12-2021

A method for identifying ancient introgression between caballine and non-caballine equids using whole genome high throughput data.

Kalpani de Silva
University of Louisville

A METHOD FOR IDENTIFYING ANCIENT INTROGRESSION BETWEEN CABALLINE AND NON-CABALLINE EQUIDS USING WHOLE GENOME HIGH-THROUGHPUT DATA

By

Kalpani de Silva

B.Sc. (Bioinformatics), University of Colombo, Sri Lanka, 2014

A Dissertation

Submitted to the Faculty of the

Graduate School of University of Louisville

in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in
Interdisciplinary Studies: Specialization in Bioinformatics

Graduate School
University of Louisville
Louisville, Kentucky

December 2021

A METHOD FOR IDENTIFYING ANCIENT INTROGRESSION BETWEEN CABALLINE AND NON-CABALLINE EQUIDS USING WHOLE GENOME HIGH-THROUGHPUT DATA

By

Kalpani de Silva
B.Sc. (Bioinformatics), University of Colombo, Sri Lanka, 2014

A Dissertation Approved on

November 19, 2021

By the following Dissertation Committee

_____
Theodore S. Kalbfleisch , Ph.D. Advisor

_____
Eric C. Rouchka, D.Sc. Co-Advisor

_____
Juw Won Park, Ph.D.

_____
Corey Watson, Ph.D.

_____
Ryan Gill, Ph.D.

DEDICATION

This dissertation is dedicated to my loving parents

Jayanthi Gunasekara and Ariyasiri de Silva

and to my dearest husband

Akila Samaraweera

ACKNOWLEDGEMENTS

I would like to thank several people who helped in many ways to make this dissertation a reality.

First, I am grateful for my principal advisor Dr. Theodore S. Kalbfleisch for his excellent support and guidance during my doctoral journey. I am thankful for his patience, friendly manner, and the opportunity to learn so much which made my graduate life a wholesome experience. Next, I am grateful for my co-advisor Dr. Eric C. Rouchka for his valuable advice and guidance throughout the Bioinformatics Ph.D. program. I am thankful for giving me the opportunity to join the Bioinformatics lab and always helping to overcome obstacles and move towards the successful completion of my Ph.D. journey.

I would also like to thank Dr. Juw Won Park, Dr. Corey Watson, and Dr. Ryan Gill for being on my dissertation committee and providing valuable feedback and insights on the dissertation work. I am grateful for Dr. Ernest Bailey from Gluck Equine Research Center at University of Kentucky for the continuous guidance and kind support over the years. I am also thankful for Dr. David Samuelson and everyone else in the Department of Biochemistry and Molecular Genetics for been very helpful and welcoming during my time with them. I also would like to thank Dr. Adel Elmaghraby and everyone else in the Department of Computer Science and Engineering for all their support.

My experience at UofL would not be the same without former and current students in the Bioinformatics Ph.D. program and in the Bioinformatics lab. I would like to thank Sen Yao, Ernur Saka, Mohammed Sayed, Sudhir Srivastava, Aanchal Malhotra, Aryan

Neupane, Jonah Daneshmand, Jae Hwang, Emily Duderstadt, Sivarchana Mareedu and many others. I am also sincerely thankful for Dr. Julia Chariker for her friendly caring nature. My deepest gratitude also goes to my Sri Lankan friends I met in Louisville who helped me in many ways through the years.

Most of all, I am grateful to my family, my parents, my two sisters and in-laws for their unconditional love and support for everything I do in my life. I am especially thankful for my husband, Akila who was by my side through thick and thin from the day one of my PhD journey and life in USA. I am grateful for his sacrifices, strength, love and understanding.

ABSTRACT

A METHOD FOR IDENTIFYING ANCIENT INTROGRESSION BETWEEN CABALLINE AND NON-CABALLINE EQUIDS USING WHOLE GENOME HIGH-THROUGHPUT DATA

Kalpani de Silva

November 19, 2021

Introgression is one of the main mechanisms that transfer adapted alleles between species. The advantageous variants will get positively selected and retained in the recipient population while rest of the variants undergo negative selection. When analyzing horse genome, two alleles were found in *CXCL16* gene, one associated with susceptibility and one with resistance to developing persistent shedding of the Equine Arteritis Virus. The two alleles differ by 4 non-synonymous variants in exon 1 of the gene. Comparison with 3 non-caballine equids (zebras, asses and hemiones) revealed that one haplotype was almost identical to the haplotype found in non-caballines while the other had differences characteristic of 4.5 million years since a common ancestor. Based on this observation, we project that an ancient introgression event occurred between caballine and non-caballine equids. If so, we should be able to find more instances of introgression between these species. We developed a method to identify putatively introgressed segments in the horse genome. It is estimated that non-caballine equids such as zebras and asses diverged from horses between 4 and 4.5 MYA. Genomic analysis of these animals vs. equine reference

genome reveals the divergence at both the nucleotide and chromosomal level. Whole genome data for the non-caballine equids when mapped to the caballine (*Equus caballus*) reference genome show a greater frequency of single nucleotide differences than horses have relative to the same reference. We have created a Likelihood Estimate framework that uses this difference in single nucleotide frequencies to predict whether a haplotype evolved along the caballine or non-caballine lineage. Our results demonstrated that these haplotypes are between 0.5 and 2kb in length and are detectable at a rate of several hundred loci per horse. About 1.1% of the equine genome was introgressed and 64% of the identified putative regions were associated with either structural elements, regulatory regions, or both. These regions were responsible for gene products involved in regulation of response to stimuli, signal transduction, integral components of cell membrane and important metabolism pathways such as purine metabolism and thiamine metabolism. Furthermore, these haplotypes occur at high frequency in the horse population suggesting that they are positively selected by evolution.

TABLE OF CONTENTS

DEDICATION **iii**

ACKNOWLEGDEMENTS **iv**

ABSTRACT **vi**

LIST OF TABLES **xii**

LIST OF FIGURES **xiv**

1. INTRODUCTION **1**

1.1 Overview……1

1.2 Motivation……..….1

1.3 Objectives…….. 6

1.4 Dissertation outline…….6

2. BACKGROUND **9**

2.1 Genetic variation……....9

2.1.1 Signal Nucleotide Polymorphisms……..……..…12

2.1.2 Equine reference genome……..15

2.1.3 Genomic tools and resources…….16

2.1.4 Next Generation Sequencing Technology…….18

2.2 What is introgression……....19

2.2.1 Examples of adaptive introgression……...23

2.2.2 Ghost introgression……26

2.3 Horse domestication…….....27

2.3.1 Gene flow in equids……...28

2.4 Introgression detection methods……...32

2.4.1 *ABBA-BABA* or Patterson's *D*-statistics……32

2.4.2 Hidden Markov Model (HMM)……34

3. MAPPING AND VARIANT CALLING 35

3.1 Overview ……35

3.2 Sample information ……36

3.3 Data acquisition……37

3.4 Quality analysis of the data……...39

3.4.1 Quality control……...40

3.5 Mapping data……41

3.6 Variant calling……..44

3.7 Results and Discussion……47

4. INTROGRESSION DETECTION ALGORITHM 52

4.1 Overview…….…….52

4.2 Variant filteration…….53

4.3 Ancestral alleles vs horse-specific alleles…….……...56

4.4 Pair distribution function…….59

4.5 Maximum Likelihood Estimation……62

4.6 Haplotype phasing……....64

4.7 Phylogenetic inference…….66

4.8 Evaluating predicted regions……....67

4.9 Results and Discussion……69

4.9.1 Ancestral alleles vs species-specific alleles……...69

4.9.2 MLE based introgression detection…….……..70

4.9.3. Comparison to sequenced archaic genomes…….…....72

4.9.4 Evaluating predicted regions ……..…73

5. FUNCTIONAL ANNOTATION OF PUTATIVE INTROGRESSED REGIONS 75

5.1 Overview……..75

5.2 Annotation work done on domesticated animal genomes……....75

5.3 Understanding the putative regions based on structural annotation of equine genome……...78

5.4 Understanding the putative regions based on histone modificationsites…....79

5.5 Understanding the putative regions based on functional annotation……...…83

5.5.1 Start GO workflowusing NCBI Blast……....85

5.5.1.1 Continue GO workflow with OmicsBox……87

5.5.1.2 InterProScan……...88

5.5.1.3 Gene Ontology mapping……90

5.5.1.4 Blast2GO annotation…….……90

5.6 Results and Discussion……91

5.6.1 Comparison against histone modification sites……..91

5.6.2 Comparison against structural annotation of equine genome…..……92

5.6.3 GO annotation results……93

5.6.3.1 NCBI blastx search results……..……93

5.6.3.2 OmicsBox InterProScan results……..94

5.6.3.3 GO mapping results……..103

5.6.3.4 GO annotation results…….……..104

5.6.3.4.1 Biological Process aspect of GO……...104

5.6.3.4.2 Cellular Component aspect of GO……...…..107

5.6.3.4.3 Molecular Function aspect of GO ……..…108

5.6.3.4.4 Enzyme code mapping and KEGG pathway analysis……109

6. DISCUSSION **111**

7. CONCLUSIONS AND FUTURE **115**

REFERENCES **117**

APPENDIX A **126**

APPENDIX B **128**

APPENDIX C **131**

APPENDIX D **133**

CURRICULUM VITAE **149**

LIST OF TABLES

Table 1: Non-synonymous nucleotide substitutions responsible for the *CXCL16* allelic variants. The coordinates are relative to EquCab3…….....3

Table 2: Summary of the sample information. The animals were sequenced using paired-end, short read, Illumina HiSeq technology……36

Table 3: The mapping fractions for each animal after mapping against equine reference genome……...42

Table 4: The class of filtered variants according to its component alleles and its mapping to the equine reference genome. …….....54

Table 5: Variation in number of regions identified with different MLE cutoff values ……63

Table 6: Introgressed regions identified per horse genome ……70

Table 7: Annotation of the putative introgressed regions based on the structural annotation of the equine genome ……..……...78

Table 8: Regulatory regions at histone modification sites overlap with putative introgressed regions at respective tissues ……...82

Table 9: List of protein domains identified for the putative introgressed regions by IntroProScan ..……96

Table 10: List of protein families identified for the putative introgressed regions by IntroProScan ..……98

Table 11: List of specific Biological Process (BP) GO-terms obtained from annotation ……..105

Table 12: List of specific Cellular Component (cc) GO-terms obtained from annotation ……..107

Table 13: List of specific Molecular Function (MF) GO-terms obtained from annotation …….....108

Table 14: Enzyme codes mapped into gene products of putative introgressed sequences ……..…109

LIST OF FIGURES

Figure 1: The 240 bp region in *CXCL16* locus which gives rise to the suspicion of introgression between caballines and non-caballines……...3

Figure 2: Phylogenetic relationships among equids …….....5

Figure 3: Movement of genes from one species to another by recurrent backcrossing of hybrids to their parental species.……..…..……..20

Figure 4: The terms species and hybrids in a case of introgressive hybridization. Solid black areas represent original species and first-generation hybrids. Dotted areas represent later hybrid generations and back-crosses [47]. …….22

Figure 5: Morphological evolution from *Eiohippus* to *Equus* [71]. …..……..29

Figure 6: Adaptive radiation of *Equus*. …….30

Figure 7: Events of gene flow between populations. The arrows represent the gene flow events. The Divergence time and population split times are indicated by darker and lighter ends of the colored rectangles [74]……...……..32

Figure 8: The gene trees that produce *ABBA* and *BABA* patterns ……33

Figure 9: A summary of the main steps of the introgression detection pipeline with the tools used and the output file formats. …….…….…….…….35

Figure 10: The class of a variant according to its component alleles and its mapping to the equine reference genome ……..…….…….……....49

Figure 11: The diagram showing the locations of the variants in a transcript corresponding to their severity in consequences. ……..……..……..……....49

Figure 12: The percentage of variants with consequences in the entire genome…….....50

Figure 13: Main steps of the introgression detection workflow...……...52

Figure 14: The percentage of filtered variants with consequences in the entire genome …54

Figure 15: The percentage of filtered variants with consequences only on coding sequences ……..……..55

Figure 16: Inheritance of the ancestral allele among equids through evolution and by introgressio……..……...56

Figure 17: Variant distribution along caballine (TB03, TB10, Saddlebred, FAANG TB) and non-caballine (Donkey, Grevyi, Onager) genomes. The putative introgressed region is circled in the figure. The variants are color coded based on the nucleotide. ……...……58

Figure 18: SNP distribution across the genome. The blue colored circles represent SNPs. The distances between SNPs are marked in red arrows. …….....60

Figure 19: Nearest neighbor distribution function of *Horse-Specific* alleles and *Ancestral* alleles…….61

Figure 20: Maximum likelihood estimation of putative introgressed regions at maximum cutoff value of 100……...…….....62

Figure 21: Haplotype phasing. Three bi-allelic variants can produce 2^{3-1} = 4 possible haplotype patterns ……..65

Figure 22: Evaluating the phylogenetic trees based on the traversal distance from a caballine node to where caballines and non-caballines cluster. (a) False positive tree (b) Successful tree. ……..……68

Figure 23: Allele distribution identified in horse genome. ……69

Figure 24: The putative introgressed region (MLE>100) distribution across the horse genome……...71

Figure 25: Evaluation of identified regions. The fraction of trees with no caballine haplotype clustered with a non-caballine clade (marked with red arrow). The trees with caballine haplotypes clustered with a non-caballine clade (marked with green arrows). ……..…….73

Figure 26: Histone is a protein that provides structural support to a chromosome. Some modifications in histones are associated with regulation of gene expression. ……79

Figure 27: OmicsBox Blast2GO module example workflow for GO annotation of a set of sequences ……...87

Figure 28: Putative introgressed regions overlap with tissue specific histone modification data ……91

Figure 29: Out of the total 8,951 of putative introgressed regions, 64% had either a structural annotation, a regulatory region annotation or both attached to it. ……...92

Figure 30: The e-value distribution of the blastx results vs the number of best hits obtained at each e-value. ……..93

Figure 31: Number of blastx hits identified from public databases for the query sequences ……94

Figure 32: Protein domains identified in the putative introgressed regions by InterProScan ……95

Figure 33: Protein families identified in the putative introgressed regions by InterProScan ……98

Figure 34: Distribution of GO terms vs the mapped putative introgressed regions……..103

Figure 35: Annotation score distribution across GO mapped putative introgressed sequences …….104

CHAPTER 1
INTRODUCTION

1.1 Overview

Introgression can act as a significant contributor to evolution through transfer of adapted alleles between species by hybridization and repeated backcrossing. Although much of the introgressed DNA appears to undergo negative selection it can occasionally introduce variants that are adaptive and therefore retained and spread in the recipient population. Adaptive introgression, as it is called, is well known for plant species but less commonly known for animal species. The issue has garnered interest because of recent discoveries of introgressed genes from other hominid species into *Homo sapiens* during the last 100,000 years. Yet there are no extensive studies conducted on horse whole genome introgression. This research aims to bridge this gap by investigating models to identify introgression in this domesticated animal. We can use the knowledge and advance computational, statistical methods generated from this work to study other domesticated economically important animals.

1.2 Motivation

Researchers at the University of Kentucky identified an allele in the horse that confers susceptibility for persistent shedder status of Equine Arteritis Virus (EAV) following the infection of virus [1]. The susceptible haplotype consists of four specific

nucleotide changes in the *CXCL16* gene in equine chromosome 11. The stallions possessing the susceptibility haplotypes are significantly more likely to remain permanent carriers of the virus in their reproductive system than the horses that possess the resistant haplotype. The stallions homozygous for the resistant *CXCL16* haplotype initially show shedding of virus in their semen following infection. But in most cases, they could clear the virus from the reproductive tract within months following infection. Even having one copy of the susceptible haplotype (heterozygotes) is a greater risk for becoming permanent shedders of EAV.

The four non-synonymous nucleotide substitutions responsible for the *CXCL16* allelic variants among horse were identified through a Genome Wide Association study (GWAS) (Table 1). They were found within exon 1 and resulted in four amino acid changes producing two proteins (Eq*CXCL16*S and Eq*CXCL16*R). This encoded a protein product that works as a cellular receptor for EAV infection and exerts a dominant mode of inheritance. The resistant protein Eq*CXCL16*R does not function as an EAV cellular receptor. Sarkar et al. has genotyped *CXCL16* sequences from 240 horses across four horse breeds (Thoroughbred [n=67], Standardbred [n=60], Quarter Horse [n=53], Saddlebred Horse [n=60]). This provided an estimate of the frequencies of the two alleles across horse breeds. They identified 85 horses that were resistant to EAV and 155 were susceptible. All resistant horses were homozygous for the A, G, T, G alleles. The susceptible horses were either homozygous for the T, C, A, A alleles or were heterozygous A/T, G/C, T/A, G/A. This provides strong evidence the *CXCL16S* allele (T, C, A, A) is responsible for EAV susceptibility in the horse population.

Table 1: Non-synonymous nucleotide substitutions responsible for the *CXCL16* allelic variants. The coordinates are relative to equcab3.

Coordinates	**Non-synonymous nucleotide substitutions**	
	Resistant haplotype	Susceptible haplotype
Chr11:50,087,128	A	T
Chr11:50,087,154	G	C
Chr11:50,087,157	T	A/G
Chr11:50,087,163	G	A

We identified eight polymorphisms in a 240 bp region, which gives rise to the suspicion that this is an introgressed region (Figure 1). We inspected a 2kb region surrounding this site until where the haplotype ended. A region of 947 bases appeared strongly conserved for the susceptibility allele and the haplotype identified in non-caballine. These are two haplotypes likely diverged from one another over several million years, as we would not expect to see that much fixed difference over a short period of time.

Figure 1: The 240 bp region in *CXCL16* locus which gives rise to the suspicion of introgression between caballines and non-caballines

There were no intermediate haplotypes that would suggest a path that can get from one haplotype to the other. So, we analyzed the other non-caballine equids (donkey, grevyi and onager) and identified that the haplotype found in their genomes were nearly identical to the EAV susceptibility haplotype that was found in horse. This provides evidence that this haplotype was introduced into horse through an introgression event. In agriculture, different species are crossed intentionally. The female F1 progeny of these crosses are usually fertile and when these F1 get crossed with other animals it will spread the genetic material from the first species to the second species.

But the horse cannot produce fertile offspring with closely related species due to reproductive isolation. Usually, hybrid animals are rare in nature. Mayr proposed the gene exchange during hybridization breaks up the internally balanced chromosome sections producing sterile offspring resulting in eliminating these individuals [2]. There are several other examples in nature for sterile hybrids. The mating of a male lion and a female tiger result in a liger [3]. This is called post-zygotic isolation. The cross between a female horse and a male donkey is a mule [4]. A mule is viable, but it is sterile. Unlike tigers and lions, horse and donkey do not even have the same number of chromosomes. Horse has 32 chromosomes while zebra, donkeys and other equids have chromosomes ranging from 16-31 (Figure 2). This is called chromosomal plasticity. We do not know when these chromosome differences appeared. But we imagine we are talking about an event that occurred millions of years ago. Perhaps a now extinct Equid had a similar number of chromosomes in their genomes and were able to produce fertile offspring with the horse.

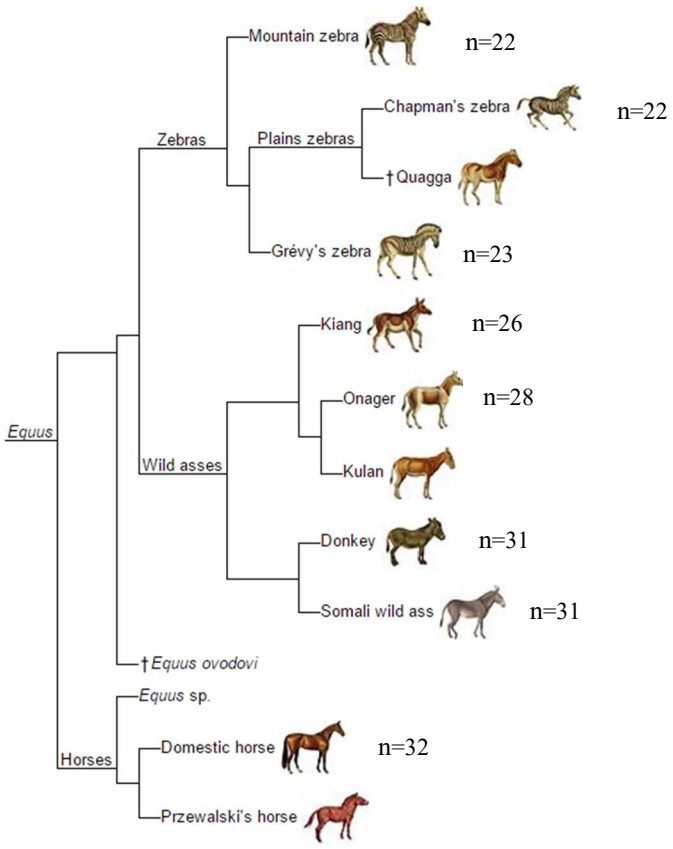


Figure 2: Phylogenetic relationships among equids

As one of the earliest domesticated species horses have been historically, culturally, and economically important animal to humans. Over the years they have been used extensively for military purposes, transportation, agriculture, and entertainment. Horses have been selectively bred for performance traits like speed, strength, endurance, gait and for appearance traits like size, color and for their temperament. Unlike model organisms such as yeast, fruit fly, and mouse which are widely studied because they are easy to maintain and breed in a laboratory setting, the non-model organism horse has not been extensively studied, until now. This dissertation describes the first comprehensive study

conducted on investigating potential introgression between caballine and non-caballine equid genomes. There are methods proposed to detect introgression events in modern humans that occurred about 40,000 – 50,000 years ago, which is a quite recent event compared to any estimate of equine introgression we address in this project. Hence there is a need to develop methods to detect ancient introgression involving animal species more divergent than previously reported in the literature. These reasons provide the rationale for the development of methodologies that enable detection of ancient introgression.

1.3 Objectives

The scope of this dissertation is focused on fulfilling the following objectives.

1. Mapping and variant calling of a cohort of caballine and non-caballine equids
2. Developing a workflow to detect ancient introgression
3. Annotating and identifying the biological importance of identified putative introgressed regions.

1.4 Dissertation outline

This dissertation is organized into seven chapters. Chapter 2 provides a brief overview of the concept of introgression and a summary of the current literature on introgression detection in different species. It begins with an introduction to genetic variation followed by the importance of studying variation in genomes. Next it explains the available tools and resources for horse genomic studies including the horse reference genome and equine Single Nucleotide Polymorphism (SNP) chip array. It continues with a brief explanation of exponential growth of high throughput sequencing technologies and

their utilization in whole genome sequence analysis. Furthermore, the chapter talks about the examples of adaptive introgression, introduces the concept of ghost introgression, horse domestication and overall gene flow in equids. The chapter ends with a brief description of available introgression detection methods. Chapter 3 explains the mapping and variant calling process. It provides information about the samples used in the analysis. Next the quality analysis and quality control of data followed by steps involved in mapping the data and variant calling of the mapped data. Finally, in this chapter we discuss the consequences of the identified variants. Chapter 4 is where we describe the introgression detection algorithm. The chapter begins with a detailed description of the input variants for the algorithm and next defines the terms *Ancestral* alleles and *Horse-Specific* alleles. Then it explains the pair distribution function and Maximum Likelihood Estimation (MLE) leading to detection of putative haplotypes. Next, it describes the haplotype phasing process and phylogenetic inferences which helps to identify putative introgressed region and concludes with an evaluation of identified regions. Chapter 5 is dedicated to the functional annotation of the putative introgressed regions. In the beginning, it points out how most of the non-model reference genomes are still not fully annotated and mentions some of the recent annotation work done on domesticated animal genomes. It also describes the ongoing attempt of annotating the horse genome and compares the identified regions against structural annotation of the equine reference genome. Next, the data was compared against histone modification data to identify potential overlaps with regulatory regions. Finally, it describes how the data was run through the Gene Ontology (GO) annotation pipeline at the end of the chapter. Chapter 6 is where we discuss the results obtained throughout the dissertation and provides insight into what new information, we add to scholarly literature

on ancient ghost introgression in equine genome. Ultimately chapter 7 is devoted for conclusions and future directions of the project.

CHAPTER 2
BACKGROUND

2.1 Genetic variation

For centuries humans have been improving phenotypes of animals and plants via artificial selection and inbreeding. In the beginning, selection was performed considering only visible traits like patterns in flowers, coat color in animals, and measurable traits, for example body weight of chicken, volume of milk produced by a cow, performance in domestic horse (*Equus caballus*) etc. [5] This variation observed in phenotype is a result of both genetic factors and environmental factors. For example, the milk production of a cow depends on both its genetic content and the environment such as quality of feed provided to the cow [6].

Genetic variation is mostly a random phenomenon that is responsible for the differences in individuals of a species. It is what makes individuals unique in each population. Mutations are the ultimate source of genetic variation. A major contributor to inter-individual variation within a population is recombination [7]. This is the process of rearranging genomic segments between homologous chromosomes during meiosis, which is a major step in the production of gametes in sexually reproducing organisms. Hundreds of mutations can occur spontaneously within a single cycle of reproduction [7]. As a result of these mutations there will be different forms of the same gene (called alleles) among individuals of the population. A population with many different alleles at a single

chromosome locus will have a high amount of genetic variation for that locus. Genetic variation is essential for the evolution of a species because it introduces new alleles into a population and gives the ability for the species to adapt into new environments through natural selection [8]. For example, Yakutian horses live in one of the coldest places on earth, far northern Siberia and they have well adapted compact body conformations, extremely hairy winter coats, and acute seasonal differences in metabolic activities. After comparing genomes of these present-day domesticated horses with wild Przewalski's horses and ancient horse genomes researchers have found the contemporary Yakutian horses do not descend from the native Przewalski's horses in that region but are a result of domestication following migration of the Yakut people [9]. So, Yakutian horses report one of the fastest cases of adaptation through genetic variation into the extreme temperatures of the Arctic region.

Similarly Tibetan horses have evolved important genetic mutations to adapt into high altitudes. As elevation increases the barometric pressure decreases resulting in fewer oxygen molecules in the air, which causes hypoxia. When oxygen is low, the body produces more red blood cells through angiogenesis and long-term exposure to hypoxic environments will cause pulmonary hypertension. As a result, living in a hypoxic environment can be detrimental over time. In order to overcome these difficulties Tibetan horse need adaptations. Researchers have found two missense mutations in the *EPAS1* gene that appear to be strongly associated with oxygen transportation and blood circulation in Tibetan horse at hypoxic conditions [10]. These mutations increase the stability of *EPAS1* gene and its affinity to *HIF1B* transcription factor. The Hypoxia Inducible Factor (HIF) pathway gets activated as a response to low oxygen concentration and it modulates

the hypoxic response in the cell. Humans and domesticated animals have flourished in hypoxic environments for centuries. So, they have done GWAS in animals that are adapted to high altitudes vs animals that live in low altitudes. In addition to the Tibetan horse, numerous other domesticated animals like dog [11], pig [12], chicken [13], goat [14], and cattle [15] have shown similar adaptations. All these animals including humans live in these areas have independently developed the same response to hypoxia through different mutations in the same gene (*EPAS1*) [16]. The genes (*EPAS1*, *EGLN1*, *PPARA*) involved in the HIF pathway diverge slightly at the nucleotide and amino acid level among populations, but function of the gene remains conserved [17, 18]. This is a good example of convergent evolution, where organisms that are distantly related and living in the same geographic regions develop different genetic variation which produced the same or similar phenotypic adaptations. Apart from genetic variation in animal genomes a considerable amount of literature has been published on genetic variation as an aid for adaptation in plant genomes. Namely, tropical bean (*Phaseolus vulgaris L.*) genotypes have variants associated with special adaptation to efficient phosphorus absorption [19] and different populations of glasswort (*Salicornia*) plants have different variations corresponding to flooding levels in salt marshes resulting in adaptations to different degrees of flooding [20].

Out of all the variations, advantageous mutations are most likely to remain and will propagate to the next generations. The mutations which are deleterious for the organism will be removed by natural selection. If these deleterious mutations have major effects on a population selection will quickly remove them. Although they affect the fitness of individuals, they will stay for a shorter period of time resulting in lower frequency. On the other hand, selection will act slowly on mutations that slightly reduce the fitness of a

population. So, they will stay for a longer period of time at higher frequencies and will end up affecting more individuals. Apart from selection, genetic drift is the other important phenomena that affects allele frequency. It is the change in allele frequency in a population due to random sampling of organisms [21]. Unlike natural selection, genetic drift does not depend on the beneficial or harmful nature of alleles. Instead, it solely depends on the random sample of individuals chosen to maintain the next generation. The effect of genetic drift is larger when the frequency of an allele is low and the population is small, the effect is smaller when the frequency of an allele is high, and the population is large. Therefore, genetic drift can reduce genetic variation by causing gene variants to disappear completely [22]. It also can cause initially rare variants to become more frequent and get fixed in a population. In a natural population both genetic drift and natural selection act on the mutations.

Studies of genetic variation can be utilized to understand the evolutionary history of populations. In the case of domesticated animals breeding and selection have played an important role in shaping their genomes. With the arrival of new molecular technologies and analysis methods researchers have been able to successfully employ genetic variation studies to evaluate these genomes.

2.1.1 Signal Nucleotide Polymorphisms

One of the most significant discoveries in modern genetics is that the genetic variations are the key to successful evolution and the reason for some diseases. The major challenge in studying them is understanding the functional consequences of these variants. The Genome Aggregation Database (gnomAD) was formed eight years ago with the

intention of collecting and studying human genomes around the world to better understand the human genetic diversity [23]. As of 2020, the researchers have identified a total of 241 million variants in the human genome and most of these are in non-coding regions. The majority of these variants are unlikely to impact human health but some of them are connected to disease. It is an ongoing process to identify which variants are important for disease, development, and survival. Tracking the progress in the human genetic studies is the best way to have an overview of what has been going on lately in the field of molecular genetics. It will give an insight about what might be the future for animal genetics.

There are different types of genetic polymorphisms studied so far at the nucleotide level such as Restriction Fragment Length Polymorphism (RFLP) and microsatellites. RFLPs utilize the variation in length of DNA fragments produced by a given restriction enzyme [24]. The RFLP profile is different among individuals. They were used for hereditary disease diagnosis, paternity tests and forensics etc. RFLP markers were mainly used for the first large scale effort of producing a human genetic map [25]. Later this technique was replaced by microsatellites, which are short 6 to 10 base pair tandem repeats throughout the genome. Microsatellites became popular because of the possibility to generate genotypes by simple Polymerase Chain Reaction (PCR), followed by allele sizing on polyacrylamide gels and the high number of alleles present at a single microsatellite locus. Due to relatively high mutation rates this technique has been used to assess the genetic diversity within and between species and to develop linkage maps for species of agricultural interest, with the main ones being the cow [26], pig [27], chicken [28], sheep , goat [29], and horse [30]. Among horses, Thoroughbreds are the oldest domesticated horse breed derived from few founders. The microsatellite diversity in this closed population was

used to study the genetic variation, and the contribution of founder animals for this variation [31]. Similarly, this marker technique was used for numerous other studies and dominated the field of molecular genetics for a decade.

Nowadays with the development of High Throughput Sequencing (HTS) technologies and increasing availability of Whole Genome Sequence (WGS) data SNPs are the most frequently studied genetic variant. A SNP is a single nucleotide change in a specific locus between individuals of a population [32]. This can be either a transition or a transversion of the nucleotide bases. A transition is an interchange between purines (Adenine, Guanine) or pyrimidines (Cytosine, Thymine). A transversion is an interchange of purine for pyrimidine or vice versa. Over the years SNPs became more popular than microsatellites due to their high density across the genomes, both coding and non-coding regions, and for their ability to suggest functional impact. This feature of SNPs is important when studying closely related populations, testing candidate genes, testing candidate polymorphisms in exons, promoters, or other important regions such as splice sites or other regulatory regions.

During past few years numerous studies have been conducted on the horse genome, domestication process together with demographic changes, for example population expansion, reduction, and admixture. Introduction of efficient genotyping tools, their reproducibility and automation of the process had a huge impact. A variety of SNP arrays (50K, 170K, 600K and 770K etc.) have been developed and they are commercially available for several important domesticated animals including the horse.

2.1.2 Equine reference genome

All these successful attempts on developing horse SNP arrays were inspired by the availability of a complete horse reference genome. The horse genome is about 2.7 Giga bases (Gb), which is larger than dog genome (2.5 Gb) and smaller than bovine genome (2.9 Gb) and human genome (3.3 Gb) [33], [34]. The sequencing of the first horse draft genome started in 2006 which was known as Horse Genome Project. It was the result of a 10-year long collaborative effort of a group of international horse genome researchers. As mentioned before Thoroughbreds are the essence of horse athletics and is an economically important breed. Genomic DNA from a Thoroughbred mare named "Twilight" from Cornell University in Ithaca, NY was used to generate the equine genome using a whole genome shotgun sequencing approach. Twilight was selected from a panel of candidate horses because of the high level of homozygosity observed in its Major Histone Compatibility Complex (MHC). Otherwise, it would be difficult to assemble the genome. Because MHC is a set of highly polymorphic genes essential for adaptive immune system. The DNA was sheared into fragments, 1kb, 4kb, and 10kb since Sanger sequencing methods was used. The complete sequence (EqCab2.0, 6.8x coverage) was published in 2009 along with a horse genetic variation map of one million SNPs [35]. This provided scientist a view of genome wide genetic variability in horses and they could identify more than 90 hereditary conditions related to infertility, inflammatory, musculoskeletal, neuromuscular, cardiovascular and respiratory diseases in the horse which were genetically similar to diseases in humans [36]. This provides evidence that horse is a good option for a model species.

In 2012, whole genome shotgun sequence data was published using an American Quarter Horse mare [37]. This time Next Generation Sequencing (NGS) technologies were used to sequence the genome at a coverage of 24.7x. The availability of the Quarter horse genome data increased the number of known polymorphisms to 3.1 million. This was the first attempt of sequencing a horse genome using NGS. Other than the SNPs they were able to identify 193K insertions, deletions (INDEL) and 282 Copy Number Variants (CNV). The subsequent annotation of the variants disclosed functions related to sensory perception, signal transduction, and immunity. After this more horses from, different breeds were sequenced to have a better understanding about their genetic basis. In 2018, the Twilight genome was re-sequenced and re-assembled to produce newest version of Equine reference genome EqCab 3.0 [38]. It was built on the base of the Sanger data used to create the reference genome EqCab 2.0. The new assembly improved both contiguity and accuracy by using new sequencing technologies like Chicago®, Hi-C proximity ligation data and 16-fold long-read PacBio data, and 80-fold Illumina short read data. They have been able to achieve 10-fold reduction in the gaps observed in complex genomic regions in previous assembly and have achieved a 3% increase in the number of assembled bases.

2.1.3 Genomic tools and resources

The horse reference genome has been a valuable tool for subsequent genomic studies. Later several other genomic resources were developed, and millions of sequence polymorphisms from diverse horse breeds were discovered. Among those are three generations of SNP panels. The first-generation SNP panel Illumina EquineSNP50 BeadChip was made available since 2008 parallel to the publication of EqCab 2.0 genome. This SNP panel contained 54,602 SNP markers. The utility of the panel was assessed by

successfully genotyping 354 horses from 14 different breeds and conducting genome-wide linkage disequilibrium analysis, inbreeding and genetic distance measurements within breeds, as well as multidimensional scaling and parsimony analysis among domestic horse breeds and Przewalski's Horse [39]. About 98% of SNPs were validated by the tests. The SNPs spanned regions with an average separation of 43.1 kb along the autosomes. There were also few gaps larger than 500 kb and the SNP chip was not very powerful for across the breeds analysis. This SNP panel was replaced by the development of the second-generation SNPs panel, Illumina Equine SNP70 BeadChip. It includes 53,500 SNPs (with MAF > 0.05) from the previous array and the rest of the SNPs were from RNAseq data, the Twilight sequence data used to build her genome, and seven breeds used in the initial development of SNP panel. Altogether there are 74,500 SNPs resulting in increased coverage across genome (average 1.5 SNPs per 50 kb) and filling up many gaps of the previous array [40]. This array also contained SNPs from chromosome X and Y. The first- and second-generation SNP arrays were used to identify several disease traits and phenotypic traits of interest. Identification of candidate genes for lavender foal syndrome, genetic mutations for foal immunodeficiency syndrome and alternate gait are few of the examples [41].

