LEVERAGING BIG DATA FOR BIOINFORMATIC ANALYSIS IN MODERN POPULATION GENOMICS.

by

WALTER WOLFSBERGER

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN BIOLOGICAL AND BIOMEDICAL SCIENCES

2023

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:

Dr. Taras Oleksyk, Ph.D., Chair
Dr. Fabia Batistuzzi, Ph.D.
Dr. Randal Westrick, Ph.D.

To my parents, Maria and Walter, whose love and friendship influenced me and my choices.

To my wife, Karina, for her love, support, and patience.

To my colleagues and friends, the ones who have shared this journey with me.

And to all the brave Ukrainians who stand defiantly in the face of Russian invasion.

Слава Україні!

ACKNOWLEDGMENTS

I want to express my deepest gratitude to the following individuals and organizations whose support, guidance, and contributions have been instrumental in the completion of this thesis/dissertation:

My academic advisor, Dr. Taras Oleksyk, for helping me rediscover the wonders of discovery and research. Your mentorship and support are the high bar I will measure successful scientists, mentors, and friends with. I would also like to thank my committee members, Dr. Fabia Batistuzzi and Dr. Randal Westrick, for their advice and support throughout my time at OU.

My friends and colleagues in our lab at OU - Kenneth, Khrystyna, and Stephanie, for offering support, encouragement, and valuable discussions. Your friendship made the journey more enjoyable, meaningful, and fun.

Everyone in the Department of Biology at OU and the Bioengineering Department at UPRM whose knowledge and insights have helped me to grow as a scientist.

All the funding agencies and organizations that have contributed to our research. To BGI, Helmsley Foundation, Horizon 2020 EU program for donating services and funding our projects.

Lastly, I extend my gratitude to the broader academic community and scholars whose works have paved the way for this research.

Thank you for your unwavering support, encouragement, and trust in my abilities. Your contributions have made this achievement possible.

Walter Wolfsberger

ABSTRACT

LEVERAGING BIG DATA FOR BIOINFORMATIC ANALYSIS OF MODERN POPULATION GENOMICS.

by

Walter Wolfsberger

Adviser: Dr. Taras Oleksyk, Ph.D.

The technological advancements and the cost reduction of genomic sequencing provide novel capabilities to pose and answer biological questions on a grand scale. The initial efforts of establishing genomic references for many species of the globe serve as a foundation for the projects that aim to analyze the populations of said species. This expansion of the number of samples involved in individual studies is often connected with increased infrastructural costs for data storage and analysis.

Population genomics methods are expanding on the existing genetic approaches, leveraging our ability to automate big data processing and introducing comparative methods to our collection of scientific instruments. It has extensive applications in animal and wildlife research, conservation, and human population analyses.

These unique opportunities are associated with emerging challenges related to the nature of the approaches and their relative novelty in the field. Analysis of a multitude of individuals often increases requirements in terms of bioinformatics expertise and resources. An increase in complexity and data volume means researchers often need access to high-performance computing facilities and specific training to utilize them. The field

still grows, with new instruments or tests frequently introduced and outdated approaches depreciating.

This work reviews population genomics methods and their application related to wildlife research, conservation, and human populations research, employs them to provide answers in multiple studies, and presents a newly developed analysis suite. The suite is aimed to facilitate reproducible, accessible population genomic testing across various fields of application, seeking to address the growing expertise challenges in the field.

TABLE OF CONTENTS

ACKNOWLEDGMENTS i

ABSTRACT v

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xiii

CHAPTER ONE
BACKGROUND AND SIGNIFICANCE 1

1.1 Introduction to population genomics 1

1.2 Challenges of population genomics 3

CHAPTER TWO
POPULATION GENOMICS IN ANIMAL SPECIES AND CONSERVATION 8

2.1 Introduction to population genomics for animal species and conservation 8

2.2 Prerequisites to population genomics analysis 10

2.3 Methods that infer population structure and history 12

2.3.1 Principal Component Analysis (PCA) and linkage disequilibrium (LD) pruning of genomic data 12

2.3.2 Admixture/Structure: model approaches for estimation of ancestry in unrelated individuals 16

2.3.3 Runs of Homozygosity(ROH) and Inbreeding Analysis via the Froh Statistic 20

2.4 Methods for detection of signatures of selection along the genome using population data. 24

TABLE OF CONTENTS—Continued

2.4.1 Extended haplotype homozygosity(EHH), site-specific Extended Haplotype Homozygosity(EHHS), and variant call data phasing 25

2.4.2 Single population selection signature - Integrated Haplotype Homozygosity Score(iHS) 28

2.4.3 Pairwise population selection signature - ratio of EHHS between populations(Rsb) and cross-population EHH (XP-EHH) 29

2.5 Case study: Application of population genomics to infer the genetic diversity and selection in Puerto Rican horses 31

2.5.1 Background 31

2.5.2 Methodology 33

2.5.3 Results 38

2.5.4 Discussion 46

2.5.5 Conclusion 50

CHAPTER THREE
POPULATION GENOMICS IN HUMAN POPULATIONS 52

3.1 Introduction to human population genomics 52

3.2 Databases as a force multiplier in human population research 54

3.3 Computational burden of human population genetics 56

3.4 Case study: genomic deserts, blank spots on the map of genome diversity in Europe 57

3.4.1 Background 58

3.4.2 Discoveries from sequencing of underrepresented population 59

TABLE OF CONTENTS—Continued

3.4.3 Benefits, challenges, and conclusions 61

3.5 Employing population genomics analysis to characterize Genome diversity in Ukraine 62

3.5.1 Background 63

3.5.2 Methodology 64

3.5.3 Results 68

CHAPTER FOUR
KNOWLEDGE DISSEMINATION AND DEVELOPMENT OF ACCESSIBLE POPULATION GENOMICS ANALYSES SOFTWARE TOOL 75

4.1 Bioinformatics as the backbone for modern genomics 75

4.2 Scalable, reproducible workflows: Snakemake overview 76

4.3 PopGen Playground – accessible workflow for population genomics analysis 77

4.4 Bioinformatics knowledge dissemination 78

4.5 Supporting scientists worldwide - scientists without borders: lessons from Ukraine: 80

4.5.1 Background 80

4.5.2 Summary of the global response to the war in Ukraine 82

4.5.3 Analysis of opportunities for Ukrainian scientists listed in the #ScienceForUkraine 84

4.5.4 Effective mechanisms of remote support scientists in countries facing hardship 87

CHAPTER FIVE
OVERALL CONCLUSIONS AND FUTURE DIRECTIONS 89

5.1 Conclusions 89

TABLE OF CONTENTS—Continued

5.2 Future directions 90

REFERENCES 91

APPENDIX 107

LIST OF TABLES

Table 1	Summary of variation in the 97 whole-genome sequences from Ukraine	70
Table 2	Analyses incorporated in PopGen Playground suite.	78
Table 3	Mechanisms to support scientists and students from countries experiencing hardship	88

LIST OF FIGURES

Figure 1	Growth of DNA sequencing.	4
Figure 2	Human population structure within Europe.	15
Figure 3	Genetic Structure of Human Populations.	17
Figure 4	Demographic origins of ROH.	22
Figure 5	Phasing visual guide.	26
Figure 6	EHH plot around a focal marker rs6 for two alleles.	28
Figure 7	Whole genome iHS plot for cattle breed population.	29
Figure 8	Distance-based, neighbor joining tree calculated from SNP frequencies in 38 horse populations.	35
Figure 9	Principal component analysis plot of the merged dataset of horse breeds.	36
Figure 10	Admixture graph of the shared dataset of horse breeds	40
Figure 11	Runs of homozygosity (ROH) in the genomes of the two Puerto Rican horse breeds	41
Figure 12	Integrated extended haplotype homozygosity scores (iHs)	42
Figure 13	Extended haplotype homozygosity (EHH) decay and shared haplotype length	43
Figure 14	Signatures of selection based on the ratio of EHHs between populations	45
Figure 15	Public availability of whole-genome sequences in Europe	60
Figure 16	Principal component (PC) analysis of merged dataset, containing European populations.	72
Figure 17	Admixture of ancestral components for Ukrainian and neighboring populations	73

Figure 18 Locations of opportunities for Ukrainian scientists in the #ScienceForUkraine database in one year since the start of the conflict 85

LIST OF ABBREVIATIONS

AIMs	Ancestry Informative Markers
DMRT3	Doublesex and Mab-3 Related Transcription Factor 3
DNA	Deoxyribonucleic Acid
ECA23	Equus caballus chromosome 23
EDTA	Ethylenediaminetetraacetic Acid
EHH	Extended Haplotype Homozygosity
EHHS	Site-specific Extended Haplotype Homozygosity
Froh	Inbreeding based on ROH
GRCh38	Genome Reference Consortium Human Build 38
GWAS	Genome-Wide Association Studies
HPC	High Performance Computing
HW	Hardy-Weinberg equilibrium
IACUC	Institutional Animal Care and Use Committee
iHH	Integrated Extended Haplotype Homozygosity Score
iHS	Integrated Haplotype Score
InDel	Insertion/Deletion variant
LD	Linkage Disequilibrium
MCMC	Markov chain Monte Carlo
MEI	Mobile Element Insertion
mtDNA	Mitochondrial DNA
NROH	Total Number of Runs of Homozygosity
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
pVCF	Processed Variant Call Format
PRPF	Puerto Rican Paso Fino horse breed
PR NPB	Puerto Rican Non Pure-Bred horse breed(“Criolo”)
QC	Quality Control
r2	Coefficient of Determination
ROH	Runs of Homozygosity
Rsb	Ratio of Extended Haplotype Homozygosity between populations
SNP	Single Nucleotide Polymorphisms
SROH	Total Length of Runs of Homozygosity
TITV	Transition/Transversion Ratio
UPRM	University of Puerto Rico at Mayagüez
VCF	Variant Call File
WGS	Whole Genome Sequencing
XP-EHH	Cross-Population Extended Haplotype Homozygosity

CHAPTER ONE

BACKGROUND AND SIGNIFICANCE

1.1 Introduction to population genomics

The Human Genome Project, launched in 1990, stands as a turning point in modern genomics. Initially focused on our own genome, its impact has been felt across a multitude of species(Hood & Rowen, 2013). Immediately after, the sequencing of several model organisms, including the fruit fly (*Drosophila melanogaster*) and the mouse (*Mus musculus*), produced important comparative datasets(Abzhanov et al., 2008). These advances sparked a surge in genome sequencing efforts across all life forms. Notable efforts include the 1,000 Genomes Project(Birney, 2021), which aimed to document human genetic variation thoroughly, and the Genome 10K project(Haussler et al., 2009), which sought to sequence the genomes of 10,000 vertebrate species. The genomes of many crop species, such as rice and maize, have been sequenced to benefit agricultural practices. Today, genomics remains a vital aspect of numerous biological disciplines, ranging from ecology to medicine, and initiated an ambitious project to sequence the genomes of all known species - the Earth BioGenome Project(Lewin et al., 2018).

Population genomics is a rapidly growing discipline that traces its origins to the history of genetics research, along with a combination of technological advancements and the development of modern analysis methods. The original genetics studies established the foundation for our understanding of genetic variation and its implications, often focusing on individual genes or markers. Since then, it has gradually evolved, driven by the technological advancements in genomic sequencing and analytical methods, to a point at

which we can approach genetic variation studies on the population scale.(Luikart et al., 2003)

Population genomics aims to understand the distribution and causes of genetic diversity among and within populations, including the various factors contributing to these variations. These factors include demographic elements such as population history, age, and size, as well as evolutionary forces like mutation, migration, selection, nonrandom mating, and genetic drift. (Romiguier et al., 2014)

The rise of high-throughput sequencing is crucial in shaping the modern genetic research landscape. Access to massive whole genome sequencing (WGS) technology did not only provide us with the ability to answer more questions but also deepened our understanding of the complex mechanisms and interactions within the genomes of multitudes of species. WGS data provides access to the whole organism's genome, including coding and non-coding regions, and allows a more comprehensive ability to assess the whole breadth of the genetic variation(Rhie et al., 2021). It is crucial for the discovery of the genetic basis of complex traits, which are often connected to structural changes and non-coding variation. Multiple animal and wildlife projects were able to generate advanced insights and propagate more advanced projects by generating reference genomes. A reference sequence is fundamental for population analyses, as it establishes a standardized framework for comprehending variation and allows us to detect variation by comparing two or more individuals within a population. This is particularly important in animal genomics as it enables the exploration of population history, detect signatures of adaptation, and develop conservation tools. (Ekblom et al., 2018; Gordon et al., 2016; Johnson et al., 2018; Oleksyk et al., 2010) In this context, human research is ahead, as there

is a comprehensive genome reference, and multiple populations were already sequenced and analyzed. However, these efforts are biased by the "pioneer advantage" – where countries that had their populations sequenced first are benefitting from all the breadth of genomic advancements disproportionately to countries that were unable to afford such initiatives due to various reasons.(Oleksyk et al., 2022) Together, the technological and analytical advancements have enabled the investigation of complex genetic patterns and processes, positioning population genomics as one of the leading fields in modern biological research.

1.2 Challenges of population genomics

While the adoption of population genomics techniques in modern research comes with many opportunities to better our understanding of the genetic basis for complex traits and mechanisms that shape variation, they also come with their own challenges shaped by the same advancements in technology and analysis. Population studies rely on processing large volumes of complex data for a growing number of samples, and their generation in genomics is growing faster than humanity's other significant information domains: astronomy, social media, and video hosting. According to the estimates of Illumina company (Illumina Inc., San Diego, CA), the genome sequencing rate doubles approximately every twelve months, and if it continues at that rate, we will venture into exabytes of bases sequenced and stored per year (Figure 1). In this sense, genomics has already entered the Big Data domain, where large-scale projects require specific data handling techniques to process the information, organize, and maintain its coherency.

By comparison, the computation-oriented Moore's law suggests the doubling of the computational capacity rate to be 18 months. This means that we might reach a moment

when we will be bottlenecked by our ability to generate insight from the data produced in the near future. This volume and complexity of genomic data production presents significant analytical challenges, promoting the need for further advancements and optimization of the analytical approaches. (Stephens et al., 2015a)

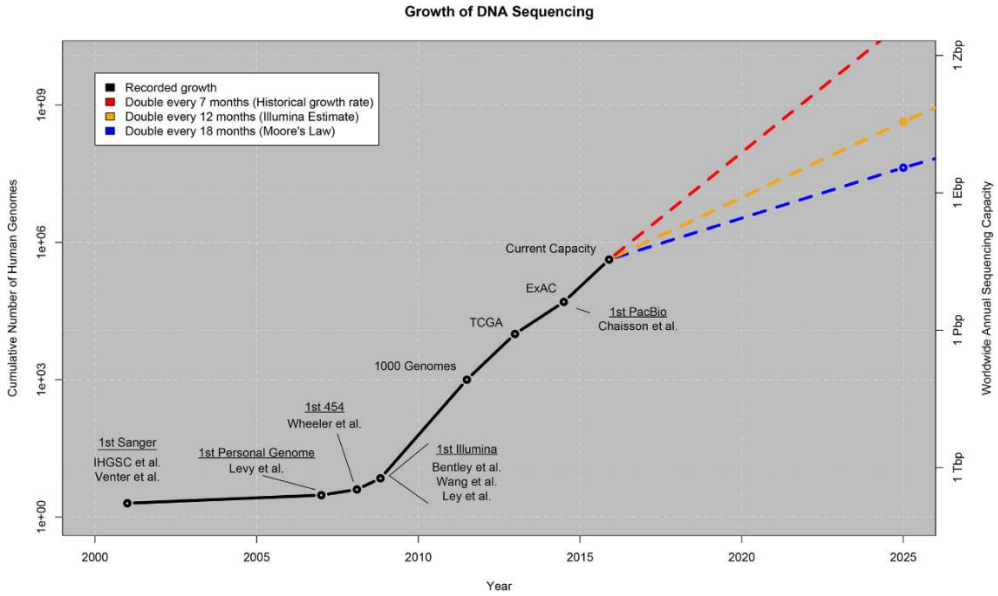

Figure 1. Growth of DNA sequencing.
The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)). The yellow values beyond 2015 represent the Illumina projection of growth.

Population genomics routinely intersects with several other biological domains due to its broad applicability in understanding genetic variation within and among populations. Evolutionary biology, ecology, conservation, biogeography, medicine, and public health can integrate the inferences of population genomics into their findings. The data analysis in population genomics often incorporates multiple heterogeneous steps, a problem that becomes even more critical in projects intersecting with other biological domains. The same analytical test is often incorporated in multiple packages, each with unique hardware requirements. It requires advanced computational expertise, which may not be available in

all research settings. While data analyses should ideally be conducted in a reproducible way, a significant portion of intermediate data manipulations or non-analysis-related steps like plotting are often omitted from the methodology section of manuscripts.(Fuller et al., 2013). This creates limitations regarding the validation and recreation of the results on the original data. Additionally, it means that existing studies often can not be used to analyze new data. This issue is not exclusive to population genomics and is encountered in every evolving discipline that relies on information volume to generate insight.

To address these issues, the data analysis community is developing the concept of sustainable data science, which consists of guidelines and instruments that serve the community. Designed to be employed at the data analysis stage, they ensure that questions of reproducibility, scalability, and transparency are given due attention. With this aim, a practical solution lies in implementation of automated workflow management systems such as Snakemake (Köster et al., 2021). With its flexible framework, complex bioinformatics pipelines can be created to handle large genomic datasets, and enables researchers to develop scalable and reproducible data analyses. Snakemake workflows are written in Python programming language, allowing for easy integration with other widely used Python libraries and tools (Karthikeyan & Jose, 2021). They are highly supported, portable, and can be deployed on various computing platforms, from laptops to high-performance computing clusters. (Mölder et al., 2021)

In summary, as sequencing technologies and computational throughput become more accessible to research groups, the challenge lies in streamlining the effort and expertise required to analyze the datasets produced by sequencing(Stephens et al., 2015b). Bioinformatics faces heterogeneity in technical and software solutions, coupled with the

lack of strict data storage standards, which produces more information than can be effectively interpreted. The relative novelty of this field often means a lack, or sometimes an abundance, of "best practices" and standards capable of processing the data, which contributes to the confusion for scientists entering the field and having basic bioinformatics tools but lacking access to advanced bioinformatics expertise. This lack of standards results in significant amounts of time spent on data management and wrangling instead of actual research.(Fuller et al., 2013)

To promote the availability of population genomics projects to the broader scientific community, concentrated effort must be directed towards producing streamlined, reduced-effort repeatable data analysis and handling solutions. Significant progress in this area will not only provide opportunities for smaller-scale projects but also lift the analysis burden, freeing up time and expertise that can be used to work on other aspects of the research.

The chapters of this dissertation are structured to provide an in-depth understanding of population genomics for animals and humans and the significance of bioinformatics in this field. Chapter two lays the groundwork by addressing prerequisite analyses that enable population genomics and describing the analyses in the field. It concludes with the application of these concepts in a case study of Puerto Rican horses. Chapter three introduces population genomics in application to human populations, exploring its unique aspects. It covers topics like the power of established databases, computational challenges related to the wide adoption of WGS techniques, and biases in existing genomic studies. It concludes with a thorough description of the analysis of genome diversity in Ukraine. Chapter four emphasizes the role of bioinformatics in modern genomics, the development

of accessible workflow software, and the importance of knowledge dissemination. It also explores the support of scientists in peril, exemplified by the situation in Ukraine. Chapter 5 provides overall conclusions and describes the future directions of our research.

CHAPTER TWO

POPULATION GENOMICS IN ANIMAL SPECIES AND CONSERVATION

This chapter examines the role of population genomics in the research of animal species conservation. It will cover various analyses and their practical applications and highlight a first-author paper that utilizes many of these techniques.

2.1 Introduction to population genomics for animal species and conservation

Every species possesses an exclusive program – the genome, that evolves through natural selection and other evolutionary processes, shaped by its surroundings or ecological niche. Each species maintains a unique habitat, and the genetic code is a constantly rearranging combination of code that reflects its demography and adaptation to the niche. While some species can adapt to changing environments, others cannot keep pace and ultimately face extinction. Genetic diversity is achieved through a balance of mutation, migration and drift, contributing to diversity and enabling adaptation through natural selection. (Ellegren & Galtier, 2016; Oleksyk et al., 2010). Population genomics is a valuable instrument in the study of animal species genomes and is effectively used to answer questions related to both domesticated animals and wildlife. It is used to identify the genetic variation between species, within and between populations, offers insights into evolutionary processes, adaptation, and biodiversity, and has a variety of applications. (Bourgeois & Warren, 2021)

The advancements in the availability of WGS have contributed to our ability to infer the histories of species, even with little prior knowledge, purely based on genomic information. The discovery of variation in multiple genomes across whole populations or

related species became a standard method of inferring the evolutionary processes(Bourgeois & Warren, 2021). This process has expanded our knowledge base from the initially limited number of model species(Abzhanov et al., 2008), which became the foundation for further research. The scope of genomic studies in animal species and conservation proliferated in recent years, with studies covering a variety of topics such as exploration of the origins and history of domesticated species (Kavar & Dovč, 2008), the impact of domestications on a species(Axelsson et al., 2013), the genetic basis of local adaptation(Tsartsianidou et al., 2021), and even adaptation related to the environmental effects of climate change (Axelsson et al., 2013).

Despite its lower costs, genomic research still requires significant resources such as time, expertise, and computational power. Understanding each genome is a complex task that involves delving deep into its structure, function, and evolution. This task is even more challenging when studying multiple species. Therefore, researchers tend to prioritize eukaryotic species of high interest or importance (Del Campo et al., 2014). This includes "charismatic" species like pandas and whales, which can attract public support and funding for genomic research(Árnason et al., 2018; H. Fan et al., 2019). Economically important species such as crop plants, livestock animals, and medicinal species are also prioritized for their practical benefits in enhancing agricultural productivity, disease resistance, and drug development (Haberer et al., 2005; Tsartsianidou et al., 2021). However, this narrow focus on a limited number of species leaves many other species unstudied, resulting in gaps in our understanding of biodiversity and evolution. As genomic research progresses, the challenge will be to expand the scope of research to include a broader range of species.(Lewin et al., 2018; McGuire et al., 2020)

By their nature, the population genomic analysis methods can be divided into two categories: those that aim to infer population structure and history and those which are covering the detection of signatures of evolutionary processes along the genome in the population. Ideally, a typical population genomics study would aim to employ and compare the results of various methods from both categories to reach a comprehensive answer on the processes investigated. It is important to note that multiple methods and tools provide similar answers using different signals or input information. If the results of those approaches generally agree – it gives the research group a better idea of the robustness of the discoveries. Otherwise, their disagreement can be valuable to reach a meaningful conclusion regarding the methodology, update the prior hypothesis, and provide areas of additional interest for future research. (Bourgeois & Warren, 2021)

2.2 Prerequisites to population genomics analysis

While this dissertation focuses on population genomics, it is essential to provide an overview of the fundamental prerequisites for these projects in the broader realm of genomics. To successfully proceed to the step of analyzing populations based on Whole Genome Sequencing(WGS) data, a series of important preliminary data collection and analysis steps are required. These preliminary steps serve as groundwork, laying the foundation for a comprehensive and accurate exploration of population genomics.

Reference genome: Population genomics inferences ultimately rely on our ability to record variation in individuals. The reference genome is a representative template that serves as a standard for comparing and analyzing genetic differences in individuals. While there is no issue for projects that involve human genomes due to the presence of advanced reference sequence(Nurk et al., 2022), many animal species still lack one. In this case, the

ideal approach would be assembling the genome in question, and the proliferation of WGS technologies allows many conservation projects to follow that strategy. (Formenti et al., 2022). Alternatively, an assembly of a closely related species can be used as a substitute, but, depending on the study's goals and desired sensitivity, it might not provide acceptable quality results.(Valiente-Mullor et al., 2021).

Variant calling is a data analysis step that involves identifying genetic variants, such as single nucleotide polymorphisms (SNPs) and structural variations, within an individual's genome. Variant calling requires the presence of a reference genome. Individuals' genomic sequences are mapped to the reference, and places of difference are recorded. (Van der Auwera et al., 2013a) This step is usually storage and computation intensive, as a single human sample in this step would have a file size of approx. 100 Gbytes.(Stephens et al., 2015a).

Phylogenetic analysis is essential for understanding the genetic relationships among various species or populations. Phylogenetics is the investigation of evolutionary relationships, and the tests allow researchers to create trees, networks, or cladograms that represent these relationships based on genetic data. (Yang & Rannala, 2012) These cladograms are constructed using different techniques, such as distance-based methods, maximum likelihood methods, and Bayesian inference. Various data types are suitable for phylogenetic analysis - such as genomic sequence alignments, or files containing information on genetic variation, each having unique advantages and constraints.(Yang & Rannala, 2012). Generally, phylogenetics focuses on relationships among species. However, it can be applied at the population level, where gene flow between populations is minimal or non-existent. For example, phylogenetic trees can guide conservation efforts

by identifying genetically distinct populations and help prioritize conservation resources and strategies. For example, a genetically distinct population may represent a unique evolutionary lineage worth preserving. (Xue, 2022) In the presence of significant gene flow, phylogenetic trees are often replaced by genetic networks that can incorporate additional information about genetic connections within the population. For many population genomics studies, phylogenetic analyses are often the initial step. They enable researchers to explore the genetic relationships and evolutionary history of different species or populations. However, as whole genome data continues to grow in volume and complexity, their importance in our understanding of biodiversity and evolution is only increasing. (Rannala & Yang, 2008)

2.3 Methods that infer population structure and history

2.3.1 Principal Component Analysis (PCA) and linkage disequilibrium (LD) pruning of genomic data.

Principal Component Analysis (PCA) is a widely used and powerful statistical method for reducing the dimensionality of independent variables in large datasets in an interpretable way to prevent significant data loss. It is an adaptive technique applicable to various kinds of data, and many scientific fields use their variation of the method to perform exploratory data analysis. In most of them – it is used to identify patterns in data and express the multi-dimensional data points in such a way as to highlight their similarities and differences.(Jollife & Cadima, 2016) PCA works by finding the directions, or 'principal components'(PCs), along which the variation in the data is maximized. The first principal component is the direction that explains the most variation, the second principal component explains the second most, and so on.

