://github.com/jmurga/Analytical.jl. It is based on Julia language and support multi-threading, interactive environments as well as Command Line Interface usage. Tutorials and examples are available at https://jmurga.github.io/Analytical.jl/dev/.

Results and Discussion

Evolutionary processes affecting adaptation inference

We used our forward simulations to retrieve polymorphism and divergence data with a priori known adaptation parameters. To input the data in our model, we pooled the SFS and number of fixations of $5 \cdot 10^4$ genes. Note that this is about 2.5 times as many genes as appear in the human genome -we use this larger number of genes such that the trends in the simulated data will be clear and not dominated by noise, while noting that noise may play a larger role in real datasets for some species with limited proteome coverage. As demonstrated previously, the frequency spectrum (Tataru et al., 2017) and $\alpha_{(x)}$ (Uricchio and Hernandez, 2014) are substantially affected by the presence of weakly beneficial alleles (Figure 1). The asymptote of the $\alpha_{(x)}$ curve bends below the true value of $\alpha_{(x)}$, which is caused by an excess of high frequency nonsynonymous variants under weak positive selection. This results in downward bias of α estimates when aMKT (Messer and Petrov, 2013a) is applied to the data (see Table 5.2). In real sequencing datasets we cannot a priori separate beneficial and deleterious alleles, but in our simulated datasets we can remove the weakly beneficial alleles and test whether this will fix the downwards bias in aMK. When removing weakly beneficial alleles we observed an increase in α estimates from aMK which tend towards the true value of α (see Figure 5.2, Table 5.2).

In addition, $\alpha_{(x)}$ can be substantially affected by BGS, especially when weakly beneficial alleles contribute to the frequency spectrum (see Figure 5.2). In cases where α is dominated by strong adaptation, both asymptotic values (accounting for all alleles, or just neutral and deleterious) tend to be similar, because strongly beneficial alleles are not substantially impeded by selective interference with linked deleterious variation. Similar results were reported in Uricchio et al. (2019), and we include them here for completeness. We also tested the effect of recent demographic events with a simulation of the Tennesen demographic model, specifically for the African continental group (Uricchio et al., 2019). We used demographic parameters following Adrion et al. (2020a). To improve performance we simulated the African population in isolation, rather than including the full multi-population model –consequently our simulations do not include the effects of migration on the frequency spectrum. We observed similar patterns to equilibrium simulation regarding the overall shape and asymptotic values of $\alpha_(x)$ (Figure 5.3). Since this model includes a recent and rapid population expansion, we observe distortions to the frequency spectrum at extremely low and high frequencies due to the excess of rare alleles relative to an equilibrium demographic model.

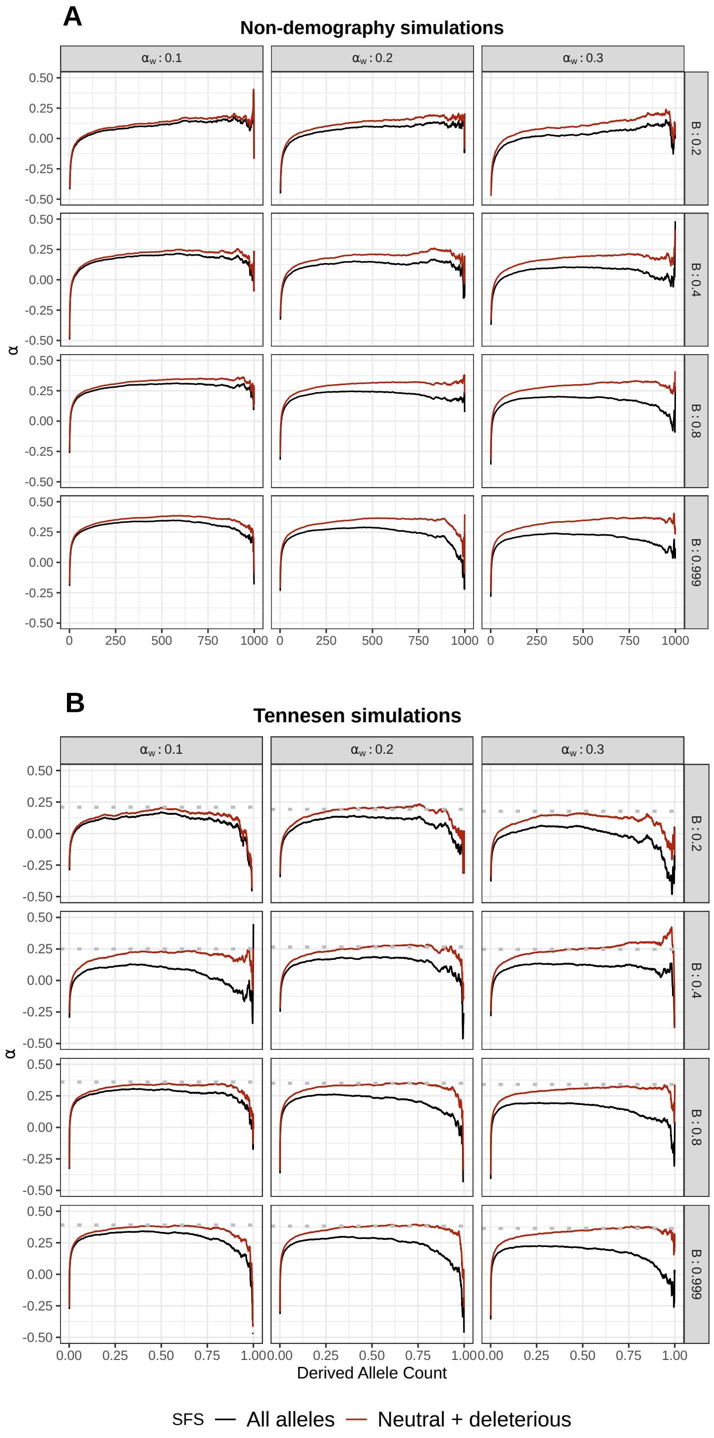


Figure 5.2: BGS worsens the distorting effect of weakly advantageous mutations on the α curve. We simulated the effect of weakly advantageous allele and BGS effect on $\alpha_{(x)}$ using `SLiM 3` (Methods). Each row represents a BGS value. Each column represents a proportion of α_W. We assumed a proportion of adaptive substitutions of $\alpha = 0.4$ in the absence of BGS. A. Simultions at demographic equilibrium. B. Simulations under Tennessen et al. (2012) demographic model.

ABC analysis

Equilibrium demography

To compare true parameter values to inferred values, we calculated the Maximum-A-Posteriori (MAP) estimate for each posterior distribution that we obtained from ABC-MK, following the manual of Thornton (2009). Our method can distinguish both weak and strong adaptive contributions to α while performing reasonably accurate estimations (Table 5.3). Table 5.3 shows inferred values and the associated error for each parameter and simulation. In all cases, the posterior distribution overlaps the distribution of the true values from bootstrapped datasets. Figure 5.3 presents the simulated values and MAP estimates. Table 5.3 shows inference parameters and associated error.

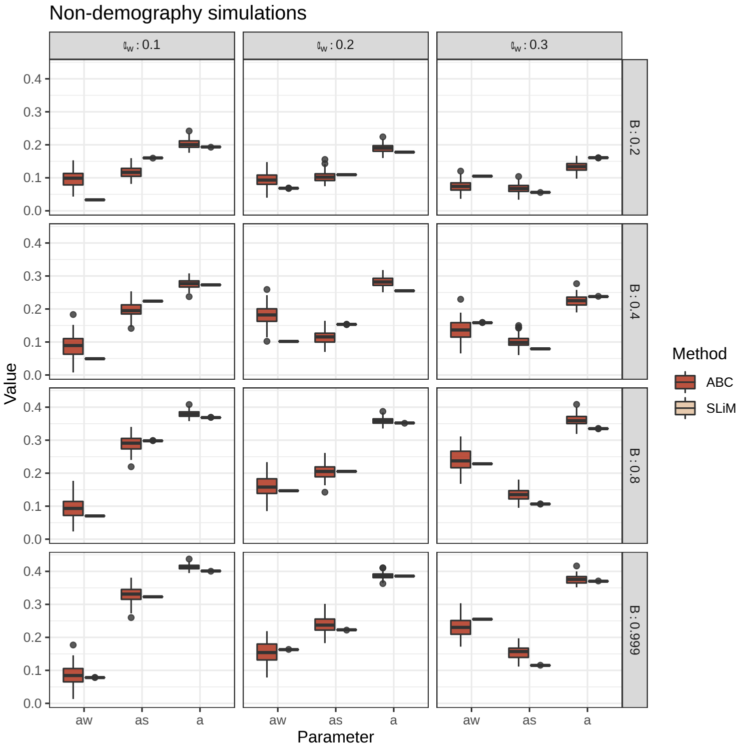


Figure 5.3: ABC-MK inference at equilibrium. MAP distribution of 100 sets of summary statistics per parameter value. ABC inferences were performed using ABCreg.

Non-equilibrium demography

We tested our method using simulations performed under the demographic model of Tennessen et al. (2012). Although these demographic events (which include an ancient expansion in the ancestral population and a period of recent rapid growth) affect the number of segregating sites and the shape of the SFS (which is not modeled in our calculations), our method is reasonably robust to these distortions when we exclude low frequency variants (DAC < 5 i.e. DAF < 0.0038; Figure 5.4). When we include variants at which the SFS is most distorted, such as singletons and very high frequency variants, the inference is biased towards weak selection, although the overall value of α is not strongly affected (Figure 5.4A-C). This likely reflects the qualitative similarity of recent growth events and weakly beneficial alleles in terms of their effects on the SFS, as both will disproportionately increase the number of nonsynonymous variants at low frequency relative to an equilibrium model. In Figure 5.5 we explore a wider range of parameters, using the best performing set of summary statistics (i.e., excluding derived allele counts under 5). Table 5.3 shows inference parameters and associated error.

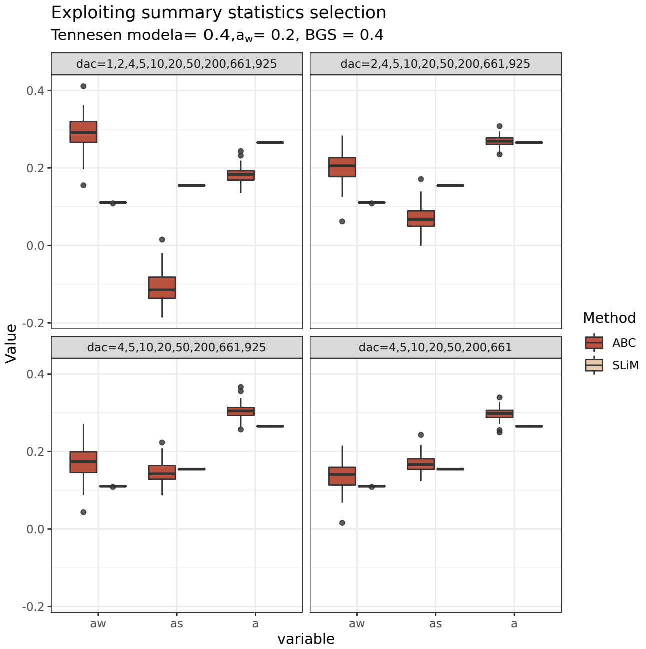


Figure 5.4: Summary statistics selection. ABC inference excluding low-frequency variants in Tennessen et al. (2012) demographic simulations.

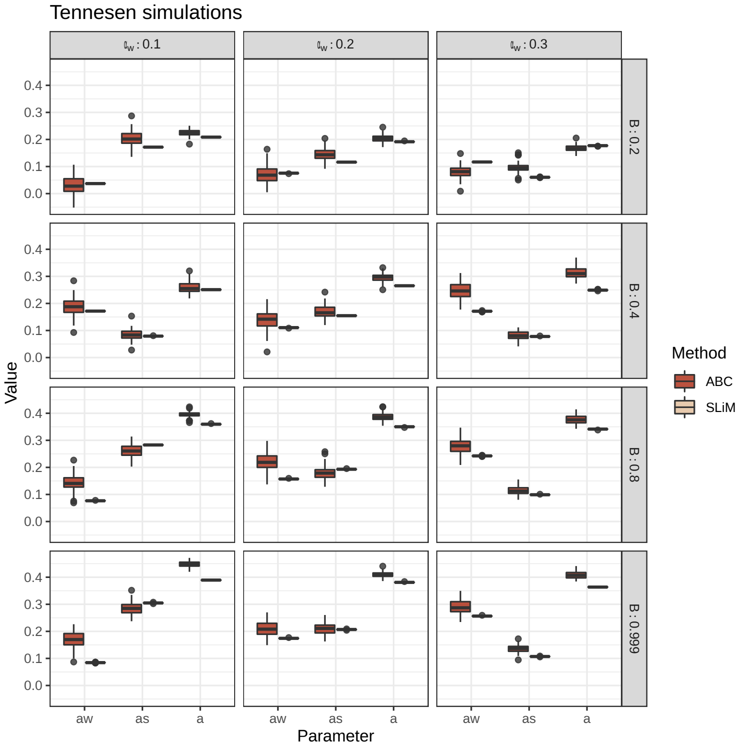


Figure 5.5: ABC inference of Tennessen et al. (2012) demographic model simulations. MAP distribution of 100 datasets. ABC inference was performed using ABCreg

Table 5.3: ABC inference

Scenario	True α_W	B	α_W	α_s	α	$\Delta\alpha_W$	$\Delta\alpha_S$	$\Delta\alpha$
Non-demography	0.1	0.2	0.095 [0.065-0.123]	0.117 [0.092-0.141]	0.203 [0.187-0.222]	0.062	0.043	0.009
	0.1	0.4	0.089 [0.052-0.13]	0.198 [0.173-0.229]	0.276 [0.259-0.292]	0.04	0.026	0.003
	0.1	0.8	0.095 [0.057-0.14]	0.289 [0.26-0.317]	0.38 [0.368-0.391]	0.024	0.009	0.011
	0.1	0.999	0.084 [0.05-0.121]	0.331 [0.303-0.36]	0.413 [0.403-0.424]	0.006	0.008	0.012
	0.2	0.2	0.094 [0.065-0.125]	0.103 [0.087-0.121]	0.189 [0.173-0.207]	0.026	0.006	0.012
	0.2	0.4	0.182 [0.142-0.218]	0.114 [0.093-0.141]	0.283 [0.264-0.304]	0.08	0.039	0.027
	0.2	0.8	0.162 [0.125-0.2]	0.205 [0.175-0.228]	0.358 [0.345-0.371]	0.015	0.001	0.006
	0.2	0.999	0.153 [0.113-0.189]	0.239 [0.209-0.271]	0.387 [0.377-0.399]	0.01	0.016	0.001
	0.3	0.2	0.074 [0.054-0.094]	0.067 [0.046-0.084]	0.133 [0.117-0.15]	0.031	0.011	0.028
	0.3	0.4	0.138 [0.104-0.175]	0.101 [0.084-0.119]	0.224 [0.204-0.243]	0.021	0.021	0.013
	0.3	0.8	0.239 [0.204-0.277]	0.136 [0.115-0.16]	0.36 [0.342-0.377]	0.01	0.029	0.025
	0.3	0.999	0.231 [0.195-0.273]	0.154 [0.129-0.177]	0.375 [0.36-0.39]	0.024	0.039	0.005
Tennesen model	0.1	0.2	0.03 [-0.012-0.069]	0.204 [0.172-0.235]	0.225 [0.211-0.238]	0.007	0.032	0.016
	0.1	0.4	0.189 [0.15-0.23]	0.084 [0.065-0.106]	0.259 [0.235-0.286]	0.017	0.005	0.008
	0.1	0.8	0.142 [0.103-0.181]	0.261 [0.231-0.289]	0.395 [0.385-0.407]	0.066	0.022	0.036
	0.1	0.999	0.169 [0.133-0.207]	0.285 [0.258-0.314]	0.449 [0.436-0.462]	0.084	0.02	0.059
	0.2	0.2	0.071 [0.029-0.108]	0.145 [0.116-0.18]	0.205 [0.191-0.219]	0.004	0.029	0.013
	0.2	0.4	0.139 [0.088-0.184]	0.17 [0.142-0.202]	0.295 [0.277-0.311]	0.028	0.015	0.03
	0.2	0.8	0.222 [0.175-0.265]	0.18 [0.151-0.21]	0.388 [0.373-0.407]	0.064	0.014	0.037
	0.2	0.999	0.21 [0.173-0.247]	0.21 [0.185-0.236]	0.41 [0.396-0.423]	0.035	0.003	0.029
	0.3	0.2	0.08 [0.055-0.104]	0.095 [0.073-0.115]	0.168 [0.153-0.184]	0.036	0.035	0.009
	0.3	0.4	0.246 [0.209-0.285]	0.082 [0.064-0.101]	0.312 [0.286-0.337]	0.075	0.004	0.063
	0.3	0.8	0.279 [0.246-0.32]	0.114 [0.097-0.132]	0.377 [0.358-0.395]	0.037	0.015	0.035
	0.3	0.999	0.29 [0.26-0.324]	0.135 [0.114-0.155]	0.408 [0.39-0.428]	0.034	0.028	0.044

Human viral interacting proteins

As an example of application, we estimate coding adaptation in human genes that interact with viruses. Several studies have argued that viral infections have driven adaptation in the human genome (e.g., Nédélec et al. (2016); Castellano et al. (2019b)). Genomic analysis of patterns of variation within experimentally determined Viral Interacting Proteins (VIPs) has repeatedly uncovered signals of both frequent and strong adaptation (Enard et al., 2016; Uricchio et al., 2019). The selective pressure imposed by virus on hosts appears to be a strong driver of adaptation during human evolution at different time scales or adaptive regimes (Deschamps et al., 2016; Racimo et al., 2017; Enard and Petrov, 2018).

We applied our approach to an augmented set of VIPs and non-VIPs (proteins not known to interact with any virus) that were previously studied in Uricchio et al. (2019) and estimated adaptation rates using our new ABC-MK implementation. We extended this analysis by partitioning specific VIPs into interaction partners with RNA and DNA viruses. We followed previous studies of RNA-VIPs to test at a deeper scale if the RNA-VIPs virus exhibits stronger adaptation rates than DNA-VIPs, to test whether our method provides any additional support to the hypothesis that RNA viruses are important drivers of human adaptation (Enard and Petrov, 2020).

We used genomic data from the same 661 individuals of African descent that were studied in Uricchio et al. (2019), whose genomes were sampled in the Thousand Genomes Project, while increasing the total number analyzed of curated VIPs from 4,066 to 5,310. From this, 1,258 annotations correspond to DNA-VIPs, whereas 3,471 correspond to RNA-VIPs. In this case, we bootstrap each dataset 100 times following the polyDFE manual Tataru et al. (2017) to get MAP estimates from posteriors distributions following the previously described ABC scheme.

We inferred values of α, α_W, and α_S that were very similar to those inferred in Uricchio et al. (2019) regarding VIPs and non-VIPs datasets, despite the increased number of analyzed VIPs (see Figure 5.6 and Table 5.4) and the simplified, streamlined ABC-MK implementation. Interestingly, when distinguishing between DNA and RNA-VIPs, we found higher strong adaptation rates in RNA-VIPs in both α and α_W (see Figure 5.6 and Table 5.4). Uricchio et al. (2019) noted that the higher adaptation rate for VIPs cannot be explained by the BGS effect, because VIPs undergo slightly stronger BGS than non-VIPs. The same occurs for RNA-VIPs vs DNA-VIPs here, as the mean BGS strength at RNA-VIPs is 0.556 compared to 0.616 for DNA-VIPs. These results may reflect biological differences between RNA and DNA viruses, especially in terms of zoonosis frequency, and suggest that RNA-viruses may have played an especially important role in human adaptation.

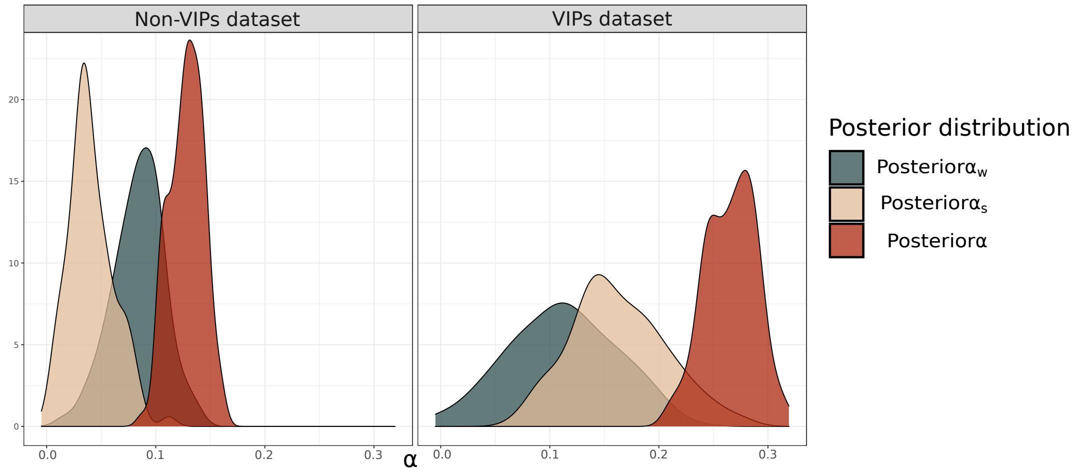


Figure 5.6: Non-VIPs and VIPs posterior distributions. Inference was performed using human lineage data from Uricchio et al. (2019). A total of 1301 non-VIPs and 4729 VIPs were analyzed.

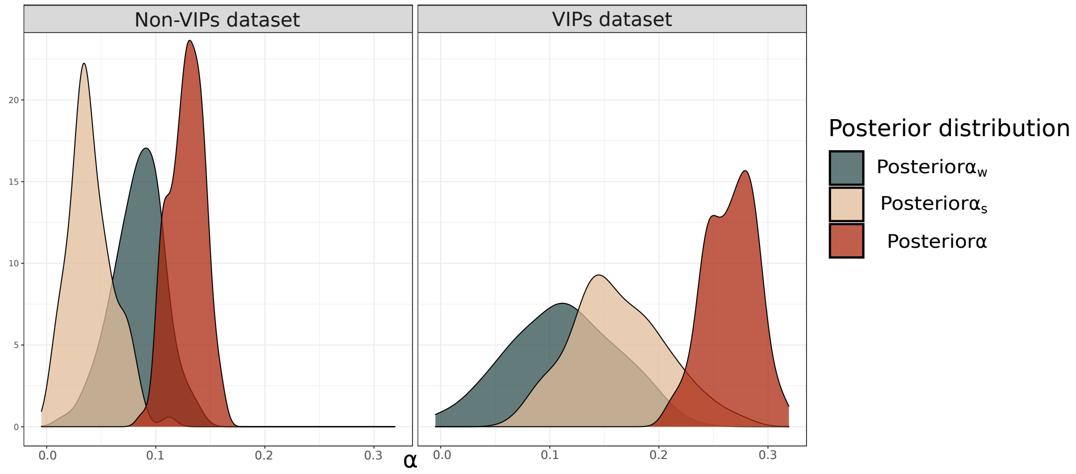


Figure 5.7: DNA-VIPs and RNA-VIPs posterior distributions. Inferences were performed using human lineage data from Uricchio et al. (2019). A total of 1258 DNA-VIPs and 3471 RNA-VIPs were analyzed.

Table 5.4: α_W, α_S, α, negative selection coefficient ($2N_es_-$) and shape parameter (β) in human datasets

Dataset	α_W	α_S	α	$2N_es_-$	β
Whole-genome	0.11 [0.067-0.157]	0.047 [0.021-0.078]	0.152 [0.131-0.173]	674.821 [421.362-853.764]	0.142 [0.136-0.148]
Non-VIPs	0.084 [0.053-0.109]	0.041 [0.017-0.071]	0.128 [0.106-0.146]	1320.172 [756.934-1635.432]	0.129 [0.122-0.135]
DNA-VIPs	0.06 [0.024-0.1]	0.133 [0.082-0.206]	0.205 [0.155-0.261]	554.467 [393.956-807.322]	0.188 [0.17-0.208]
RNA-VIPs	0.137 [0.045-0.230]	0.176 [0.112-0.247]	0.304 [0.275-0.334]	531.974 [380.099-808.089]	0.207 [0.194-0.22]

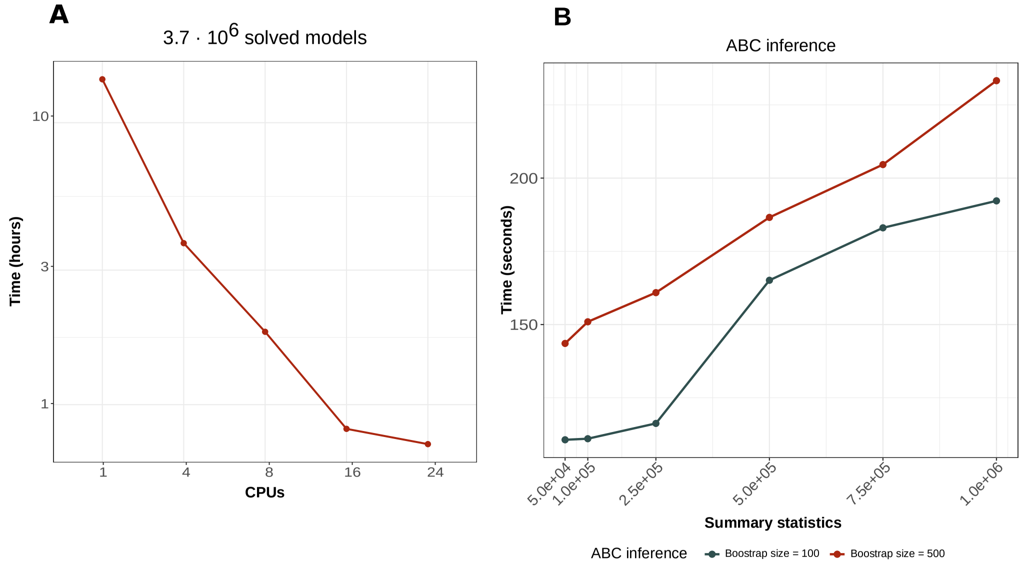


Figure 5.8: ABC-MK average running times. Both graphics represent the average time of 10 independent replicas. A. Average running time to solve $3.7 \cdot 10^6$ independent models solved from prior distribution to get the analytical fixation and polymorphic rates as a function of the number of CPU threads. B. Average running time to subset summary statistics given the random subset of the analytical rates and α inference using the empirical data.

Conclusions

The new ABC-MK approach works much more efficiently than the previous procedure in Uricchio et al. (2019), it can efficiently be executed on a workstation, while Uricchio et al. (2019) one requires the use of an HPC. Together, our analytical procedure, computation workflow and the Poisson sampling scheme allow us to avoid the expensive requirements of forward simulations, dramatically reducing execution time. Therefore, while for a given empirical dataset the Uricchio et al. (2019) approach takes several days to complete the estimation, the new procedure can be fully executed in less than 1 hour, depending on the number of CPU threads employed. More importantly, considering that we have made independent the analytical estimation of fixation and polymorphic rates and the Poisson sampling scheme, once the analytical rates have been estimated, the Poisson sampling scheme and the α inference can be performed in a few minutes given any dataset even without parallelizing (see figure 5.8). In addition, our software does not require a priori knowledge of B from McVicker et al. (2009) measuring BGS strength while efficiently relaxing our model assumptions to any selection coefficient at ABC estimations. We suggest that this new approach, as well as the exposed results, replace the previous version of the ABC-MK (Uricchio et al., 2019) software.

Chapter 6

iMKT: the integrative McDonald and Kreitman test

Abstract

The McDonald and Kreitman test (MKT) is one of the most powerful and widely used methods to detect and quantify recurrent natural selection using DNA sequence data. Here we present iMKT (acronym for integrative McDonald and Kreitman test), a novel web-based service performing four distinct MKT types. It allows the detection and estimation of four different selection regimes -adaptive, neutral, strongly deleterious and weakly deleterious- acting on any genomic sequence. iMKT can analyze both user's own population genomic data and pre-loaded Drosophila melanogaster and human sequences of protein-coding genes obtained from the largest population genomic datasets to date. Advanced options in the website allow testing complex hypotheses such as the application example showed here: do genes located in high recombination regions undergo higher rates of adaptation? We aim that iMKT will become a reference site tool for the study of evolutionary adaptation in massive population genomics datasets, especially in Drosophila and humans. iMKT is a free resource online at https://imkt.uab.cat.

Introduction

One of the most striking evidence of the power of natural selection is the characteristic footprints that it leaves on the patterns of genetic variation. A growing number of statistical methods to analyze genomic data allows us to detect and quantify adaptation and other selection regimes in the genome at different temporal scales (reviewed in Casillas and Barbadilla (2017)).

The McDonald and Kreitman test (MKT) (McDonald and Kreitman, 1991) is one of the most powerful and robust methods we have to detect the action of natural selection at the DNA level. MKT tests for the presence of recurrent positive (adaptive) selection on a gene or genome region. Unlike the $\omega = d_N/d_S$ ratio (Kimura, 1977), which uses only divergence data among species to compute the quotient of the number of non-synonymous (d_N) to synonymous (d_S) substitutions, the MKT uses both polymorphic and divergence data. Polymorphic data allows taking into account purifying selection on divergent non-synonymous sites, significantly increasing the detection power of recurrent positive selection. The MKT covers the evolutionary period spanning from the divergence of the outgroup species to the present. The null model of MKT is the neutral hypothesis (Kimura, 1968, 1983). Because infrequent adaptive mutations fix fast relatively to common neutral mutations, they contribute almost exclusively to divergence and not to polymorphism; therefore, an excess of the divergence ratio relative to the polymorphism ratio is the signal of positive selection. The fraction of adaptive nonsynonymous substitutions (α) can be estimated from the MKT data (Charlesworth, 1994; Smith and Eyre-Walker, 2002).

The main drawback of MKT is that it assumes strict neutrality of segregating sites. Because weak negative selection abounds in the genomes (Casillas and Barbadilla, 2017), α estimates are biased downward. Several MKT methodological extensions try to correct the bias. In the next section, four MKT approaches are listed: (i) the standard (original) MKT; (ii) the Fay, Wyckoff and Wu correction (fwwMKT) (Fay et al., 2001); (iii) the extended MKT (eMKT) (Mackay et al., 2012) and (iv) the asymptotic MKT (aMKT) (Messer and Petrov, 2013a). Each method has pros and cons are discussed, and for the comparison of their different outputs, it would be very convenient to have a web service to perform at once the four MKT. Existing web servers compute either the standard MKT (Egea et al., 2008; Vos et al., 2013) or more recently the aMKT (Haller and Messer, 2017). None of them contains pre-loaded population genomics data of representative species as Drosophila melanogaster or humans.

Here we present iMKT (acronym for integrative McDonald and Kreitman test), a web-based service performing the four MKT types described in the next section and Figure 6.1. It allows the detection and estimation of four selection regimes (adaptive, neutral, strongly deleterious and weakly deleterious) acting on protein-coding DNA sequences. The benefit of this tool is fourfold.