The third generation SNP array was introduced a few years ago overcoming the limitations in BeadChip technology. Previous arrays were bias towards a single horse (Twilight) and limited additional horse breeds. They also had a low SNP density and not enough SNPs from Y chromosome. This latest array was developed using sequence data of 156 horses from 24 distinct breeds. Initial analysis discovered 23 million SNPs and eventually 670,805 SNP markers were filtered into Equine 670 k SNP Array (MNEc670k)

[42]. This SNP array was successfully used to identify a nonsense variant responsible for Naked Foal Syndrome in the Akhal-Teke and to identify variants responsible for curly coat in horse [43, 44]. With the introduction of enhanced equine reference genome SNP array coordinates were remapped against EqCab 3.0. Several other genomic tools called whole exon array and genome tiling arrays were later replaced by more comprehensive whole genome shotgun sequencing. The importance of WGS is evident by its role at developing above mentioned genotyping tools. Today equine researchers have access to lot of horse WGSs leading to diverse successful genetic studies. Development of cheaper and faster HTS technologies have paved the way for this advancement in horse genomic studies.

2.1.4 Next Generation Sequencing Technology

As mentioned earlier efficient NGS platforms have opened the possibility of generating large-scale sequence data of non-model organisms in a surprisingly short period of time. It is more economical as a very large amount of sequence reads can be produced through a single sequencing cycle for a reasonable cost. It is convenient since the technology is capable of sequencing, discovering variation, and genotyping several thousands of markers even in genomes where there is no prior genetic information, because primer designing, and cloning are iterative difficult processes. Many studies that initially employed genome wide association techniques are currently using NGS to further investigate regions of association [41]. It is also used for RNA sequencing data which help to compare gene expression patterns, identify novel transcripts, and genes.

There are several NGS platforms with varied read length, run times and costs. Among those Illumina technology is the most popular platform. Once extracting genomic

DNA from cells, it is sheared into small fragments. Two distinct universal adaptors are ligated at each end of the fragments to prepare them for sequencing libraries. These adaptors enable the library fragments to be amplified by the same pair of primers and enable the library fragments to bind with the oligonucleotides attached to the flow cell floor. A cluster of the same fragment will be generated through bridge amplification and these clusters will be sequenced simultaneously. Custom barcodes are attached to the universal primers used for amplification allowing for high order multiplexing of many different samples. The resulting sequence reads will be de-multiplexed, quality filtered, mapped against a reference will undergo variant calling step to identify SNPs, followed by genotyping.

These SNPs will be utilized in a variety of important genomic analysis such as marker-assisted selection, genomic selection, mapping of quantitative trait loci, and GWAS. In this dissertation, we describe how SNPs were utilized to study ancient introgression by investigating patterns and distribution of homozygosity across equids. Before moving forward, a general idea about introgression and horse evolution will be presented below.

2.2 What is introgression

When an advantageous mutation occurs randomly in a population, with time it will spread across most of the individuals in that population. In about $10^5 – 10^6$ years this new advantageous allele will get fixed in the population (i.e., become the only allele shared by the population) [45]. This is the process of evolution through natural selection which takes millions of years and it's very slow. But if an organism gets introduced into a new environment or its current environment changes this process will not help it to adapt for

survival at genetic level. But if there is a closely related species already successfully evolved with well adapted genes for the new environment and if these two species interbreed the progeny will receive a copy of the well adapted genes from the second species. This is where the concept of hybridization come into play. It is a much faster method to introduce alleles to a second species.

Detailed morphological analysis of hybrid plant populations have shown that the inter-species hybridizations have various ultimate effects on populations [46]. The most common result was the infiltration of germplasm of one species into that of another through repeated back crossing of the hybrids to their parental species. If two species A and B come to effective contact, they usually do so under conditions which greatly favor either A or B. If A and B prefer different habitats, seldom or never is there a habitat equally acceptable to both. So, if they meet at all it will usually be in a situation quite favorable to one of the

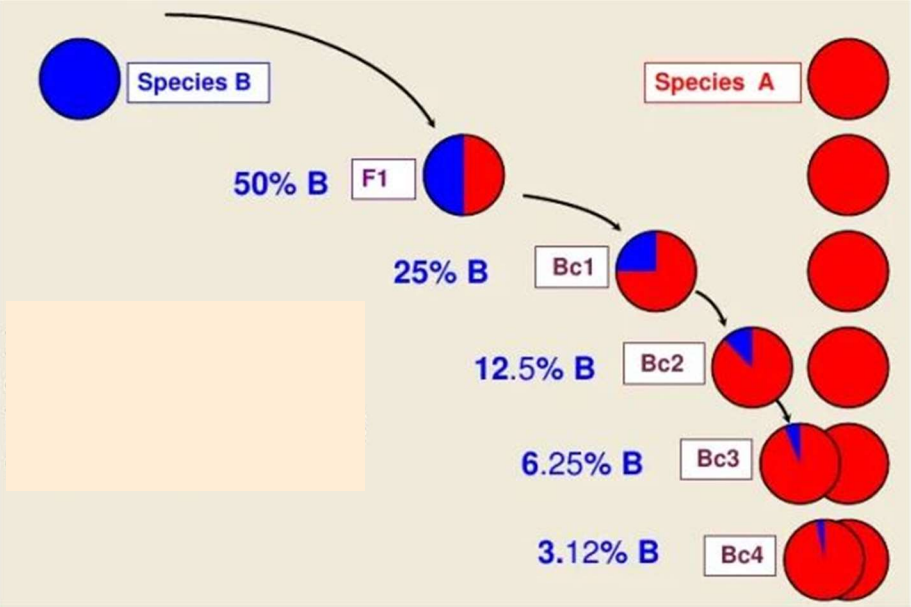


Figure 3: Movement of genes from one species to another by recurrent backcrossing of hybrids to their parental species[1]

[1] https://www.gnxp.com/WordPress/2016/01/08/admixture-vs-introgression-is-there-a-difference/

species but just fairly to the other. Therefore, if hybrids are produced, they tend to cross back to the more abundant species. The progenies of these secondary hybrids are likewise crossed back again as shown in Figure 3. The final result will depend upon the balance between the deleterious effects of the foreign germplasm and its advantageous effects in the environment where the hybridization has taken place or to which the hybrids may spread. Over the years after studying a number of flowering plants, it has been shown hybridization between species is a common effect. So it was given the term introgressive hybridization by Anderson and Hubricht in 1938 [47].

Today we define introgression as the incorporation of novel alleles from one species into another. It is called adaptive introgression if the gene transfer result in an increase in fitness of recipient species. This cannot be just one isolated event where one organism from a population breed with another organism in a second population. There must be several of these events of inter-breeding which will lead to fast propagation of these well adapted genes into the ill equipped first species. Now the selection can act on segregating genomic blocks instead of one or two new alleles produced by random mutations. This will increase the rate of evolution for species introduced into new environment. Introgression is different from simple hybridization which is a uniform mixture of genetic materials from two parents. Contrarily, introgression results in a complex, highly variable mixture of genes, and only involve a minimal percentage of the parent genomes.

The application of the terms hybrid and species becomes controversial with successive back crosses between an original first-generation hybrid and one of the parental species. F1 is clearly entitled to the term hybrid but the progeny of its first back cross and

progeny after few back crosses cannot be distinguished morphologically from the unhybridized animals from the same species. So, the term hybrid is safe to be used for obvious intermediates between recognized species. When describing introgression, the term species will be used in a broad enough sense to include individuals with hybrid ancestry as shown in Figure 4. Species A contains individuals with genes introgressed from species B as well as original species A individuals, who are not introgressed. With time the advantageous introgressed genes spread across entire species A.

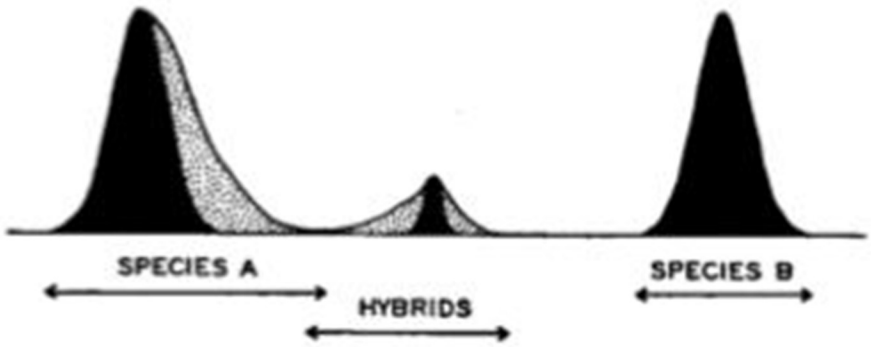


Figure 4: The terms species and hybrids in a case of introgressive hybridization. Solid black areas represent original species and first-generation hybrids. Dotted areas represent later hybrid generations and back-crosses [47].

In animal husbandry we always see animals get artificially crossed to get better meat, temperament, carcass traits or coat color. The species that are known to productively interbreed are cattle, bison, yak from genus *Bos*, and domestic dog, coyote, wolf from genus *Canid* and all the cats except leopards in genus *Felis*. They all productively interbreed and produce fertile offspring. And exchange important genetic information between genomes. We talk about natural introgression events in this dissertation.

2.2.1 Examples of adaptive introgression

Introgression is an important source of genetic variations in both plants and animals. However, introgression in plants is extensively studied compared to animals. This is due to the increased, indiscriminate spreading capacity and shorter breeding cycles of plants compared to complex tissues and organ systems of animals which require integration of larger genomic segments, as well as species specific preferences in sexual reproduction. These features restrict the frequency of hybridization events between animal species. Therefore, reducing the number of adaptive recombinants and increasing the reproductive isolation in animals. Initially, zoologists were not supportive of the idea of animal introgression because hybrid animal taxa appear to be rare due to above reasons and the difficulty in detecting them [48]. But later some zoologist became more open to the concept based on experimental results on Australian fruit fly and evidence from avian evolution [49, 50]. The Australian fruit fly (*Dacus tryoni*) seems to have adapted to extreme temperatures through introgression with closely related species. It was tested by maintaining pure and hybrid populations. After two years the hybrid lineage started to increase rapidly under higher temperatures indicating effect of adaptive introgression. In a 2012 study, the role of introgression was identified in fixation of mimicry genes in two butterfly subspecies *H. melpomene amaryllis* and *H. melpomene timareta ssp. nov*. in the genus *Heliconius* [51]. There are 43 species in this genus with different color patterns. The two subspecies in the study shared similar color pattern despite being evolutionarily distant. The researchers were able to find 2% to 5% introgression in mimicry loci between them.

In some situations, these events can be discovered by sequencing ancient DNA from fossil records. One such situation is discovery of ancient introgression between

humans and their extinct relatives. In 2010 researchers found the first evidence of gene flow from Neandertals into ancestors of non-African humans [52]. Neandertals are the evolutionarily closest relative of present-day humans and they lived in Europe and western Asia about 30,000 years ago. The results suggested between 1 to 4% of the genomes of Eurasian people are derived from Neandertals. Out of these genomic segments certain segments are found to be positively selected in ancestral modern humans. This initial study found such positive selected genes involved in metabolism and in cognitive and skeletal development. Denisovans were identified as the other branch of archaic hominids who lived in large parts of East Asia while Neanderthals were present in Europe and western Asia [53]. Their fossil records revealed that present Melanesians and East Asians share 3% and 0.2% genomic segments respectively with Denisovans [54]. There have been numerous studies over the last decade regarding the archaic adaptive introgression on modern human populations [55-57].

Interestingly not all the introgressed archaic haplotypes were fixed in the modern human populations. There has been evidence that selection has acted against majority of the introgressed regions in human genome. Particularly regions with high gene density had low Neandertal introgression, mitochondria and Y chromosome also had no Neandertal ancestry [58]. A 2016 study suggests this might be due to the differences of effective populations sizes of doner and receiver populations [59]. The Neandertals had accumulated many weak deleterious alleles over time which were neutral in their small population. Once these got introduced into larger human populations they were strongly selected against and removed.

Although most alleles were removed from human population few were left behind due to their beneficial nature in humans. *EPAS1* locus is one such example for adaptive introgression. It was mentioned earlier how mutations in *EPAS1* locus contribute to adaptation of Tibetans into high altitudes. After scanning large set of worldwide human populations researchers have identified the responsible haplotype is only found in Denisovans and Tibetans [60]. Not only Tibetan humans but dogs also have the same response for long term hypoxia. The studies have shown *EPAS1* locus in Tibetan dogs was nearly identical to what was found in Tibetan grey wolves and was introgressed into dogs from the grey wolves [61]. This gene is a clear example of adaptive introgression in the wild as a mechanism to improve species abilities.

Another group of researchers were able to identify possible ancient interspecies introgression in pigs (*Sus scrofa*) by analyzing WGS of 69 animals [62]. Local pigs from southern and northern parts of China have developed distinct thermoregulatory mechanisms to adapt for hot and cold temperatures. In this study the researchers were able to find new loci responsible for regional adaptations for low- and high-latitude environments and they also found an X-linked, large (14 Mb) region in northern Chinese pigs which seems to be introgressed from an extinct *Sus* species about 8.5 million years ago (MYA). This is much earlier than the known evolutionary history of this pig species which is about 5 MYA. Their findings provided new insights into the evolutionary history of pigs at that time and an example for the role of ancient ghost introgression in adaptation in mammals.

2.2.2 Ghost introgression

Ancient introgression has become an interesting topic nowadays since it opens doors to the methods of gene flow into extant species from distant past. Evolution is a continuous process of supporting the continuous existent of the fittest. So many lineages go extinct in the process, and some might not even leave fossil remains. But many of these now extinct lineages would be closely related and possibly hybridized with the lineages that have given rise to extant species. This is called ghost introgression [63]. Some studies have found genetic signatures of these ancient ghost introgression events in modern day genomes. The study on Chinese pigs, mentioned in above section is a good example for this phenomenon.

Similar to the study on Chinese pigs, Kuhlwilm and group has recently found exceptionally divergent haplotypes in the bonobo (*Pan paniscus*) genome [64]. Through an extensive study combining several available methods they tried to find whether these haplotypes were introduced into bonobo through introgression with chimpanzees (*P. troglodytes*) or by an extinct primate lineage. Their findings revealed that bonobos received between 0.9% and 4.2% genetic material from an extinct species whose fossil records are not yet found. Fossilization is a rare phenomenon. Not all the ancient species are preserved in fossil records and most fossils do not yield sufficient genetic material. So, we are probably missing numerous ancient introgression events that left traces of extinct species in present-day genomes.

The 2018 study on rorquals provided a good example that introgressed regions could significantly speed up the rate of adaptation and speciation [65]. In this study they analyzed Genomic DNA from extant cetaceans and have been able to reveal an interlaced

evolutionary history with high levels of introgression. Most of these introgression events between cetacean species occurred around 10.5 to 7.5 MYA during the fast radiation of rorquals. These ancient radiation events need to be further investigated, but there is enough evidence that ghost introgression could significantly speed up the adaptation process by harnessing the power of genetic variation introduced by hybridization with other taxa rather than waiting for *de novo* mutations. In addition, these studies can provide insights into the morphology of extinct lineages. It is important to develop robust methods to identify ghost introgression where extant species will benefit, and we could eventually learn the genomic composition of extinct species.

In this dissertation we describe our efforts to detect ghost introgression in the domestic horse. To do this we need to have an understanding about gene flow between equids and the timeline of their dispersal and account for the role horse domestication played in guiding their genetic composition.

2.3 Horse domestication

According to current literature domestication of plants and animals started at least 12,000 to 15,000 years ago with the domestication of dogs [66]. In this process humans have intentionally selected and altered the members of a species possessing desirable attributes. Some of the oldest evidence suggest early horses were hunted by Middle and Upper Paleolithic hominins in Europe for their meat and bones [67]. They have found remains of horse bones and wooden spears from a site in Schöningen, Germany, dated around 400,000 years ago. Among these discoveries there are thousands of deep cave wall paintings of horse, portrayed in their natural poses as well as various ways of hunting them.

Although it is not clear when or where exactly domestication occurred, the evidence suggests it could be in Eurasian steppes about 5,500 years ago [68].

Horse is the last domesticated animal of importance. Goat, sheep, chicken, pig, cattle, dog are examples of animals that were domesticated before horse. The reduction in their average body sizes, teeth and reduction in skull size are proof for living in domestic environments for a longer time than horse. Later horses were used for warfare, transportation, and agriculture and have played an important role in human civilization. Even today they continue to play a key role in leisure industry and are selectively bred for performance traits and appearance traits. There are about 500 different horse breeds today. This large genetic diversity is the result of multiple domestication event throughout Eurasia for an extended period. Studies show that the horse did not go through a domestication bottleneck as the result of continuous gene flow between domesticated and wild horses [69].

2.3.1 Gene flow in equids

The first ancestors of the horse appeared on earth about 55 million years ago during early Eocene era [70]. It was about 13 million years after the great extinction of dinosaurs. This makes the horse one of the oldest mammals in north America that survived to modern time. This earliest known ancestor was a small dog like animal called *Eohippus* [71]. Over the course of about 45 million years this small animal evolved into the horse we see today (Figure 5). Based on the fossil records *Eohippus* has had a greater number of functional toes, four in the front and three in the hind feet while modern horse has only one functional digit in each foot. Also, *Eohippus* has a teeth structure which suggest omnivore lifestyle

feeding on a variety of plant foods while horse today has grinding teeth suitable for gazing on grass fields. Moreover, this ancient member of the horse family was probably more a browsing animal in a forest environment.

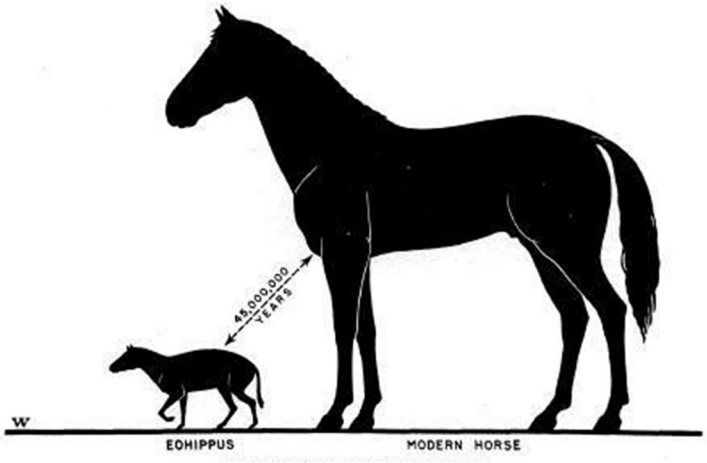


Figure 5: Morphological evolution from *Eiohippus* to *Equus* [71].

Paleontologists have discovered that the horse evolution is a very complex tree with branches leading to extinct species and some leading to species closely related to genus *Equus* [72]. According to fossil records Miocene and Pliocene eras are considered the time of the true equine. Merychippus who lived during Miocene era is considered one of the most successful early horses (Figure 6). In fact, scientists believe Merychippus was the ancestor for at least 19 other species. Hipparion was one such successful species that existed for roughly 22 million years and went extinct about 781,000 years ago.

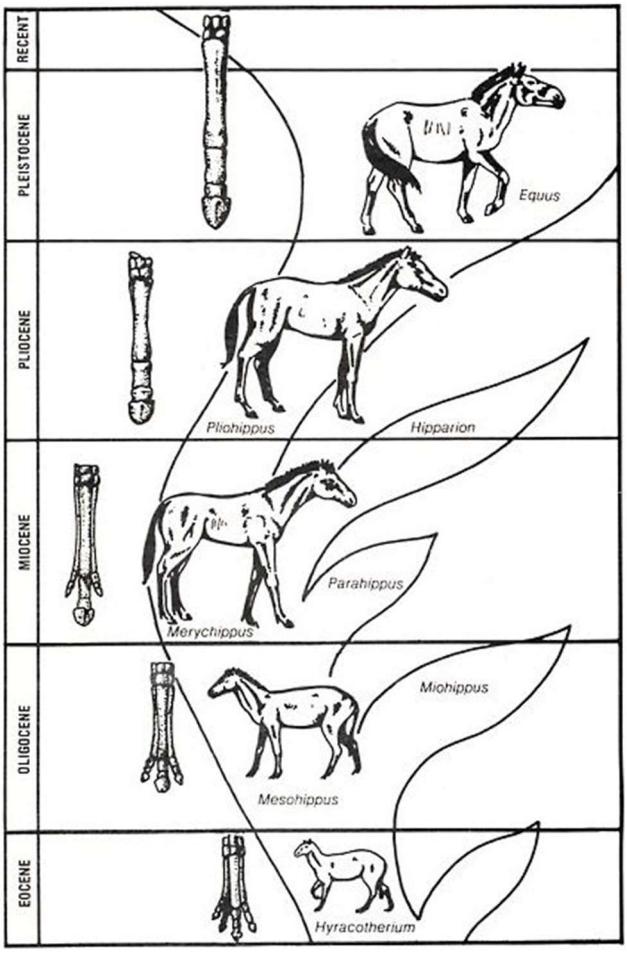


Figure 6: Adaptive radiation of *Equus*[2]

Horses belong to family *Equidae* within order *Perissodactyla* which includes two other extant families *Tapiridae* and *Rhinocerodae.* Genus *Equus* is divided into three clades including horses, zebras, and wild asses. The domestic horses are also called caballines because of the name *caballus* and the rest of the Equus are called non-caballines. We know *Equus caballus* diverged from other equids about 4.6 million years ago [73], and they continued to thrive and evolve in North America.

[2] http://www.bio.miami.edu/dana/dox/equus_evolution.html

Meanwhile early horses expanded their range by crossing bridge of Beringia from North America to Eurasia. The ancestor of modern horse crossed the bridge about 800,000 to 1 million years ago. Interestingly Equus caballus became extinct in the New World around 11,000 years ago. A variety of reasons might have caused this extinction including climate changes and been hunted by early humans. Luckily, they survived in Old World thanks to earlier migrations and were domesticated in central Asia as described above. Eventually they were reintroduced to the New World by early settlers about 500 years ago. We can imagine the introgression events we are interest in occurred between 4.6 million to 1 million years ago in North America. Because at this time horses, other equids and their common ancestors were all alive and living together. They must have had the opportunity to interbreed and transfer genetic material back into the horse genome without the barrier of chromosomal plasticity we discussed earlier.

In 2014 Haken and his group found evidence for gene flow between equine species despite the differences in their chromosomal numbers [74]. They have found evidence for four main gene flow events (Figure 7). Three of these events occurred among the non-caballines within the last 350,000 years. One from kiang (*E. kiang*) into the donkey (*E. a. asinus*) lineage, one between the Somali wild ass (*E. a. somaliensis*) and Grevy's zebra (*E. grevyi*), and one between African asses and the mountain zebra (*E. z. hartmannae*). They also detected evidence for gene flow from caballines into non-caballines during the first divergence within *Equus*. It is estimated that this gene flow ceased 2.1–3.4 MYA, which closely matches the paleontological evidence for the non-caballine dispersal out of America. Unlike for the humans, fossil records for horse are rare for times before domestication. Hence there will be more historical introgression events hidden from us. In

this project we are endeavoring to identify these ancient putative introgressed regions in the equine genome.

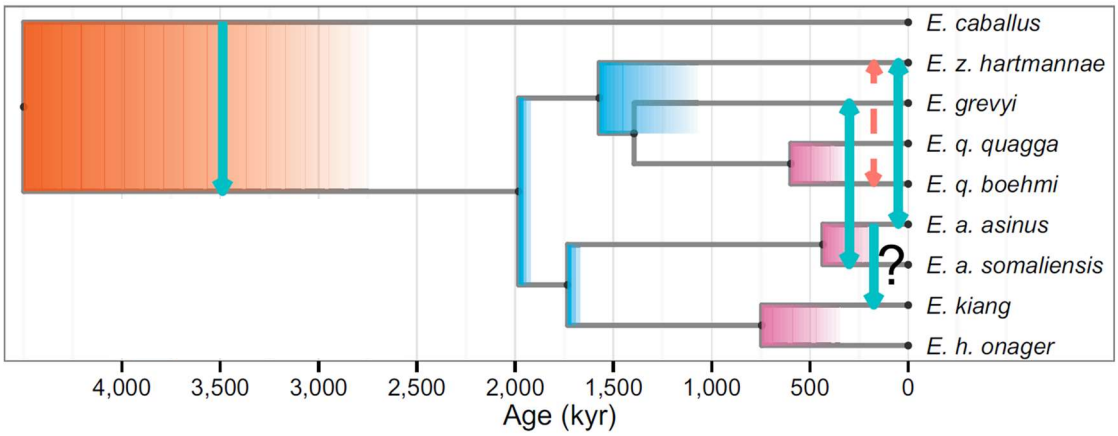


Figure 7: Events of gene flow between populations. The arrows represent the gene flow events. The Divergence time and population split times are indicated by darker and lighter ends of the colored rectangles [74]

2.4 Introgression detection methods

As mentioned earlier, analysis of genome sequences from archaic and modern humans have revealed multiple episodes of admixture between highly diverged hominids. The methods developed to locate DNA segments introgressed during these events provided a better insight to human evolution. Introgression detection methods rely on diverse population genetic statistics based on allele frequency across populations, sequence divergence, and Linkage Disequilibrium (LD). The following is a summary of few of the methods.

2.4.1 *ABBA-BABA* or Patterson's *D*-statistics

ABBA-BABA [52] on single nucleotide sites or Patterson's *D*-statistics [75] was originally developed to test for admixture between Neanderthals and modern humans. In this test sequencing data of three individuals from three populations are compared using

their allelic differences relative to an out-group. More specifically it is tested if the data are consistent with the tree (((*H1*, *H2*), *H3*), out-group). In the case of human introgression detection, *H1* and *H2* are individuals from modern human populations and *H3* is a Neandertal. They have used a chimpanzee as the out-group (*O*). The symbol *A* is defined as the ancestral allele state seen in *O*, and *B* as the derived allele at any polymorphic site in the alignment. Only the sites with two types of alleles in which *H1* and *H2* differ from each other and *H3* has the derived allele *B* were used for the analysis. For each of these sites one randomly sampled allele was selected for each of the individuals (*H1*, *H2*, *H3* and the *O*). So, the sites will have one of the two polymorphic patterns "*ABBA*" or "*BABA*" (Figure 8). *ABBA* patterns are where *H1* and *O* have the ancestral allele, and *H2* and *H3* have the derived alleles. *BABA* patterns are where *H2* and *O* have the ancestral allele.

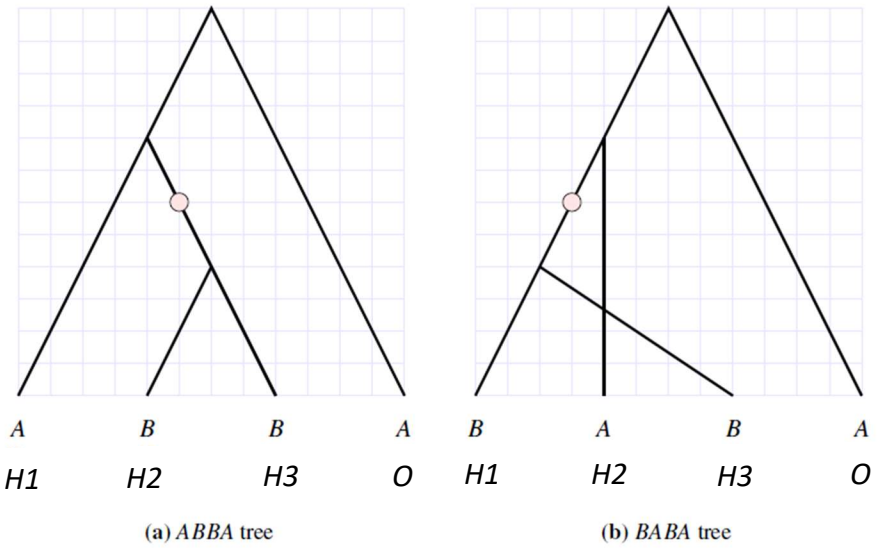


Figure 8: The gene trees that produce *ABBA* and *BABA* patterns

The null hypothesis is that there is no gene flow between species and tree is correct (2.1). Under this hypothesis *ABBA* and *BABA* patterns are equally likely to occur (*nABBA*

is the number of *ABBA* sites and *nBABA* is the number of *BABA* sites) and we can decide both patterns are due to incomplete lineage sorting.

$$D = \frac{(nABBA - nBABA)}{(nABBA + nBAB\ \ )} = 0 \qquad (2.1)$$

A deviation from this equilibrium is expected if there has been gene flow between *H3* and *H1* or *H2* or the tree is not correct. A positive value of *D* can be interpreted as *H2* being closer to *H3* than *H1* is to *H3*. A negative value of *D* can be interpreted as *H1* being closer to H3 than H2 is. *D* statistic that differs significantly from 0 is evidence of gene flow between species. This approach provides an estimate of putative admixture between closely related species where the admixing population *H3* is known and genomic data is available for *H3* [52, 73, 74]. This method is not beneficial in detecting ghost introgression where the admixing population (*H3*) in not known and genetic material is not available.

2.4.2 Hidden Markov Model (HMM)

HMMs are the other commonly found method in literature. They are statistical models that capture hidden information from observable sequential symbols (e.g., a nucleotide sequence). They have many applications in sequence analysis, in particular to predict exons and introns in genomic DNA, identify functional domains in proteins (profile HMM) and align two sequences (pair HMM). Racimo and their group have utilized the association between ancestral SNPs to detect candidate regions using HMM [76]. Here the hidden state represents the ancestry of an individual at a certain genomic location while observed states model the data.

CHAPTER 3

MAPPING AND VARIANT CALLING

3.1 Overview

This chapter provides an overview of the sample data set used for the analysis, their mapping, and variant calling process. The following flow chart in Figure 9 summarizes the main steps of the pipeline. The steps up to variant calling will be described in detail along this chapter.

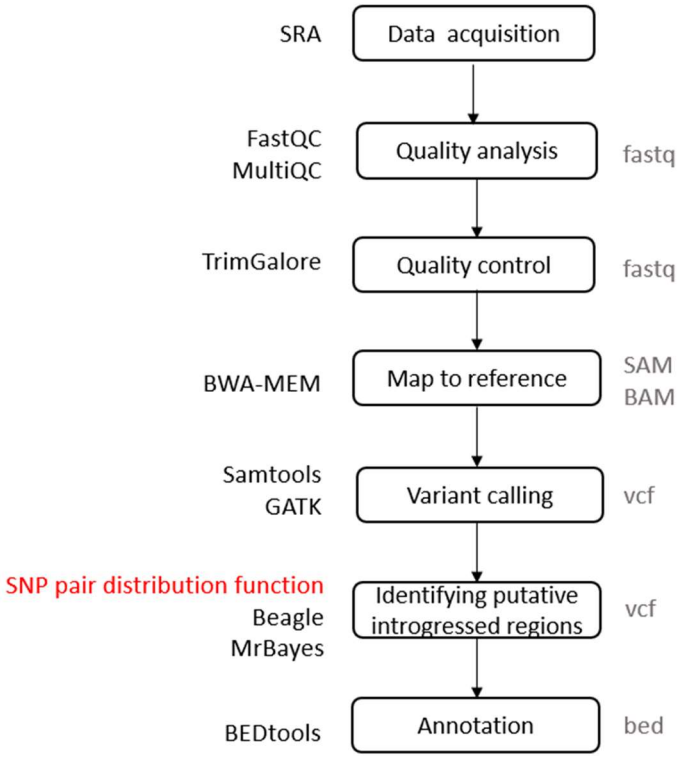


Figure 9: A summary of the main steps of the introgression detection pipeline with the tools used and the output file formats.

3.2 Sample information

In this study, all extant equine species are analyzed at whole genome scale. Other than that, one Quagga zebra genome which went extinct in early 1900s is also analyzed [77]. There are nine non-caballine genomes involved in the study and seven of them are previously published by Jónsson in 2014 [74]. They have collected blood samples from zoo animals for extant species, and hairs from a Quagga specimen preserved at the Musée des Confluences, Lyon, France. In addition, fourteen caballine genomes are also included in the current study. Out of these the standardbred and two of the thoroughbreds were from a previous study on equine *CXCL16* gene [1]. Additional information regarding the samples is provided in Table 2.

Table 2: Summary of the sample information. The animals were sequenced using paired-end, short read, Illumina HiSeq technology

Latin name	Common name	Biosample	Depth of coverage	Phred quality score	Source
		Non-caballines			
E. Quagga quagga	Quagga	SAMEA3166709	19.5x	35	Jonsson 2014
E. Quagga boehmi	Plain's zebra/boehmi	SAMEA2797679	29x	37	Jonsson 2014
E. Grevyi	Grevy's zebra	SAMEA3166710	24.6x	38	Jonsson 2014
E. Zebra hartmannae	Mountain zebra	SAMEA2802528	25x	37	Jonsson 2014
E. Kiang	Tibetan wild ass/kiang	SAMEA2802529	24.2x	37	Jonsson 2014
E. Hemionus onager	Onager	SAMEA2802530	28.8x	38	Jonsson 2014
E. Africanus somaliensis	Somali wild ass	SAMEA2802531	34.7x	39	Jonsson 2014
E. Asinus asinus	Donkey D_1989	N/A	47.9x	36	Doug antczak

E. Asinus asinus	Donkey D_3611	N/A	44.2x	36	Doug antczak
		Caballines			
E. Caballus	Thoroughbred twilight	SAMN02179858	35.7x	41	Orlando 2013
E. Caballus	Thoroughbred TB03	SAMN03838867	30.9x	37	Sarkar 2016
E. Caballus	Thoroughbred TB10	SAMN03838868	31.6x	37	Sarkar 2016
E. Caballus	Thoroughbred 686521	SAMEA104728877	26.7x	41	Kingsley 2019
E. Caballus	Thoroughbred 683610	SAMEA104728862	18.3x	41	Kingsley 2019
E. Caballus	Thoroughbred H_2158	N/A	39.8x	36	Doug antczak
E. Caballus	Thoroughbred H_3958	N/A	41.8x	36	Doug antczak
E. Caballus	Arabians AR03	SAMN06820322	15.5x	42	N/A
E. Caballus	Arabians AR04	SAMN06820323	14.7x	42	N/A
E. Caballus	Arabians AR05	SAMN06820324	14.9x	42	N/A
E. Caballus	Saddlebreds 3517	N/A	21.3x	36	Ernest bailey
E. Caballus	Saddlebreds 3519	N/A	23.1x	36	Ernest bailey
E. Caballus	Standardbred ST22	SAMN03838869	32x	37	Sarkar 2016
E. Caballus	Haflinger	SAMEA5721589	23.1x	41	Singer-berk 2019

3.3 Data acquisition

Most of the data used in this analysis are publicly available at National Center for Biotechnology Information (NCBI) [78]. The data we generated for this project will be made available prior to the publication. NCBI consists of many databases including sequence data, biomedical literature, and bioinformatics tools relevant for biological research. The NCBI Sequence Read Archive (SRA) is the largest publicly available repository of high throughput sequencing data [79]. As of Spring 2021 it contains peta-

bases of raw, short read sequence data and alignment information from high-throughput sequencing platforms including the Roche-454 GS and FLX, Illumina Genome Analyzer, Applied Biosystems SOLiD System, Helicos Heliscope, and CompleteGenomics platforms. It is convenient as the archives contain data from all branches of living things. SRA stores data in SRA format. The fasterq-dump tool from sratoolkit.2.9.6-1[3] was used to download the data from the SRA and convert it to fastq format (3.1). It uses multi-threading and temporary files to rapidly download voluminous datasets. Data storage and analysis was performed on the Cardinal Research Cluster at University of Louisville and Lipscomb Compute Cluster at University of Kentucky.

fasterq-dump [accession] --split-spot --split-files -O [path to output file] (3.1)

A spot is a data object corresponding roughly to a cluster on an Illumina sequencing chip, or a well on the 454 or PacBio platforms. This spot, in the SRA file contains biological information which is essentially sequence reads and technical information such as adapters, barcodes for multiplexing etc. The --split-spot option in fasterq-dump splits spots into the read pairs produced by the corresponding Illumina cluster. The --split-files option separates the forward and reverse reads in to two separate files, *_1.fastq and *_2.fastq. All the other unmatched reads are stored in *.fastq file. It is good to perform a simple check on any accession number before the conversion from SRA to fastq. The command $ vdb-dump --info [accession] gives a lot of information about an accession including the data size, number of sequences etc. After the conversion from SRA to fastq, the number of reads in the fastq files were calculated and compared against the number given by the $ vdb-dump --info command to make sure all the reads were downloaded successfully.

[3] www.ncbi.nlm.nih.gov/sra/

3.4 Quality analysis of the data

The advance in NGS technologies has resulted in high throughput sequence data in hitherto unimaginable volumes. This created challenges for biological analysis in assessing the quality of the data due to introduction of errors and bias. Most high throughput sequencers generate output in fastq format. This format combines the base calls for the sequence with an encoded quality value indicating how confident the base caller is that the called base generated was correct. Before proceeding with a study, it is a good idea to do some basic quality analysis on the data to ensure that there are no hidden problems which might be more difficult to detect at a later stage.