In population genomics, the adaptation of PCA has become a cornerstone technique for understanding the genetic structure of populations and identifying patterns of genetic variation of individuals within and between populations(Price et al., 2006a). For example, if used for a large dataset of animal genetic data, a PCA plot could display how individuals are grouped based on their habitat or niche. This structure might reflect the history of migration, gene flow, and genetic isolation.

Typical usage of PCA involves the input of every individual variant call for every individual in the population. Before the PCA analysis, the data files are pruned for sites with missing information, those of low variability, and sites in linkage disequilibrium (LD). It means that some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies. This phenomenon is one of the most important descriptors of the variation along the chromosomes and can occur for several reasons, including physical proximity on the chromosome (linked genes are often inherited together), low rates of recombination between the loci, population bottlenecks, non-random mating, and natural selection. Finding extended LD in the genome is an important indicator of recent selection (Oleksyk et al., 2010), but including LD loci in population analysis, especially in the association studies, could bias results, and they can show up as associated with a trait or disease simply because they are linked to a different variant that is the actual causal variant. Pruning for LD can also reduce the number of tests done and thereby reduce the number of false positives.

Identification of sites with LD and their pruning are common data-wrangling tasks in population genomics. Due to the nature of the recombination process happening during

meiosis, genomic sites in the physical vicinity of the chromosome are less likely to be separated and are more likely to be inherited together. While there is an idealized hypothetical scenario where sites in LD do not occur, most populations will have some levels of population-wide LD due to factors like finite population size, genetic drift, non-random mating, selection, or population structure.(Pritchard & Przeworski, 2001) The linkage disequilibrium pruning step selects only independent markers using various methods. The most common approach in the field utilizes the coefficient of determination (r^2, a statistical measure that describes the degree of correlation between pairs of alleles at two different loci in a population) as a filter for every pair of variants within a sliding window of a specified size. Correct implementation of LD pruning leads to the retainment of only independent sites with genetic variation. (Calus & Vandenplas, 2018)

The analysis output usually contains two files: a file that reflects every identified principal component and the percentage of the total variability of the original dataset it explains per every identified principal component, from descending order, and the file with every individual score for every principal component. Due to the ordering of principal components, PC1 and PC2 will describe the dominant components that describe the structure of variation within individuals. Subsequent analysis and plotting of the results of PCA will provide insight into the structure of the samples provided. An example PCA plot is provided in Figure 2.

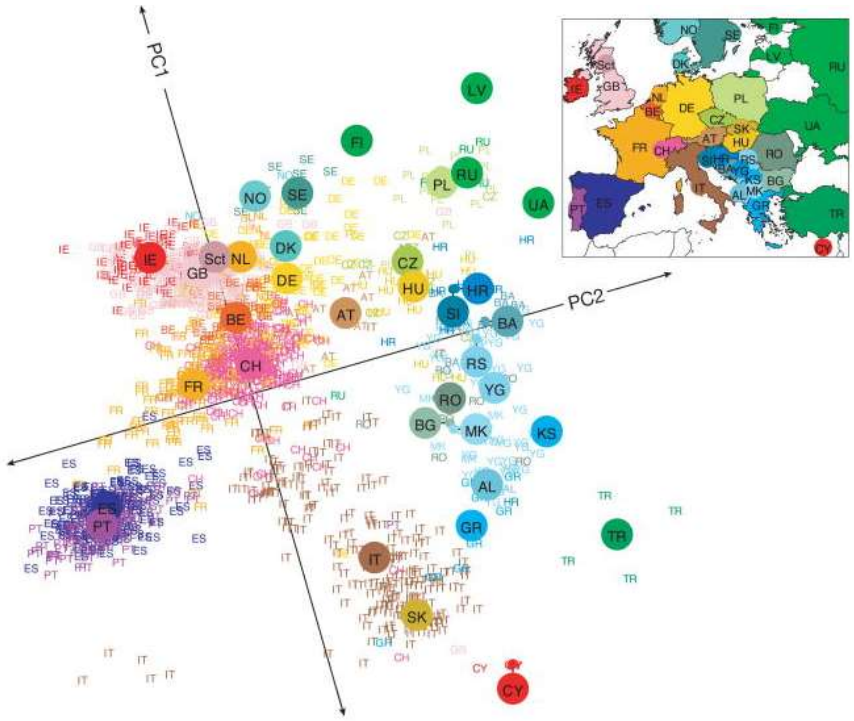

Figure 2. Human population structure within Europe.
A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small colored labels represent individuals, and large colored points represent the median PC1 and PC2 values for each country. The inset map provides the key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe. (Novembre et al., 2008)

It is important to mention that some investigators have questioned the reliability and robustness of PCA results and claimed PCA outcomes could be data artifacts, susceptible to manipulation, and potentially unfavorable in association studies (Elhaik, 2022). Moreover, PCA can be heavily influenced by population structure, heavily influenced by rare variants, which may disproportionately drive the patterns. In addition, this approach produces combinations of all the input variables assuming a linear relation between them, which is not the case and can make the results challenging to interpret in a biological, evolutionary, and historical context. Despite these limitations, PCA can still be a useful tool in population genetics if used carefully. For instance, it is a powerful exploratory tool to identify potential patterns or outliers in the data, which can then be

investigated further using other methods. Also, by being aware of these limitations, researchers can take steps to mitigate them, for example, by carefully considering the impact of population structure and rare variants in their analyses.

2.3.2 Admixture/Structure: model approaches for estimation of ancestry in unrelated individuals

The concept of admixture, referring to the genetic material mixing of separate populations due to interbreeding, plays a vital role in population genomics. Admixture is characterized as the sharing of specific gene variants (alleles) between previously isolated populations. The study of admixture is essential to understanding the genetic makeup of various species, their migration patterns, and interbreeding history. Admixture significantly impacts the genetic makeup of populations, influencing their diversity and potentially affecting their ability to adapt to environmental changes. (Y. Liu et al., 2013)

The estimation of individual proportions of ancestral components is one of the most used analyses in population genomics ever since the original publication of the genetic structure of human populations using STRUCTURE (Rosenberg et al., 2002) (Figure 2). STRUCTURE plot showing individual genomes, you would see each individual represented by a single vertical line divided into K colored segments. The lengths of the segments show the proportions of the individual's genome that are inferred to come from the K populations. The colors themselves (e.g., whether a population is shown in red or blue) do not have intrinsic meaning; they merely differentiate between the inferred ancestral populations. However, the relative proportions of colors across individuals or populations can provide insight into population structure, shared ancestry, migration events, and other aspects of evolutionary history (**Figure 3**).

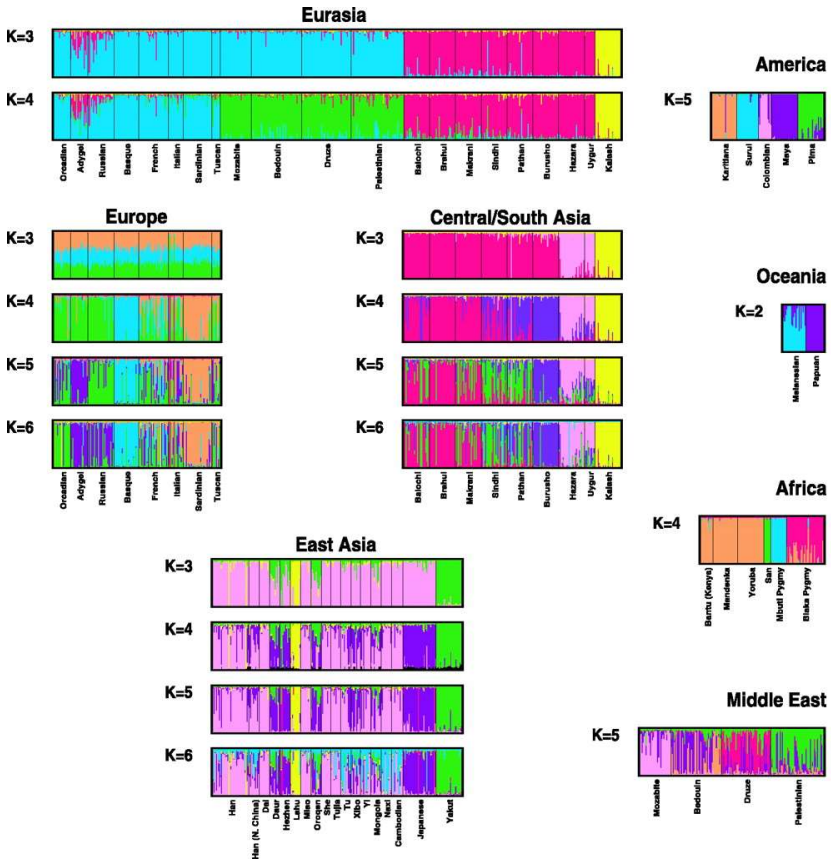

Figure 3. Genetic Structure of Human Populations.
Modified from Rosenberg et al. 2002. Each color in the plot corresponds to one of the inferred ancestral populations. The number of colors equals the number of ancestral populations inferred by the analysis (K). The height of each color in an individual's bar in the plot represents the proportion of that individual's genome inferred to originate from the ancestral population corresponding to that color. The entire bar represents the genome of an individual or a population (if individual genomes are averaged). While the relative proportions of colors across individuals or populations depend on the underlying assumptions and are subject to limitations, they can provide insight into population structure, shared ancestry, migration events, and other aspects of evolutionary history.

With all its usefulness, the admixture approach has its limitations. Deciding the number of ancestral populations (K) is challenging and can significantly impact results. Too few and important differences may be missed; too many and you may detect differences that don't have biological significance. To avoid this problem, several different K values are displayed at once (Figure 3). Admixture coefficients can be difficult to interpret biologically. They give proportions of individual genomes that come from each

of the K populations, but determining what those populations represent can be non-trivial. STRUCTURE assumes that allele frequencies in the ancestral populations are in Hardy-Weinberg equilibrium and that loci are at linkage equilibrium. These assumptions may not always hold, which can affect the accuracy of the results (Rosenberg et al., 2002).

This approach is tackled by a family of highly (Bourgeois & Warren, 2021)concordant methods of instruments, and for this research, we will concentrate on one of the most widely used instruments – ADMIXTURE. It is a model-based approach using maximum-likelihood estimation, a statistical method that estimates the parameters of a model by optimizing a likelihood function. This optimized approach makes ADMIXTURE well-suited for analyzing large-scale genomic data frequently produced in contemporary genomics studies.(Alexander et al., 2009a)

In general, ADMIXTURE calculates the ancestral proportions in individuals based on a user-defined number of ancestral populations, known as the K-value. The algorithm identifies genetic clusters based on allele frequencies in individual variant calls. It initializes ancestry proportions randomly and updates them iteratively to maximize the likelihood of observing genotypes, assuming Hardy-Weinberg equilibrium and linkage equilibrium. (Alexander et al., 2009a)

As an input, ADMIXTURE uses a file with variant calls in “ped” format, which is easily convertible using the PLINK(Chang et al., 2015) toolkit from Variant Call Files(VCFs). The data requires similar pruning as PCA methods – filtering for missing data, removal of sites with low variability, and LD pruning. To run, it also requires a given number of ancestral populations (K). The tool outputs a tabular file where every row represents an individual from the original dataset, and the following K columns are

assigned the proportions of the ancestral components, which all summarize to 1, reflecting on the admixture rates for each of the ancestral populations.

The selection of the optimal K value is a well know statistical problem that can significantly influence the interpretation of results. There is no guaranteed methodology, and the solution to K often involves analyzing various lines of evidence using a comprehensive methodology. Prior information about the population, results of PCA, or estimation of consistency between multiple runs, can guide the selection, in addition to ADMIXTURE's built-in cross-validation procedure for estimating prediction error to select the best K. Ultimately, it is important to note that there are complex admixture scenarios that any single discrete K value will not perfectly model.

A thoughtful interpretation of the ADMIXTURE results is always required.(Alexander et al., 2009a) Admixture analysis can be useful in genetic association studies to control for population stratification, a situation where allele frequencies differ between populations, which can lead to false-positive associations if not properly accounted for. Using mapping by admixture disequilibrium has shown to be a powerful tool in population-based association studies to uncover disease loci (Smith & O'Brien, 2005). However, there are also three major limitations to the current admixture approaches. First, this analysis assumes random mating within populations, which is not always accurate. Second, the correct interpretation depends on the reference populations used in the analysis. These are often absent, so the inclusion of an assumed reference population could bias admixture estimates (Oleksyk et al., 2022). If the reference populations do not accurately capture the actual ancestral populations of the individuals being studied, the admixture estimates can be inaccurate. An example of this instance is

using East Asians instead of Native Americans to estimate the Native American admixture in the Caribbean (Gravel et al., 2013). While the majority of sites change due to neutral processes, natural selection can influence the frequency of alleles in a population and could confound admixture estimates. As with any of the genomics tools, admixture analysis can be a powerful tool for understanding population history and individual ancestry, but it needs to be applied carefully, and its assumptions and limitations must be considered.

2.3.3 Runs of Homozygosity(ROH) and Inbreeding Analysis via the Froh Statistic

Runs of Homozygosity are consecutive segments of homozygous genotypes that are identical by descent, indicating they have been inherited from a common ancestor recombination breaking them up. When an individual inherits two copies of an ancestral haplotype, it results in an ROH. The length and frequency of ROH in individual genomes can provide context about the history of a population. Shorter Runs are usually ancient and indicate the shared ancestry of all individuals in the population, while long ROH are most of the time recent and demonstrate recent inbreeding or a diminished population size.(Ceballos et al., 2018)

There are multiple approaches to detect the presence of runs of homozygosity, but surprisingly, the simple sliding window approach seems to outperform the later-developed, model-based approaches, with the latter having minor advantages in the detection of the smaller ROH. The classical method for detecting Runs of Homozygosity (ROH) involves scanning each chromosome by moving a fixed-size window along its length to search for stretches of consecutive homozygous Single Nucleotide Polymorphisms (SNPs). ROHs are labeled by calculating the proportion of entirely homozygous windows that include the SNP under scrutiny.(Ceballos et al., 2018) If it meets a pre-defined threshold, the SNP is

tagged as part of ROH. The approach allows for a user-defined tolerated error rate that can ignore errors in sequencing or spurious rare new mutations from affecting the analysis. An ROH is called when a homozygous segment contains a certain number of consecutive SNPs that exceed a predetermined threshold regarding SNP count and/or chromosomal length. The specific length threshold used to identify an ROH varies depending on the study. However, it is usually set to reflect the expected length of ROH based on the population's demographic history.

The input of a typical ROH is a variant call file with multiple individuals and a set of threshold parameters to tune the sensitivity of the analysis correctly. This method outputs a list of runs of homozygosity per individual provided and their length, the percentage of SNPS in complete heterozygosity, etc. There are multiple visualization types for these tests, with the common being a scatterplot of Number vs Total length of ROH. An example plot and interpretation guidelines are demonstrated in Figure 4.

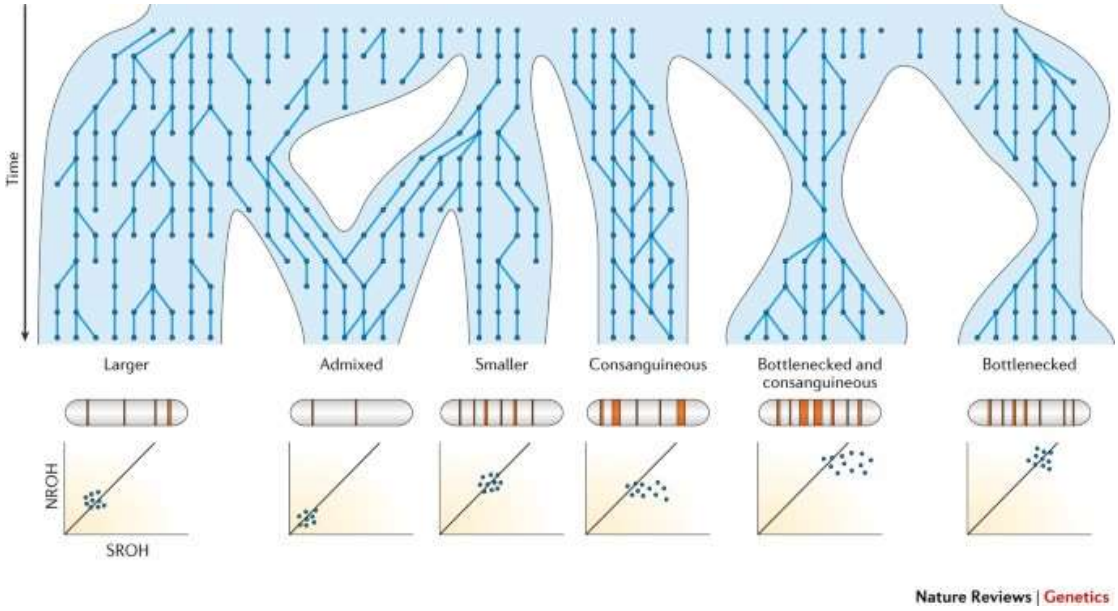

Figure 4. Demographic origins of ROH.
The demographic history of six diverse hypothetical populations is represented in the upper part of the plot. Representative pedigrees are indicated by dark blue lines connecting individuals (dots), loops show inbreeding, and the width of the light blue areas represents the population size. Thus, bottlenecks are shown by a narrowing, which necessarily reduces the number of ancestral lineages present in the population; conversely, larger populations contain more ancestral lineages. Admixture is shown by a confluence of two hitherto separate populations and mating between the pedigree lineages. The consequences of each demographic scenario are illustrated below schematic chromosomes showing the typical distribution of runs of homozygosity (ROH) in each and at the bottom, a plot of the total length of ROH (SROH) versus the total number of ROH (NROH) expected in each scenario. As can be seen, the burden of ROH relates to the size of the population, with smaller populations having more and longer ROH than larger populations. Admixture brings together different haplotypes and typically reduces the number of ROH to very few short ROH, whereas bottlenecks increase the number of ROH, which are typically still relatively short. Consanguinity, on the other hand, adds a small number of very long ROH for those who are the offspring of cousin marriage, thus increasing the variance in the sum of ROH, visible as a right shift in the NROH versus SROH plot. Some populations are both bottlenecked and practice consanguineous marriage, hence having many short and some long ROH, resulting in the highest burden of ROH. (Ceballos et al., 2018)

ROH can produce a *Froh*(inbreeding based on ROH) statistic. Inbreeding refers to mating between individuals who are related to each other. It can occur in small, isolated populations or through intentional breeding methods. Inbreeding results in increased homozygosity in the genome when an individual inherits two identical copies of an allele from both parents at a specific locus(Kardos et al., 2015). Inbreeding reduces the diversity within a population and, over time, leads to a loss of variation. This can lead to inbreeding depression – a reduction of fitness among offspring of related parents due to loss of

masking for the harmful recessive alleles that would usually happen with offspring of unrelated individuals. High levels of inbreeding often threaten small or isolated populations as the loss of diversity means a reduction of their adaptation toolkit to answer changes in environmental changes or resistance to various pathogens and diseases.(Caballero et al., 2020)

Estimation of inbreeding can provide insight into the history and the structure of populations under scrutiny. High levels of inbreeding often indicate a small effective population size, lack of migration, or a recent population bottleneck. For conservation genomics, inbreeding research forms the foundation for the development of effective breeding programs, translocation decisions, and other activities to ensure the long-term survival of specific populations or species in general.

Based on the ROH statistic, the *Froh* is calculated using the following formula:

$$Froh = \frac{\sum L_{ROH}}{L_{genome}}$$

Where $\sum L_{ROH}$ is the sum of the length of all ROH detected in an individual, and L_{genome} is the total length of the genome that was used for the analysis. Similarly, *Froh* can be used to calculate the metric per chromosome. In this case, $\sum L_{ROH}$ stands for the sum of the lengths of all ROH in an individual on a chromosome, and L_{genome} represents the length of the chromosome.(Marras et al., 2015; Pilon et al., 2021)

The Froh statistic provides a score of 0 to 1 for each individual. A score of 0 indicates no long stretches of the genome where both copies of each gene are identical by descent from a common ancestor. This suggests the individual is outbred or has completely

unrelated ancestors. On the other hand, a score of 1 indicates complete autozygosity, meaning the entire genome is made up of ROH. This would suggest extreme inbreeding, such as self-fertilization. However, such a high score is unlikely in most natural populations.

ROH are increasingly used in population genomics, particularly in human populations, as a way of understanding historical demographic events and their impact on genetic variation. They can provide insights into an individual's genetic history, including inbreeding, recombination, and natural selection. However, it is important to interpret the results of ROH with caution and within the context of several limitations. The history and characteristics of each population can significantly influence ROH patterns. The estimates of ROH are highly dependent on recombination rates, which can vary across the genome and between populations. Inferences made using ROH are often population-specific and so may not be generalizable across different populations. Moreover, detection of short ROH can be difficult, especially with lower-density genotyping data. This can lead to potential underestimation of the inbreeding coefficients (Wolfsberger et al., 2022a). Finally, the very definition of what constitutes a "run" of homozygosity can be subjective and is influenced by factors such as the density of the genotyping platform and the level of linkage disequilibrium in the population.

2.4 Methods for detection of signatures of selection along the genome using population data

Recent detection methods of selected 'genomic footprints' are essential for discovering disease genes and understanding the events that shaped the current population genetic structure. These footprints in human and animal genomes allow the interpretation

of modern and ancestral gene origins and changes. Current methods to reveal selected regions in genome-wide selection scans fall into eight categories, including phylogenetics, evaluating mutation rates and polymorphism, assessing linkage disequilibrium and genetic variation, inspecting the frequency distribution of genetic variation, assessing population differentiation, and examining admixture contributions (Oleksyk et al., 2010). In this section, we will review methods that can be applied to recent population events with the help of genomic data.

2.4.1 Extended haplotype homozygosity(EHH), site-specific Extended Haplotype Homozygosity(EHHS), and variant call data phasing

Due to their nature –analysis of haplotypes - the following methods work with phased variant calls. Current high-throughput sequencing technologies and variant calling approaches cannot directly assign alleles to specific chromosomes of a heterozygous diploid (or multiploid) individual as demonstrated in Figure 5. Instead, this task of phasing is typically performed by specialized, complex bioinformatic tools like SHAPEIT and fastPHASE. For example, SHAPEIT begins with random assignment of the haplotype arrangement for every sample. After, it employs the Markov chain Monte Carlo (MCMC) algorithm to refine the haplotype configuration in many iterative steps. During each iteration, the algorithm suggests a new haplotype configuration and evaluates its fit based on a likelihood function that corresponds to how suitable it is to explain the observed data. The function combines the information about the recombination and mutation rates between genetic markers with the frequencies of different haplotypes in the population or a reference panel. The likelihood of a given state is higher if the pairs of haplotypes are consistent with the observed genotypes and the prior known properties of genetic

recombination and mutation. This iterative process is repeated multiple times, allowing the algorithm to explore a wide range of possible haplotype configurations. Ultimately, after numerous iterations, the algorithm converges to a final stationary haplotype configuration with the highest likelihood score from the multitude of random starting haplotype arrangements. This configuration represents the most probable arrangement of alleles on each chromosome(Browning & Browning, 2011; Delaneau et al., 2019a). Instruments like SHAPEIT4 can incorporate reference panels or prior information about known haplotypes, which can be produced using long-read sequencing or other methods to improve the accuracy of its phase estimation (Delaneau et al., 2019b). All the algorithms benefit in terms of accuracy and ability to resolve complex recombination regions if a sufficiently large sample size is provided, making them well-suited for population genomics. Although exceptionally computationally demanding, the application of these tools is straightforward, and the results are usually of sufficient quality for calculating EHH-based statistics.

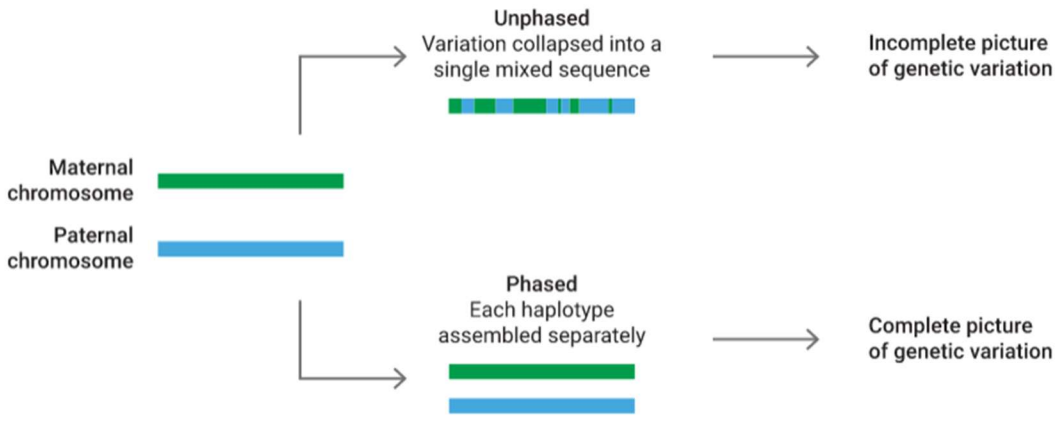

Figure 5. Phasing visual guide
Phasing involves separating maternally and paternally inherited copies of each chromosome into haplotypes to get a complete picture of genetic variation.

With phased variant calls, we can perform the first analysis in the family of EHH tests. While runs of homozygosity investigate the genome for stretches of complete biallelic (in the case of diploid organisms) homozygosity for both chromosomes, extended haplotype homozygosity operates under the increased resolution of phased variation to investigate similar patterns across each of the pair of chromosomes for every individual allele. EHH is defined as the probability that two randomly chosen chromosomes carrying the core allele are homozygous over a given surrounding chromosomal region (Sabeti et al., 2002a). An example EHH visualization for an ancestral and derived allele and its interpretation is provided in Figure 6.