1. Four MKTs, two of which were not previously available as open software packages, can be performed at once to analyze user's own population genomic data in a simple interface offered by a web-based service.

2. It allows the simultaneous comparisons of the results of the different MKTs, which behave differently according to different properties of the data.

3. Taking advantage of the copious information gathered in previous population genome browsers, PopFly (Hervas et al., 2017) and PopHuman (Casillas et al., 2018), it offers a fast tool to estimate the different selective regimes on thousands of *D. melanogaster* and human protein-coding genes on several worldwide populations.

4. It allows comparing the selective regimes of a set of coding genes (selected according to the user's criterion, such as recombination rate bins or chromosome

localization) with those of the genome-wide distribution in both humans and D. melanogaster.

The incessant accumulation of massive genome data makes this website a timely resource to describe and quantify natural selection for any biological species at the genome level.

Material and methods

MKT methodologies

McDonald and Kreitman test (MKT). The standard McDonald and Kreitman test (MKT) (McDonald and Kreitman, 1991) was developed to be applied to protein-coding sequences, combining both divergence (D) and polymorphism (P) sites, and categorizing mutations as synonymous (P_S, D_S) and non-synonymous (P_N, D_N). If all mutations are either strongly deleterious or neutral, then D_N/D_S is expected to roughly equal P_N/P_S. In contrast, if positive selection is operating in the region, adaptive mutations rapidly reach fixation and thus contribute relatively more to divergence than to polymorphism when compared to neutral mutations, and then $D_N/D_S > P_N/P_S$ (Figure 6.1-A). Assuming that adaptive mutations contribute little to polymorphism but substantially to divergence, the proportion of non-synonymous substitutions that have been fixed by positive selection can be inferred as $\alpha = 1 - (\frac{P_N}{P_S} \cdot \frac{D_S}{D_N})$ (Smith and Eyre-Walker, 2002)(Figure 6.1-B). The main limitation of the test is the presence in the population of non-synonymous slightly deleterious variants, biasing downward the estimates of adaptive evolution (α). Below are three proposed methods to correct the bias.

Fay, Wyckoff and Wu correction (fwwMKT). Because slightly deleterious variants tend to segregate at lower frequencies than do neutral mutations, Fay, Wyckoff and Wu or FWW correction (Figure 6.2-C) propose to remove low-frequency polymorphisms from the analysis (Fay et al., 2001). α is estimated using the standard MKT equation but considering only those polymorphic sites (for both neutral and selected classes) with a frequency above an established cutoff. Charlesworth and Eyre-Walker (2008) showed that even removing low-frequency variants, the estimate of α is still downwardly biased. Only these estimates are reasonably accurate when the rate of adaptive evolution is high and the distribution of fitness effects of slightly deleterious mutations is leptokurtic (because leptokurtic distributions have a smaller proportion of polymorphisms that are slightly deleterious).

Extended MKT (eMKT). Mackay et al. (2012) proposed the extended MKT (Figure 6.1-D). Instead of simply removing low-frequency polymorphism below a given threshold, the count of segregating sites in non-synonymous sites is partitioned in the number of neutral variants (using neutral sites as a proxy) and the number of weakly deleterious variants. This increases the power of detecting adaptive selection (as it does not remove as much data as the fwwMKT) and allows the independent estimation of both adaptive and weakly deleterious substitutions. PN, the count of segregating sites in the non-synonymous class, is discomposed into the number of neutral variants and the number of weakly deleterious variants, $P_N = P_{N_{neutral}} + P_{N_{weakly\ del}}$ (Mackay et al., 2012). The estimation of both numbers allows estimating positive (adaptive) and negative selection independently. α is estimated from the standard MKT table discounting weakly deleterious variants: P_N is substituted by the expected number of neutral segregating sites, $P_{N_{neutral}}$. The correct estimate of α is then $\alpha = 1 - (\frac{P_{N_{neutral}}}{P_S} \cdot \frac{D_N}{D_S})$.

Asymptotic MKT (aMKT). Messer and Petrov (2013a) proposed an asymptotic extension of MKT that takes slightly deleterious mutations into account and yields accurate estimates of α (Figure 6.1-E). This method, named asymptotic MKT, is robust to the presence of selective sweeps (hitchhiking) and to the segregation of slightly deleterious substitutions (BGS). In this method, α is estimated in different frequency intervals (x) and these values are then adjusted to an exponential function, of the form: $\alpha_{fit(x)} = a + be^{cx}$. The asymptotic α estimate is obtained by extrapolating the value of this function to $\alpha_{asymptotic} = \alpha_{fit(x=1)}$.

The asymptotic MKT has been extended to estimate both positive (adaptive) and negative selection. aMKT requires a high volume of polymorphic data to fit the asymptotic function, being a suitable method in the case of concatenating numerous variants of multiple genes.

Input data

The iMKT server can analyze both user's own population data and pre-loaded data of *D. melanogaster* or human protein-coding genes.

In the first case, the user can upload as input either polymorphism and divergence data or aligned multi-FASTA files. For polymorphism and divergence data, the user must upload two files: (i) a tab-delimited file containing the distribution of Derived Allele Frequencies (DAF) (Ronen et al., 2013) of all segregating (polymorphic) variants for two types of sites (putatively under selection and putatively neutral), and (ii) a file containing the counts of divergent positions for the two site types. For aligned multi-FASTA files, the user needs to enter one or more files containing aligned protein-coding sequences for at least two sequences of the same species to estimate polymorphism counts, and one orthologous sequence from an outgroup species to estimate divergence and infer ancestral alleles. Examples of such files are provided at the website.

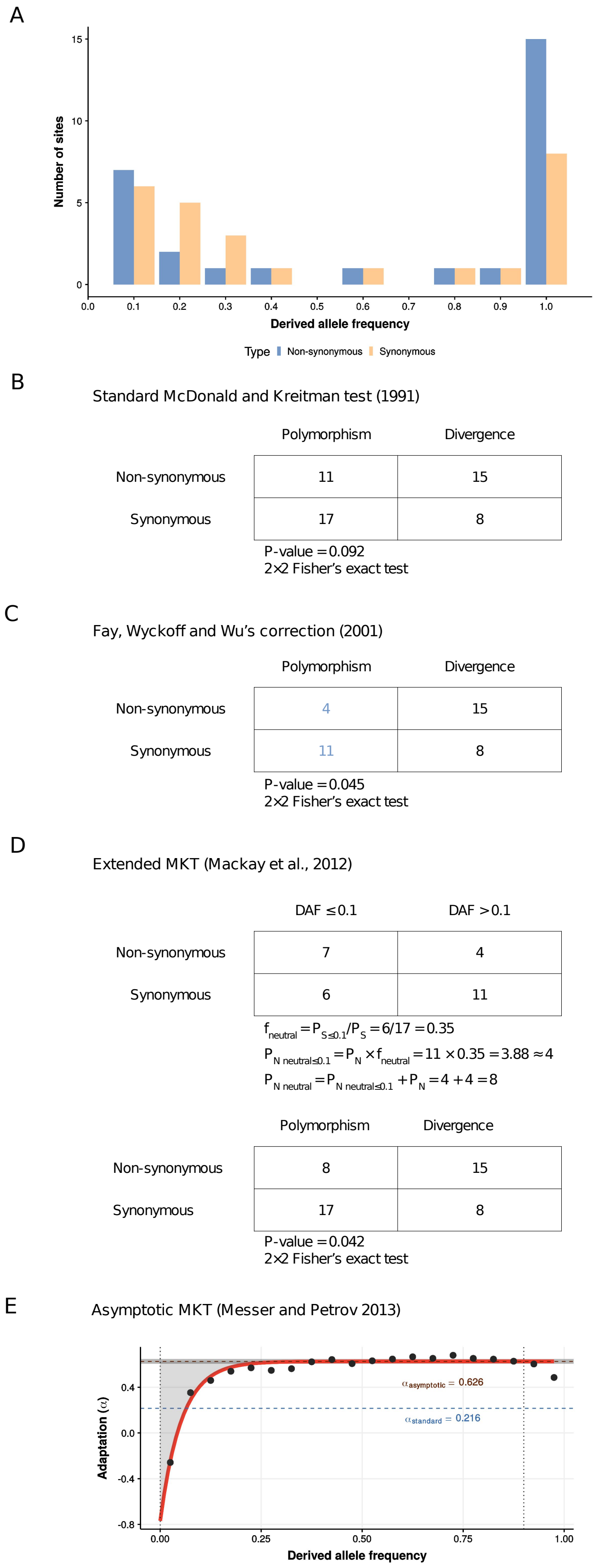
For analyzing *D. melanogaster* or human protein-coding genes, the user can use the population genomic data available in the web server. In this case, the user can either submit a list of protein-coding genes or select them from the list provided, and select the population(s) and preferred method(s) to analyze the selective regimes on a group of protein-coding genes.

Population genetics pipeline for *D. melanogaster* and human data

We have designed and implemented a custom pipeline for analyzing the Drosophila Genome Nexus (Lack et al., 2015, 2016) and Human 1000GP Phase III (Auton et al., 2015) data, which could potentially be escalated to any available genomic data source. The pipeline pre-calculates the DAF and number of divergent synonymous and nonsynonymous sites, which are needed to further perform on-the-fly MKTs. A total of 13,753 protein-coding genes for 16 *D. melanogaster* populations (Lack et al., 2015, 2016) and 20,643 protein-coding genes for 26 human populations of distinct geographical origin (Auton et al., 2015) were analyzed. Pre-calculated DAF and divergence values are stored in the server. The complete pipeline is available as a Jupyter Notebook at https://github.com/BGD-UAB/iMKTData to allow its reproducibility.

Data retrieval

***D. melanogaster* population genomic data.** Variation data generated by the Drosophila Genome Nexus, together with divergence data between *D. melanogaster* and *D. simulans*, was retrieved from PopFly (Hervas et al., 2017) in FASTA format. Only populations with at least four genome sequences and less than 20% of missing or ambiguous nucleotides each (after filtering by identity by descent, admixture, and



B

Standard McDonald and Kreitman test (1991)

	Polymorphism	Divergence
Non-synonymous	11	15
Synonymous	17	8

P-value = 0.092
2×2 Fisher's exact test

C

Fay, Wyckoff and Wu's correction (2001)

	Polymorphism	Divergence
Non-synonymous	4	15
Synonymous	11	8

P-value = 0.045
2×2 Fisher's exact test

D

Extended MKT (Mackay et al., 2012)

	DAF ≤ 0.1	DAF > 0.1
Non-synonymous	7	4
Synonymous	6	11

$f_{neutral} = P_{S\leq 0.1}/P_S = 6/17 = 0.35$

$P_{N\ neutral\leq 0.1} = P_N \times f_{neutral} = 11 \times 0.35 = 3.88 \approx 4$

$P_{N\ neutral} = P_{N\ neutral\leq 0.1} + P_N = 4 + 4 = 8$

	Polymorphism	Divergence
Non-synonymous	8	15
Synonymous	17	8

P-value = 0.042
2×2 Fisher's exact test

E

Asymptotic MKT (Messer and Petrov 2013)

Figure 6.1: (Caption in next page)

Figure 6.1: Comparison of the four MKT methods implemented in iMKT. (A) The hypothetical derived allele frequency (DAF) spectrum of synonymous and non-synonymous classes for a gene exhibiting an excess of both slightly deleterious and fixed non-synonymous differences with $n = 10$ sampled chromosomes. (B) The standard MKT for this gene ($P - value = 0.09$, $2x2$ Fisher's exact test). (C) The 22 table by Fay, Wyckoff and Wu's correction (Fay et al., 2001)) taking into account only polymorphism found on more than one chromosome ($P - value = 0.045$, $2x2$ Fisher's exact test). (D) Extended MKT (Mackay et al., 2012). The count of segregating sites in non-synonymous sites is partitioned into the number of neutral variants and the number of weakly deleterious variants. PN is substituted with the number of nonsynonymous polymorphisms that is neutral ($P - value = 0.042$, $2x2$ Fisher's exact test). (E) Asymptotic MKT. Example of the result of asymptotic MKT using*D. melanogaster* 2R chromosome and *D. simulans* as outgroup. The two vertical lines show the limits of the x cutoff interval used (in the example [0,0.9]). Black dots indicate the binned values for each DAF category. The solid red curve shows the fitted $fit(x)$. The dashed red line is the final asymptote. The dark gray band indicates the 95% CI around the estimation. The blue dashed line shows the estimated using the standard MKT for comparison. For MKT methods definitions, see Section 6. Adapted and expanded from (Hahn, 2018)

heterozygosity) were included. DAF spectrum by functional classes was estimated by resampling a number of lines with nucleotide information (excluding undetermined sites, *N* bases) at each position without replacement. This procedure maximizes the number of informative sites to analyze. The number of lines resampled for each population was chosen depending on the number of lines sequenced and the quality of those sequences (Table C.1). Positions and genes without valid information for at least this defined number of lines were discarded for the analysis. The ancestral state of each polymorphic site was inferred from the comparison with the outgroup species *D. simulans*. The genome reference sequence and annotations correspond to the 5.57 FlyBase release (Thurmond et al., 2019). Gene-associated recombination rate for 100 kb non-overlapping windows were retrieved from Comeron et al. (2012).

Human population genomic data. Genome variation data and ancestral state of variants generated by the 1000GP Phase III (Auton et al., 2015), together with divergence estimates between humans and chimpanzees, were retrieved from PopHuman (Casillas et al., 2018) in Variant Call Format (VCF). The dataset included 84.4 million variants detected across 2,504 individuals from 26 different populations, which were mapped to the human reference genome version GRCh37/hg19. Reportedly inbred individuals (Gazal et al., 2015) and non-accessible nucleotides (Auton et al., 2015) were discarded following the PopHuman methodology (Casillas et al., 2018). Genome annotations were retrieved from GENCODE (release 27). Recombination rate values associated with each protein-coding gene were obtained from Bhérer et al. (2017) and correspond to the sex-average estimates.

Estimation of the number of synonymous and nonsynonymous changes

Inferring the action of natural selection on coding sequences relies on the computation of polymorphism and divergence data on two distinct types of sites in the genome: one putatively selected (usually non-synonymous coding sites), and one putatively neutral (usually synonymous coding sites) (McDonald and Kreitman, 1991). This implies assigning a selective class for each nucleotide site in the genome. This task is not trivial when different transcripts overlap a genomic region. For example, one nucleotide site can be a non-synonymous site for one transcript but a synonymous site for another nested gene transcript. In these cases, we assign the most selective constrained class to the nucleotide site. In the example, the site is considered non-synonymous.

Exclusion of low-frequency variants

Slightly deleterious variants are mainly segregating at low frequency (Fay et al., 2001; Templeton, 1996; Charlesworth and Eyre-Walker, 2008). These rare polymorphisms can be excluded from the analyses by specifying one or several threshold frequency values depending on the fwwMKT, the eMKT or the aMKT method. In addition, the aMKT allows removing high-frequency variants that might be due to polarization errors (Messer and Petrov, 2013a; Haller and Messer, 2017).

Statistical analysis

For analyses including several protein-coding genes, users are recommended to select the option Concatenate genes. In this case, iMKT analyzes the selective regimes for the whole gene set instead of for each gene separately and applies a statistical test of heterogeneity of the selection acting among the analyzed genes (Cochran-Mantel-Haenszel statistic). In addition, the iMKT web server allows performing statistical enrichment analyses to assess whether a group of genes is either enriched or depleted of positively selected genes when compared to the complete genome distribution or to a second group of genes submitted by the user. In this case, the user should choose also the option Compare against whole-genome distribution or Compare against a second dataset. A resampling 95% confidence interval (CI) is generated by estimating α with the chosen MKT test for 100 bootstrap replicates by sampling genes with replacement within each group. In the asymptotic MKT, 95% CI intervals around the α estimation are already provided in the output.

Output

The output of iMKT is an extensive report displayed as an HTML page. It contains several sections, starting with a summary table with the input parameters, a table with descriptive statistics, and the standard MKT table. Finally, the tests selected by the user are displayed below.

Practical guide to the iMKT website

The iMKT site allows performing four MK-derived tests as a web-based service. The website is divided into different sections, each of which allows performing different types of analyses.

MKT analysis

This page allows performing diverse MK-derived tests and estimating different selective regimes in your own data. The input can either be polymorphism and divergence data in two separate files, as described in the Methods section, or protein-coding sequences as aligned multi-FASTA files. When a multi-FASTA file is uploaded, the server outputs the DAF spectrum and the divergence calculations, which can be downloaded by the user and used in subsequent analyses. Note, however, that the former input type gives more flexibility to analyze any sort of functional site. As an example, you might want to test for selection at nonsynonymous coding sites (N) compared to synonymous coding sites (S) as the classical MKT was formulated, or to test for selection at Conserved Noncoding Sequences (CNS, N) compared to non-CNS (S) (Casillas et al., 2007), etc. The choices are unlimited according to the user's needs.

PopFly/PopHuman data analysis

If you want to analyze *D. melanogaster* or human protein-coding data, iMKT contains readily available variation data obtained from the largest genome variation datasets in each species (see Methods). The first step is to select the genes to be analyzed in the table displaying all the available genes. Genes are identified by either the Gene symbol or the FlyBase/Ensembl ID. Genes in the table can be sorted/filtered by chromosome and recombination rate, in addition to the Gene symbol and Flybase/Ensembl ID. In case the user needs to analyze a specific list of genes that cannot be easily filtered from the provided table (e.g. genes related to a specific pathway, as obtained from a search in KEGG (Kanehisa et al., 2019), a list with those genes, were genes are identified by symbol or FlyBase/Ensembl ID, can be uploaded. Second, one or more populations on which to perform the analysis need to be specified. Third, one or more MK-derived methods can be chosen. Finally, advanced options are available to analyze all the genes as a group instead of analyzing them separately (option Concatenate genes), and to compare the results of this gene set against all the genes of the genome (option Compare against whole-genome distribution) or against a second group of genes provided by the user (option Compare against a second dataset). Potential applications include analyzing a single protein-coding gene or exploring different selective regimes in genes that are expressed tissues, anatomic structures, or developmental stages (Salvador-Martínez et al., 2018)).

Other sections of the website

The iMKT website includes extensive methodological and technical documentation (see the section Documentation in the website), as well as a complete tutorial on the usage of iMKT, with step-by-step examples (see the section Help and tutorial from the main page). The website also contains sample files for each available type of analysis and links to related resources such as PopFly, PopHuman, and the iMKT R package.

Application example of the iMKT website

The iMKT website is designed to help testing evolutionary hypotheses from a population genetics perspective. The online tutorial, apart from guiding you in the usage of this resource, contains some worked-out cases that can be addressed using iMKT. In the application example developed here, we want to assess whether recombination rate limits the adaptive potential of protein-coding genes. The specific hypothesis is that genes located in high recombination regions undergo higher rates of adaptation. To test the hypothesis, we start by entering the PopFly data analysis page of iMKT. Next, we use the filtering options below the table to select 475 genes having a recombination rate higher than 7 cM/Mb (Min recombination rate: 7). After selecting the genes, we select one or more populations (United States (RAL)) and an MKT test (eMKT). Finally, we choose the option Compare against whole-genome distribution, which compares the distribution of α for the selected 475 genes located in regions of high recombination against the corresponding distribution for all *D. melanogaster* genes. As part of an extensive output report, an illustrative box plot shows a pronounced difference in the level of adaptation (α) between genes located in regions of high recombination (blue; α mean = 0.602; ±SD = 0.032) and all 13,753 *D. melanogaster* genes (orange; m α mean = 0.44; ±SD = 0.055) (Figure 6.2-A).

We can repeat the same procedure for the *D. melanogaster* ancestral population from Zambia (Zambia (ZI)). As previously, the output report uncovers a much higher level of adaptation (α) in genes located in regions of high recombination (blue; α mean = 0.633; ±SD = 0.028) compared to the total 13,753 *D. melanogaster* genes (orange; α mean = 0.457; ±SD = 0.053) (Figure 6.2-A).

Finally, the same analysis in humans for a colonizing population (Utah Residents (CEU)) and an ancestral population (Yoruba (YRI)) reveals negative α adaptation values in most cases and differences between the two groups of genes compared (Figure 6.2-B). The results of this straightforward analysis show that: (i) *D. melanogaster* undergoes higher rates of adaptation than humans; and (ii) genes located in regions of high recombination undergo higher rates of adaptation in both *D. melanogaster* and humans.

The example application developed here illustrates the power of iMKT to reveal new knowledge about evolutionary processes in Drosophila and humans without the need for labor-intensive data retrieval and/or processing by the user. The wide range of potential queries that can be performed using the searching capabilities of the iMKT website remarkably facilitates comprehensive analyses of evolutionary adaptation and constraint, even for non-bioinformaticians. As such, iMKT is a comprehensive reference site for the study of protein adaptation in massive population genomics datasets, especially in Drosophila and humans. Finally, we want to emphasize that the flexibility of iMKT to input custom data allows analyzing diversity data outside protein-coding regions. This expands, even more, the hypotheses that can be tested and makes iMKT a key tool to test for recurrent adaptation in the genome of any species.

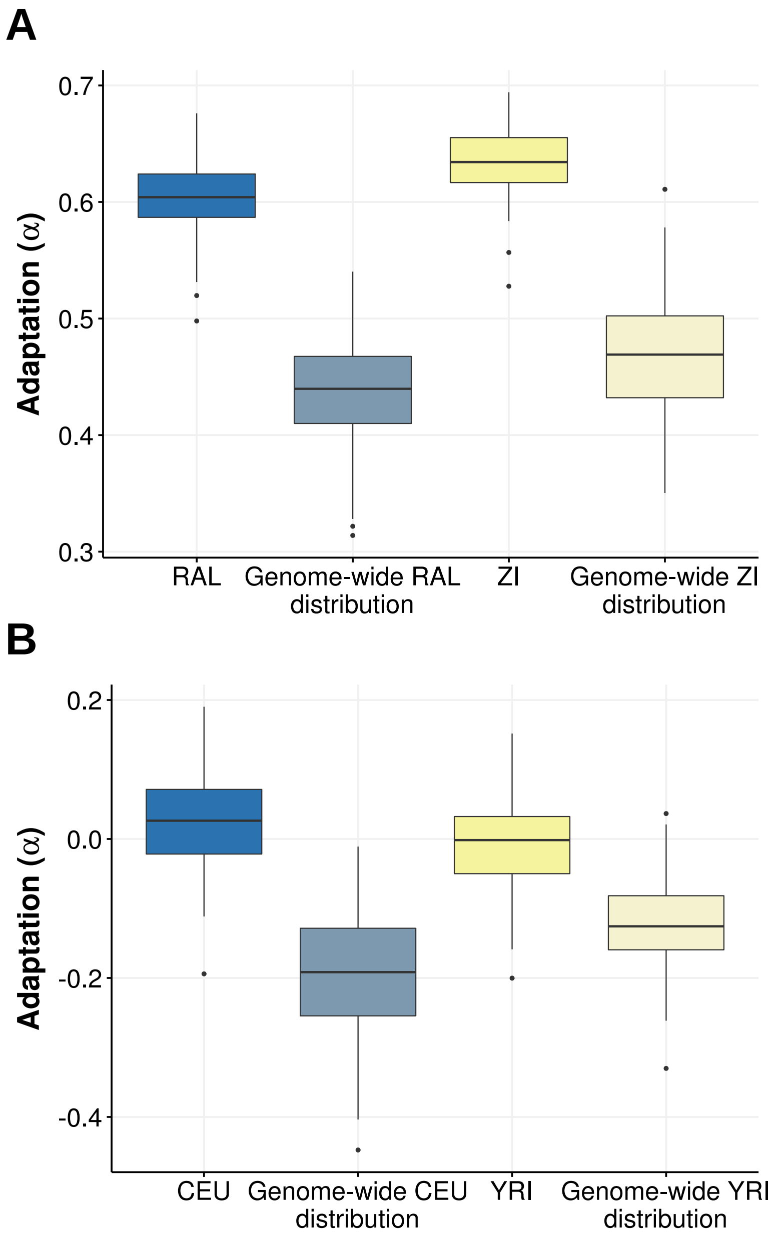


Figure 6.2: iMKT graphical output of an application example. Sampling distribution of α values for protein-coding genes located in regions of high recombination (recombination rate $> 7\ cM/Mb$) compared to all protein-coding genes in the genome for (A) the *D. melanogaster* Raleigh (RAL) population (blue) and the *D. melanogaster* Zambia (ZI) population (yellow) and (B) the human Utah Residents (CEU) population (blue) and the human Yoruba (YRI) population (yellow). The distribution was calculated by randomly sampling 400 genes 100 times from the two lists of genes with replacement and estimating α in each bin. Polymorphisms with a frequency below 0.05 in the analyzed population were discarded (see main text).

Chapter 7

Discussion

This thesis outlines and compiles the evidence of positive selection events in the human lineage using data from the 1000 Genome Project (1000GP). The work done can be divided into two parts. The first, a comprehensive study at the genomic and populational level of the 1000GP data to catalog positive selection events. The second, the creation of statistics that improve the existing methodologies for detecting recurrent positive selection through the revision of the MKT applied at both the genetic and genomic levels. The exhaustive review of the MKT approaches was prompted by the scarcity of results we encountered in Chapter 3.

We brought together the collective effort of the last 15 years aiming to catalog positive selection in the human lineage. Nevertheless, in this thesis, the processing of the data, the statistical methodologies employed and the new extensions of the MKT all have been carried out taking into account the current population genomic data. Thus, both the bioinformatic framework and the theoretical framework have been examined using not only data from the 1000GP project, but also data from the Drosophila Genome Nexus (DGN) project. Table 1.2 shows how the number of nucleotide variation catalogs has increased since the launch of the 1000GP and the DGRP projects. With this in mind, we opted to make the existing methodologies as flexible as possible. Accordingly, this thesis encompasses the largest catalogs of nucleotide variation in order to extrapolate our analyses to other species; and while we are aware that the methodologies presented here are unlikely to be fully transferable (especially with regard to viruses and bacteria), our aim was not to solely focus on the specific case of humans. Taking into account the quality and intrinsic characteristics of the data deposited in other variation catalogs, such as 1001G (Alonso-Blanco et al., 2016), ag1000G (Miles et al., 2017), SGDP (Mallick et al., 2016) or HGPD (Bergström et al., 2020), the pipelines and statistics of recurrent positive selection described in this thesis should be fully reproducible on these datasets.

In addition, a significant proportion of the thesis is based on data we have generated *in silico* through large-scale forward-in-time simulations. These simulations ultimately attempt to reproduce, in the most reliable way possible, empirical data, under complex demographic models and different modes of selection. In doing so, our simulations assess underlying patterns of nucleotide variation beyond those exhibited in the human and *D. melanogaster* genomes.

7.1 Genome-wide scan of positive selection in the human lineage

As described in the introduction (see Section 1.4.2), the outlier approach involves an integrative search for positive selection events in which the whole genome can be systematically examined. Thus, the analysis takes into account the genomic context, ultimately elucidating the role of natural selection, distinguishing it from other confounding variables, such as the population demographic history. Over the last two decades, the number of Genome-wide Scans (GWS) has been growing steadily. Table 1.2 gives an approximate idea of the number of studies carried out, as well as the data and methodologies used in the detection of positive selection. It is a compendium that reviews and adds to previous studies, such as those found in Akey (2009), Haasl and Payseur (2016), or Lohmueller and Nielsen (2021). Several important facts should be noted from Table 1.2, as well as from the original table compiled by Haasl and Payseur (2016).

As stated in Haasl and Payseur (2016), between 1999 and 2009, 35 of 49 (71%) of GWSs focused on humans, while from 2010 to present, only 38 of 83 (46%) of GWSs focused on humans, indicating an increased focus on other model and non-model organisms. Furthermore, the trend of sequencing is likely to continue to increase, thanks to the decreasing costs and progress of new sequencing technologies. A clear example of this is the spatio-temporal catalog developed by Kapun et al. (2021) as a result of advances in Pool-Sequencing. Furthermore, population genetics has shown over the last 50 years that model organisms are crucial in understanding the mechanisms promoting natural variation in populations.

Considering this trend and the extensive search for selection patterns in humans in the last decade, at present we have several positive selection maps pinpointing the proportion of the genome putatively selected (Voight et al., 2006; Pickrell et al., 2009; Johnson and Voight, 2018). During the last few years, there has been a move towards other unresolved questions in the field, including inferences of historical recombination (Adrion et al., 2020b; Barroso et al., 2019), mutation rate (DeWitt et al., 2021; Barroso and Dutheil, 2021), or the history and population structure of present and past human populations (Speidel et al., 2021; Wohns et al., 2021).

Nonetheless, returning to Table 1.2, several features shared by the previous studies should be noted. In the first place, only a small percentage ($\approx 10\%$) of these incorporate Whole-Genome Sequencing (WGS) data, and only. About 13% incorporate either WGS or Whole-Exome Sequencing (WES) data. Hence, a large percentage of regions putatively targeted to date have been identified through genotyping technologies. As described in detail in Clark et al. (2005), genotyping techniques have led to a persistent problem of SNPs ascertainment bias. This problem is demonstrated in the Clark et al. (2005) study through the allele frequency distributions of the most relevant datasets, on which most of the studies in Table 1.2 are based. For example, as shown in Figure 1 from Clark et al. (2005), in the HapMap dataset the Site Frequency Spectrum (SFS) shows an absence of rare alleles and an overrepresentation of intermediate frequency alleles compared to what is expected by strict neutrality or what is found in the Perlegen dataset. Such ascertainment bias affects the inferences of natural selection, specially those scans that rely on statistics testing distortions in the SFS, but also affects the False Discovery Rate and coherence among GWSs' candidate regions. In addition, as extensively explored in Nielsen and Signorovitch (2003), ascertainment bias results in lower apparent LD too. While these panels certainly provided the first genomic data of human populations, we cannot ignore the fact that they were not explicitly designed to be applied to population genetics but for testing association between common SNPs and risk of complex disorders (Clark et al., 2005). Therefore, the description of nucleotide variability and inference of natural selection or demography have been biased due to the lack of randomness in the discovered SNPs. This is especially true if we consider that genotyping techniques were biased toward identifying polymorphism within European ancestry (Clark et al., 2005; Lohmueller and Nielsen, 2021).

The use of WGS data has overcome, at least in part, this problem, and one can expect that:

1. We now have a more complete representation of the putatively selected regions;
2. Due to higher resolution and less biased technique, there is a greater concordance between the events detected among different WGS studies.

Nonetheless, again Table 1.2 shows that most studies often include the same number of population groups incorporating one African (YRI), one European (CEU) and one Asian (CHB+JPT) population, and none devotes any particular attention to previous studies beyond the classic examples such as LCT, EDAR or G6DP. Therefore, despite providing new metrics and essential conclusions, previous studies, including WGS analysis, usually ignore how many of these events were already known, detected, undetected or in which populations using the previous methodologies.

Our GWS of positive selection throughout the human lineage is the most comprehensive study of positive selection events, including a total of 22 populations from phase 3 of the 1000GP project, as well as recurrent positive selection. We show:

1. The lack of concordance between previous studies.
2. The need for a catalog that brings together these events according to the detected selection patterns, populations studied and methodologies employed.