A freely available software tool was used to assess sequence quality. FastQC[4] (version 0.11.8) is an application which takes BAM, SAM or fastq files as input and runs a series of analysis to generate a comprehensive quality report. The FastQC report contains information about basic statistics such as total number of sequences found in the fastq file, number of poor-quality sequences, sequence length and % GC. There are also graphical representations of sequence length distribution, duplication level, per base sequence quality, sequence content, N content, per sequence quality score and GC content. Since there are several animals in the study, and FastQC generates output files for forward and reverse read files separately, there are many quality reports left to analyze before drawing a conclusion about read quality. So we used MultiQC tool to summarize the quality reports for each species [80]. This tool assesses the quality analysis output files across samples and generates a summary report which is much easier to analyze.

[4] www.bioinformatics.babraham.ac.uk/projects/fastqc/

3.4.1 Quality control

Quality control plays a crucial role at each step since poor quality data can affect the mapping and adapter contaminations can lead to low mapping efficiency. Also, the alignment quality is important in downstream successful variant detection. The tool Trim Galore version: 0.6.4[5] was used to trim the reads. It is a wrapper script which functions with FastQC and the publicly available adapter trimming tool Cutadapt (version: 2.9) [81].

There are several other solutions available for adapter trimming. HTSeq [82] and Biostrings [83] from Bioconductor are two such algorithms which provide fast, error tolerant processing of large biological data sets. But they require the user to write custom scripts to use these tools. They are more suitable for projects that deviate from standard workflow. In this project the Burrows-Wheeler Aligner (BWA) was used for read mapping [84]. Although BWA is fast and efficient it does not have trimming capabilities. So, it is best to use a stand-alone adapter trimmer like Cutadapt which is flexible and fast.

First, low quality reads were trimmed at the Cutadapt default phred score cutoff of 20. As the next step Cutadapt was used to search for adapter sequences in the 3' end of the reads. The auto-detection mode of the tool was used for this. There are three different standard adapter sequences Illumina universal, Illumina small RNA, Nextra transposase. The information on which standard adapters (13bp) were detected and removed by the tool can be found in the trimming report. According to the default stringency level assigned by the tool the reads that overlap with at least one of the adapter sequence bases were trimmed. Although the default setting may look too stringent, the adapter contamination can lead to misalignments and removal of the reads as a whole because of too many mismatches during

[5] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

the alignment process. The default maximum trimming error rate is 10% [81]. After quality filtering and adapter removal the reads can result in very short sequences. So, the sequences were filtered and the reads shorter than 20 bp were removed. For paired end reads both reads were removed if one of the reads were shorter than the threshold length. The resulting files were stored in compressed gzip format. All the information about thresholds and trimming results can be found in the summary report. Applying these steps to downloaded data ensure only the high-quality portion of data was used for mapping and further analysis.

3.5 Mapping data

After getting rid of the artifacts in the sequence reads, they need to be mapped against a reference genome. The goal of mapping is to align the reads into their respective locations in the genome. It is a key step in any analysis and reveals information about the sequence variants for the animal. There are multiple methods in use depending on the availability of a reference genome. If there is a high-quality reference genome for the organism of interest, reads can be mapped directly to it. But if no reference assembly is available or if the available assembly is fragmented and poorly annotated, there are other options. One option is to perform a *de novo* assembly. The other, easier option is to map the reads against a reference genome of a closely related species. The second option has been successfully used to determine inter-species and intraspecies variations in sheep and cattle by mapping WGS data from sheep to cattle genome [85]. They were able to identify 83 million variants out of which 78 million were homozygous in all the sheep analyzed. Those are the interspecies variation. Nearly 3.7 million heterozygous variants were identified out of which 41% mapped with orthologous positions in sheep genome and 80% of those corresponded to actual heterozygotes which means they were species specific.

This inter-species mapping technique is a good alternative since construction of a new reference genome is a time-consuming difficult task.

According to molecular phylogenetic and taxonomic studies there are 17 living species of odd-toed ungulates (perissodactyls) [86]. Out of these 17 species only horse from family Equidae and white rhinoceros from family Rhinocerotidae have publicly available reference genomes. So, the rest of the equids used for this analysis were mapped against Equine reference genome which has a 4.5 million years of evolutionary distance from the rest. The new EquCab3 assembly is better in both composition and contiguity, and has produced better results for universal ortholog analysis and comparative annotation studies based on pig, cow, white rhinoceros, elephant and human genomes [38]. All these data suggest it is the best candidate for our purpose and the following high mapping fractions of the non-caballines confirm that (Table 3).

Table 3: the mapping fractions for each animal after mapping against equine reference genome

Animal	Read count	Mapped count	Mapped in proper pair	Mapping fraction (mapped count/read count)
Non-caballines				
Quagga	2,339,672,584	1,977,936,125	1,971,155,466	84.54%
Boehmi	853,539,724	848,679,877	818,500,054	99.43%
Zebra	679,150,486	675,058,853	653,976,056	99.40%
Onager	802,301,712	796,311,543	766,915,194	99.25%
Kiang	370,670,580	367,755,287	338,193,396	99.21%
Grevyi	687,760,936	683,507,739	661,104,392	99.38%
Somali ass	1,073,371,966	1,066,337,806	1,029,601,978	99.34%
D_1989	1,078,215,552	1,073,389,228	1,009,591,524	99.55%
D_3611	627,453,154	625,734,210	606,346,194	99.73%
Caballines				
Twilight	627,453,154	625,734,210	606,346,194	99.73%
TB03	837,211,334	836,069,673	812,137,682	99.86%
TB10	866,335,638	865,173,937	844,601,818	99.87%

683610	410,252,530	408,497,172	402,927,902	99.57%
686521	426,677,664	424,951,180	419,341,632	99.60%
H_2158	939,884,241	938,029,403	914,543,560	99.80%
H_3958	977,961,911	976,286,980	955,996,198	99.83%
AR03	309,098,565	308,070,201	299,626,136	99.67%
AR04	290,848,088	289,843,305	282,616,902	99.65%
AR05	299,972,162	298,988,828	291,211,278	99.67%
3517	413,437,907	412,065,045	403,274,162	99.67%
3519	451,408,512	450,000,156	439,405,954	99.69%
ST22	873,480,996	872,155,484	842,938,014	99.85%
Haflinger	405,773,614	405,085,279	395,020,112	99.83%

As the first step for mapping, the reference fasta file was indexed with Burrows-Wheeler Alignment (BWA)-0.7.17 [87]. This step needs to be done only once per reference genome. The resulting index files can be stored and reused every time a new data set is mapped to that reference. For accession numbers which have several sequencing runs associated with them, the quality filtered fastq files corresponding to R1 and R2 read pairs for all the lanes, and all runs for a given animal were merged into two separate files MergedReads_R1.fq.gz and MergedReads_R2.fq.gz using a custom program. Five to 10 years ago, it was often necessary to run a sample over multiple lanes, and multiple runs to achieve a sufficient number of reads and depth of coverage for accurate variant detection. These two files were mapped against the reference genome using BWA-MEM algorithm (3.2). BWA-MEM align 70bp-1Mbp query sequences using seeding alignments with maximal exact matches (MEM) and then extending seeds with local alignment. Default minimum seed length of 19bp was used.

bwa mem -t 8 -R "@RG\[read group header line]" [reference fasta file path] [read1.fq.gz] [read2.fq.gz] > [output.sam] (3.2)

All reads for each animal were assigned a read group in the mapping command. A read group can contain information about the sample, library preparation and flow cell

lanes. This is important in order to identify subset of reads generated from separate libraries on the same lane during multiplexing. Each alignment record will be marked by read group ID. This will allow for identification of technical errors behind sequence artifacts in mapped data. The variant detection software will use the sample name defined in the read group to identify the samples at later stages in variant calling.

3.6 Variant calling

This is the step where we identified SNPs and small indels in our mapped data and genotype the samples at those sites. We have already mentioned that the sequencers produce large amounts of data. But sequence fragments are randomly selected from within the library for sequencing, and this stochastic process will necessarily produce artifactual low coverage. This and other problems such as assembly errors in the reference genome can make it difficult to identify true variants from machine errors. Because of that, variant identification and genotyping is a critical step in any study as the downstream analysis and interpretations depend on the accuracy of this step. Just as with the NGS technologies, variant detection tools and approaches have evolved and improved dramatically over the years. In this project we used currently documented best practices for variant calling with Genome Analysis ToolKit (GATK) [88].

The alignment file produced by BWA in the previous step was in Sequence Alignment Map (SAM) format. It is a large plain text file containing the alignment information of the mapped short reads. Because it takes a lot of space in the disk, this SAM file was converted into its compressed binary format, a Binary Alignment Map (BAM) file using SAMtools (version: 1.9) (3.3) [89].

samtools view -F 2048 -Sb [output.sam] -o [output.bam] (3.3)

While doing this, supplementary reads were removed. Supplementary reads are assigned a bitwise flag of 2048 in the SAM file, which means these reads map to different locations elsewhere in the genome beyond an obvious primary mapping. They are removed from the analysis to reduce the ambiguity. Next, the BAM file records were updated with SAMtools fixmate tool. It reconciles any inconsistencies in the bitwise flag because sometimes BWA can leave unusual flags in SAM records that will be difficult to reconcile for any subsequent processing that relies on the bitwise flag. The SAMtools fixmate tool checked whether the reads were mapped in proper pair, if the mate was mapped in the reverse orientation, was that appropriately represented and made sure the mate mapping position and chromosome are correct. During this process mate score tags were assigned for the records that will be used during marking duplicate alignments. The mate pair fixed BAM file records were sorted using SAMtools before marking duplicates (3.4).

samtools sort -T tmp -O BAM --reference [reference genome] –th reads 7 [output.bam] -o [sorted output.bam] (3.4)

picard.jar MarkDuplicates I=[sorted output.bam] O=[marked output.bam] (3.5)

The reads are coordinate sorted and indexed for the ease of access. Both optical duplicates and PCR duplicates were marked with Picard tools (version 2.7.1) [90] in order to avoid false positive variant calls (3.5). The resulting BAM file will have only one fragment from each duplicate group unchanged others will be marked with 0x400 flag and will not be used downstream. Next the files were checked for indel alignment.

Aligners like BWA works with one read at a time and does not make decisions looking at all the aligned reads to a certain region of the reference genome. So, the aligner

might assign indels to some reads and SNPs to other reads aligned to the same region. This is sometimes because aligners cannot identify indels near the end of a read and a mismatch is cheaper than a gap according to scoring matrices. A line of unreal adjacent SNPs can occur in the alignment due to this and we might lose real SNPs. We can fix this by shifting the indels around. Often same indel can be placed at multiple positions and still represent the same haplotype. So, they can be either right-aligned or left-aligned. But the standard convention is to left-align the indels. GATK LeftAlignIndels tool was used to locally realign the reads and place the indels at their left- most position (3.6) [91].

gatk LeftAlignIndels --input [marked output.bam] --OUTPUT [realigned.output.bam] --reference [reference genome] (3.6)

The realigned bam file from the above step was used in identifying SNPs and indels with GATK HaplotypeCaller in -erc gvcf mode[92]. The HaplotypeCaller is an assembly based variant caller which shows higher accuracy for indel variants. On the other hand, the first-generation position based variant callers are good at detecting SNPs but not very accurate with indels.

gatk -R [reference genome] -T HaplotypeCaller -I [realigned.output.bam] --emitRefConfidence GVCF -o [output.g.vcf.gz] (3.7)

When HaplotypeCaller comes across a region with potential variation from the reference genome it discards the already available mapping information for that region and reassembles the reads via de Bruijn-like graph. This allows the caller to identify potential haplotypes for that region. Next, PairHMM is used to determine likelihood of the haplotypes by aligning the reads against each haplotype. Based on these likelihood values the variants are assigned genotypes. We can derive physical phasing from the reads as an

added advantage. Assembly based tools are accurate but graph complexity and computational power increase exponentially as the number of samples in the study increase. As a solution for this problem HaplotypeCaller -erc gvcf mode was activated in order to produce an intermediate genomic vcf file per sample. A gvcf file is a compressed text file which stores information about both variant records and non-variant block records. All the per sample gvcf files were combined into a multi-sample gvcf file using CombineGVCFs tool (3.8). Finally, GenotypeGVCFs tool was used to do joint genotyping for the samples producing a single raw vcf file. This file was used for further downstream analysis.

gatk -T CombineGVCFs -R [reference genome] --variant [sample1.g.vcf.gz] --variant [sample2.g.vcf.gz] -O [combined.g.vcf.gz] (3.8)

gatk -T GenotypeGVCFs -R [reference genome] -V [combined.g.vcf.gz] -o [output.vcf.gz] (3.9)

3.7 Results and Discussion

The sequence data for each individual reached between 19.5x to 47.9x depth of coverage (Table 1) and more than 99% mapping rate except for the quagga (Table 2). Quagga had a reduction in mapping fraction because the sample was obtained from a museum specimen and the sample quality was low. These results allowed us to call variants on the mapped data with high confidence. After following current best practices, we could identify a total of 79,673,821 SNPs across 23 genomes spanning caballine and non-caballine species. We compared the SNP calls for the whole-genome sequencing against the publicly available Equine SNPs at Ensembl. As of March 2021, Ensembl had

20,355,589 SNPs stored for Equine genome. Out of these 67% (13,671,537) were consistent with the SNPs identified from our data. Next, we compared the SNPs to EquineAxiomHD_Chip genotyping array. Of the 735,272 polymorphic loci in the genotyping array, 69% (504,096) were consistent with the SNPs identified from sequencing data.

We have further categorized these variants using the command line version of Variant Effect Predictor (VEP version 104) from Ensembl with the Ensembl Equus caballus database (version 104) [93]. According to the VEP analysis these variants affect 29,277 genes and 57,512 transcripts that they overlap. Based on variant class classification (Figure 10) 89% of them are Single Nucleotide Variants (SNV), 5% were deletions, 3% insertions, 2% sequence alterations where and 1% were indels (an insertion and a deletion affecting 2 or more nucleotides). Next the variants were compared against the Ensembl transcripts by VEP tool to find overlaps in order to find the consequences of these variants on the protein sequences. It uses a rule-based approach to predict the effects of each allele on each transcript corrosponding to a variant. Each allele of a variant may have a different effect on the transcripts. The following figure shows the possible locations of the variants in a transcript corresponding to their consequences (Figure 11).

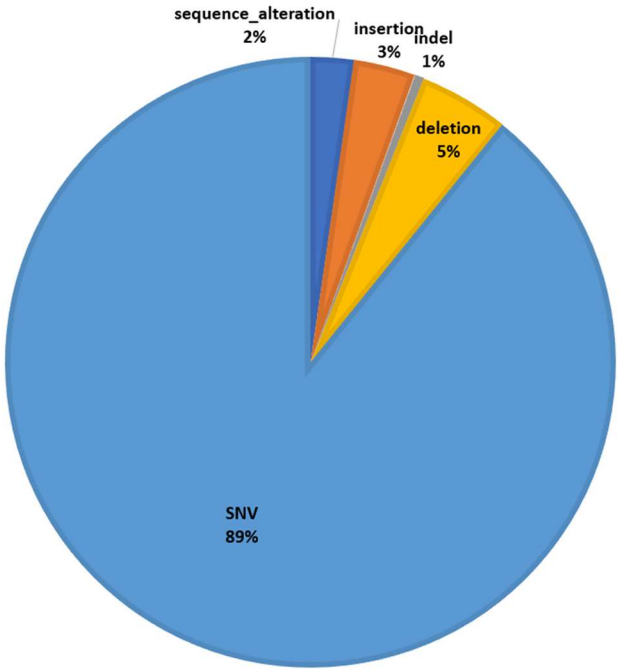


Figure 10: The class of a variant according to its component alleles and its mapping to the equine reference genome

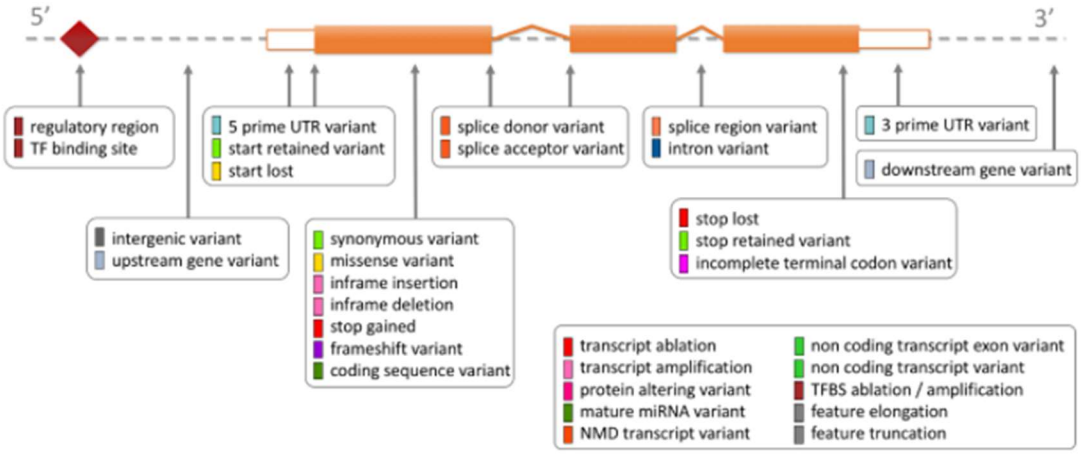


Figure 11: The diagram showing the locations of the variants in a transcript corresponding to their severity in consequences.

The variants identified in the exonic regions, start or end of a transcript and splice junctions can cause severe consequences in the resulting transcripts altering the protein products. The varants are categorized according to their impact on the proteins. The high impact variants are assumed to have a disruptive effect on the proteins. The start loss varants, stop gain varants, stop lost varants and splice variants are candidates of this category who can cause protein truncation, loss of function or triggering nonsense mediated decay. The moderate variants are the non-disruptive ones that might change effectiveness of the proteins. The low impact variants are unlikely to change protein behavior. The number of identified variants in each variant class and their categorization based on the impact of consequences can be found in Appendix B. The following pie chart shows that most of the variants identified in this analysis are in the modifier category which are usually

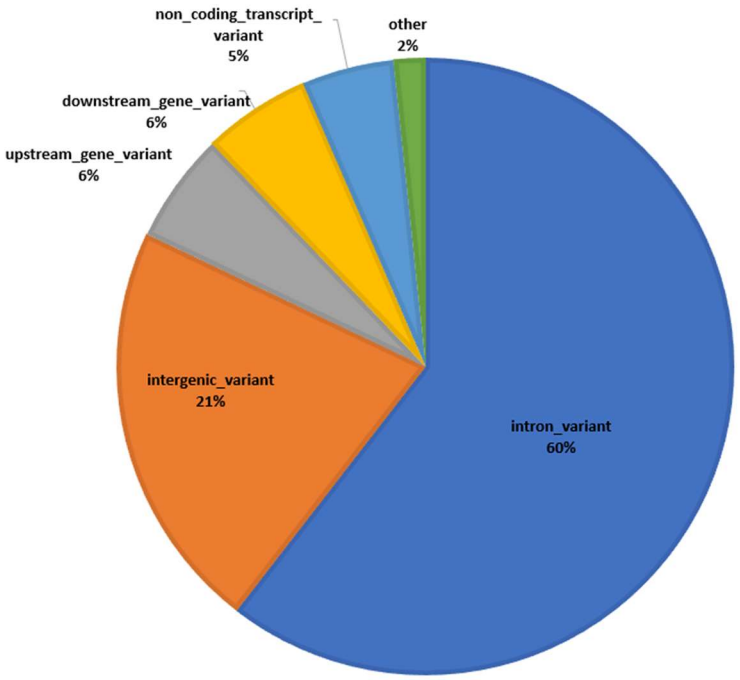


Figure 12: The percentage of variants with consequences in the entire genome

non-coding variants or varaints affecting non-coding genes where predictions are difficult (Figure 12).

Most were intronic (60%) or intergenic (21%) variants while some were found in 5' and 3' untranslated regions and in non-coding RNA genes. Compartively a small amount (0.03%) of variants were found in high impact areas of the gene structure (Appendix B, Table 2).

CHAPTER 4

INTROGRESSION DETECTION ALGORITHM

4.1 Overview

This chapter describes how the variants identified in the previous chapter were used for the introgression detection algorithm. The question we tried to answer was how to identify whether, a region has come from a different species based on genomic signatures. In Sarkar et al. [1], two haplotypes were identified in horses (caballine) for an ~900 base pair locus centered on exon 1 of *CXCL16*. These two haplotypes differed by 14 single nucleotide variants. The presence of only two, very diverged alleles, with no evidence of intermediate versions was difficult to explain. It was more interesting that the corresponding locus in non-caballines (zebras, donkeys, and asses) was identical to one of the two haplotypes. We developed a method to identify loci in horses whose haplotypes could be grouped into two phylogenetic groups, one shared with non-caballine, and the

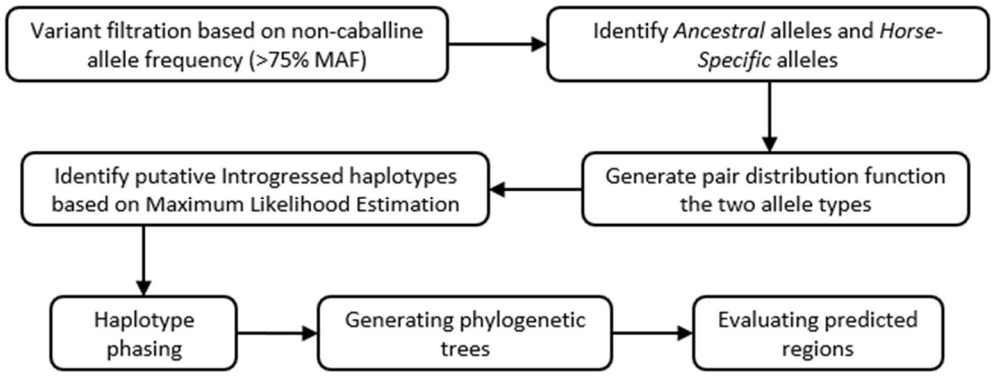


Figure 13: Main steps of the introgression detection workflow.

other distinct from it. The flow diagram for this novel workflow is shown in Figure 13 and will be described in this chapter.

4.2 Variant filtration

Based on the preliminary studies on *CXCL16* locus, we hypothesized that there are genomic segments in the horse with two distinct groups of haplotypes, one group unique to the horse, and the other with great phylogenetic similarity to its closely related non-caballines equid species. After mapping and cohort variant calling on the combined caballine and non-caballine genomic data, we started out to identify introgression at the haplotype level. We were looking for alleles that were fixed at high frequency in non-caballines and were different from the equine reference. The assumption we were making is that either the horse specific allele appeared and became fixed in the horse after the caballine and non-caballine species diverged, or the alleles that are ancestral to both caballines and non-caballines were polymorphic, and the horse was selected for one allele and the non-caballines were selected for the other allele. Alleles of this nature that introgressed back into the horse, will be either fixed, or in very high frequency in the non-caballines, and at varying frequencies in the horse.

Thus, for downstream analysis, variants with more than 75% minor allele frequency across non-caballine genomes were filtered out from the 73 million SNPs identified in the previous chapter. After the filtering step we were left with 22,353,415 SNPs and out of these 3,925,189 (17%) were consistent with the publicly available SNPs for the Equine genome at Ensembl as of March 2021. 166,823 (1%) were consistent with the

EquineAxiomHD_Chip genotyping array. The filtered variants were structurally annotated again using VEP tool as described in the previous chapter (Table 4).

Table 4: the class of filtered variants according to its component alleles and its mapping to the equine reference genome

Variant class	Count
SNV	20,312,859 (90.9%)
Insertion	801,711 (3.6%)
Deletion	649,292 (3%)
Sequence alteration	466,178 (2.1%)
Indel	123,375 (0.6%)

Most of the remaining variants were intronic and intergenic (non-coding) where it is difficult to predict and not enough evidence of impact (Figure 14). Appendix B (Table

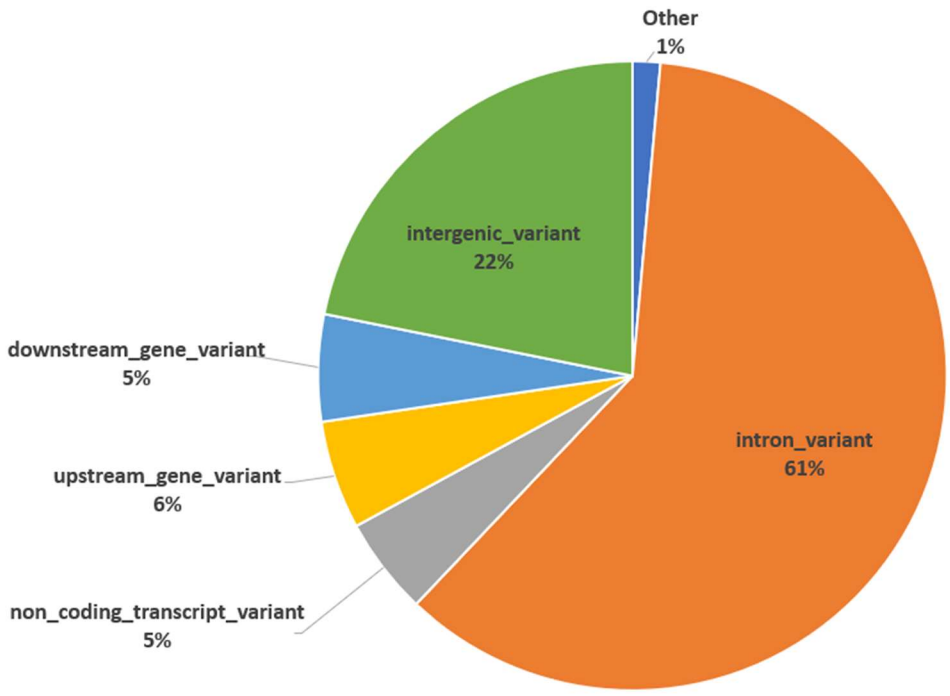


Figure 14: The percentage of filtered variants with consequences in the entire genome

3) has more information about the count of filtered variants responsible for different consequences on the protein sequences.

When analyzing the variants in coding sequences we could see most of them are synonymous variants (63%) (Figure 15). A variant synonymous where there is no resulting change to the encoded amino acid. These variants will not affect the protein function. Second most abundant are missense variants (35%), where the variant changes one or more bases resulting in a different amino acid sequence, but the protein length is preserved. In some situations, these variants will disrupt the protein function. The frame shift variants are caused by indels which results in disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three. Protein sequences

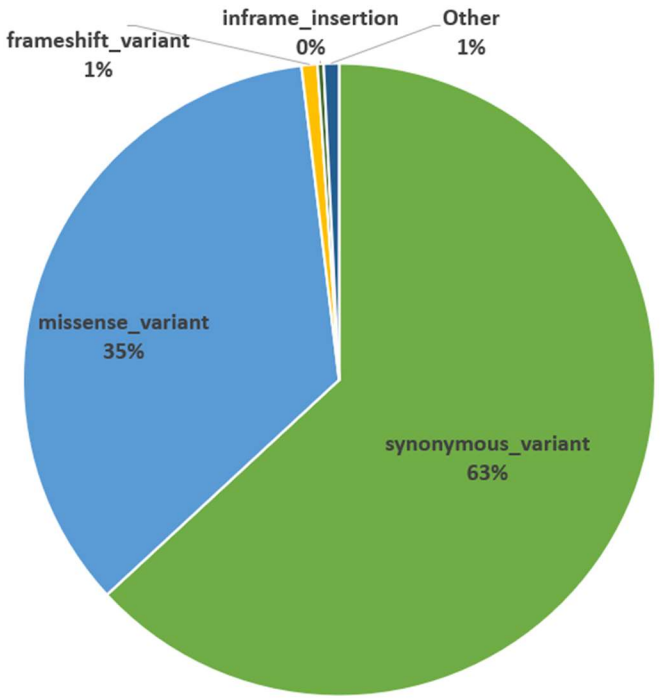


Figure 15: The percentage of filtered variants with consequences only on coding sequences

are expressed as triplets of nucleotides called codons. The information about rest of the coding region variants can be found in Appendix B (Table 4).

4.3 Ancestral alleles vs horse-specific alleles

First, we need to identify the concept of an ancestral allele to understand introgression detection. As discussed in chapter 2, the Most Recent Common Ancestor (MRCA) of these animals existed about 4 million years ago [73]. Somewhere in its genome existed a nucleotide *A* and it is the ancestral allele for this location in the genome (Figure 16). Approximately 4.6 MYA caballines split from non-caballines and 2 million years later assess split from zebras in the non-caballine clade. But all along the *A* allele from common ancestor was conserved and retained by all these animals. In this scenario, caballines have derived a new allele that became fixed in their genomes and that is labeled as *G* in figure 4. This is called a species-specific allele. Because if we see a *G* at this place of an equid genome, we can identify the animal as a horse. But if we see an ancestral allele *A* in a horse

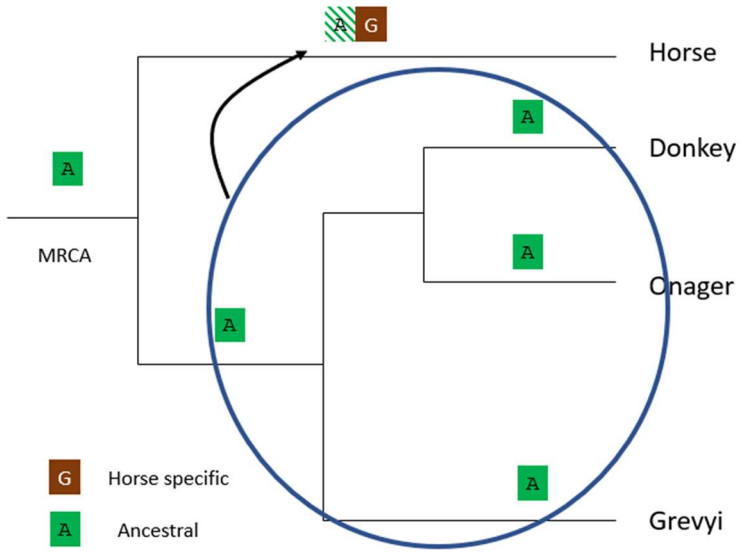


Figure 16: Inheritance of the ancestral allele among equids through evolution and by introgression.

there are two different ways that it could have got in there. One is through evolution, where the *G* allele never became fixed in the horse and the other is through introgression where *G* was fixed, and *A* was brought back into horse from a non-caballine cross.

In our analysis a variant in the multi vcf file was considered as a candidate for *Ancestra*l alleles if at least one horse out of the 14 horse samples had an alternative allele (0/1 or 1/1) at this locus matching the common non-caballine allele. This is a locus where ancestral alternative allele that evolved along the non-caballine lineage has entered and is present in the horse population as well as a horse derived new allele. On the other hand, a variant was considered *Horse-Specific* where only horses show homozygous reference (0/0) with the reference. This is where horse has evolved a horse specific allele that differs from all other equids. Any alternative alleles found in these variants could be potential candidates of introgression. The *Ancestral* alleles are a subset of the *Horse-Specific* alleles.

Ancestral alleles: 3,870,974

Horse-Specific: 22,353,415

Mapping the non-caballine genomes against the equine reference genome revealed a lot of *Horse-Specific* alleles where non-caballines are fixed for ancestral homozygous variants that do not agree with equine reference. These variants stretch in dense clusters along the non-caballine genomes. When observing their respective bam files through Integrative Genomics Viewer (IGV) it is evident that some of these dense non-caballine allele clusters are shared with some of the horse genomes along with *Horse-Specific* alleles. The region circled in figure 5 is an example of one such region found on chromosome 2, between chromosome positions 64,279,746 and 64,304,334 on EquCab3. It is plausible that these are introgressed regions that came into the horse genome from non-caballine

genomes. The gray colored regions on the IGV wiggle tracks represent the nucleotides that agrees with EquCab 3.0 reference. The solid-colored lines represent variants that are homozygously different from the reference and the two-colored lines represent heterozygous variants. After a speciation event, by definition, interbreeding becomes more difficult, in most cases first generation F1 crosses do not produce fertile males. In extant equids, both male and female F1s created between horses and non-caballine equids are unable to reproduce. As a result, their genomes will drift from one another without much gene flow. The longer the two animals are separated and are evolving to adapt to different environments, the more their genomes will diverge from each other and the more variants between the species will occur. The density of variants between any horse and the equine

Figure 17: Variant distribution along caballine (TB03, TB10, Saddlebred, FAANG TB) and non-caballine (Donkey, Grevyi, Onager) genomes. The putative introgressed region is circled in the figure. The variants are color coded based on the nucleotide.

reference genome will be much less than the density of variants between any zebra, donkey or ass and the equine reference. In Figure 17, it is clear that there is a higher density of variants in the non-caballines relative to the horses. There is, however one small region (circled) for TB10 where the SNP density is quite high, and upon further inspection, the one of the two haplotypes for that region was identical to the haplotype observed in the non-caballines. The method described here looks at the density of SNV alleles in a horse that are shared with non-caballines to designate a locus is putatively introgressed based on whether that region of the genome has a SNV density consistent with that of horses, or that of another species.

4.4 Pair distribution function

In probability and statistics, the nearest neighbor distribution function *H(r)* is one of the basic quantities used to model randomly positioned points in time or space. They are used to describe the probability of a second point existing within some distance *r* from the reference point. *H(r)* is successfully applied in various fields like biology, geology, and physics. In 1990 Torquato [94] introduced theoretical formalism to calculate *H(r)* for random distributions of finite-sized non-interacting particles in *D*-dimensional hard spheres with $D = 1, 2$ and 3. Here we are incorporating the *H(r)* determination for the case of hard rods ($D = 1$) as shown in (4.1) where σ is the diameter, η is the density of particles and x is the distance between particles. This results in an exponential decay based on the density of the particles in the space.

$$\sigma H(x) = \frac{2\eta}{1-\eta} \exp\left(\frac{-2\eta(x-1)}{1-\eta}\right) \qquad x > 1 \qquad (4.1)$$

For example, if there are eleven SNPs spread along the genome as depicted in Figure 18. There are ten distance measurements between each adjacent pair of SNPs. The distances between SNPs will be smaller in regions where there is greater SNP density. And the distances between SNPs will be larger in regions with lower SNP density. If this information was represented in a histogram of distances, the histogram for a denser system would have a higher peak at short distance, and decay more rapidly than that of a less dense system. Denser the SNPs are the more rapidly the graph will decay, suggesting SNPs will more likely find a closer neighbor.

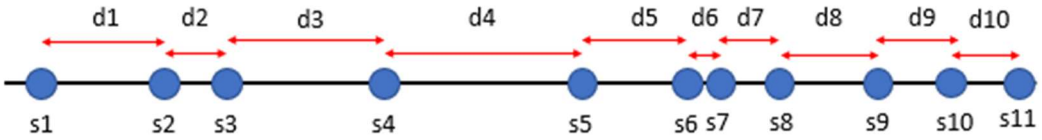


Figure 18: SNP distribution across the genome. The blue colored circles represent SNPs. The distances between SNPs are marked in red arrows.

The histograms shown in Figure 19 were derived empirically from the vcf file produced from the caballine and non-caballine animals (23) used in this study, and both curves, 1) the horse specific alleles, and 2) variant sites with alleles shared by horses and non-caballines show a near exponential decay. They do not fit the exponential decay exactly providing evidence that variant sites in a genome do not behave as non-interacting pairs. The resulting distribution functions plot is presented in Figure 19. The blue colored graph represents the *Horse-Specific* alleles, i.e., sites where an allele found in the horses differs from that present in all non-caballines which are denser across the genome and have a rapid decay while the *Ancestral* alleles, i.e., those that are found in all species and may

have been inherited from the most common recent ancestor are represented by the orange-colored graph which has a slower decay and a longer tail. *Ancestral* alleles are less frequent compared to *Horse-Specific* alleles in smaller distances and they are frequent in larger distances.