In contrast to allele-specific EHH, the site-specific Extended Haplotype Homozygosity (EHHS) considers all alleles, and measures the overall decay of homozygosity per site, extending outwards until it breaks down. Integrated *EHHS* scores are labeled *iHH*. Rarely used by itself, it is instrumental for the methods described further.(Sabeti et al., 2002a; Tang et al., 2007a)

The most important limitation of the EHH-based tests is the age of the selection event: they are most effective at detecting recent, strong positive selection. Their power to detect older or weaker selection events declines after approximately 1,200 generations or 30K years for humans (Oleksyk et al., 2010). Even for the relatively recent selection events, EHH tests can help identify regions of the genome under selection, not the specific causal variants within these regions, because selection can cause a large region of homozygosity surrounding the selected variant that includes many hitchhiking alleles. How big this region also depends on the local recombination rate that should be incorporated in a selection test, but most often is not (Oleksyk et al., 2008).

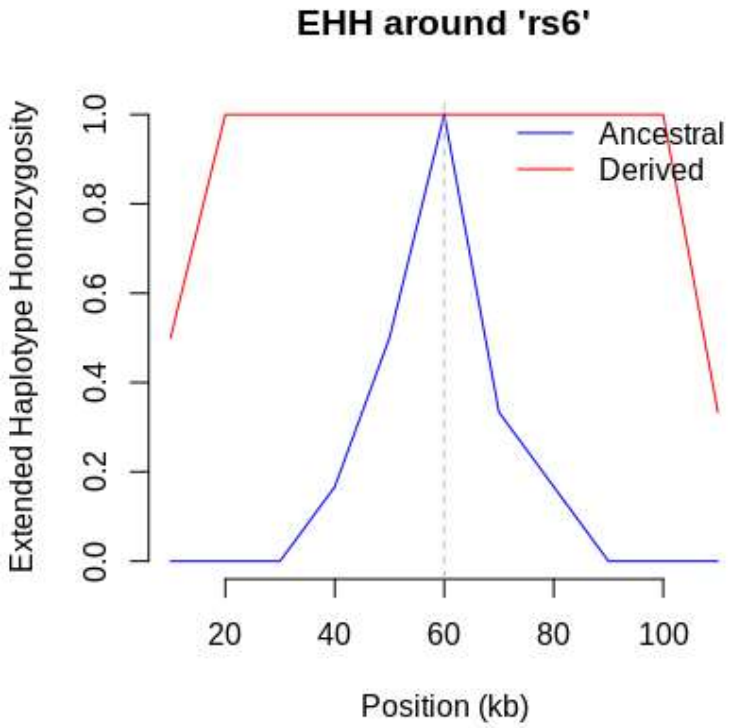

Figure 6. EHH plot around a focal marker rs6 for two alleles.
EHH of the ancestral allele decays more rapidly than that of the derived allele, represented by an extended horizontal line. If the line decays slowly, it suggests that there is an extended haplotype carrying the derived allele, which could be a sign of recent positive selection if the haplotype is unusually long compared to what would be expected under neutral evolution.

2.4.2 Single population selection signature - Integrated Haplotype Homozygosity Score (iHS)

When extended to whole chromosomes and the whole genome of an organism, these metrics let us detect a direction score to identify signals of recent positive selection in the data. The Integrated Haplotype Score (iHS) metric is produced by calculating EHH scores across all the sites in the whole sequence separately for ancestral and derived alleles and further transformations. The iHS measure is symmetrical around 0 and represents the amount of extended haplotype homozygosity at each site along the ancestral allele relative to the derived allele. This symmetry simplifies the interpretation of the iHS and allows for

more straightforward comparisons across different loci within the population, with positive values indicating selection on the derived allele and negative values indicating selection on the ancestral allele, as demonstrated in Figure 7. (Voight et al., 2006)

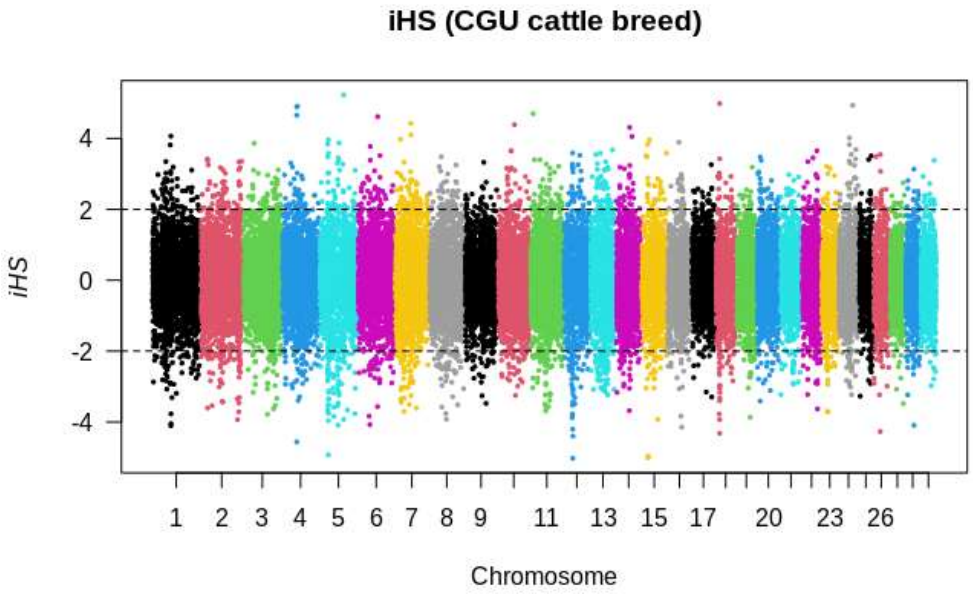

Figure 7 Whole genome *iHS* plot for cattle breed population.
Every dot represents an individual site with extremely positive values indicating selection on the derived allele and extremely negative - selection on the ancestral allele.

2.4.3 Pairwise population selection signature - *Ratio of EHHS between populations (Rsb) and Cross-Population Extended Haplotype Homozygosity (XP-EHH)*

While the previous method (*iHS*) aims to discover signatures of positive selection within a population for a specific allele – *Rsb* and *XP-EHH* are similar methods that expand on the Extended Haplotype Homozygosity concept to produce a site-specific comparison of two different populations.(Sabeti et al., 2002a; Tang et al., 2007a). By comparing the site-specific EHH between two populations, the *Rsb* method can identify loci that have undergone recent positive selection in one population but not the other. The *Rsb* values are computed for all markers present in both data frames. They are coded by chromosome name and position, not their identifiers, so each population must have unique chromosomal

positions. (Tang et al., 2007a) *XP-EHH* test is based on a similar idea, but instead of using the site-specific information, it uses an integrated score *iHH* (which considers a specific allele at the focal locus for every locus).(Sabeti et al., 2002a)

In contrast to the within-population *iHS,* the methods which compare two populations have a set of advantages. They can identify signals of selection that are specific to one population but not the other. This can provide insights into population-specific adaptations, especially when combined with the ecological context of the difference within the environment the populations are occupying. The between-population methods can detect complete selective sweeps or older selection events that have driven an allele to near fixation or high frequency in one population but not the other. (Eydivandi et al., 2021)

Similarly to other EHH methods, *Rsb* and *XP-EHH* also use variant files of two separate populations as input and produce a symmetrical output statistic. A positive *Rsb* value suggests that the focal site has a stronger positive selection signature in the first population, while a negative – suggests the same for the second population. The magnitude of the value indicates the strength of the signal - a stronger signal of selection is indicated by a larger absolute value.

Finally, methods like the *Rsb* and *XP-EHH* that involve comparing EHH between populations, which can be challenging if the populations have different demographic histories or different underlying patterns of variation. Consequently, it is important to use a combination of methods and metrics and to consider additional information when studying selection in the genome rather than relying solely on the EHH-based tests statistics(Guiblet et al., 2015; Oleksyk et al., 2008, 2010). Some haplotype-based tests also incorporate F_{ST}, a fixation index proposed by Sewall Wright to assess the proportion of

total genetic variance that can be attributed to differences between populations (WRIGHT, 1949). For example, the *XP-CLR* (Cross Population Composite Likelihood Ratio) test combines information on allele frequency differentiation (measured by F_{ST} or similar statistics) and haplotype structure to detect selection (Chen et al., 2010).

2.5 Case study: Application of population genomics to infer the genetic diversity and selection in Puerto Rican horses

This section will overview a first-author manuscript produced by OU Genomics lab, "Genetic diversity and selection in Puerto Rican horses", published in Scientific Reports (Wolfsberger et al., 2022b). The sample collection and part of the preliminary analysis that led to this study were published as part of an MS thesis in our lab (Ayala, 2016). As the first author and primary contributor to the peer-reviewed paper, I was responsible for implementing all bioinformatic analyses in this work. Below is the review of the general context of the study focused on the bioinformatic approaches reviewed earlier and employed to better our understanding of the history of Puerto Rican horse breeds.

2.5.1 Background

After their domestication, horses have been selectively bred for various traits, including the ability to perform additional gaits beyond the standard walk, trot, and gallop. One such is the four-beat ambling gait, known for being comfortable for riders. This gait is exhibited by gaited horses found worldwide, suggesting that gaitedness is an ancient trait selected independently across different horse breeds(Promerová et al., 2014a). Researchers believe that this altered gait phenotype is linked to a specific 'gait-

keeper' mutation in the DMRT3 gene, which regulates the spinal circuits that control stride in vertebrates.(Kristjansson et al., 2014)

The Puerto Rican Paso Fino (PRPF) is a naturally gaited horse breed prized for its specific phenotypic characteristics, including smooth, natural, four-beat, lateral ambling gait referred to as “paso”(Hendricks, 2007). PRPF’s origin is directly connected to the history of the West Indies, which were the destination for immigrants and their livestock in the Caribbean from the initial arrival of the Spanish settlers until the conquest of Mexico. During this time, various Iberian horses, likely related to the modern Spanish breeds, were introduced to Puerto Rico and other Caribbean islands. Those imported horses likely represented the genetic variation that existed in their countries of origin, which, in this case, meant Spain(Luís et al., 2006).

Puerto Rico became the breeding ground for horses that were later exported from the island by the Spanish. The resulting admixture of the imported breeds on the island eventually resulted in the local mixed variety called “the Criollo”. These small but powerful horses with a variety of gaits are quite capable of carrying big cargo with little effort. Even as traditional uses for horses have drastically declined in the last century, the non-purebred (NPB) Criollo horses are still ubiquitous on the island of Puerto Rico. Although it is logical to assume that PRNPB horses were the initial source of the highly valued purebred PRPF, the genetic link between them has never been established.

The PRPF that arose in Puerto Rican farms were used by the landowners and their foremen that supervised their plantations on horses, riding on horses selected for their smooth and secure footing on the challenging, uneven, and slippery mountain terrain of the island's interior. This led to the development of an additional isometric, short step in which

the hoof barely rises and softly lands for a gentle footstep. Other features – some related to the ride's comfort and some purely cosmetic (Mack et al., 2017), were selected for by breeders, and by 1840, the term “paso fino” was established.

In this study, we sought to understand the origins of the famed Puerto Rican breed. We examined mitochondrial markers in Puerto Rico and compared genetic diversity in nuclear markers to other horse breeds, and looked at the diversity between the purebred PRPF and Puerto Rican Criollo horses to understand their connection, with a particular focus on the frequencies of the alleles in the DMRT3 gene. Finally, we have used population genomics approaches to study the genetic diversity at the chromosomal level and identify the structure of populations of horses on the island and identified signatures of selection across the whole genome for PRPF and PRNPB populations.

2.5.2 Methodology

Sample collection and data generation: Prior to this study, hair samples were gathered from 200 distinct horses across 27 different locations throughout Puerto Rico, resulting in 162 Puerto Rican Non-Purebred (PRNPB) and 38 Puerto Rican Paso Fino (PRPF) horses. The horse owners confirmed that the samples were unrelated and gave informed consent for their animals to participate in the study. All hair collections were conducted under the supervision or assistance of the horse owner or trainer. The methods used adhered to all relevant guidelines and regulations, and the collection procedure received approval from the Institutional Animal Care and Use Committee (IACUC) at the University of Puerto Rico at Mayagüez (UPRM). (Ayala, 2016)

We genotyped all 200 samples for the DMRT3 Ser301STOP "gait-keeper" mutation and successfully produced the genotypes of 180 of them. Additionally, we

genotyped extra 24 samples (3 PRPF and 21 PR NPB) with whole-genome Illumina Neogen Equine Community Array (65,157 markers evenly distributed across 31 autosomes)(McCue et al., 2012). The genotypes were converted to Variant Call and Plink(Chang et al., 2015) formats for downstream analysis.

Data preparation: We merged the data obtained from our samples with a publicly available dataset of a broader study that included 795 samples of 37 horse breeds all around the world genotyped on a similar array(Petersen et al., 2013). After merging, filtering by the genotyping rate, and soft pruning the results for LD (–geno 0.02, –indep-pairwise 50 10 0.2), a total of 819 samples and 22,347 variants were retained. Notably, the dataset contained genome-wide genotypes of 20 additional PRPF horses that were merged with our data. The approximately equal number of genome-wide genotypes from each breed (23 PRPF and 21 PRNPB) were included in all the downstream genome-wide analyses. This combined dataset will be referred to as a "merged dataset" in the future to avoid confusion.

The dataset with 22,345 sites covering the whole genome was used to perform PCA, Admixture, and ROH analysis, iHS, XP-EHH, and Rsb selection tests. In the ancestral/derived EHH test, the 966 markers on the Equus caballus chromosome 23 (ECA23) were used to illustrate the haplotype homozygosity around the putatively selected locus.

Principal component analysis (PCA). A statistical procedure used for simplifying the complexity in multi-dimensional genome data while retaining trends and patterns (PCA) was performed using the EIGENSOFT package(Price et al., 2006b), while the plotting of the principal components was performed with custom Python scripts and libraries for data processing (pandas, matplotlib, seaborne). The visualization for PCA

plots was streamlined by assigning colors to each breed according to the phylogeny previously calculated from SNP frequencies in 38 breeds (Figure 8)(Petersen et al., 2013).

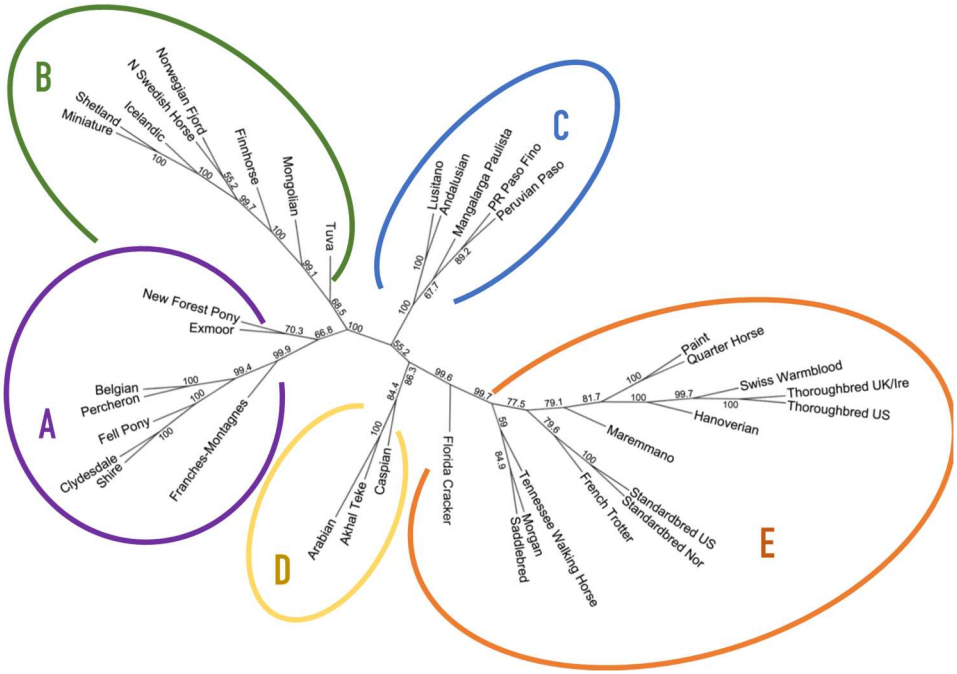

Figure 8. Distance-based, neighbor joining tree calculated from SNP frequencies in 38 horse populations.
Majority rule, neighbor-joining tree created from 10,536 SNP makers using Nei's genetic distance and allele frequencies within each population. Percent bootstrap support for all branches calculated from 1,000 replicates is shown. (modified from Petersen et al., 2013). For easy reference, the branches are grouped into five clades (A, B, C, D, and E), shown by different colors in the plot.

The two most significant components (PC1 and PC2) were used to visualize genetic distances between all horse samples across the studies (Figure 9). For easy reference, the colors in the graph correspond to the groupings dendrogram of horse breeds from prior studies, demonstrated in Figure 8. The area containing Puerto Rican horses and related breeds is shown in purple (clade C) and enlarged in Figure 9.

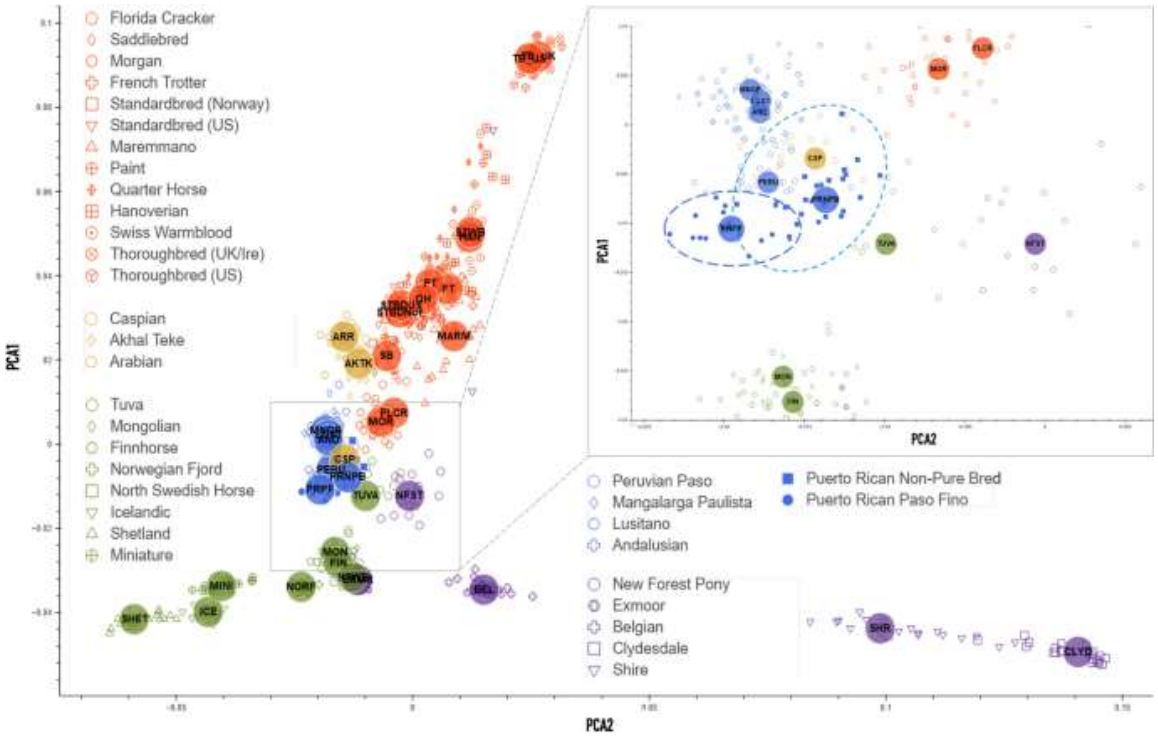

Figure 9 Principal component analysis plot of the merged dataset of horse breeds.
Puerto Rican Paso Fino (PRPF) and Puerto Rican Criollo (PRNPB) from this study are encircled in the enlarged area. The color scheme of this plot corresponds to the color-coded groupings in Figure 6

Estimation of admixture of ancestral components: Admixture analysis was performed on a merged dataset using ADMIXTURE software. The optimal number of subpopulations was selected using ADMIXTURE's cross-validation, with k = 25 on a 20-fold repetition of the procedure.(Alexander et al., 2009b) I have developed a custom script to streamline the structure visualization plot for horse breeds. Due to high number of ancestral components, the plot used many colors to represent various admixture, and my script masks the components that are rare or absent in Puerto Rican horses. This modification highlights only the components that may be shared between other horses and the Puerto Rican breeds. The threshold for shared components can be adjusted as needed and shows only components that comprise less than 1% of the shared population structure(Figure 10).

Runs of homozygosity and inbreeding: We conducted a Runs of Homozygosity (ROH) analysis on PRPF and PRNPB using a sliding window method(Ceballos et al., 2018). Due to the low density of SNP coverage, we selected a sliding window size of 15 loci, with a minimum of 5 consecutive SNPs to begin the run and a maximum distance of 10 Mbp between SNPs in each run. We first calculated the ROH numbers and sizes for each individual separately and then averaged them for each breed of Puerto Rican origin.

The coefficient of inbreeding was derived from the ROH analysis (Pilon et al., 2021). Inbreeding estimates were determined using two methods: individual inbreeding coefficients (Fmax) measure observed versus expected genetic diversity in an individual in a population. The sum of runs of homozygosity (SROH) estimates the proportion of the genome covered by fragments lacking any variation in SNPs (heterozygosity).

Selection signatures via extended haplotype heterozygosity: We analyzed the Puerto Rican horse's genetic makeup to identify signs of recent selection. Using the rehh package for R(Gautier et al., 2017), we pinpointed regions with high local haplotype homozygosity. We only used PRPF and PRNPB samples from our merged dataset for these tests. We phased our genotypes using SHAPEIT4 tool(Delaneau et al., 2019a) and discarded loci that were absent in both populations. The iHS, the within-population genome-wide analysis, was calculated for PRPF and PRNRB as well as other breeds using rehh 2.0. We also calculated the pairwise population statistics, Rsb and XP-EHH, comparing the breeds in their phylogenetic context. For this and the following analysis, in addition to PRPF and PRNPB samples, we included genotype data from the Andalusian, Lusitano, and Mangalarga Paulista breeds, which are closely related to Puerto Rican horses.

To visualize recent selection and contrast haplotypes associated with the selected trait in the genomic region, we used extended haplotype homozygosity (EHH)(Tang et al., 2007b). In this approach, the haplotype frequency and decay of haplotype length were evaluated for the chromosomal neighborhood flanking the candidate marker under selection, for example, the "gait-keeper" mutation BIEC2 620109 (DMRT3Stop).

2.5.3 Results

DMRT3Stop mutation rate in PRPF and PR NPB: All 200 samples originally collected were genotyped for DMRT3Stop mutation previously associated with the paso gait and combined with the information available on other horse breeds from previous studies. (Promerová et al., 2014b). Surprisingly, the frequency of the mutant allele in the PRNPB population is very high at 87.4%, and the paso gait has never been associated with PRNPB in previous literature (Ayala, 2016). This discovery guided us to explore the history of selection and any potential connection between the two breeds in Puerto Rico.

Population structure and admixture: We used the PCA approach to visualize the genetic difference between Puerto Rican horses and other Iberian and American breeds. In the PCA, we used the merged dataset of 819 samples and 22,347 variants. The components explaining the most variation between samples - PC1 and PC2 were then used to visualize the data.

The resulting distribution shows genetic differences between Puerto Rican horses and other breeds, with PRPF and PRNPB forming a distinct spread along a "vector." The Iberian breeds (Lusitano, Andalusian and Mangalanga Paulista) that share the same branch on the phylogenetic tree (group "C", Figure 6) are clustered nearby. However, horses from other branches, specifically the Caspian and Tuvan horses, also cluster in the vicinity. Even

though some of the diversity in the non-purebred (NPB) Puerto Rican Criollo horses is shared with the Peruvian Paso, the genotypes of PRPF horses form a distinct cluster, which is different from all other breeds (Figure 7).

In order to shed light on the genetic origin of the PRPF horses, we used the admixture analysis of the merged dataset to find the shared genetic components between Puerto Rican horses and other breeds. The initial visualization of all the horse breeds resulted in a crowded plot, but we were able to highlight specific segments that were shared among various breeds.

Based on the admixture analysis, Puerto Rican horses share population structure components with several horse breeds from around the world. Specifically, the PRNPB horses share genetic diversity components with New Forest Pony, Tuva, Mongolian, Caspian horses, and Florida Cracker. These breeds exhibit multiple genetic components, as indicated by the presence of various colors in the plot, which displays diversity within the breeds. On the other hand, the PRPF genome is mainly dominated by a single (purple) component that appears to be mostly unique to this breed. Traces of this component can also be observed in New Forest Pony, Tuva, and Mongolian horses. Interestingly, while the Peruvian Paso also shares a genetic component with PRNPB, it appears to be dominated by a different (orange), more common component than that prevalent in the PRPF. (Figure 8)

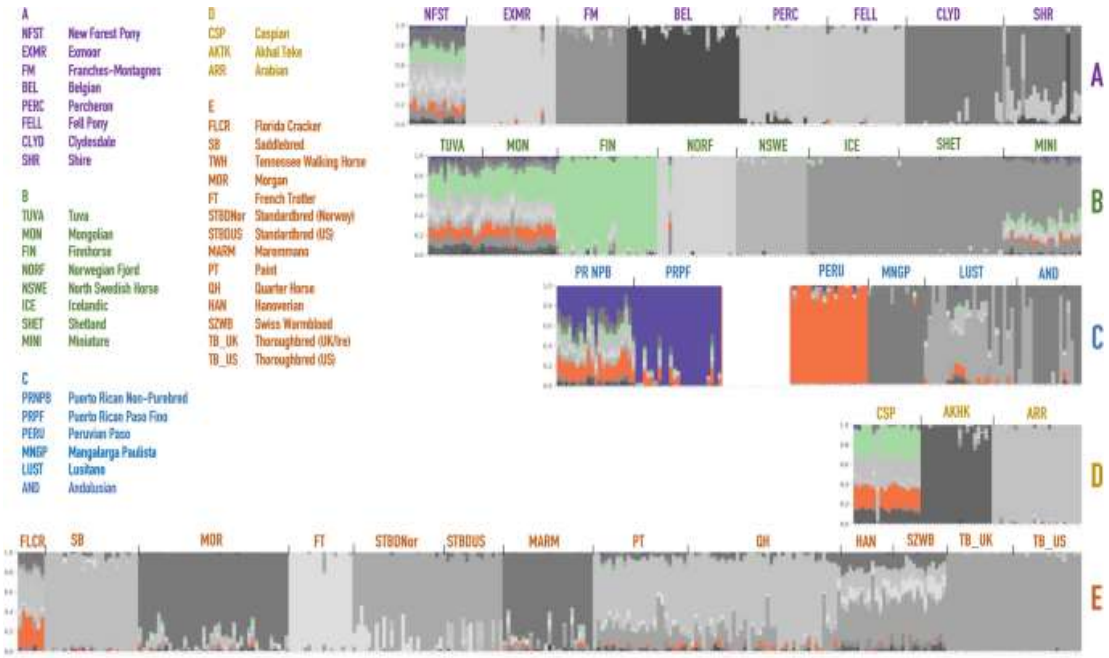

Figure 10 Admixture graph of the shared dataset of horse breeds
The components that are rare in the Puerto Rican Non-Purebred (PRNPB) Criollo horses (< 1% of the total proportion of ancestral components) are masked to improve readability by our own script. As a result, the population components that may have been shared are clearly visible.