Finally, to the compendium included in Haasl and Payseur (2016), and more recently in Lohmueller and Nielsen (2021), we must now add and emphasize the works of Garud et al. (2021) and Schrider and Kern (2017). These, for the first time, focus their analyses on the detection of selective events on standing variation and use methodological advances that have not been included in this thesis. These developments will be discussed in the next section.

7.1.1 Summary of statistic selection

At least three previous works have a similar scope of this thesis: Schrider and Kern (2017), Sugden et al. (2018) and Johnson and Voight (2018). These works, together with the one by Grossman et al. (2013), are probably the most important and comprehensive GWSs to date. As in our case, the authors conducted their research using the 1000GP data. In the case of Schrider and Kern (2017), the research was conducted across six populations. Sugden et al. (2018) employs its methodology using the 1000GP phase I data. Finally, Johnson and Voight (2018) incorporates the entire 1000GP phase III panel. These studies were carried out almost simultaneously to the one presented in Chapter 3.

Both Schrider and Kern (2017) and Johnson and Voight (2018) studies focus on soft and hard sweeps detection, respectively, whereas Sugden et al. (2018) does not use the 1000GP phase III data. On the one hand, Schrider and Kern (2017) focuses on the detection of complete sweeps through a machine learning method, which is also able to classify between hard and soft sweeps. On the other hand, Johnson and Voight (2018) focuses its research on ongoing or partial sweeps, reviewing the selection measures provided by iHS and normalizing the calculation by the local recombination ratio.

In our analysis, not only do we examine both cases, but we also include other statistics and other selection regimes. Clear examples are the XP-EHH and Fay and Wu's H statistics. Similarly to iHS, XP-EHH can detect ongoing or partial sweeps. However, XP-EHH is of particular interest for cases where the selective sweep is on the verge of becoming fixed, thereby increasing the power of iHS if the selected mutation is at a frequency higher than 80% (Pickrell et al., 2009). In turn, Fay and Wu's H statistical test is specially designed to detect nearly or recently fixed sweeps. So, the selection of the summary statistics used in Chapter 3 is not arbitrary. By combining such statistics, we get a complete description of the positive selection events in the human lineage. Hence, our study examines the main neutrality statistics (specially designed to measure unexpected patterns of high frequency derived and rare alleles), population differentiation, the profiles of unusually long patterns of linkage disequilibrium, as well as recurrent positive selection signals. In this way, our study not only explores partial or complete sweeps independently but also a broader spectrum of adaptive signals.

It is important to note that there is no single test capable of measuring all types of selection. Some significant attempts to provide a single measure capable of determining a selective event, regardless of its nature, include the methodologies proposed by Sugden et al. (2018) and Schrider and Kern (2017). These studies result from technological advances in machine learning. Other machine learning methodologies (or classification methodologies, which also make use of the statistics described above) have also been proposed in the past, such as Pybus et al. (2015) (a prime example applied to the 1000GP phase I) or Ronen et al. (2013). Machine learning methodologies are becoming particularly relevant, as they are proving tremendously helpful not only in unraveling the characteristic genome-wide footprint of selection, but also other population genetics events, such as introgression, demography or recombination (Kern and Schrider, 2018; Mondal et al., 2019; Adrion et al., 2020b; Gower et al., 2021).

Most of these approaches train their models by computing a set of statistics on in silico data, usually generated through coalescent simulations. Then, the model produces vectors or images, through which it learns to distinguish the presence of genomic regions subjected to selection, or between hard and soft sweeps. Although the aim of this thesis is not to discuss the role that machine learning is playing, neither the technical issues of each methodology nor their application to 1000GP phase III data, given the studies of Schrider and Kern (2017), Sugden et al. (2018) (or even more recently Torada et al. (2019) and Hejase et al. (2021) among others), it seems almost inevitable to compare the machine learning approaches with the selection of statistical tests employed here.

Firstly, as we did in our study, a machine learning approach may seem more powerful than the simple analysis and cataloging of events according to each signal described. However, most of the named methodologies use a combination of statistics equal or very similar to the one chosen in Chapter 3. The difference, of course, is the integration and automated learning based on the features collected from the entire selected set. Therefore, as with ABC approaches, the selection of summary statistics is a key factor in these methodologies. However, even simple selection can pose a problem. Many of the early studies that used machine learning technologies or integrated different summary statistics required the pre-calculation of all statistics in a particular region. As stated in Sugden et al. (2018): *we found that more than half of variant sites had at least one undefined component statistic (...) This poses a particular problem when scanning for complete sweeps, defined here as sweeps in which the beneficial allele has fixed in the population of interest.* In this case Sugden et al. (2018) refers to the probabilistic Composite Multiple Signal methodology (Grossman et al., 2010), which will also be discussed below. However, it applies to other methodologies such as the one presented in Pybus et al. (2015).

On the other hand, there is the bias of the data on which the model is trained. In most cases, in silico data is generated through coalescent simulators, including simple positive selection models, such as `msms`, `cosi2` or `discoal`. Therefore, because of the coalescence simulator, although complex demographic models can be included, the data can hardly replicate the characteristics of the human genome, which ultimately introduces biases in the training. Thus, the reliance on simulations can be considered one of the weaknesses of these methodologies. In the following sections, we will extend the discussion about these simulations.

Despite the novelty of automatic classification and learning approach, it is impossible not to note the similarities between machine learning methods and the methodology proposed by Grossman et al. (2010). As with machine learning methodologies, CMS requires simulations and the calculation of a set of tests. CMS, employing purely statistical processing, combines the characteristics of each test, and using the simulated data, provides the probability of finding a selective sweep in the observed data, in the same way as other machine learning methods such as the one proposed by Sugden et al. (2018) do. The methodology of Grossman et al. (2010) was reviewed and applied to the first phase of the 1000GP phase I data (Grossman et al., 2013) and, as demonstrated by the number of citations, can be considered the most comprehensive and complex GWS carried out to date. Thus, Grossman et al. (2013) results continue to be the basis on which any study on positive selection in humans is based, including the results of this thesis.

Notwithstanding, the hard work and technological advances described above, and regardless of the debate concerning the simulations and/or the summary statistics used, it is essential to highlight that none of these studies pay attention to recurrent positive selection events. Hence, our choice of statistics and the rest of the chapters in this thesis are more pertinent to describing positive selection events in their multifold facets, independently of the nature of the signal. Thus, as expressed in Section 1.2.6 and Sabeti et al. (2007), we can expect that the patterns captured by the summary statistics here employed correlate with the expected fixation times and mostly overlap the periods, as is shown in Figure 1.9. Thus, although we do not summarize the results in a single value, our study covers the description of selective events in the human lineage starting with the separation from the chimpanzee and including the main Out-of-Africa (OoA) demographic events. In addition, our results also address other types of selection due to SFS distortion, although these results are secondary to the main purpose of the thesis.

7.1.2 Outlier approach and arbitrary cutoff

Except for the evidence of recurrent positive selection events, the putative regions evaluated in Chapter 3 were discovered following an outlier approach. We described the basics and main features of the outlier approach in Section 1.4.2. Akey (2009) provides a complete overview of the technique. However, as seen in Chapter 3, the outlier approach establishes an arbitrary cutoff. This decision might increase the already described problems related to outlier approaches. In any case, our decision to establish a cutoff is not likely to reflect a further increase of false positives. Our study sets a threshold *P-value* of $5 \cdot 10^{-4}$, which is lowered to $5 \cdot 10^{-3}$ in contiguous regions. Such a constraint limits the detection of the strongest signals along the genome and leads to a high number of false negatives, which could account for our inability to detect known events, such as G6PD or the HBB cluster, as reported in Johnson and Voight (2018).

The outlier approach performance has been extensively examined, especially in the studies of Kelley et al. (2006) and Teshima et al. (2006), as stated in the Introduction. By investigating various neutrality tests and simulations in coalescence, both studies show that, if the purpose of the study is to identify a restricted set of candidates for a more in-depth description, the outlier approach is the optimal procedure. Kelley et al. (2006) suggests that an outlier approach is a reasonable study design as long as one accepts that a substantial proportion of candidates may be false positives or false negatives. For us, the aim is not to study certain regions in detail, but to establish a methodology capable of creating a catalog that identifies putative regions in new populations and recovering those already described. Since the studies carried out by Kelley et al. (2006) and Teshima et al. (2006), the role of simulations and the cutoff has not been revisited comprehensively. However, some studies extensively review the role of other types of selection, especially the role of background selection (BGS) in such approaches. This type of linked selection may increase the heterogeneity of the statistics used, creating patterns and thus unusual distributions that could increase the number of false positives and false negatives.

There has been much debate about how other processes, such as positive selection or BGS, may affect the different applied summary statistics and analyzed empirical distributions. While a model which incorporates the effect of linked selection could, in principle, provide much more accurate cutoffs, the role of linked selection on the levels of variation (and consequently the summary statistics) remains a significant subject of debate. Moreover, over the last decades, it has been postulated that such a model could even provide a solution to Lewontin's paradox (Lewontin, 1974; Leffler et al., 2012; Buffalo, 2021). Nevertheless, so far, none of even the most comprehensive studies on this topic have succeeded in yielding more than a few conclusions (Corbett-Detig et al., 2015; Buffalo, 2021). Consequently, none of the GWSs described incorporate BGS or complex selection models through which to establish more precise threshold values. Therefore, just as we accept that the demographic effect is ubiquitous, we accept that the BGS effect is also ubiquitous and then unusual patterns associated with linked selection have also to be considered in the empirical distribution.

If we adopt the neutral model in its entirety, we can obtain null distributions through simulators in coalescence that incorporate complex demographic histories. Furthermore, to generate these distributions, we can use simulators that include selection models, meaning we could also discern the nature of the selective event. As one might expect, this is the standard approach taken by all GWS of positive selection, as well as the basis of the exploration carried out by Kelley et al. (2006) and Teshima et al. (2006).

However, little attention has been paid to the effects of variation in recombination rates on the distribution under the neutral model. Booker et al. (2020) discusses at length how GWSs discern the role of selection with regard to deviations from a null model (nearly-neutral theory), which explains natural variation primarily by the action of evolutionary drift. What Booker et al. (2020) points out is how variation in the rate of recombination also has important effects on the behavior of the null model. In their study, the exploration of various statistics according to the recombination values shows much more spurious distributions in regions with low recombination rates and narrower ones in regions with high recombination rates (Booker et al., 2020). To our knowledge, only the study of Johnson and Voight (2018) incorporates the role of local recombination. In this case, Johnson and Voight (2018) does correct the iHS values through the estimation of the local recombination rate, but fails to correct the F_{ST} values.

Similarly, the null model is likely to be flawed due to the simplification of the mutation rate. Simulations commonly incorporate a fixed mutation rate. In contrast, we have obtained empirical evidence over the past several years that the mutation rate is a complex and dynamic process and should be considered on a time scale. This is demonstrated by the $TCC \rightarrow TTC$ enrichment in European populations, as described by Harris and Pritchard (2017). Among the findings obtained from the methodology of (Spence and Song, 2019), the ARG inference was able to date such an event to 5,000-30,000 years. This pulse in European populations is one of the great current debates. Recently, DeWitt et al. (2021) has dated the event to 80kyears. DeWitt et al. (2021) proposes an inference method that makes use of the genomic context, condensing the SFS information into k-mer nucleotide context. The DeWitt et al. (2021) approach, for the first time, shows the human mutation rate over time and associates the error in the TTC pulse dating due to the demographic model and the de novo mutation rate values themselves, which cannot adequately account for the observed SFS.

Briefly, synthesizing the recent breakthroughs of the Booker et al. (2020) and DeWitt et al. (2021) studies, we would like to stress that the simple act of simplifying both the recombination and the mutation rate in the null model leads to the biasing of the empirical distributions produced by the simulations. Ultimately, this will affect any window-based approach to genome scans in more general terms. By taking this bias into account, we can determine that there is not much difference between our approach and the approaches which determine a cutoff from the data generated by coalescent simulations. Nonetheless, once the simulations incorporate accurate models in mutation rates, recombination, and selection coefficients, the increase in variance in summary statistics due to perturbations will be similar to that of natural populations. So, the question is, can we define a cutoff that, without taking into account simulated data, can reliably show us (at the very least) the strongest signals of adaptation?

As described by Akey (2009), one way to assess the validity of the GWS results is to examine the overlap of outlier loci between studies. This is nonetheless a simple approach. Both Akey (2009), and ourselves were able to do so thanks to the rapid accumulation of genomic maps of positive selection. In total, Akey (2009) identifies 5110 distinct regions in one or more studies. These regions span 4Mb of sequence (14% of the genome) and contain 4243 genes (23% of all genes). Strikingly, only 722 regions (14.1%) were identified in two or more studies, 271 regions (5.3%) were identified in three or more studies, and 129 regions (2.5%) were identified in four or more studies. In addition, the integrated positive selection map does not include several of the most compelling genes with well-substantiated positive selection claims, such as G6PD and DARC. Proceeding in the same way, our study, for the first time, manages to compile 70% of the detected regions, including not only the studies chosen by Akey (2009), but also later studies that included information from the 1000GP data, as well as the dbPSHP database (Li et al., 2014). Such results demonstrate two things. First: the efficiency of the chosen test set, which captures most of the features produced by the different types of selective events. Second: the importance of the quality of the initial data and its rigorous statistical treatment. However, the reliability of our study comes at a cost, which is that it only detects the strongest positive selection signals.

As in Akey (2009), we also found that the bulk of studies detect regions uniquely even considering the increased number of studies exposed in Table 1.2. Nonehteless, the number of regions detected by two or more studies increased from 21.9% to 36%. As shown in Chapter 3, the vast majority of candidate regions are cross-referenced with one publication, which ultimately means that only our analysis was able to detect previous candidate signals. It is important to note that the Akey (2009) study reports a total of 5110 regions putatively under selection, while ours is limited to a total of 2859. As mentioned above, our study sets a cutoff of $5 \cdot 10^{-4}$, which is probably the most astringent cutoff ever used in such a study. This value is usually set at around 99.9% percentile of the distribution, as we can see in one of the most recent studies of this nature, which explores selection signals on the X chromosome (Villegas-Mirón et al., 2021). As we have already stressed throughout this section, our chosen cutoff value has somewhat constricted the scope of our study and served to incorporate a large number of false negatives, which have prevented us from detecting some of the most classical examples. Nevertheless, considering the purpose of our study and the increased overlap between regions already described, PopHumanScan represents a first step in the field towards a system in which we can maintain updated selection signals across the genome, regardless of methodologies and datasets.

The information deposited in the database for different summary statistics shows consistent results regarding previous analysis not only in the overall number of regions. As extensively described in Chapter 3, we maintained information concerning populations, metapopulations, and signal types in creating the database. In addition, we performed comprehensive profiling, including information regarding functional information (especially regarding regulatory information), evolutionary conservation, as well as archaic introgression. Thus, in line with Johnson and Voight (2018), we found a clear overlap of signals in the regions of the candidate regions at the metapopulation level. As stated in Chapter 3, this percentage is higher for candidate regions showing both LD and SFS signatures (52.7%). It is 33.6% for candidate regions showing only LD signatures and 27.1% for candidate regions showing only SFS signatures. These results would indicate that the statistics we used in our GWS test different features of the region's genetic variability and are mainly complementary. Similarly, we find that most candidate regions overlap with various regulatory elements, such as TFBS, cis-regulatory modules, or enhancers (90%), results which are consistent with observations made by Enard et al. (2014), or more recently by Villegas-Mirón et al. (2021) in the X-chromosome analysis.

Thus, we can conclude here that:

1. It was necessary to perform another GWS of positive selection over the 1000GP phase III.
2. The outlier approach has a high exploratory power, especially considering the thorough treatment of the data, as well as the statistical rigor to pinpoint the strongest positive selection signals.
3. The set of summary statistics capturing different adaptive events allows us to represent the wide range of previously described signals of positive selection

Nevertheless, we would like to highlight what we consider to be the two major weaknesses of the analyses performed in Chapter 3. Firstly, not including the simulations is undoubtedly a shortcoming, even considering the arguments exposed, the selected cutoff in the empirical distributions was arbitrary. Despite the controversy, coalescence simulations have become one of the main tools in this type of analysis. Forward-in-time simulators, or approaches such as the one presented by Wang et al. (2021), are likely to be able to recreate the patterns of diversity, recombination and LD observed in human populations. Therefore, we could establish more accurate thresholds in combination with more complex selection models, demography, and recombination.

Secondly, throughout this section, we have emphasized that we have provided a catalog that collects and typifies the strongest signals of positive selection in the 1000GP data. However, no proactive search for selection signals on standing variation (soft sweeps) has been performed in this catalog. Certainly, as we explained in the Introduction, Garud et al. (2015) and Garud et al. (2021) describe extensively how the strongest iHS signals overlap with the H12 approach (the primary approach to detect soft sweeps to date). However, we believe that the non-inclusion of a statistic that implicitly detects soft sweeps, such as H12 or $\mathrm{nS_L}$, limits the overall catalog presented here.

7.2 MKT on human lineage

Three of out the four chapters of this thesis are based on the detection of recurrent positive selection, and centered around the MKT.

The detailed review at the gene level of the MKT arises from the lack of results in Chapter 3, in which only 21 human genes were detected under positive selection. These results showed a lack of agreement both in number and signals concerning the articles that map the recurrent positive selection events through the MKT or similar methodologies such as Bustamante et al. (2005) (304 genes under positive selection), Nielsen (2005) (top 50 genes showing evidence of positive selection), Arbiza et al. (2006) (104 genes under positive selection) or Gayà-Vidal and Albà (2014) (241 and 24 genes under positive selection considering Branch-site and MK-test). Likewise, we do not find any functional enrichment applying Gene Ontology analyzes (Bustamante et al., 2005; Nielsen, 2005; Arbiza et al., 2006), although this could simply occur due to the low number of genes analyzed. The differences in the GO analysis and in the number of genes detected under positive selection when comparing PopHumanScan with previous studies can be accounted for by: i) the nature of the data used in our study; ii) the MKT correction used.

We have studied the effect of the sample size on the number of recorded SNPs within the 1000GP phase III data. With this purpose, we performed a resampling analysis. Figure 7.1 shows the number of non-synonymous and synonymous SNPs as a function of the number of individuals sampled. Similar to the extrapolation performed by Gravel and National Heart (2014), we observed that the variable number of SNPs follows an exponential distribution. Considering the results of Gravel and National Heart (2014) and our analysis, we observe that the sample size at the population level shown in Chapter 3, as well as studies like Bustamante et al. (2005)

or Nielsen (2005), barely comprise 19% of the total non-synonymous/synonymous polymorphism described in the 1000GP data. Therefore, we followed Uricchio et al. (2019) and performed the analysis in Chapters 4 and 5 using the complete African lineage data from the 1000GP, instead of specific populations. In this way, we increased the sample size from 85 individuals (average number of individuals per population) to 661 individuals (total number of African individuals). A sample size from 85 to 661 individuals would increase the detection of the total polymorphism to 51%. Considering the level of polymorphism in humans, the sample size could limit the statistical power when performing the MKT at the gene level. The analysis presented in Chapter 4 (Table 4.4) using data from African lineage showed an increase in the number of genes putatively evolving under positive selection, considering MKT and eMKT (from 21 to 72 and 66, respectively). Nonetheless, since eMKT cannot deal with slightly deleterious (SDM) as discussed in Chapter 4, we decided to develop a new approach to MKT specially designed for gene-by-gene analysis.

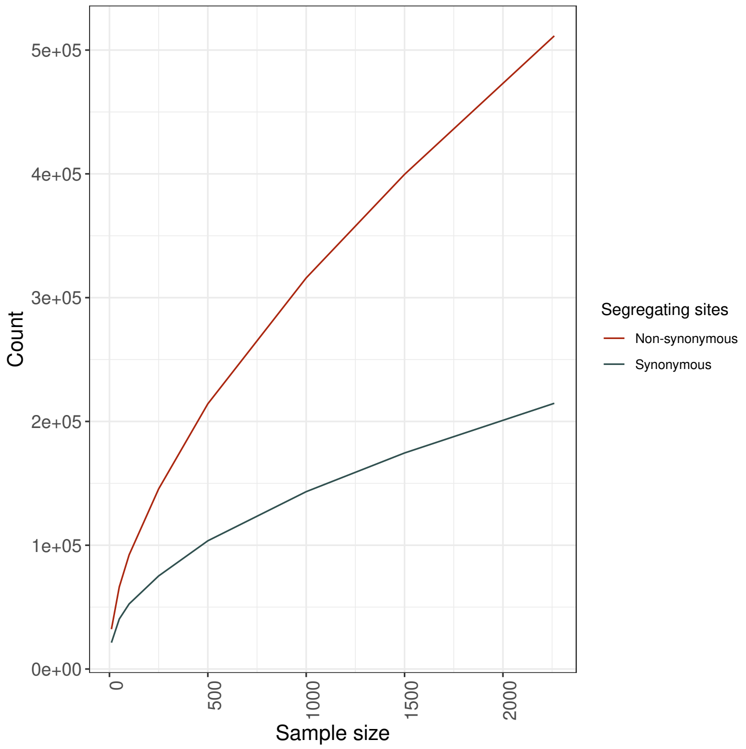


Figure 7.1: Number of non-synonymous/synonymous sites at 1000GP data regarding the the sample size.

As explained in depth in Chapter 4, despite the numerous corrections proposed for the MKT, the vast majority of them overcome its limitations by using genomic information or large gene pools (Eyre-Walker and Keightley, 2009; Messer and Petrov, 2013b; Galtier, 2016; Tataru et al., 2017; Uricchio et al., 2019). Only the original test, the correction by Fay et al. (2001) (fwwMKT) and the correction by Mackay et al. (2012) (eMKT) are designed to perform the analyses at the gene-by-gene level (i.e. allowing to test the statistical significance for each gene independently). In order to fulfill our objectives, we discarded the original MKT due to its limitations regarding SDM, especially relevant in humans considering the shape of the DFE inferred from the synonymous and non-synonymous polymorphism (Boyko et al., 2008; Racimo and Schraiber, 2014). The extensive exploration of MKT methodologies in Chapter 4 revealed the lack of results, the limitations of fwwMKT and eMKT, and the necessity for a new MKT correction at the gene level. As shown in Table 4.4, the number of genes under positive selection detected by both eMKT and fwwMKT is reduced compared to the original MKT. Considering fwwMKT, the obvious explanation for such a trend is the loss of data associated with the methodology, which makes it only capable of detecting the strongest signals. Extending the examples presented in Chapter 4, when applying the original MKT vs the fwwMKT over a pool of all protein-coding genes in Drosophila and humans, we found that the non-synonymous polymorphism is reduced by 86% and 92%, and the synonymous polymorphism is reduced by 73% and 87%, respectively. Because of the amount of data excluded by applying fwwMKT, we considered that it is not suitable for analyzing human gene sequence data.

On the other hand, as shown in Chapter 4, the average data loss when applying the eMKT is minimal ($\approx 5\%$), so we would expect that the number of genes found to be evolving under positive selection will increase, considering the amount of data used and the characteristics of the correction. Nonetheless, we observed two anti-intuitive results considering the eMKT: first, from simulations, we noted that the α estimate decreases when the frequency cutoff is increased; second, the eMKT detected fewer genes under positive selection than the original MKT. As a result, the eMKT poorly improves the results of the original MKT. Conceptually, the impMKT and the eMKT follow the same principles: 1) to maintain P_S information, since these sites are considered neutral; and 2) to remove the proportion of SDM segregating at P_N and performing the MKT using neutral or effectively neutral polymorphism. In this way, α is summarized similarly for both extensions. We can generalize both extensions as follows:

$$\alpha = 1 - \frac{D_S}{D_N} \frac{P_{effectively\ neutral}}{P_S} \tag{7.1}$$

Hence, the differences of both estimates differ in the calculation of the effectively neutral polymorphism. The eMKT assumes that a continuous form of the DFE does not drive the selected alleles. Instead, it considers a fixed proportion of nearly neutral variants given the expected proportion of synonymous polymorphisms to estimate SDM. This becomes unrealistic depending on the underlying DFE and leads to the cutoff used to be detrimental. Attending to the definition of $P_{effectively\ neutral}$ proposed by Mackay et al. (2012), the expected proportion of neutral polymorphism below the frequency threshold is supposed to be determining the fraction of effectively neutral sites:

$$P_{effectively\ neutral} = P_N \cdot f_{neutral<5\%} = P_N \cdot \frac{P_{S(j<5\%)}}{P_S} \tag{7.2}$$

nonetheless, considering that $P_{S(j<5\%)}/P_S$ converges to 1 as the frequency cutoff increases, $P_{effectively\ neutral}$ will be equal to P_N finally providing similar estimation to the original MKT.

In this way, in Chapters 3 and 4 we determined that eMKT underestimates the signals of recurrent positive selection. Most of the drawbacks regarding MKT analysis are discussed in Chapters 4 and Chapter 5, including the impact of slightly deleterious and beneficial mutation, DFE shape, or pooled datasets. Nonetheless, a few more aspects should be considered in MKT analysis that will be exposed in the following sections. Finally, we are planning to add in the PopHumanScan catalog the new signals detected through the new proposed approach, the impMKT, for gene-by-gene analyses.

7.2.1 Folded SFS vs unfolded SFS

We note that the eMKT and fwwMKT were developed using the folded site frequency spectrum (fSFS), while the exploration presented in Chapter 4 was performed using the unfolded site frequency spectrum (uSFS). The uSFS provides more evolutionary information, and, as shown in Chapter 4, such information can be used to extend the imputation for an excess of high-frequency alleles. Such extension can be especially relevant to study genes located in regions subjected to high levels of background selection (BGS) or low recombination, where Hill-Robertson interference (HRi) can be significant. Thus, HRi and α estimations can be corrected by adding the new imputed alleles to the non-synonymous divergence count. Nonetheless, uSFS requires determined ancestral alleles to be precisely estimated. Unlike the fSFS, which can be observed directly from the polymorphism data, the inference of ancestral states requires genetic data for outgroup species and the application of maximum parsimony methods. The misattribution of ancestral alleles can also affect α estimation. On one hand, an excess of high-frequency alleles can be attributed to hitchhiking with linked selected substitutions or weak adaptation, which finally can affect ML methods that infer the DFE by over-estimating the role of positive selection. On the other hand, an excess of high-frequency allele will affect the asymptotic fit in aMKT, under-estimating α. To date, the method proposed by Keightley et al. (2016) is the most sophisticated approximation to estimate uSFS. The approach was extended by Keightley and Jackson (2018) input information of more than two outgroups species, phylogenetic tree topology, and reviewing multiple nucleotide substitution models. Because having one or more outgroups is not always possible, we explored the impMKT using fSFS instead of uSFS, proposing it as an alternative to eMKT and fwwMKT when uSFS is not available.

The fSFS analysis causes a slight decrease in the mean estimates of α (see 7.2). In addition, such a decrease is more pronounced when the frequency cutoff is increased. It should be noticed that by applying the frequency cutoff on the fSFS, both low-frequency and high-frequency derived alleles are removed from the analyses, which reduces the data to estimate α and finally the statistical power. The same trend occurs when using the fSFS in gene-by-gene analyses for both Drosophila and humans, decreasing the number of genes found to evolve under positive selection by 25% and 44%, respectively (see Table 7.1).

Nevertheless, using the fSFS and focusing only on the central part of the frequency spectrum can be especially interesting in two cases. First, it is a better choice in cases of mispolarization, a situation which would add an additional bias to SDM. The cutoff will potentially eliminate fictitious (due to mispolarization) derivate alleles at a high frequency that would deviate the ratio $P_{N(j>15\%)}/P_{S(j>15\%)}$ used in the imputation. Second, the cutoff will eliminate the accumulation of SDMs at high frequencies due to interference between positively selected and slightly deleterious alleles.

Table 7.1: Number of detected genes under positive selection when using the fSFS.

		uSFS		fSFS	
Population	Set	α	N	α	N
ZI	Analyzable	-0.032 ± (1.664)	7588	-0.032 ± (1.664)	5780
ZI	Negative	-4.698 ± (4.888)	339	-4.698 ± (4.888)	1690
ZI	Positive	0.775 ± (0.121)	2244	0.775 ± (0.121)	318
AFR	Analyzable	-0.679 ± (2.21)	3230	-0.679 ± (2.21)	1756
AFR	Negative	-5.375 ± (4.676)	244	-5.375 ± (4.676)	115
AFR	Positive	0.759 ± (0.121)	205	0.759 ± (0.121)	140

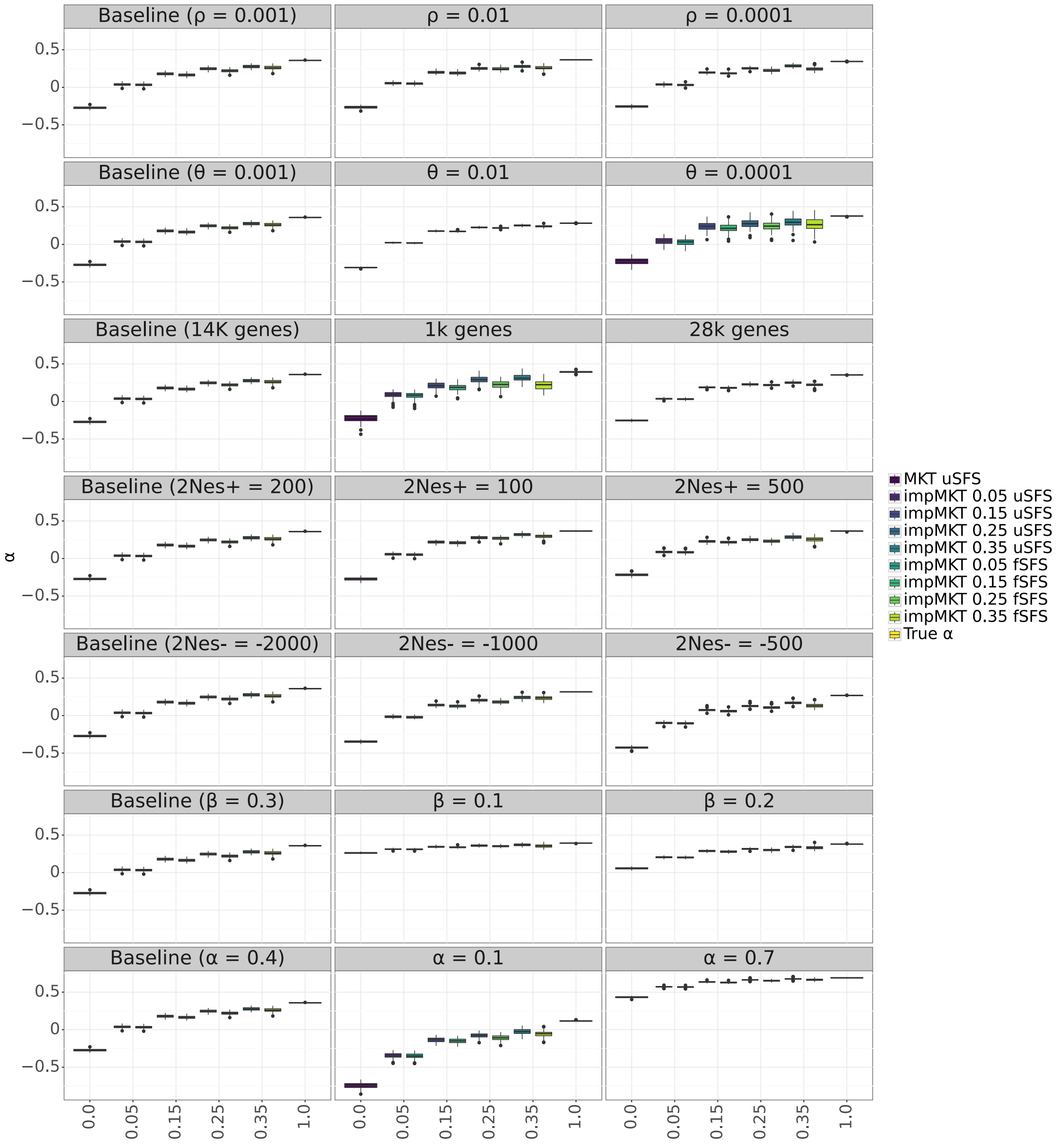


Figure 7.2: Replicas of the analysis performed in Chapter 4. We used the fSFS and the uSFS to test the impMKT.