The distances between *Ancestral* and *Horse-Specific* SNPs were calculated up to 5,000 bp using a custom program. About 99% of the *Horse-Specific* alleles are found in a 500 bp window. Only two thirds of the *Ancestral* alleles are found in the same 500 bp window. The two graphs cross over each other at about 193 bp distance. Based on the graph if two SNPs are separated by 100 bp, it is more consistent with the *Horse-Specific* distribution and provides evidence the region may have been introgressd. If separated by 300bp it is more consistent with the *Ancestral* distribution and suggests both alleles came down the normal evolutionary path. During introgression these regions get recombined and inserted into the genome in different lengths of genomic segments [95]. These regions are

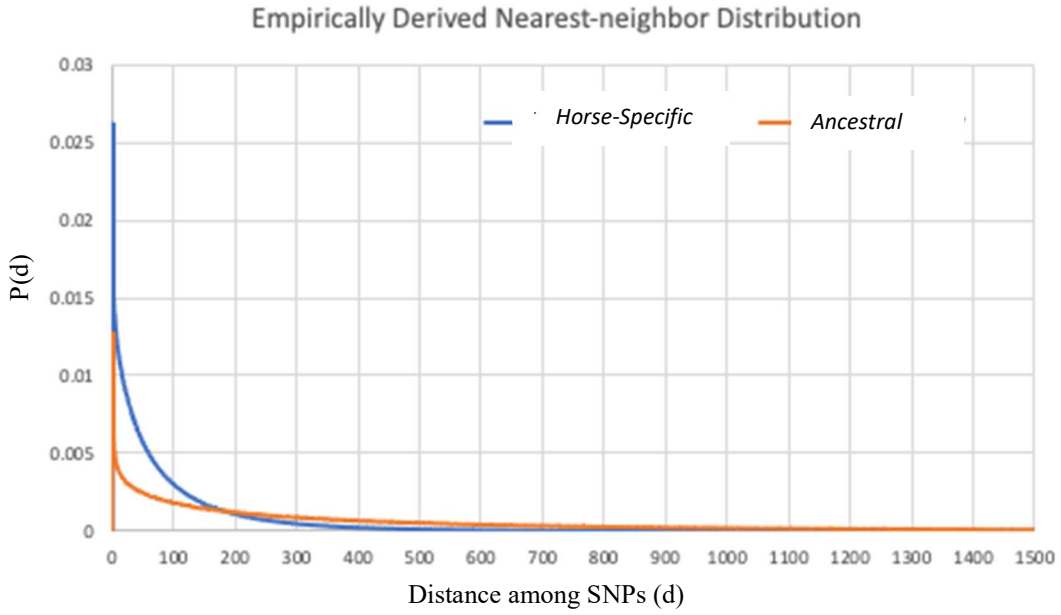


Figure 19: Nearest neighbor distribution function of *Horse-Specific* alleles and *Ancestral* alleles.

referred to as haplotypes. As the next step we scanned along the EquCab 3.0 genome for *Horse-Specific* alleles using a Maximum Likelihood Estimation (MLE) approach in an effort to identify these haplotypes.

4.5 Maximum Likelihood Estimation

The algorithm we derived for introgressed haplotype detection calculated a MLE for the potential haplotypes in a 20 SNP window. The goal of the MLE here was to find the SNP arrangements that maximizes the likelihood of being on an introgressed haplotype. A haplotype is a set of alleles that tend to be inherited together because of their proximity on the same chromosome [95]. Detection of haplotypes is the first step in a genetic study like this. The algorithm was executed along each horse genome in the analysis using a 20 SNP sliding window over the homozygous alternative alleles (1/1). The analysis was restricted to homozygous alleles such that we could be confident of the composition of the haplotype as no phasing would be required. The MLE fraction was calculated for each potential haplotype by taking the ratio of products of distance probabilities for each SNP

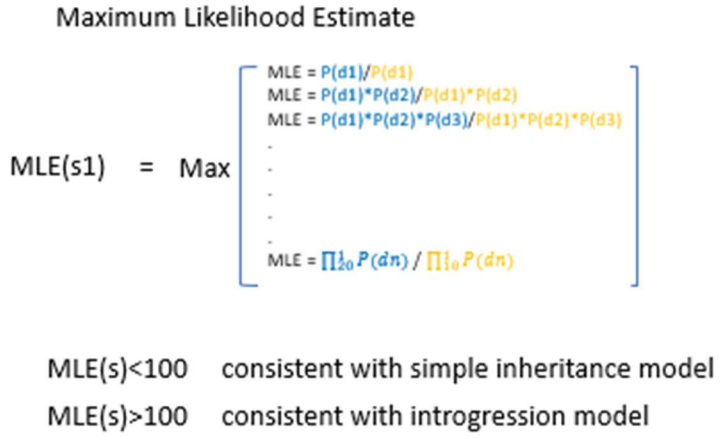


Figure 20: Maximum likelihood estimation of putative introgressed regions at maximum cutoff value of 100.

pair in the haplotype, from the *Horse-Specific* alleles nearest neighbor distribution and the *Ancestral* alleles nearest neighbor distribution (Figure 20). The maximum for each subset of contiguous SNPs was determined for every product ratio going all the way up to 20 SNPs. The region with the largest value for the ratio of the products for cumulative distances from d1 to d20, was the one used to evaluate for the introgression signal. Large MLE values indicated the SNP distances in the haplotype are consistent with the *Horse-Specific* distribution and lower MLE calculations indicated the SNPs more likely came down the *Ancestral* path. We have employed an MLE cutoff of 100 for each 20 bp haplotype window. This value was derived experimentally through trial-and-error to maximize the number of identified introgressed regions while minimizing the potential noise (Table 5). The identified haplotypes were used for further analysis, where horse derived a new allele and the non-caballines are fixed for an ancestral allele. The regions where SNPs are less dense were not good candidates of introgression.

Table 5: variation in number of regions identified with different MLE cutoff values

MLE threshold	**Number of identified putative regions**
100	15,583
300	11,188
1000	8,314
3000	6,485

Examples of MLE calculations for a positive result of introgression and a negative result can be found in Appendix C. Putative introgressed regions were identified along the fourteen horse genomes. The identified regions were merged and overlapping regions were coalesced to form a single list of regions without redundancies. There were 15,583 regions in the final list. The regions from both caballines and non-caballines were extracted from

the initial multi vcf file, (the one before filtering for the variants with more than 75% minor allele frequency across non-caballines) based on the genomic coordinates identified from the horse genome. As the next step in our method these genomic regions were phased to catalog the haplotypes present in the animals for the region.

4.6 Haplotype phasing

We are working with diploid organisms, and they inherit two copies of chromosomes, one copy from each parent. These copies are mostly identical to each other and only differ at a small fraction of nucleotide positions. Sequencing technologies obtain genotype information on variant sites which typically convolves the two haplotypes from both chromosomes. For a chromosome region with n bi-allelic variants, there would be 2^{n-1} number of possible haplotype patterns available for that region (Figure 21).

The objective of phasing is to recover the two haplotypes out of the group of possible haplotypes and assign them to the paternal and maternal chromosomes they came from. Many genetic analyses like association studies, detecting positive selection, estimating recombination rate, understanding gene function, and studying regions of the genome that are functionally related require phased haplotypes. It is a cumbersome task to unequivocally identify which two haplotypes are actually present in a heterozygous region out of the 2^{n-1} possible variant combinations.

Therefore, we have calculated MLE only for regions where a horse homozygously differs from the horse reference genome to generate the coalesced list of putative introgressed regions. In this way we unequivocally know the composition of the two haplotypes for these regions. An important consequence of this approach is that these haplotypes need to be in sufficiently high frequency in the population in order to be homozygous in at least one of the horses in the study.

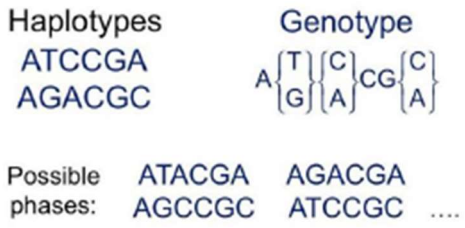


Figure 21: Haplotype phasing. Three bi-allelic variants can produce $2^{3-1} = 4$ possible haplotype patterns

Previously we were just using *Horse-Specific* alleles when looking to identify for introgression. Now for phasing, we have used all the variants present in these regions in every animal. The software program Beagle (version 5.0) was utilized for the phasing step [96]. It is the most computational efficient tool available for genotype calling, phasing, and genotype imputation. Once the phased haplotypes were generated, they were labeled based on their caballine or non-caballine origin. Next the unique sets were collapsed and used for phylogenetic analysis. Assuming we have two clades, the trees will be positive if a caballine haplotype is found in non-caballine clade and it will be false positive if no caballine haplotype is found in non-caballine clade.

4.7 Phylogenetic inference

Bayesian phylogenetic methods were first introduced in the 1990s [97] and soon became popular among researchers due to its computational efficiency in applying evolutionary models to genomic sequence data. Also, Bayesian analyses do not suffer the over-parameterization problems suffered by Maximum Likelihood methods. Other than the development of powerful models Bayesian methods became famous as user-friendly computer programs. Here we have incorporated MrBayes-3.2.6 a free software tool that performs Bayesian inference of phylogeny written by John P. Huelsenbeck and Frederik Ronquist in 2001 [98]. MrBayes uses the assigned substitution model and the tree topology together to specify the suitable statistical model for the sequence data in the analysis. The substitution models are used to estimate of evolutionary distances based on the number of substitutions that have occurred since a pair of sequences diverged from a common ancestor. We have used General Time Reversible substitution model with gamma-distributed rate variation (GTR+G) as all other models are nested within it [99].

Time reversible substitution models are very useful since they do not care which sequence is the ancestor and which is the descendant as long as all other parameters such as the number of substitutions per site that is expected between the two sequences are held constant. So, under the time reversible model, a phylogenetic tree can be rooted using any of the species and can be re-rooted based on the new information. Real world biological data sets like ours do not usually have access to sequence data of ancestral species. As there was no given special ancestral species all species would be derived from each other with the same probability. We have used the default number of 1000,000 cycles for the Markov

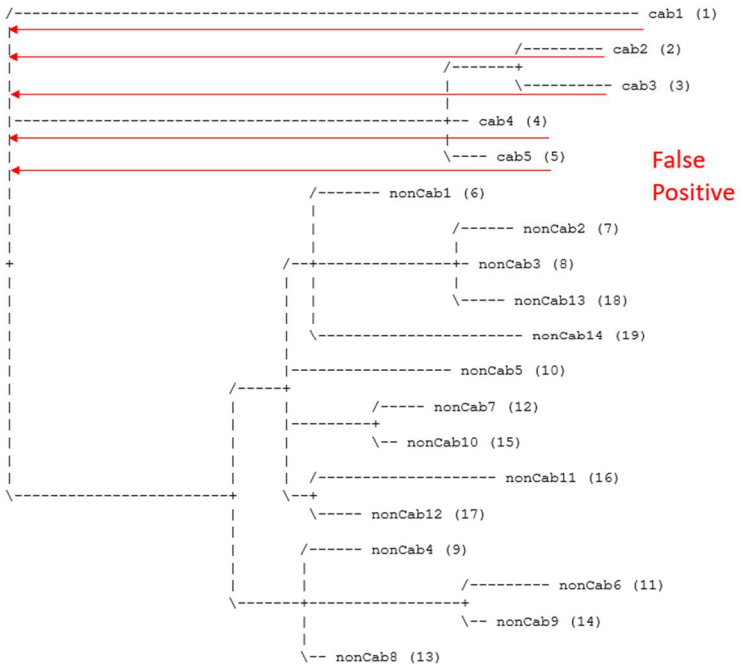
chain Monte Carlo (MCMC) algorithm and the Markov chain was set to sample at every 500 cycles.

4.8 Evaluating Predicted Regions

Some of these regions could be false positives. Since we extracted them based on the SNP density along the horse genome, they might not have the evolutionary relationship hypothesized for them to have. To solve this, we tried to identify false positive results, which did not have a signature of introgression, i.e., where the caballines did not cluster with the non-caballines. The generated trees were evaluated based on their branch lengths. The branches in a phylogenetic tree represent the transmission of genetic information from one generation to the next generation [100]. The length of a branch indicates the extent of genetic change based on the number of nucleotide substitutions in the sequence data. The longer the branch, the more extensive genetic change has occurred. We measured how far across the phylogenetic tree we need to traverse from a caballine to find a non-caballine. In a false positive tree, we had to go to top node to get into non-caballines clade (Figure 22a). That means the identified haplotypes were genetically distant from each other and were not candidates for introgression. Therefore, we ruled out these phylogenetic trees/regions from consideration. In a correct tree we needed to traverse a shorter distance from the beginning of a caballine node to where the caballine and non-caballine haplotypes clustered (Figure 22b).

The custom programs used to implement the introgression detection workflow are available at the following link. https://github.com/kalbflei/IntrogressionAnalysis

(a)



(b

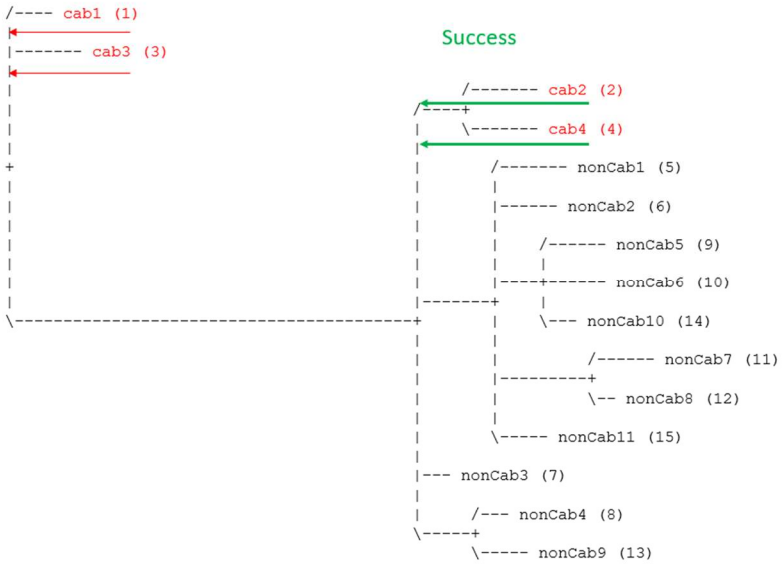


Figure 22: Evaluating the phylogenetic trees based on the traversal distance from a caballine node to where caballines and non-caballines cluster. (a) False positive tree (b) Successful tree.

4.9 Results and Discussion

4.9.1 Ancestral alleles vs species-specific alleles

The horse genome contains approximately 2.7 billion bases[6] and out of these 29,815,688 bases were identified as SNPs in the horse with alleles that are distinct from non-caballines, comprising about 1.1% of the horse genome. There were 23 million *Horse-Specific* alleles (Figure 23). These are the alleles that can be used to distinguish horse from the non-caballines. It is again about 1% of the genome. According to these results horses are 99% similar to all of the other animals in the study. Out of these 22 million alleles we found about 4 million ancestral SNPs which had one allele distinct to horses, and the second inherited either through evolution, or plausibly through introgression.

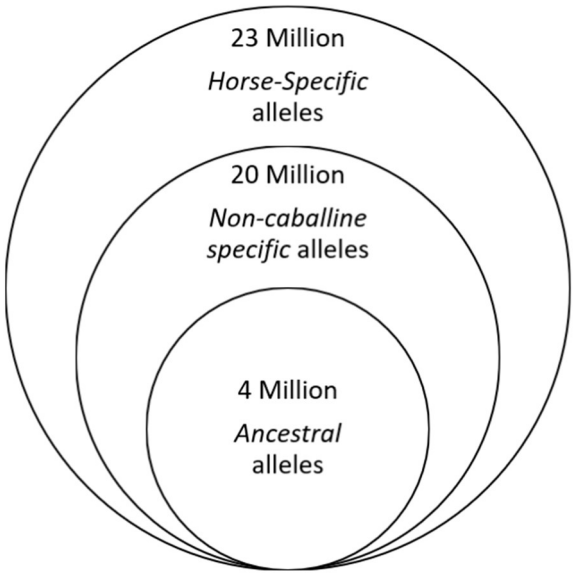


Figure 23: Allele distribution identified in horse genome.

[6] https://www.nih.gov/news-events/news-releases/horse-genome-assembled

4.9.2 MLE based introgression detection

We executed our bioinformatics workflow through all the caballine genomes incorporating the above-mentioned SNPs and identified putative introgressed regions along the horse genome. After running the algorithm over these animals, we saw Standardbreds, Haflinger and Arabians have higher number of putative introgressed regions than the thoroughbreds (Table 6). As we were looking for haplotypes, the animal had to be homozygous for a particular haplotype in order for us to consider it. Twilight could not homozygously differ from herself. That is why she has exactly zero signal across the regions. Probably that is also the reason for thoroughbreds to have a lower number of putative regions. This is evidence for existence of introgressed regions that Twilight does not have, and odds are other thoroughbreds share the same introgressed regions that are simply invisible because she has them. As a solution if we use a different reference genome from another horse, from another breed these numbers will mostly likely change.

Table 6: introgressed regions identified per horse genome

Animal ID	Breed	Putative introgressed regions
ST22	Standardbred	4164
3517	Saddlebred	4329
3519	Saddlebred	4243
SAMEA5721589	Haflinger	4781
AR03	Arabian	4111
AR04	Arabian	4090
AR05	Arabian	4381
683610	Thoroughbred	2769
686521	Thoroughbred	2849
H_2158	Thoroughbred	2491
H_3958	Thoroughbred	3127
TB03	Thoroughbred	2663
TB10	Thoroughbred	2641
Twilight	Thoroughbred	0

A genome wide signal for introgression was obtained once we coalesced the regions from all the caballines (Figure 24). The putative regions were spread across the entire genome, but there was an intense signal at chromosome 20. It was identified that this particular region in chromosome 20 corresponded to the MHC (Major Histone Compatibility) locus (Figure 24). Interestingly MHC genes code for cell surface proteins called MHC molecules that are essential for the functioning of the adaptive immune system [101]. The MHC molecules bind with and display the peptide fragments derived by the pathogens on the cell surface, the appropriate T cells can recognize them and react accordingly. As mentioned in chapter 2 if a new species comes in contact with a new environment and if this species is ill equipped to deal with the challenges of the new environment, crossing them with one of the already adapted species will bring in the alleles like MHC that are advantageous for the new species. We could identify a total 15,583 such putative regions with MLE greater than 100 and they were spanning across about 1.1% of the genome.

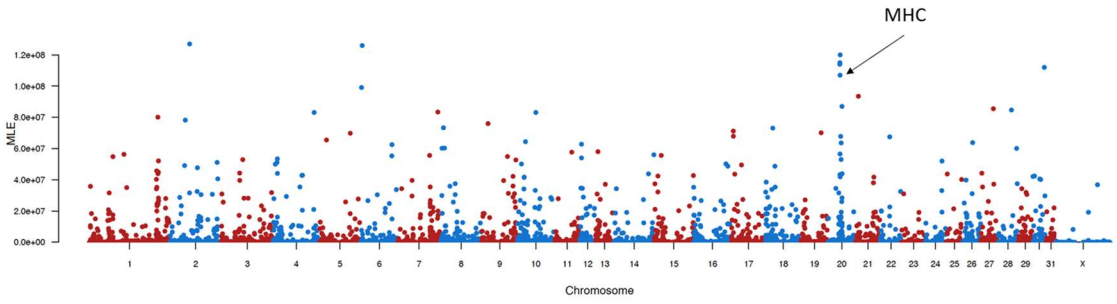


Figure 24: The putative introgressed region (MLE>100) distribution across the horse genome.

The length of a haplotype can give an idea about the time of introgression because each round of recombination may break down haplotype blocks into shorter fragments. Thus, short shared haplotypes tend to be older and are therefore more likely to be caused by ancient introgression events [102]. Meanwhile recent introgression can be identified by long shared haplotypes. In the current analysis the average length of identified introgressed haplotypes was nearly 2kb. Which is a much shorter nucleotide length compared to 57 kb introgressed haplotypes found in modern humans from Neanderthals which estimated to be introgressed about 47k–65k years ago [56]. This proves that caballine introgression was much older than the well-studied human introgression event.

4.9.3 Comparison to sequenced archaic genomes

The most ancient horse for which genetic information is available is a 1.12x draft genome sequence deciphered in 2013 from a fossilized bone fragment found near Thistle Creek, Canada [73]. This fossil belongs to an extinct prehistoric horse who lived more than 700,000 years ago in Middle Pleistocene era. This analysis of the Thistle Creek horse genome revealed that the origin of genus Equus was 4.0–4.5 MYA which was twice the time of origin that was accepted by the researchers at that time. We managed to execute our introgression detection workflow on the Thistle Creek horse genome and found ~25x fewer introgressed regions compared to rest of the horses. This could be due to the low coverage and other artifacts in the ancient genome. In spite of that we could still identify putatively introgressed regions in the Thistle creek horse, and interestingly that provides evidence that the introgression events have occurred at least that far back in time.

4.9.4 Evaluating Predicted Regions

In our attempt to identify false positives we traversed along all the generated phylogenetic trees measuring the branch lengths from a caballine to a non-caballine. Based on the results we generated the curve in Figure 25, where the x axis represents the number of caballine haplotypes that were found in the non-caballine clade and y axis represents the fraction of trees. About 25% of the identified regions did not have a caballine haplotype clustered with a non-caballine clade (marked with red arrow). There was about 25% false positive rate. This rate begins to increase if the MLE threshold was set below 100. The important regions were the ones where 1, 2 or 3 caballine haplotypes cluster with the non-caballine clade (marked with green arrows). The regions with higher numbers (more than three) of putative haplotypes that cluster with non-caballines are much more complicated, and difficult to understand or explain because there is lot more variation. Typically, the regions that get positively selected after introgression stay fixed and largely unchanged.

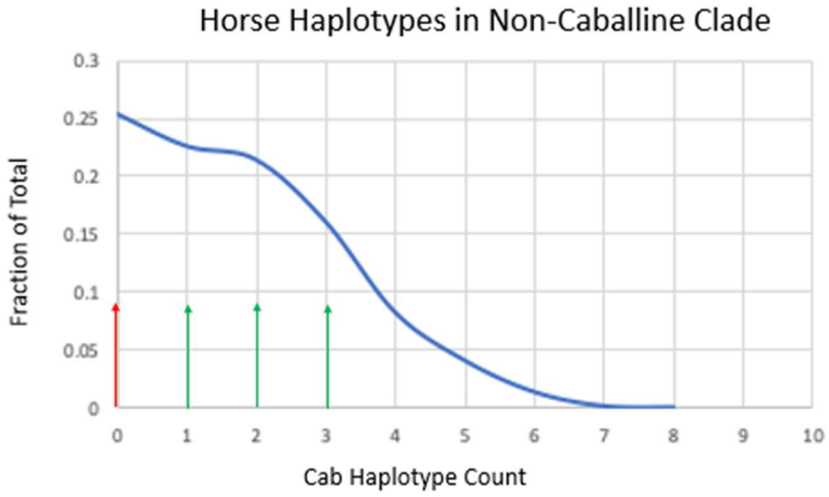


Figure 25: Evaluation of identified regions. The fraction of trees with no caballine haplotype clustered with a non-caballine clade (marked with red arrow). The trees with caballine haplotypes clustered with a non-caballine clade (marked with green arrows).

An arbitrary threshold of 0.5 was assigned to filter out the false positive results meaning that if the caballine haplotype was in a clade where it needed to go to less than half the origin before it was able to go down to the non-caballine clade, then it was a positive introgression event. We were left with 8,951 putative introgressed regions.

Usually when generating phylogenetic trees, it is expected to see horse haplotypes clustering away from non-caballine haplotypes. But often, in the regions identified using this method, there are one or more haplotypes found in the horse that group with non-caballine haplotypes owing to introgression. Due to the differences in allelic composition, these haplotypes appear to have diverged because of speciation rather than drift within species, and as such suggest an inter-species transfer event. This is the signature we have used in our phylogenetic analysis of equids to identify introgressed regions. After the evaluation of the identified putative introgressed regions the next step was to identify their biological significance.

CHAPTER 5

FUNCTIONAL ANNOTATION OF PUTATIVE INTROGRESSED REGIONS

5.1 Overview

The putative introgressed regions identified in the previous chapter were annotated to have an idea about their biological importance. This chapter will provide details on the annotation process and the findings.

5.2 Annotation work done on domesticated animal genomes

The studies conducted on domesticated animal genomes have led researchers to a better understanding of their genetic selection, adaptation, and evolution. A few examples of economically important domesticated animals who are subjected to research over the years other than horse are chicken, pig, cattle, and sheep. Although reference genomes for these animals have been available for almost a decade, the annotation of these genomes remain an ongoing process. Chicken (*Gallus gallus*) was the first agriculturally important animal to have its genome published in 2004 [103]. The chicken genome is approximately one third of the human genome at about one billion base pairs [104]. The alignment of the chicken and human genomes revealed there are protein coding segments in long blocks of conserved synteny among two genomes, which allowed researchers to better understand their common ancestor. Next there was a detailed analysis done on the pig transcriptome with the intention of providing better knowledge about uncharacterized genes in pigs [105]. They clustered the gene transcripts from several tissues based on their expression level. This is a great step towards recapitulating known expression patterns and recognizing

functional relationships between genes as well as inferring the function of new genes. In another study, a group of researchers who were working on the sheep genome identified higher expression for genes involved in keratin cross-linking and lipid metabolism [106]. This discovery revealed an interaction between lipid metabolism and wool synthesis in sheep. These are a few examples of how genome wide analysis of these animals and their ancestors have revealed relationships between sequence and function. But in comparison to human, mouse, and other model organisms their transcriptomes are still not adequately characterized. Although a fair number of coding regions were identified, there is still little to no information on noncoding regions and regulatory sequences. Much of the identified genetic variations underlying expression patterns in the above mentioned studies are likely to be a result of epigenetic mechanisms in regulatory sequences [107].

Due to this complex nature of gene expression, it is impossible to predict phenotypes based on just genotype without the understanding of the underlying biological mechanisms. In an effort to expand the understanding of these mechanisms, internationally coordinated Functional Annotation of Animal Genomes (FAANG) project was established [108]. The aim of the project is to annotate the major functional elements in the genomes of domesticated animal species based on common standardized procedures and pipelines. The project was mainly influenced by the Encyclopedia of DNA Elements (ENCODE) Consortium which was established in 2003 with the aim of identifying functional elements in the human genome [109]. Later the project was expanded from the human genome to classic model species like mouse, *Drosophila*, *Caenorhabditis elegans* and zebrafish as the researchers saw the potential improvement of knowledge and reliability of the cross-species genome analysis. Similarly, the FAANG project is also an international

collaboration among domestic animal research groups. It is designed as a cost-effective approach that will produce faster results with minimum data redundancy. The FAANG pilot projects were aimed at identifying regulatory elements in sheep, chicken, pig, cattle and goat genomes [110]. They have produced genome-wide data sets on RNA expression, DNA methylation, chromatin modification and chromatin accessibility for the above animal genomes. With the development of a new reference genome, the equine research community has also joined this effort and is currently making progress towards annotating the horse genome [111]. As an initial step to facilitate equine researchers, the horse FAANG effort has created a biobank consisting of 80 tissue samples from two cell lines and six body fluids [112]. The tissue samples were obtained from two adult thoroughbred mares (SAMEA104728862/683610, SAMEA104728877/686521) whom we have also incorporated in introgression detection workflow. It is reported that this is the first non-human biobank with extensive phenotypic data and also the first published equine specific biobank.

The identified putatively introgressed regions have not only been retained but have been identified in all the horses we have analyzed. This gave us enough reason to believe they have an important role related to gene regulation. So, we compared the putative introgressed regions against Chromatin Immunoprecipitation Sequencing (ChIP-Seq) data obtained for histone markers in 8 equine tissues as a part of FAANG project. Interestingly almost 44% of our regions overlapped with the regulatory regions identified by histone markers. This provide evidence that there is a connection between the identified regions in equine genome and the surrounding genomic regulation. These relationships can be very complex as they are important for the genome. Furthermore, we conducted a homology

search and Gene Ontology (GO) annotation in order to have a better understanding about these regions.

5.3 Understanding the putative regions based on structural annotation of equine genome

Genomic annotation includes the identification of elements like genes, introns, and exons within the genomic sequence. This is called structural annotation. Attaching the functional regions relevant for these same elements such as promoters and regulatory regions is called functional annotation. The equine genome was structurally annotated by Ensembl gene annotation pipeline after the EquCab 3.0 genome was submitted in 2018. The structural gene annotation was carried out using a combination of homology mapping, RNA-seq alignments and mapping annotation from a suitable reference species. They have been able to identify 20,955 coding genes and 9,014 non-coding genes in the equine genome[7]. We compared the putative introgressed regions against the available structural annotation of equine genome. The results are summarized in the Table 7.

Table 7: Annotation of the putative introgressed regions based on the structural annotation of the equine genome

Structural element	Number of overlapped introgressed regions	Number of overlapped filtered introgressed regions
Coding dna sequences (CDS)	935	577
Exon	345	206
MRNA	4,127	2,328
Gene	9	5
UTR	71	42
lnc_RNA	887	521
Total	6,374	3,679

[7] https://uswest.ensembl.org/Equus_caballus/Info/Annotation

Functional annotation is not generally done automatically as part of the genome annotation process. Typically, functional annotation is done as a separate, focused effort.

5.4 Understanding the putative regions based on histone modification sites

DNA is a complex double stranded molecule. Each strand is composed of a chain of nucleotides connected to each other through a sugar phosphate backbone. If the DNA from all the chromosomes in a human were uncoiled and placed end to end the resulting strand would be 67 million miles long[8]. This is where the histone proteins come into play.

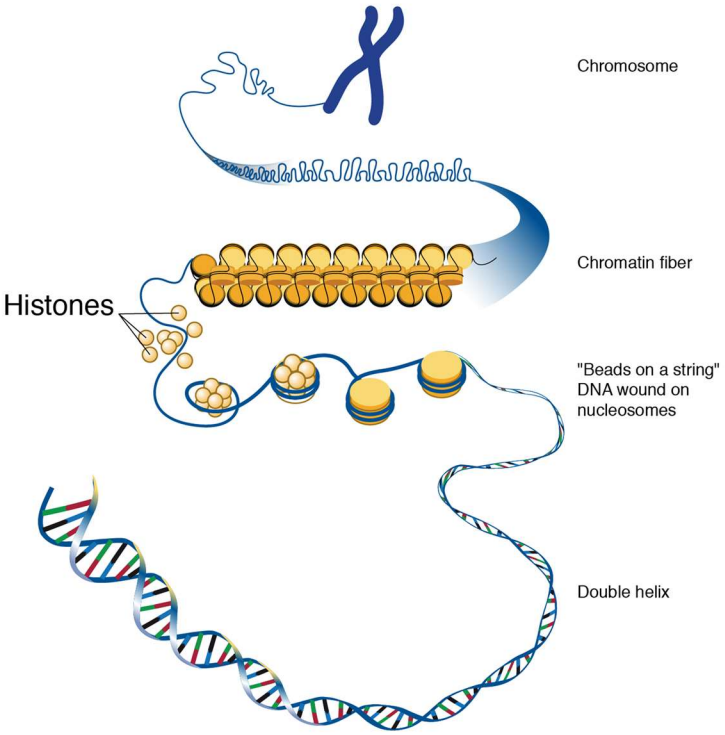


Figure 26. Histone is a protein that provides structural support to a chromosome. Some modifications in histones are associated with regulation of gene expression[9]

[8] https://www.nigms.nih.gov/education/Inside-Life-Science/Pages/genetics-by-the-numbers.aspx
[9] https://www.genome.gov/genetics-glossary/Chromatin

Histones are critical in the packing of DNA into chromatin and next into chromosomes (Figure 26).

DNA strands wrap around eight histones, forming a structure called a nucleosome. Nucleosomes are made with two subunits of core histones called H2A, H2B, H3, H4, and a linker histone called H1 acting as a stabilizer. Modifications in histone proteins can cause changes in the chromatin structure making DNA accessible for transcription or making DNA more compact and inaccessible for transcription [113].

Most common histone modifications are methylation, acetylation, and phosphorylation. During methylation certain amino acids get modified by the addition of methyl groups. Only the lysine (K) and arginine (R) residues in histones can get methylated, most commonly the lysine residues in H3 and H4 histone tails get methylated [114]. Usually, DNA methylation inhibits gene expression by reducing the accessibility of proteins into regulatory regions. The other important modification, histone acetylation is the addition of an acetyl group (COCH3) from acetyl coenzyme A. Again, the Lysine residues in histone H3 and H4 are the preferred candidates of acetylation [115]. Histone acetylation is involved in the regulation of many important cellular processes, but the most common function of acetylation is the modifications in chromatin structure. This opens chromatin structure and makes it accessible to transcription factors increasing gene expression significantly.

Since 1964 researchers have hypothesized these histone tail modifications have a role in genomic regulations [116]. Another idea is that these modifications are the by-products of regulatory activities and are signatures left to create cellular memory of frequent transcriptional activity [117]. However, many studies have been conducted over

a decade providing evidence that these markers are strongly associated with regulatory elements. An assay comprised of chromatin immuno-precipitation followed by next generation sequencing (ChIP-Seq assay) has been the primary method used to examine histone modifications in different regulatory regions. In 2019 Kingsley and the Equine FAANG group performed ChIP-Seq on eight equine tissues namely adipose, brain (parietal cortex), heart, lamina, liver, lung, (skeletal) muscle, and ovary as an attempt to fill the gap in the current equine genome annotation on functions of non-coding regions [118]. The FAANG consortium has selected the following four histone modification marks in these tissues which were found to be the most informative by the ENCODE projects. These modifications cause transcription repression or activation depending on their target site.

- Histone H3 lysine 4 trimethylation (H3K4me3) - Found at promoters of active genes and transcription start sites.
- Histone H3 lysine 27 trimethylation (H3K27me3) - Found at transcriptionally silenced genes. A temporary repressive signal that controls development regulators and found
- Histone H3 lysine 27 acetylation (H3K27ac) - Found at transcription start sites and marks regulatory elements like active enhancers and promoters.
- H3 lysine 4 monomethylation (H3K4me1) - Found at downstream of transcription start sites and marks regulatory elements associated with enhancers and other distal elements.

The raw and processed data are available on the FAANG website[10] under the accession PRJEB35307. For this study we have analyzed the combined peak-calls that represent the

[10] https://data.faang.org/home

enriched peaks shared between the biological replicates (SAMEA104728862/683610, SAMEA104728877/686521) for each tissue sample. The following table shows the number of regulatory regions identified for each tissue at the respective histone modification site and out of these regions how many regulatory regions overlapped with the putative introgressed regions (Table 8). Forty four percent (3,964) of the filtered putative introgressed regions overlapped with regulatory regions based on the histone modification data.

Table 8: Regulatory regions at histone modification sites overlap with putative introgressed regions at respective tissues

	Tissue	Number of combined peaks	Number of combined peaks overlapped with putative introgressed regions	% Of peaks	Number of introgressed regions	% Of regions
H3K4ME1	Adipose	107,318	1,058	0.99	889	5.70
	Brain	95,918	875	0.91	744	4.77
	Heart	121,663	1,188	0.98	972	6.24
	Lamina	114,708	1,239	1.08	990	6.35
	Liver	116,760	1,040	0.89	790	5.07
	Lung	92,972	703	0.76	579	3.72
	Muscle	95,816	801	0.84	640	4.11
	Ovary	102,986	1,047	1.02	911	5.85
H3K4ME3	Adipose	26,905	331	1.23	308	1.98
	Brain	27,101	346	1.28	322	2.07
	Heart	26,475	313	1.18	300	1.93
	Lamina	29,380	355	1.21	330	2.12
	Liver	28,498	328	1.15	309	1.98
	Lung	28,546	343	1.20	332	2.13
	Muscle	28,110	335	1.19	311	1.99
	Ovary	28,378	373	1.31	358	2.30
H3K27AC	Adipose	79,620	777	0.98	615	3.95
	Brain	78,823	885	1.12	711	4.56
	Heart	68,728	724	1.05	569	3.65
	Lamina	82,394	829	1.01	652	4.18
	Liver	87,589	849	0.97	626	4.02

	Lung	69,054	651	0.94	519	3.33
	Muscle	76,495	751	0.98	567	3.64
	Ovary	64,318	723	1.12	608	3.90
H3K27ME3	Adipose	25,183	572	2.27	412	2.64
	Brain	24,243	431	1.78	323	2.07
	Heart	68,113	1,794	2.63	1328	8.52
	Lamina	37,366	968	2.59	726	4.66
	Liver	63,874	1,419	2.22	1146	7.35
	Lung	30,191	547	1.81	445	2.86
	Muscle	42,610	1,232	2.89	877	5.63
	Ovary	40,825	1,043	2.55	765	4.91

5.5 Understanding the putative regions based on functional annotation

Functional annotation is the association of biological information with regions identified by structural annotation. Traditional functional annotation approaches were mainly focused on protein-coding genes. But there has been a new flow of studies on different functions of noncoding genes and untranslated transcripts, long non-coding RNAs, microRNAs, and the like [119]. Thus, the demand increases for well annotated genomes of non-model organisms. Initially the assignment of functional information to gene products was done manually. This provided accurate annotation but was a complex, laborious, and time-consuming task.

Today automatic functional annotation alternatives have been invented to keep up with the continued exponential rate of sequence data generation. Local alignment tools like, Basic Local Alignment Search Tool (BLAST) can be used to automate the predictions by comparing reads against protein databases [120]. The functions will be assigned to the query sequences based on the resulting sequence hits produced by high-scoring alignments. These tools were built on the assumption that the similar sequences were evolved from a single ancestor and therefore retain similar functions. BLAST identifies these evolutionary relationships by comparing orthologous and paralogous sequences from closely related

species. Local alignment tools are easy to use and provide reliable output in most situations. Sometimes these results can be comparatively highly error prone and are of low sensitivity or specificity due to errors in public databases [121].