Runs of homozygosity and inbreeding:

We have performed runs of homozygosity tests and have calculated the coefficient of inbreeding *Froh* based on the ROH. The coefficient of inbreeding was higher in PRPF horses compared to the non-purebred (NPB) Criollo horses, with PRPF showing a value of 0.23 and NRB Criollo horses showing a value of 0.16. Paso Fino horses had more ROH regions than Criollo horses, and these regions were, on average, almost three times larger in size in PRPF compared to PRNPB. This suggests that PRPF founders were originally selected from the PRNPB genetic pool. The results of the ROH scan are reflected in Figure 11.

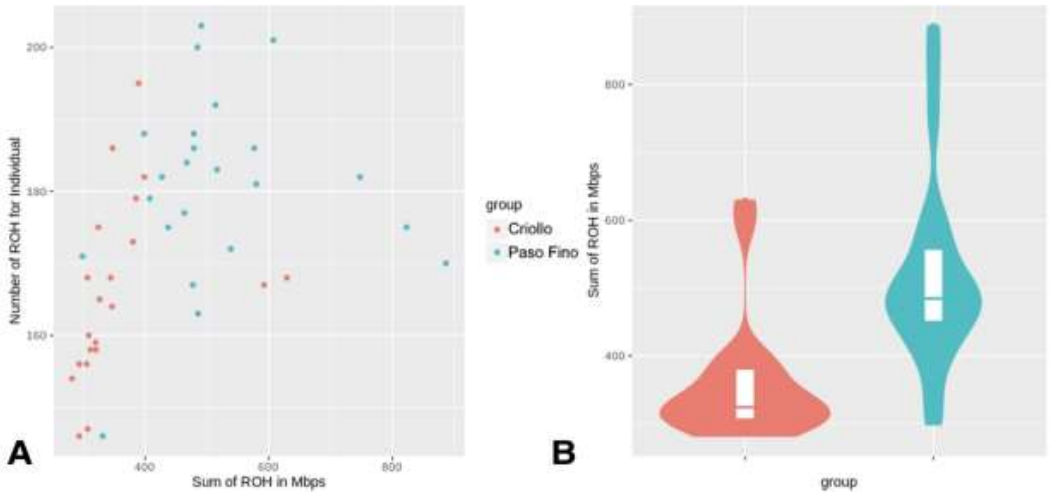

Figure 11 Runs of homozygosity (ROH) in the genomes of the two Puerto Rican horse breeds
(A) Number vs. sum of length included in the ROH per individual in PRPF vs PRNPB (Criollo) horses. (B) A significant difference in the distribution of ROH extent (sum of length) between PRNPB (Criollo, red) and PRPF (Paso Fino, teal) horses.

Signatures of selection within horse population: EHH tests – iHs and EHH

We performed a genome-wide analysis for extended haplotype homozygosity (EHH) using *iHs* statistics across all the Puerto Rican samples from the merged dataset. (Tang et al., 2007b). A strong genome-wide selection signature is detected in the PRNPB horses in the *DMRT3*_Ser301STOP region (Figure 5). The exact site of the *DMRT3*_Ser301STOP mutation showed some indication of the presence of a signature of selection ($p < 0.05$). However, it was directly adjacent to three other markers (BIEC2 627539, BIEC2 621347, and BIEC2 620774) that showed the highest significance level in the genome adjusted for multiple testing (above the red line). This is explained by the limitations of the *iHs* approach, which lies in its inability to identify selection signatures in sites close to fixation(Tang et al., 2007b), which is, unfortunately, the situation with the DMRT3_Ser301STOP mutation. Thus, we focused on the variation of SNPs located within ±2 Mb proximity of the DMRT3_Ser301STOP mutation on ECA23 in PRNPB and PRPF horses, using the rehh 2.0 package in R(Gautier et al., 2017).

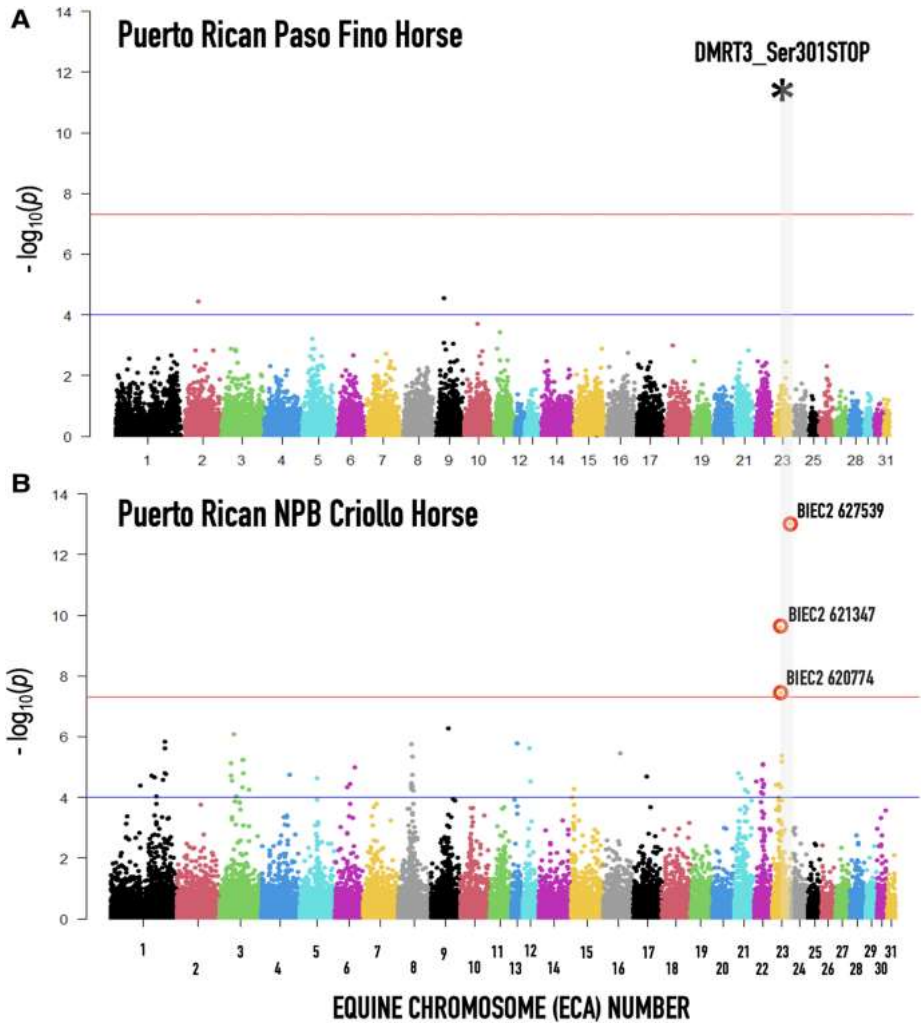

Figure 12 Integrated extended haplotype homozygosity scores (*iHs*)
iHs across the horse genome show selection signature near the paso fino gait locus in the PRNPB but not in the PRPF horses
(A) Puerto Rican Paso Fino (PRPF) (B) Puerto Rican Criollo (PRNPB) horse. The blue line represents SNPs showing recent selection signatures corrected for the multiple testing significant at the chromosome level, with the red line cutoff—SNPs significant at the whole genome ledvel. The shaded area indicates the location and the region of our marker of interest, DMRT3_Ser301STOP mutation (BIEC2 620109, marked by an asterisk). Genotyping was performed with the whole-genome Illumina Neogen Equine Community Array (65,157 markers evenly distributed across 31 autosomes)(McCue et al., 2012).

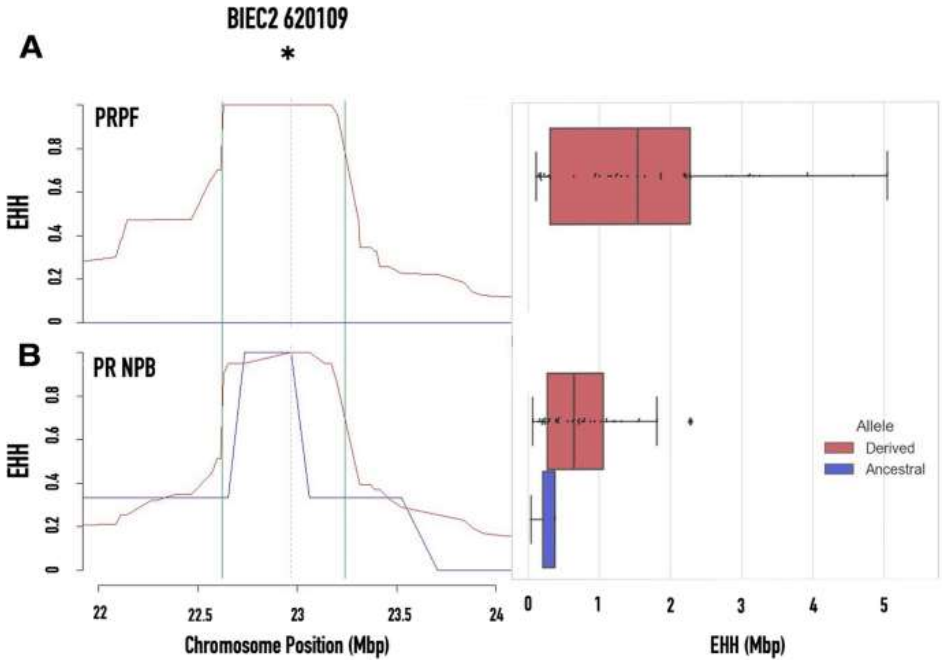

Figure 13 Extended haplotype homozygosity (EHH) decay (left) and shared haplotype length (right) Graph constructed for the regions around the focal DMRT3_Ser301STOP mutation on chromosome 23 associated with the paso gait (locus BIEC2 620109).

The decay of extended haplotype homozygosity in loci BIEC2 627539, BIEC2 621347, and BIEC2 620774 shows the highest significance level in the genome adjusted for multiple testing in PRNPB as shown by the *iHs* plot in Figure 12. Additionally, we plotted the EHH values across the ancestral and derived alleles (with the presence of *DMRT3*_Ser301STOP considered as a derived allele) for both PRPF and PR NPB. The decay of homozygosity near the *DMRT3*_Ser301STOP (BIEC2-620109 locus) stretches further along the chromosome for haplotypes with the gait-keeper A allele in Puerto Rican NPB Criollo horses compared to those with the ancestral allele. The PRPF horses are fixed for the derived allele, and the EHH around it is even more extended than in PRNPB horses. This suggests there may have been selection at this locus in the past. (Figure 13)

Given access to multiple populations data, we have leveraged the advantages of XP-EHH and *Rsb* tests. We have observed strong signatures of selection in PRNPB horses

compared to other non-PR horse breeds from the merged datasets. Our analysis involved examining the haplotype homozygosity and divergence between PRPF and PRNPB, as well as comparing their genomes with other closely related breeds – Lusitano, Andalusian, and Mangalanga Paulista, labeled as Clade "C" to detect selection signatures (Figure 6 and 7). We identified strong signatures near the *DMRT3* gene in pairwise tests with the C group. No tests demonstrated selection signature in PRNPB compared to PRPF, but signatures in the opposite direction were identified. The results of the *Rsb* and XP-EHH test using genome-wide data from horses is provided in Figure 12.1-2. We have cataloged the regions with signals of selection, annotated them with the closest genes using the RefSeq annotation database (O'Leary et al., 2016) with equine reference genome v3 (Kalbfleisch et al., 2018), and aim to use them as potential targets for further selection signature testing. Our research highlights the importance of whole genome population genomics approaches in helping researchers unravel complex histories of populations. We have employed these methods to bolster our understanding of the relationship between the two horse breeds in Puerto Rico and were able to find strong support for a hypothesis of their differentiation and history. Our findings support the scenario where the more common PRNPB or "*Criollo*" horses are the direct descendants of the original genetic pool that was transported from the Iberian Peninsula and other parts of Europe by the early Spanish settlers. It is also likely that at least some of the original founders of the PRNRB population carried the "gait-keeper" *DMRT3* allele when they arrived on the island. Over time, this allele was gradually selected by the Puerto Rican farmers for its contribution to the gait, and eventually, the allele rose to the state near fixation in the PRNRB population (Ayala 2016, Wolfsberger et al., 2022).

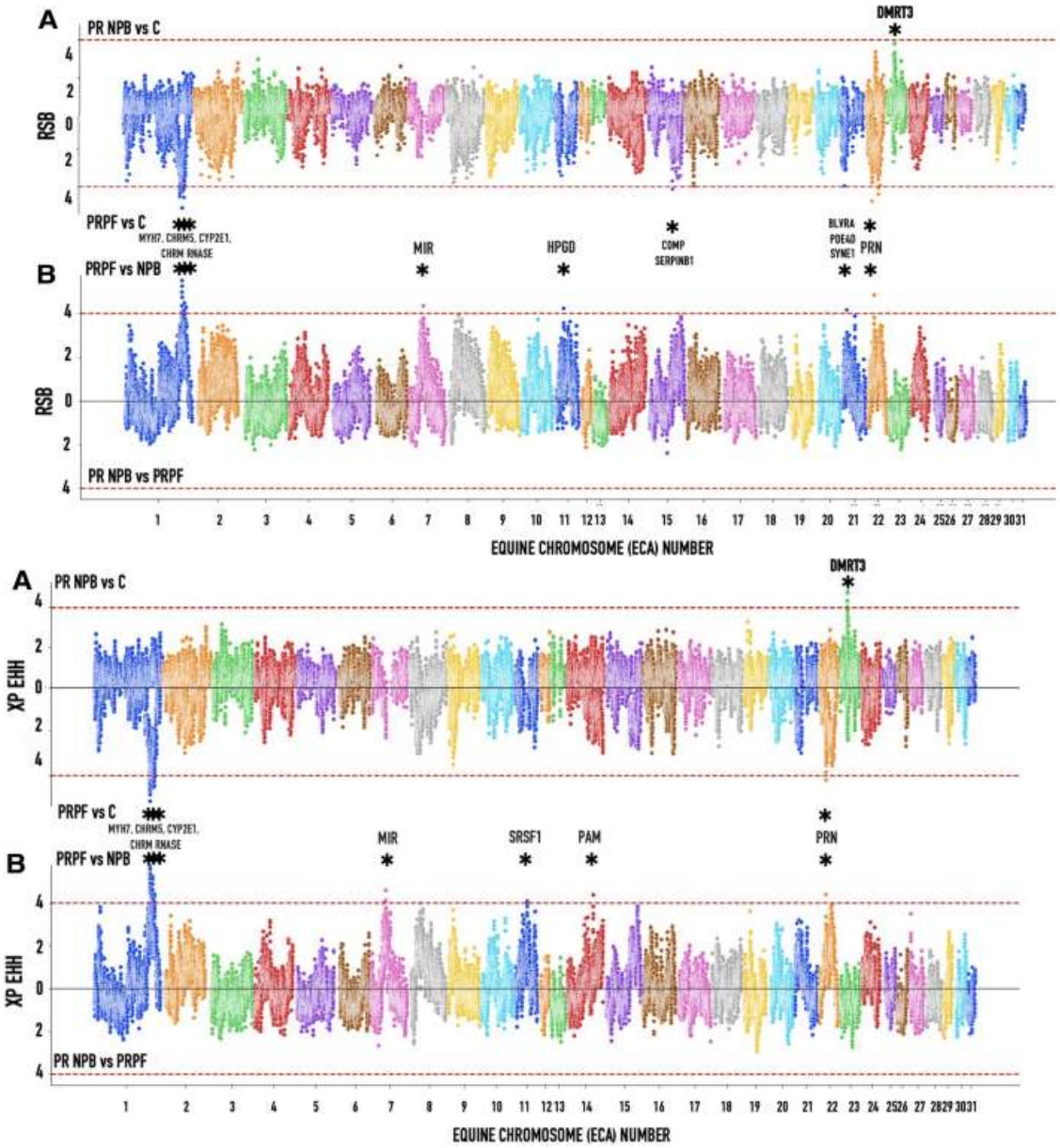

Figure 14. Signatures of selection based on the ratio of EHHs between populations do not show selection for DMRT3 but suggest that a number of other genes could be undergoing recent selection in the PRPF breed.

(1) EHHS between populations *(*Rcb) (A) Selection tests comparing PRPF and PRNPB horse genomes to the other breeds in the C clade (Lusitano, Andalusian, and Mangalanga Paulista). (B). Comparing PRPF to PRNPB. **(2). Cross-Population Extended Haplotype Homozygosity (XP EHH).** Genotyping was performed with the whole-genome Illumina Neogen Equine Community Array (65,157 markers evenly distributed across 31 autosomes)(McCue et al., 2012). Genes with RSB values > 4 are displayed on the graph with the closest genes according to RefSeq track on equine reference genome v3.

2.5.4 Discussion

An earlier analysis of the mtDNA variation in the two horse breeds of Puerto Rico was consistent with previous research, the hypothesis that the two horses had common origins (Ayala, 2016). Specifically, the mtDNA network analysis revealed a predominant Iberian lineage, with numerous shared haplotypes between the two breeds indicating their common descent. These findings suggest that the ancestral population on the island was likely established by the arrival of a mixed group consisting of several Iberian horse breeds. The analysis of the genome-wide array data allowed us to leverage modern analysis techniques and provide additional context to our findings. The PCA analysis of Puerto Rican horses indicates that PRPF is a distinct and native breed of American Horses with Iberian ancestry. The analysis also showed that PRPF is closely related to the PRNPB population but is completely distinct from all other breeds, except Peruvian Paso, which is likely to descend from PRNPB (Figures 9 and 10).

We used the estimation of ancestral components analysis via ADMIXTURE to look at genome contributions in the context of population variation among the worldwide horse breeds (Figure 10). Both breeds of horses from Puerto Rico share a common characteristic that appears to be exclusive to the local island horses. PRNPB horses are much more diverse, and its ADMIXTURE plot appears to be a mixture of genetic components that share an origin with the Iberian, Northern European, and Asian horse breeds (Figure 10). At the same time, PRPF horses are dominated by a single genetic component found to a lesser extent but almost exclusively in PRNPB horses. This distinguishing feature has not been observed in any other horse breeds surveyed. The most likely explanation of this observation is that while the PRNPB horse has a unique mixture incorporating variation

from a diverse set of lines brought on the ships to the island, the PRPF must have been selected for a specific set of variation from the admixed pool of the PRNPB.

If our conclusion is accurate, and PRPF is selected from PRNPB in Puerto Rico, there should have less genetic diversity in the derived vs the original breed. Consistent with this statement, the ROH analysis of the Puerto Rican breeds indicates that much of the PRPF is depleted of heterozygosity compared to the PRNPB (Figure 11). The extended ROHs in PRPF are consistent with the acting artificial selection in this highly selected breed. At the same time, the levels of inbreeding in both breeds was high and are likely to be the consequence of genetic drift and widespread consanguineous mating in the small populations of island horses. Therefore, while the ROH approaches are often used to test hypotheses for artificial selection in domesticated animals(Bosse et al., 2012), in our case, their ability to detect signatures of selection is greatly diminished due to high rates of homozygosity across their genomes overall. Instead, we used EHH-based selection tests: iHs (Figure 12), EHH (Figure 13), as well as *Rsb* and *XP-EHH* (Figure 14) that involve comparing EHH between breeds (horse populations) to evaluate the hypothesis that *DMRT3*_Ser301STOP (BIEC2-620109 locus) was artificially selected in PRPF horse.

The DMRT3 mutant allele, also referred to as the "gait-keeper" or allele A, is commonly found in many horse breeds known for their unique gaits or bred for harness racing. In contrast, other horse breeds usually carry the wild-type allele (allele C). The A allele is highly prevalent in Northern European breeds, like the Icelandic horse, where selective breeding for specific gaits could potentially lead to a complete disappearance of the C allele. However, this mutation is rare in Iberian horses, with only a single low-frequency report in the Pura Raza Galega breed. (Promerová et al., 2014a) In contrast,

many horse breeds in the New World possess this allele, likely due to crossbreeding of the Iberian horses with non-Iberian, Northern European breeds. While some studies indicated that the selection of the DMRT3 gene could account for differences in gait within a breed for Colombian Paso, a similar study in Mangalarga Marchador and French Trotter indicated that while DMRT3 is associated with gait, it may not be the only gene that influences the trait.(Promerová et al., 2014b; Voight et al., 2006)

Using genotyping essay for *DMRT3*_Ser301STOP (BIEC2-620109 locus), we discovered that the mutant *DMRT3Stop* allele is at very high frequency in the PRNPB population. Surprisingly, 142 out of the 143 genotyped PRNPB horses had at least one *DMRT3* mutant allele (Ayala, 2016). This was unexpected because the PRNPB horses would also be expected to have the prized gait. While PRPF horses are known to perform this gait from birth and do not require any special training to develop it, this trait does not define the PRNPB horses.

When compared to other "criollo" horses mentioned in the literature that have a low frequency of the mutant "gait-keeper" allele (such as Brazilian, Venezuelan, and Columbian breeds) (Promerová et al., 2014b), PRNPB stands out. These breeds also originated from a combination of various Iberian breeds, including a significant influence of Portuguese breeds.

Given the shared mtDNA variation (Ayala, 2016), common genetic structure (Figure 8), and the much larger extent of ROH in PRPF and PRNPB (Figure 11), the most likely hypothesis for the origin of PRPF horse is that it was chosen from the genetic pool of the local NPB horse population and not bought to the island separately. To investigate the possibility of this recent artificial selection (selective breeding), the genomic

neighborhood of the candidate allele must be evaluated in a formal test. Given the history of Puerto Rico, the selection must be recent and not older than the historic horse's arrival on the island. To detect selection signatures in this time range, it is possible to use extended haplotype homozygosity (EHH), population differentiation tests, or a combination of both approaches. (Oleksyk et al., 2010)

Our *iHs* test detected a signature of selection in PRNRB horses at DMRT3_Ser301STOP mutation (BIEC2 620109), a clear signature of the selected region is present on ECA23, (Figure 9). However, this test did not detect any selection signatures in PRPF. The *iHs* test - a recombination-based test that relies solely on variation within specific horse breeds with a major drawback of haplotype-based selection tests is their poor performance in genomes with low genetic diversity (Voight et al., 2006). Therefore, the result is to be expected as there are only a few variable markers on chromosome 23 in PRPF. The illustration of the results of the other single-population test (EHH, (Sabeti et al., 2002b)) illustrates this issue: there is a single haplotype associated with *DMRT3* allele A, and none associated with the allele C (Figure 10). In contrast, the PRNPB horses have haplotypes with both alleles, and those associated with the allele C are both shorter and less frequent, pointing to the possibility of the recent selection pressure on this locus in PRNPB (Figure 10). This does not support the hypothesis of selection for the *paso fino* gait in the PRPF. Alternatively, the possible interpretation of this result is that the DMRT3 haplotype selection has already happened in NPB, and the continuous selection in PRPF would not be possible due to the complete lack of genetic diversity in the locus.

To further explore this hypothesis, we conducted the pairwise between-population analysis and produced a genome-wide distribution of the pairwise EHH test statistics: *Rsb*

and XP-EHH (Figure 12). To provide the reference for the inter-population comparisons, we included a PRPF, PRNPB, and a group consisting of the available genomes from a group of closely related breeds - Lusitano, Andalusian, and Mangalanga Paulista, labeled as “C.” Consequently, three different comparisons were made: PRPF vs. PRNPB, PRPF vs. Clade C, and PRNPB vs. Clade C (Figure C).

The pairwise between-population analysis detected a significant selection signature around the DMRT3 locus in PR NRB horses compared to the outgroup Clade C. This signature is unique to the non-purebred horses, and no selection signatures were found in either test when comparing PRNRB to PRPF (Figure 12). Instead, we have identified several regions with selection signals in the PRPF genome compared to PRNPB and other horse breeds. None of these sites were near the *DMRT3* gene and might correspond to other characteristics selected in PRPF horses, but some may still be related to gait selection. Notably, the strongest signatures were found near the *MYH7* muscle myosin gene on chromosome 1 and the PRNP prion protein gene on chromosome 22.

Given these results and observations, we proposed a new hypothesis for the PRPF origin where they are not selected for the "gait-keeper" mutation, which was already prevalent in the island population of non-purebred horses, likely due to a combination of founder effect and selective breeding by local farmers. Currently, PRPF are selected for secondary characteristics encoded by the genes indicated in Figure 12, and no further selection on the DMRT3 locus is possible due to the lack of genomic diversity in the breed.

2.5.5 Conclusion

In summary, our research indicates that PRPF and PRNPB horses are closely related, and the "gait-keeper" mutation is almost as prevalent in PRNPB as it is in PRPF.

Interestingly, we did not find any selection signatures focused on this gene in PRPF, but we did find a strong signature associated with this gene in PRNPB.

According to our findings, the most likely scenario is that PRPF is a distinctive horse breed derived from local non-purebred horses (PR NPB). The PRNPB genetic pool was formed through a mix of horses brought to Puerto Rico from Spain and other Old World regions, with some of the original horses carrying the "gait-keeper" DMRT3 mutant allele. This allele was then chosen for by local farmers through selective breeding for various traits over generations and has since become more prevalent in the non-purebred horse population across the island.

The founders of PRPF were recently selected from the existing PRNPB pool; however, since the DMRT3 mutant allele was already nearly fixed, the selection in purebred horses was focused on other genes, which may or may not be related to paso gait. Many of these have functions that relate to neuronal and muscular development. The candidate genes identified by our selection scans include MYH7, PRN, and others, and the table is included in the appendix.