7.2.2 Chosen outgroup

How the estimate of α (the adaptive evolutionary rate) can be affected by the chosen outgroup is extensively discussed in Keightley and Eyre-Walker (2012). These estimates can be essentially affected when the divergence time between two species is short and the rates of adaptive evolution is high, and the bias would depend on the strength of adaptation, the true value of α and the DFE. Despite the correction stated in Keightley and Eyre-Walker (2012), the authors do not recommend outgroup species whose branch lengths to the common ancestor is lesser than $10N_e$ generations (Keightley and Eyre-Walker, 2012). Thus, in those analyses where this recommendation is ignored, two things could happen: i) if $\alpha \rightarrow 0$, the estimation can be potentially highly overestimated; ii) if $\alpha > 0$, the estimation can potentially be underestimated, particularly if few SDMs contribute to nucleotide divergence. In addition, advantageous mutations reach fixation more quickly, depending on their selection coefficients, than the neutral ones, increasing also α in the short-branch estimates (Keightley and Eyre-Walker, 2012).

The bias introduced in these estimates may be mainly due to three reasons. First, a misattribution of the polymorphism to divergence. Since divergence is usually estimated using one focal outgroup, some differences can be attributed to polymorphism but not substitutions (Keightley and Eyre-Walker, 2012), increasing the divergence count and finally overestimating α. Second, the ancestral polymorphism that contributes to divergence. For example, if a slightly deleterious mutation is polymorphic at the time of the divergence between the two species, it may be lost in one lineage but remain polymorphic in the other. Third, the fixation of this mutation will be affected by the selection force itself, which may not be strong enough to eliminate it, and the demographic history of both species. Thus, reducing the population's effective size in one of the lineages could determine that the effect of drift overpowers the force of selection (as we explained in Section 1.2.3), finally contributing to the divergence. Consequently, there will be a greater contribution of SDM to the divergence than expected, which leads to an overestimation of α (Keightley and Eyre-Walker, 2012).

There are several ways to overcome the bias produced by these three situations, considering that the correction described by Keightley and Eyre-Walker (2012) does not resolve this bias if the divergence time is short. On the one hand, using nucleotide variation catalogs of the outgroup species would reduce the bias associated with polymorphisms contributing to apparent divergence or mispolarization errors. However, polymorphism may still appear to be fixed in a sample of sequences. An example is the use of the catalogs described in Signor et al. (2018) or de Manuel et al. (2016). The approach proposed by de Manuel et al. (2016) is interesting, since they provided the *Pan troglodyte* nucleotide information mapped to the human reference genome. On the other hand, the DFE inference methods proposed by Galtier (2016) and Tataru et al. (2017) are capable of inferring α only from polymorphism data, although obviously, the power is increased if we use divergence data Tataru et al. (2017). These approximations suppose an alternative to estimate α that is independent of an external group.

A third possible option is to incorporate information from an ancestral species close to the study population in addition to the outgroup. An example would be to incorporate ancient DNA data into human analysis, an approximation that has not yet been explored in MKT methodologies or DFE inference. In recent years, aDNA sequencing has increased exponentially, especially in hominins. aDNA polymorphism data can provide information on the direction and strength of selection from the split with the outgroup, considering whether the mutation continues to be polymorphic, derived, ancestral, or fixed in this additional population. Future studies are required to see how to incorporate this aDNA polymorphism into MKT approaches.

7.2.3 Neutral evolving sites and the null hypothesis

All the MKT analyses carried out in this thesis have followed the initial proposal of McDonald and Kreitman (1991), in which synonymous sites are considered neutral mainly due to the degeneracy of the genetic code. The vast majority of the studies use four-fold degenerate sites as a proxy for the mutation rate. However, this assumption could affect the measure of the adaptive evolution rate under the evidence that synonymous mutations are also subject to selection. Lawrie et al. (2013) is one of the clearest examples of how the detection of recurrent positive selection can be overestimated. Lawrie et al. (2013) shows that up to 22% of synonymous sites can be under the effect of strong purifying selection. Ultimately, the constraint of these sites will lead to assume that the number of non-synonymous substitutions is higher than the expected neutral hypothesis, leading to a misguiding interpretation of the d_N/d_S ratio.

One possible solution is to use short introns as a neutral proxy instead of synonymous sites. This approach has been commonly used in *D. melanogaster* (Keightley and Eyre-Walker, 2007; Eyre-Walker and Keightley, 2009; Castellano et al., 2016), thanks to the studies of Parsch et al. (2010) and Halligan and Keightley (2006), where the evolutionary characteristics and forces in the intronic structure of Drosophila are widely described. On the one hand, Halligan and Keightley (2006) found that the most rapidly evolving intron sites are around bases 8-30 of the 65 bp introns. On the other hand, Parsch et al. (2010) shows that the high divergence observed in short introns is not due to adaptive evolution. Both findings suggest that short intron sequences may be the most appropriate proxy for the neutral mutation rate. However, the use of short introns is not always possible. First, while orthologous protein sequences are easy to identify, the intronic content of genes can differ significantly. Second, one would expect that the splicing sequences of the introns were also under the action of natural selection.

During the last few years, studies like Frigola et al. (2017) in humans or Monroe et al. (2022) in the plant *Arabidopsis thaliana* have challenged the assumption of randomness of mutations. Such studies state that the genomic mutation rate may depend on confounding factors such as epigenomic landscape, accessibility of DNA repair machinery, or sequence function. This proposal not only compromises the use of introns as a neutrality proxy, explained here, but it challenges a major tenet of population genetics. In this way, mutation rate could partially account for the role attributed to natural selection, since a fraction of the constraint of the functional elements could be due to the non-randomness of mutations and not exclusively to purifying selection. Rodriguez-Galindo et al. (2020) proved that the human mutation rate in introns and exons is unaffected in the germline considering both sequence context and multiple histone marks. Nonetheless, Monroe et al. (2022) have found in Arabidopsis lower nucleotide diversity around genes, lower mutation rates in coding regions, lower mutation rates in important functional genes, as well as associations between mutation rate, epigenomic landscape and DNA repair machinery (Monroe et al., 2022). Although these studies do not deny the role of natural selection as a nucleotide diversity driver, they also show a significant role of the epigenomic landscape and the mutation rate in the explanation of patterns of genome diversity, or difference in the DFE. Extending the computational approach proposed by Barroso et al. (2019) to disentangle the role of demography and recombination on genome-wide, Barroso and Dutheil (2021) incorporate the mutation rate and found that the mutation landscape is the major driver of the distribution of diversity in *D. melanogaster*. Barroso and Dutheil (2021) deeply argue about the role of mutation rate and the incorporation of the mutational landscape into the null model, and more importantly, their debate is not focused on the assumption of molecular population genetics, but rather in the next questions: *under what conditions can the shape of the mutation landscape itself be selected for? (...) how conserved is the mutation landscape across species?* The study proposed by Frigola et al. (2017) and Monroe et al. (2022) pinpointed a new role of the mutation rate and the epigenomic landscape, but further studies as required to understand the main drivers of adaptation. Nonetheless, as discussed in Section 7.1.2, we need a null model that incorporates a more complex mutational landscape to better infer the role of mutational input in population genomics (Johri et al., 2020; DeWitt et al., 2021; Barroso and Dutheil, 2021; Johri et al., 2021).

7.2.4 Linked selection and the MKT

The estimates of the adaptive evolution rate have been carried out mainly through extensions of the Poisson Random Field (PRF). As we explained in Section 1.3.4, studies such as Boyko et al. (2008), Eyre-Walker and Keightley (2009), Racimo and Schraiber (2014) or, more recently, Zhen et al. (2021), assume models in which sites segregate independently. Uricchio et al. (2019) search for the first time if the low levels of adaptation in humans compared to other species may be due to the role of linked selection. BGS can lead to patterns of fixation not considered by strict neutrality due to linkage with deleterious segregating alleles (Charlesworth et al., 1993; Charlesworth, 1994; Hudson and Kaplan, 1995; Nordborg et al., 1996; Pouyet et al., 2018). Hence, a method that interrogates linked selection is crucial to understanding the shape of genetic diversity and the adaptation rate

Uricchio et al. (2019) developed ABC-MK, a method that exploits the impact of BGS on the fixation rate. BGS can be summarized by the B value (McVicker et al., 2009) and varies across the genome. ABC-MK interrogates α as a function of BGS, inferring not only the rate of adaptation but also the strength of beneficial alleles (Uricchio et al., 2019). As shown in Chapter 3, we developed an extension of the previous ABC-MK. In our approach, we avoid expensive forward-in-time simulations benefitting from the empirically observed data through a novel sampling process and extend the estimation using expected BGS values. The estimation of the B value is not mandatory to perform the analyses. The B value estimation requires information about recombination rates, which is not available for all the species. Our primary goal in developing the ABC-MK extension was to circumvent the computational resources required to execute the original ABC-MK on the 1000GP data, which requires running the analyses in an HPC.

We showed that our results are consistent with the results presented in Uricchio et al. (2019) regarding equilibrium demographic model and recent human demography events. Nonetheless, since our extension is based on the exact diffusion approximation to model the SFS, while accounting for BGS presented in Uricchio et al. (2019), we do not test the effect of genetic draft in our approach (see Section 1.2.4). The impact of linked positive selection on the frequency trajectories of linked alleles also drives systematic variation in diversity genome-wide. Uricchio et al. (2019) performed simulations to test ABC-MK sensitivity to hitchhiking, scaling the strength and rate of positive selection much higher than expected in humans. The simulation results showed that hitchhiking can decrease α estimations. The main explanation is the increase in slightly deleterious fixations, as well as the interference between strongly-beneficial alleles.

Nonetheless, Uricchio et al. (2019) tested the effect of recurrent hitch-hicking at the previous ABC-MK estimations showing robust estimations. Thus, although the basis and mechanism of BGS and draft are different, and their expected diversity patterns highly differ (as exposed in Schrider (2020)), both processes should result in reducing the expected π/π_0 ratio due to linkage, therefore affecting allele frequency trajectories and fixation. Although we did not simulate genetic draft in Chapter 5, our implementation can perform similar estimations in cases of strong recurrent hitch-hicking because: i) Uricchio et al. (2019) demonstrated the negligible effect on recurrent hitch-hiking on α inference; ii) the new ABC-MK implementation showed similar accuracy and precision than the previous implementation; iii) the new ABC-MK implementation can simulate virtually any expected reduction in diversity due to the performance. In addition, as exposed Uricchio et al. (2019), only regions where the genome experiences strong recurrent sweeps can slightly affect α inference. Nonetheless, since ABC-MK is a software designed to input a pool of genes, it is unlikely that strong recurrent hitch-hicking will affect all genes equally. Thus, such an effect probably should be diluted in real, empirical data.

7.2.5 RNA-VIPs vs. DNA-VIPs

In Chapter 5, we showed that RNA Viral Interacting Proteins (VIPs) exhibit stronger adaptation rates than DNA-VIPs. There are several examples providing evidence that the host-pathogen race imposed by the virus is an essential driver of human genome adaptation at different time scales or adaptive regimes (Deschamps et al., 2016; Enard et al., 2016; Racimo et al., 2017; Enard and Petrov, 2018; Uricchio et al., 2019; Castellano et al., 2019b). We extended the analysis by Uricchio et al. (2019), differentiating between RNA and DNA-VIPs. As a result, we provide evidence that RNA viruses are an essential driver of human adaptation (Enard and Petrov, 2018, 2020).

Following Uricchio et al. (2019), we note that the effect of BGS cannot explain the higher rate of adaptation of the RNA-VIPs because the DNA-VIP pool experiences a slightly stronger BGS effect. Although we stated that these results might reflect the role of RNA-viruses in human adaptation and zoonosis frequencies, a few points should be considered. First, it should be noted that the analysis includes 3,471 RNA-VIPs and only 1,258 DNA-VIPs. Second, although our model considers different levels of BGS and the average recombination rate, measurements could be affected provided that other highly-correlated characteristics with these protein sets promote adaptation. As aforementioned in Chapter 4, the approach proposed by Huang (2021), based on multiple regression models, can jointly evaluate the effects of these correlated genomic features on α estimation, while showing the primary feature modulating adaptation. A posterior analysis considering other important features involved in adaptation (such as mutation rate, relative solvent accessibility or Protein-Protein Interaction) (Moutinho et al., 2019b; Huang, 2021), in addition to BGS and recombination rate, can determine whether RNA-VIPs experienced a higher rate of adaptation. Third, the higher rate of adaptation in RNA-VIPs could be directly related to the size of the viral genome and its mutation rate. A possible solution would be to normalize the adaptation rate taking into account the genome size. However, RNA and DNA-VIPs are an extension of the annotations presented in Enard and Petrov (2020) and Souilmi et al. (2021). Although the RNA viruses are smaller than DNA viruses, we have found that the size of DNA viruses for which VIPs were annotated is much more variable. An alternative option, not involving highly heterogeneous genome size, would be to check if levels of variation, together with levels of BGS, can be explained by the mutation rate of the human genome, which ultimately could also provide information on the mutation rate of viral genomes and the host-pathogen interaction. As aforementioned in Section 7.2.3, the method developed by Barroso and Dutheil (2021) states that mutation rate variation mainly explains the levels of nucleotide diversity in *D. melanogaster*. In future analysis, the mutation rate estimation needs to be incorporated both in our analysis and in the approach by Huang (2021) to unravel if the adaptation levels in RNA and DNA-VIPs can be explained by the mutation rate or not.

Chapter 8

Conclusions

1. The outlier approach combining eight different statistics to detect candidate regions under selection in 22 non-admixed human populations has been able to locate distinct signatures in 2859 regions that stand out from the background genomic variability.

2. The combination of statistics, including abnormally long haplotypes, SFS deviations and, for the first time in conjunction with the excess of non-synonymous substitutions between our species and chimpanzees, manifest the footprints of selective sweeps at different historical ages, or recurrent selection that has been taking place during the last millions of years.

3. PopHumanScan online database facilitates the thorough analysis of candidate regions under selection in the human genome and the incorporation wiht new data from the scientific community, while automatically putting together the evidence of selection with structural and functional annotations of the regions and cross-references to previously published articles.

4. Unlike the rest of positive selection GWS, our approach, although limited to the strongest positive selection signals, detects for the first time 70% of the signals from previous studies, also describing for the first time a total of 873 new regions.

5. The impMKT is an straightforward, unbiased and efficient approach especially designed to test recurrent positive selection at the gene level.

6. We reviewed the number of genes under positive selection using impMKT in the human lineage using 1000GP data, increasing the signals exposed in PopHumanScan from 21 to 205.

7. We developed three statistics measuring the components of the DFE using the imputation of SDM, as well as a nuisance parameter testing for population contraction, mispolarization or sequencing error.

8. The new ABC-MKT version can reproduce previous analyses in 1000GP data, increasing the performance by several orders of magnitude, without the necessity of an HPC

9. We found stronger signals of positive selection in RNA-VIPs than DNA-VIPs, providing additional support to the hypothesis that RNA viruses are important drivers of human adaptation.

10. We developed a user-friendly web server that tests recurrent positive selection in DGN and 1000GP data. The iMKT performs analysis at the gene level, genome level, and custom pools of genes, benefiting the heuristic MKT extensions explored in this thesis.

Bibliography

Adrion, J.R., Cole, C.B., Dukler, N., et al. A community-maintained standard library of population genetic models. *eLife*, 9:e54967, 2020a. ISSN 2050-084X. doi: 10.7554/eLife.54967.

Adrion, J.R., Galloway, J.G., and Kern, A.D. Predicting the Landscape of Recombination Using Deep Learning. *Molecular Biology and Evolution*, 37(6):1790–1808, 2020b. ISSN 0737-4038. doi: 10.1093/molbev/msaa038.

Akashi, H. Inferring the Fitness Effects of DNA Mutations From Polymorphism and Divergence Data: Statistical Power to Detect Directional Selection Under Stationarity and Free Recombination. *Genetics*, 151(1):221–238, 1999. ISSN 1943-2631. doi: 10.1093/genetics/151.1.221.

Akey, J.M. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research*, 19(5):711–722, 2009. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.086652.108.

Akey, J.M., Zhang, G., Zhang, K., et al. Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research*, 12(12):1805–1814, 2002. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.631202.

Albrechtsen, A., Moltke, I., and Nielsen, R. Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. *Genetics*, 186(1):295–308, 2010. ISSN 1943-2631. doi: 10.1534/genetics.110.113977.

Ali, M., Liu, X., Pillai, E.N., et al. Characterizing the genetic differences between two distinct migrant groups from Indo-European and Dravidian speaking populations in India. *BMC Genetics*, 15(1):86, 2014. ISSN 1471-2156. doi: 10.1186/1471-2156-15-86.

Allison, A.C. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *Br Med J*, 1(4857):290–294, 1954. ISSN 0007-1447, 1468-5833. doi: 10.1136/bmj.1.4857.290. Publisher: British Medical Journal Publishing Group Section: Article.

Alonso-Blanco, C., Andrade, J., Becker, C., et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, 166(2):481–491, 2016. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2016.05.063. Publisher: Elsevier.

Altshuler, D., Donnelly, P., and The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005. ISSN 1476-4687. doi: 10.1038/nature04226.

Altshuler, D.M., Gibbs, R.A., Peltonen, L., et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010. ISSN 1476-4687. doi: 10.1038/nature09298.

Amato, R., Pinelli, M., Monticelli, A., et al. Genome-Wide Scan for Signatures of Human Population Differentiation and Their Relationship with Natural Selection, Functional Pathways and Diseases. *PLOS ONE*, 4(11):e7927, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0007927. Publisher: Public Library of Science.

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1):D789–D798, 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1205.

Andersen, K.G., Shylakhter, I., Tabrizi, S., et al. Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1590):868–877, 2012. doi: 10.1098/rstb.2011.0299. Publisher: Royal Society.

Andrés, A.M., Hubisz, M.J., Indap, A., et al. Targets of Balancing Selection in the Human Genome. *Molecular Biology and Evolution*, 26(12):2755–2764, 2009. ISSN 0737-4038. doi: 10.1093/molbev/msp190.

Arbiza, L., Dopazo, J., and Dopazo, H. Positive Selection, Relaxation, and Acceleration in the Evolution of the Human and Chimp Genome. *PLOS Computational Biology*, 2(4):e38, 2006. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0020038. Publisher: Public Library of Science.

Auton, A., Abecasis, G.R., Altshuler, D.M., et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. ISSN 1476-4687. doi: 10.1038/nature15393.

Bailey, J.A., Gu, Z., Clark, R.A., et al. Recent Segmental Duplications in the Human Genome. *Science*, 297(5583):1003–1007, 2002. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1072047.

Balloux, F. and Lehmann, L. Substitution Rates at Neutral Genes Depend on Population Size Under Fluctuating Demography and Overlapping Generations. *Evolution*, 66(2):605–611, 2012. ISSN 1558-5646. doi: 10.1111/j.1558-5646.2011.01458.x.

Bao, W., Kojima, K.K., and Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):11, 2015. ISSN 1759-8753. doi: 10.1186/s13100-015-0041-9.

Barbadilla, A., King, L.M., and Lewontin, R.C. What does electrophoretic variation tell us about protein variation? *Molecular Biology and Evolution*, 13(2):427–432, 1996. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a025602.

Barroso, G.V. and Dutheil, J.Y. Mutation rate variation shapes genome-wide diversity in Drosophila melanogaster. Technical report, 2021. doi: 10.1101/2021.09.16.460667. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Barroso, G.V., Puzović, N., and Dutheil, J.Y. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, 15(11):e1008449, 2019. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008449.

Barton, N.H. Linkage and the limits to natural selection. *Genetics*, 140(2):821–841, 1995. ISSN 0016-6731, 1943-2631.

Beall, C.M., Cavalleri, G.L., Deng, L., et al. Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*, 107(25):11459–11464, 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1002443107. Publisher: National Academy of Sciences Section: Biological Sciences.

Begun, D.J. and Aquadro, C.F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature*, 356(6369):519–520, 1992. ISSN 1476-4687. doi: 10.1038/356519a0.

Begun, D.J., Holloway, A.K., Stevens, K., et al. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLOS Biology*, 5(11):e310, 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050310.

Beja-Pereira, A., Luikart, G., England, P.R., et al. Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics*, 35(4):311–313, 2003. ISSN 1546-1718. doi: 10.1038/ng1263.

Beleza, S., Johnson, N.A., Candille, S.I., et al. Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. *PLOS Genetics*, 9(3):e1003372, 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003372. Publisher: Public Library of Science.

Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 1999. ISSN 0305-1048. doi: 10.1093/nar/27.2.573.

Bergman, C.M. and Haddrill, P.R. Strain-specific and pooled genome sequences for populations of *Drosophila melanogaster* from three continents. Technical Report 4:31, F1000Research, 2015. doi: 10.12688/f1000research.6090.1. Type: article.

Bergström, A., McCarthy, S.A., Hui, R., et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484), 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aay5012. Publisher: American Association for the Advancement of Science Section: Research Article.

Bersaglieri, T., Sabeti, P.C., Patterson, N., et al. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*, 74(6):1111–1120, 2004. ISSN 0002-9297, 1537-6605. doi: 10.1086/421051.

Bhatia, G., Patterson, N., Pasaniuc, B., et al. Genome-wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. *The American Journal of Human Genetics*, 89(3):368–381, 2011. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2011.07.025. Publisher: Elsevier.

Bhérer, C., Campbell, C.L., and Auton, A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature Communications*, 8(1):14994, 2017. ISSN 2041-1723. doi: 10.1038/ncomms14994.

Bierne, N. and Eyre-Walker, A. The Genomic Rate of Adaptive Amino Acid Substitution in Drosophila. *Molecular Biology and Evolution*, 21(7):1350–1360, 2004. ISSN 0737-4038. doi: 10.1093/molbev/msh134.

Bigham, A., Bauchet, M., Pinto, D., et al. Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. *PLOS Genetics*, 6(9):e1001116, 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001116. Publisher: Public Library of Science.

Booker, T.R. Inferring Parameters of the Distribution of Fitness Effects of New Mutations When Beneficial Mutations Are Strongly Advantageous and Rare. *G3: Genes, Genomes, Genetics*, 10(7):2317–2326, 2020. ISSN 2160-1836. doi: 10.1534/g3.120.401052.

Booker, T.R. and Keightley, P.D. Understanding the Factors That Shape Patterns of Nucleotide Diversity in the House Mouse Genome. *Molecular Biology and Evolution*, 35(12):2971–2988, 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy188.

Booker, T.R., Yeaman, S., and Whitlock, M.C. Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Molecular Ecology*, 29(22):4274–4279, 2020. ISSN 1365-294X. doi: https://doi.org/10.1111/mec.15501.

Boyko, A.R., Williamson, S.H., Indap, A.R., et al. Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLOS Genetics*, 4(5):e1000083, 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000083.

Brandt, D.Y.C., Wei, X., Deng, Y., et al. Evaluation of methods for the inference of ancestral recombination graphs. Technical report, 2021. doi: 10.1101/2021.11.15.468686. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Brown, G.R., Hem, V., Katz, K.S., et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1):D36–D42, 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1055.

Bubb, K.L., Bovee, D., Buckley, D., et al. Scan of Human Genome Reveals No New Loci Under Ancient Balancing Selection. *Genetics*, 173(4):2165–2177, 2006. ISSN 1943-2631. doi: 10.1534/genetics.106.055715.

Buffalo, V. Why do species get a thin slice of π? Revisiting Lewontin's Paradox of Variation. *bioRxiv*, page 2021.02.03.429633, 2021. doi: 10.1101/2021.02.03.429633.

Bustamante, C.D., Fledel-Alon, A., Williamson, S., et al. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–1157, 2005. ISSN 1476-4687. doi: 10.1038/nature04240.

Bustamante, C.D., Nielsen, R., and Hartl, D.L. A Maximum Likelihood Method for Analyzing Pseudogene Evolution: Implications for Silent Site Evolution in Humans and Rodents. *Molecular Biology and Evolution*, 19(1):110–117, 2002a. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a003975.

Bustamante, C.D., Nielsen, R., Sawyer, S.A., et al. The cost of inbreeding in Arabidopsis. *Nature*, 416(6880):531–534, 2002b. ISSN 1476-4687. doi: 10.1038/416531a. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6880 Primary_atype: Research Publisher: Nature Publishing Group.

Byrska-Bishop, M., Evani, U.S., Zhao, X., et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Technical report, 2021. doi: 10.1101/2021.02.06.430068. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Cai, Z., Camp, N.J., Cannon-Albright, L., et al. Identification of regions of positive selection using Shared Genomic Segment analysis. *European Journal of Human Genetics*, 19(6):667–671, 2011. ISSN 1476-5438. doi: 10.1038/ejhg.2010.257. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;Genetic variation;Genomic analysis Subject_term_id: computational-biology-and-bioinformatics;genetic-variation;genomic-analysis.

Campo, D., Lehmann, K., Fjeldsted, C., et al. Whole-genome sequencing of two North American Drosophila melanogaster populations reveals genetic differentiation and positive selection. *Molecular Ecology*, 22(20):5084–5097, 2013. ISSN 1365-294X. doi: 10.1111/mec.12468. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12468.

Campos, J.L. and Charlesworth, B. The Effects on Neutral Variability of Recurrent Selective Sweeps and Background Selection. *Genetics*, 212(1):287–303, 2019. ISSN 1943-2631. doi: 10.1534/genetics.119.301951.

Campos, J.L., Zhao, L., and Charlesworth, B. Estimating the parameters of background selection and selective sweeps in Drosophila in the presence of gene conversion. *Proceedings of the National Academy of Sciences*, 114(24):E4762–E4771, 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1619434114.

Cannings, C. The latent roots of certain Markov chains arising in genetics: A new approach, II. Further haploid models. *Advances in Applied Probability*, 7(2):264–282, 1975. ISSN 0001-8678, 1475-6064. doi: 10.2307/1426077. Publisher: Cambridge University Press.

Carlson, C.S., Thomas, D.J., Eberle, M.A., et al. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11):1553–1565, 2005. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.4326505. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Carneiro, M., Albert, F.W., Melo-Ferreira, J., et al. Evidence for Widespread Positive and Purifying Selection Across the European Rabbit (Oryctolagus cuniculus) Genome. *Molecular Biology and Evolution*, 29(7):1837–1849, 2012. ISSN 0737-4038. doi: 10.1093/molbev/mss025.

Casillas, S. and Barbadilla, A. PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. *Nucleic Acids Research*, 34(suppl_2):W632–W634, 2006. ISSN 0305-1048. doi: 10.1093/nar/gkl080.

Casillas, S. and Barbadilla, A. Molecular Population Genetics. *Genetics*, 205(3):1003–1035, 2017. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.116.196493.

Casillas, S., Barbadilla, A., and Bergman, C.M. Purifying Selection Maintains Highly Conserved Noncoding Sequences in Drosophila. *Molecular Biology and Evolution*, 24(10):2222–2234, 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm150.

Casillas, S., Mulet, R., Villegas-Mirón, P., et al. PopHuman: the human population genomics browser. *Nucleic Acids Research*, 46(D1):D1003–D1010, 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx943.

Casillas, S., Petit, N., and Barbadilla, A. DPDB: a database for the storage, representation and analysis of polymorphism in the Drosophila genus. *Bioinformatics*, 21(suppl_2):ii26–ii30, 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti1103.

Casper, J., Zweig, A.S., Villarreal, C., et al. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, 46(D1):D762–D769, 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1020.

Castellano, D., Coronado-Zamora, M., Campos, J.L., et al. Adaptive Evolution Is Substantially Impeded by Hill–Robertson Interference in Drosophila. *Molecular Biology and Evolution*, 33(2):442–455, 2016. ISSN 0737-4038. doi: 10.1093/molbev/msv236.

Castellano, D., James, J., and Eyre-Walker, A. Nearly Neutral Evolution across the Drosophila melanogaster Genome. *Molecular Biology and Evolution*, 35(11):2685–2694, 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy164.

Castellano, D., Macià, M.C., Tataru, P., et al. Comparison of the Full Distribution of Fitness Effects of New Amino Acid Mutations Across Great Apes. *Genetics*, 213(3):953–966, 2019a. ISSN 1943-2631. doi: 10.1534/genetics.119.302494.

Castellano, D., Uricchio, L.H., Munch, K., et al. Viruses rule over adaptation in conserved human proteins. *bioRxiv*, page 555060, 2019b. doi: 10.1101/555060.

Charlesworth, B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetics Research*, 63(3):213–227, 1994. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672300032365.

Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205, 2009. ISSN 1471-0064. doi: 10.1038/nrg2526. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 3 Primary_atype: Reviews Publisher: Nature Publishing Group.

Charlesworth, B. Molecular population genomics: a short history. *Genetics Research*, 92(5-6):397–411, 2010. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672310000522.

Charlesworth, B. and Charlesworth, D. Population genetics from 1966 to 2016. *Heredity*, 118(1):2–9, 2017. ISSN 1365-2540. doi: 10.1038/hdy.2016.55.

Charlesworth, B., Morgan, M.T., and Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993. ISSN 0016-6731, 1943-2631.

Charlesworth, J. and Eyre-Walker, A. The Rate of Adaptive Evolution in Enteric Bacteria. *Molecular Biology and Evolution*, 23(7):1348–1356, 2006. ISSN 0737-4038. doi: 10.1093/molbev/msk025.

Charlesworth, J. and Eyre-Walker, A. The McDonald–Kreitman Test and Slightly Deleterious Mutations. *Molecular Biology and Evolution*, 25(6):1007–1015, 2008. ISSN 0737-4038. doi: 10.1093/molbev/msn005.

Chen, C.H., Chuang, T.J., Liao, B.Y., et al. Scanning for the Signatures of Positive Selection for Human-Specific Insertions and Deletions. *Genome Biology and Evolution*, 1:415–419, 2009. ISSN 1759-6653. doi: 10.1093/gbe/evp041.

Chen, H., Patterson, N., and Reich, D. Population differentiation as a test for selective sweeps. *Genome Research*, 20(3):393–402, 2010. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.100545.109. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Clark, A.G., Eisen, M.B., Smith, D.R., et al. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–218, 2007. ISSN 1476-4687. doi: 10.1038/nature06341.