It is a good idea to combine these local BLAST results with Gene Ontology (GO) workflow which is currently the most widely used most comprehensive functional annotation schema [122]. Three aspects of a gene's function are covered by GO, namely Biological Process (BP) which is a biological program in a cell which uses the function of a gene, Cellular Component (CC) which is the location of the gene product in a cell, and Molecular Function (MF) which is the activity of a gene product at the molecular level. The GO Consortium developed and maintains the ontology standards to one which is consistently described to allow a comprehensive coverage of biological concepts of a particular gene product across species [123]. The standard GO annotation can provide several GO-terms for the same sequence based on the known or predicted function of its protein product. GO-terms are organized in a hierarchical manner and a particular term may have more than one parent terms. Today Gene Ontology is a vocabulary of about 50,000 terms that describe cellular components, molecular functions, and biological processes.

We have used OmicsBox (v1.4.12) to annotate the identified putative introgressed regions. OmicsBox/Blast2GO provides an interface to create, edit and run workflows based on the Common Workflow Language (CWL) specification [124]. This interface allows one to describe all analysis steps using the functions and tools offered by Blast2GO and connect them in a workflow to perform a complete analysis in a single run. OmicsBox/Blast2GO has been successfully used for annotation of non-model organisms [125]. OmicsBox has

access to additional functional databases such as Kyoto Encyclopedia of Gene and Genomes (KEGG) pathway mapping engine, plus a KEGG pathway visualization module [126]. It allows a user to have higher-order functional information, which are stored in the GENES and PATHWAY databases, respectively. Users can quickly obtain enzyme codes for their datasets and the metabolic pathways linked to these enzyme codes.

5.5.1 Start GO workflow using NCBI Blast

In order to assign function to identified regions, GO workflow was executed as described below. The genomic regions corresponding to the coordinates of the identified putative introgressed regions were extracted from EquCab 3.0 genome with the help of samtools. The faidx tool in samtools extract the subsequences from indexed reference fasta file.

samtools faidx [path to Ec_build-3.0.fa file] [genomic coordinates] >> [path to IntrogressedSequences.fasta file] (5.1)

After extracting the fasta sequences, the NCBI nonredundant database was downloaded using wget command from ftp://ftp.ncbi.nlm.nih.gov/blast/db on 2-9-2021 to generate a non-redundant equine protein database. NCBI has compiled the nr database as a protein database for Blast searches. It contains non-identical sequences from GenBank CDS translations, PDB, Swiss-Prot, PIR, and PRF. The nr database is very useful since it is frequently updated and comprehensive. The downside is that this is a huge database. The downloaded NCBI nr database contained about 158 Gb of sequence data. Next the GenIdentifier (GI) accessions for family *Equidae* (taxonomy ID 9788) proteins were downloaded from the NCBI protein database at http://www.ncbi.nlm.nih.gov/protein. The query ((all [filter])) AND "equidae"[porgn: __txid9788] was used to search for the protein

IDs for family *Equidae*, which was 180,416 sequences. The sequence IDs were downloaded and sorted into sequence.gi.txt file.

The Basic Local Alignment Search Tool (BLAST) is used to find regions of local similarity between sequences [120]. This program can compare nucleotide sequences or protein sequences to available sequence databases to find statistically significance matches. Here BLAST was used to infer functional relationships between sequences and to sequence gene families. So as the next step, the NCBI blast executable (2.11.0) was downloaded and installed from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.11.0+-x64-linux.tar.gz. Then the downloaded NCBI nr database was filtered for only the family *Equidae* proteins from the sequence.gi.txt and output file Equidaefromnrdb.faa was created.

blastdbcmd -db nr -entry_batch [path to sequence.gi file] -dbtype prot -out [path to Equidaefromnrdb.faa file] -logfile [path to err file] (5.2)

A total of 156,875 sequences were found in the Equidaefromnrdb.faa file. There was a redundancy of sequences with similar headers. Therefore, the sequences with redundant headers were removed using a custom script. The remaining FilteredEquidaefromnrdb.faa file with 120,381 sequence records were used to construct a blastable database called Equidae_db from family *Equidae* protein sequences.

makeblastdb -in [path to FilteredEquidaefromnrdb.faa file] -input_type fasta -dbtype prot -out [path to Equidae_db] (5.3)

The fasta sequences extracted from EquCab 3.0 genome in the first step were compared against the Equidae_db database using following blastx command. The blastx

tool can search protein databases using translated nucleotide queries. The results of the blastx command were stored in an XML file.

blastx -db [path to Equidae_db] -query [path to IntrogressedSequences.fasta -evalue 1e-05 - outfmt 16 -max_target_seqs 20 -out [path to IntrogressedSequences.xml] (5.4)

5.5.1.1 Continue GO workflow with OmicsBox

The above blast search can be also done with CloudBlast search provided by OmicsBox tools. Rest of the steps in the functional analysis process was performed in the OmicsBox. As mentioned below there are few key steps in the OmicsBox functional analysis module (Figure27).

- High-Throughput Blast and InterProScan
- Gene Ontology Mapping
- Blast2GO Annotation
- Functional Interpretation

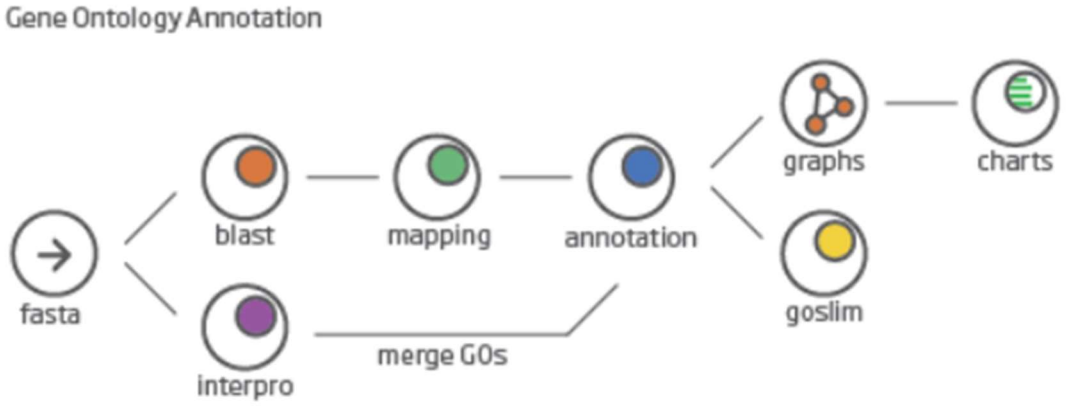


Figure 27: OmicsBox Blast2GO module example workflow for GO annotation of a set of sequences[11]

[11] https://www.biobam.com/omicsbox/

5.5.1.2 InterProScan

InterProScan provides functional analysis of proteins by classifying them into protein families, protein domains and important sites. In order to do this InterProScan scans a collection of important sites and databases enumerated below.

- Conserved Domain Database (CDD), a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins.
- High-quality Automated and Manual Annotation of Proteins (HAMAP), a collection of manually curated family profiles for protein classification.
- Protein Analysis Through Evolutionary Relationships (HMMPanther), a large curated biological database of gene/protein families and their functionally related subfamilies.
- Protein Information Resource (HMMPIR) scans the HMMs that are present in the PIR Protein Sequence Database (PSD) of functionally annotated protein sequences.
- FPrintScan scans against the fingerprints in the PRINTS database. These fingerprints are groups of motifs that together are more potent than single motifs by making use of the biological context inherent in a multiple motif method.
- ProfileScan scans against PROSITE profiles to find structural and sequence motifs in protein sequences.
- Simple Modular Architecture Research Tool (HMMSmart) scans the HMMs that are present in the SMART domain/domain families database.
- The Institute for Genomic Research (HMMTigr) scans the HMMs that are present in the TIGRFAMs protein families database.

- PatternScan, a new version of the PROSITE pattern search software which uses new code developed by the PROSITE team.
- Gene3D a database of globular domain annotations for millions of available protein sequences.
- Structure-Function Linkage (DatabaseSFLD) a database links evolutionarily related sequences and structures from diverse superfamilies of enzymes to their corresponding chemical function.
- SuperFamily a library of hidden Markov models that represent all proteins of known structure.
- Coils a program that compares a sequence to a database of known parallel two-stranded coiled-coils and derives a similarity score. By comparing this score to the distribution of scores in globular and coiled-coil proteins, the program then calculates the probability that the sequence will adopt a coiled-coil conformation.
- Phobius a program for prediction of transmembrane topology and signal peptides from the amino acid sequence of a protein.
- SignalPHMM predicts the presence of signal peptides and the location of their cleavage sites in proteins from Archaea, Gram-positive Bacteria, Gram-negative Bacteria and Eukarya.
- Transmembrane Helices (TMHMM) a membrane protein topology prediction method based on a hidden Markov model. It predicts transmembrane helices and discriminate between soluble and membrane proteins with high degree of accuracy.

InterProScan will automatically translate nucleotides into amino acids. The results of this step (InterProScan IDs and GO-terms) will be merged with the blast mapping results during the annotation step.

5.5.1.3 Gene Ontology mapping

The GO Mapping step retrieved the GO-terms associated with the hits obtained by the previous blast search. Based on the blast result accessions it retrieved gene names and ID s from GO annotation database. The GO-terms retrieved from this step were merged with the GO-terms obtained from the InterProScan step before annotating a sequence.

5.5.1.4 Blast2GO annotation

This step evaluated the information obtained during blast mapping and InterProScan steps and assigned the most reliable GO-terms to the input sequences. The annotation was carried out by applying the Annotation Rule (AR) to all the identified GO-terms. This rule seeks to find the most specific annotations with a certain level of reliability. The annotation rule is described in detail in their 2008 publication [124]. To apply this rule, an Annotation Score (AS) was calculated for each GO-term based on multiple parameters such as sequence similarity, BLAST highest scoring pair (HSP) length, e-values, the GO hierarchical structure and GO-term Evidence Codes (EC). The rule selected only the most specific GO-term per branch that lies above a user-defined cut-off value (threshold). Once the AS was calculated for each GO, the AR selected the lowest term per branch that lies over a certain threshold (default=55). Default values for the EC and threshold were chosen to provide a good balance between quantity and quality of annotation.

5.6 Results and Discussion

5.6.1 Comparison against histone modification sites

All the tissue specific ChIP-Seq data sets had overlapping regions with the coordinates of the introgressed regions in our comparison of putative introgressed regions with histone modification sites. The following figure shows a visual representation of the comparison of overlapping introgressed regions with histone modification sites across the tissues (Figure 28). According to the results in Table 8, between 5 and 6% of the putative introgressed regions overlapped with histone modification sites associated with enhancers (H3K4me1) in every tissue type in the analysis with the highest overlap being with the tissues from lamina and heart. About 2% of the putative introgressed regions overlapped with H3K4me3 sites with the highest overlap with the ovaries, and 3-4% of the putative introgressed regions overlapped with H3K27ac sites, the highest being the brain. Both

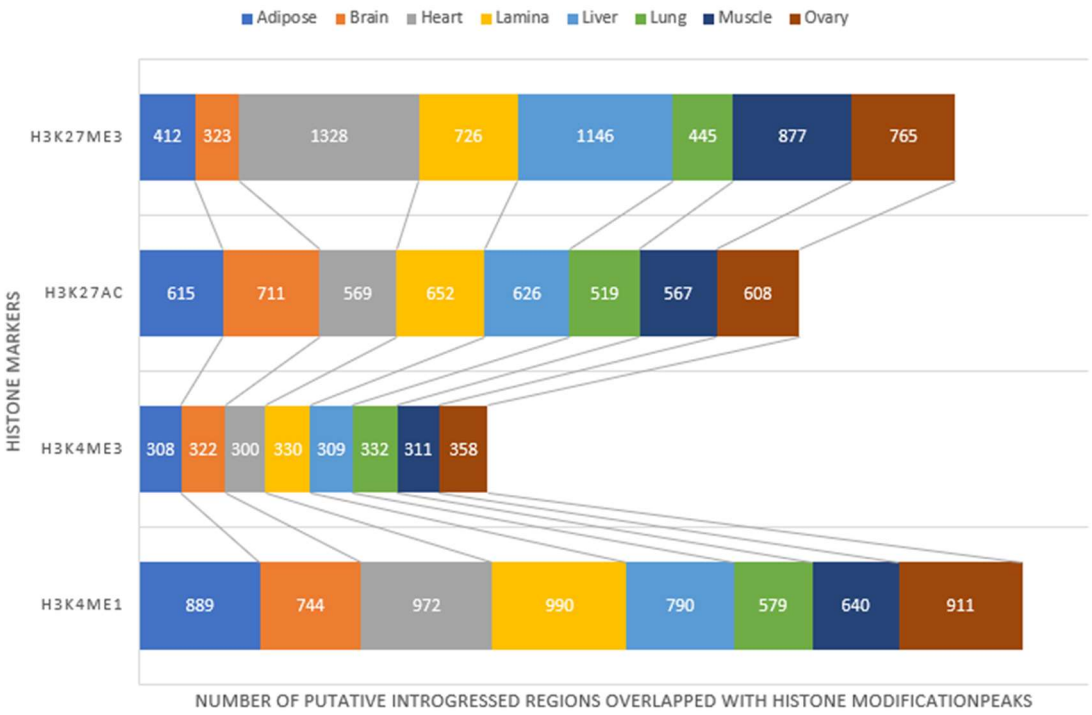


Figure 28: Putative introgressed regions overlap with tissue specific histone modification data

these histone modifications are associated with transcription start sites and promoters of active genes. Two to nine percent of these regions overlapped with gene silencing histone modification sites (H3K27me3) highest overlap been with heart tissue. A total of 3,964 (44%) of the regions identified in the equine genome were associated with the regulatory regions.

5.6.2 Comparison against structural annotation of equine genome

About 41% (3,679) of the identified putative introgressed regions were structurally annotated. Based on the EquCab 3.0 annotation, 577 putative regions were CDS, 206 were from exonic regions, 5 were from genic regions, 42 were from UTR regions and 521 were from non-coding regions. The following Venn diagram shows the comparison between

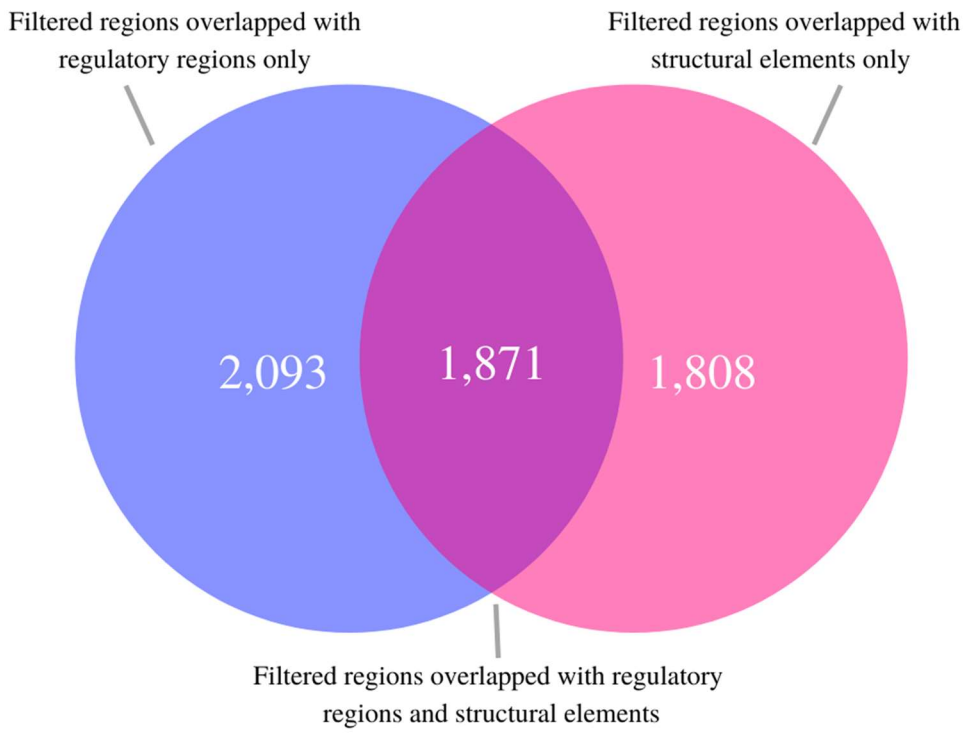


Figure 29: Out of the total 8,951 of putative introgressed regions, 64% had either a structural annotation, a regulatory region annotation or both attached to it.

structural element annotations and regulatory region annotations for the filtered regions (Figure 29). Based on the above comparisons, interestingly a 64 % (5,772) of the identified putative introgressed regions were associated with either a structural element, a regulatory function or both.

5.6.3 GO annotation results

5.6.3.1 NCBI blastx search results

A blastx search provides information about the best hit protein sequence found for the provided nucleotide query sequence based on translated sequence similarity. The blast best hit result has the lowest e-value and the highest bit-score [127]. The e-value is a statistically significant threshold for reporting matches against database sequences. It indicates the likeliness that sequence similarity is not by random chance. The smaller the e-value better the match. The matches will be less stringent at higher e-values. Most of the best hits identified by the blast search had perfectly low e-values (Figure 30).

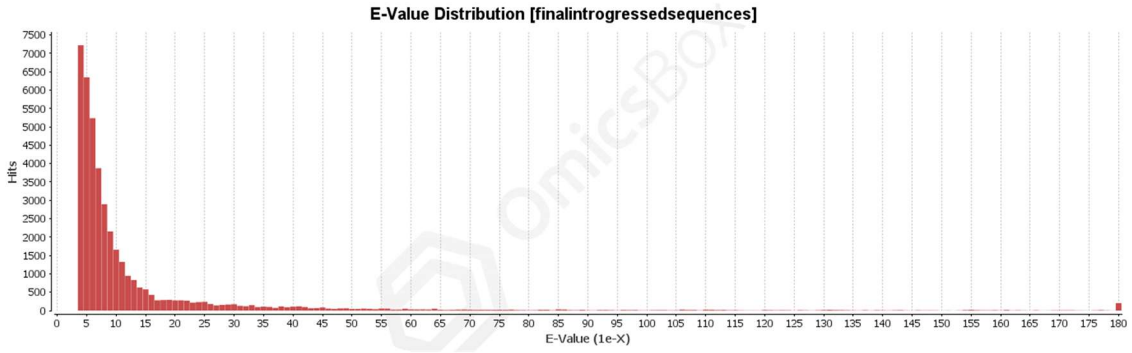


Figure 30: The e-value distribution of the blastx results vs the number of best hits obtained at each e-value.

The bit-score is inversely proportional to the e-value. A larger bit-score is, the less likely it is to obtain by chance than a smaller bit-score. Around 850 query sequences had just one hit while the rest had more than one hit where blastx algorithm calculated the best hit for those sequences (Figure 31). Blastx provides a unique accession number for the best hit protein sequence, based on the sequence similarity. This gives us a hint of functionality of the respective query sequence. According to the blastx results 5,422 (35%) of the total regions identified had a hit with a protein sequence.

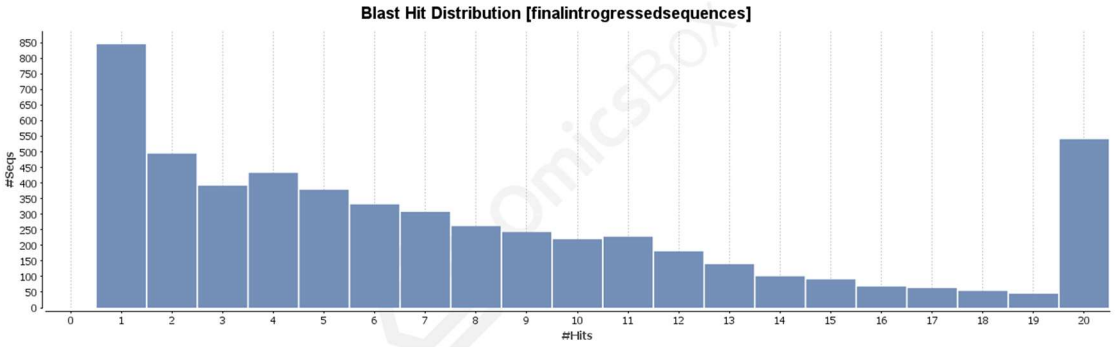


Figure 31: Number of blastx hits identified from public databases for the query sequences

5.6.3.2 OmicsBox InterProScan results

After scan through a list of protein databases InterProScan has retrieved domain/family information for 4,043 (26%) total input sequence regions. According to the results the regions were spread across 87 protein domains and some of the most frequent domains are shown in the Figure 32.

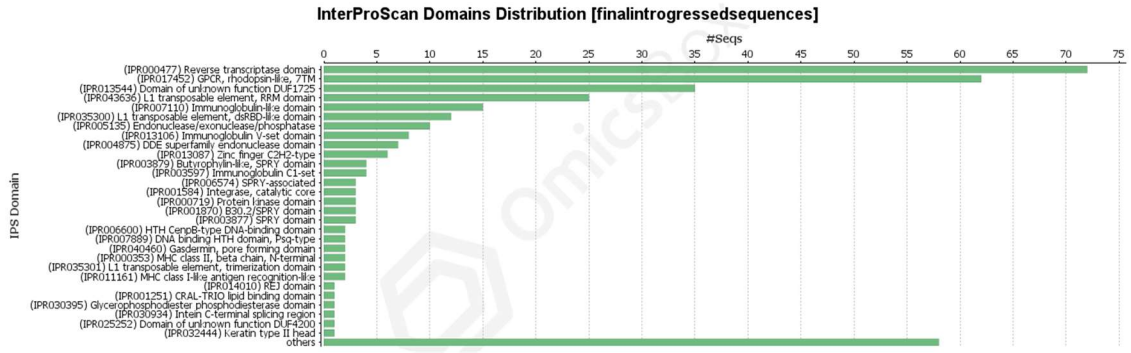


Figure 32: Protein domains identified in the putative introgressed regions by InterProScan

Reverse transcriptase domain (IPR000477) is the most frequently identified protein domain among 72 putative introgressed regions. Reverse transcriptase (RT) is also known as RNA-dependent DNA polymerase which is a DNA polymerase enzyme that transcribes single-stranded RNA into DNA [128]. This enzyme is used by retroviruses, bacterial retrons to integrate their genomic material into host genomes for replication of retroid elements. These are DNA received from an infection in a long-forgotten ancestor.

Interestingly the second most common protein domain is the G protein-coupled receptors (GPCR), rhodopsin-like, 7TM (IPR017452) where 62 of the putative introgressed regions were categorized. The rhodopsin-like GPCRs represent a protein family that includes hormones, neurotransmitters and light receptors, all of which transduce extracellular signals into intracellular pathways through interaction with guanine nucleotide-binding (G) proteins [129]. GPCRs are major drug targets, and therefore are a subject of significant research interest. The identified regions in our study are associated with Olfactory receptors (ORs) which are GPCRs that signal by elevating intracellular cAMP. There are only a handful of studies carried out by examining the role of olfaction

in horses. The available results show that it is an evolutionary beneficial trait which aids equids in social recognition of other animals including rivals, reproduction, for flight from predators all based on the body odors and fecal odors [130].

There were also regions related with immunoglobulin-like domains that consist of sequence structures found in several diverse protein families. Ig-like domains are involved in a variety of functions, including cell-cell recognition, cell-surface receptors, muscle structure and the immune system [131]. There were many other important protein domains identified such as MHC which were attached to putative introgressed regions, which are listed in the following table (Table 9).

Table 9: List of protein domains identified for the putative introgressed regions by IntroProScan

IPS ID	Protein Domain Name
IPR000477	Reverse transcriptase domain
IPR017452	GPCR, rhodopsin-like, 7TM
IPR013544	Domain of unknown function DUF1725
IPR043636	L1 transposable element, RRM domain
IPR007110	Immunoglobulin-like domain
IPR035300	L1 transposable element, dsRBD-like domain
IPR005135	Endonuclease/exonuclease/phosphatase
IPR013106	Immunoglobulin V-set domain
IPR004875	DDE superfamily endonuclease domain
IPR013087	Zinc finger C2H2-type
IPR003879	Butyrophylin-like, SPRY domain
IPR003597	Immunoglobulin C1-set
IPR006574	SPRY-associated
IPR001584	Integrase, catalytic core
IPR000719	Protein kinase domain
IPR001870	B30.2/SPRY domain
IPR003877	SPRY domain
IPR006600	HTH CenpB-type DNA-binding domain
IPR007889	DNA binding HTH domain, Psq-type
IPR040460	Gasdermin, pore forming domain
IPR000353	MHC class II, beta chain, N-terminal
IPR035301	L1 transposable element, trimerization domain
IPR011161	MHC class I-like antigen recognition-like
IPR014010	REJ domain

IPR001251	CRAL-TRIO lipid binding domain
IPR030395	Glycerophosphodiester phosphodiesterase domain
IPR030934	Intein C-terminal splicing region
IPR025252	Domain of unknown function DUF4200
IPR032444	Keratin type II head
IPR010473	Formin, GTPase-binding domain
IPR001594	Palmitoyltransferase, DHHC domain
IPR001208	MCM domain
IPR001254	Serine proteases, trypsin domain
IPR001279	Metallo-beta-lactamase
IPR039509	SPATA31/FAM205
IPR000488	Death domain
IPR006703	AIG1-type guanine nucleotide-binding
IPR013098	Immunoglobulin I-set
IPR039008	Intermediate filament, rod domain
IPR021930	Heparan sulphate-N-deacetylase
IPR007125	Histone H2A/H2B/H3
IPR013845	Ribosomal protein S4e, central region
IPR020454	Diacylglycerol/phorbol-ester binding
IPR001039	MHC class I alpha chain, alpha1 alpha2 domains
IPR041982	Ribosomal protein S4, KOW domain
IPR030386	GB1/RHD3-type guanine nucleotide-binding
IPR009056	Cytochrome c-like domain
IPR001660	Sterile alpha motif domain
IPR006153	Cation/H+ exchanger
IPR000742	EGF-like domain
IPR041697	Zinc-finger C2H2-type 11
IPR013122	Polycystin cation channel, PKD1/PKD2
IPR002018	Carboxylesterase, type B
IPR040878	C17orf99, Ig domain
IPR012461	FAM83, N-terminal
IPR000315	B-box-type zinc finger
IPR000008	C2 domain
IPR002156	Ribonuclease H domain
IPR004046	Glutathione S-transferase, C-terminal
IPR001433	Oxidoreductase FAD/NAD
IPR002219	Protein kinase C-like, phorbol ester/diacylglycerol-binding domain
IPR001478	PDZ domain
IPR015894	Guanylate-binding protein, N-terminal
IPR032281	40S ribosomal protein SA, C-terminal domain
IPR005160	Ku70/Ku80 C-terminal arm
IPR001841	Zinc finger, RING-type
IPR004114	THUMP domain
IPR001818	Peptidase M10, metallopeptidase
IPR012972	NLE
IPR000885	Fibrillar collagen, C-terminal

IPR001627	Sema domain
IPR007707	Transforming acidic coiled-coil-containing protein, C-terminal
IPR004841	Amino acid permease/ SLC12A domain
IPR039478	Protein FAM184A/B, N-terminal
IPR001565	Synaptotagmin
IPR022049	FAM69, protein-kinase domain
IPR000195	Rab-GTPase-TBC domain
IPR028073	PTHB1, N-terminal domain
IPR000436	Sushi/SCR/CCP domain
IPR000834	Peptidase M14, carboxypeptidase A
IPR037615	Kazrin, SAM domain repeat 2
IPR037958	Butyrophilin subfamily 1/2, SPRY/PRY domain
IPR000328	Retroviral envelope protein GP41-like
IPR001079	Galectin, carbohydrate recognition domain
IPR002791	Damage-control phosphatase ARMT1-like, metal-binding domain
IPR001214	SET domain
IPR001073	C1q domain

The regions which were successfully scanned through IntroProScan process were related to 151 protein families. The top few families are shown in Figure 33 and the rest are in Table 10.

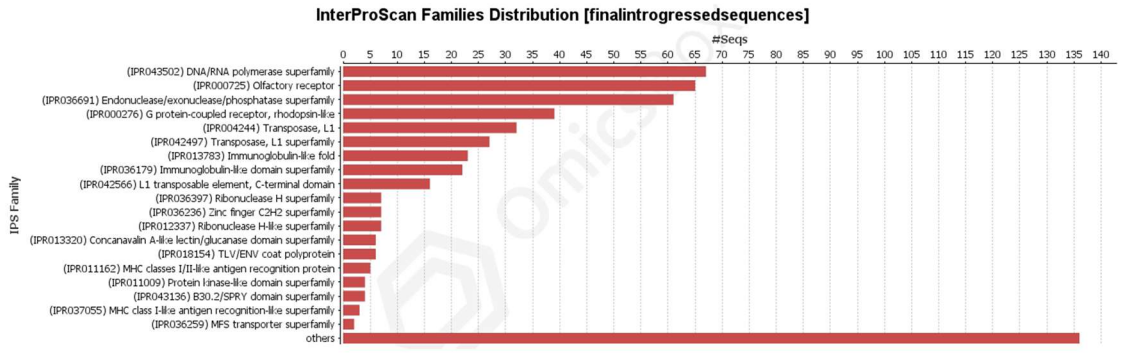


Figure 33: Protein families identified in the putative introgressed regions by InterProScan

Table 10: List of protein families identified for the putative introgressed regions by IntroProScan

IPS ID	Protein Family Name
IPR043502	DNA/RNA polymerase superfamily
IPR000725	Olfactory receptor
IPR036691	Endonuclease/exonuclease/phosphatase superfamily
IPR000276	G protein-coupled receptor, rhodopsin-like
IPR004244	Transposase, L1
IPR042497	Transposase, L1 superfamily
IPR013783	Immunoglobulin-like fold
IPR036179	Immunoglobulin-like domain superfamily
IPR042566	L1 transposable element, C-terminal domain
IPR036397	Ribonuclease H superfamily
IPR036236	Zinc finger C2H2 superfamily
IPR012337	Ribonuclease H-like superfamily
IPR013320	Concanavalin A-like lectin/glucanase domain superfamily
IPR018154	TLV/ENV coat polyprotein
IPR011162	MHC classes I/II-like antigen recognition protein
IPR011009	Protein kinase-like domain superfamily
IPR043136	B30.2/SPRY domain superfamily
IPR037055	MHC class I-like antigen recognition-like superfamily
IPR036259	MFS transporter superfamily
IPR014745	MHC class II, alpha/beta chain, N-terminal
IPR027417	P-loop containing nucleoside triphosphate hydrolase
IPR043128	Reverse transcriptase/Diguanylate cyclase domain
IPR007677	Gasdermin
IPR035908	ATP synthase, F0 complex, subunit A superfamily
IPR001907	ATP-dependent Clp protease proteolytic subunit
IPR042772	SH3 domain and tetratricopeptide repeat-containing protein SH3TC1/SH3TC2
IPR029246	Protein TALPID3
IPR037359	Heparan sulfate sulfotransferase
IPR032825	FRAS1-related extracellular matrix protein 1
IPR023395	Mitochondrial carrier domain superfamily
IPR028708	72kDa type IV collagenase
IPR035983	HECT, E3 ligase catalytic domain
IPR000568	ATP synthase, F0 complex, subunit A
IPR029606	Protein FAM184B
IPR005462	Transient receptor potential channel, canonical 6
IPR018422	Cation/H+ exchanger, CPA1 family
IPR033618	Ras association domain-containing protein 2
IPR033056	POU domain, class 6, transcription factor 2
IPR036927	Cytochrome c oxidase-like, subunit I superfamily
IPR000876	Ribosomal protein S4e
IPR042360	Aminoacyl tRNA synthase complex-interacting multifunctional protein 2

IPR030610	Protein-tyrosine kinase 2-beta
IPR005599	GPI mannosyltransferase
IPR028677	Synaptotagmin-10
IPR036396	Cytochrome P450 superfamily
IPR026511	Parathyroid hormone-responsive B1
IPR029627	Serine-rich coiled-coil domain-containing protein
IPR013083	Zinc finger, RING/FYVE/PHD-type
IPR030770	Signal-induced proliferation-associated 1-like protein 1
IPR015502	Glypican-1
IPR035892	C2 domain superfamily
IPR036282	Glutathione S-transferase, C-terminal domain superfamily
IPR026704	Katanin-interacting protein
IPR016335	Receptor-type tyrosine-protein phosphatase C
IPR011993	PH-like domain superfamily
IPR002153	Transient receptor potential channel, canonical
IPR036352	Sema domain superfamily
IPR036909	Cytochrome c-like domain superfamily
IPR011989	Armadillo-like helical
IPR037955	Inactive histone-lysine N-methyltransferase 2E
IPR030718	Protocadherin-15
IPR036770	Ankyrin repeat-containing domain superfamily
IPR040090	Thioredoxin domain-containing protein 16
IPR029605	FAM184 family
IPR029569	Calcium homeostasis modulator family
IPR033028	Whirlin
IPR030641	Grifin
IPR028082	Periplasmic binding protein-like I
IPR039915	TACC family
IPR024079	Metallopeptidase, catalytic domain superfamily
IPR009057	Homeobox-like domain superfamily
IPR035976	Sushi/SCR/CCP superfamily
IPR038237	Ribosomal protein S4e, central domain superfamily
IPR033614	C-terminal RASSF family
IPR026234	Mas-related G protein-coupled receptor family
IPR043519	Nucleotidyltransferase superfamily
IPR023591	Ribosomal protein S2, flavodoxin-like domain superfamily
IPR033237	BMP/retinoic acid-inducible neural-specific protein
IPR036034	PDZ superfamily
IPR030053	Taste receptor type 2 member 7
IPR000215	Serpin family
IPR042178	Serpin superfamily, domain 1
IPR011029	Death-like domain superfamily
IPR004203	Cytochrome c oxidase subunit IV family
IPR030267	Nesprin-3
IPR029045	ClpP/crotonase-like domain superfamily
IPR013761	Sterile alpha motif/pointed domain superfamily

IPR007775	Leukocyte-specific transcript 1, LST-1
IPR029021	Protein-tyrosine phosphatase-like
IPR029571	Calcium homeostasis modulator protein 3
IPR002293	Amino acid/polyamine transporter I
IPR026774	2'-5'-oligoadenylate synthase
IPR031751	Protein of unknown function DUF4735
IPR005707	Ribosomal protein S2, eukaryotic/archaeal
IPR001675	Glycosyl transferase family 29
IPR032675	Leucine-rich repeat domain superfamily
IPR037614	Kazrin
IPR021950	Transcription factor Spt20
IPR044156	Galectin-like
IPR039484	Alpha-1,2-mannosyltransferase ALG9-like
IPR014722	Ribosomal protein L2, domain 2
IPR036186	Serpin superfamily
IPR043504	Peptidase S1, PA clan, chymotrypsin-like fold
IPR008983	Tumour necrosis factor-like domain superfamily
IPR008405	Apolipoprotein L
IPR016024	Armadillo-type fold
IPR001894	Cathelicidin-like
IPR006539	P-type ATPase, subfamily IV
IPR028563	MICAL-like protein
IPR035969	Rab-GTPase-TBC domain superfamily
IPR001019	Guanine nucleotide binding protein
IPR029628	Serine-rich coiled-coil domain-containing protein 1
IPR036383	Thrombospondin type-1
IPR028054	Protein of unknown function DUF4481
IPR039752	F-box only protein
IPR039763	Protein-glutamate O-methyltransferase
IPR026197	Secretogranin III
IPR027075	Cleavage and polyadenylation specificity factor subunit 2
IPR036400	Cytochrome b5-like heme/steroid binding domain superfamily
IPR024606	Protein of unknown function DUF3827
IPR009072	Histone-fold
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase
IPR027309	P2X purinoreceptor extracellular domain superfamily
IPR039261	Ferredoxin-NADP reductase
IPR031327	Mini-chromosome maintenance protein
IPR031128	E3 ubiquitin-protein ligase RNF14
IPR007960	Taste receptor type 2
IPR036045	Sec1-like superfamily
IPR031127	E3 ubiquitin ligase RBR family
IPR036116	Fibronectin type III superfamily
IPR023562	Clp protease proteolytic subunit /Translocation-enhancing protein TepA
IPR036866	Ribonuclease Z/Hydroxyacylglutathione hydrolase-like
IPR036865	CRAL-TRIO lipid binding domain superfamily

IPR030349	Phospholipid-transporting ATPase IK
IPR028131	Tubulinyl-Tyr carboxypeptidase
IPR036639	Cytochrome c oxidase subunit IV superfamily
IPR042942	Laforin
IPR029058	Alpha/Beta hydrolase fold
IPR026915	Usherin
IPR003049	P2X6 purinoceptor
IPR000164	Histone H3/CENP-A
IPR023299	P-type ATPase, cytoplasmic domain N
IPR002327	Cytochrome c, class IA/ IB
IPR027862	Protein of unknown function DUF4534
IPR038578	GT29-like superfamiliy
IPR009003	Peptidase S1, PA clan
IPR038819	Basic immunoglobulin-like variable motif-containing protein
IPR029723	Integral membrane protein GPR137
IPR005828	Major facilitator, sugar transporter-like
IPR001863	Glypican
IPR015482	Syntrophin

5.6.3.3 GO mapping results

Mapping retrieved the GO-terms associated with the hits obtained by BLAST search. Out of the 5,422 sequences that had a BLAST hit only 704 were able to acquire GO-terms. This step did not have any user defined parameters. The obtained GO-term distribution is shown in Figure 34. In theory every gene and gene product should have at least one entry in each of CC, MF, BP. Every gene has multiple functions; hence every gene can have more than one GO-term attached to it. Some genes have dozens if not hundreds of entries in each of the categories.