The major limitation of our study was the absence of phenotypic data and the small sample size of the individuals with the whole genome data. For future research, we suggest a comprehensive phenotype-genotype analysis based on horse pedigrees and sequencing data from these candidate genes combined with a large-scale population genomics analysis based on low-pass WGS sequencing.

CHAPTER THREE

POPULATION GENOMICS IN HUMAN POPULATIONS

This chapter explores the specifics of population genomics research in human populations. First, I will cover the background and the recent advancements in the field. Then I will use two case studies that resulted in two shared first-author publications describing issues in the current state of human population genomics overall and specifically in uncovering the distribution of genetic diversity in Ukraine.

3.1 Introduction to the field of human population genomics

There is overwhelming support for the idea that environmental factors constantly shape the genomic diversity of human populations (Fumagalli et al., 2011; Hancock et al., 2011). Different environmental niches, each with their unique sets of pressures, have led to the selection of specific genetic combinations that provide survival advantages. Humans have adapted to a wide range of pressures, including dairy consumption (Itan et al., 2009), arctic environments (Cardona et al., 2014), tropical rainforests (Perry et al., 2014), high altitude (Simonson et al., 2010), local toxins (Takser et al., 2016), and ultraviolet exposure (Jablonski & Chaplin, 2010). Diseases have been playing a major role in this process, making up some of the most prevalent environmental agents shaping natural selection in the populations of our ancestors (Barreiro & Quintana-Murci, 2009; Fumagalli et al., 2011; Karlsson et al., 2014; Kwiatkowski, 2005). As a result, a complex and diverse set of local adaptations has emerged, reflecting the great variety of environments that humans have inhabited. (Fan et al., 2016)

One of the factors contributing to the rapid advances in understanding genomic diversity and the evolution of our own species is the availability of the massive volumes

of genomic data generated by ongoing research. In contrast with many other animal species where often not a single reference genome has been assembled to date, the human reference genome is one of most technologically up-to-date and state-of-the-art sequence ever produced, leveraging the latest long read sequencing approaches, with the current version boasting telomere-to-telomere "complete" sequence addressing the most complex regions of our genome (Nurk et al., 2022).

At the same time, the practical and computational approaches in population genomics approaches for human populations are similar to those described for animal species in Chapter Two and generally either infer population structure and history or detect signatures of evolutionary processes along the genome in the population. Generally, both human and animal population studies greatly benefit from the incorporation of results of prior research from either field, and most require the presence of high-quality reference genome and results of previous projects that uncover variation in human populations (Totikov et al., 2021). Finally, human population genomics naturally integrates with and contributes to the practical field of personalized medicine, which aims to tailor medical treatment to the individual characteristics of each patient (Manolio et al., 2009; Need & Goldstein, 2009; Relling & Evans, 2015). By studying variation within each human population, researchers can identify population-specific disease risks and treatment responses, which often promote discovery in personalized medicine or guide specific disease treatment or prevention strategies within the population.(Oleksyk et al., 2022; Polimanti et al., 2014)

3.2 Databases as a force multiplier in human population research

Since the inception of modern genomics approaches for human population research, multiple initiatives have aimed to provide the scientific community with access to high-quality data from diverse human populations. These initiatives led to groundbreaking discoveries in human evolution, population genetics, and personalized medicine and accelerated the pace of research in those fields. One of the most impactful of these is The International Genome Sample Resource (IGSR), formerly known as the 1,000 Genomes Project (D. L. Altshuler et al., 2010; D. M. Altshuler et al., 2012; Auton, Abecasis, Altshuler, Durbin, Schloss, et al., 2015). By the time it was concluded in 2015, this international research effort had established a detailed, publicly available catalog of human genetic variation and currently provides public access to 3,202 sequenced individual genomes variant files from more than 26 populations, making it one of the most widely used human variation databases today (Birney, 2021). The European Bioinformatics Institute currently supports IGSR, hosts and maintains data from additional genomic projects, such as Human Genome Diversity Project (HGDP)(Bergström et al., 2020), the Simons Genome Diversity Project (SGDP)(Mallick et al., 2016) and other smaller projects (Fairley et al., 2020)

Several countries have launched their own initiatives to showcase the diversity among their citizens. These efforts include the 100,000 UK Genome Project(Smedley et al., 2021), the "All of Us" research program in the USA("The 'All of Us' Research Program," 2019), the Asian Genome Project(Wall et al., 2019), the African Genome Sequence Variation(Gurdasani et al., 2014) project, as well as whole-genome sequence population studies in the Netherlands(Boomsma et al., 2013), Korea(Jeon et al., 2020),

Ukraine(Oleksyk et al., 2021), and other nations. These initiatives have replaced global surveys and now serve as a significant global reference resource for human genetic variation, providing valuable resources for discovering disease variants.

Alternatively, projects like The Genome Aggregation Database (gnomAD) aggregate and harmonize both exome and genome sequencing data from a wide variety of large-scale sequencing projects (Karczewski et al., 2020). While not providing access to individual-level variant information, this database is useful for investigating the worldwide allele frequency of any allele for more than 76,156 individual genomes and more than 140 000 exome samples(Karczewski et al., 2020). The publicly accessible datasets have played a crucial role in enhancing our comprehension of human genetic variation by creating remote opportunities to incorporate the global scientific community into genomic research (Karishma Chhugani et al., 2022)

Massive databases of confirmed and predicted trait-causative genomic variants are established and updated constantly. These sources serve as critical tools for researchers investigating the genetic basis of human traits and diseases and are invaluable in studies that profile the human population. ClinVar and InterVar databases provide a freely accessible archive of reports of clinically proven relationships among human variations and phenotypes following robust guidelines(Landrum et al., 2014; Q. Li & Wang, 2017), while dbNSFP provides a comprehensive catalog of functional predictions and annotations, combining 36 deleteriousness prediction scores, nine conservation scores, and one loss of function score to assess the potential deleteriousness and for all human nonsynonymous, or protein-altering single nucleotide variants. (X. Liu et al., 2020) They allow researchers to identify genetic variants associated with specific traits or diseases, predict the impact of

a newly discovered mutation, understand their distribution across different populations, and predict their potential functional impact. This, in turn, has implications for understanding human evolution, identifying population-specific adaptations, and informing personalized medicine approaches. Overall, the well-cataloged body of prior research datasets of human populations provides additional value for every new research project in this domain.

3.3 Computational burden of human population genetics

Whole Genome Sequencing (WGS) is a vital tool for population genomics studies as it provides a comprehensive view of existing genetic variation, including single nucleotide variants and structural variants. Its ability to identify common and rare variants, as well as complex genetic patterns, makes it useful in understanding human evolution, migration, and disease susceptibility. Human population projects leverage the WGS technologies, enabling researchers to capture the full image of genetic variation in human populations (1000 Genomes Consortium, 2015). New projects that utilize WGS offer both opportunities and challenges, particularly in the areas of bioinformatics and data management (Goodwin et al., 2016).

However, using WGS in such projects generates an enormous amount of data. Due to the technological constraints of modern sequencing approaches a single genome sequence of an individual can often be over 100 gigabytes (Stephens et al., 2015a). This poses a significant bioinformatics burden, as the data requires processing, analysis, and storage in a way that allows for efficient retrieval and re-analysis if needed. The bioinformatics challenges associated with WGS data include sequence alignment, variant

calling, annotation, and interpretation. Ensuring data quality and reproducibility further adds complexity to the task.

In addition to the bioinformatics challenges, managing the vast amounts of data generated by WGS is a significant undertaking. This involves not only the physical storage of the data but also ensuring data security, maintaining data integrity, and providing access to researchers. The inclusion of previously discussed datasets of human variation and medical annotations provides invaluable information but also increases the bioinformatics burden on any study. For example, including only the gnomAD database requires at least extra 30 TBytes of additional storage (Karczewski et al., 2020).

Despite these challenges, the advantages of utilizing WGS in human population genomics are vast. The extensive scope of WGS data enables a thorough comprehension of genetic variation in human populations, which can yield crucial insights into human health and disease and aid in the creation of innovative therapeutic approaches. But it's important to note that such undertakings are not available for research groups without dedicated bioinformatics expertise and access to high-performance computing facilities.

3.4. Case study: Genomic deserts, blank spots on the map of genome diversity in Europe

This section will provide an overview of a first-author manuscript produced by OU Genomics lab, "The Pioneer Advantage: Filling the blank spots on the map of genome diversity in Europe", published on Sep 09, 2022, in GigaScience (Oleksyk et al., 2022). As a major contributor to the paper, I have led data collection and participated in its analysis. In addition, a second collaborative manuscript entitled "Genome Deserts", describing the availability of genome data around the world and surveying communities of scientists on

the challenges of the national genome projects is currently being prepared for submission (summer 2023) and has been encouraged to be submitted by the editors of Cell. I have been leading the statistical analysis on this manuscript and will be sharing the first authorship on this paper (Oleksyk Taras et al., 2023)

3.4.1 Background

Efforts to map global genome diversity were initially led by international groups such as the Human Genome Diversity Panel (HGDP)(Cavalli-Sforza, 2005) and the 1,000 Genomes (G1K)(Auton, Abecasis, Altshuler, Durbin, Abecasis, et al., 2015) project. These groups aimed to create a comprehensive map of human genetic diversity. However, while they had some success, their efforts have since dwindled, leaving gaps in our knowledge of human diversity, especially concerning the local and rare variants. In response, national projects have emerged to fill these gaps and provide locally-based annotations of disease variants (Oleksyk et al., 2022). These projects are backed by various entities, including governments, international collaborations, and enthusiast groups, and they continue to contribute to our understanding of human genetic diversity. In addition, efforts were also made to remove the ascertainment bias in genome data caused by the majority of genomes available from countries dominated by European ancestry. For example, in the African Genome Variation Project, as reported by (Gurdasani et al., 2014), approximately 13.4 million variants were characterized across 1,481 individuals from 18 ethnolinguistic groups, of which about 3 million were previously unreported. However, they lack a unified global strategy and provide uneven geographic and population coverage, resulting in a fragmented picture of genome diversity across the world.

Geographic genome surveys have confirmed the existence of population genetic structure in Europe, a concept first proposed by Luca Cavalli-Sforza in the 1990s(Cavalli-Sforza et al., 1996). Despite ongoing debates about the exact distribution of these genetic patterns, the correlation between geography and genetic distance in Europe is well-established. However, early surveys focused on Western Europe and European Americans and largely overlooked Northeastern and Eastern European populations. Subsequent studies in these regions revealed distinct phylogeographic patterning, further underscoring the importance of local genome variation for biomedical studies.(Novembre et al., 2008)

As of July 2022, there have been over 3,089 publicly accessible whole-genome sequences from various continental European populations in addition to 2,638 genomes released by research groups in Iceland(Gudbjartsson et al., 2015) and 204,109 in the United Kingdom(Smedley et al., 2021). However, most of the data comprises populations from the United Kingdom and developed countries of the European Union. Other regions, such as Eastern Europe and the Balkans, have only been represented by a few publicly available genomes distributed across more extensive global sequencing databases (Figure 15).

3.4.2 Discoveries from sequencing of an underrepresented population

Efforts like the Genome of Europe (Goe, n.d.) aim to fill the gaps in the genetic map of the continent by establishing a network of national genomic reference cohorts encompassing a minimum of 500,000 European individuals. However, as of 2022, this goal has yet to be realized. Even though the focus has shifted from international partnerships

to domestic projects, inequities persist in the dissemination of accessible human genome data from Europe.(Smetana & Brož, 2022)

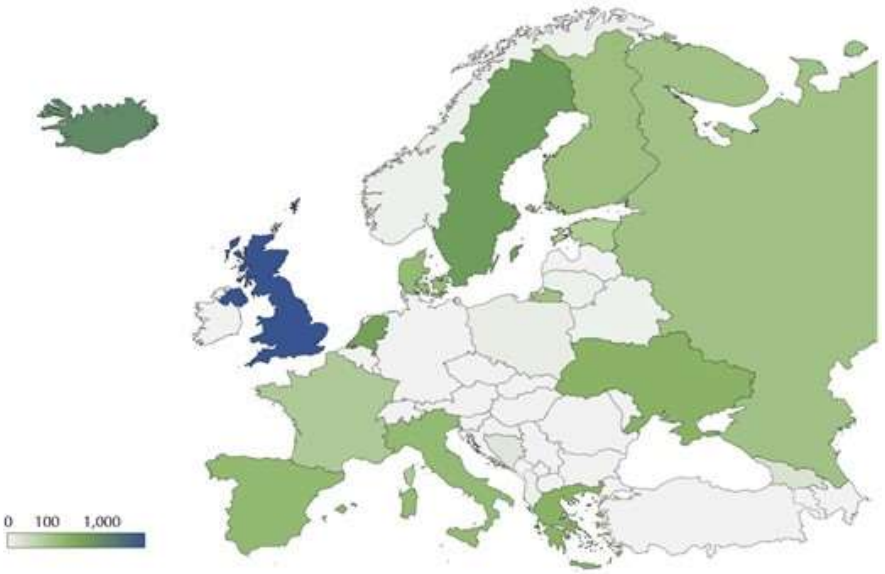

Figure 15 Public availability of whole-genome sequences in Europe.

Our recent whole-genome study conducted in Ukraine highlighted the lack of representation of Eastern European genomic diversity in the current data. The study revealed significant differences in SNP allele frequency, offering new insights into population structure and showing a unique Ukrainian cluster, distinct from other publicly available European populations or ethnic Russian subpopulations in the European part of Russia (Oleksyk et al., 2021). This study of fewer than 100 Ukrainian genomes revealed 478,000 novel genomic SNPs, a significant number even when compared to the most genetically diverse populations in sub-Saharan Africa, where three million novel variants were reported by the African Genome Variation Project (Gurdasani et al., 2014). The abundance of novel genomic variation is not due to high variation within the Ukrainian population but rather the lack of variation representation from the countries neighboring

Ukraine. This phenomenon is known as "first-mover" or "pioneer advantage," emphasizes the importance of comprehensive coverage of worldwide genome variation and the necessity to eliminate remaining "genome deserts." (Oleksyk et al., 2022)

3.4.3 Benefits, challenges, and conclusions

A comprehensive understanding of the geographic dimensions of local genome diversity in Europe is essential for the biomedical community. Countries would benefit from having their own national genome database to inform regionally relevant and objective public health policies due to their unique histories and socioeconomic structures. Public availability of data is essential to facilitate comparative discoveries and help nourish local communities of data-oriented scientists and researchers.

National genome projects are often funded from a mix of private and public sources, but countries with smaller economies tend to have less successful genomic initiatives. Ukraine stands out as an exception, with 254 genomes publicly available due to the successful pursuit of international collaborations with entities like the BGI, the National Institutes of Health (USA), and European Cross-Border cooperation grants. Generally, the establishment of national genomic projects requires addressing ethical considerations, raising awareness, securing funding, establishing legal frameworks, and building necessary infrastructure (reviewed in (Oleksyk Taras et al., 2023)). Countries that decide to pursue a national project could learn from or replicate different strategies to help fill the genome data gap.

National genomic projects provide an effective platform for training genome scientists and bioinformaticians at the national level. These projects should involve a multidisciplinary and interinstitutional team, including policymakers, lawyers, data

scientists, human geneticists, and humanities experts who understand relevant ethical and social issues. International collaborations can be beneficial in overcoming the complexity of such projects, not only for protocol development and sequencing but also for the ultimate objective of improving public health in each country.

The lack of established data-sharing practices in the scientific community exacerbates the uneven distribution of genome data in Europe. Commercial companies often compile massive amounts of data but restrict access to it. Data from ancestry testing and diagnostic sequencing is often challenging to retrieve, citing privacy concerns. Publishing individual genome data requires appropriate levels of informed consent and a combination of technical and societal conditions. The lack of consensus or legal precedents involving ownership or open release of biomedical materials and derivative data further complicates the issue.

Despite these challenges, public access to consented and open data is crucial. The best practice is to deposit collected genome sequences, informed consent details, and accompanying data into an international database, as demonstrated by the G1K consortium. This approach serves the global research community while protecting the interests of participating individuals and communities. Publicly available genome data from a country's general population is vital in unlocking the potential of genomic-based personal medicine, benefiting everyone.

3.5. Employing population genomics analysis to characterize genome diversity in Ukraine

This section will provide an overview of a shared first-author article produced by OU Genomics lab, “Genome diversity in Ukraine”, published on Jan 13, 2021, in

GigaScience (Oleksyk, Wolfsberger et al., 2021). As a major author and first equal contributor to the paper, I have led the data analysis step of the article and participated in conceptualization, data curation, and software development.

3.5.1. Background

Located at a historical crossroads of migration and trade routes, Ukraine is the largest country located entirely within Europe. Millennia of migration and admixture have shaped its population, resulting in a diverse genetic makeup with unique variations across the country. While most of the population is ethnically Ukrainian, there is a significant Russian minority in the southeast, as well as smaller minority groups historically present in different parts of the country, such as Belarusians, Bulgarians, Crimean Tatars, Greeks, Gagauz, Hungarians, Jews, Moldovans, Poles, Romanians, and Roma.

To fill the gaps in our understanding of genomic variation in Eastern Europe, which is often overlooked in global genomic surveys, we sequenced and provided genome data from 97 Ukrainians from Ukraine (UAU97). This is the first attempt to describe and evaluate genome-wide diversity in Ukraine. We aim to highlight the importance of studying local variation in the region and reveal the Ukrainian population's distinct and unique genetic components. Extra focus was directed towards the annotation of medically related variants, particularly those with allele frequencies that differ from neighboring populations, resulting in an annotated dataset of genome-wide variation in the genomes of healthy adults sampled across Ukraine.

3.5.2. Methodology

Sampling and DNA extraction: As part of the "Genome Diversity in Ukraine", we collected blood samples from healthy volunteers across different regions of Ukraine. Medical professionals oversaw the collection process at hospitals, and volunteers were familiarized with the study and collection procedure before giving full consent for their genotypic and phenotypic data to be publicly available. Each participant completed a questionnaire detailing their self-reported region of origin, grandparents' birthplaces, sex, and several phenotypic features. All personal identifiers were removed after the interview and sample collection, leaving only an alphanumeric identifier and a barcode.

Blood samples were collected into two 5-mL EDTA tubes by a certified nurse or phlebotomist, assigned a barcode number, and shipped on dry ice to a certified biomedical laboratory in Uzhhorod, Ukraine, for immediate DNA extraction. The remaining blood and DNA samples are stored at the biobank of the Biology Department, Uzhhorod National University, Ukraine. In total, we collected blood samples from 113 individuals.

After arriving at the laboratory, DNA was isolated from a blood sample of 200 µL using the innuPREP DNA Blood Minikit (IST Innuscreen GmbH, Berlin, Germany). The quality of the DNA was visually verified on a 2% agarose gel. Out of the total 113 samples submitted, we successfully extracted DNA from 97 samples and subsequently normalized concentrations to 20–30 ng/µL for downstream use. Samples were recorded and sent to NIH for genotyping, and a second sample was sent to a BGI facility (Shenzhen, PRC) or Psomagen Inc. (Gaithersburg, MD, USA) for whole-genome sequencing. Finally, we maintained the remaining ~2 mL of each blood sample at -80ºC for backup and future use.

Sequencing and genotyping: The genomic DNA of the 97 samples underwent quality and concentration checks upon arrival at the BGI facility in Shenzhen, PRC. Fragmentation, end-repair, and ligation of adaptors were performed on the DNA, followed by PCR product purification and heat denaturation. The resulting double-stranded PCR products were circularized to form the final library. High-density DNA nanochip technology was used to load the DNA nanoballs (DNBs) into a patterned nanoarray, where pair-end 100-bp reads were obtained through combinatorial probe-anchor synthesis. Adaptor sequences, contamination, and low-quality reads were removed from the raw data. All 97 whole genome samples submitted were successfully sequenced.

In addition, we created a reference database to control for technical errors associated with the use of specific technology. First, a single individual was resequenced at a depth of approximately 64x, at Psomagen Inc. (Gaithersburg, MD, USA) using the TruSeq DNA PCR Free 350 bp protocol by Illumina producing 150-bp-long reads. Second, the Illumina Infinium Global Screening BeadChip Array-24 v1.0 (GSAMD-24v1–0) was used to genotype all 97 samples for 700,078 loci at the National Cancer Institute's Division of Cancer Epidemiology and Genetics in Bethesda, MD, USA. The data were analyzed using the standard Illumina microarray data analysis workflow, with samples filtered for contamination, completion rate, and relatedness during quality control (QC). Ancestry assessment was performed using the SNPweights(Mayakonda et al., 2018) software and a reference panel consisting of European, West African, and East Asian populations. All samples were attributed to the European ancestry group. After QC and sample exclusion, 87 samples with 689,918 loci and a completion rate of 99.9% were retained for further analysis.

Variant calling and quality control: We used the BGISEQ-500 platform to produce sequencing data for 97 samples and utilized Sentieon tools to analyze the data. The reads were aligned to the GRCh38 human reference genome through BWA-MEM, and mapped reads were prepared for variant calling with GATK (Van der Auwera et al., 2013a). This included marking duplicates, indel realignment, and base quality score recalibration. GATK HaplotypeCaller was used for SNP and indel discovery for each individual, which were then merged into a single pVCF using bcftools (Danecek et al., 2021). We also performed mobile element discovery with *MELT* (Gardner et al., 2017)and structural variant discovery with lumpy-sv(Layer et al., 2014) and Smoove. (https://github.com/brentp/smoove) GangSTR(Mousavi et al., 2019) was used to call short tandem repeats, and dinumt(Van der Auwera et al., 2013b) to call nuclear mitochondrial DNA.

To ensure consistency, we compared variant files across three platforms: BGISEQ-500 sequencing, Illumina genotyping, and Illumina NovaSeq6000 S4 sequencing. Illumina genotyping was conducted on 86 of the 97 samples previously sequenced with BGISEQ-500. One sample was also sequenced with Illumina NovaSeq6000 S4. For comparison purposes, the variant detection programs were rerun without joint calling for the BGISEQ-500 sequencing.

We compared the SNPs from the WGS platforms to those identified using the Illumina SNP array for both matching position and matching genotype. Structural variants and MEIs were compared between the WGS platforms. Variants were considered the same if they had 95% reciprocal overlap. We found that Illumina identified a higher number of larger variants than BGISEQ-500, possibly due to its higher coverage or because the

variant identification tools were built to detect variation from Illumina sequencing data and, therefore, may not be able to detect variants in BGISEQ-500 data as accurately.

Variant annotation: We used ANNOVAR(Wang et al., 2010) and SNPEff (Cingolani et al., 2012)software to annotate sequence variant files with the reference databases of GRCh38. RefSeq Gene, 1GP superpopulation, and dbSNP150 with allelic splitting and left normalization were utilized for ANNOVAR annotations. ClinVar Version 20200316, InterVar gnomeAd Version 3.0, and dbnsfp Version 35c(Karczewski et al., 2020; Landrum et al., 2018; Q. Li & Wang, 2017; X. Liu et al., 2016) were used to annotate medically related and functional variants. We also supplemented the default GRCh38 annotation database with ClinVar and GWAS catalog database annotation using the snpSift tool (Cingolani et al., 2012).

Population analysis: We analyzed the Whole Genome Sequencing (WGS) variants using Principal Component Analysis (PCA), merging them with samples from neighboring countries from IGSR and HGDP datasets. Eigensoft(Patterson et al., 2006) was used to conduct the analysis, filtering the resulting dataset by genotyping rate and pruning for variants in linkage disequilibrium. We remained with 677 samples and 208,945 variants. Custom Python programming language was used to visualize PC1 and PC2.

For the model-based structure analysis, we used the same dataset as in the PCA and performed the analysis using ADMIXTURE(Alexander et al., 2009b) software. The optimal K parameter was identified using the 10-fold cross-validation function of ADMIXTURE, with K = 3 resulting in the lowest error. Python programming language was also used to construct a population structure plot.

To estimate inbreeding coefficients, we pruned the samples for genotyping rate and linkage disequilibrium and used the resulting dataset containing the remaining 117,641 loci from 84 samples. Various inbreeding estimates were performed.

3.5.3. Results

Variant calling and confirmation: We estimated the number of passing bi-allelic SNP calls for each sample in the database, filtering out about 12% of these based on excess heterozygosity and low variant quality scores. We also removed 4% of indels that did not pass the quality filters. We found a good correspondence between the SNP calls made using BGISEQ-500 and NovaSeq 6000 S4 data, with a strong overlap across all three technologies. The TITV ratio for novel SNPs was lower than that for SNPs in the dbSNPs database, which may indicate our improved ability to detect small insertions in newer sequencing technologies.

In addition to SNPs, we identified and quantified major classes of structural variations in the Ukrainian population, including small indels, large structural variants, and mobile element insertions (MEI). A significant proportion of large variants and MEIs were not previously reported in the 1GP Database. We found a significant correspondence between the calls made using BGISEQ-500 and Illumina NovaSeq 6000 S4 data, with both platforms showing a significant overlap in calling indels.

We compared our database to existing global resources of population variation, such as gnomAD (Karczewski et al., 2020) and the IGSR database ((Fairley et al., 2020)). Small variants were considered "novel" if they were absent from all the samples in the two global datasets. Large structural variants and MEIs were considered novel if they were not

present in the gnomAD and IGSR databases. A summary of the variation is presented in Table 1.

Annotation for functional and medically relevant variants: Our study focused on the distribution of functional variation on phenotypes, particularly those with medical relevance. Our analysis revealed an average of over 8,000 mutations within the coding regions for every individual. Additionally, we identified numerous regulatory variants, including 2,229 transcription factor binding site ablation (TFBS) mutations. Our findings included 43,892 benign mutations in medically related genes, 189 unique pathogenic or likely pathogenic variants, and 20 protective or likely protective alleles, as defined by ClinVar (Landrum et al., 2018).

Many of the identified variants have been previously reported in genome-wide association studies or ClinVar and are known to be medically significant. On average, each individual in our study carried 19 pathogenic and 12 protective mutations. We found more shared variants with the GWAS than the ClinVar catalog, but it is important to note that GWAS variants are not always useful for predicting pathogenic phenotypes. We created a list of known disease variants with different frequencies in Ukrainians compared to the EUR superpopulation from IGSR and the RUS population from HGDP. The variants we identified are linked to various medical conditions, in hyperglycinemia/iminoglycinuria, bisphosphonate response efficacy, autism, Leber congenital amaurosis, and breast cancer susceptibility in carriers of *BRCA1* and *BRCA2* genes.