Clark, A.G., Glanowski, S., Nielsen, R., et al. Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios. *Science*, 302(5652):1960–1963, 2003. doi: 10.1126/science.1088821. Publisher: American Association for the Advancement of Science.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., et al. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15(11):1496–1502, 2005. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.4107905. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Clemente, F.J., Cardona, A., Inchley, C.E., et al. A Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. *The American Journal of Human Genetics*, 95(5):584–589, 2014. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2014.09.016. Publisher: Elsevier.

Colomer-Vilaplana, A., Murga-Moreno, J., Canalda-Baltrons, A., et al. PopHumanVar: an interactive application for the functional characterization and prioritization of adaptive genomic variants in humans. *Nucleic Acids Research*, page gkab925, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab925.

Colonna, V., Ayub, Q., Chen, Y., et al. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology*, 15(6):R88, 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-6-r88.

Comeron, J.M., Ratnappan, R., and Bailin, S. The Many Landscapes of Recombination in Drosophila melanogaster. *PLOS Genetics*, 8(10):e1002905, 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002905.

Consortium, T..G.P. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061, 2010. ISSN 1476-4687. doi: 10.1038/nature09534.

Consortium, T..G.P. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012. ISSN 1476-4687. doi: 10.1038/nature11632.

Consortium, T.A.g..G., Clarkson, C.S., Miles, A., et al. Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. *Genome Research*, 30(10):1533–1546, 2020. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.262790.120. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Consortium, T.I.H. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007. ISSN 1476-4687. doi: 10.1038/nature06258.

Corbett-Detig, R.B., Hartl, D.L., and Sackton, T.B. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS Biology*, 13(4):e1002112, 2015. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002112.

Currat, M., Trabuchet, G., Rees, D., et al. Molecular Analysis of the β-Globin Gene Cluster in the Niokholo Mandenka Population Reveals a Recent Origin of the βS Senegal Mutation. *The American Journal of Human Genetics*, 70(1):207–223, 2002. ISSN 0002-9297. doi: 10.1086/338304.

de Manuel, M., Kuhlwilm, M., Frandsen, P., et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354(6311):477–481, 2016. doi: 10.1126/science.aag2602. Publisher: American Association for the Advancement of Science.

DeGiorgio, M., Huber, C.D., Hubisz, M.J., et al. S weep F inder 2: increased sensitivity, robustness and flexibility. *Bioinformatics*, 32(12):1895–1897, 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw051.

DeGiorgio, M. and Szpiech, Z.A. A spatially aware likelihood test to detect sweeps from haplotype distributions. Technical report, 2021. doi: 10.1101/2021.05.12.443825. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Derrien, T., Johnson, R., Bussotti, G., et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, 22(9):1775–1789, 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.132159.111.

Deschamps, M., Laval, G., Fagny, M., et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *The American Journal of Human Genetics*, 98(1):5–21, 2016. ISSN 0002-9297. doi: 10.1016/j.ajhg.2015.11.014.

DeWitt, W.S., Harris, K.D., Ragsdale, A.P., et al. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, 118(21), 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2013798118. Publisher: National Academy of Sciences Section: Biological Sciences.

Di, C., Murga Moreno, J., Salazar-Tortosa, D.F., et al. Decreased recent adaptation at human mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants. *eLife*, 10:e69026, 2021. ISSN 2050-084X. doi: 10.7554/eLife.69026. Publisher: eLife Sciences Publications, Ltd.

Dobzhansky, T. A REVIEW OF SOME FUNDAMENTAL CONCEPTS AND PROBLEMS OF POPULATION GENETICS. *Cold Spring Harbor Symposia on Quantitative Biology*, 20(0):1–15, 1955. ISSN 0091-7451, 1943-4456. doi: 10.1101/SQB.1955.020.01.003.

Dobzhansky, T. *Genetics and the Origin of Species: Columbia Classics edition*. Columbia University Press, 1982. ISBN 978-0-231-05475-1. Pages: 364 Pages.

Dutheil, J.Y. Towards more realistic models of genomes in populations: the Markov-modulated sequentially Markov coalescent. *arXiv:2010.08359 [q-bio]*, 2020. ArXiv: 2010.08359.

Egea, R., Casillas, S., and Barbadilla, A. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, 36(suppl_2):W157–W162, 2008. ISSN 0305-1048. doi: 10.1093/nar/gkn337.

Eichstaedt, C.A., Antão, T., Pagani, L., et al. The Andean Adaptive Toolkit to Counteract High Altitude Maladaptation: Genome-Wide and Phenotypic Analysis of the Collas. *PLOS ONE*, 9(3):e93314, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0093314. Publisher: Public Library of Science.

Enard, D., Cai, L., Gwennap, C., et al. Viruses are a dominant driver of protein adaptation in mammals. *eLife*, 5:e12469, 2016. ISSN 2050-084X. doi: 10.7554/eLife.12469.

Enard, D., Depaulis, F., and Crollius, H.R. Human and Non-Human Primate Genomes Share Hotspots of Positive Selection. *PLOS Genetics*, 6(2):e1000840, 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000840. Publisher: Public Library of Science.

Enard, D., Messer, P.W., and Petrov, D.A. Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6):885–895, 2014. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.164822.113.

Enard, D. and Petrov, D.A. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell*, 175(2):360–371.e13, 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.08.034.

Enard, D. and Petrov, D.A. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1812):20190575, 2020. doi: 10.1098/rstb.2019.0575.

Evans, S.N., Shvets, Y., and Slatkin, M. Non-equilibrium theory of the allele frequency spectrum. *Theoretical Population Biology*, 71(1):109–119, 2007. ISSN 0040-5809. doi: 10.1016/j.tpb.2006.06.005.

Ewing, G. and Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq322.

Eyre-Walker, A. and Keightley, P.D. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, 26(9):2097–2108, 2009. ISSN 0737-4038. doi: 10.1093/molbev/msp119.

Eyre-Walker, A., Woolfit, M., and Phelps, T. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. *Genetics*, 173(2):891–900, 2006. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.106.057570.

Fagny, M., Patin, E., Enard, D., et al. Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets. *Molecular Biology and Evolution*, 31(7):1850–1868, 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu118.

Fan, R., Lange, K., Zhang, Y., et al. A cross-population extended haplotype-based homozygosity score test to detect positive selection in genome-wide scans. *Statistics and Its Interface*, 4(1):51–63, 2011. ISSN 1938-7997. doi: 10.4310/SII.2011.v4.n1.a6. Publisher: International Press of Boston.

Fan, S., Hansen, M.E.B., Lo, Y., et al. Going global by adapting local: A review of recent human adaptation. *Science*, 354(6308):54–59, 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaf5098.

Fay, J.C. Weighing the evidence for adaptation at the molecular level. *Trends in Genetics*, 27(9):343–349, 2011. ISSN 0168-9525. doi: 10.1016/j.tig.2011.06.003.

Fay, J.C. and Wu, C.I. Hitchhiking Under Positive Darwinian Selection. *Genetics*, 155(3):1405–1413, 2000. ISSN 0016-6731, 1943-2631.

Fay, J.C., Wyckoff, G.J., and Wu, C.I. Positive and Negative Selection on the Human Genome. *Genetics*, 158(3):1227–1234, 2001. ISSN 0016-6731, 1943-2631.

Fay, J.C., Wyckoff, G.J., and Wu, C.I. Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature*, 415(6875):1024–1026, 2002. ISSN 1476-4687. doi: 10.1038/4151024a.

Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981. ISSN 1432-1432. doi: 10.1007/BF01734359.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., et al. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution*, 31(5):1275–1291, 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu077.

Field, Y., Boyle, E.A., Telis, N., et al. Detection of human adaptation during the past 2000 years. *Science*, 354(6313):760–764, 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aag0776.

Fisher, R.A. *The genetical theory of natural selection.* The genetical theory of natural selection. Clarendon Press, Oxford, England, 1930. doi: 10.5962/bhl.title.27468. Pages: xiv, 272.

Frichot, E., Schoville, S.D., Bouchard, G., et al. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, 30(7):1687–1699, 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst063.

Frigola, J., Sabarinathan, R., Mularoni, L., et al. Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics*, 49(12):1684–1692, 2017. ISSN 1546-1718. doi: 10.1038/ng.3991. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cancer;Genome informatics;Genomics Subject_term_id: cancer;genome-informatics;genomics.

Fu, Y.X. and Li, W.H. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, 1993. ISSN 0016-6731, 1943-2631.

Fumagalli, M., Moltke, I., Grarup, N., et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254):1343–1347, 2015. doi: 10.1126/science.aab2319. Publisher: American Association for the Advancement of Science.

Fumagalli, M., Sironi, M., Pozzoli, U., et al. Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLOS Genetics*, 7(11):e1002355, 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002355. Publisher: Public Library of Science.

Galtier, N. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLOS Genetics*, 12(1):e1005774, 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005774.

Galtier, N. and Rousselle, M. How Much Does Ne Vary Among Species? *Genetics*, 216(2):559–572, 2020. ISSN 1943-2631. doi: 10.1534/genetics.120.303622.

Gardiner-Garden, M. and Frommer, M. CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282, 1987. ISSN 0022-2836. doi: 10.1016/0022-2836(87)90689-9.

Garud, N.R., Messer, P.W., Buzbas, E.O., et al. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLOS Genetics*, 11(2):e1005004, 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005004.

Garud, N.R., Messer, P.W., and Petrov, D.A. Detection of hard and soft selective sweeps from Drosophila melanogaster population genomic data. *PLOS Genetics*, 17(2):e1009373, 2021. ISSN 1553-7404. doi: 10.1371/journal.pgen.1009373.

Gayà-Vidal, M. and Albà, M.M. Uncovering adaptive evolution in the human lineage. *BMC Genomics*, 15(1):599, 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-599.

Gazal, S., Sahbatou, M., Babron, M.C., et al. High level of inbreeding in final phase of 1000 Genomes Project. *Scientific Reports*, 5(1):17453, 2015. ISSN 2045-2322. doi: 10.1038/srep17453. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Inbreeding;Population genetics Subject_term_id: inbreeding;population-genetics.

Genovese, G., Friedman, D.J., Ross, M.D., et al. Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. *Science*, 329(5993):841–845, 2010. doi: 10.1126/science.1193032. Publisher: American Association for the Advancement of Science.

Gillespie, J.H. *The Causes of Molecular Evolution*. Oxford University Press, 1994. ISBN 978-0-19-509271-4.

Gillespie, J.H. *Population Genetics: A Concise Guide*. JHU Press, 2004. ISBN 978-1-4214-0170-6.

Good, B.H., Walczak, A.M., Neher, R.A., et al. Genetic Diversity in the Interference Selection Limit. *PLOS Genetics*, 10(3):e1004222, 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004222. Publisher: Public Library of Science.

Gossmann, T.I., Song, B.H., Windsor, A.J., et al. Genome Wide Analyses Reveal Little Evidence for Adaptive Evolution in Many Plant Species. *Molecular Biology and Evolution*, 27(8):1822–1832, 2010. ISSN 0737-4038. doi: 10.1093/molbev/msq079.

Gower, G., Picazo, P.I., Fumagalli, M., et al. Detecting adaptive introgression in human evolution using convolutional neural networks. *eLife*, 10:e64669, 2021. ISSN 2050-084X. doi: 10.7554/eLife.64669. Publisher: eLife Sciences Publications, Ltd.

Granka, J.M., Henn, B.M., Gignoux, C.R., et al. Limited Evidence for Classic Selective Sweeps in African Populations. *Genetics*, 192(3):1049–1064, 2012. ISSN 1943-2631. doi: 10.1534/genetics.112.144071.

Gravel, S. and National Heart, L. Predicting Discovery Rates of Genomic Features. *Genetics*, 197(2):601–610, 2014. doi: 10.1534/genetics.114.162149. Publisher: Oxford Academic.

Grenier, J.K., Arguello, J.R., Moreira, M.C., et al. Global Diversity Lines–A Five-Continent Reference Panel of Sequenced Drosophila melanogaster Strains. *G3: Genes, Genomes, Genetics*, 5(4):593–603, 2015. ISSN 2160-1836. doi: 10.1534/g3.114.015883. Publisher: G3: Genes, Genomes, Genetics Section: Investigations.

Grossman, S.R., Andersen, K.G., Shlyakhter, I., et al. Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell*, 152(4):703–713, 2013. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2013.01.035.

Grossman, S.R., Shylakhter, I., Karlsson, E.K., et al. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science*, 327(5967):883–886, 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1183863.

Haasl, R.J., Johnson, R.C., and Payseur, B.A. The Effects of Microsatellite Selection on Linked Sequence Diversity. *Genome Biology and Evolution*, 6(7):1843–1861, 2014. ISSN 1759-6653. doi: 10.1093/gbe/evu134.

Haasl, R.J. and Payseur, B.A. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, 25(1):5–23, 2016. ISSN 1365-294X. doi: https://doi.org/10.1111/mec.13339.

Hahn, M. *Molecular Population Genetics*. Oxford University Press, Oxford, New York, 2018. ISBN 978-0-87893-965-7.

Hahn, M.W. Toward a Selection Theory of Molecular Evolution. *Evolution*, 62(2):255–265, 2008. ISSN 1558-5646. doi: https://doi.org/10.1111/j.1558-5646.2007.00308.x.

Haldane, J.B.S. *The causes of evolution*. Longmans, Green and Co., London; New York, 1932. OCLC: 5006266.

Haller, B.C. and Messer, P.W. asymptoticMK: A Web-Based Tool for the Asymptotic McDonald–Kreitman Test. *G3: Genes, Genomes, Genetics*, 7(5):1569–1575, 2017. ISSN 2160-1836. doi: 10.1534/g3.117.039693. Publisher: G3: Genes, Genomes, Genetics Section: Investigations.

Haller, B.C. and Messer, P.W. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, 36(3):632–637, 2019. ISSN 0737-4038. doi: 10.1093/molbev/msy228.

Halligan, D.L. and Keightley, P.D. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Research*, 16(7):875–884, 2006. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.5022906. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Halligan, D.L., Oliver, F., Eyre-Walker, A., et al. Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. *PLOS Genetics*, 6(1):e1000825, 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000825. Publisher: Public Library of Science.

Hancock, A.M. and Di Rienzo, A. Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. *Annual Review of Anthropology*, 37(1):197–217, 2008. doi: 10.1146/annurev.anthro.37.081407.085141.

Hancock, A.M., Witonsky, D.B., Alkorta-Aranburu, G., et al. Adaptations to Climate-Mediated Selective Pressures in Humans. *PLOS Genetics*, 7(4):e1001375, 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001375. Publisher: Public Library of Science.

Hancock, A.M., Witonsky, D.B., Ehler, E., et al. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):8924–8930, 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0914625107. Publisher: National Academy of Sciences Section: Colloquium Paper.

Hardy, G.H. Mendelian Proportions in a Mixed Population. *Science*, 28(706):49–50, 1908. doi: 10.1126/science.28.706.49. Publisher: American Association for the Advancement of Science.

Harris, A.M. and DeGiorgio, M. Identifying and Classifying Shared Selective Sweeps from Multilocus Data. *Genetics*, 215(1):143–171, 2020a. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.120.303137. Publisher: Genetics Section: Investigations.

Harris, A.M. and DeGiorgio, M. A Likelihood Approach for Uncovering Selective Sweep Signatures from Haplotype Data. *Molecular Biology and Evolution*, 37(10):3023–3046, 2020b. ISSN 0737-4038. doi: 10.1093/molbev/msaa115.

Harris, A.M., Garud, N.R., and DeGiorgio, M. Detection and Classification of Hard and Soft Sweeps from Unphased Genotypes by Multilocus Genotype Identity. *Genetics*, 210(4):1429–1452, 2018. ISSN 1943-2631. doi: 10.1534/genetics.118.301502.

Harris, H. C. Genetics of Man Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 164(995):298–310, 1966. doi: 10.1098/rspb.1966.0032.

Harris, K. From a database of genomes to a forest of evolutionary trees. *Nature Genetics*, 51(9):1306–1307, 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0492-x.

Harris, K. and Pritchard, J.K. Rapid evolution of the human mutation spectrum. *eLife*, 6:e24284, 2017. ISSN 2050-084X. doi: 10.7554/eLife.24284. Publisher: eLife Sciences Publications, Ltd.

Harrow, J., Frankish, A., Gonzalez, J.M., et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.135350.111.

Hartl, D.L. *A primer of population genetics and genomics*. 2020. ISBN 978-0-19-886229-1 978-0-19-886230-7. OCLC: 1231712803.

Haygood, R., Fedrigo, O., Hanson, B., et al. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics*, 39(9):1140–1144, 2007. ISSN 1546-1718. doi: 10.1038/ng2104.

Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Research Publisher: Nature Publishing Group.

Hejase, H.A., Dukler, N., and Siepel, A. From Summary Statistics to Gene Trees: Methods for Inferring Positive Selection. *Trends in Genetics*, 36(4):243–258, 2020. ISSN 0168-9525. doi: 10.1016/j.tig.2019.12.008. Publisher: Elsevier.

Hejase, H.A., Mo, Z., Campagna, L., et al. SIA: Selection Inference Using the Ancestral Recombination Graph. Technical report, 2021. doi: 10.1101/2021.06.22.449427. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Hernandez, R.D. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24(23):2786–2787, 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn522.

Herráez, D.L., Bauchet, M., Tang, K., et al. Genetic Variation and Recent Positive Selection in Worldwide Human Populations: Evidence from Nearly 1 Million SNPs. *PLOS ONE*, 4(11):e7888, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0007888. Publisher: Public Library of Science.

Hervas, S., Sanz, E., Casillas, S., et al. PopFly: the Drosophila population genomics browser. *Bioinformatics*, 33(17):2779–2780, 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx301.

Hider, J.L., Gittelman, R.M., Shah, T., et al. Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evolutionary Biology*, 13(1):150, 2013. ISSN 1471-2148. doi: 10.1186/1471-2148-13-150.

Hill, W.G. and Robertson, A. The effect of linkage on limits to artificial selection. *Genetics Research*, 8(3):269–294, 1966. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672300010156. Publisher: Cambridge University Press.

Hinds, D.A., Stuve, L.L., Nilsen, G.B., et al. Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science*, 307(5712):1072–1079, 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1105436.

Huang, Y.F. Dissecting genomic determinants of positive selection with an evolution-guided regression model. *Molecular Biology and Evolution*, (msab291), 2021. ISSN 0737-4038. doi: 10.1093/molbev/msab291.

Hudson, R.R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.2.337. Publisher: Oxford Academic.

Hudson, R.R. and Kaplan, N.L. Deleterious background selection with recombination. *Genetics*, 141(4):1605–1617, 1995. ISSN 0016-6731, 1943-2631.

Hudson, R.R., Kreitman, M., and Aguadé, M. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics*, 116(1):153–159, 1987. ISSN 1943-2631. doi: 10.1093/genetics/116.1.153.

Huerta-Sánchez, E., Jin, X., Asan, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513):194–197, 2014. ISSN 1476-4687. doi: 10.1038/nature13408. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7513 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic variation Subject_term_id: genetic-variation.

Huttley, G.A., Smith, M.W., Carrington, M., et al. A Scan for Linkage Disequilibrium Across the Human Genome. *Genetics*, 152(4):1711–1722, 1999. ISSN 1943-2631. doi: 10.1093/genetics/152.4.1711.

Hvilsom, C., Qian, Y., Bataillon, T., et al. Extensive X-linked adaptive evolution in central chimpanzees. *Proceedings of the National Academy of Sciences*, 109(6):2054–2059, 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1106877109. Publisher: National Academy of Sciences Section: Biological Sciences.

James, J.E., Piganeau, G., and Eyre-Walker, A. The rate of adaptive evolution in animal mitochondria. *Molecular Ecology*, 25(1):67–78, 2016. ISSN 1365-294X. doi: 10.1111/mec.13475. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.13475.

Jarvis, J.P., Scheinfeldt, L.B., Soi, S., et al. Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. *PLOS Genetics*, 8(4):e1002641, 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002641. Publisher: Public Library of Science.

Jensen, J.D., Payseur, B.A., Stephan, W., et al. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution*, 73(1):111–114, 2019. ISSN 1558-5646. doi: https://doi.org/10.1111/evo.13650.

Johansson, Å. and Gyllensten, U. Identification of local selective sweeps in human populations since the exodus from Africa. *Hereditas*, 145(3):126–137, 2008. ISSN 1601-5223. doi: 10.1111/j.0018-0661.2008.02054.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0018-0661.2008.02054.x.

Johnson, K.E. and Voight, B.F. Patterns of shared signatures of recent positive selection across human populations. *Nature Ecology & Evolution*, 2(4):713–720, 2018. ISSN 2397-334X. doi: 10.1038/s41559-018-0478-6.

Johri, P., Aquadro, C.F., Beaumont, M., et al. Statistical inference in population genomics. Technical report, 2021. doi: 10.1101/2021.10.27.466171. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Johri, P., Charlesworth, B., and Jensen, J.D. Toward an Evolutionarily Appropriate Null Model: Jointly Inferring Demography and Purifying Selection. *Genetics*, 215(1):173–192, 2020. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.119.303002.

Jukes, T.H. and Cantor, C.R. CHAPTER 24 - Evolution of Protein Molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969. ISBN 978-1-4832-3211-9. doi: 10.1016/B978-1-4832-3211-9.50009-7.

Kanehisa, M., Sato, Y., Furumichi, M., et al. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1):D590–D595, 2019. ISSN 0305-1048. doi: 10.1093/nar/gky962.

Kao, J.Y., Zubair, A., Salomon, M.P., et al. Population genomic analysis uncovers African and European admixture in Drosophila melanogaster populations from the south-eastern United States and Caribbean Islands. *Molecular Ecology*, 24(7):1499–1509, 2015. ISSN 1365-294X. doi: 10.1111/mec.13137. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.13137.

Kapun, M., Nunez, J.C.B., Bogaerts-Márquez, M., et al. Drosophila Evolution over Space and Time (DEST): A New Population Genomics Resource. *Molecular Biology and Evolution*, (msab259), 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab259.

Karasov, T., Messer, P.W., and Petrov, D.A. Evidence that Adaptation in Drosophila Is Not Limited by Mutation at Single Sites. *PLOS Genetics*, 6(6):e1000924, 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000924. Publisher: Public Library of Science.

Karczewski, K.J., Francioli, L.C., Tiao, G., et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2308-7.

Kayser, M., Brauer, S., and Stoneking, M. A Genome Scan to Detect Candidate Regions Influenced by Local Natural Selection in Human Populations. *Molecular Biology and Evolution*, 20(6):893–900, 2003. ISSN 0737-4038. doi: 10.1093/molbev/msg092.

Keightley, P.D., Campos, J.L., Booker, T.R., et al. Inferring the Frequency Spectrum of Derived Variants to Quantify Adaptive Molecular Evolution in Protein-Coding Genes of Drosophila melanogaster. *Genetics*, 203(2):975–984, 2016. ISSN 1943-2631. doi: 10.1534/genetics.116.188102.

Keightley, P.D. and Eyre-Walker, A. Joint Inference of the Distribution of Fitness Effects of Deleterious Mutations and Population Demography Based on Nucleotide Polymorphism Frequencies. *Genetics*, 177(4):2251–2261, 2007. ISSN 1943-2631. doi: 10.1534/genetics.107.080663.

Keightley, P.D. and Eyre-Walker, A. Estimating the Rate of Adaptive Molecular Evolution When the Evolutionary Divergence Between Species is Small. *Journal of Molecular Evolution*, 74(1):61–68, 2012. ISSN 1432-1432. doi: 10.1007/s00239-012-9488-1.

Keightley, P.D. and Jackson, B.C. Inferring the Probability of the Derived vs. the Ancestral Allelic State at a Polymorphic Site. *Genetics*, 209(3):897–906, 2018. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.118.301120.

Kelleher, J., Wong, Y., Wohns, A.W., et al. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0483-y.

Kelley, J.L., Madeoy, J., Calhoun, J.C., et al. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, 16(8):980–989, 2006. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.5157306.

Kern, A.D. and Hahn, M.W. The Neutral Theory in Light of Natural Selection. *Molecular Biology and Evolution*, 35(6):1366–1371, 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy092.

Kern, A.D. and Schrider, D.R. Discoal: flexible coalescent simulations with selection. *Bioinformatics*, 32(24):3839–3841, 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw556.

Kern, A.D. and Schrider, D.R. diploS/HIC: An Updated Approach to Classifying Selective Sweeps. *G3 Genes|Genomes|Genetics*, 8(6):1959–1970, 2018. ISSN 2160-1836. doi: 10.1534/g3.118.200262.

Kim, B.Y., Wang, J.R., Miller, D.E., et al. Highly contiguous assemblies of 101 drosophilid genomes. *bioRxiv*, page 2020.12.14.422775, 2020. doi: 10.1101/2020.12.14.422775.

Kim, Y. and Wiehe, T. Simulation of DNA sequence evolution under models of recent directional selection. *Briefings in Bioinformatics*, 10(1):84–96, 2009. ISSN 1467-5463. doi: 10.1093/bib/bbn048.

Kimura, M. Stochastic Processes and Distribution of Gene Frequencies Under Natural Selection. *Cold Spring Harbor Symposia on Quantitative Biology*, 20:33–53, 1955. ISSN 0091-7451, 1943-4456. doi: 10.1101/SQB.1955.020.01.006. Publisher: Cold Spring Harbor Laboratory Press.

Kimura, M. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964. ISSN 0021-9002, 1475-6072. doi: 10.2307/3211856.

Kimura, M. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129):624–626, 1968. ISSN 1476-4687. doi: 10.1038/217624a0.

Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608):275–276, 1977. ISSN 1476-4687. doi: 10.1038/267275a0.

Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980. ISSN 1432-1432. doi: 10.1007/BF01731581.

Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983. ISBN 978-0-521-31793-1. Google-Books-ID: olIoSumPevYC.

Kimura, M. and Ohta, T. Protein Polymorphism as a Phase of Molecular Evolution. *Nature*, 229(5285):467–469, 1971. ISSN 1476-4687. doi: 10.1038/229467a0. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5285 Primary_atype: Research Publisher: Nature Publishing Group.

Kimura, R., Fujimoto, A., Tokunaga, K., et al. A Practical Genome Scan for Population-Specific Strong Selective Sweeps That Have Reached Fixation. *PLOS ONE*, 2(3):e286, 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0000286.

Kimura, R., Ohashi, J., Matsumura, Y., et al. Gene Flow and Natural Selection in Oceanic Human Populations Inferred from Genome-Wide SNP Typing. *Molecular Biology and Evolution*, 25(8):1750–1761, 2008. ISSN 0737-4038. doi: 10.1093/molbev/msn128.

Kingman, J.F.C. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982. ISSN 0304-4149. doi: 10.1016/0304-4149(82)90011-4.

Koscielny, G., An, P., Carvalho-Silva, D., et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research*, 45(D1):D985–D994, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1055.

Kousathanas, A. and Keightley, P.D. A Comparison of Models to Infer the Distribution of Fitness Effects of New Mutations. *Genetics*, 193(4):1197–1208, 2013. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.112.148023. Publisher: Genetics Section: Investigations.

Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. *Nature*, 304(5925):412–417, 1983. ISSN 1476-4687. doi: 10.1038/304412a0.

Kwiatkowski, D.P. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *The American Journal of Human Genetics*, 77(2):171–192, 2005. ISSN 0002-9297. doi: 10.1086/432519.

Lack, J.B., Cardeno, C.M., Crepeau, M.W., et al. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. *Genetics*, 199(4):1229–1241, 2015. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.115.174664.

Lack, J.B., Lange, J.D., Tang, A.D., et al. A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. *Molecular Biology and Evolution*, 33(12):3308–3313, 2016. ISSN 0737-4038. doi: 10.1093/molbev/msw195.

Lanfear, R., Kokko, H., and Eyre-Walker, A. Population size and the rate of evolution. *Trends in Ecology & Evolution*, 29(1):33–41, 2014. ISSN 0169-5347. doi: 10.1016/j.tree.2013.09.009.

Langley, C.H., Stevens, K., Cardeno, C., et al. Genomic Variation in Natural Populations of Drosophila melanogaster. *Genetics*, 192(2):533–598, 2012. ISSN 1943-2631. doi: 10.1534/genetics.112.142018.

Lappalainen, T., Salmela, E., Andersen, P.M., et al. Genomic landscape of positive natural selection in Northern European populations. *European Journal of Human Genetics*, 18(4):471–478, 2010. ISSN 1476-5438. doi: 10.1038/ejhg.2009.184. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 4 Primary_atype: Research Publisher: Nature Publishing Group.

Lawrie, D.S., Messer, P.W., Hershberg, R., et al. Strong Purifying Selection at Synonymous Sites in D. melanogaster. *PLOS Genetics*, 9(5):e1003527, 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003527. Publisher: Public Library of Science.

Leffler, E.M., Bullaughey, K., Matute, D.R., et al. Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? *PLOS Biology*, 10(9):e1001388, 2012. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001388. Publisher: Public Library of Science.

Lesurf, R., Cotto, K.C., Wang, G., et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Research*, 44(D1):D126–D132, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1203.

Lewontin, R.C. *The Genetic Basis of Evolutionary Change*. Columbia University Press, 1974. ISBN 978-0-231-03392-3. Google-Books-ID: rLMTAQAAIAAJ.

Lewontin, R.C. Twenty-five years ago in Genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone? *Genetics*, 128(4):657–662, 1991. ISSN 0016-6731, 1943-2631.

Lewontin, R.C. and Hubby, J.L. A MOLECULAR APPROACH TO THE STUDY OF GENIC HETEROZYGOSITY IN NATURAL POPULATIONS. II. AMOUNT OF VARIATION AND DEGREE OF HETEROZYGOSITY IN NATURAL POPULATIONS OF DROSOPHILA PSEUDOOBSCURA. *Genetics*, 54(2):595–609, 1966. ISSN 0016-6731, 1943-2631.

Li, M.J., Wang, L.Y., Xia, Z., et al. dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Research*, 42(D1):D910–D916, 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1052.

Li, N. and Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4):2213–2233, 2003. ISSN 1943-2631. doi: 10.1093/genetics/165.4.2213.

Liu, X., Ong, R.T.H., Pillai, E.N., et al. Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations. *The American Journal of Human Genetics*, 92(6):866–881, 2013. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2013.04.021.

Lohmueller, K.E. and Nielsen, R., editors. *Human Population Genomics: Introduction to Essential Concepts and Applications*. Springer International Publishing, 2021. ISBN 978-3-030-61644-1. doi: 10.1007/978-3-030-61646-5.

Lynch, M. The Origins of Eukaryotic Gene Structure. *Molecular Biology and Evolution*, 23(2):450–468, 2006. ISSN 0737-4038. doi: 10.1093/molbev/msj050. Publisher: Oxford Academic.

Lyne, R., Sullivan, J., Butano, D., et al. Cross-organism analysis using InterMine. *genesis*, 53(8):547–560, 2015. ISSN 1526-968X. doi: https://doi.org/10.1002/dvg.22869.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., et al. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(D1):D986–D992, 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt958.