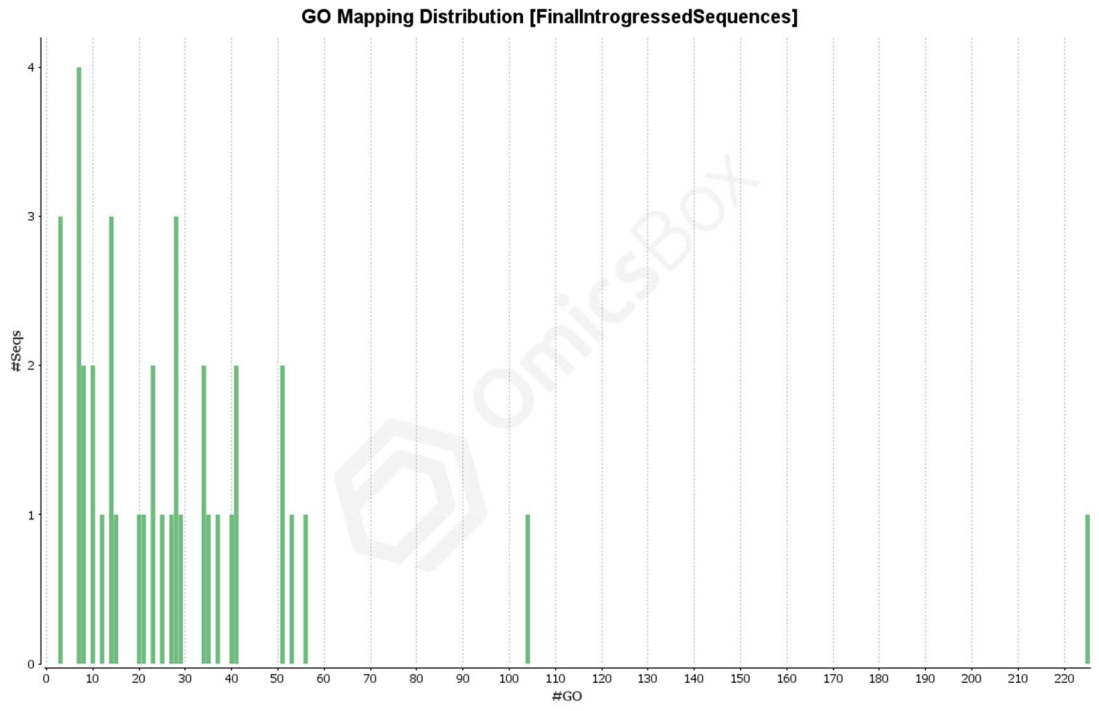


Figure 34: Distribution of GO terms vs the mapped putative introgressed regions

5.6.3.4 GO annotation results

The annotation step selected the most appropriate GO-terms for each sequence obtained by BLAST mapping and InterProScan steps. This is done by applying annotation rule based on the selected annotation score as discussed before. Figure 35 shows the annotation scores calculated by the tool and the number of sequences annotated at each score. Based on the distribution, the default annotation threshold was selected as the cutoff for annotation rule. Out of the 704 Go mapped sequences 535 were successfully annotated.

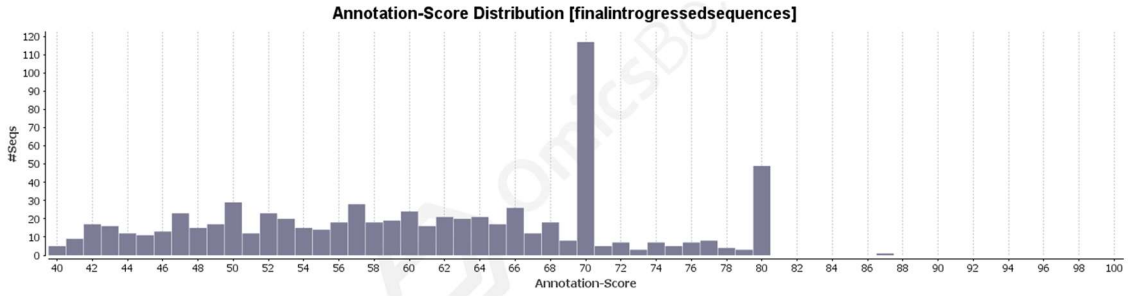


Figure 35: Annotation score distribution across GO mapped putative introgressed sequences

5.6.3.4.1 Biological Process aspect of GO

The successfully annotated sequences were explicitly and implicitly matched with 788 BP GO-terms, 205 CC GO-terms and 320 MF GO-terms (Appendix D, Figure 5-7). The implicit GO-terms are usually omitted in OmicsBox because there would be too much information on the screen and the information is considered. The tables 4, 5 and 6 shows only the most specific GO-terms in BP, CC and MF categories. The majority of annotated putative introgressed gene products in the BP category are involved in important biological processes such as regulation of response to stimuli like stress and detection of chemical stimulus involved in sensory perception, cell communication in signal transduction through

G protein-coupled receptor signaling pathways and regulation of cellular processes like cell differentiation and regulation in primary metabolic processes etc. More information about the BP annotated proteins and their interactions through GO graphs can be found at Appendix D (Figure 8, Figure 14) and the table below (Table 11).

Table 11: List of specific Biological Process (BP) GO-terms obtained from annotation

GO-terms	**#seqs**
Cellular process	402
Biological regulation	256
Regulation of biological process	240
Response to stimulus	237
Regulation of cellular process	234
Cellular response to stimulus	204
Metabolic process	198
Organic substance metabolic process	190
Cell communication	188
Signaling	186
Signal transduction	178
Cellular metabolic process	178
Nitrogen compound metabolic process	174
Multicellular organismal process	168
Primary metabolic process	144
Response to chemical	143
Organonitrogen compound metabolic process	131
Macromolecule metabolic process	128
Cellular nitrogen compound metabolic process	122
G protein-coupled receptor signaling pathway	121
System process	119
Nervous system process	110
Sensory perception	100
Cellular macromolecule metabolic process	95
Detection of stimulus	94
Sensory perception of chemical stimulus	93
Detection of chemical stimulus	93
Detection of stimulus involved in sensory perception	92
Detection of chemical stimulus involved in sensory perception	92
Localization	91
Sensory perception of smell	90
Detection of chemical stimulus involved in sensory perception of smell	90
Protein metabolic process	79
Establishment of localization	73

Organic cyclic compound metabolic process	73
Cellular component organization or biogenesis	73
Transport	73
Cellular component organization	72
Gene expression	71
Heterocycle metabolic process	70
Cellular aromatic compound metabolic process	69
Nucleobase-containing compound metabolic process	67
Developmental process	67
Biosynthetic process	66
Organic substance biosynthetic process	66
Positive regulation of biological process	66
Cellular biosynthetic process	65
Regulation of metabolic process	63
Immune system process	62
Ion transport	61
Nucleic acid metabolic process	60
Regulation of macromolecule metabolic process	60
Anatomical structure development	60
Cellular protein metabolic process	59
Positive regulation of cellular process	57
Regulation of cellular metabolic process	56
Cellular amide metabolic process	56
Macromolecule biosynthetic process	55
Regulation of primary metabolic process	55
Regulation of nitrogen compound metabolic process	54
Peptide metabolic process	54
Multicellular organism development	54
Cellular nitrogen compound biosynthetic process	53
Cellular macromolecule biosynthetic process	53
Regulation of biological quality	52
Response to stress	52
System development	50
Negative regulation of biological process	49
Regulation of gene expression	48
Rna metabolic process	48
Regulation of response to stimulus	48
Immune response	48
Organelle organization	46
Protein modification process	44
Macromolecule modification	44
Response to external stimulus	44
Cellular developmental process	44
Organic cyclic compound biosynthetic process	44
Cellular protein modification process	44
Response to organic substance	43

Heterocycle biosynthetic process	43
Cell differentiation	43

5.6.3.4.2 Cellular Component aspect of GO

The CC category predicted the locations of an annotated putative introgressed gene product relative to cellular structure. The genes were categorized as cellular compartments or stable macromolecular complexes of which they are working as a part. The annotated gene products in the analysis were categorized as cellular anatomical entities such as integral components of plasma membrane, membrane bound organelles, intracellular organelles, intracellular membrane-bound organelles like cytoskeleton, cytoplasm. Detailed on CC Go graphs can be found at Appendix D (Figure 9, 15) and table below (Table 12).

Table 12: List of specific Cellular Component (cc) GO-terms obtained from annotation

GO-terms	**#seqs**
Cellular anatomical entity	458
Membrane	301
Intrinsic component of membrane	248
Integral component of membrane	247
Intracellular anatomical structure	245
Organelle	231
Membrane-bounded organelle	199
Intracellular organelle	183
Cell periphery	170
Cytoplasm	163
Plasma membrane	159
Intracellular membrane-bounded organelle	151
Nucleus	108
Endomembrane system	91
Protein-containing complex	71
Organelle membrane	68
Intracellular non-membrane-bounded organelle	64
Non-membrane-bounded organelle	64

Vesicle	60
Intracellular vesicle	58
Cytoplasmic vesicle	58
Bounding membrane of organelle	53
Cytosol	47
Vesicle membrane	45

5.6.3.4.3 Molecular Function aspect of GO

Molecular function GO-terms describe the gene activities at a molecular level rather than entities like biological process or cellular components. Most of the annotated putative introgressed genes had a catalytic activity such as hydrolase activity and oxidoreductase activity. Some function as transmembrane signaling receptors, like olfactory receptor activity and G protein-coupled receptor activity. Other proteins had binding activities like nucleic acid binding, heterocyclic compound binding, cytoskeletal protein bounding etc. More information of MF GO graph is in Appendix D (Figure 10, 16) and the following table (Table 13).

Table 13: List of specific Molecular Function (MF) GO-terms obtained from annotation

GO-terms	**#seqs**
Binding	306
Catalytic activity	176
Ion binding	151
Organic cyclic compound binding	143
Heterocyclic compound binding	141
Signaling receptor activity	120
Molecular transducer activity	120
Transmembrane signaling receptor activity	119
Olfactory receptor activity	104
Metal ion binding	102
G protein-coupled receptor activity	102
Cation binding	102
Anion binding	98
Small molecule binding	98

Protein binding	94
Catalytic activity, acting on a protein	93
Nucleic acid binding	91
Hydrolase activity	70
Transition metal ion binding	64
Dna binding	60
Carbohydrate derivative binding	56
Oxidoreductase activity	56
Nucleoside phosphate binding	54
Nucleotide binding	54
Ribonucleotide binding	50
Purine nucleotide binding	50
Purine ribonucleotide binding	48
Purine ribonucleoside triphosphate binding	47

5.6.3.4.4 Enzyme code mapping and KEGG pathway analysis

Enzymes are a specialized class of proteins responsible for catalyzing chemical reactions within a cell [132]. They accelerate the chemical reactions by converting substrate molecules into different molecules referred to as products. According to the type of reactions they catalyze, these enzymes are classified into seven categories namely Oxidoreductases, Transferases, Hydrolases, Lyases, ligases, isomerases, and translocases. The Enzyme Commission numerical nomenclature classifies enzymes based on the overall reaction catalyzed. The OmicsBox Enzyme Code mapping step identified classes of enzymes catalyzing similar reactions rather than the individual enzymes. These Enzyme Codes were used to load KEGG pathways and following are few of the important metabolic pathways identified based on the annotated Enzyme Codes.

Table 14: Enzyme codes mapped into gene products of putative introgressed sequences

EC classes	**#seqs**
1.- oxidoreductases	56
2.- transferases	44

3.- hydrolases	70
4.- lyases	41
5.- isomerases	2
6.- ligases	2
7.- translocases	33

The Appendix D (Table 1, 2) contains more information about rest of the identified 57 pathways and Enzyme Codes. The most frequent pathways identified were from nucleotide metabolism, cofactors and vitamin metabolism, amino acids metabolism, and carbohydrate metabolism.

Purine metabolism – This nucleotide metabolic pathway occurs actively in the cytosol of the liver where all the necessary steps for Purine metabolism occur, including de novo purine biosynthetic pathway, purine salvage pathway, and degradation [133].

Thiamine metabolism – Thiamine is a vitamin which is released by the action of phosphatase and pyrophosphatase in the upper small intestine. Humans usually store thiamine in skeletal muscle, heart, brain, liver, and kidneys. Metabolism of this molecule helps convert carbohydrates into energy [134].

Alanine, aspartate and glutamate metabolism - This pathway describes the metabolism of the amino acids alanine and aspartate. Alanine is broken down by oxidative deamination, the inverse reaction of the reductive amination biosynthesis, catalyzed by the same enzymes [135].

Fructose and mannose metabolism - This carbohydrate metabolism pathway utilizes fructose, and fructose-6-phosphate, a glycolysis intermediate to produce important nucleotide sugars such as GDP-D-mannose and GDP-L-fucose. These compounds are essential substrates for glycosylphosphatidylinositol (GPI) anchor biosynthesis and synthesis of the anchors for N-glycans (N-glycans biosynthesis) [136].

CHAPTER 6

DISCUSSION

Introgression studies are gaining popularity and revealing important evolutionary relationships among closely related species. Likewise, this is the first comprehensive population genomics study conducted on investigating potential introgression between caballine and non-caballine equid genomes. In order to do this, we have analyzed high coverage whole genome sequencing data from a sample of 23 equids. The results demonstrated that about 1.1% of the equine genome consists of introgressed regions. Most of the introgressed regions get negatively selected with time and only a small number of regions get fixed in the recipient genome. The low percentage of introgressed Neandertal (1-4%) genomic regions in the modern human is an example for this [53].

The suggested introgression event between caballines and non-caballines is much older than the ~45,000 year-old introgression identified in the modern human genome [52]. The comparatively short ~2kb caballine haplotypes found in the equine genome is evident for this. Because modern humans have much longer introgressed haplotypes ~57kb long. The haplotype length tends to reduce with long term exposure to recombination [76]. So, shorter shared haplotypes tend to be much older than the longer ones. To further analyze this the genomic data from the oldest available equine fossil record was studied through the proposed introgression detection workflow. Even though the introgressed regions were ~25x fewer than rest of the horses, the fact that we could still identify introgression suggested that the introgression event is at least 700,000 years old. Since there is a

reproductive barrier between the caballines and the non-caballines we could imagine the introgression between them happened several million years ago but several million years after their divergence from a common ancestor 4.5 MYA.

Interestingly 64% of the putative introgressed regions were associated with structural elements responsible for coding, non-coding regions and the regulatory regions in the equine genome. Based on the FANNG initiative annotation data the introgressed regions were responsible for enhancing of gene transcription, regulation in promoter sites and transcription repression. They have been conserved in the equine genome for ~4.5 million years implying that their function is essential for the organism. Sixty-one percent of the variants used for introgression detection algorithm were from intronic regions and another 21% were from intergenic regions. So, there were non-coding genes involved in the results. They were initially considered to be non-functional evolutionary dead-ends. But due to recent research now we know some of these non-coding RNA genes (snRNA, snoRNA, lncRNA) participate in gene regulation [137].

Furthermore, we have carried out a GO annotation for the gene products associated with these regions. A 35% of the regions had a corresponding blast hit and 26% produced InterProScan results. Both steps could map 871 GO-terms with sequence products. After the final annotation step GO-terms explicitly assigned for the gene products were extracted. There were 82 GO-terms related to biological process, 28 for molecular functions, and 24 cellular components. The biological processes are large cellular programs accomplished by multiple activities. Few such important processes like regulation of response to stimuli, cell communication in signal transduction, regulation of cell differentiation and regulations in primary metabolic processes were attached to the identified introgressed regions

(Appendix D). Some of the regions in the study were associated with olfactory receptors which are smell sensory recognition units, function through G protein-coupled receptor signaling pathways. There were also regions related with immunoglobulin-like domains and MHC domains which are directly related with the immune system of the organism. Interestingly our results align with the extensive gene flow studies conducted by Orlando et al. They have also found evidence for continuous selection on olfaction and the immune system and throughout horse evolution [73].

Our preliminary studies found evidence that the allelic variation responsible for *CXCL16* cell surface receptor is due to introgression [1]. And in this study, we found the protein products annotated in the cellular component aspect are mainly responsible for integral components of plasma membrane. Other identified protein products have molecular functions, binding activities like nucleic acid binding, heterocyclic compound binding, cytoskeletal protein bounding etc. Several sequences were coding for enzymes involved in metabolic pathways such as Purine metabolism, Thiamine metabolism which are quite important for generation of necessary energy and cofactors to promote cell survival and proliferation [138]. These pathways provide the energy required specially in horse endurance exercises. Mutations in purine metabolism could cause muscle weakness, decreased muscle mass, cramping after exercise [139]. These are all essentially evolutionarily beneficial traits which would have assisted early horses to adapt and survive in new environments. These traits aid equids in social recognition of other animals for reproduction, to identify rivals, for flight from predators. There is a good chance they were introgressed into horse genome giving them an adaptive advantage.

We have developed a novel method to detect a special branch of introgression which is ghost introgression where we identify the potential introgressed regions in an extant species without the genomic information from the extinct/unknown doner species. The other available methods make strong assumptions about population history. For example, when detecting Neandertal segments in a target non-African genome, HMM assume Neandertal genome as an archaic reference and west African genomes as a modern human reference that does not share the introgression event [140]. Our method is based on MLE and phylogenetic inference and was presented at several scientific meetings over the years [141-143]. This has been one of the first attempt to identify genomic introgression in horse and it was challenging since horse genome is still going under annotation, there are still ongoing attempts to improve the current annotation. Because current genomic annotations for the horse have limited information about the functions of non-coding regions, making it difficult to identify variants that alter gene regulation. Notably, many conserved non-coding sequences are far from genes and cannot be assigned to defined functional classes.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTION

Several studies including ours have provided evidence that extinct species can imprint their genomic signature in the genomes of closely related present-day species through introgression. This was possible due to the increasing availability of high-throughput sequencing methods and WGS data to investigate the question. Going forward with our analysis, we need to incorporate more equine genomes into the analysis since large samples offer rich source of informative population summary statistics. Using reference genomes from other horse species will also remove the bias introduced by Twilight reference genome. There are active efforts going on to generate two phased genomes from a Shire and an Arabian F1 cross. The availability of a complete reference genome sequence enables researchers to address important genomic questions in their research. But in order to derive value from these high-throughput genomics datasets, we require good genomic annotation. This will allow us to translate functional genomics results into practical solutions for equine health and production and answer questions about their evolution.

By April 2021 there were 4,276 horse proteins in uniprotkb, 154 (swiss-prot) are reviewed and manually annotated and 4,430 (TrEMBL) are computationally analyzed and await full manual annotation. Bioinformatics tools in this area needs to be optimized for the difficult task of functional mining and providing a useful balance between quality and quantity of the transferred knowledge. With the Functional Annotation of Animal Genomes

(FAANG) initiative to annotate the equine genome, in the future we will be able to do the additional association studies that needed to be done in order to establish the impact of some of these regions. Some of the protein domains, biological process and metabolic pathways identified by the GO annotations are not extensively studied yet in horse genome.

However, these regions are still present at high frequency (by virtue of their homozygous genotype in at least one horse) and because they have not been driven out of the horse population, we believe they must be important for these animals, and we were able to provide a promising overview of the function of sequence products using the currently available tools. It would be useful to develop methods that can infer introgression jointly with models of selection, because these regions which are fixed in the horse genome must have been positively selected by evolution. Researchers have only recently been able to detect and quantify ghost introgression. Hence it has not been taken into account in studying evolutionary history of many organisms. Further studies will provide information about adaptation and speciation, and in depicting phylogenetic relationships. Future work is necessary in evaluating the functional and phenotypic consequences of introgressed sequences and refining estimates on the timing of these events.

REFERENCES

1. Sarkar, S., et al., *Allelic Variation in CXCL16 Determines CD3+ T Lymphocyte Susceptibility to Equine Arteritis Virus Infection and Establishment of Long-Term Carrier State in the Stallion.* PLOS Genetics, 2016. **12**(12): p. e1006467.
2. Mayr, E., *Animal Species and Evolution*. 1963, Cambridge: Harvard Univ. Press. .
3. Li, G., et al., *Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae).* Genome Res, 2016. **26**(1): p. 1-11.
4. BENIRSCHKE, K., L.E. BROWNHILL, and M.M. BEATH, *Somatic chromosomes of the horse, the donkey and their hybrids, the mule and the hinny.* Reproduction, 1962. **4**(3): p. 319-326.
5. Robertson, A., *Artificial selection in plants and animals.* Proceedings of the Royal Society of London. Series B. Biological Sciences, 1966. **164**(995): p. 341-349.
6. Bovenhuis, H., J.A. Van Arendonk, and S. Korver, *Associations between milk protein polymorphisms and milk production traits.* Journal of dairy science, 1992. **75**(9): p. 2549-2559.
7. Zawierta, M., et al., *The role of intragenomic recombination rate in the evolution of population's genetic pool.* Theory in Biosciences, 2007. **125**(2): p. 123-132.
8. Barrett, R.D.H. and D. Schluter, *Adaptation from standing genetic variation.* Trends in Ecology & Evolution, 2008. **23**(1): p. 38-44.
9. Librado, P., et al., *Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments.* Proceedings of the National Academy of Sciences, 2015. **112**: p. E6889-E6897.
10. Liu, X., et al., *EPAS1 Gain-of-Function Mutation Contributes to High-Altitude Adaptation in Tibetan Horses.* Molecular Biology and Evolution, 2019. **36**(11): p. 2591-2603.
11. Li, Y., et al., *Population Variation Revealed High-Altitude Adaptation of Tibetan Mastiffs.* Molecular Biology and Evolution, 2014. **31**(5): p. 1200-1205.
12. Ai, H., et al., *Population history and genomic signatures for high-altitude adaptation in Tibetan pigs.* BMC Genomics, 2014. **15**(1): p. 834.
13. Zhang, H., et al., *Blood Characteristics for High Altitude Adaptation in Tibetan Chickens1.* Poultry Science, 2007. **86**(7): p. 1384-1389.
14. Song, S., et al., *Exome sequencing reveals genetic differentiation due to high-altitude adaptation in the Tibetan cashmere goat (Capra hircus).* BMC Genomics, 2016. **17**(1): p. 122.

15. Wu, D.-D., et al., *Pervasive introgression facilitated domestication and adaptation in the Bos species complex.* Nature Ecology & Evolution, 2018. **2**(7): p. 1139-1145.
16. Witt, K.E. and E. Huerta-Sánchez, *Convergent evolution in human and domesticate adaptation to high-altitude environments.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2019. **374**(1777): p. 20180235.
17. Simonson, T.S., et al., *Genetic Evidence for High-Altitude Adaptation in Tibet.* Science, 2010. **329**: p. 72-75.
18. Beall, C.M., et al., *Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders.* Proceedings of the National Academy of Sciences, 2010. **107**: p. 11459-11464.
19. Yan, X., J.P. Lynch, and S.E. Beebe, *Genetic Variation for Phosphorus Efficiency of Common Bean in Contrasting Soil Types: I. Vegetative Response.* Crop Science, 1995. **35**(4): p. cropsci1995.0011183X003500040028x.
20. Davy, A.J., S.M. Noble, and R.P. Oliver, *Genetic variation and adaptation to flooding in plants.* Aquatic Botany, 1990. **38**(1): p. 91-108.
21. Masel, J., *Genetic drift.* Curr Biol, 2011. **21**(20): p. R837-8.
22. Star, B. and H.G. Spencer, *Effects of genetic drift and gene flow on the selective maintenance of genetic variation.* Genetics, 2013. **194**(1): p. 235-244.
23. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
24. Jarcho, J., *Restriction fragment length polymorphism analysis.* Curr Protoc Hum Genet, 2001. **Chapter 2**: p. Unit 2.7.
25. Donis-Keller, H., et al., *A genetic linkage map of the human genome.* Cell, 1987. **51**(2): p. 319-337.
26. Kappes, S.M., et al., *A second-generation linkage map of the bovine genome.* Genome Res, 1997. **7**(3): p. 235-49.
27. Rohrer, G.A., et al., *A comprehensive map of the porcine genome.* Genome Res, 1996. **6**(5): p. 371-91.
28. Groenen, M.A., et al., *A consensus linkage map of the chicken genome.* Genome research, 2000. **10**(1): p. 137-147.
29. Vaiman, D., et al., *A genetic linkage map of the male goat genome.* Genetics, 1996. **144**(1): p. 279-305.
30. Swinburne, J., et al., *First comprehensive low-density horse linkage map based on two 3-generation, full-sibling, cross-bred horse reference families.* Genomics, 2000. **66**(2): p. 123-34.
31. Cunningham, E.P., et al., *Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses.* Anim Genet, 2001. **32**(6): p. 360-4.
32. Vignal, A., et al., *A review on SNP and other types of molecular markers and their use in animal genetics.* Genetics Selection Evolution, 2002. **34**(3): p. 275.

33. Lindblad-Toh, K., et al., *Genome sequence, comparative analysis and haplotype structure of the domestic dog.* Nature, 2005. **438**(7069): p. 803-19.
34. Elsik, C.G., et al., *The genome sequence of taurine cattle: a window to ruminant biology and evolution.* Science, 2009. **324**(5926): p. 522-8.
35. Wade, C.M., et al., *Genome sequence, comparative analysis, and population genetics of the domestic horse.* Science, 2009. **326**(5954): p. 865-7.
36. Chowdhary, B.P., N. Paria, and T. Raudsepp, *Potential applications of equine genomics in dissecting diseases and fertility.* Anim Reprod Sci, 2008. **107**(3-4): p. 208-18.
37. Doan, R., et al., *Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare.* BMC genomics, 2012. **13**: p. 78-78.
38. Kalbfleisch, T.S., et al., *Improved reference genome for the domestic horse increases assembly contiguity and composition.* Communications Biology, 2018. **1**(1): p. 197.
39. McCue, M.E., et al., *A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies.* PLoS genetics, 2012. **8**(1): p. e1002451-e1002451.
40. McCue, M. and J. Mickelson, *Genomic tools and resources: Development and applications of an equine SNP genotyping array*. 2013: Wiley-Blackwell.
41. Finno, C.J. and D.L. Bannasch, *Applied equine genetics.* Equine Veterinary Journal, 2014. **46**(5): p. 538-544.
42. Schaefer, R.J., et al., *Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds.* BMC Genomics, 2017. **18**(1): p. 565.
43. Bauer, A., et al., *A nonsense variant in the ST14 gene in akhal-teke horses with naked foal Syndrome.* G3: Genes, Genomes, Genetics, 2017. **7**(4): p. 1315-1321.
44. Thomer, A., et al., *An epistatic effect of KRT25 on SP6 is involved in curly coat in horses.* Scientific reports, 2018. **8**(1): p. 1-12.
45. Fages, A., et al., *Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series.* Cell, 2019. **177**(6): p. 1419-1435.e31.
46. Suarez-Gonzalez, A., C. Lexer, and Q.C. Cronk, *Adaptive introgression: a plant perspective.* Biology letters, 2018. **14**(3): p. 20170688.
47. Anderson, E. and L. Hubricht, *HYBRIDIZATION IN TRADESCANTIA. III. THE EVIDENCE FOR INTROGRESSIVE HYBRIDIZATION.* American Journal of Botany, 1938. **25**(6): p. 396-402.
48. Miller, A.H., *Mayr's 'Systematics and the Origin of Species'.* The Auk, 1943. **60**(2): p. 289-291.
49. Lewontin, R. and L. Birch, *Hybridization as a source of variation for adaptation to new environments.* Evolution, 1966: p. 315-336.

50. GOULD, S.J. and D.S. WOODRUFF, *History as a cause of area effects: an illustration from Cerion on Great Inagua, Bahamas.* Biological Journal of the Linnean Society, 1990. **40**(1): p. 67-98.
51. Heliconius Genome, C., *Butterfly genome reveals promiscuous exchange of mimicry adaptations among species.* Nature, 2012. **487**(7405): p. 94-98.
52. Green, R.E., et al., *A draft sequence of the Neandertal genome.* Science, 2010. **328**(5979): p. 710-722.
53. Reich, D., et al., *Genetic history of an archaic hominin group from Denisova Cave in Siberia.* Nature, 2010. **468**(7327): p. 1053-1060.
54. Meyer, M., et al., *A High-Coverage Genome Sequence from an Archaic Denisovan Individual.* Science, 2012. **338**: p. 222-226.
55. Hammer, M.F., et al., *Genetic evidence for archaic admixture in Africa.* Proceedings of the National Academy of Sciences, 2011. **108**(37): p. 15123-15128.
56. Vernot, B. and J.M. Akey, *Resurrecting surviving Neandertal lineages from modern human genomes.* Science, 2014. **343**(6174): p. 1017-1021.
57. Xu, D., et al., *Archaic hominin introgression in Africa contributes to functional salivary MUC7 genetic variation.* Molecular Biology and Evolution, 2017. **34**(10): p. 2704-2715.
58. Sankararaman, S., et al., *The combined landscape of Denisovan and Neanderthal ancestry in present-day humans.* Current Biology, 2016. **26**(9): p. 1241-1247.
59. Juric, I., S. Aeschbacher, and G. Coop, *The Strength of Selection against Neanderthal Introgression.* PLOS Genetics, 2016. **12**(11): p. e1006340.
60. Huerta-Sánchez, E., et al., *Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA.* Nature, 2014. **512**(7513): p. 194-197.
61. Miao, B., Z. Wang, and Y. Li, *Genomic Analysis Reveals Hypoxia Adaptation in the Tibetan Mastiff by Introgression of the Gray Wolf from the Tibetan Plateau.* Molecular Biology and Evolution, 2016. **34**(3): p. 734-743.
62. Ai, H., et al., *Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing.* Nature Genetics, 2015. **47**(3): p. 217-225.
63. Ottenburghs, J., *Ghost Introgression: Spooky Gene Flow in the Distant Past.* BioEssays, 2020. **42**(6): p. 2000012.
64. Kuhlwilm, M., et al., *Ancient admixture from an extinct ape lineage into bonobos.* Nature Ecology & Evolution, 2019. **3**(6): p. 957-965.
65. Árnason, Ú., et al., *Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow.* Science Advances, 2018. **4**.
66. Zeder, M.A., *Central questions in the domestication of plants and animals.* Evolutionary Anthropology: Issues, News, and Reviews, 2006. **15**(3): p. 105-117.
67. Guthrie, R.D., *The nature of Paleolithic art*. 2005: University of Chicago Press.
68. Outram, A.K., et al., *The Earliest Horse Harnessing and Milking.* Science, 2009. **323**: p. 1332-1335.

69. Vilà, C., et al., *Widespread Origins of Domestic Horse Lineages.* Science, 2001. **291**: p. 474-477.
70. MacFadden, B.J., *Fossil Horses: Evidence for Evolution.* Science, 2005. **307**(5716): p. 1728-1730.
71. Stock, C., *The Dawn Horse or Eohippus.* Engineering and Science, 1947. **10** (4): p. 4-5.
72. MacFadden, B.J., *Fossil Horses: Systematics, Paleobiology, and Evolution of the Family Equidae*. 1992, New York: Cambridge Univ. Press.
73. Orlando, L., et al., *Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse.* Nature, 2013. **499**(7456): p. 74-78.
74. Jónsson, H., et al., *Speciation with gene flow in equids despite extensive chromosomal plasticity.* Proceedings of the National Academy of Sciences, 2014. **111**: p. 18655-18660.
75. Patterson, N., et al., *Ancient Admixture in Human History.* Genetics, 2012. **192**: p. 1065-1093.
76. Racimo, F., D. Marnetto, and E. Huerta-Sánchez, *Signatures of Archaic Adaptive Introgression in Present-Day Human Populations.* Molecular Biology and Evolution, 2016. **34**(2): p. 296-317.
77. Weddell, B.J., *Conserving living natural resources: in the context of a changing world*. 2002: Cambridge University Press.
78. Coordinators, N.R., *Database resources of the National Center for Biotechnology Information*, in *Nucleic Acids Res*. 2018. p. D8-d13.
79. Leinonen, R., et al., *The sequence read archive.* Nucleic acids research, 2011. **39**(Database issue): p. D19-D21.
80. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-3048.
81. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* 2011, 2011. **17**(1): p. 3.
82. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2014. **31**(2): p. 166-169.
83. Pagès H, A.P., Gentleman R, DebRoy S. *Biostrings: Efficient manipulation of biological strings. R package version 2.58.0*. 2020; Available from: https://bioconductor.org/packages/Biostrings.
84. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.
85. Kalbfleisch, T. and M. Heaton, *Mapping whole genome shotgun sequence and variant calling in mammalian species without their reference genomes [version 2; peer review: 2 approved]*. F1000Research, 2014. **2**(244).

86. Solari, S. and R.J. Baker, *Mammal Species of the World: A Taxonomic and Geographic Reference by D. E. Wilson; D. M. Reeder.* Journal of Mammalogy, 2007. **88**(3): p. 824-830.
87. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
88. Koboldt, D.C., *Best practices for variant calling in clinical sequencing.* Genome Medicine, 2020. **12**(1): p. 91.
89. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
90. Broad Institute, G.R. *"Picard Toolkit."*. 2019; Available from: https://broadinstitute.github.io/picard/.
91. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**(1110): p. 11.10.1-11.10.33.
92. Poplin, R., et al., *Scaling accurate genetic variant discovery to tens of thousands of samples.* bioRxiv, 2017: p. 201178.
93. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biology, 2016. **17**(1): p. 122.
94. Torquato, S., B. Lu, and J. Rubinstein, *Nearest-neighbour distribution function for systems on interacting particles.* Journal of Physics A: Mathematical and General, 1990. **23**(3): p. L103-L107.
95. Delaneau, O., et al., *Haplotype Estimation Using Sequencing Reads.* The American Journal of Human Genetics, 2013. **93**(4): p. 687-696.
96. Browning, S.R. and B.L. Browning, *Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering.* The American Journal of Human Genetics, 2007. **81**(5): p. 1084-1097.
97. Rannala, B. and Z. Yang, *Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference.* Journal of Molecular Evolution, 1996. **43**(3): p. 304-311.
98. Huelsenbeck, J.P. and F. Ronquist, *MRBAYES: Bayesian inference of phylogenetic trees.* Bioinformatics, 2001. **17**(8): p. 754-5.
99. Yang, Z., *Estimating the pattern of nucleotide substitution.* J Mol Evol, 1994. **39**(1): p. 105-11.
100. Hillis, D.M., *Approaches for Assessing Phylogenetic Accuracy.* Systematic Biology, 1995. **44**(1): p. 3-16.
101. Janeway Jr, C.A., et al., *The major histocompatibility complex and its functions*, in *Immunobiology: The Immune System in Health and Disease. 5th edition*. 2001, Garland Science.