Estimation of population structure: We conducted several population analyses to highlight the unique and valuable nature of our new dataset. Our findings suggest that the genetic diversity of the Ukrainian population has been uniquely shaped by evolutionary

and demographic forces, making it an important consideration for future genetic studies. However, we did not evaluate any historical hypotheses regarding the origins, founding, migration, and admixture of this population.

Table 1. Summary of variation in the 97 whole-genome sequences from Ukraine

	All samples			Mean per sample	
Sequencing results			**Novel % gnomAD (1000 Genomes)**	**All**	**Novel %**
Total sequence reads	99.8 Bn			1.03 Bn	
Mean coverage	97 Samples at 30×			30×	
Variation					
	No. of total unique variants	Novel gnomAD count			
SNPs	13,010,979	477,564	3.7 (12.6)	3,488,083	0.1 (0.7)
Bi-allelic	12,667,283	470,667	3.7 (12.7)	3,340,557	0.3 (0.6)
Multi-allelic	343,696	6,897	2.0 (7.4)	146,340	0.8 (4.7)
Small indels	2,727,604	76,484	2.8 (7.4)	917,731	0.3 (1.0)
Deletions	1,805,739	55,599	3.1 (9.0)	624,919	0.3 (2.4)
Insertions	14,459,87	30,453	2.1 (6.7)	571,461	0.2 (2.1)
Structural variants					
Large deletions	16,078	10,914	67.9 (48.3)	3,524	52.6 (19.1)
Large duplications	1,845	1,356	73.5 (42.3)	562	89.4 (35.2)
Inversions	337	314	93.2 (47.8)	185	94.1 (48.6)
Mobile element insertions					
Alu	2,316	1,805	77.9 (38.1)	473	68.1 (18.0)
L1	451	289	64 (50.1)	79	60.8 (27.8)
SVA	100	75	75 (52.0)	20	70 (50)
NUMT	714			16	

To illustrate how our dataset contributes to the genetic map of Europe, we examined the genetic relationships between Ukrainian individuals within our sample and assessed the

genetic differences between this population and its immediate neighbors in Europe. We conducted a principal component analysis (PCA) of a merged dataset of 654 samples, which included European populations from the 1KG (Utah residents [CEU] with Northern and Western European ancestry, Toscani in Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], Iberian population in Spain [IBS] and French (FRA) and Russian (RUS) populations from the HGDP, as well as the relevant high-coverage human genomes from the Estonian Biocentre Human Genome Diversity Panel (EGDP: Croatians [CRO], Estonians [EST], Germans [GER], Moldovans [MOL], Polish [POL], and Ukrainians [UKR]) [43] and Simons Genome Diversity Project (Czechs [CZ], Estonians [EST], French [FRA], Greeks [GRE], and Polish [POL])(Fairley et al., 2020)

The Ukrainian genomes from this and other studies form a single cluster positioned between Northern (Russians, Estonians) and Western European populations (Utah residents with Northern and Western European ancestry, French, British, and Germans). There was significant overlap with other Central and Eastern European populations, such as Czechs, Polish, and people from the Balkans. This overlap is not surprising given the close geographic proximity of these populations and may also reflect the incomplete representation of samples from surrounding populations. The PCA analysis results were plotted and shown in Figure 16.

We used the ADMIXTURE package to broaden our understanding of the genetic makeup of the Ukrainian population. By doing so, we were able to construct a preliminary image of the potential ancestry contributions and population admixture. We identified the optimal K by using a 10-fold cross-validation function within the K = 2-6 range. With the optimal K being 3, we were able to highlight the similarities and differences of the

Ukrainian population compared to other populations in Central and Eastern Europe. The results of admixture testing are presented in Figure 17.

Even though all samples were collected from individuals who identified as ethnic Ukrainians, we noticed two outliers. Sample EG600048 clustered with Southern Europeans, which highlights the importance of recognizing the unique composition of a population, as overlooking this can result in biased biomedical studies. While genetics is not a reliable determinant of ethnicity, it can help assess individual ancestry contributions. In order to aid future ancestry studies, we provide a complete list of candidates for Ancestry Informative Markers (AIMs) that differentiate Ukrainians from neighboring populations in Europe.

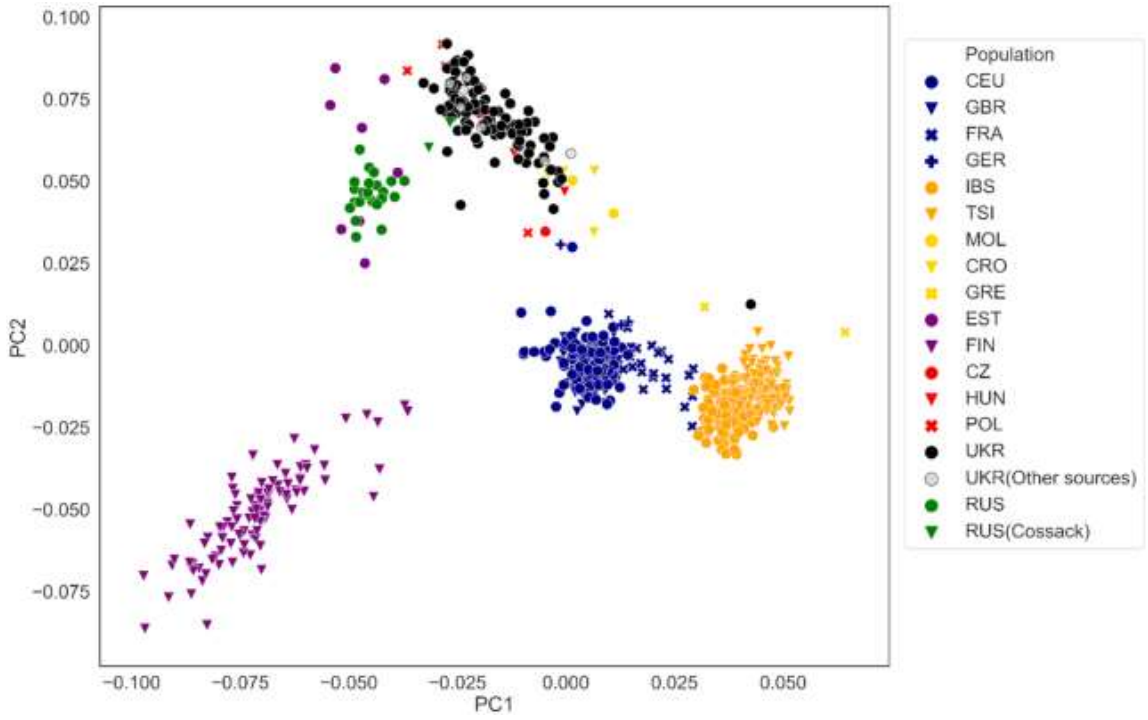

Figure 16 Principal component (PC) analysis of merged dataset containing European populations. Colors reflect prior population assignments from the European samples from the 1KG and HGDP. The Ukrainian samples of this study are marked as black circles. The other samples are color-coded according to their clustering and tentatively represent Western, Southern, South-Eastern, North-Eastern, and Eastern European populations.

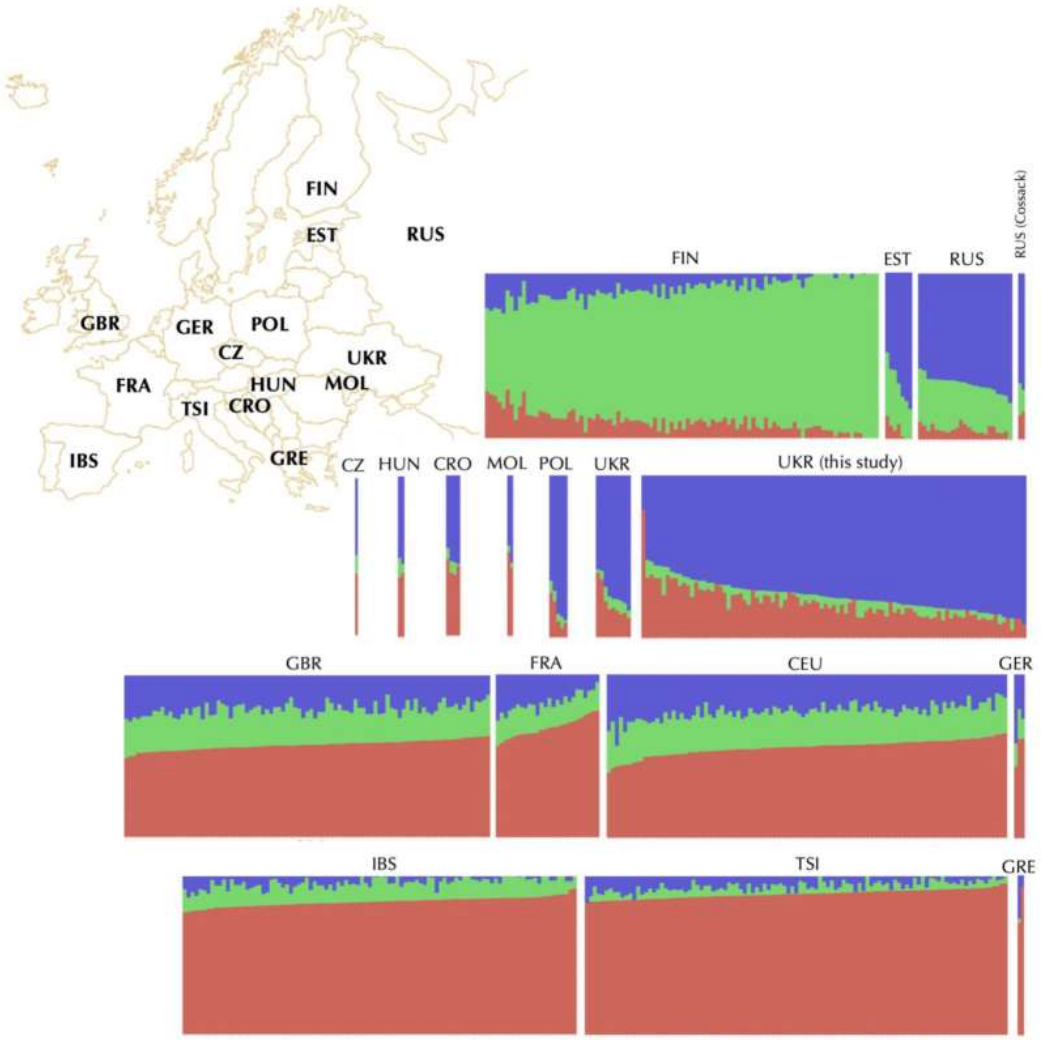

Figure 17 Admixture of ancestral components for Ukrainian and neighboring populations
Admixture plot constructed using ADMIXTURE package at K = 3 illustrates similarities and differences between genomes from this study as well as samples from the 1KG (Utah residents [CEU] with Northern and Western European ancestry, Toscani in Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], and Iberian population in Spain [IBS]) and French (FRA) and Russians (RUS) from HGDP, as well as the relevant high-coverage human genomes Croatian (CRO), Czech (CZ), Estonian (EST), German (GER), Greek (GRE), Hungarian (HUN), Moldovan (MOL), Polish (POL), Russian Cossack (RUS), and Ukrainian (UKR) from the Estonian Biocentre Human Genome Diversity Panel (EGDP) as well as Simons Genome Diversity Project

Discussion and Conclusions: Ukrainians possess many known and several novel genetic variants that have clinical and functional significance. These variants often display allele frequencies that are different from neighboring populations in the rest of Europe. While several large genome projects contribute to understanding global genetic variation,

many rare and endemic alleles have not yet been identified by international databases such as 1KG. We expect that future sampling and sequencing will continue to enhance and complete our understanding of the genomic diversity in people across Ukraine. This will significantly contribute to the development of genetic approaches in biomedical research and applications.

.

CHAPTER FOUR

KNOWLEDGE DISSEMINATION AND DEVELOPMENT OF ACCESSIBLE POPULATION GENOMICS ANALYSES SOFTWARE TOOL

4.1 Bioinformatics as the backbone for modern genomics

Population genomics, as a field, would likely not exist in its current form if not for the technological advancements in sequencing and data analysis. Bioinformatics and the development of computational genomic data analysis have originated due to the proliferation of computing and are at the core of most of the approaches and discoveries in modern population genomics. (Bartlett et al., 2017) They has provided the necessary tools to handle and analyze the massive amounts of raw data of data generated by genomic studies via high-throughput sequencing technologies. Without bioinformatics, managing and making sense of these large datasets would be extremely time consuming, making it nearly impossible. With modern software and algorithms, researchers can efficiently analyze raw sequence data, identify variants, predict their functional effects, and examine patterns of genetic variation within and across populations. Bioinformatics has paved the way for the creation of advanced statistical techniques that aid in the analysis of population genomic data. These techniques include the ability to deduce population structure, identify natural selection, and estimate demographic parameters such as population size and migration rates. As a result, these methods have provided invaluable insights into the evolutionary processes that influence genetic variation in populations.

Additionally, these advances has facilitated the integration of different types of data in population genomics studies and opened the door to answering multi-domain research questions. For example, it is now possible to integrate genomic data with environmental,

phenotypic, and geographic(Van Assche et al., 2015) information to study how various factors influence genetic variation within and among populations. This integrative approach has greatly enriched our understanding of population genomics. However, the growing complexity of modern research combined with the relative novelty and lack of standardized approaches for analyses establishes complex barriers to research groups starting their population genomics projects. Essentially, while bioinformatics has made it possible it didn't necessarily made it accessible.

The development of streamlined, reproducible, and scalable workflows, combined with the dissemination of knowledge, can significantly enhance the accessibility of bioinformatics. These efforts can help to democratize bioinformatics, enabling a wider range of individuals to contribute to research and benefit from its findings.

4.2 Scalable, reproducible workflows: Snakemake Overview

This dissertation presents an open-access workflow to perform population genomics analyses based on the Snakemake engine(Köster et al., 2021). A powerful tool in computational data analysis, workflow management systems have gained popularity due to the various advantages they offer. In this regard, we will briefly overview its features. Snakemake is a language-agnostic system that can integrate code from different programming languages, such as Python, R, and shell scripting, allowing bioinformaticians to utilize the best tools and packages available across various languages. This enhances the efficiency and effectiveness of their workflow.

Snakemake emphasizes reproducibility, which is crucial in scientific research. It defines the dependencies between different steps in a workflow, ensuring that each step is executed with the correct input and in the correct order. This explicitness makes workflows

easier to understand and debug and ensures that they can be reliably reproduced by other researchers. Additionally, it supports containerization technologies like Docker and Singularity, which further enhance reproducibility by ensuring that the same software environment is used every time a workflow is run.

Scalability is another advantage of Snakemake workflows, making them suitable for large datasets. The engine can automatically parallelize tasks where possible, allowing for efficient use of computational resources. The workflows can be executed on a wide range of computing platforms, from single-core laptops to multi-core servers and high-performance computing clusters, providing scalability in terms of computational resources. Lastly, it takes a modular approach to workflow development, breaking them into smaller, reusable components. This promotes code reuse, making workflows easier to maintain and extend. The modularity also makes it easier to share and collaborate on, which is crucial in a field like bioinformatics, where collaboration is essential. (Köster et al., 2021)

4.3 PopGen Playground – accessible workflow for population genomics analysis

Utilizing Snakemake and the knowledge accumulated via previous research, we developed and incorporated the majority of population genomics approaches described in this dissertation into a single destination suite - PopGen Playground, hosted on GitHub(https://github.com/wwolfsberger/OU_popgen_playground). It aims to perform population genomics analysis of Variant Call Files. It addresses the questions not only of specific analyses, automates data wrangling, visualization, and modification steps, such as format conversion and phasing, and removes the need to produce exploratory plots. Incorporating isolated environments via the Conda package management system("Anaconda Software Distribution," 2020) for every stage of analysis ensures

streamlined computation and prevents unknown variables from interrupting the run. The system is set up by specifying a checklist of desired analyses via a configuration file and supports a scalable approach for high-performance computing, like Matilda OU Cluster.

Supported by a detailed manual explaining the analyses involved, it provides a robust system for performing population genomics research for both animal and human populations. The complete list of available analyses and their overview is presented in Table 2.

Table 2, Analyses incorporated in PopGen Playground suite.

Software\tool	**Class**	**Purpose**
PLINK(Chang et al., 2015)	Pedigree, Identity by descent/state, data wrangling, PCA test	Estimating inbreeding and relatedness, estimating structure, LD pruning, data conversion
ADMIXTURE(Alexander et al., 2009a)	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci
detectRUNS (R package)(Chang et al., 2015; Marras et al., 2015)	Runs of Homozygosity, Inbreeding based on runs of homozygosity	Estimating inbreeding and relatedness, detection of signals of selection
Shapeit4(Delaneau et al., 2008)	Variant call phasing,	haplotype detection, missing data imputation
Rehh (R package)(Gautier et al., 2017)	Extended Haplotype Homozygosity, iHS, Rsb, XP-EHH	Within population and between population site-specific or alleles-specific signals of selection
snpSift(Cingolani et al., 2012)	IGSR, dbGap, ClinVar, GWAS catalog, dbNSFP databases	Automated human database data retrieval and annotation, prediction of deleteriousness
snpEFF(Cingolani et al., 2012)	Variant annotation	Functional prediction of variant deleteriousness

4.4 Bioinformatics knowledge dissemination

The adaptability of bioinformatics for remote work is mainly attributed to the digital format of the data and the software employed for analysis. Once genomic data is sequenced and digitized, it can be stored and accessed from anywhere globally. Additionally, bioinformatics tools are commonly software programs that can be

downloaded and executed on any computer connected to the internet. This allows bioinformaticians to work remotely from various locations, such as a research lab, home office, or café, as long as they have a computer and internet connection.

Despite the setbacks, challenges, and complexity modern bioinformatics faces due to technological advancements, it is, ironically, one of the most accessible fields in modern genomics worldwide, as bioinformatics is naturally poised for remote work (Marx, 2013). This has opened exciting opportunities for global collaboration and training in the field. Accessibility has significant implications for global training opportunities in bioinformatics. With the proliferation of online learning platforms, it is now possible to offer quality bioinformatics training to individuals around the world. These platforms can provide interactive, hands-on training in a variety of bioinformatics skills, from basic programming to advanced genomic data analysis (Via et al., 2011). For example, Coursera has major universities offering comprehensive bioinformatics training for free with a paid certificate for a relatively minor fee. This can help to democratize access to bioinformatics training, making it possible for individuals in low- and middle-income countries to gain the skills they need to contribute to the field (Welch et al., 2014).

From a global perspective, the ability to work remotely can facilitate collaborations in bioinformatics, combined with a worldwide build-up in bioinformatics expertise worldwide (Mangul et al., 2019). Researchers from different parts of the world can work together on shared datasets, combining their expertise to tackle complex bioinformatics challenges. These collaborations can lead to new insights and innovations, advancing our understanding of biology and medicine.

4.5 Supporting scientists worldwide - Scientists without Borders: Lessons from Ukraine

This section will provide an overview of a first-author manuscript produced by OU Genomics lab, “Scientists without Borders: Lessons from Ukraine”, currently in press at GigaScience.

4.5.1 Background

Countries worldwide face a variety of challenges, such as political upheaval, wars, economic crises, and natural disasters. These difficulties often result in scholars and students being displaced, cutting off their ties to professional communities and creating uncertainty about their futures. Established researchers may lose funding, workplaces, laboratories, and communication lines, potentially losing invaluable samples and data. Both undergraduate and graduate students may experience career interruptions, leading to disengagement and dropout from educational opportunities and research careers. However, the international community can aid these groups, and every success or failure in these efforts offers valuable lessons for addressing future global challenges.

The increasing interconnectedness of the scientific world has made it potentially more accessible for the international community to assist researchers facing hardships. This is largely due to investment and engagement in global projects. Collaborative initiatives such as the Global Alliance for Genomics and Health(Health*, 2016), the International Brain Initiative(Quaglio et al., 2021), the Human Genome Project(Gibbs, 2020), and the Large Hadron Collider(Horyn, 2022) have achieved significant success by facilitating the exchange of resources, expertise, and data on a global scale.

The collective knowledge of these collaborative groups is widely disseminated through scientific conferences and open publishing. The advent of online preprints, journals, and open-access publishing has revolutionized knowledge distribution by making information more accessible. Global research networks have emerged across various fields, and initiatives like the Open Science Framework and the Open Access movement(Miedema, 2022) have further promoted knowledge sharing.

Web-based technologies in projects, especially in the emerging omics sciences, reduce the need for in-person engagement. The major challenges in the early stages of the Human Genome Project and the 1,000 Genomes Project were data acquisition, which required specific laboratory skills and on-site resources for sample collection and sequencing(J. Li et al., 2021). Now, researchers and institutions can easily accumulate and share medical and other phenotypic data digitally. There is an increasing availability of low-cost DNA, RNA, and protein sequencing data from research and high-throughput screening, which international groups can jointly analyze.(Cantelli et al., 2021)

The main challenge these groups faced in the decades following the publication was the lack of access to powerful hardware and software to analyze the influx of data. This problem is now solvable with advanced clusters, increased storage, and cloud computing. However, to grow, collaborations that rely on building large international networks must maximize global human engagement. Providing opportunities to scientists disconnected due to history, war, political, economic, or ecological crises in countries facing hardship is crucial.

To support researchers and students in countries facing hardship, we suggest several effective and feasible mechanisms based on our experience during the year of

conflict in Ukraine. These include establishing direct remote research collaborations and joint publication, donating funds and directly sharing resources, sharing information about research opportunities, participating in joint online events inside the country in crisis, promoting more accessible alternative mechanisms to publish in scientific journals, promoting more straightforward access to international conferences, enabling opportunities for remote education and training, and encouraging funding agencies to provide opportunities, specifically targeting the research communities facing hardship.

4.5.2 Summary of the global response to the war in Ukraine

The global scientific community responded to the war in Ukraine, swiftly creating online resources to support affected colleagues. These resources, often in the form of open spreadsheets, listed various forms of support, including short and long-term positions and internships. Initiatives such as the University of Oregon's list of laboratories ready to assist displaced Ukrainian scientists and the larger collaborative initiative, #ScienceForUkraine, emerged to provide information on opportunities for Ukrainian researchers. Other groups focused on preserving Ukrainian cultural heritage, such as Saving Ukrainian Cultural Heritage Online (SUCHO), which engaged in web-archiving, digitizing, and preserving heritage from Ukrainian cultural institutions.

Various organizations and institutions stepped in to provide support. The US National Academies' Policy and Global Affairs Division, in collaboration with the Polish Academy of Sciences, created the Scientists and Engineers in Exile or Displaced initiative, providing short-term stipends for Ukrainian scholars studying outside Ukraine. The Shevchenko Scientific Society, a publicly funded organization, also created opportunities for displaced scientists. National academies in the EU issued supportive statements to

Ukraine, and the European Federation of Academies of Sciences and Humanities (ALLEA) began hosting scholars displaced by the war.

The widespread support from the scientific community is crucial for the recovery and mitigation of the war's impact on academic and scientific communities in Ukraine. The networks established at the start of the war laid a foundation for the further integration of Ukrainian scientists into the global community. These networks are instrumental in the success of more formal organizational initiatives involving Ukraine that are currently under development.

Many Ukrainian scholars could not take advantage of global opportunities by moving abroad due to martial law restrictions, visa complications, and family responsibilities. Most of them stayed in Ukraine, where they faced instability and vulnerability as universities in the east closed or relocated. The war effort led to budget cuts for state-run universities, affecting salaries and research funding. Despite these obstacles, many Ukrainian researchers remained committed to their academic and research pursuits. (Gaind, 2022)

The first year of the war offered various opportunities to Ukrainian scientists, as demonstrated by a survey of the *#ScienceForUkraine* database. This initiative aimed to support and mitigate setbacks for the Ukrainian academic community during the crisis. As the largest publicly accessible data source, it sheds light on the types of opportunities extended since the conflict's inception. The scientific community's initial response to the crisis was to support refugees fleeing Ukraine, drawing parallels to previous situations where scholars had to escape political instability, such as in Venezuela, Afghanistan, and

Syria. The expectation was that the Ukrainian war would likewise generate a significant influx of displaced researchers requiring assistance.

4.5.3 Analysis of opportunities for Ukrainian scientists listed in the #ScienceForUkraine

Our analysis of the #ScienceForUkraine database revealed a dominant trend towards in-person opportunities, with a noticeable absence of remote options (Figure 18). The majority of the listed 1,871 opportunities were classified as "paid positions", while "educational" and "funding programs" were distributed in nearly equal proportions. These opportunities were almost evenly distributed between researchers and students, with professionals forming a minor fraction. Most were based in Europe, especially in Germany, which offered paid positions, education, and funding programs. Despite these concerted efforts, the success of such initiatives largely depended on the readiness of Ukrainian scientists to relocate abroad, emphasizing the stark lack of remote opportunities for those unable to leave the country. Consequently, many opportunities remained unutilized.

In order to assist scholars, especially those in conflict-ridden areas like Ukraine, the international community must develop both short-term and long-term strategies. These strategies should cater to those who have fled their country and those who have decided to remain. It is essential to focus on scalable mechanisms that can benefit a large number of affected individuals. Remote positions should be expanded to provide more opportunities. The global scientific community should be more involved, and effective training should be offered to bridge any knowledge gaps. These measures would help Ukrainian science survive difficult times and integrate into the global scientific community.