Machado, H.E., Bergland, A.O., Taylor, R., et al. Broad geographic sampling reveals the shared basis and environmental correlates of seasonal adaptation in Drosophila. *eLife*, 10:e67577, 2021. ISSN 2050-084X. doi: 10.7554/eLife.67577.

Mackay, T.F.C., Richards, S., Stone, E.A., et al. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384):173–178, 2012. ISSN 1476-4687. doi: 10.1038/nature10811.

Macpherson, J.M., Sella, G., Davis, J.C., et al. Genomewide Spatial Correspondence Between Nonsynonymous Divergence and Neutral Polymorphism Reveals Extensive Adaptation in Drosophila. *Genetics*, 177(4):2083–2099, 2007. ISSN 1943-2631. doi: 10.1534/genetics.107.080226.

Mallick, S., Li, H., Lipson, M., et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016. ISSN 1476-4687. doi: 10.1038/nature18964.

Martínez-Fundichely, A., Casillas, S., Egea, R., et al. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Research*, 42(D1):D1027–D1032, 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1122.

Masel, J. Genetic drift. *Current Biology*, 21(20):R837–R838, 2011. ISSN 0960-9822. doi: 10.1016/j.cub.2011.08.007.

Mathieson, S. and Mathieson, I. FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution*, 35(12):2957–2970, 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy180.

Mattiangeli, V., Ryan, A.W., McManus, R., et al. A genome-wide approach to identify genetic loci with a signature of natural selection in the Irish population. *Genome Biology*, 7(8):R74, 2006. ISSN 1474-760X. doi: 10.1186/gb-2006-7-8-r74.

McDonald, J.H. and Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654, 1991. ISSN 1476-4687. doi: 10.1038/351652a0.

McVean, G.A. and Cardin, N.J. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005. doi: 10.1098/rstb.2005.1673. Publisher: Royal Society.

McVicker, G., Gordon, D., Davis, C., et al. Widespread genomic signatures of natural selection in hominid evolution. *PLOS Genetics*, 5(5):1–16, 2009. doi: 10.1371/journal.pgen.1000471.

Messer, P.W. and Petrov, D.A. Frequent adaptation and the McDonald–Kreitman test. *Proceedings of the National Academy of Sciences*, 110(21):8615–8620, 2013a. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1220835110. ISBN: 9781220835115 Publisher: National Academy of Sciences Section: Biological Sciences.

Messer, P.W. and Petrov, D.A. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, 28(11):659–669, 2013b. ISSN 0169-5347. doi: 10.1016/j.tree.2013.08.003.

Metspalu, M., Romero, I.G., Yunusbayev, B., et al. Shared and Unique Components of Human Population Structure and Genome-Wide Signals of Positive Selection in South Asia. *The American Journal of Human Genetics*, 89(6):731–744, 2011. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2011.11.010. Publisher: Elsevier.

Mi, H., Huang, X., Muruganujan, A., et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1138.

Migliano, A.B., Romero, I.G., Metspalu, M., et al. Evolution of the Pygmy Phenotype: Evidence of Positive Selection from Genome-wide Scans in African, Asian, and Melanesian Pygmies. *Human Biology*, 85(1/3):251–284, 2013. ISSN 0018-7143, 1534-6617. doi: 10.3378/027.085.0313. Publisher: Wayne State University Press.

Miles, A., Harding, N.J., Bottà, G., et al. Genetic diversity of the African malaria vector Anopheles gambiae. *Nature*, 552(7683):96–100, 2017. ISSN 1476-4687. doi: 10.1038/nature24995. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7683 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic variation;Malaria;Next-generation sequencing Subject_term_id: genetic-variation;malaria;next-generation-sequencing.

Minster, R.L., Hawley, N.L., Su, C.T., et al. A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nature Genetics*, 48(9):1049–1054, 2016. ISSN 1546-1718. doi: 10.1038/ng.3620. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome-wide association studies;Obesity Subject_term_id: genome-wide-association-studies;obesity.

Mondal, M., Bertranpetit, J., and Lao, O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, 10(1):246, 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-08089-7. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic variation Subject_term_id: genetic-variation.

Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., et al. Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, pages 1–5, 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04269-6. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Epigenomics;Genetic variation;Molecular evolution;Mutation Subject_term_id: epigenomics;genetic-variation;molecular-evolution;mutation.

Moran, P.a.P. The statistical processes of evolutionary theory. *The statistical processes of evolutionary theory.*, 1962. Publisher: Clarendon Press; Oxford University Press.

Moutinho, A.F., Bataillon, T., and Dutheil, J.Y. Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology*, 2019a. ISSN 1573-8477. doi: 10.1007/s10682-019-10026-z.

Moutinho, A.F., Trancoso, F.F., and Dutheil, J.Y. The Impact of Protein Architecture on Adaptive Evolution. *Molecular Biology and Evolution*, 36(9):2013–2028, 2019b. ISSN 0737-4038. doi: 10.1093/molbev/msz134.

Murga-Moreno, J., Coronado-Zamora, M., Bodelón, A., et al. PopHumanScan: the online catalog of human genome adaptation. *Nucleic Acids Research*, 47(D1):D1080–D1089, 2019a. ISSN 0305-1048. doi: 10.1093/nar/gky959.

Murga-Moreno, J., Coronado-Zamora, M., Hervas, S., et al. iMKT: the integrative McDonald and Kreitman test. *Nucleic Acids Research*, 47(W1):W283–W288, 2019b. ISSN 0305-1048. doi: 10.1093/nar/gkz372.

Murphy, D., Elyashiv, E., Amster, G., et al. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. *bioRxiv*, page 2021.07.02.450762, 2021. doi: 10.1101/2021.07.02.450762. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Myles, S., Tang, K., Somel, M., et al. Identification and analysis of genomic regions with large between-population differentiation in humans. *Annals of Human Genetics*, 72(Pt 1):99–110, 2008. ISSN 0003-4800. doi: 10.1111/j.1469-1809.2007.00390.x.

Nédélec, Y., Sanz, J., Baharian, G., et al. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*, 167(3):657–669.e21, 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.09.025.

Nei, M. and Li, W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273, 1979.

Nevo, E., Beiles, A., and Ben-Shlomo, R. The Evolutionary Significance of Genetic Diversity: Ecological, Demographic and Life History Correlates. In G.S. Mani, editor, *Evolutionary Dynamics of Genetic Diversity*, Lecture Notes in Biomathematics, pages 13–213. Springer, Berlin, Heidelberg, 1984. ISBN 978-3-642-51588-0. doi: 10.1007/978-3-642-51588-0_2.

Nicolaisen, L.E. and Desai, M.M. Distortions in Genealogies due to Purifying Selection and Recombination. *Genetics*, 195(1):221–230, 2013. ISSN 1943-2631. doi: 10.1534/genetics.113.152983.

Nielsen, R. Molecular Signatures of Natural Selection. *Annual Review of Genetics*, 39(1):197–218, 2005. doi: 10.1146/annurev.genet.39.073003.112420.

Nielsen, R., Akey, J.M., Jakobsson, M., et al. Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310, 2017. ISSN 1476-4687. doi: 10.1038/nature21347.

Nielsen, R., Hubisz, M.J., Hellmann, I., et al. Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, 19(5):838–849, 2009. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.088336.108. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Nielsen, R. and Signorovitch, J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, 63(3):245–255, 2003. ISSN 0040-5809. doi: 10.1016/S0040-5809(03)00005-4.

Nielsen, R. and Slatkin, M. *An Introduction to Population Genetics: Theory and Applications*. Sinauer, 2013. ISBN 978-1-60535-153-7.

Nielsen, R., Williamson, S., Kim, Y., et al. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11):1566–1575, 2005. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.4252305.

Nordborg, M., Charlesworth, B., and Charlesworth, D. The effect of recombination on background selection*. *Genetics Research*, 67(2):159–174, 1996. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672300033619.

Ohashi, J., Naka, I., Patarapotikul, J., et al. Extended Linkage Disequilibrium Surrounding the Hemoglobin E Variant Due to Malarial Selection. *The American Journal of Human Genetics*, 74(6):1198–1208, 2004. ISSN 0002-9297. doi: 10.1086/421330.

Ohta, T. Population size and rate of evolution. *Journal of Molecular Evolution*, 1(4):305–314, 1972. ISSN 1432-1432. doi: 10.1007/BF01653959.

Ohta, T. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246(5428):96–98, 1973. ISSN 1476-4687. doi: 10.1038/246096a0.

Ohta, T. Mutational pressure as the main cause of molecular evolution and polymorphism. *Nature*, 252(5482):351–354, 1974. ISSN 1476-4687. doi: 10.1038/252351a0. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5482 Primary_atype: Reviews Publisher: Nature Publishing Group.

Ohta, T. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, 23:263–286, 1992. ISSN 0066-4162. Publisher: Annual Reviews.

Ohta, T. and Gillespie, J.H. Development of Neutral and Nearly Neutral Theories. *Theoretical Population Biology*, 49(2):128–142, 1996. ISSN 0040-5809. doi: 10.1006/tpbi.1996.0007.

Oleksyk, T.K., Zhao, K., Vega, F.M.D.L., et al. Identifying Selected Regions from Heterozygosity and Divergence Using a Light-Coverage Genomic Dataset from Two Human Populations. *PLOS ONE*, 3(3):e1712, 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0001712. Publisher: Public Library of Science.

O'Reilly, P.F., Birney, E., and Balding, D.J. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Research*, 18(8):1304–1313, 2008. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.067181.107. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Papatheodorou, I., Fonseca, N.A., Keays, M., et al. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Research*, 46(D1):D246–D251, 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1158.

Parsch, J., Novozhilov, S., Saminadin-Peter, S.S., et al. On the Utility of Short Intron Sequences as a Reference for the Detection of Positive and Negative Selection in Drosophila. *Molecular Biology and Evolution*, 27(6):1226–1234, 2010. ISSN 0737-4038. doi: 10.1093/molbev/msq046.

Payseur, B.A., Cutter, A.D., and Nachman, M.W. Searching for Evidence of Positive Selection in the Human Genome Using Patterns of Microsatellite Variability. *Molecular Biology and Evolution*, 19(7):1143–1153, 2002. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a004172.

Peck, J.R. A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. *Genetics*, 137(2):597–606, 1994. ISSN 1943-2631. doi: 10.1093/genetics/137.2.597.

Pennacchio, L.A., Ahituv, N., Moses, A.M., et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, 2006. ISSN 1476-4687. doi: 10.1038/nature05295.

Pickrell, J.K., Coop, G., Novembre, J., et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19(5):826–837, 2009. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.087577.108.

Piras, I.S., De Montis, A., Calò, C.M., et al. Genome-wide scan with nearly 700 000 SNPs in two Sardinian sub-populations suggests some regions as candidate targets for positive selection. *European Journal of Human Genetics*, 20(11):1155–1161, 2012. ISSN 1476-5438. doi: 10.1038/ejhg.2012.65. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 11 Primary_atype: Research Publisher: Nature

Publishing Group Subject_term: Genetic variation;Genome-wide association studies Subject_term_id: genetic-variation;genome-wide-association-studies.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., et al. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.097857.109.

Pool, J.E., Corbett-Detig, R.B., Sugino, R.P., et al. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. *PLOS Genetics*, 8(12):e1003080, 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003080. Publisher: Public Library of Science.

Pouyet, F., Aeschbacher, S., Thiéry, A., et al. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, 7:e36317, 2018. ISSN 2050-084X. doi: 10.7554/eLife.36317.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., et al. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013. ISSN 1476-4687. doi: 10.1038/nature12228. Bandiera_abtest: a Cc_license_type: cc_y Cg_type: Nature Research Journals Number: 7459 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Evolution;Evolutionary biology Subject_term_id: evolution;evolutionary-biology.

Pybus, M., Dall'Olio, G.M., Luisi, P., et al. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Research*, 42(D1):D903–D909, 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1188.

Pybus, M., Luisi, P., Dall'Olio, G.M., et al. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, 31(24):3946–3952, 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv493.

Racimo, F., Marnetto, D., and Huerta-Sánchez, E. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Molecular Biology and Evolution*, 34(2):296–317, 2017. ISSN 0737-4038. doi: 10.1093/molbev/msw216.

Racimo, F., Sankararaman, S., Nielsen, R., et al. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371, 2015. ISSN 1471-0064. doi: 10.1038/nrg3936.

Racimo, F. and Schraiber, J.G. Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms. *PLOS Genetics*, 10(11):e1004697, 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004697.

Raj, T., Kuchroo, M., Replogle, J.M., et al. Common Risk Alleles for Inflammatory Diseases Are Targets of Recent Positive Selection. *The American Journal of Human Genetics*, 92(4):517–529, 2013. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2013.03.001. Publisher: Elsevier.

Rand, D.M. and Kann, L.M. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Molecular Biology and Evolution*, 13(6):735–748, 1996. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a025634.

Rasmussen, M.D., Hubisz, M.J., Gronau, I., et al. Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genetics*, 10(5):e1004342, 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004342.

Robertson, A., Hill, W.G., Ewens, W.J., et al. Population and quantitative genetics of many linked loci in finite populations. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 219(1216):253–264, 1983. doi: 10.1098/rspb.1983.0073.

Rodriguez-Galindo, M., Casillas, S., Weghorn, D., et al. Germline de novo mutation rates on exons versus introns in humans. *Nature Communications*, 11(1):3304, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17162-z. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational models;Genetic variation;Molecular evolution;Mutagenesis Subject_term_id: computational-models;genetic-variation;molecular-evolution;mutagenesis.

Ronen, R., Udpa, N., Halperin, E., et al. Learning Natural Selection from the Site Frequency Spectrum. *Genetics*, 195(1):181–193, 2013. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.113.152587.

Rousselle, M., Laverré, A., Figuet, E., et al. Influence of Recombination and GC-biased Gene Conversion on the Adaptive and Nonadaptive Substitution Rate in Mammals versus Birds. *Molecular Biology and Evolution*, 36(3):458–471, 2019. ISSN 0737-4038. doi: 10.1093/molbev/msy243.

Rousselle, M., Simion, P., Tilak, M.K., et al. Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. *PLOS Genetics*, 16(4):e1008668, 2020. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008668.

Sabeti, P.C., Reich, D.E., Higgins, J.M., et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002. ISSN 1476-4687. doi: 10.1038/nature01140.

Sabeti, P.C., Schaffner, S.F., Fry, B., et al. Positive Natural Selection in the Human Lineage. *Science*, 312(5780):1614–1620, 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1124309.

Sabeti, P.C., Varilly, P., Fry, B., et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918, 2007. ISSN 1476-4687. doi: 10.1038/nature06250.

Sackton, T.B., Kulathinal, R.J., Bergman, C.M., et al. Population Genomic Inferences from Sparse High-Throughput Sequencing of Two Populations of Drosophila melanogaster. *Genome Biology and Evolution*, 1:449–465, 2009. ISSN 1759-6653. doi: 10.1093/gbe/evp048.

Salvador-Martínez, I., Coronado-Zamora, M., Castellano, D., et al. Mapping Selection within Drosophila melanogaster Embryo's Anatomy. *Molecular Biology and Evolution*, 35(1):66–79, 2018. ISSN 0737-4038. doi: 10.1093/molbev/msx266.

Sawyer, S.A. and Hartl, D.L. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, 1992. ISSN 0016-6731, 1943-2631.

Sawyer, S.A., Kulathinal, R.J., Bustamante, C.D., et al. Bayesian Analysis Suggests that Most Amino Acid Replacements in Drosophila Are Driven by Positive Selection. *Journal of Molecular Evolution*, 57(1):S154–S164, 2003. ISSN 1432-1432. doi: 10.1007/s00239-003-0022-3.

Scheinfeldt, L.B., Soi, S., Thompson, S., et al. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biology*, 13(1):R1, 2012. ISSN 1474-760X. doi: 10.1186/gb-2012-13-1-r1.

Schlebusch, C.M., Gattepaille, L.M., Engström, K., et al. Human Adaptation to Arsenic-Rich Environments. *Molecular Biology and Evolution*, 32(6):1544–1555, 2015. ISSN 0737-4038. doi: 10.1093/molbev/msv046.

Schraiber, J.G. and Akey, J.M. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, 2015. ISSN 1471-0064. doi: 10.1038/nrg4005.

Schrider, D.R. Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by Selective Sweeps. *Genetics*, 216(2):499–519, 2020. ISSN 1943-2631. doi: 10.1534/genetics.120.303469.

Schrider, D.R. and Kern, A.D. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution*, 34(8):1863–1877, 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx154.

Schrider, D.R., Shanku, A.G., and Kern, A.D. Effects of Linked Selective Sweeps on Demographic Inference and Model Selection. *Genetics*, 204(3):1207–1223, 2016. ISSN 1943-2631. doi: 10.1534/genetics.116.190223.

Shlyakhter, I., Sabeti, P.C., and Schaffner, S.F. Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429, 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu562.

Shriver, M.D., Kennedy, G.C., Parra, E.J., et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics*, 1(4):274, 2004. ISSN 1479-7364. doi: 10.1186/1479-7364-1-4-274.

Signor, S.A., New, F.N., and Nuzhdin, S. A Large Panel of Drosophila simulans Reveals an Abundance of Common Variants. *Genome Biology and Evolution*, 10(1):189–206, 2018. ISSN 1759-6653. doi: 10.1093/gbe/evx262.

Simonson, T.S., Yang, Y., Huff, C.D., et al. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science*, 329(5987):72–75, 2010. doi: 10.1126/science.1189406. Publisher: American Association for the Advancement of Science.

Simpson, E.H. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. ISSN 2517-6161. doi: https://doi.org/10.1111/j.2517-6161.1951.tb00088.x.

Sjöstrand, A.E., Sjödin, P., and Jakobsson, M. Private haplotypes can reveal local adaptation. *BMC Genetics*, 15(1):61, 2014. ISSN 1471-2156. doi: 10.1186/1471-2156-15-61.

Smith, J.M. and Haigh, J. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35, 1974. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672300014634. Publisher: Cambridge University Press.

Smith, N.G.C. and Eyre-Walker, A. Adaptive protein evolution in Drosophila. *Nature*, 415(6875):1022–1024, 2002. ISSN 1476-4687. doi: 10.1038/4151022a.

Somel, M., Wilson Sayres, M.A., Jordan, G., et al. A Scan for Human-Specific Relaxation of Negative Selection Reveals Unexpected Polymorphism in Proteasome Genes. *Molecular Biology and Evolution*, 30(8):1808–1815, 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst098.

Soni, V., Moutinho, A.F., and Eyre-Walker, A. Site level factors that affect the rate of adaptive evolution in humans and chimpanzees; the effect of contracting population size. Technical report, 2021. doi: 10.1101/2021.05.28.446098. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Souilmi, Y., Lauterbur, M.E., Tobler, R., et al. An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. *Current Biology*, 31(16):3504–3514.e9, 2021. ISSN 0960-9822. doi: 10.1016/j.cub.2021.05.067.

Speidel, L., Cassidy, L., Davies, R.W., et al. Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. *Molecular Biology and Evolution*, 38(9):3497–3511, 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab174.

Speidel, L., Forest, M., Shi, S., et al. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0484-x.

Spence, J.P. and Song, Y.S. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10):eaaw9206, 2019. ISSN 2375-2548. doi: 10.1126/sciadv.aaw9206.

Stoletzki, N. and Eyre-Walker, A. The Positive Correlation between dN/dS and dS in Mammals Is Due to Runs of Adjacent Substitutions. *Molecular Biology and Evolution*, 28(4):1371–1380, 2011. ISSN 0737-4038. doi: 10.1093/molbev/msq320.

Storz, J.F., Payseur, B.A., and Nachman, M.W. Genome Scans of DNA Variability in Humans Reveal Evidence for Selective Sweeps Outside of Africa. *Molecular Biology and Evolution*, 21(9):1800–1811, 2004. ISSN 0737-4038. doi: 10.1093/molbev/msh192.

Strasburg, J.L., Kane, N.C., Raduski, A.R., et al. Effective Population Size Is Positively Correlated with Levels of Adaptive Divergence among Annual Sunflowers. *Molecular Biology and Evolution*, 28(5):1569–1580, 2011. ISSN 0737-4038. doi: 10.1093/molbev/msq270.

Sugden, L.A., Atkinson, E.G., Fischer, A.P., et al. Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nature Communications*, 9(1):703, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03100-7.

Suo, C., Xu, H., Khor, C.C., et al. Natural positive selection and north–south genetic diversity in East Asia. *European Journal of Human Genetics*, 20(1):102–110, 2012. ISSN 1476-5438. doi: 10.1038/ejhg.2011.139. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group.

Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989. ISSN 0016-6731, 1943-2631.

Tamura, K. and Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526, 1993. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040023.

Tang, K., Thornton, K.R., and Stoneking, M. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLOS Biology*, 5(7):e171, 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050171.

Tataru, P., Mollion, M., Glémin, S., et al. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. *Genetics*, 207(3):1103–1119, 2017. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.117.300323.

Templeton, A.R. Contingency Tests of Neutrality Using Intra/Interspecific Gene Trees: The Rejection of Neutrality for the Evolution of the Mitochondrial Cytochrome Oxidase II Gene in the Hominoid Primates. *Genetics*, 144(3):1263–1270, 1996. ISSN 0016-6731, 1943-2631.

Tennessen, J.A. and Akey, J.M. Parallel Adaptive Divergence among Geographically Diverse Human Populations. *PLOS Genetics*, 7(6):e1002127, 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002127. Publisher: Public Library of Science.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, 337(6090):64–69, 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1219240.

Tennessen, J.A., Madeoy, J., and Akey, J.M. Signatures of positive selection apparent in a small sample of human exomes. *Genome Research*, 20(10):1327–1334, 2010. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.106161.110. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Teshima, K.M., Coop, G., and Przeworski, M. How reliable are empirical genomic scans for selective sweeps? *Genome Research*, 16(6):702–712, 2006. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.5105206.

The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1108.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1099.

Thornton, K.R. Automating approximate Bayesian computation by local linear regression. *BMC Genetics*, 10(1):35, 2009. ISSN 1471-2156. doi: 10.1186/1471-2156-10-35.

Thornton, K.R. A C++ Template Library for Efficient Forward-Time Population Genetic Simulation of Large Populations. *Genetics*, 198(1):157–166, 2014. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.114.165019. Publisher: Genetics Section: Investigations.

Thurmond, J., Goodman, J.L., Strelets, V.B., et al. FlyBase 2.0: the next generation. *Nucleic Acids Research*, 47(D1):D759–D765, 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1003.

Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., et al. Haplotype Diversity and Linkage Disequilibrium at Human G6PD: Recent Origin of Alleles That Confer Malarial Resistance. *Science*, 293(5529):455–462, 2001. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1061573.

Torada, L., Lorenzon, L., Beddis, A., et al. ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, 20(9):337, 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2927-x.

Torres, R., Stetter, M.G., Hernandez, R.D., et al. The Temporal Dynamics of Background Selection in Nonequilibrium Populations. *Genetics*, 214(4):1019–1030, 2020. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.119.302892. Publisher: Genetics Section: Investigations.

Torres, R., Szpiech, Z.A., and Hernandez, R.D. Human demographic history has amplified the effects of background selection across the genome. *PLOS Genetics*, 14(6):e1007387, 2018. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007387.

Uricchio, L.H. and Hernandez, R.D. Robust Forward Simulations of Recurrent Hitchhiking. *Genetics*, 197(1):221–236, 2014. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.113.156935.

Uricchio, L.H., Petrov, D.A., and Enard, D. Exploiting selection at linked sites to infer the rate and strength of adaptation. *Nature Ecology & Evolution*, 3(6):977–984, 2019. ISSN 2397-334X. doi: 10.1038/s41559-019-0890-6.

Vernot, B., Stergachis, A.B., Maurano, M.T., et al. Personal and population genomics of human regulatory variation. *Genome Research*, 22(9):1689–1697, 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.134890.111. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Vernot, B., Tucci, S., Kelso, J., et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*, 352(6282):235–239, 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aad9416.

Villegas-Mirón, P., Acosta, S., Nye, J., et al. Chromosome X-wide analysis of positive selection in human populations: from common and private signals to selection impact on inactivated genes and enhancers-like signatures. Technical report, 2021. doi: 10.1101/2021.05.24.445399. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Vitti, J.J., Grossman, S.R., and Sabeti, P.C. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, 47(1):97–120, 2013. doi: 10.1146/annurev-genet-111212-133526.

Voight, B.F., Kudaravalli, S., Wen, X., et al. A Map of Recent Positive Selection in the Human Genome. *PLOS Biology*, 4(3):e72, 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040072.

Vos, M., Beek, T.A.H.t., Driel, M.A.v., et al. ODoSE: A Webserver for Genome-Wide Calculation of Adaptive Divergence in Prokaryotes. *PLOS ONE*, 8(5):e62447, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0062447. Publisher: Public Library of Science.

Wang, E.T., Kodama, G., Baldi, P., et al. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences*, 103(1):135–140, 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0509691102.

Wang, Z., Wang, J., Kourakos, M., et al. Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, 21(8):2689–2705, 2021. ISSN 1755-0998. doi: 10.1111/1755-0998.13386. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13386.

Ward, L.D. and Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Research*, 44(D1):D877–D881, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1340.

Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975. ISSN 0040-5809. doi: 10.1016/0040-5809(75)90020-9.

Weir, B.S., Cardon, L.R., Anderson, A.D., et al. Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, 15(11):1468–1476, 2005. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.4398405.

Weir, B.S. and Cockerham, C.C. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution; International Journal of Organic Evolution*, 38(6):1358–1370, 1984. ISSN 1558-5646. doi: 10.1111/j.1558-5646.1984.tb05657.x.

Williamson, S. Adaptation in the env Gene of HIV-1 and Evolutionary Theories of Disease Progression. *Molecular Biology and Evolution*, 20(8):1318–1325, 2003. ISSN 0737-4038. doi: 10.1093/molbev/msg144.

Williamson, S.H., Hubisz, M.J., Clark, A.G., et al. Localizing Recent Adaptive Evolution in the Human Genome. *PLOS Genetics*, 3(6):e90, 2007. ISSN 1553-7404. doi: 10.1371/journal.pgen.0030090.

Wohns, A.W., Wong, Y., Jeffery, B., et al. A unified genealogy of modern and ancient genomes. Technical report, 2021. doi: 10.1101/2021.02.16.431497. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Wright, S. EVOLUTION IN MENDELIAN POPULATIONS. *Genetics*, 16(2):97–159, 1931. ISSN 1943-2631. doi: 10.1093/genetics/16.2.97.

Wright, S. The Distribution of Gene Frequencies Under Irreversible Mutation. *Proceedings of the National Academy of Sciences*, 24(7):253–259, 1938.

Wright, S. Genetical Structure of Populations. *Nature*, 166(4215):247–249, 1950. ISSN 1476-4687. doi: 10.1038/166247a0. Number: 4215 Publisher: Nature Publishing Group.

Wuren, T., Simonson, T.S., Qin, G., et al. Shared and Unique Signals of High-Altitude Adaptation in Geographically Distinct Tibetan Populations. *PLOS ONE*, 9(3):e88252, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0088252. Publisher: Public Library of Science.

Xu, S., Li, S., Yang, Y., et al. A Genome-Wide Search for Signals of High-Altitude Adaptation in Tibetans. *Molecular Biology and Evolution*, 28(2):1003–1011, 2011. ISSN 0737-4038. doi: 10.1093/molbev/msq277.

Yi, X., Liang, Y., Huerta-Sanchez, E., et al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, 329(5987):75–78, 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1190371.

Zeng, K., Fu, Y.X., Shi, S., et al. Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. *Genetics*, 174(3):1431–1439, 2006. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.106.061432.

Zerbino, D.R., Achuthan, P., Akanni, W., et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1098.

Zhang, C., Bailey, D.K., Awad, T., et al. A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. *Bioinformatics*, 22(17):2122–2128, 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl365.

Zhang, Y.B., Li, X., Zhang, F., et al. A Preliminary Study of Copy Number Variation in Tibetans. *PLOS ONE*, 7(7):e41768, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0041768. Publisher: Public Library of Science.

Zhen, Y., Huber, C.D., Davies, R.W., et al. Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and Drosophila melanogaster. *Genome Research*, 31(1):110–120, 2021. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.256636.119.

Zuckerkandl, E. and Pauling, L. Evolutionary Divergence and Convergence in Proteins. In V. Bryson and H.J. Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, 1965. ISBN 978-1-4832-2734-4. doi: 10.1016/B978-1-4832-2734-4.50017-6.

Appendices

Appendix A

PopHumanScan: the online catalog of human genome adaptation

Table A.1: [

Number of lines resampled in each Drosophila population available at iMKT]Number of lines resampled in each Drosophila population available at iMKT

Population code	**Population description**	**Metapopulation**	**Sample size**
CDX	Chinese Dai in Xishuangbanna, China	🟢 East-Asian (EAS)	93
CDX	Chinese Dai in Xishuangbanna, China	🟢 East-Asian (EAS)	93
CHB	Han Chinese in Beijing, China	🟢 East-Asian (EAS)	103
CHS	Southern Han Chinese	🟢 East-Asian (EAS)	105
JPT	Japanese in Tokyo, Japan	🟢 East-Asian (EAS)	104
KHV	Kinh in Ho Chi Minh City, Vietnam	🟢 East-Asian (EAS)	99
CEU	Utah residents (CEPH) with Northern and Western European ancestry	🔵 European (EUR)	99
GBR	British in England and Scotland	🔵 European (EUR)	91
FIN	Finnish in Finland	⚫ European (EUR)	99
IBS	Iberian Populations in Spain	🔵 European (EUR)	107
TSI	Toscani in Italia	🔵 European (EUR)	107
ESN	Esan in Nigeria	🟡 African (AFR)	99
GWD	Gambian in Western Division, Mandinka	⚫ African (AFR)	113
LWK	Luhya in Webuye, Kenya	🟡 African (AFR)	99
MSL	Mende in Sierra Leone	🟡 African (AFR)	85
YRI	Yoruba in Ibadan, Nigeria	🟡 African (AFR)	108
ACB	African Caribbean in Barbados	🟡 African (AFR)	96
ASW	People with African Ancestry in Southwest USA	🟣 African (AFR)	61
BEB	Bengali in Bangladesh	🟣 South-Asian (SAS)	86
GIH	Gujarati Indians in Houston, TX, USA	⚫ South-Asian (SAS)	103
ITU	Indian Telugu in the UK	🟣 South-Asian (SAS)	102
PJL	Punjabi in Lahore, Pakistan	🟣 South-Asian (SAS)	96
STU	Sri Lankan Tamil in the UK	🟣 South-Asian (SAS)	102

Table A.2: Compendium of candidate regions under selection extracted from 268 publications. (Table in XLS format, see supplementary material https://academic.oup.com/nar/article/47/D1/D1080/5134333).

Table A.3: Statistical over-representation test of Gene Ontology terms in 1,447 GENCODE protein-coding genes overlapping our candidate regions under selection, according to the GO Molecular Function classification (released 2018/07/03) and the PANTHER over-representation test (released 2017/12/05). Significance was tested with Fisher's Exact Test with FDR multiple test correction.