102. Racimo, F., et al., *Evidence for archaic adaptive introgression in humans.* Nature Reviews Genetics, 2015. **16**(6): p. 359-371.
103. Hillier, L.W., et al., *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**(7018): p. 695-716.
104. Morton, N.E., *Parameters of the human genome.* Proceedings of the National Academy of Sciences, 1991. **88**: p. 7474-7476.
105. Freeman, T.C., et al., *A gene expression atlas of the domestic pig.* BMC Biology, 2012. **10**(1): p. 90.
106. Jiang, Y., et al., *The sheep genome illuminates biology of the rumen and lipid metabolism.* Science, 2014. **344**: p. 1168-1173.
107. Schaub, M.A., et al., *Linking disease associations with regulatory information in the human genome.* Genome Research, 2012. **22**(9): p. 1748-1759.
108. Andersson, L., et al., *Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project.* Genome Biology, 2015. **16**(1): p. 57.
109. ENCODE_Project_Consortium, *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**: p. 636-640.
110. Tuggle, C.K., et al., *GO-FAANG meeting: a Gathering On Functional Annotation of Animal Genomes.* Animal Genetics, 2016. **47**(5): p. 528-533.
111. Giuffra, E., C.K. Tuggle, and F. Consortium, *Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap.* Annual Review of Animal Biosciences, 2019. **7**(1): p. 65-88.
112. Burns, E.N., et al., *Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project.* Animal Genetics, 2018. **49**(6): p. 564-570.
113. Bannister, A.J. and T. Kouzarides, *Regulation of chromatin by histone modifications.* Cell Research, 2011. **21**(3): p. 381-395.
114. Ng, S., et al., *Dynamic protein methylation in chromatin biology.* Cellular and molecular life sciences, 2009. **66**(3): p. 407-422.
115. Berger, S.L., *Histone modifications in transcriptional regulation.* Curr Opin Genet Dev, 2002. **12**(2): p. 142-8.
116. Allfrey, V.G., R. Faulkner, and A.E. Mirsky, *ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS.* Proceedings of the National Academy of Sciences, 1964. **51**: p. 786-794.
117. Howe, F.S., et al., *Is H3K4me3 instructive for transcription activation?* Bioessays, 2017. **39**(1): p. 1-12.
118. Kingsley, N.B., et al., *Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq.* Genes (Basel), 2019. **11**(1).

119. Wang, M., et al., *Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits.* Genetics Selection Evolution, 2017. **49**(1): p. 56.
120. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
121. Jones, C.E., A.L. Brown, and U. Baumann, *Estimating the annotation error rate of curated GO database sequence annotations.* BMC Bioinformatics, 2007. **8**: p. 170.
122. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.
123. Consortium, T.G.O., *Gene Ontology Consortium: going forward.* Nucleic Acids Research, 2014. **43**(D1): p. D1049-D1056.
124. Götz, S., et al., *High-throughput functional annotation and data mining with the Blast2GO suite.* Nucleic Acids Res, 2008. **36**(10): p. 3420-35.
125. Jung, H., et al., *Twelve quick steps for genome assembly and annotation in the classroom.* PLOS Computational Biology, 2020. **16**(11): p. e1008325.
126. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
127. Pertsemlidis, A. and J.W. Fondon, *Having a BLAST with bioinformatics (and avoiding BLASTphemy).* Genome Biology, 2001. **2**(10): p. reviews2002.1.
128. Inouye, S. and M. Inouye, *Structure, function, and evolution of bacterial reverse transcriptase.* Virus genes, 1995. **11**(2-3): p. 81-94.
129. Vassilatis, D.K., et al., *The G protein-coupled receptor repertoires of human and mouse.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(8): p. 4903-4908.
130. Rørvang, M.V., B.L. Nielsen, and A.N. McLean, *Sensory Abilities of Horses and Their Importance for Equitation Science.* Frontiers in Veterinary Science, 2020. **7**(633).
131. Potapov, V., et al., *Protein--protein recognition: juxtaposition of domain and interface cores in immunoglobulins and other sandwich-like proteins.* J Mol Biol, 2004. **342**(2): p. 665-79.
132. Rost, B., *Enzyme Function Less Conserved than Anticipated.* Journal of Molecular Biology, 2002. **318**(2): p. 595-608.
133. Yin, J., et al., *Potential Mechanisms Connecting Purine Metabolism and Cancer Therapy.* Frontiers in Immunology, 2018. **9**(1697).
134. Rindi, G. and U. Laforenza, *Thiamine intestinal transport and related issues: recent aspects.* Proc Soc Exp Biol Med, 2000. **224**(4): p. 246-55.
135. Curien, G., et al., *Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters.* Molecular Systems Biology, 2009. **5**(1): p. 271.
136. Rolfsson, Ó., et al., *Mannose and fructose metabolism in red blood cells during cold storage in SAGM.* Transfusion, 2017. **57**(11): p. 2665-2676.

137. Xiao, J., et al., *Pseudogenes and Their Genome-Wide Prediction in Plants.* International Journal of Molecular Sciences, 2016. **17**(12): p. 1991.
138. Harkness, R.A., *Purine metabolism in the horse--are evolutionary differences linked to muscular performance?* Equine veterinary journal, 1986. **18**(1): p. 5-6.
139. Westermann, C.M., et al., *Equine metabolic myopathies with emphasis on the diagnostic approach. Comparison with human myopathies. A review.* Vet Q, 2007. **29**(2): p. 42-59.
140. Sankararaman, S., et al., *The genomic landscape of Neanderthal ancestry in present-day humans.* Nature, 2014. **507**(7492): p. 354-7.
141. De Silva, K., Heaton, M.P., Kalbfleisch, T.S. , *Identification of conserved genomic regions and variation therein amongst Cetartiodactyla species using next generation sequencing.* UT-KBRIN Bioinformatics Summit (April 8-10, 2016), Lake Barkley State Resort Park, Cadiz, KY, 2016: p. 8-10.
142. De Silva, K., Bailey, E., Kalbfleisch, T, S.,, *Identify Shared and Species-Specific k-mers in Equids and Caballines for Identification of Evolutionary Features.* Plant and Animal Genome XXVII Conference (January 12-16, 2019), 2019.
143. De Silva, K., *Investigating Ancient Introgression between Caballine and Non-Caballine Equids.* Plant and Animal Genome XXVI Conference (January 13-17, 2018), 2018.

APPENDIX A

LIST OF ABBREVIATIONS

BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Alignment
CDS	Coding DNA Sequences
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
CNV	Copy Number Variation
DNA	Deoxyribonucleic acid
EAV	Equine Arteritis Virus
ENCODE	Encyclopedia of DNA Elements
FAANG	Functional Annotation of Animal Genomes
GATK	Genome Analysis Tool Kit
Gb	Giga bases
GI	GenIdentifier
GnomAD	Genome Aggregation Database
GO	Gene Ontology
GTR	General Time Reversible
GWAS	Genome Wide Association Studies
HIF	Hypoxia Inducible Factor
HMM	Hidden Markov Model

HTS	High Throughput Sequencing
IGV	Integrative Genomics Viewer
INDEL	Insertion or Deletion
MAF	Minor Allele Frequency
MCMC	Markov chain Monte Carlo
MEM	maximal exact matches
MHC	Major Histone Compatibility
MHC	Major Histone Compatibility Complex
MLE	Maximum Likelihood Estimate
MRCA	Most Recent Common Ancestor
MYA	Million Years Ago
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NY	New York
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
SAM	Sequence Alignment Map
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
SRA	Sequence Read archive
VEP	Variant Effect Predictornv
WGS	Whole Genome Sequence

APPENDIX B

VARIANT EFFECT PREDICTION

B.1 Variant Classes

Table B1: The class of a variant according to its component alleles and its mapping to the equine reference genome

Variant class	Variant count
SNV	71,062,566 (89.2%)
Deletion	3,814,073 (4.8%)
Insertion	2,600,510 (3.3%)
Sequence alteration	1,809,148 (2.3%)
Indel	387,524 (0.5%)

B.2 Variant consequences on the transcripts

Table B2: The count of variants responsible for different consequences on the protein sequences in the entire genome

Consequence type	Variant count	Impact
Splice_donor_variant	10,778	High
Splice_acceptor_variant	4,131	High
Stop_gained	10,911	High
Frameshift_variant	35,802	High
Stop_lost	915	High
Start_lost	1,767	High
Inframe_insertion	8,540	Moderate
Inframe_deletion	9,468	Moderate
Missense_variant	670,547	Moderate
Protein_altering_variant	1,694	Moderate
Splice_region_variant	192,850	Low
Synonymous_variant	849,456	Low
Stop_retained_variant	590	Low
Start_retained_variant	35	Low
Coding_sequence_variant	1,252	Modifier
Mature_mirna_variant	314	Modifier
5_prime_utr_variant	194,695	Modifier
3_prime_utr_variant	383,143	Modifier

Non_coding_transcript_exon_variant	473,808	Modifier
Intron_variant	104,123,570	Modifier
Non_coding_transcript_variant	8,264,780	Modifier
Upstream_gene_variant	9,983,826	Modifier
Downstream_gene_variant	9,793,809	Modifier
Intergenic_variant	37,158,105	Modifier

Table B3: The count of filtered variants responsible for different consequences on the protein sequences in the entire genome

Consequence type	**Variant count**	**Impact**
Splice_acceptor_variant	694	High
Splice_donor_variant	956	High
Stop_gained	927	High
Frameshift_variant	2,960	High
Stop_lost	184	High
Start_lost	380	High
Inframe_insertion	1,128	Moderate
Inframe_deletion	930	Moderate
Protein_altering_variant	57	Moderate
Missense_variant	127,243	Moderate
Splice_region_variant	50,872	Low
Start_retained_variant	14	Low
Synonymous_variant	229,612	Low
Stop_retained_variant	171	Low
Coding_sequence_variant	176	Modifier
Mature_mirna_variant	59	Modifier
5_prime_utr_variant	43,134	Modifier
3_prime_utr_variant	95,662	Modifier
Non_coding_transcript_exon_variant	131,055	Modifier
Intron_variant	29,249,942	Modifier
Non_coding_transcript_variant	2,412,251	Modifier
Upstream_gene_variant	2,689,349	Modifier
Downstream_gene_variant	2,649,437	Modifier
Intergenic_variant	10,533,847	Modifier

Table B4: the count of filtered variants with consequences only on coding sequences

Consequence type	**Count**	**Impact**
Stop_gained	927	High
Frameshift_variant	2,960	High
Stop_lost	184	High

Start_lost	380	High
Inframe_insertion	1,128	Moderate
Inframe_deletion	930	Moderate
Missense_variant	127,243	Moderate
Protein_altering_variant	57	Moderate
Start_retained_variant	14	Low
Stop_retained_variant	171	Low
Synonymous_variant	229,612	Low
Coding_sequence_variant	176	Modifier

B.3 Variant count by chromosome

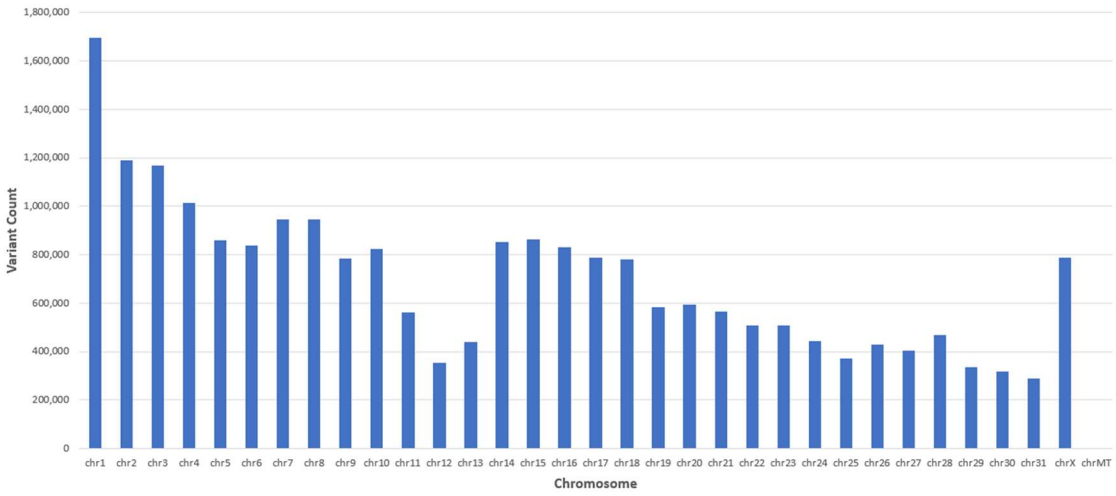


Figure B1: Distribution of variant counts along the chromosomes

APPENDIX C

MLE CALCULATION OF A NEGATIVE REGION AND A POSITIVE REGION

MLE = P(824)*P(209)*P(35)*P(59)*P(89)*P(60)/P(824)*P(209)*P(35)*P(59)*P(89)*P(60)
MLE = 0.630 (Likely not introgressed)

Figure C.1 MLE calculation of a negative result of introgression in a genomic region. The ratio of the products of distance probabilities from the *Twilight-Specific* alleles nearest neighbor distribution (blue) and the *Ancestral* alleles nearest neighbor distribution (orange) for cumulative distances between the SNPs is less than the experimental threshold of 100

Figure C.2 MLE calculation of a positive result of introgression in a genomic region. The ratio of the products of distance probabilities from the *Twilight-Specific* alleles nearest neighbor distribution (blue) and the *Ancestral* alleles nearest neighbor distribution (orange) for cumulative distances between the SNPs is greater than the experimental threshold of 100

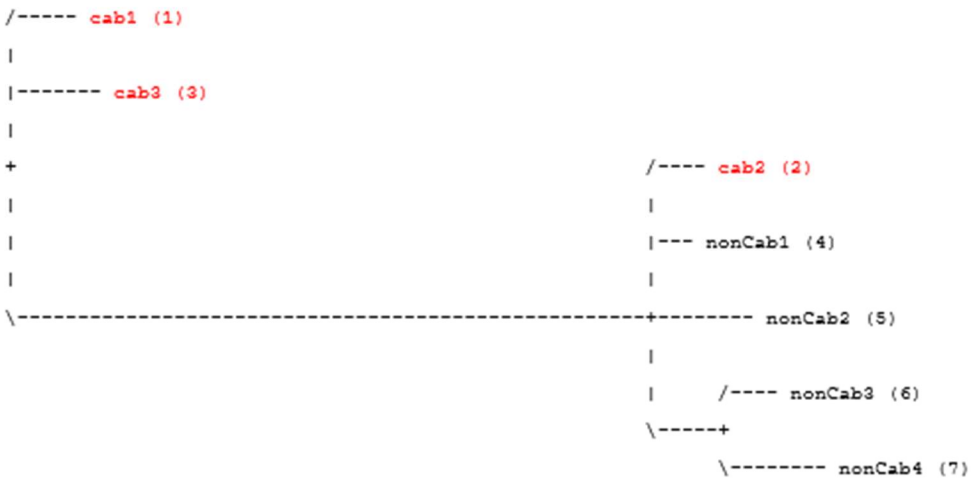


Figure C.3 The phylogenetic tree for the caballine and non-caballine haplotypes in the positively introgressed region. One of the caballine haplotypes (red) cluster with the non-caballine haplotypes.

APPENDIX D

FUNCTIONAL ANNOTATION OF PUTATIVE INTROGRESSED REGIONS

D.1 NCBI Blastx search results

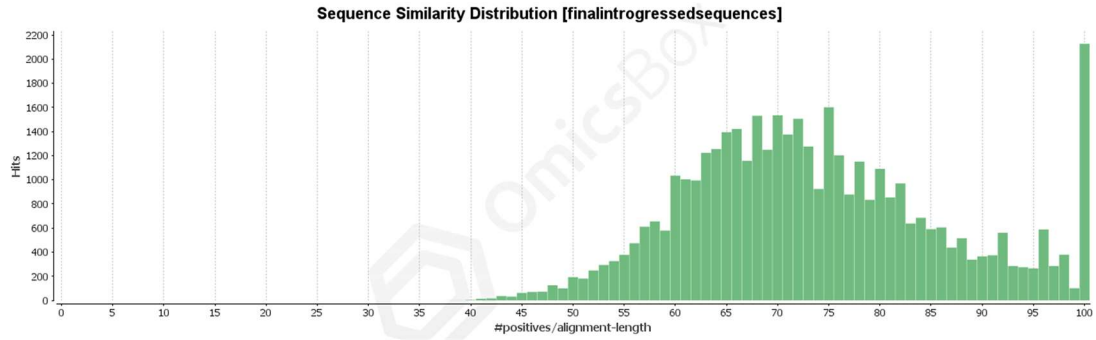


Figure D1: Sequence similarity distribution of blast bits.

D.2 OmicsBox GO Annotation results

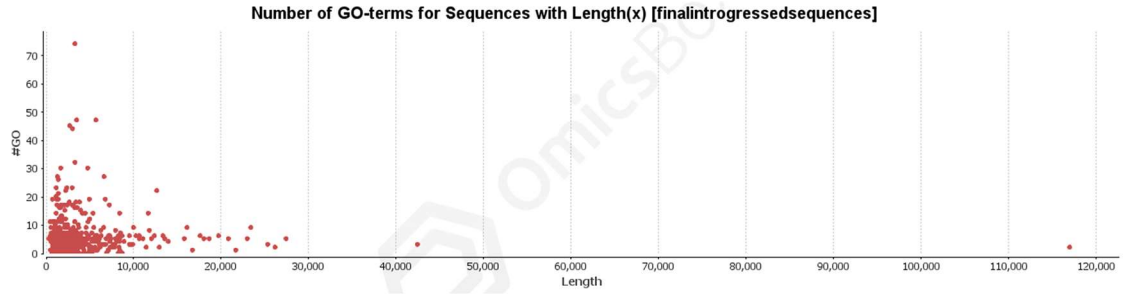


Figure D2: Number of GO-terms assigned for sequences at different lengths.

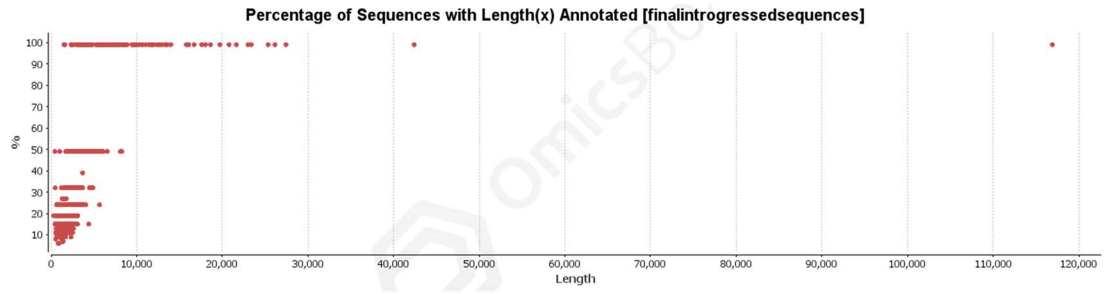


Figure D3: Percentage of annotated sequences based on the sequence length.

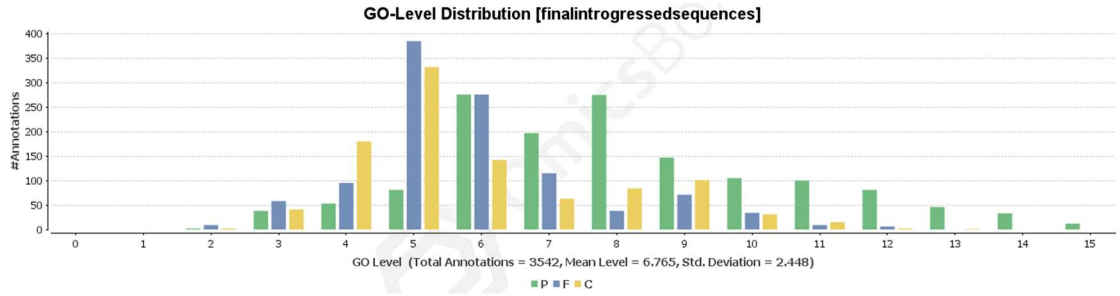


Figure D4: GO level vs the number of annotations for each aspect of gene ontology. P – Biological Process, F – Molecular Function, C – Cellular Component

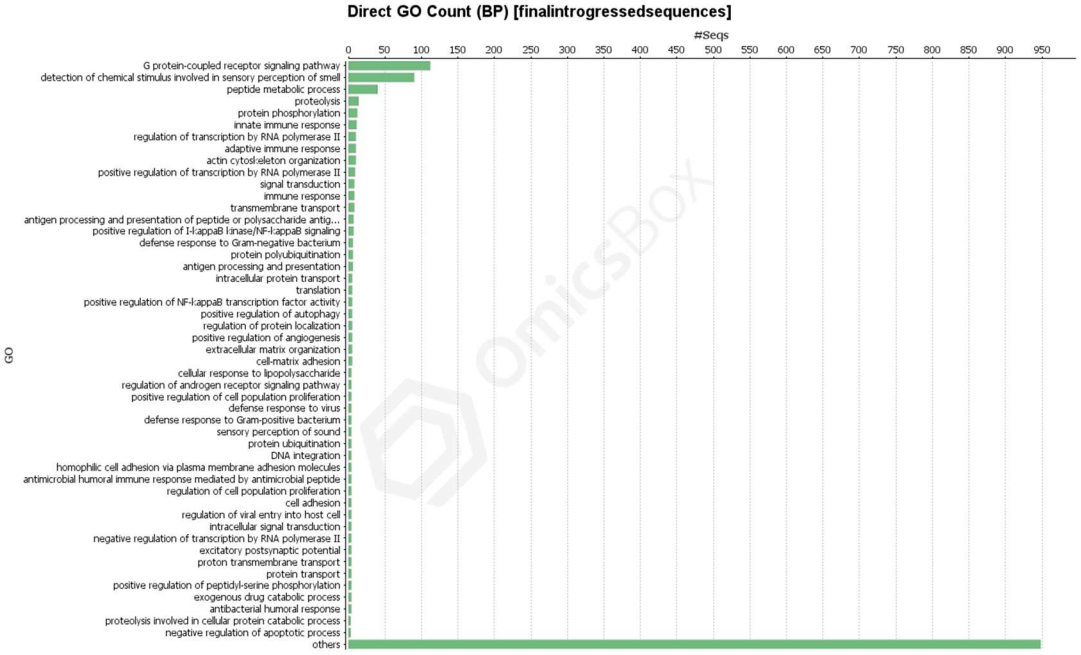


Figure D5: Distribution of 788 Biological Process (BP) GO-terms assigned explicitly and implicitly for the sequences.

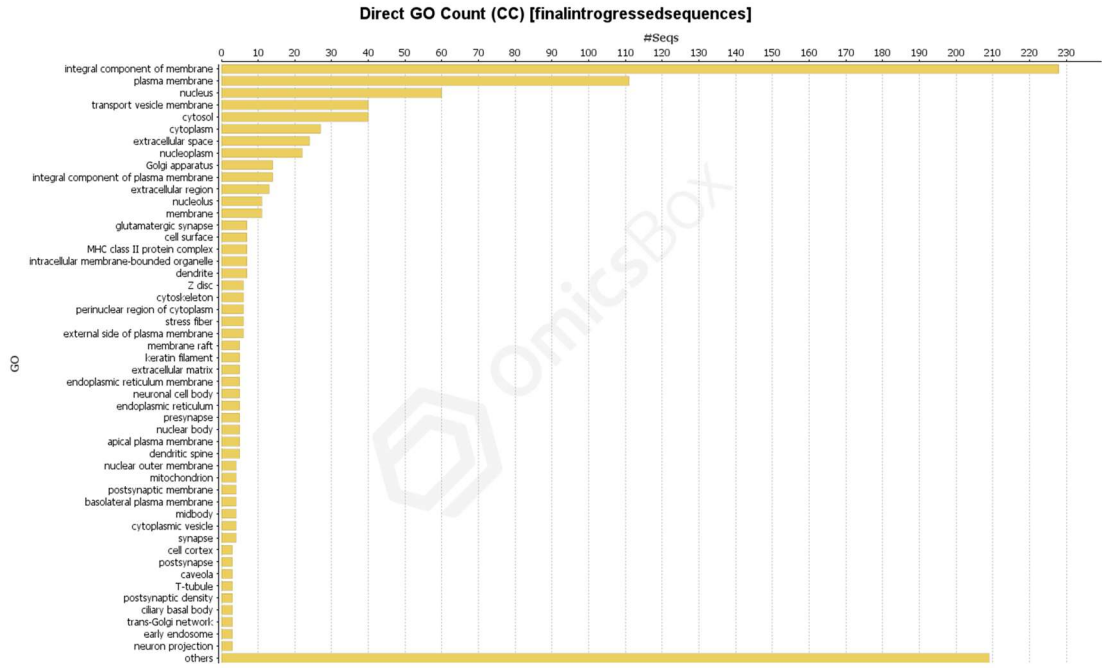


Figure D6: Distribution of 205 Cellular Component (CC) GO-terms assigned explicitly and implicitly for the sequences.

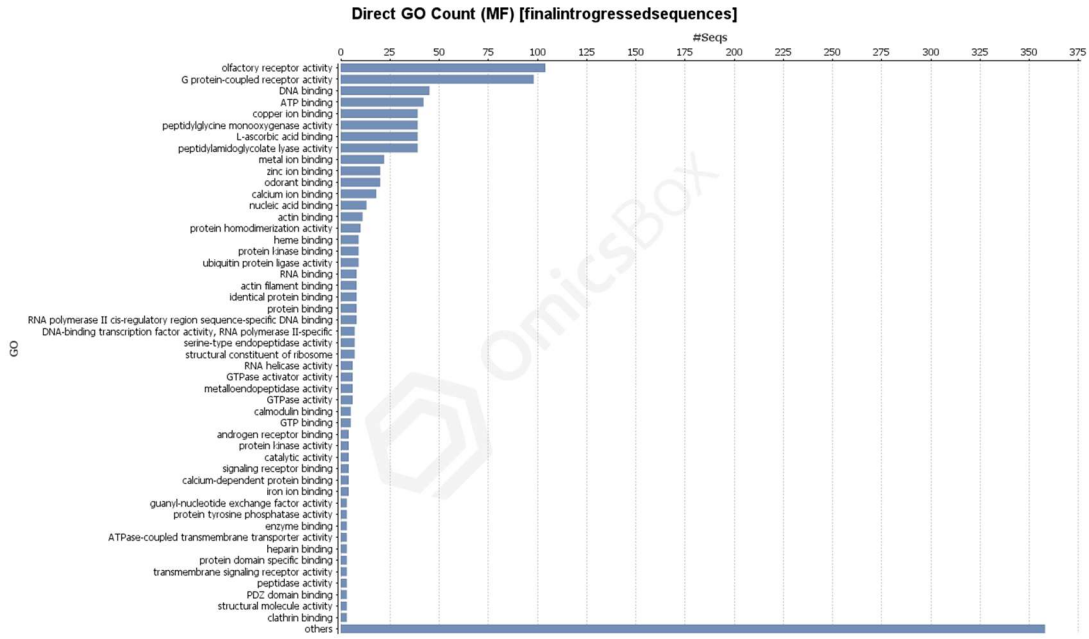


Figure D7: Distribution of 320 Molecular Function (MF) GO-terms assigned explicitly and implicitly for the sequences.

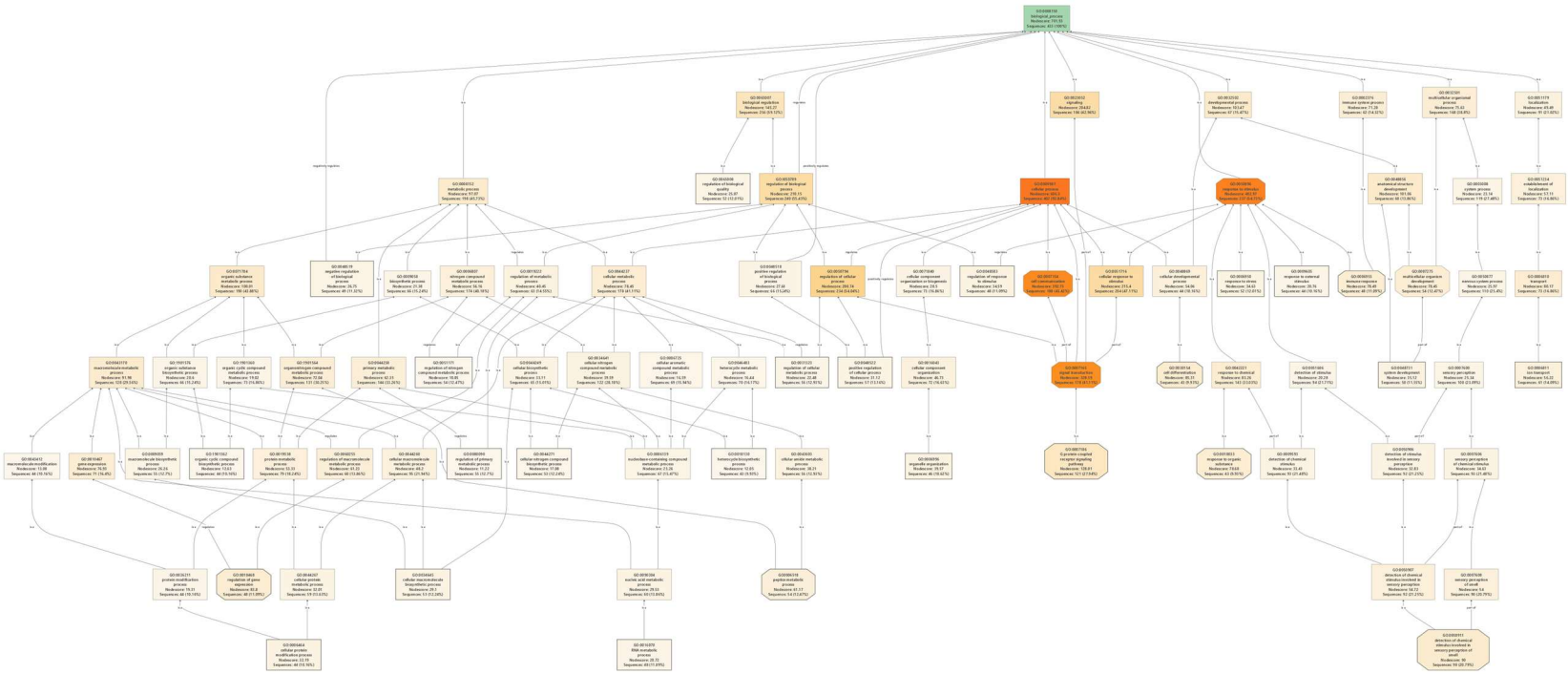


Figure D8: The GO graph showing relationships between 82 explicit BP GO-terms annotated against 433 sequences.

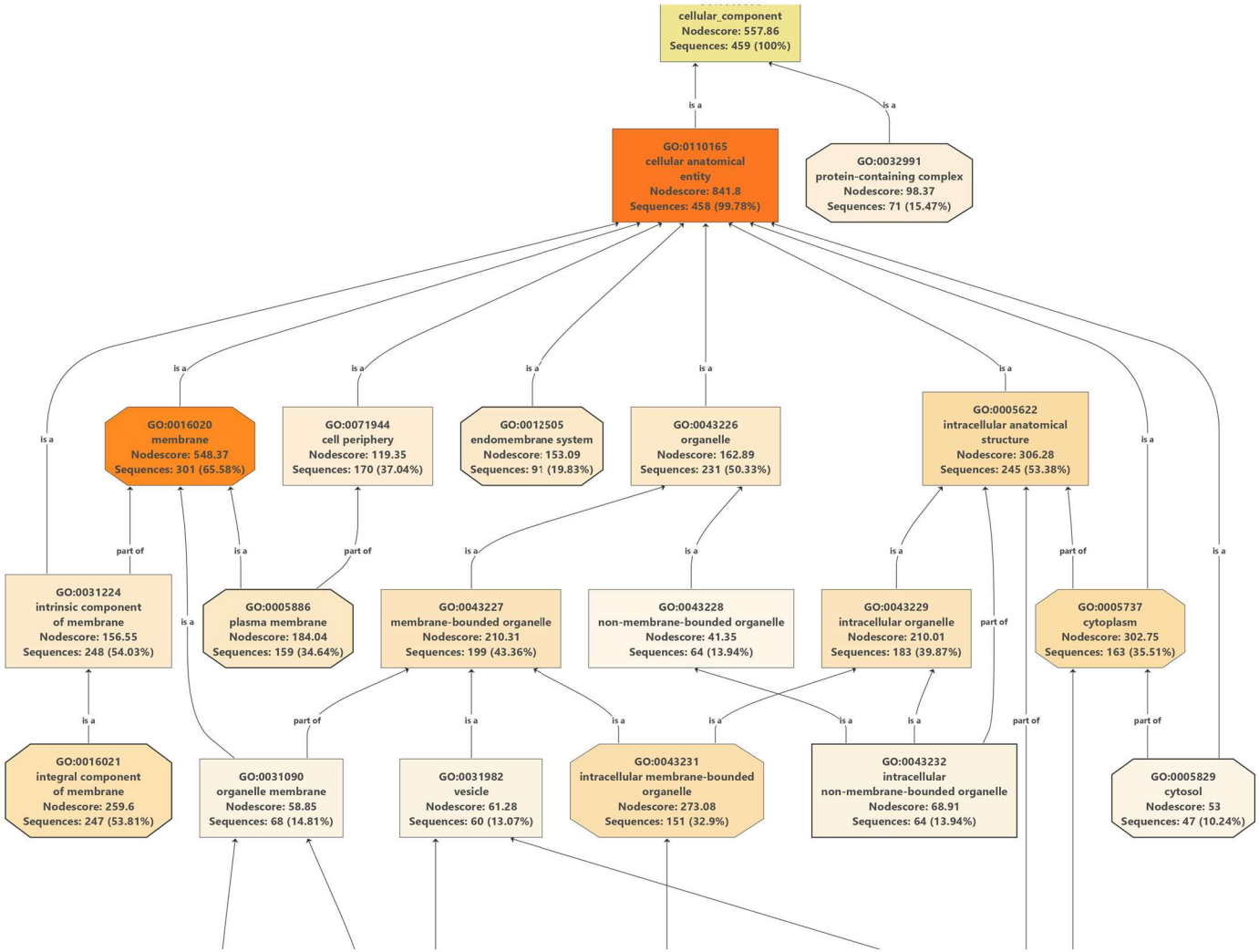


Figure D9: The GO graph showing relationships between 24 explicit CC GO-terms annotated against 459 sequences.

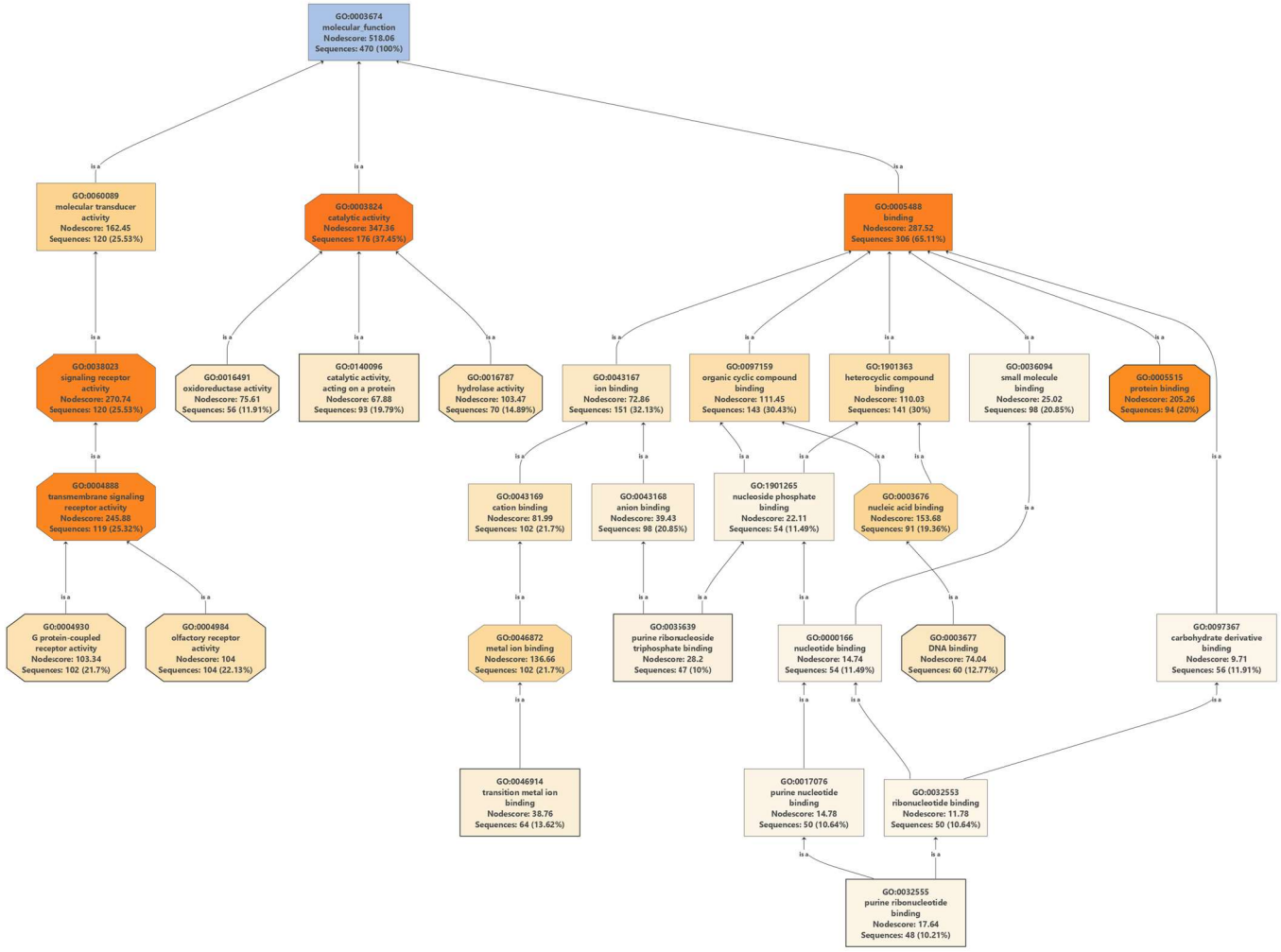


Figure D10: The GO graph showing relationships between 24 explicit CC GO-terms annotated against 459 sequences.