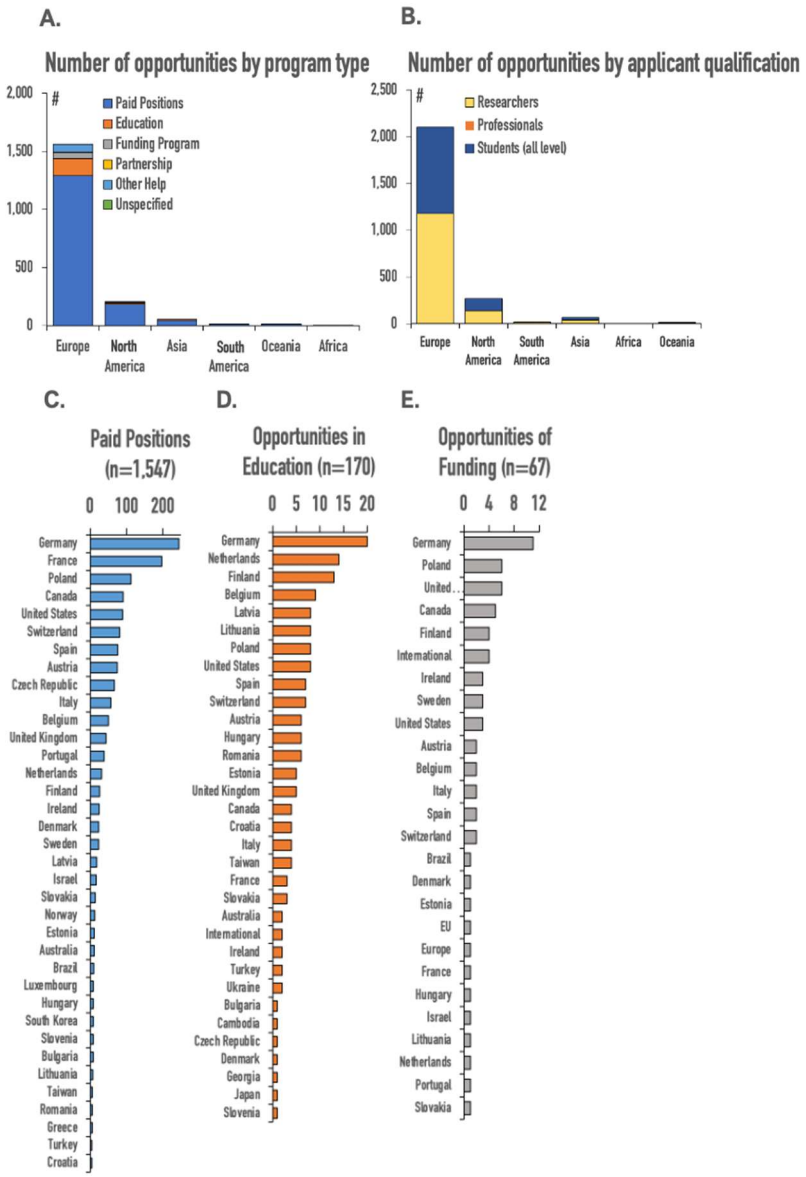

Figure 18. Locations of opportunities for Ukrainian scientists in the *#ScienceForUkraine* database in one year since the start of the conflict (February 2022-February 2023).

A. Number of opportunities by program type. The majority of the 1,871 listed opportunities can be classified as paid positions. These include research and postdoctoral fellowships, as well as some researcher jobs. Most of these are in Europe (1,293), with North America as a distant second (181). Educational opportunities and funding programs were distributed in similar proportions.

B. Number of opportunities by applicant qualification. Opportunities were distributed roughly equally between researchers and students of every level, with very few indicated as "*professional*".

C, D, and E. Distribution of different types of opportunities (listed in **Figure 1.A**) by country. The total number of opportunities per category is given in parentheses. Germany had the most opportunities in all categories.

There are two types of effective remote opportunities: those that involve establishing contact with specific scientists or groups in Ukraine and those that offer support indirectly without requiring contact with specific researchers. Remote collaborations such as data sharing, co-authoring papers, and offering expertise in a specific field could be beneficial. To make these collaborations work, individual scientists and research groups should invite their Ukrainian counterparts to take part in research projects and author joint publications. Funds could also be donated to research institutions in Ukraine, or resources could be shared, such as access to scientific journals, databases, online tools, and software. Modifications to the policies of scientific publishers, research foundations, and academic institutions can also aid Ukrainian researchers.

Training is essential to equip Ukrainian researchers with the skills needed for international collaboration and remote work. Funding initiatives should be put in place to train scientists in computational and bioinformatics skills, thereby facilitating their remote engagement with international consortiums and establishing collaborations with the global community. Lastly, the development of a platform matching scientists and computer experts with laboratories and faculties offering remote mentorships or employment could prove invaluable. This platform would present a feasible solution to supporting at-risk researchers and students.(Mangul et al., 2019)

Timely implementation of a coordinated assistance plan targeting scientists at risk in invaded countries by decision-making authorities is essential. This would involve establishing new research funding opportunities for international collaborations, including joint research and academic projects and programs. Existing organizations like the United States Agency for International Development (USAID) should play a central role, but the

implementation must go beyond them. A combined effort from governments, the private sector, and the scientific community is necessary to provide the support that these scientists desperately need.

4.5.4. Effective mechanisms of remote support scientists in countries facing hardship

The current crisis in Ukraine, precipitated by Russia's unprovoked invasion, presents vital lessons that can be applied to similar situations worldwide. By examining Ukraine's circumstances, we have identified several potential pathways to assist scientific communities facing adversity. These methods are not exclusive to Ukraine but can be adapted to other situations borne out of war, persecution, or natural disasters, as witnessed in Venezuela, Afghanistan, Syria, Puerto Rico, and Turkey. We have outlined effective and feasible avenues to support scientists and students disadvantaged by military, political, economic, or ecological crises (Table 3). However, this list of opportunities is by no means exhaustive and should act as a conversation starter rather than a conclusion.

An urgent dialogue needs to commence regarding a feasible long-term plan to rebuild science and research opportunities and establish new data and computational science directions. We need to create programs enabling scientists who cannot leave their home countries to acquire novel skills remotely and disseminate them within their local scientific communities. For those researchers remaining in their country, transitioning to computational, data-driven research can provide a viable solution. This approach, which relies on remote teaching and the accessibility of resources from anywhere in the world, necessitates systematic training in cutting-edge bioinformatics and data analysis skills. To achieve this, we should establish collaborations with world-leading institutions and create

remote mechanisms for world-class specialists to train undergraduate and graduate students. As science becomes increasingly interconnected globally, we can integrate scientists from hardship-stricken countries into the international community, allowing them to contribute to global research efforts remotely and help tackle global challenges.

Table 3. Mechanisms to support scientists and students from countries experiencing hardship.

Group 1 (A-D) includes efforts that are directed towards and imply establishing contact with a specific scientist or a group of scientists in Ukraine. Group 2 (E-H) In the second group, there are venues to help Ukrainian researchers indirectly, without knowing or contacting specific researchers or scientific groups.

	#	Mechanism	Description
Group 1	A	**Establish direct remote research collaborations**	A scientist or researcher can collaborate remotely with Ukrainian scientists directly on common research projects.
	B	**Donate funds and provide access to resources**	Funds can be donated directly to research institutions to support their work
	C	**Share research opportunities**	Informing colleagues about opportunities that could be missed because of limited access or language barriers.
	D	**Participate in online events inside the country**	Taking part in webinars, online conferences, and workshops put together by local organizers
Group 2	E	**Promote publications in the scientific journals**	Encouraging journals to support researchers by waiving publication fees, and providing translations and editorial services, and including versions of published papers in authors' native languages.
	F	**Promote participation in international conferences**	Encourage organizers of scientific conferences to provide free or discounted access to registration and memberships.
	G	**Enabling opportunities for remote education and training**	Provide and encourage other groups to provide online training to provide have adequate competitive for the multitude of positions offered by the global scientific community.
	H	**Encourage funding agencies to provide funding opportunities for the research communities facing hardship**	Encourage funding agencies, non-profits, foundations, governmental and academic institutions to establish new research funding opportunities for international collaborations, prioritizing joint research and academic projects and programs.

CHAPTER FIVE
OVERALL CONCLUSIONS AND FUTURE DIRECTIONS

5.1. Conclusions

In conclusion, this dissertation addresses the challenges in the rapidly expanding field of population genomics by providing a comprehensive review and structuring a methodical approach toward extensive population genomics analysis. The study focusing on the origins and evolutionary history of Puerto Rican horses offers a valuable model for the bioinformatic analysis of animal populations under intense evolutionary pressures, whether from adaptation or drift. This research not only resolves historical questions regarding island horse breeds but also sheds light on the genetic and evolutionary contexts behind the unique gait of the Puerto Rican Paso Fino, thus enhancing our comprehension of selected genes in island horses.

The genomics research aimed at identifying rare, localized, and endemic functional variants of biomedical relevance in specific human populations presents an example of bioinformatic analysis of big data in human populations. This research raises attention to the presence of gaps in the profiling of genomic variation within Europe. Despite significant progress in the field of genomics, specific European populations remain under-represented, thereby creating a disparity in our understanding of genetic diversity. The study addresses one such population - Genome Diversity in Ukraine project, which produced publicly available genome annotations and profiles. These resources provide invaluable insights into historical questions and lay the foundation for further genomic and medical research in these under-studied regions.

Lastly, the software suite PopGen Playground provides access to streamlined population genomics will make big data analyses more accessible to a broader scientific and medical community. This development will promote future research and encourage scalable, repeatable research, significantly improving the process of conducting population genomic studies in research groups with limited bioinformatics expertise. It discusses the accessibility of bioinformatics for remote training and work and provides an overview of methods to support scientists in need effectively.

5.2. Future directions

Moving forward, we aim to continually refine and enhance the PopGen Playground suite, expanding its capabilities to support a broader range of population genomics analyses. Our vision encompasses accommodating novel methodologies and algorithms, which are rapidly emerging in this dynamic field. Beyond mere software development, a core aspect of our future efforts lies in bolstering the worldwide dissemination of bioinformatics knowledge. Recognizing the power of digital platforms, we plan to leverage online resources to make this knowledge accessible globally. This would entail creating and promoting online educational materials, workshops, and interactive sessions, helping to bridge the gap between complex genomic data and its interpretation. We anticipate that this dual strategy of software improvement and knowledge dissemination will facilitate scientific discovery and democratize bioinformatics learning, further advancing the field of population genomics.

REFERENCES

Abzhanov, A., Extavour, C. G., Groover, A., Hodges, S. A., Hoekstra, H. E., Kramer, E. M., & Monteiro, A. (2008). Are we there yet? Tracking the development of new model systems. *Trends in Genetics*, *24*(7), 353–360. https://doi.org/10.1016/j.tig.2008.04.002

Alexander, D. H., Novembre, J., & Lange, K. (2009a). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/GR.094052.109

Alexander, D. H., Novembre, J., & Lange, K. (2009b). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., … Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature 2010 467:7319*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature 2012 491:7422*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Anaconda Software Distribution. (2020). In *Anaconda Documentation*. Anaconda Inc. https://docs.anaconda.com/

Árnason, Ú., Lammers, F., Kumar, V., Nilsson, M. A., & Janke, A. (2018). Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Science Advances*, *4*(4). https://doi.org/10.1126/SCIADV.AAP9873/SUPPL_FILE/AAP9873_SM.PDF

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. In *Nature*. https://doi.org/10.1038/nature15393

Axelsson, E., Ratnakumar, A., Arendt, M. L., Maqbool, K., Webster, M. T., Perloski, M., Liberg, O., Arnemo, J. M., Hedhammar, Å., & Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature 2013 495:7441*, *495*(7441), 360–364. https://doi.org/10.1038/nature11837

Ayala, N. M. (2016). *Assessing origin and genetic diversity through maternal lineages and "Gait Keeper" (DMRT3) frequency of Puerto Rican horses*. https://hdl.handle.net/20.500.11801/79

Barreiro, L. B., & Quintana-Murci, L. (2009). From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics 2010 11:1*, *11*(1), 17–30. https://doi.org/10.1038/nrg2698

Bartlett, A., Penders, B., & Lewis, J. (2017). Bioinformatics: Indispensable, yet hidden in plain sight? *BMC Bioinformatics*, *18*(1), 1–4. https://doi.org/10.1186/S12859-017-1730-9/METRICS

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484). https://doi.org/10.1126/SCIENCE.AAY5012/SUPPL_FILE/AAY5012-BERGSTROM-SM.PDF

Birney, E. (2021). The International Human Genome Project. *Human Molecular Genetics*, *30*(R2), R161. https://doi.org/10.1093/HMG/DDAB198

Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., Van Dijk, F., Francioli, L. C., Hottenga, J. J., Laros, J. F. J., Li, Q., Li, Y., Cao, H., Chen, R., Du, Y., Li, N., … Van Duijn, C. M. (2013). The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics 2014 22:2*, *22*(2), 221–227. https://doi.org/10.1038/ejhg.2013.118

Bosse, M., Megens, H. J., Madsen, O., Paudel, Y., Frantz, L. A. F., Schook, L. B., Crooijmans, R. P. M. A., & Groenen, M. A. M. (2012). Regions of homozygosity in the porcine genome: Consequence of demography and the recombination landscape. *PLoS Genet.*, *8*(11), e1003100. https://doi.org/10.1371/journal.pgen.1003100

Bourgeois, Y. X. C., & Warren, B. H. (2021). An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Molecular Ecology*, *30*(23), 6036–6071. https://doi.org/10.1111/MEC.15989

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics 2011 12:10*, *12*(10), 703–714. https://doi.org/10.1038/nrg3054

Caballero, A., Villanueva, B., & Druet, T. (2020). On the estimation of inbreeding depression using different measures of inbreeding from molecular markers. *Evol Appl*, *00*(2), 1–13. https://doi.org/10.1111/eva.13126

Calus, M. P. L., & Vandenplas, J. (2018). SNPrune: An efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. *Genetics Selection Evolution*, *50*(1), 1–11. https://doi.org/10.1186/S12711-018-0404-Z/FIGURES/3

Cantelli, G., Cochrane, G., Brooksbank, C., McDonagh, E., Flicek, P., McEntyre, J., Birney, E., & Apweiler, R. (2021). The European Bioinformatics Institute: empowering cooperation in response to a global health crisis. *Nucleic Acids Research*, *49*(D1), D29. https://doi.org/10.1093/NAR/GKAA1077

Cardona, A., Pagani, L., Antao, T., Lawson, D. J., Eichstaedt, C. A., Yngvadottir, B., Shwe, M. T. T., Wee, J., Romero, I. G., Raj, S., Metspalu, M., Villems, R., Willerslev, E., Tyler-Smith, C., Malyarchuk, B. A., Derenko, M. V., & Kivisild, T. (2014). Genome-Wide Analysis of Cold Adaptation in Indigenous Siberian Populations. *PLOS ONE*, *9*(5), e98076. https://doi.org/10.1371/JOURNAL.PONE.0098076

Cavalli-Sforza, L. L. (2005). The Human Genome Diversity Project: past, present and future. *Nature Reviews Genetics 2005 6:4*, *6*(4), 333–340. https://doi.org/10.1038/nrg1596

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1996). The History and Geography of Human Gene. (Abridged paperback edition). *The Journal of the Royal Anthropological Institute*, *2*(1), 413.

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, *19*(4), 220–234. https://doi.org/10.1038/NRG.2017.109

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research*, *20*(3), 393–402. https://doi.org/10.1101/GR.100545.109

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80. https://doi.org/10.4161/FLY.19695

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/GIGASCIENCE/GIAB008

Del Campo, J., Sieracki, M. E., Molestina, R., Keeling, P., Massana, R., & Ruiz-Trillo, I. (2014). The others: Our biased perspective of eukaryotic genomes. *Trends in Ecology and Evolution*, *29*(5), 252–259. https://doi.org/10.1016/j.tree.2014.03.006

Delaneau, O., Coulonges, C., & Zagury, J. F. (2008). Shape-IT: New rapid and accurate algorithm for haplotype inference. *BMC Bioinform.*, *9*, 1–14. https://doi.org/10.1186/1471-2105-9-540

Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019a). Accurate, scalable and integrative haplotype estimation. *Nat. Commun.*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-13225-y

Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019b). Accurate, scalable and integrative haplotype estimation. *Nature Communications 2019 10:1*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-13225-y

Ekblom, R., Brechlin, B., Persson, J., Smeds, L., Johansson, M., Magnusson, J., Flagstad, Ø., & Ellegren, H. (2018). Genome sequencing and conservation genomics in the Scandinavian wolverine population. *Conserv Biol*, *32*(6), 1301–1312. https://doi.org/10.1111/cobi.13157

Elhaik, E. (2022). Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports 2022 12:1*, *12*(1), 1–35. https://doi.org/10.1038/s41598-022-14395-4

Eydivandi, S., Roudbar, M. A., Karimi, M. O., & Sahana, G. (2021). Genomic scans for selective sweeps through haplotype homozygosity and allelic fixation in 14 indigenous sheep breeds from Middle East and South Asia. *Scientific Reports 2021 11:1*, *11*(1), 1–18. https://doi.org/10.1038/s41598-021-82625-2

Fairley, S., Lowy-Gallego, E., Perry, E., & Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, *48*(D1), D941–D947. https://doi.org/10.1093/NAR/GKZ836

Fan, H., Wu, Q., Wei, F., Yang, F., Ng, B. L., & Hu, Y. (2019). Chromosome-level genome assembly for giant panda provides novel insights into Carnivora chromosome evolution. *Genome Biology*, *20*(1), 1–12. https://doi.org/10.1186/S13059-019-1889-7/FIGURES/3

Fan, S., Hansen, M. E. B., Lo, Y., & Tishkoff, S. A. (2016). Review of Recent Human Adaptation. *Science*, *354*(6308), 54–59. https://doi.org/10.1126/science.aaf5098

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., … Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution*, *37*(3), 197–202. https://doi.org/10.1016/J.TREE.2021.11.008

Fuller, J. C., Khoueiry, P., Dinkel, H., Forslund, K., Stamatakis, A., Barry, J., Budd, A., Soldatos, T. G., Linssen, K., Rajput, A. M., & HUB Participants, H. (2013). Biggest challenges in bioinformatics. *EMBO Reports*, *14*(4), 302–304. https://doi.org/10.1038/embor.2013.34

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLOS Genetics*, *7*(11), e1002355. https://doi.org/10.1371/JOURNAL.PGEN.1002355

Gaind, N. (2022). How three Ukrainian scientists are surviving Russia's brutal war. *Nature*, *605*(7910), 414–416. https://doi.org/10.1038/D41586-022-01272-3

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Stephen Pittard, W., Mills, R. E., & Devine, S. E. (2017). The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Research*, *27*(11), 1916–1929. https://doi.org/10.1101/GR.218032.116

Gautier, M., Klassmann, A., & Vitalis, R. (2017). rehh 20: A reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.*, *17*(1), 549. https://doi.org/10.1111/1755-0998.12634

Gibbs, R. A. (2020). The Human Genome Project changed everything. *Nature Reviews Genetics 2020 21:10*, *21*(10), 575–576. https://doi.org/10.1038/s41576-020-0275-3

Goe, (. (n.d.). *Beyond One Million Genomes The Genome of Europe Realising a population genomic reference cohort of at least 500,000 citizens across Europe by 2022 Beyond One Million Genomes*.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., Malig, M., Raja, A., Fiddes, I., Hillier, L. W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R. K., Haussler, D., Chin, C.-S., & Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science (New York, N.Y.)*, *352*(6281), aae0344. https://doi.org/10.1126/science.aae0344

Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J. K., Muzzio, M., Rodriguez-Flores, J. L., Kenny, E. E., Gignoux, C. R., Maples, B. K., Guiblet, W., Dutil, J., Via, M., Sandoval, K., Bedoya, G., Oleksyk, T. K., Ruiz-Linares, A., Burchard, E. G., Martinez-Cruzado, J. C., & Bustamante, C. D. (2013). Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLOS Genetics*, *9*(12), e1004023. https://doi.org/10.1371/JOURNAL.PGEN.1004023

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G. T., Stacey, S. N., Frigge, M. L., Holm, H., Saemundsdottir, J., Helgadottir, H. T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., … Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics 2015 47:5*, *47*(5), 435–444. https://doi.org/10.1038/ng.3247

Guiblet, W. M., Zhao, K., O'Brien, S. J., Massey, S. E., Roca, A. L., & Oleksyk, T. K. (2015). SmileFinder: A resampling-based approach to evaluate signatures of selection from genome-wide sets of matching allele frequency data in two or more diploid populations. *GigaScience*, *4*(1). https://doi.org/10.1186/2047-217X-4-1

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., … Sandhu, M. S. (2014). The African Genome Variation Project shapes medical genetics in Africa. *Nature 2014 517:7534*, *517*(7534), 327–332. https://doi.org/10.1038/nature13997

Haberer, G., Young, S., Bharti, A. K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing, R. A., Rounsley, S., Birren, B., Nusbaum, C., Mayer, K. F. X., & Messing, J. (2005). Structure and Architecture of the Maize Genome. *Plant Physiology*, *139*(4), 1612. https://doi.org/10.1104/PP.105.068718

Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J. K., Coop, G., & Di Rienzo, A. (2011). Adaptations to Climate-Mediated Selective Pressures in Humans. *PLOS Genetics*, *7*(4), e1001375. https://doi.org/10.1371/JOURNAL.PGEN.1001375

Haussler, D., O'Brien, S. J., Ryder, O. A., Keith Barker, F., Clamp, M., Crawford, A. J., Hanner, R., Hanotte, O., Johnson, W. E., McGuire, J. A., Miller, W., Murphy, R. W., Murphy, W. J., Sheldon, F. H., Sinervo, B., Venkatesh, B., Wiley, E. O., Allendorf, F. W., Amato, G., … Turner, S. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.*, *100*(6), 659–674. https://doi.org/10.1093/jhered/esp086

Health*, T. G. A. for G. and. (2016). A federated ecosystem for sharing genomic, clinical data. *Science*, *352*(6291), 1278–1280. https://doi.org/10.1126/SCIENCE.AAF6162/SUPPL_FILE/AAF6162_PAGE-GLOBAL-ALLIANCE_SM.PDF

Hendricks, B. (2007). *International Encyclopedia of Horse Breeds*. University of Oklahoma Press.

Hood, L., & Rowen, L. (2013). The human genome project: Big science transforms biology and medicine. *Genome Medicine*, *5*(9), 1–8. https://doi.org/10.1186/GM483/METRICS

Horyn, L. (2022). *A Search for Displaced Leptons in the ATLAS Detector*. https://doi.org/10.1007/978-3-030-91672-5

Itan, Y., Powell, A., Beaumont, M. A., Burger, J., & Thomas, M. G. (2009). The origins of lactase persistence in Europe. *PLoS Computational Biology*, *5*(8). https://doi.org/10.1371/JOURNAL.PCBI.1000491

Jablonski, N. G., & Chaplin, G. (2010). Human skin pigmentation as an adaptation to UV radiation. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(SUPPL. 2), 8962–8968. https://doi.org/10.1073/PNAS.0914628107/ASSET/13CF7EDB-8EA8-439A-A4E6-1ABC29AAE6DA/ASSETS/GRAPHIC/PNAS.0914628107FIG02.JPEG

Jeon, S., Jeon, S., Bhak, Y., Bhak, Y., Bhak, Y., Choi, Y., Choi, Y., Jeon, Y., Jeon, Y., Kim, S., Kim, S., Jang, J., Jang, J., Jang, J., Blazyte, A., Kim, C., Kim, C., Kim, Y., Shim, J., … Bhak, J. (2020). Korean Genome Project: 1094 Korean personal genomes with clinical information. *Science Advances*, *6*(22). https://doi.org/10.1126/SCIADV.AAZ7835/SUPPL_FILE/AAZ7835_SM.PDF

Johnson, R. N., O'Meally, D., Chen, Z., Etherington, G. J., Ho, S. Y. W., Nash, W. J., Grueber, C. E., Cheng, Y., Whittington, C. M., Dennison, S., Peel, E., Haerty, W., O'Neill, R. J., Colgan, D., Russell, T. L., Alquezar-Planas, D. E., Attenbrow, V., Bragg, J. G., Brandies, P. A., … Belov, K. (2018). Adaptation and conservation insights from the koala genome. *Nat Genet*, *50*(8), 1102–1111. https://doi.org/10.1038/s41588-018-0153-5

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065). https://doi.org/10.1098/RSTA.2015.0202

Kalbfleisch, T. S., Rice, E. S., DePriest, M. S., Walenz, B. P., Hestand, M. S., Vermeesch, J. R., O′Connell, B. L., Fiddes, I. T., Vershinina, A. O., Saremi, N. F., Petersen, J. L., Finno, C. J., Bellone, R. R., McCue, M. E., Brooks, S. A., Bailey, E., Orlando, L., Green, R. E., Miller, D. C., … MacLeod, J. N. (2018). Improved reference genome for the domestic horse increases assembly contiguity and composition. *Communications Biology 2018 1:1*, *1*(1), 1–8. https://doi.org/10.1038/s42003-018-0199-z

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … Daly, M. J. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature 2020 581:7809*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kardos, M., Luikart, G., & Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity 2015 115:1*, *115*(1), 63–72. https://doi.org/10.1038/hdy.2015.17

Karishma Chhugani, Alina Frolova, Yuriy Salyha, Andrada Fiscutean, Oksana Zlenko, Sanita Reinsone, Walter W Wolfsberger, Oleksandra V Ivashchenko, Megi Maci, Dmytro Dziuba, Andrii Parkhomenko, Eric Bortz, Fyodor Kondrashov, Paweł P Łabaj, Veronika Romero, Jakub Hlávka, Taras K Oleksyk, & Serghei Mangul. (2022). Remote opportunities for scholars in Ukraine. *Science (New York, N.Y.)*, *378*(6626), 1285–1286. https://doi.org/10.1126/SCIENCE.ADG0797

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nature Reviews Genetics 2014 15:6*, *15*(6), 379–393. https://doi.org/10.1038/nrg3734

Karthikeyan, S., & Jose, D. V. (2021). Role of Data Science in the Field of Genomics and Basic Analysis of Raw Genomic Data Using Python. *Lecture Notes in Networks and Systems*, *290*, 176–181. https://doi.org/10.1007/978-981-16-4486-3_19/TABLES/2

Kavar, T., & Dovč, P. (2008). Domestication of the horse: Genetic relationships between domestic and wild horses. *Livestock Science*, *116*(1–3), 1–14. https://doi.org/10.1016/j.livsci.2008.03.002

Köster, J., Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., & Nahnsen, S. (2021). Sustainable data analysis with Snakemake. *F1000Research 2021 10:33*, *10*, 33. https://doi.org/10.12688/f1000research.29032.2