GO Molecular Function	**Homo sapiens - REFLIST (21042)**	**PopHumanScan (1476)**	**PopHumanScan (expected)**	**Over/Under**	**Fold enrichment**	**P-value**	**FDR**
adenyl ribonucleotide binding (GO:0032559)	1557	154	109.22	+	1.41	4.59E-05	3.56E-02
↦adenyl nucleotide binding (GO:0030554)	1569	154	110.06	+	1.4	7.42E-05	4.93E-02
↦binding (GO:0005488)	15105	1163	1059.55	+	1.1	2.94E-09	6.84E-06
ion binding (GO:0043167)	6239	513	437.64	+	1.17	4.37E-05	4.06E-02
protein binding (GO:0005515)	11830	947	829.82	+	1.14	2.13E-09	9.90E-06

Table A.4: Statistical over-representation test of Gene Ontology terms in 1,447 GENCODE protein-coding genes overlapping our candidate regions under selection, according to the GO Biological Process classification (released 2018/07/03) and the PANTHER over-representation test (released 2017/12/05). Significance was tested with Fisher's Exact Test with FDR multiple test correction.

GO Biological Process	Homo sapiens - REFLIST (21042)	PopHumanScan (1476)	PopHumanScan (expected)	Over/Under	Fold enrichment	P-value	FDR
regulation of neuron projection development (GO:0010975)	477	63	33.46	+	1.88	8.66E-06	1.23E-02
↦regulation of biological process (GO:0050789)	11443	885	802.67	+	1.1	3.07E-05	3.70E-02
↦biological regulation (GO:0065007)	12104	950	849.04	+	1.12	2.45E-07	1.28E-03
↦regulation of cellular process (GO:0050794)	10758	848	754.62	+	1.12	2.70E-06	6.05E-03
↦regulation of multicellular organismal process (GO:0051239)	2904	262	203.7	+	1.29	3.96E-05	4.42E-02
↦nervous system development (GO:0007399)	2245	217	157.48	+	1.38	3.74E-06	7.32E-03
↦anatomical structure development (GO:0048856)	5179	443	363.28	+	1.22	5.45E-06	9.48E-03
↦developmental process (GO:0032502)	5501	466	385.87	+	1.21	7.14E-06	1.12E-02
↦multicellular organismal process (GO:0032501)	6697	550	469.76	+	1.17	1.97E-05	2.57E-02
↦cellular process (GO:0009987)	15086	1147	1058.21	+	1.08	4.42E-07	1.73E-03
↦cellular component organization (GO:0016043)	5448	472	382.15	+	1.24	5.07E-07	1.59E-03
↦cellular component organization or biogenesis (GO:0071840)	5622	480	394.36	+	1.22	1.92E-06	5.01E-03

Table A.5: Statistical over-representation test of Gene Ontology terms in 1,447 GENCODE protein-coding genes overlapping our candidate regions under selection, according to the GO Cellular Component classification (released 2018/07/03) and the PANTHER over-representation test (released 2017/12/05). Significance was tested with Fisher's Exact Test with FDR multiple test correction.

GO Cellular Component	**Homo sapiens - REFLIST (21042)**	**PopHumanScan (1476)**	**PopHumanScan (expected)**	**Over/Under**	**Fold enrichment**	**P-value**	**FDR**
presynaptic membrane (GO:0042734)	84	16	5.89	+	2.72	7.67E-04	4.34E-02
↦neuron part (GO:0097458)	1601	162	112.3	+	1.44	9.60E-06	1.19E-03
↦cell part (GO:0044464)	17000	1284	1192.47	+	1.08	9.63E-10	9.54E-07
↦ cell (GO:0005623)	17027	1287	1194.37	+	1.08	5.53E-10	1.10E-06
↦synapse part (GO:0044456)	752	84	52.75	+	1.59	8.47E-05	6.99E-03
↦synapse (GO:0045202)	901	111	63.2	+	1.76	7.78E-08	3.08E-05
*mapsto*synaptic membrane (GO:0097060)	328	43	23.01	+	1.87	2.79E-04	1.90E-02
↦plasma membrane region (GO:0098590)	1098	115	77.02	+	1.49	6.03E-05	5.43E-03
↦membrane (GO:0016020)	9701	782	680.48	+	1.15	3.28E-07	9.28E-05
↦plasma membrane part (GO:0044459)	2850	269	199.91	+	1.35	1.27E-06	2.79E-04
↦plasma membrane (GO:0005886)	5571	460	390.78	+	1.18	1.11E-04	8.47E-03
↦cell periphery (GO:0071944)	5689	473	399.06	+	1.19	3.95E-05	3.72E-03
↦membrane part (GO:0044425)	6926	573	485.83	+	1.18	4.16E-06	6.34E-04
transport vesicle membrane (GO:0030658)	193	28	13.54	+	2.07	8.50E-04	4.67E-02

↦organelle part (GO:0044422)	9416	739	660.49	+	1.12	7.73E-05	6.66E-03
↦organelle (GO:0043226)	13417	1023	941.14	+	1.09	1.52E-05	1.77E-03
↦intracellular organelle (GO:0043229)	12628	968	885.8	+	1.09	2.21E-05	2.30E-03
↦intracellular part (GO:0044424)	14425	1099	1011.85	+	1.09	1.55E-06	3.06E-04
↦intracellular (GO:0005622)	14699	1123	1031.07	+	1.09	2.86E-07	9.44E-05
↦cytoplasm (GO:0005737)	11502	889	806.81	+	1.1	3.05E-05	3.02E-03
↦intracellular organelle part (GO:0044446)	9178	721	643.79	+	1.12	1.04E-04	8.22E-03
cytoplasmic region (GO:0099568)	483	56	33.88	+	1.65	7.47E-04	4.35E-02
cell junction (GO:0030054)	1271	133	89.15	+	1.49	1.66E-05	1.83E-03
nucleoplasm (GO:0005654)	3487	320	244.6	+	1.31	9.81E-07	2.43E-04
↦nuclear lumen (GO:0031981)	4093	362	287.11	+	1.26	4.15E-06	6.85E-04
↦intracellular organelle lumen (GO:0070013)	5219	424	366.09	+	1.16	8.84E-04	4.60E-02
↦organelle lumen (GO:0043233)	5219	424	366.09	+	1.16	8.84E-04	4.73E-02
↦membrane-enclosed lumen (GO:0031974)	5219	424	366.09	+	1.16	8.84E-04	4.49E-02
↦nuclear part (GO:0044428)	4489	393	314.88	+	1.25	3.37E-06	6.07E-04
cytoskeleton (GO:0005856)	2171	199	152.29	+	1.31	2.22E-04	1.57E-02

↦intracellular non-membrane-bounded organelle (GO:0043232)	4195	367	294.26	+	1.25	8.22E-06	1.16E-03
↦non-membrane-bounded organelle (GO:0043228)	4195	367	294.26	+	1.25	8.22E-06	1.08E-03
plasma membrane bounded cell projection (GO:0120025)	2067	188	144.99	+	1.3	4.61E-04	2.85E-02
↦cell projection (GO:0042995)	2135	195	149.76	+	1.3	2.84E-04	1.88E-02
protein-containing complex (GO:0032991)	5377	445	377.17	+	1.18	1.20E-04	8.81E-03
integral component of membrane (GO:0016021)	5467	446	383.49	+	1.16	4.30E-04	2.75E-02
↦intrinsic component of membrane (GO:0031224)	5618	456	394.08	+	1.16	5.40E-04	3.24E-02

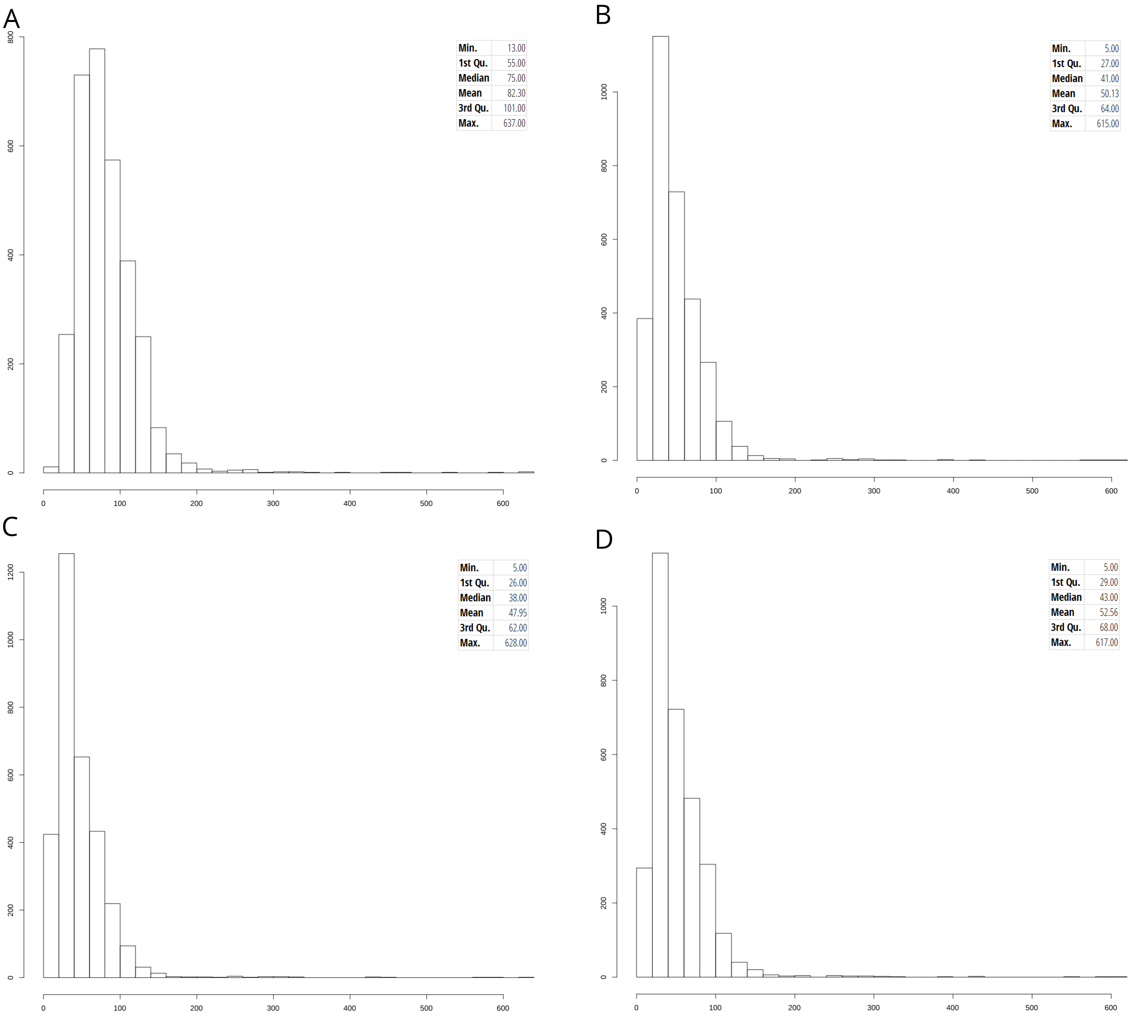


Figure A.1: Distribution of the number of SNPs in the 10-kb analyzed windows for four representative populations, one of each human meta-population: (A) YRI (African), (B) CEU (European), (C) CHB (East-Asian), and (D) GIH (outh-Asian).

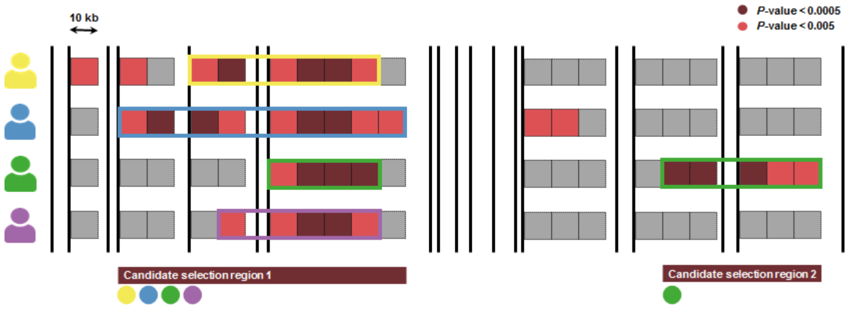


Figure A.2: Definition of candidate regions under selection. Rows represent the results of one single variation statistic calculated along the region on a population. In this example, each of the four populations represented corresponds to a different human meta-population: African, European, East-Asian, and South-Asian. Squares represent 10-kb windows analyzed along this genomic region with PopHuman, while empty regions between squares represent regions that were not analyzed because they contain non-accessible bases (black vertical lines) according to the Pilot-style Accessibility Mask of the 1000GP [see (Casillas et al., 2018) for details]. The color of the squares represents the P-value of the empirical distribution for the corresponding variation statistic and population: P-value $<$ 0.0005 (dark red), 0.0005 $<$ P-value $<$ 0.005 (red), and P-value $>$ 0.005 (gray). Rectangles spanning consecutive 10-kb windows along a row represent candidate regions under selection for the corresponding variation statistic and population, i.e., contiguous genomic regions containing at least one 10-kb significant window (P-value $<$ 0.0005) and spanning adjacent windows with P-value $<$ 0.005. In addition, they may span stretches <20 kb of contiguous nucleotides not analyzed in PopHuman. In the case of the African population, the candidate region under selection does not extend to the windows with P-value $<$ 0.005 to the left because there is a 10-kb window in the middle with P-value $>$ 0.005. In the case of the European population, the candidate region under selection does not extend to the windows with P-value $<$ 0.005 to the right because there is a stretch >20 kb of contiguous nucleotides not analyzed in PopHuman in the middle. On the contrary, regions within candidate regions under selection not analyzed in PopHuman are <20-kb long. Finally, candidate regions detected in each population are stacked to a final set of candidate regions under selection (maroon boxes at the bottom of the figure). In this example, two different candidate regions are detected: the first one with signals in the four meta-populations, and the second one with signals in the East-Asian meta-population. In the PopHumanScan analysis, empirical distributions are calculated for 7 different variation statistics and 22 populations (or 3 population pairs, depending on the variation statistic), so 116 empirical distributions are stacked simultaneously for autosomal regions (see text for details). Cited reference: Casillas et al. (2018)

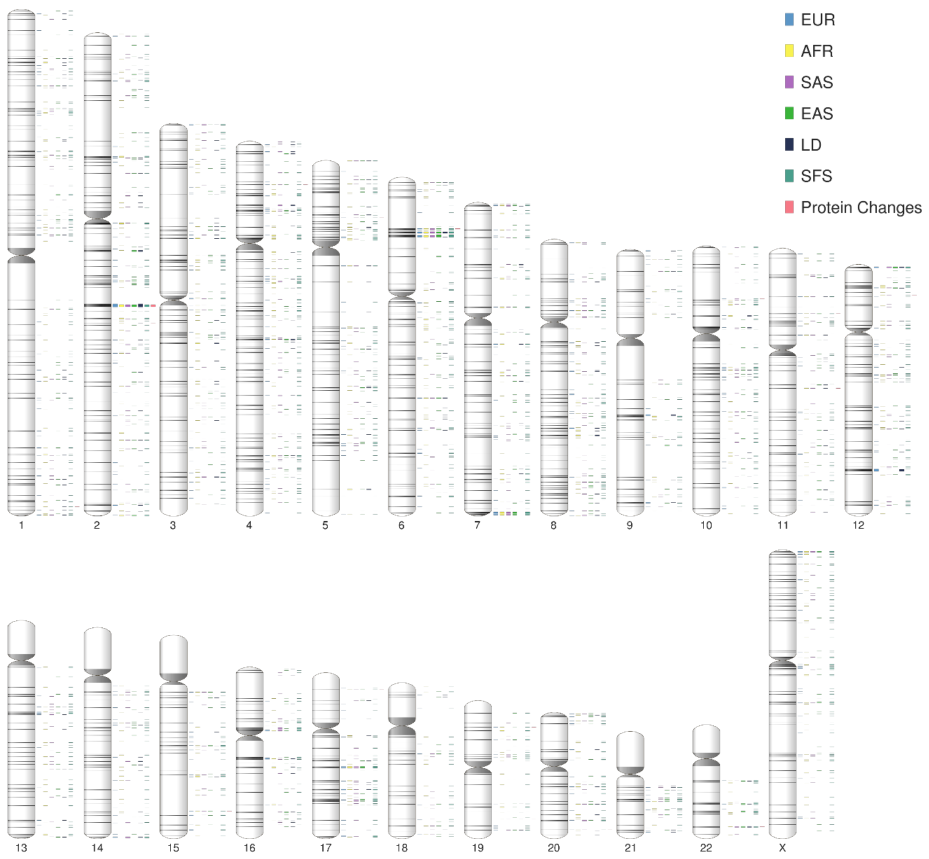


Figure A.3: Representation of the candidate regions under selection included in PopHumanScan in a chromosome ideogram. Meta-populations and signature types are color-coded. Meta-populations: African (AFR), European (EUR), East-Asian (EAS), South-Asian (SAS). Signature types: Linkage Disequilibrium (LD), Site Frequency Spectrum (SFS), and Protein Changes.

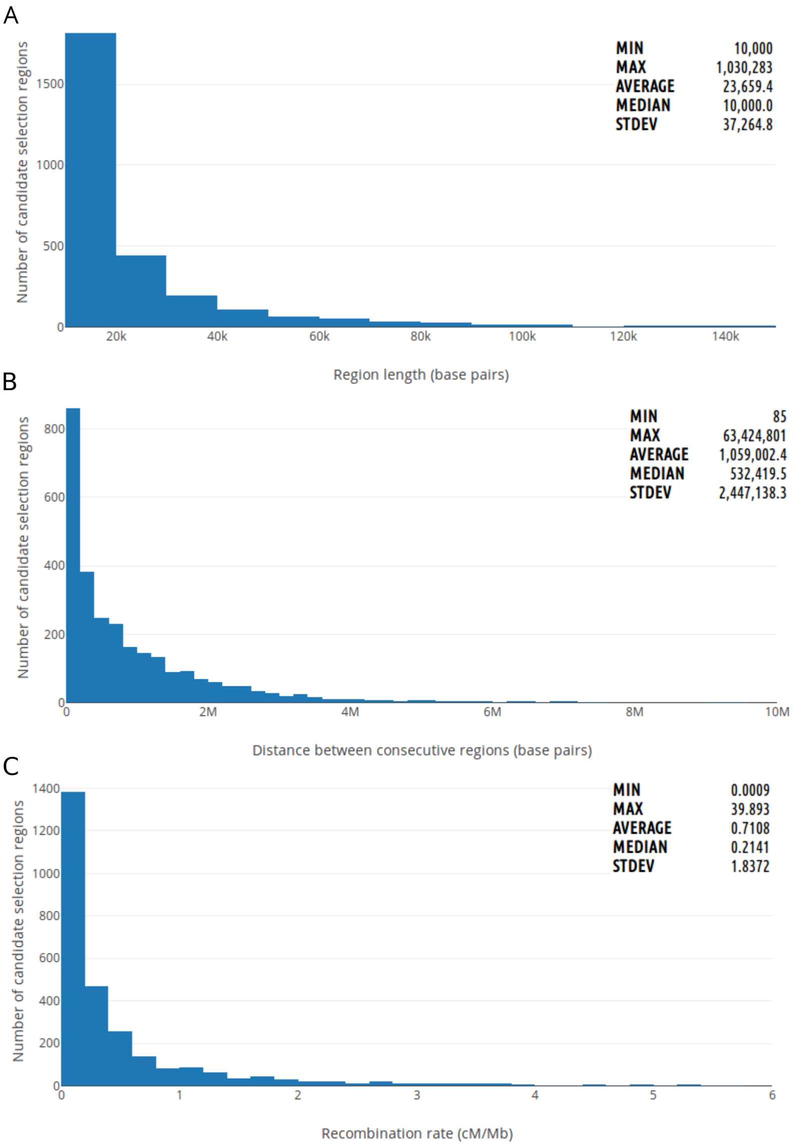


Figure A.4: Distribution of the (A) length of candidate regions under selection (bin size = 10kb), (B) distance between consecutive regions (bin size = 200kb), and (C) recombination rate of candidate regions under selection (bin size = 0.2cM/Mb).

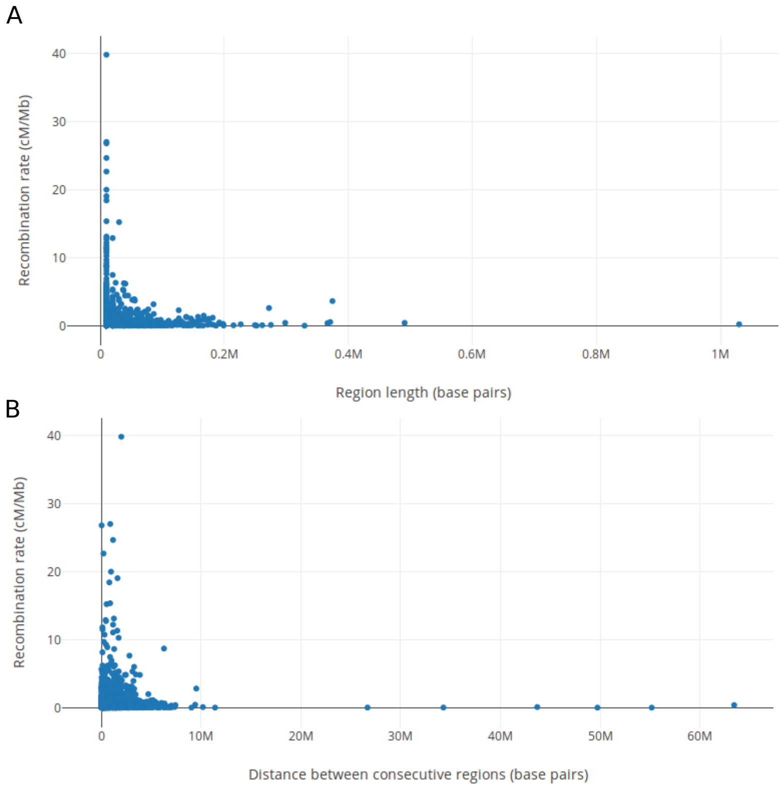


Figure A.5: Recombination rate (cM/Mb) as a function of (A) region length (base pairs), and (B) distance between consecutive regions (base pairs).

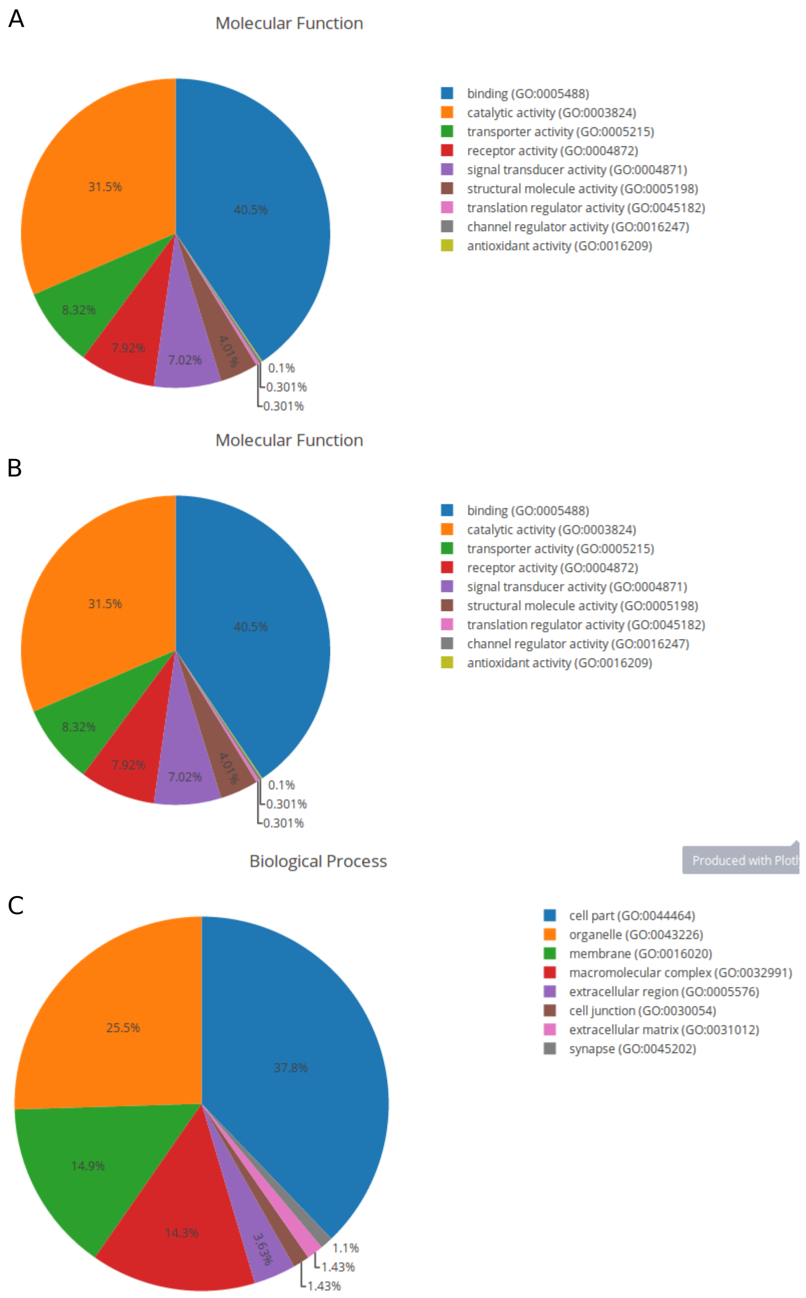


Figure A.6: Functional classification of 1,447 GENCODE protein-coding genes overlapping our candidate regions under selection, according to Gene Ontology terms. (A) Molecular Function; (B) Biological Process; (C) Cellular Component.

Appendix B

Imputed McDonald and Kreitman Test - Appendix

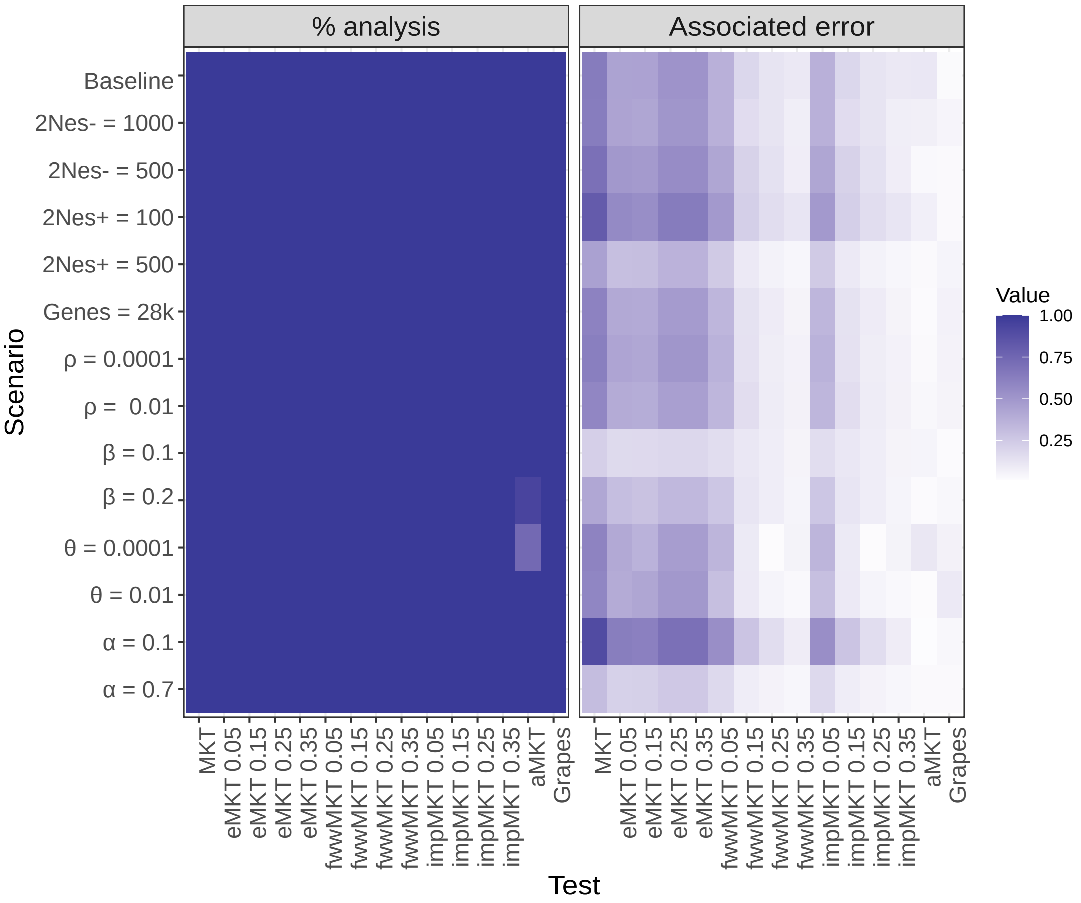


Figure B.1: Percentage of analyzed replicas and associated error bias in estimation for each scenario and MKT approach.