Kristjansson, T., Bjornsdottir, S., Sigurdsson, A., Andersson, L. S., Lindgren, G., Helyar, S. J., Klonowski, A. M., & Arnason, T. (2014). The effect of the "Gait keeper" mutation in the DMRT3 gene on gaiting ability in Icelandic horses. *Journal of Animal Breeding and Genetics*, *131*(6), 415–425. https://doi.org/10.1111/jbg.12112

Kwiatkowski, D. P. (2005). How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *American Journal of Human Genetics*, *77*(2), 171. https://doi.org/10.1086/432519

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/NAR/GKT1113

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), 1–19. https://doi.org/10.1186/GB-2014-15-6-R84/FIGURES/8

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(17), 4325–4333. https://doi.org/10.1073/PNAS.1720115115/SUPPL_FILE/PNAS.1720115115.SAPP.PDF

Li, J., Chen, H., Wang, Y., Chen, M. J. M., & Liang, H. (2021). Next-generation analytics for omics data. *Cancer Cell*, *39*(1), 3. https://doi.org/10.1016/J.CCELL.2020.09.002

Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American Journal of Human Genetics*, *100*(2), 267. https://doi.org/10.1016/J.AJHG.2017.01.004

Liu, X., Li, C., Mou, C., Dong, Y., & Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine*, *12*(1), 1–8. https://doi.org/10.1186/S13073-020-00803-9/FIGURES/4

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs HHS Public Access. *Hum Mutat*, *37*(3), 235–241. https://doi.org/10.1002/humu.22932

Liu, Y., Nyunoya, T., Leng, S., Belinsky, S. A., Tesfaigzi, Y., & Bruse, S. (2013). Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics*, *7*(1). https://doi.org/10.1186/1479-7364-7-1

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics 2003 4:12*, *4*(12), 981–994. https://doi.org/10.1038/nrg1226

Luís, C., Bastos-Silveira, C., Cothran, E. G., & Oom, M. D. M. (2006). Iberian origins of new world horse breeds. *J. Hered.*, *97*(2), 107–113. https://doi.org/10.1093/jhered/esj020

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature 2016 538:7624*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Mangul, S., Martin, L. S., Langmead, B., Sanchez-Galan, J. E., Toma, I., Hormozdiari, F., Pevzner, P., & Eskin, E. (2019). How bioinformatics and open data can boost basic science in countries and universities with limited resources. *Nature Biotechnology 2019 37:3*, *37*(3), 324–326. https://doi.org/10.1038/s41587-019-0053-y

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature 2009 461:7265*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Marras, G., Gaspa, G., Sorbolini, S., Dimauro, C., Ajmone-Marsan, P., Valentini, A., Williams, J. L., & MacCiotta, N. P. P. (2015). Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Animal Genetics*, *46*(2), 110–121. https://doi.org/10.1111/AGE.12259

Marx, V. (2013). The big challenges of big data. *Nature 2013 498:7453*, *498*(7453), 255–260. https://doi.org/10.1038/498255a

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*(11), 1747–1756. https://doi.org/10.1101/GR.239244.118

McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., Distl, O., Guérin, G., Hasegawa, T., Hill, E. W., Leeb, T., Lindgren, G., Penedo, M. C. T., Røed, K. H., Ryder, O. A., Swinburne, J. E., Tozaki, T., Valberg, S. J., Vaudin, M., … Mickelson, J. R. (2012). A high density SNP array for the domestic horse and extant perissodactyla: Utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet.*, *8*(1), e1002451. https://doi.org/10.1371/journal.pgen.1002451

McGuire, A. L., Gabriel, S., Tishkoff, S. A., Wonkam, A., Chakravarti, A., Furlong, E. E. M., Treutlein, B., Meissner, A., Chang, H. Y., López-Bigas, N., Segal, E., & Kim, J. S. (2020). The road ahead in genetics and genomics. *Nature Reviews Genetics 2020 21:10*, *21*(10), 581–596. https://doi.org/10.1038/s41576-020-0272-6

Miedema, F. (2022). Open Science: the Very Idea. *Open Science: The Very Idea*. https://doi.org/10.1007/978-94-024-2115-6

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., & Reich, M. (2021). Sustainable data analysis with Snakemake. *F1000Research 2021 10:33*, *10*, 33. https://doi.org/10.12688/f1000research.29032.1

Mousavi, N., Shleizer-Burko, S., Yanicky, R., & Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Research*, *47*(15), e90–e90. https://doi.org/10.1093/NAR/GKZ501

Need, A. C., & Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics : TIG*, *25*(11), 489–494. https://doi.org/10.1016/J.TIG.2009.09.012

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98. https://doi.org/10.1038/NATURE07331

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/SCIENCE.ABJ6987/SUPPL_FILE/SCIENCE.ABJ6987_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/NAR/GKV1189

Oleksyk, T. K., Smith, M. W., & O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. Lond. B*, *365*(1537), 185–205. https://doi.org/10.1098/rstb.2009.0219

Oleksyk, T. K., Wolfsberger, W. W., Schubelka, K., Mangul, S., & O'Brien, S. J. (2022). The Pioneer Advantage: Filling the blank spots on the map of genome diversity in Europe. *GigaScience*, *11*, 1–7. https://doi.org/10.1093/GIGASCIENCE/GIAC081

Oleksyk, T. K., Wolfsberger, W. W., Weber, A. M., Shchubelka, K., Oleksyk, O. T., Levchuk, O., Patrus, A., Lazar, N., Castro-Marquez, S. O., Hasynets, Y., Boldyzhar, P., Neymet, M., Urbanovych, A., Stakhovska, V., Malyar, K., Chervyakova, S., Podoroha, O., Kovalchuk, N., Rodriguez-Flores, J. L., … Smolanka, V. (2021). Genome diversity in Ukraine. *GigaScience*, *10*(1), 1–14. https://doi.org/10.1093/GIGASCIENCE/GIAA159

Oleksyk, T. K., Zhao, K., De La Vega, F. M., Gilbert, D. A., O'Brien, S. J., & Smith, M. W. (2008). Identifying Selected Regions from Heterozygosity and Divergence Using a Light-Coverage Genomic Dataset from Two Human Populations. *PLOS ONE*, *3*(3), e1712. https://doi.org/10.1371/JOURNAL.PONE.0001712

Oleksyk Taras, Wolfsberger Walter, Schubelka Khrystyna, Edmunds Scott, Tishkoff Sarah, Rodriguez-Flores Juan, Pellegrini Matteo, Pasaniuc Bogdan, Jane Mulder Nicola, D. Quinto-Cortés Consuelo, Huang Yu-Ning, Knyazev Sergey, & Mangul Serghei. (2023). *Genomic deserts: Global survey of underrepresented populations in genomics research (in preparation)*.

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genetics*, *2*(12), e190. https://doi.org/10.1371/JOURNAL.PGEN.0020190

Perry, G. H., Foll, M., Grenier, J. C., Patin, E., Nédélec, Y., Pacis, A., Barakatt, M., Gravel, S., Zhou, X., Nsobya, S. L., Excoffier, L., Quintana-Murci, L., Dominy, N. J., & Barreiro, L. B. (2014). Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(35), E3596-603. https://doi.org/10.1073/PNAS.1402875111

Petersen, J. L., Mickelson, J. R., Cothran, E. G., Andersson, L. S., Axelsson, J., Bailey, E., Bannasch, D., Binns, M. M., Borges, A. S., Brama, P., da Câmara Machado, A., Distl, O., Felicetti, M., Fox-Clipsham, L., Graves, K. T., Guérin, G., Haase, B., Hasegawa, T., Hemmann, K., … McCue, M. E. (2013). Genetic diversity in the modern horse illustrated from genome-wide SNP data. *PLoS ONE*, *8*(1), e54997. https://doi.org/10.1371/journal.pone.0054997

Pilon, B., Hinterneder, K., Hay, E. H. A., & Fragomeni, B. (2021). Inbreeding Calculated with Runs of Homozygosity Suggests Chromosome-Specific Inbreeding Depression Regions in Line 1 Hereford. *Animals : An Open Access Journal from MDPI*, *11*(11), 3105. https://doi.org/10.3390/ANI11113105

Polimanti, R., Iorio, A., Piacentini, S., Manfellotto, D., & Fuciarelli, M. (2014). Human pharmacogenomic variation of antihypertensive drugs: from population genetics to personalized medicine. *Http://Dx.Doi.Org/10.2217/Pgs.13.231*, *15*(2), 157–167. https://doi.org/10.2217/PGS.13.231

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006a). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006b). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.*, *69*, 1–14.

Promerová, M., Andersson, L. S., Juras, R., Penedo, M. C. T., Reissmann, M., Tozaki, T., Bellone, R., Dunner, S., Hořín, P., Imsland, F., Imsland, P., Mikko, S., Modrý, D., Roed, K. H., Schwochow, D., Vega-Pla, J. L., Mehrabani-Yeganeh, H., Yousefi-Mashouf, N., G. Cothran, E., … Andersson, L. (2014a). Worldwide frequency distribution of the "Gait keeper" mutation in the DMRT3 gene. *Animal Genetics*, *45*(2), 274–282. https://doi.org/10.1111/age.12120

Promerová, M., Andersson, L. S., Juras, R., Penedo, M. C. T., Reissmann, M., Tozaki, T., Bellone, R., Dunner, S., Hořín, P., Imsland, F., Imsland, P., Mikko, S., Modrý, D., Roed, K. H., Schwochow, D., Vega-Pla, J. L., Mehrabani-Yeganeh, H., Yousefi-Mashouf, N., G. Cothran, E., … Andersson, L. (2014b). Worldwide frequency distribution of the "Gait keeper" mutation in the DMRT3 gene. *Anim. Genet.*, *45*(2), 274–282. https://doi.org/10.1111/age.12120

Quaglio, G., Toia, P., Moser, E. I., Karapiperis, T., Amunts, K., Okabe, S., Poo, M. ming, Rah, J. C., Koninck, Y. De, Ngai, J., Richards, L., & Bjaalie, J. G. (2021). The International Brain Initiative: enabling collaborative science. *The Lancet Neurology*, *20*(12), 985–986. https://doi.org/10.1016/S1474-4422(21)00389-6

Rannala, B., & Yang, Z. (2008). Phylogenetic inference using whole genomes. *Annu. Rev. Genomics Hum. Genet.*, *9*, 217–231. https://doi.org/10.1146/annurev.genom.9.081307.164407

Relling, M. V., & Evans, W. E. (2015). Pharmacogenomics in the clinic. *Nature*, *526*(7573), 343. https://doi.org/10.1038/NATURE15817

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature 2021 592:7856*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

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

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science (New York, N.Y.)*, *298*(5602), 2381–2385. https://doi.org/10.1126/SCIENCE.1078311

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002a). Detecting recent positive selection in the human genome from haplotype structure. *Nature 2002 419:6909*, *419*(6909), 832–837. https://doi.org/10.1038/nature01140

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002b). Detecting recent positive selection in the human genome from haplotype structure. *Nature 2002 419:6909*, *419*(6909), 832–837. https://doi.org/10.1038/nature01140

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T., & Ge, R. L. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science (New York, N.Y.)*, *329*(5987), 72–75. https://doi.org/10.1126/SCIENCE.1189406

Smedley, D., Smith, K. R., Martin, A., Thomas, E. A., McDonagh, E. M., Cipriani, V., Ellingford, J. M., Arno, G., Tucci, A., Vandrovcova, J., Chan, G., Williams, H. J., Ratnaike, T., Wei, W., Stirrups, K., Ibanez, K., Moutsianas, L., Wielscher, M., Need, A., … Caulfield, M. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *New England Journal of Medicine*, *385*(20), 1868–1880. https://doi.org/10.1056/NEJMOA2035790/SUPPL_FILE/NEJMOA2035790_DISCLOSURES.PDF

Smetana, J., & Brož, P. (2022). National Genome Initiatives in Europe and the United Kingdom in the Era of Whole-Genome Sequencing: A Comprehensive Review. *Genes 2022, Vol. 13, Page 556*, *13*(3), 556. https://doi.org/10.3390/GENES13030556

Smith, M. W., & O'Brien, S. J. (2005). Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Reviews. Genetics*, *6*(8), 623–632. https://doi.org/10.1038/NRG1657

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015a). Big Data: Astronomical or Genomical? *PLOS Biology*, *13*(7), e1002195. https://doi.org/10.1371/JOURNAL.PBIO.1002195

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015b). Big Data: Astronomical or Genomical? *PLOS Biology*, *13*(7), e1002195. https://doi.org/10.1371/journal.pbio.1002195

Takser, L., Benachour, N., Husk, B., Cabana, H., & Gris, D. (2016). Cyanotoxins at low doses induce apoptosis and inflammatory effects in murine brain cells: Potential implications for neurodegenerative diseases. *Toxicology Reports*, *3*, 180. https://doi.org/10.1016/J.TOXREP.2015.12.008

Tang, K., Thornton, K. R., & Stoneking, M. (2007a). A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLOS Biology*, *5*(7), e171. https://doi.org/10.1371/JOURNAL.PBIO.0050171

Tang, K., Thornton, K. R., & Stoneking, M. (2007b). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.*, *5*(7), e171. https://doi.org/10.1371/journal.pbio.0050171

The "All of Us" Research Program. (2019). *New England Journal of Medicine*, *381*(7), 668–676. https://doi.org/10.1056/NEJMSR1809937/SUPPL_FILE/NEJMSR1809937_APPENDIX.PDF

Totikov, A., Tomarovsky, A., Prokopov, D., Yakupova, A., Bulyonkova, T., Derezanin, L., Rasskazov, D., Wolfsberger, W. W., Koepfli, K. P., Oleksyk, T. K., & Kliver, S. (2021). Chromosome-level genome assemblies expand capabilities of genomics for conservation biology. *Genes*, *12*(9). https://doi.org/10.3390/GENES12091336/S1

Tsartsianidou, V., Sánchez-Molano, E., Kapsona, V. V., Basdagianni, Z., Chatziplis, D., Arsenos, G., Triantafyllidis, A., & Banos, G. (2021). A comprehensive genome-wide scan detects genomic regions related to local adaptation and climate resilience in Mediterranean domestic sheep. *Genetics Selection Evolution*, *53*(1), 1–17. https://doi.org/10.1186/S12711-021-00682-7/TABLES/6

Valiente-Mullor, C., Beamud, B., Ansari, I., Frances-Cuesta, C., Garcia-Gonzalez, N., Mejia, L., Ruiz-Hueso, P., & Gonzalez-Candelas, F. (2021). One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Computational Biology*, *17*(1). https://doi.org/10.1371/JOURNAL.PCBI.1008678

Van Assche, R., Broeckx, V., Boonen, K., Maes, E., De Haes, W., Schoofs, L., & Temmerman, L. (2015). Integrating -Omics: Systems Biology as Explored Through C. elegans Research. *Journal of Molecular Biology*, *427*(21), 3441–3451. https://doi.org/10.1016/J.JMB.2015.03.015

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013a). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.BI1110S43

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013b). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *SUPL.43*. https://doi.org/10.1002/0471250953.BI1110S43

Via, A., de Las Rivas, J., Attwood, T. K., Landsman, D., Brazas, M. D., Leunissen, J. A. M., Tramontano, A., & Schneider, M. V. (2011). Ten Simple Rules for Developing a Short Bioinformatics Training Course. *PLOS Computational Biology*, *7*(10), e1002245. https://doi.org/10.1371/JOURNAL.PCBI.1002245

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLOS Biology*, *4*(3), e72. https://doi.org/10.1371/JOURNAL.PBIO.0040072

Wall, J. D., Stawiski, E. W., Ratan, A., Kim, H. L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E. S., Purbojati, R. W., Bhangale, T., Stepanov, V., Kharkov, V., Schröder, M. S., Ramprasad, V., Tom, J., Durinck, S., Bei, Q., Li, J., Guillory, J., … Peterson, A. S. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature 2019 576:7785*, *576*(7785), 106–111. https://doi.org/10.1038/s41586-019-1793-z

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq603

Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., & Schneider, M. V. (2014). Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. *PLOS Computational Biology*, *10*(3), e1003496. https://doi.org/10.1371/JOURNAL.PCBI.1003496

Wolfsberger, W. W., Ayala, N. M., Castro-Marquez, S. O., Irizarry-Negron, V. M., Potapchuk, A., Shchubelka, K., Potish, L., Majeske, A. J., Oliver, L. F., Lameiro, A. D., Martínez-Cruzado, J. C., Lindgren, G., & Oleksyk, T. K. (2022a). Genetic diversity and selection in Puerto Rican horses. *Scientific Reports 2022 12:1*, *12*(1), 1–14. https://doi.org/10.1038/s41598-021-04537-5

Wolfsberger, W. W., Ayala, N. M., Castro-Marquez, S. O., Irizarry-Negron, V. M., Potapchuk, A., Shchubelka, K., Potish, L., Majeske, A. J., Oliver, L. F., Lameiro, A. D., Martínez-Cruzado, J. C., Lindgren, G., & Oleksyk, T. K. (2022b). Genetic diversity and selection in Puerto Rican horses. *Scientific Reports*, *12*(1). https://doi.org/10.1038/S41598-021-04537-5

WRIGHT, S. (1949). THE GENETICAL STRUCTURE OF POPULATIONS. *Annals of Eugenics*, *15*(1), 323–354. https://doi.org/10.1111/J.1469-1809.1949.TB02451.X

Xue, H. (2022). Phylogenetics and its Application in Biodiversity Conservation. *Molecular Genetics and Genomics Tools in Biodiversity Conservation*, 1–16. https://doi.org/10.1007/978-981-16-6005-4_1/COVER

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics 2012 13:5*, *13*(5), 303–314. https://doi.org/10.1038/nrg3186

APPENDIX

Appendix 1. Data of Genome diversity of Puerto Rico

Table A1. Selected targets identified by genome-wide **XP-EHH** and **RSB** tests. Genes in bold are described after table A2. Negative values in the parentheses indicate the direction of selection in plots.

REGION	GENE(S)	CHROMOSOME	POSITION	XP EHH			RSB		
				ln PRPF*	ln PR NRB	ln PRPF vs. PR NRB	ln PRPF*	ln PR NRB	ln PRPF vs. PR NRB
PRS1	***CHRM5***	1	154,082,628	(5.12)		5.75	(4.51)		5.50
PRS2	*RNASE10, RNASE12*	1	156,642,282	(4.03)		4.59			-
PRS3	***CYP2E1**, MIR8947*	1	161,130,579	(4.52)		5.21	(4.53)		4.17
PRS4	***MYH7,** MIR208A, MIR208B*	1	161,184,230	(4.54)		4.63	(5.11)		4.48
PRS5	*MIR9047*	1	161,817,069			4.59			-
PRS6	***STRN3, HEATR5A***	1	167,046,919	(4.45)		4.39	(4.54)		4.28
PRS7	*MIR8990*	7	40,423,783			4.58			4.35
PRS8	***HPGD**, MIR9108*	11	25,516,779			-			4.23
PRS9	***SRSF1**, MIR9152*	11	30,521,866			4.08			-
PRS10	*PAM*	14	66,106,336			4.37			-
PRS11	*COMP, SERPINB1, MIR130B, MIR301B, MIR8923-1*	15	61,299,263			-	(4.11)		-
PRS12	*BLVRA, PDE4D, SYNE1*	21	12,679,828			-			4.16
PRS13	***PRND, PRNP***	22	16,867,450	(4.19)		4.39	(4.74)		4.84
PRS14	***DLD, SNRPB2,** MIR1291B*	22	31,942,581			-	(4.37)		-
PRS15	*MIRLET7D, MIRLET7F*	23	23,272,843		(4.79)	-		(4.64)	-
PRS16	*TBCCD1, MIR10A, MIR8952*	23	24,142,424			-		(4.38)	-

Table A2. A detailed description of the protein-coding genes located in the putative selection regions in the Puerto Rican Paso Fino genome

Chromosome 1

The muscarinic cholinergic receptor **CHRM5** belongs to a larger family of G protein-coupled receptors. The functional diversity of these receptors is defined by the binding of acetylcholine and includes cellular responses such as adenylate cyclase inhibition, phosphoinositide degeneration, and potassium channel mediation. Muscarinic receptors influence many effects of acetylcholine in the central and peripheral nervous system. The clinical implications of this receptor are unknown; however, stimulation of this receptor is

known to increase cyclic AMP levels. [provided by RefSeq, Jul 2008]. **Diseases associated with CHRM5 include Schizophrenia.**

Cytochrome P450 Family 2 Subfamily E Member **CYP2E1** encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. This protein localizes to the endoplasmic reticulum and is induced by ethanol, the diabetic state, and starvation. The enzyme metabolizes both endogenous substrates, such as ethanol, acetone, and acetal, as well as exogenous substrates including benzene, carbon tetrachloride, ethylene glycol, and nitrosamines which are premutagens found in cigarette smoke. Due to its many substrates, this enzyme may be involved in such varied processes as gluconeogenesis, hepatic cirrhosis, diabetes, and cancer. [provided by RefSeq, Jul 2008]

Muscle myosin **MYH7** is a hexameric protein containing 2 heavy chain subunits, 2 alkali light chain subunits, and 2 regulatory light chain subunits. This gene encodes the beta (or slow) heavy chain subunit of cardiac myosin. It is expressed predominantly in normal human ventricle. It is also expressed in skeletal muscle tissues rich in slow-twitch type I muscle fibers. Changes in the relative abundance of this protein and the alpha (or fast) heavy subunit of cardiac myosin correlate with the contractile velocity of cardiac muscle. Its expression is also altered during thyroid hormone depletion and hemodynamic overloading. Mutations in this gene are associated with familial hypertrophic cardiomyopathy, myosin storage myopathy, dilated cardiomyopathy, and Laing early-onset distal myopathy. [provided by RefSeq, Jul 2008]

HEATR5A (HEAT Repeat Containing 5A) is a Protein Coding gene. Diseases associated with HEATR5A include Bardet-Biedl Syndrome 4 and Ceroid Lipofuscinosis, Neuronal, 6. Gene Ontology (GO) annotations related to this gene include binding. An important paralog of this gene is HEATR5B.

STRN3 (Striatin 3) is a Protein Coding gene. Diseases associated with STRN3 include Syndromic Intellectual Disability. Among its related pathways are Neurophysiological process Glutamate regulation of Dopamine D1A receptor signaling. Gene Ontology (GO) annotations related to this gene include DNA-binding transcription factor activity and calmodulin binding. An important paralog of this gene is STRN.

Chromosome 11

SRSF1 gene encodes a member of the arginine/serine-rich splicing factor protein family. The encoded protein can either activate or repress splicing, depending on its phosphorylation state and its interaction partners. Multiple transcript variants have been found for this gene. There is a pseudogene of this gene on chromosome 13. [provided by RefSeq, Jun 2014]. **Diseases associated with SRSF1 include Retinitis Pigmentosa 4 and Homocystinuria.**

Chromosome 14

PAM gene encodes a multifunctional protein. The encoded preproprotein is proteolytically processed to generate the mature enzyme. This enzyme includes two domains with distinct catalytic activities, a peptidylglycine alpha-hydroxylating monooxygenase (PHM) domain and a peptidyl-alpha-hydroxyglycine alpha-amidating lyase (PAL) domain. These catalytic domains work sequentially to catalyze the conversion of neuroendocrine peptides to active alpha-amidated products. Alternative splicing results in multiple transcript variants, at least

one of which encodes an isoform that is proteolytically processed. [provided by RefSeq, Jan 2016]. Diseases associated with PAM include Menkes Disease and Spinal Muscular Atrophy, Distal, X-Linked 3.

Chromosome 22

DLD gene encodes a member of the class-I pyridine nucleotide-disulfide oxidoreductase family. The encoded protein has been identified as a moonlighting protein based on its ability to perform mechanistically distinct functions. In homodimeric form, the encoded protein functions as a dehydrogenase and is found in several **multi-enzyme complexes that regulate energy metabolism. However, as a** monomer, this protein can function as a protease. Mutations in this gene have been identified in patients with E3-deficient maple syrup urine disease and lipoamide dehydrogenase deficiency. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Jan 2014]. Mutant phenotypes in humans include **abnormalities in head and neck.**

SRNPB2 gene encodes a protein associates with stem loop IV of U2 small nuclear ribonucleoprotein (U2 snRNP) in the presence of snRNP-A'. The encoded protein may play a role in pre-mRNA splicing. Autoantibodies from **patients with systemic lupus erythematosus** frequently recognize epitopes on the encoded protein. Two transcript variants encoding the same protein have been found for this gene. [provided by RefSeq, Jul 2008]

PRNP gene codes for a membrane glycosylphosphatidylinositol-anchored glycoprotein that tends to aggregate into rod-like structures. The encoded protein contains a highly unstable region of five tandem octapeptide repeats. In humans this gene is found on chromosome 20, approximately 20 kbp upstream of a gene which encodes a biochemically and structurally similar protein to the one encoded by this gene.

Mutations in the repeat region as well as elsewhere in this gene have **been associated with Creutzfeldt-Jakob disease, fatal familial insomnia, Gerstmann-Straussler disease, Huntington disease-like 1, and kuru.** An overlapping open reading frame has been found for this gene that encodes a smaller, structurally unrelated protein, AltPrp. Alternative splicing results in multiple transcript variants.

May play a role in neuronal development and synaptic plasticity. May be required for neuronal myelin sheath maintenance. May promote myelin homeostasis through acting as an agonist for ADGRG6 receptor. May play a role in iron uptake and iron homeostasis

Appendix 2. Genome Diversity of Ukraine IRB approval

OAKLAND UNIVERSITY

Institutional Review Board

May 19, 2021

Protocol #: IRB-FY2021-346

Research Team:
Taras Oleksyk

Per federal regulations, the Oakland University IRB has determined that there is No Engagement in Research for Oakland University in the following study, "Genome diversity in Ukraine"

The IRB decision is based on the following:

The Oakland University PI, Dr. Oleksyk, will only be receiving coded biospecimens. A letter from Uzhhorod National University, assuring no codes will be released to any research collaborators has been provided in this submission.

Please note that since this project involves the use of biospecimens at Oakland University, the Biosafety Officer is copied on this IRB “Not Engaged” determination letter.

Please retain a copy of this correspondence for your records.

If you have any questions, please contact the IRB staff.