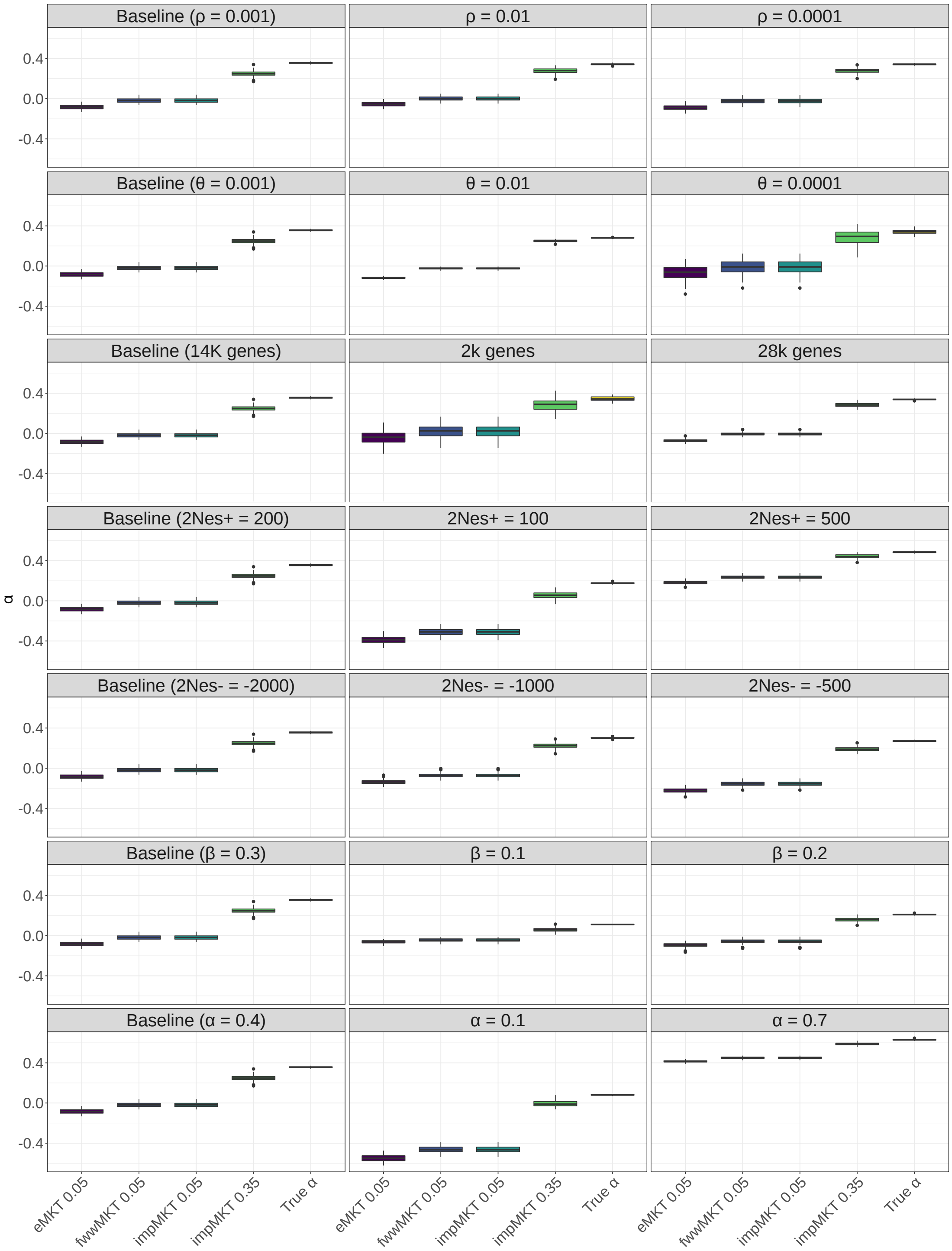


Figure B.2: impMKT α estimations using different cutoff

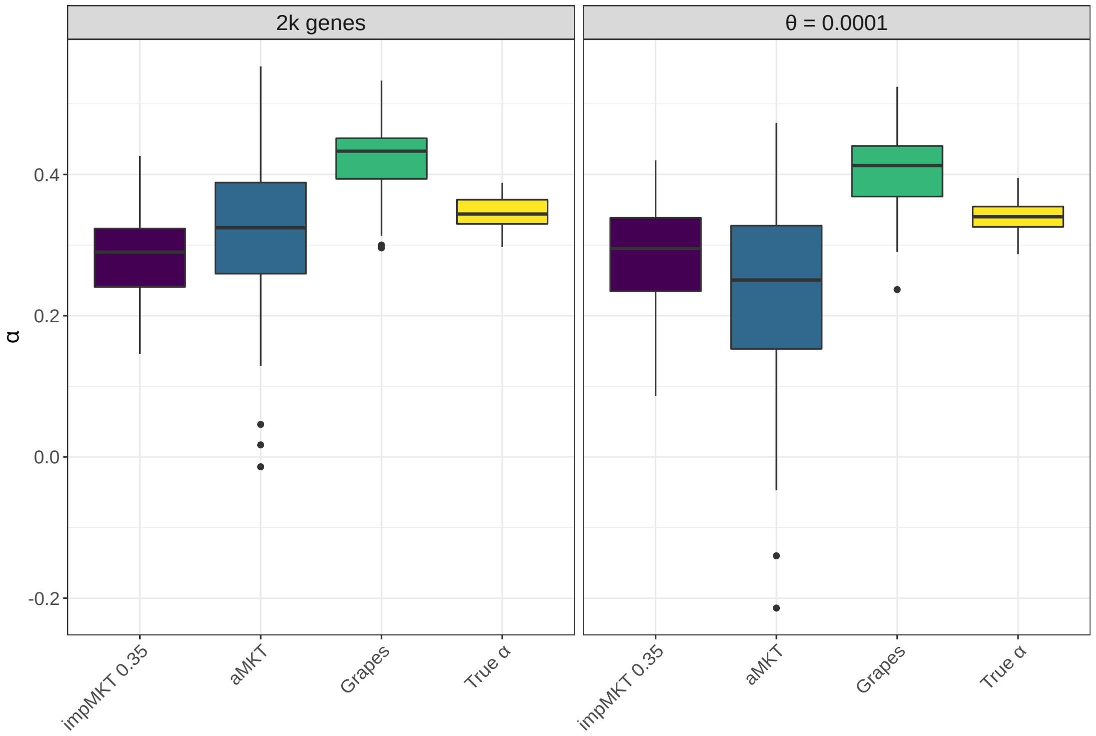


Figure B.3: Effect on α estimation reducing of the amount of segregating sites (i.e., reducing the mutation rate θ to 0.0001, and reducing the total number of simulated genes to 2000) for different MKT approaches.

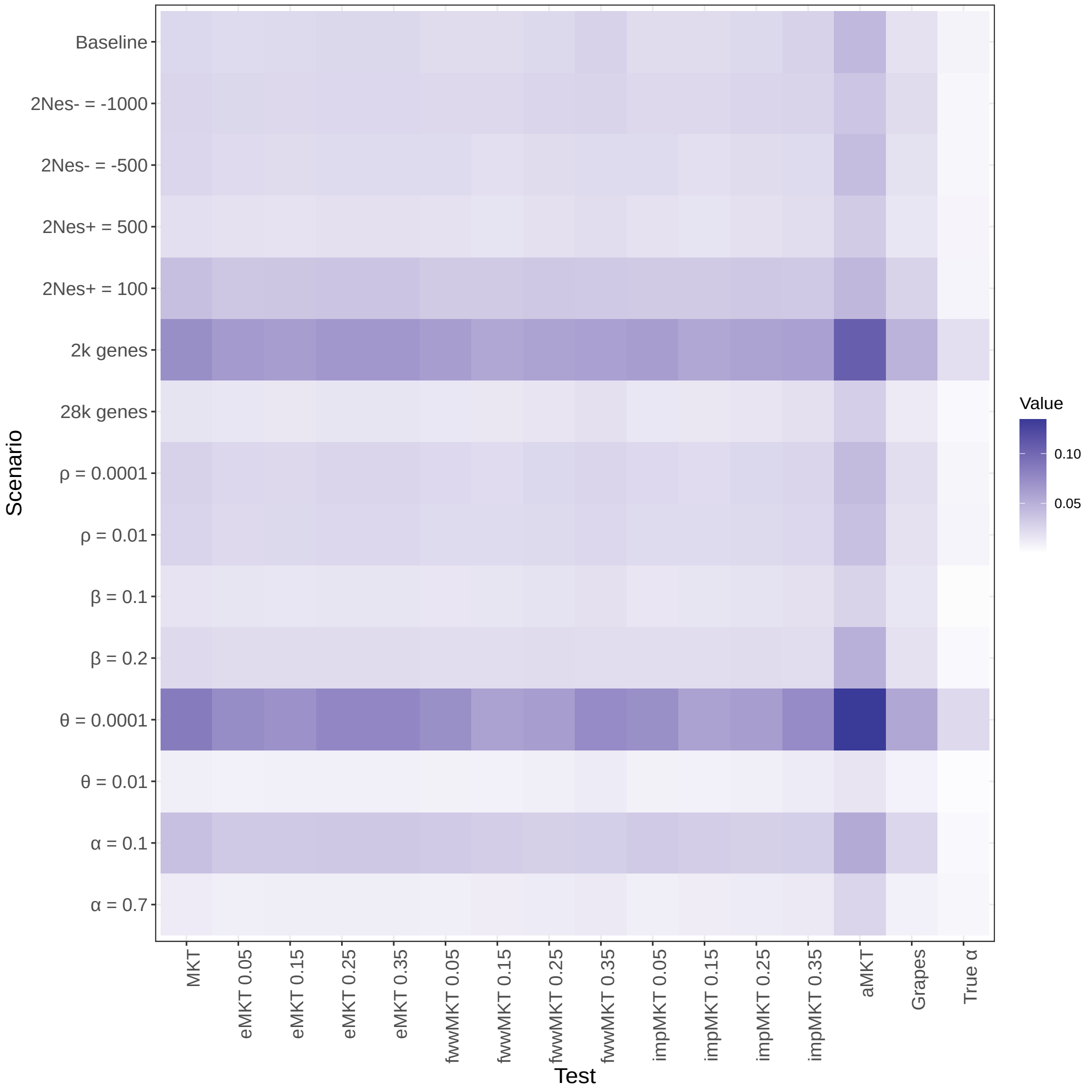


Figure B.4: Standard deviation heatmap for each scenario and MKT approach

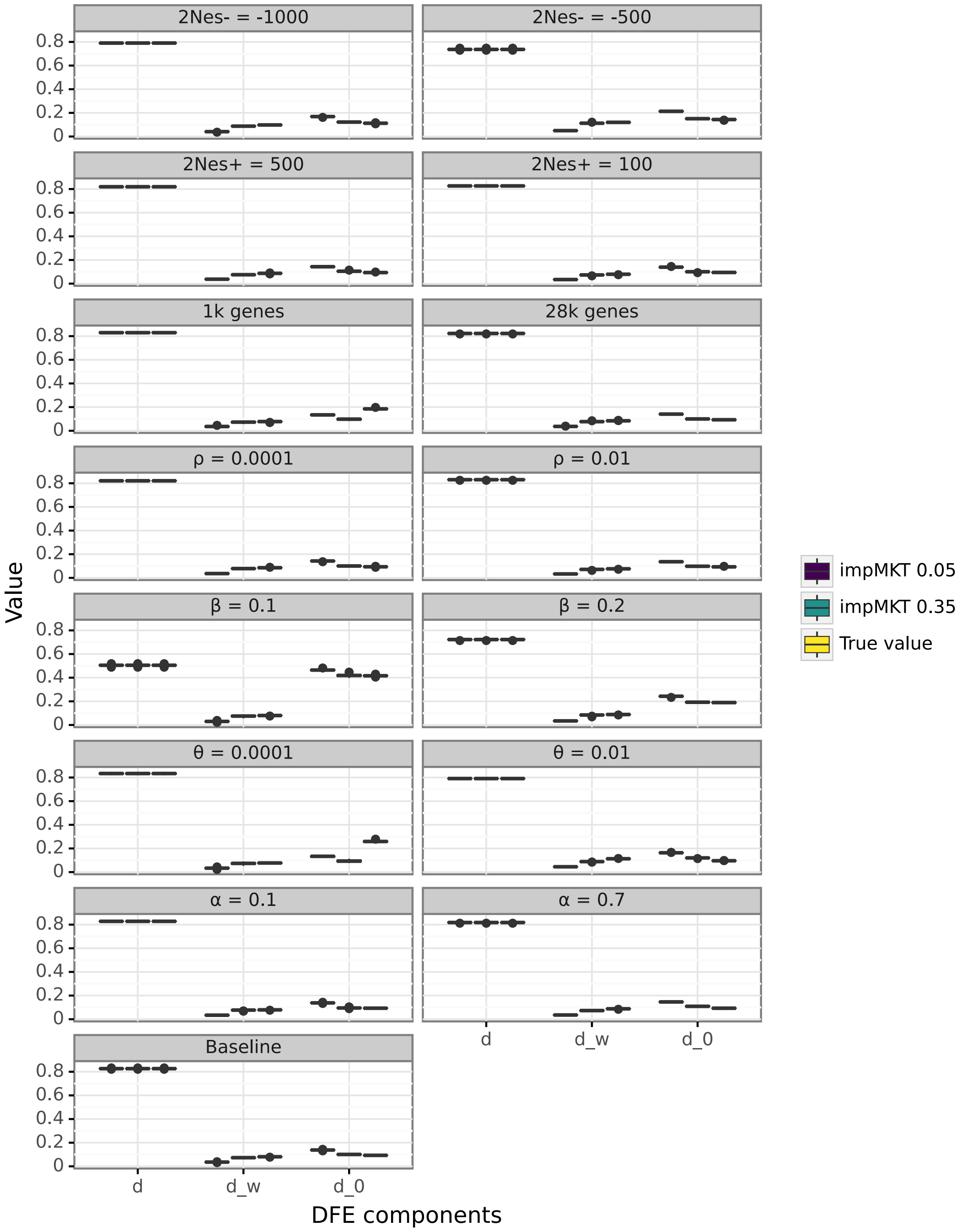


Figure B.5: d, d_w and d_0 impMKT estimations on different SLiM simulated scenarios. ρ and θ are the population-scaled recombination and mutation rates ($\rho = 4N_e r$, $\theta = 4N_e\mu$). $2N_e s$ is the scaled-population selection coefficient for beneficial and deleterious alleles. β is the shape parameter of the gamma DFE.

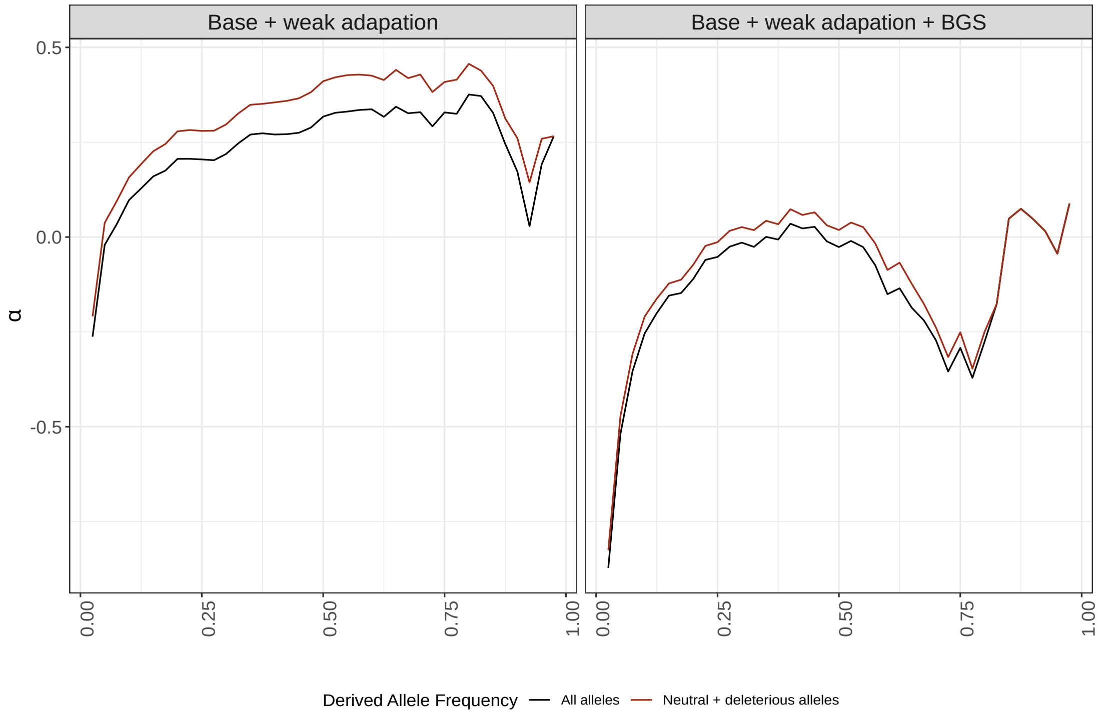


Figure B.6: BGS and weak adaptation effect on . We added weak adaptation and BGS to the baseline simulation. Weakly beneficial variants tend to segregate along with SFS. Therefore impMKT cannot deal with weak adaptation, although including another cutoff at higher frequencies.

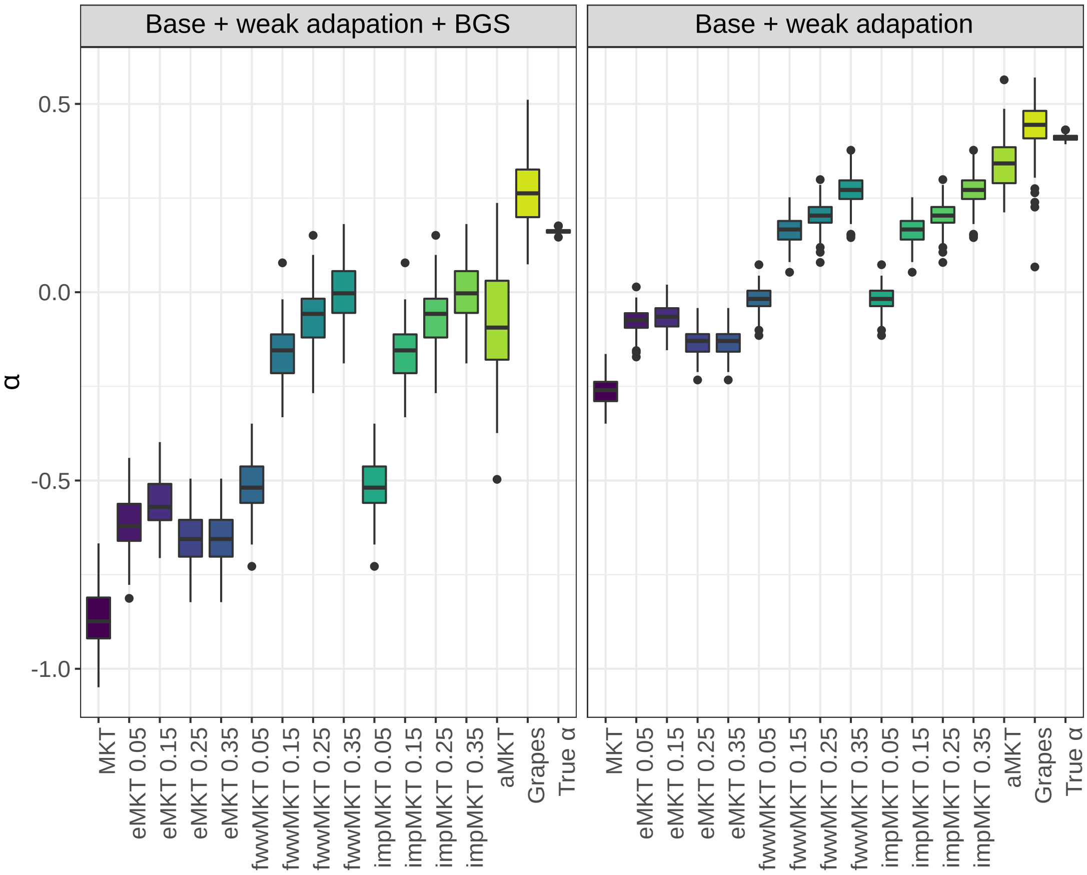


Figure B.7: α estimations at simulations accounting for weak adaptation. Any of the proposed methods can correct linkage and weak adaptation at estimations. Although the method proposed by Uricchio et al. (2019) can overcome linkage and the weak adaptation, α estimations at the gene level remain unexplored, and new approaches are required

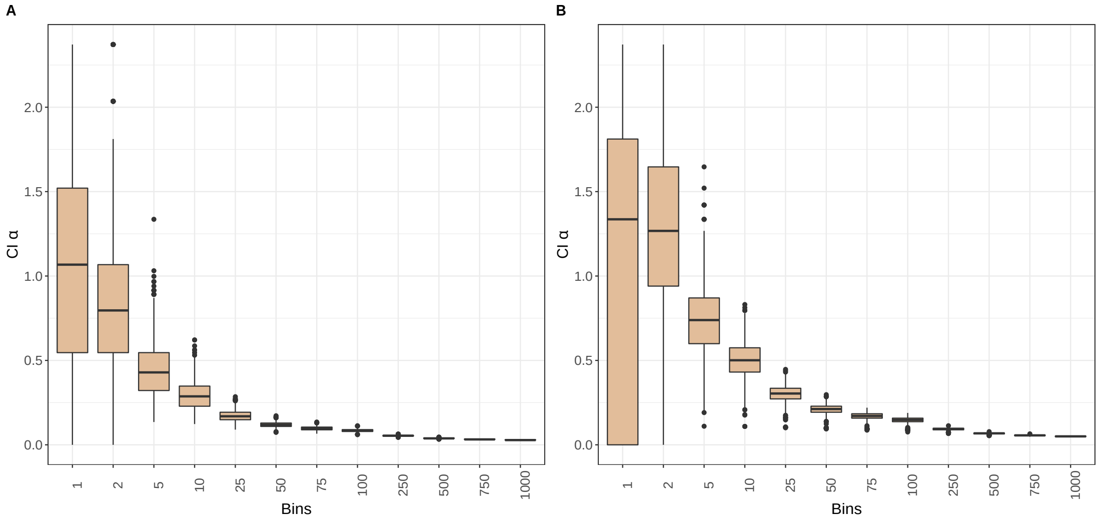


Figure B.8: Associated CI to α estimation of 1000 replicas when bootstrapping a set of 3500 random genes. A. *D. melanogaster* gene pooled dataset. B Human gene pooled dataset.

Table B.1: Mean α values for each scenario and MKT approach. Mean values were calculated using bootstrap distributions

Simulations	MKT	eMKT 0.05	eMKT 0.15	eMKT 0.25	eMKT 0.35	fwwMKT 0.05	fwwMKT 0.15	fwwMKT 0.25	fwwMKT 0.35	impMKT 0.05	impMKT 0.15	impMKT 0.25	impMKT 0.35	aMKT	Grapes	True
Baseline	-0.284	-0.086	-0.088	-0.162	-0.162	-0.019	0.167	0.229	0.251	-0.019	0.167	0.229	0.251	0.233	0.425	0.356
$2N_es- = -1000$	-0.33	-0.134	-0.12	-0.199	-0.199	-0.07	0.144	0.181	0.225	-0.07	0.144	0.181	0.225	0.243	0.406	0.301
$2N_es- = -500$	-0.423	-0.22	-0.212	-0.28	-0.28	-0.152	0.06	0.128	0.188	-0.152	0.06	0.128	0.188	0.269	0.361	0.271
$2N_es+ = 100$	-0.63	-0.385	-0.364	-0.457	-0.457	-0.306	-0.044	0.023	0.06	-0.306	-0.044	0.023	0.06	0.071	0.283	0.177
$2N_es+ = 500$	0.033	0.18	0.175	0.12	0.12	0.235	0.38	0.42	0.439	0.235	0.38	0.42	0.439	0.438	0.571	0.485
Genes 2000	-0.241	-0.037	-0.01	-0.102	-0.102	0.024	0.246	0.263	0.288	0.024	0.246	0.263	0.288	0.293	0.49	0.349
Genes 28000	-0.269	-0.072	-0.064	-0.142	-0.142	-0.006	0.203	0.248	0.285	-0.006	0.203	0.248	0.285	0.316	0.46	0.338
$\rho = 0.0001$	-0.283	-0.091	-0.08	-0.162	-0.162	-0.025	0.202	0.255	0.278	-0.025	0.202	0.255	0.278	0.295	0.461	0.34
$\rho = 0.01$	-0.239	-0.054	-0.044	-0.111	-0.111	0.003	0.192	0.259	0.283	0.003	0.192	0.259	0.283	0.285	0.454	0.344
$\beta = 0.1$	-0.114	-0.064	-0.065	-0.075	-0.075	-0.047	-0.001	0.027	0.053	-0.047	-0.001	0.027	0.053	0.055	0.123	0.111
$\beta = 0.2$	-0.212	-0.098	-0.08	-0.125	-0.125	-0.061	0.087	0.124	0.157	-0.061	0.087	0.124	0.157	0.224	0.298	0.21
$\theta = 0.0001$	-0.254	-0.061	-0.02	-0.12	-0.12	-0.003	0.244	0.327	0.293	-0.003	0.244	0.327	0.293	0.189	0.479	0.339
$\theta = 0.01$	-0.306	-0.116	-0.144	-0.214	-0.214	-0.022	0.181	0.23	0.25	-0.022	0.181	0.23	0.25	0.256	0.427	0.28
$\alpha = 0.1$	-0.821	-0.547	-0.536	-0.619	-0.619	-0.46	-0.198	-0.079	-0.006	-0.46	-0.198	-0.079	-0.006	0.06	0.199	0.08
$\alpha = 0.7$	0.317	0.415	0.409	0.373	0.373	0.452	0.548	0.571	0.589	0.452	0.548	0.571	0.589	0.596	0.678	0.63

Table B.2: Standard deviation value for each scenario and MKT approach

Simulations	MKT	eMKT 0.05	eMKT 0.15	eMKT 0.25	eMKT 0.35	fwwMKT 0.05	fwwMKT 0.15	fwwMKT 0.25	fwwMKT 0.35	impMKT 0.05	impMKT 0.15	impMKT 0.25	impMKT 0.35	aMKT	Grapes	True
Baseline	0.0253	0.0224	0.0231	0.0245	0.0245	0.0217	0.0221	0.0239	0.0282	0.0217	0.0221	0.0239	0.0282	0.0448	0.0184	0.0075
$2N_es- = 1000$	0.0264	0.0247	0.0243	0.0252	0.0252	0.0244	0.0243	0.0263	0.0271	0.0244	0.0243	0.0263	0.0271	0.0364	0.0219	0.0058
$2N_es- = 500$	0.0254	0.0231	0.0217	0.0227	0.0227	0.0228	0.0197	0.0220	0.0228	0.0228	0.0197	0.0220	0.0228	0.0417	0.0189	0.0053
$2N_es+ = 100$	0.0201	0.0188	0.0181	0.0190	0.0190	0.0187	0.0175	0.0191	0.0217	0.0187	0.0175	0.0191	0.0217	0.0327	0.0155	0.0067
$2N_es+ = 500$	0.0407	0.0350	0.0362	0.0369	0.0369	0.0336	0.0337	0.0347	0.0338	0.0336	0.0337	0.0347	0.0338	0.0452	0.0273	0.0068
Genes 2000	0.0727	0.0647	0.0630	0.0667	0.0667	0.0629	0.0559	0.0589	0.0608	0.0629	0.0559	0.0589	0.0608	0.1071	0.0480	0.0197
Genes 28000	0.0170	0.0154	0.0152	0.0161	0.0161	0.0150	0.0151	0.0166	0.0193	0.0150	0.0151	0.0166	0.0193	0.0307	0.0130	0.0046
$\rho = 0.0001$	0.0281	0.0249	0.0245	0.0261	0.0261	0.0242	0.0224	0.0252	0.0263	0.0242	0.0224	0.0252	0.0263	0.0430	0.0203	0.0065
$\rho = 0.01$	0.0268	0.0235	0.0240	0.0251	0.0251	0.0227	0.0224	0.0233	0.0251	0.0227	0.0224	0.0233	0.0251	0.0395	0.0185	0.0070
$\beta = 0.1$	0.0171	0.0160	0.0155	0.0164	0.0164	0.0159	0.0163	0.0183	0.0194	0.0159	0.0163	0.0183	0.0194	0.0273	0.0156	0.0021
$\beta = 0.2$	0.0237	0.0218	0.0217	0.0219	0.0219	0.0215	0.0214	0.0217	0.0216	0.0215	0.0214	0.0217	0.0216	0.0506	0.0187	0.0046
$\theta = 0.0001$	0.0858	0.0740	0.0702	0.0776	0.0776	0.0716	0.0599	0.0629	0.0749	0.0716	0.0599	0.0629	0.0749	0.1342	0.0559	0.0236
$\theta = 0.01$	0.0103	0.0094	0.0097	0.0101	0.0101	0.0090	0.0094	0.0104	0.0125	0.0090	0.0094	0.0104	0.0125	0.0165	0.0084	0.0019
$\alpha = 0.1$	0.0395	0.0343	0.0340	0.0343	0.0343	0.0333	0.0316	0.0302	0.0304	0.0333	0.0316	0.0302	0.0304	0.0541	0.0255	0.0045
$\alpha = 0.07$	0.0122	0.0106	0.0112	0.0114	0.0114	0.0103	0.0115	0.0127	0.0137	0.0103	0.0115	0.0127	0.0137	0.0266	0.0092	0.0054

Table B.3: Mean error bias for each scenario and MKT approach. Error bias were measured through the difference of mean values of d, d_w and d_0 for each method and the true value retrieved from `SLiM`.

Analysis	ImpMKT 0.05			ImpMKT 0.05		
	d	d_w	d_0	d	d_w	d_0
Baseline	4.43E-10	4.53E-02	4.51E-02	3.52E-10	8.08E-03	7.89E-03
$2N_es- = 1000$	5.40E-10	5.67E-02	5.64E-02	2.09E-10	1.02E-02	9.93E-03
$2N_es- = 500$	4.98E-11	7.00E-02	6.97E-02	1.23E-10	7.50E-03	7.14E-03
$2N_es+ = 100$	7.12E-10	4.49E-02	4.48E-02	2.97E-10	6.04E-03	5.91E-03
$2N_es+ = 500$	7.04E-10	4.95E-02	4.93E-02	5.61E-10	1.15E-02	1.13E-02
2000 genes	4.04E-09	4.15E-02	5.09E-02	8.43E-10	5.11E-03	8.73E-02
28000 genes	4.47E-10	4.76E-02	4.75E-02	4.49E-10	6.86E-03	6.80E-03
$\rho = 0.0001$	8.83E-10	4.92E-02	4.89E-02	1.88E-10	6.91E-03	6.53E-03
$\rho = 0.01$	2.83E-10	4.35E-02	4.32E-02	1.33E-10	4.95E-03	4.63E-03
$\beta = 0.1$	3.67E-10	4.99E-02	4.82E-02	3.03E-10	4.53E-03	2.83E-03
$\beta = 0.2$	8.56E-11	5.42E-02	5.33E-02	3.41E-10	4.20E-03	3.23E-03
$\theta = 0.0001$	4.93E-09	4.37E-02	1.26E-01	1.12E-08	4.31E-03	1.65E-01
$\theta = 0.01$	3.51E-11	6.77E-02	6.77E-02	9.04E-11	2.41E-02	2.40E-02
$\alpha = 0.1$	3.37E-10	4.52E-02	4.46E-02	6.15E-10	2.38E-03	1.86E-03
$\alpha = 0.7$	3.14E-10	5.25E-02	5.33E-02	7.40E-12	1.53E-02	1.61E-02

Table B.4: Mean α values from simulations accounting for weak adaptation and BGS

Simulations	MKT	eMKT 0.05	eMKT 0.15	eMKT 0.25	eMKT 0.35	fwwMKT 0.05	fwwMKT 0.15	fwwMKT 0.25	fwwMKT 0.35	impMKT 0.05	impMKT 0.15	impMKT 0.25	impMKT 0.35	aMKT	Grapes	True
Baseline	-0.28	-0.09	-0.09	-0.16	-0.16	-0.02	0.17	0.23	0.24	-0.02	0.16	0.23	0.25	0.24	0.38	0.35
Baseline + weak adaptation	-0.26	-0.08	-0.07	-0.13	-0.13	-0.02	0.16	0.20	0.27	-0.02	0.16	0.20	0.27	0.35	0.43	0.41
Baseline + weak adaptation + BGS	-0.87	-0.62	-0.56	-0.65	-0.66	-0.52	-0.16	-0.06	-0.003	-0.52	-0.16	-0.06	-0.003	-0.01	0.27	0.16

Appendix C

iMKT: the integrative McDonald and Kreitman test - Supplementary material

Table C.1: Number of lines resampled in each Drosophila population available at iMKT

Population	**Number of resampled lines**
RAL	160
USI	15
USW	27
CO	9
EA	10
EF	25
EG	10
GA	7
RG	21
SP	20
SD	30
ZI	154
CHB	12
FR	70
NTH	15
AUS	14