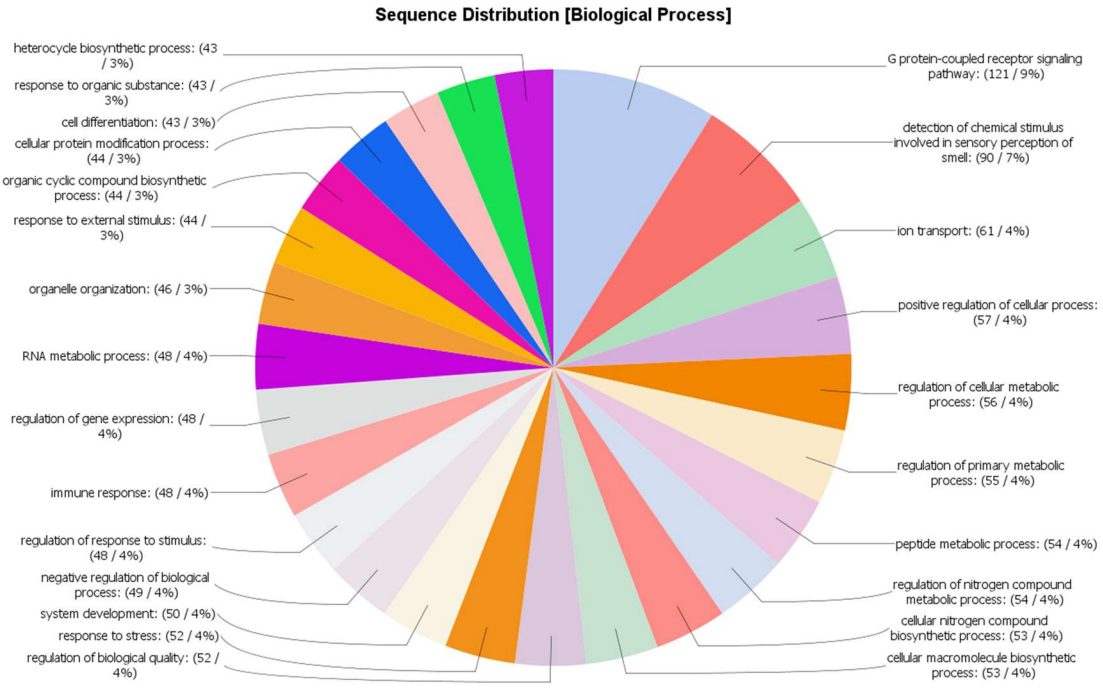


Figure D11: Multi-level pie chart generated from the number of sequences in the lowest node per branch of the Biological Process (BP) GO graph

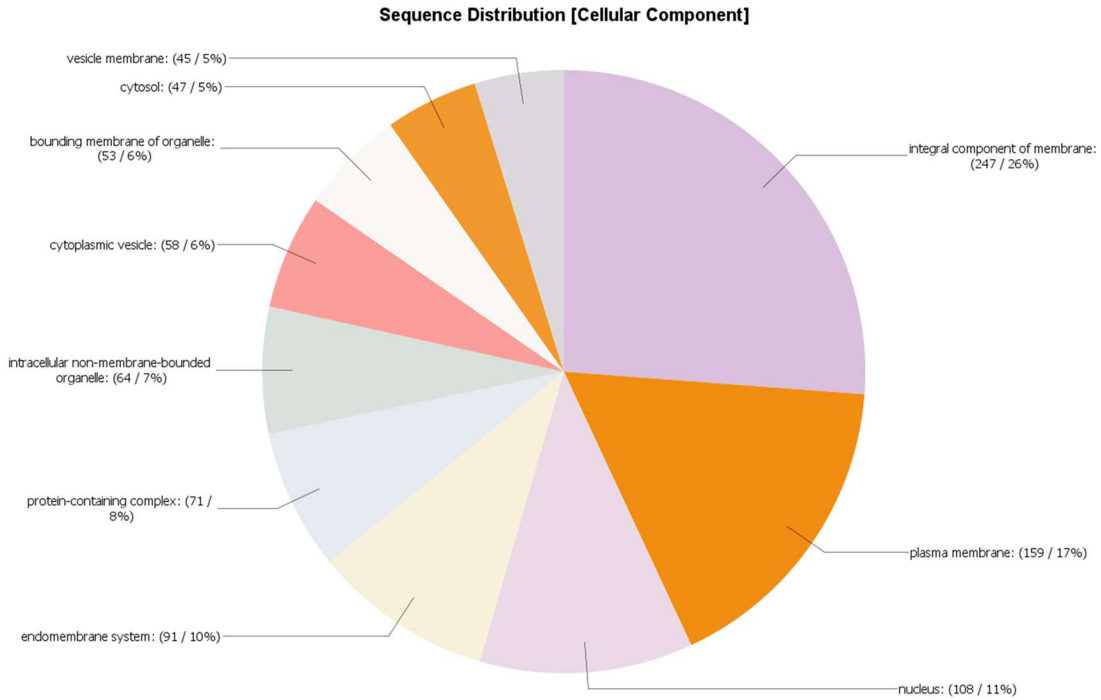


Figure D12: Multi-level pie chart generated from the number of sequences in the lowest node per branch of the Cellular Components (CC) GO graph

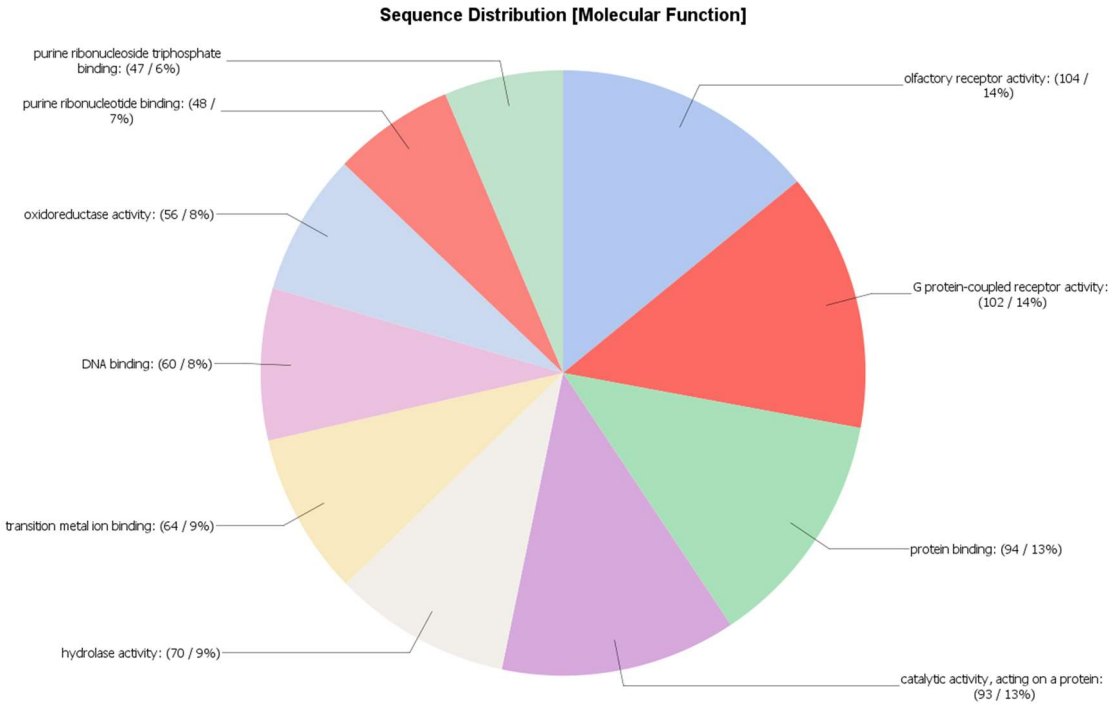


Figure D13: Multi-level pie chart generated from the number of sequences in the lowest node per branch of the Molecular Function (MF) GO graph

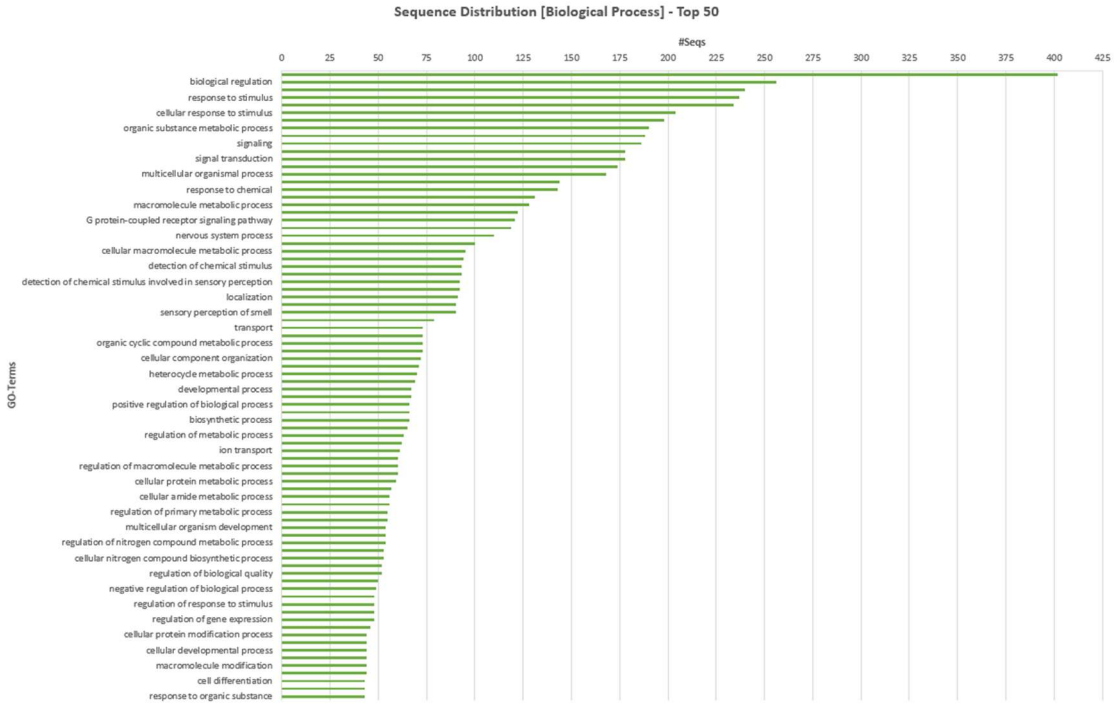


Figure D14: Distribution of specific 82 Biological Process (BP) GO-terms assigned explicitly for the sequences.

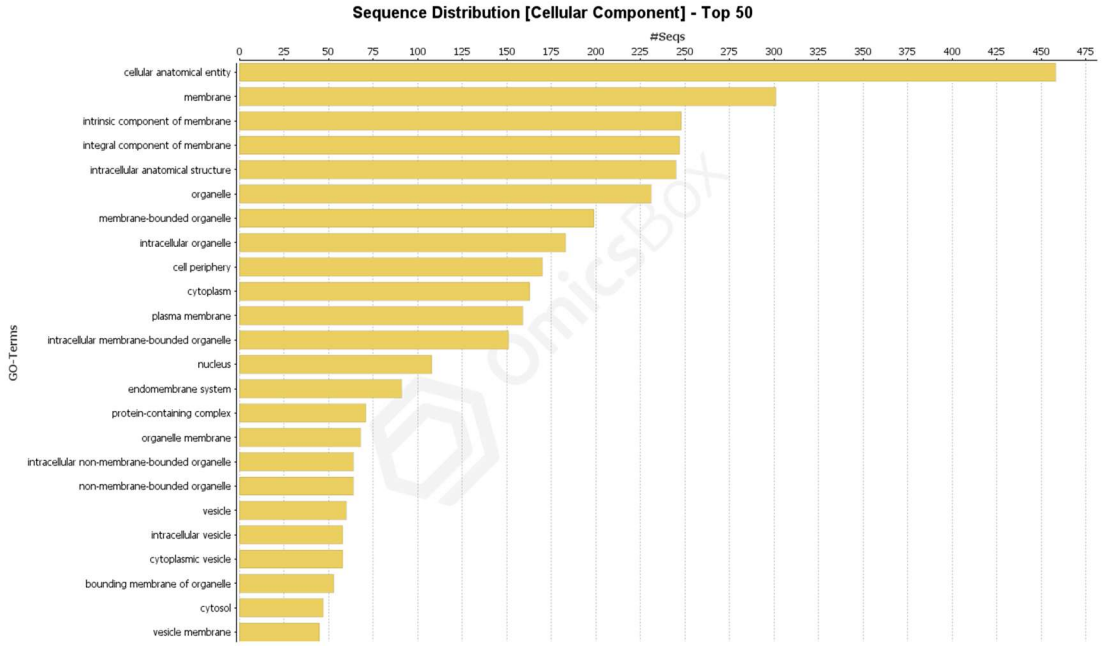


Figure D15: Distribution of specific 24 Cellular Components (CC) GO-terms assigned explicitly for the sequences.

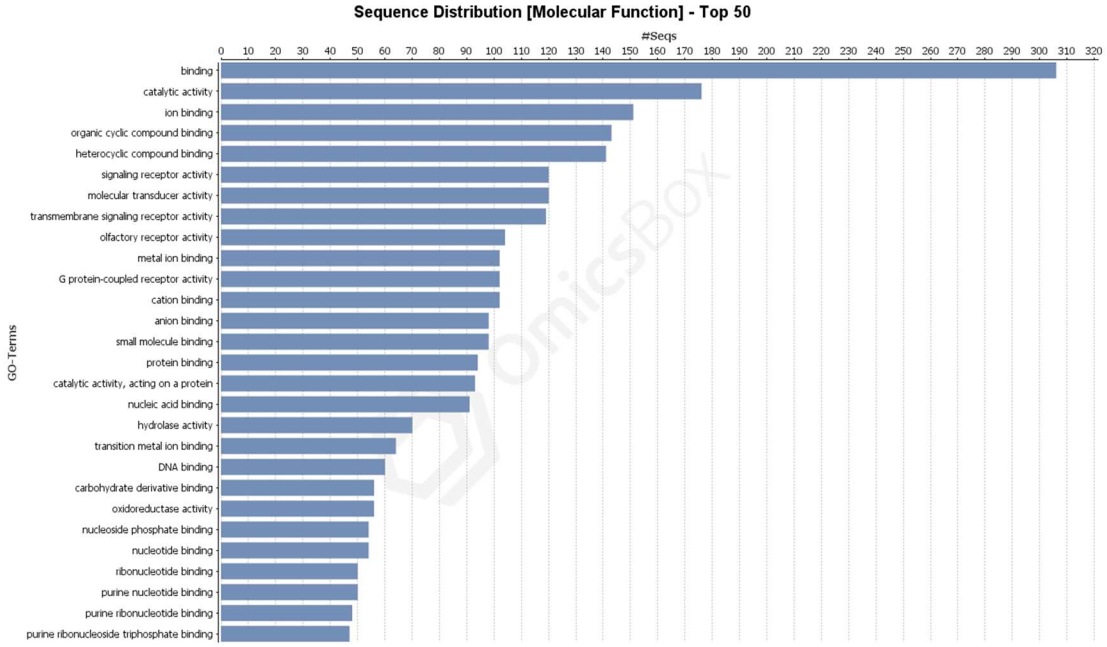


Figure D16: Distribution of specific 28 Molecular Function (MF) GO-terms assigned explicitly for the sequences.

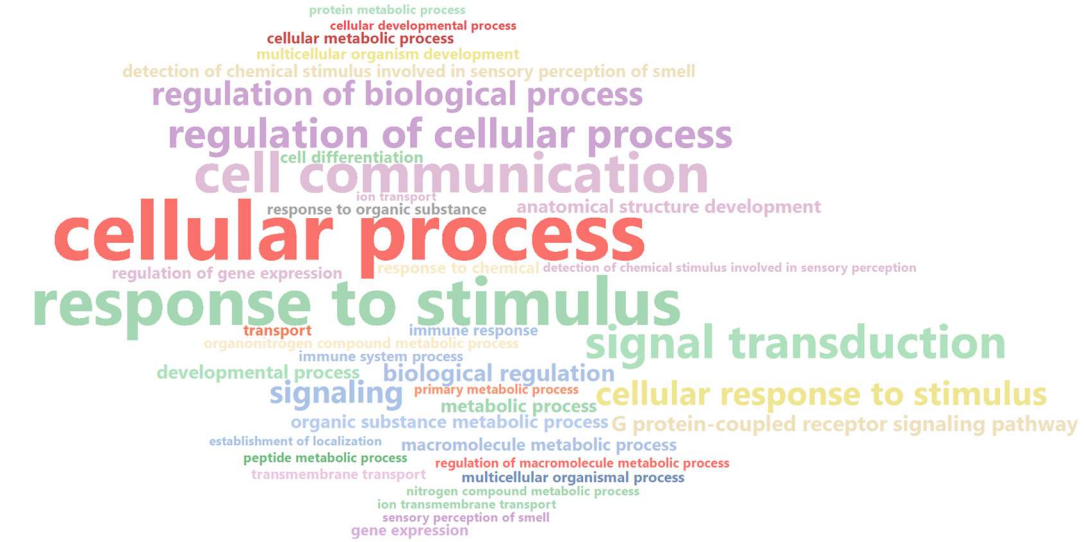


Figure D17: Top 50 Biological Process (BP) GO-terms assigned to annotated sequences

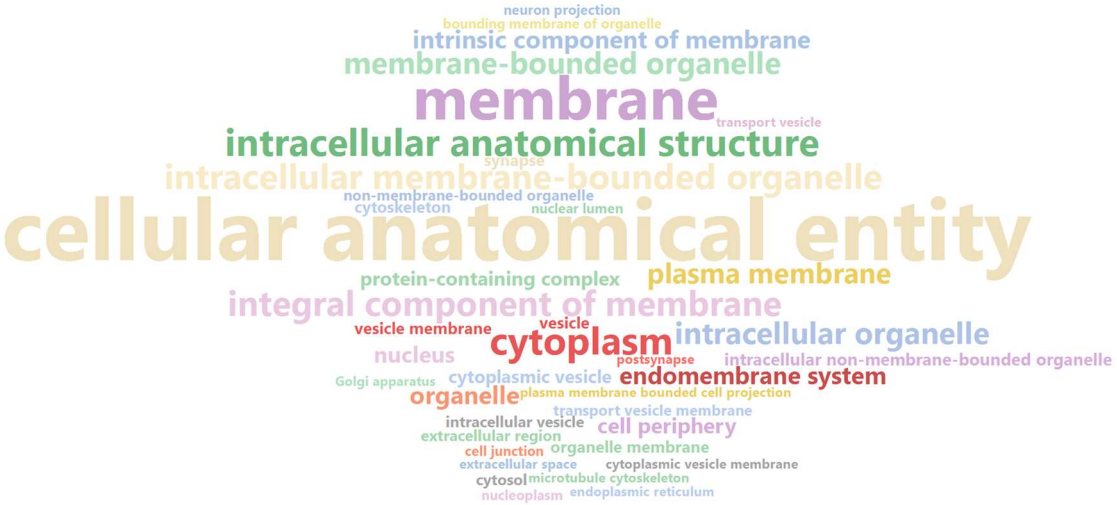


Figure D18: Top 50 Cellular Components (CC) GO-terms assigned to annotated sequences

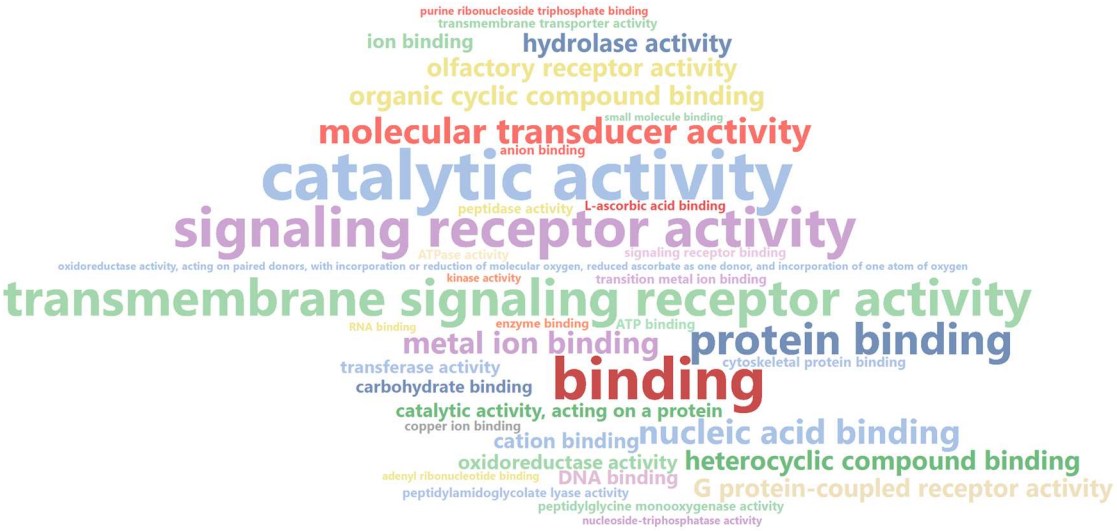


Figure D19: Top 50 Molecular Function (MF) GO-terms assigned to annotated sequences

Table D1: Enzyme Codes (EC) assigned for the annotated sequences

EC SubClasses	#Seqs
1.-.- Oxidoreductases	56
1.1.- Acting on the CH-OH group of donors	1
1.2.- Acting on the aldehyde or oxo group of donors	1
1.4.- Acting on the CH-NH(2) group of donors	1
1.6.- Acting on NADH or NADPH	4
1.9.- Acting on a heme group of donors	2
1.11.- Acting on a peroxide as acceptor	1
1.14.- Acting on paired donors, with incorporation or reduction of molecular oxygen. The oxygen incorporated need not be derived from O(2)	46
1.16.- Oxidizing metal ions	2
2.-.- Transferases	44
2.1.- Transferring one-carbon groups	1
2.3.- Acyltransferases	3
2.4.- Glycosyltransferases	5
2.7.- Transferring phosphorus-containing groups	20
2.8.- Transferring sulfur-containing groups	3
3.-.- Hydrolases	70
3.1.- Acting on ester bonds	12

3.2.- Glycosylases	1
3.4.- Acting on peptide bonds (peptidases)	26
3.5.- Acting on carbon-nitrogen bonds, other than peptide bonds	1
3.6.- Acting on acid anhydrides	30
3.7.- Acting on carbon-carbon bonds	1
4.-.- Lyases	41
4.1.- Carbon-carbon lyases	1
4.3.- Carbon-nitrogen lyases	39
4.6.- Phosphorus-oxygen lyases	1
5.-.- Isomerases	1
5.2.- Cis-trans-isomerases	1
5.6.- Isomerases altering macromolecular conformation	1
6.-.- Ligases	2
6.1.- Forming carbon-oxygen bonds	1
6.3.- Forming carbon-nitrogen bonds	1
7.-.- Translocases	33
7.1.- Catalysing the translocation of hydrons	2
7.2.- Catalysing the translocation of inorganic cations	7
7.4.- Catalysing the translocation amino acids and peptides	2
7.6.- Catalysing the translocation of other compounds	1

Table D2: KEGG metabolic pathways assigned for the annotated sequences

Pathway	Pathway ID	#Enzs in Pathway	Enzyme	#Seqs of Enzyme
Purine metabolism	map00230	4	ec:3.6.1.15 - phosphatase, ec:3.5.3.4 - ec:3.5.3.4 allantoicase, ec:4.6.1.1 - cyclase, ec:2.4.2.1 - phosphorylase	30, 1, 1, 1
Alanine, aspartate and glutamate metabolism	map00250	2	ec:3.4.17.21 - carboxypeptidase II, ec:6.3.5.5 - synthase (glutamine-hydrolysing)	3, 1
Drug metabolism - other enzymes	map00983	2	ec:3.1.1.1 - ali-esterase, ec:2.4.1.17 - 1-naphthol glucuronyltransferase	2, 2
Arginine and proline metabolism	map00330	3	ec:1.14.13.39 - synthase (NADPH), ec:1.4.3.3 - oxidase, ec:3.4.11.5 - aminopeptidase	2, 1, 2
alpha-Linolenic acid metabolism	map00592	2	ec:3.1.1.4 - A2, ec:3.1.1.32 - A1	2, 1

Arachidonic acid metabolism	map00590	2	ec:3.1.1.4 - A2, ec:1.14.99.1 - synthase	2, 1
Folate biosynthesis	map00790	2	ec:3.4.19.9 - gamma-glutamyl hydrolase, ec:1.1.1.21 - reductase	2, 1
Pentose and glucuronate interconversions	map00040	3	ec:1.1.1.2 - dehydrogenase (NADP+), ec:2.4.1.17 - 1-naphthol glucuronyltransferase, ec:1.1.1.21 - reductase	1, 2, 1
Glycerophospholipid metabolism	map00564	3	ec:2.7.8.5 - 1-phosphatidyltransferase, ec:3.1.1.4 - A2, ec:3.1.1.32 - A1	1, 2, 1
Steroid hormone biosynthesis	map00140	2	ec:1.14.15.6 - monooxygenase (side-chain-cleaving), ec:2.4.1.17 - 1-naphthol glucuronyltransferase	1, 2
Drug metabolism - cytochrome P450	map00982	2	ec:1.2.3.1 - oxidase, ec:2.4.1.17 - 1-naphthol glucuronyltransferase	1, 2
Retinol metabolism	map00830	2	ec:1.2.3.1 - oxidase, ec:2.4.1.17 - 1-naphthol glucuronyltransferase	1, 2
Ascorbate and aldarate metabolism	map00053	2	ec:1.1.1.2 - dehydrogenase (NADP+), ec:2.4.1.17 - 1-naphthol glucuronyltransferase	1, 2
Fructose and mannose metabolism	map00051	6	ec:2.7.1.7 - mannokinase (phosphorylating), ec:3.1.3.46 - 2-phosphatase, ec:2.7.1.105 - phosphofructokinase 2, ec:2.7.1.1 - hexokinase type IV glucokinase, ec:1.1.1.21 - reductase, ec:2.7.1.4 - fructokinase (phosphorylating)	1, 1, 1, 1, 1, 1
Nicotinate and nicotinamide metabolism	map00760	4	ec:2.7.7.18 - adenylyltransferase, ec:1.2.3.1 - oxidase, ec:2.7.7.1 - adenylyltransferase, ec:2.4.2.1 - phosphorylase	1, 1, 1, 1

Amino sugar and nucleotide sugar metabolism	map00520	4	ec:2.7.1.7 - mannokinase (phosphorylating), ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase, ec:2.7.1.4 - fructokinase (phosphorylating)	1, 1, 1, 1
Glycolysis / Gluconeogenesis	map00010	3	ec:2.7.1.2 - glucokinase (phosphorylating), ec:1.1.1.2 - dehydrogenase (NADP+), ec:2.7.1.1 - hexokinase type IV glucokinase	1, 1, 1
Galactose metabolism	map00052	3	ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase, ec:1.1.1.21 - reductase	1, 1, 1
Starch and sucrose metabolism	map00500	3	ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase, ec:2.7.1.4 - fructokinase (phosphorylating)	1, 1, 1
Tryptophan metabolism	map00380	2	ec:3.7.1.3 - ec:3.7.1.3 kynureninase, ec:1.2.3.1 - oxidase	1, 1
Neomycin, kanamycin and gentamicin biosynthesis	map00524	2	ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase	1, 1
Glycerolipid metabolism	map00561	2	ec:1.1.1.2 - dehydrogenase (NADP+), ec:1.1.1.21 - reductase	1, 1
Phosphatidylinositol signaling system	map04070	2	ec:2.7.1.137 - 3-kinase, ec:3.1.3.66 - 4-phosphatase	1, 1
Inositol phosphate metabolism	map00562	2	ec:2.7.1.137 - 3-kinase, ec:3.1.3.66 - 4-phosphatase	1, 1
Streptomycin biosynthesis	map00521	2	ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase	1, 1

Pyrimidine metabolism	map00240	2	ec:6.3.5.5 - synthase (glutamine-hydrolysing), ec:2.4.2.1 - phosphorylase	1, 1
Thiamine metabolism	map00730	1	ec:3.6.1.15 - phosphatase	30
Human immunodeficiency virus 1 infection	map05170	1	ec:2.7.11.1 - serine/threonine protein kinase	2
Thermogenesis	map04714	1	ec:2.7.11.1 - serine/threonine protein kinase	2
Arginine biosynthesis	map00220	1	ec:1.14.13.39 - synthase (NADPH)	2
Ether lipid metabolism	map00565	1	ec:3.1.1.4 - A2	2
Metabolism of xenobiotics by cytochrome P450	map00980	1	ec:2.4.1.17 - 1-naphthol glucuronyltransferase	2
Relaxin signaling pathway	map04926	1	ec:2.7.11.1 - serine/threonine protein kinase	2
mTOR signaling pathway	map04150	1	ec:2.7.11.1 - serine/threonine protein kinase	2
Human papillomavirus infection	map05165	1	ec:2.7.11.1 - serine/threonine protein kinase	2
Oxidative phosphorylation	map00190	1	ec:7.1.1.9 - oxidase	2
Human cytomegalovirus infection	map05163	1	ec:2.7.11.1 - serine/threonine protein kinase	2
Linoleic acid metabolism	map00591	1	ec:3.1.1.4 - A2	2
Porphyrin and chlorophyll metabolism	map00860	1	ec:2.4.1.17 - 1-naphthol glucuronyltransferase	2
PI3K-Akt signaling pathway	map04151	1	ec:2.7.11.1 - serine/threonine protein kinase	2
D-Arginine and D-ornithine metabolism	map00472	1	ec:1.4.3.3 - oxidase	1
Caprolactam degradation	map00930	1	ec:1.1.1.2 - dehydrogenase (NADP+)	1

Pyruvate metabolism	map00620	1	ec:1.1.1.2 - dehydrogenase (NADP+)	1
Vitamin B6 metabolism	map00750	1	ec:1.2.3.1 - oxidase	1
Glycosphingolipid biosynthesis - ganglio series	map00604	1	ec:2.4.99.4 - alpha-2,3-sialyltransferase	1
Valine, leucine and isoleucine degradation	map00280	1	ec:1.2.3.1 - oxidase	1
Aminoacyl-tRNA biosynthesis	map00970	1	ec:6.1.1.19 - ligase	1
Tyrosine metabolism	map00350	1	ec:1.2.3.1 - oxidase	1
Penicillin and cephalosporin biosynthesis	map00311	1	ec:1.4.3.3 - oxidase	1
Mucin type O-glycan biosynthesis	map00512	1	ec:2.4.99.4 - alpha-2,3-sialyltransferase	1
Glycosaminoglycan biosynthesis - keratan sulfate	map00533	1	ec:2.4.99.4 - alpha-2,3-sialyltransferase	1
Cysteine and methionine metabolism	map00270	1	ec:2.4.2.28 - phosphorylase	1
Glycine, serine and threonine metabolism	map00260	1	ec:1.4.3.3 - oxidase	1
Glycosaminoglycan degradation	map00531	1	ec:3.2.1.35 - hyaluronidase	1
T cell receptor signaling pathway	map04660	1	ec:2.7.10.2 - protein-tyrosine kinase	1
Glycosphingolipid biosynthesis - globo and isoglobo series	map00603	1	ec:2.4.99.4 - alpha-2,3-sialyltransferase	1
Glycosphingolipid biosynthesis - lacto and neolacto series	map00601	1	ec:2.4.99.4 - alpha-2,3-sialyltransferase	1

CURRICULUM VITAE

KALPANI DE SILVA

775 Theodore Burnett CT APT 2, Louisville, KY 402017
kmkdesilva@gmail.com 502 242 8621

EDUCATION

- **PhD in Interdisciplinary Studies:** **2015-2021**
  Specialization in Bioinformatics (GPA: 3.675/4.00)
  University of Louisville, KY, USA

- **B.Sc. (Hon's) in Bioinformatics** **2014**
  Second Class upper division (GPA: 3.34/4.00)
  University of Colombo, Colombo 03, Sri Lanka

- **Higher Diploma in Information Technology** (GPA: 3.25/4.00) **2014**
  University of Colombo School of Computing, Colombo 03, Sri Lanka

- **Diploma in Information Technology** (GPA: 2.47/4.00) **2013**
  University of Colombo School of Computing, Colombo 03

- **General Certificate in Education, Advance Level** **2008**
  Physics A, Chemistry B, Biology B, English A ***Dharmasoka College, Ambalangoda,***
  Sri Lanka

PROFESSIONAL EXPERIENCE

- **Research Assistant** **2019-2021**
  Computer Engineering and Computer Science Department
  Speed School of Engineering
  University of Louisville, KY, USA

- **Graduate Fellow** **2015-2019**
  Interdisciplinary Studies: Specialization in Bioinformatics
  School of Interdisciplinary and Graduate Studies
  University of Louisville, KY, USA

- **Teaching/Administrative experience**
  Graduate Teaching Assistant Academy **2019**
  University of Louisville, KY, USA

  Teaching Assistant at Department of Plant Sciences **2014**
  University of Colombo, Sri Lanka

- **Industrial training** **2013**
  Research Assistant, Bioactivity Laboratory, Herbal Technology Section
  Industrial Technology Institute (ITI), Sri Lanka

RESEARCH EXPERIENCE

- **Graduate Research Projects** **2015-2021**
  - Investigating Ancient Introgression between Caballine and Non-caballine Equids
  - Comparative genomics study on FAANG species
  - Identifying mutational load in North American Cattle
  - Extracting a set of high-quality Bovine SNPs to use as known variants for Base Quality Recalibration

  Supervision Dr. Theodore S. Kalbfleisch (*Gluck Equine Research Center, University of Kentucky*)

- **Undergraduate Research** **2013**
  Resolving phylogenetic relationships of flameback woodpeckers in South Asia using GBS/RAD data generated with high throughput Next Generation Sequencing (NGS)
  Supervision Dr. Sampath S. Seneviratne (*Department of Zoology, University of Colombo*)

SKILLS AND COMPETENCIES

- Ability to work in High Performance Computing cluster
- Experience in working with Next Generation Sequencing data (large data sets)
- Programming skills in Python, R, SAS and Bash
- Experience with packages numpy, scipy, scikit-allel, pandas, ggplot, matplotlib
- Proficiency in PowerPoint & Excel
- Good verbal and written communication skills
- Ability and willingness to work with a team

CONFERENCE PRESENTATIONS

Oral presentations

- **Kalpani de Silva,** Ernest Bailey, Theodore S. Kalbfleisch (2019). **Investigating ancient introgression between Caballine and Non-caballine equids, Southeast Regional IDeA Conference** held in Louisville, KY, USA from 6-8, November 2019

- **Kalpani de Silva,** Ernest Bailey, Theodore S. Kalbfleisch (2019). **Identification of shared and species-specific k-mers in Equids and Caballines to characterize adaptive introgression events, 37th International Society for Animal Genetics Conference** held in Lleida, Spain from 7-12, July 2019

- **Kalpani de Silva** (2014). **Biogeographic history of Flameback woodpeckers in Sri Lanka**. **8th Annual Bird Watchers' Conference. Field Ornithology Group of Sri Lanka** (FOGSL), Colombo, Sri Lanka.

Poster presentations

- **Kalpani de Silva,** Ernest Bailey, Theodore S. Kalbfleisch (2020). **A likelihood estimate method for detecting introgressed alleles from non-caballine equids in horses**, **Plant and Animal Genome XXVIII (PAG) Conference** held in San Diego, CA, USA from 11-15, January 2020

- **Kalpani de Silva,** Ernest Bailey, Theodore S. Kalbfleisch (2019). **Identification of shared and species-specific k-mers in Equids and Caballines to characterize adaptive introgression events, 37th International Society for Animal Genetics Conference** held in Lleida, Spain from 7-12, July 2019

- **Kalpani de Silva,** Ernest Bailey, Theodore S. Kalbfleisch (2019). **Identify Shared and species-specific k-mers in Equids and Caballines for identification of Evolutionary Features**, **Plant and Animal Genome XXVII (PAG) Conference** held in San Diego, CA, USA from 12-16, January 2019.

- **Kalpani de Silva,** Ernest Bailey, Joel Claiborne Stephens, Theodore S. Kalbfleisch (2018). **Investigating Ancient Introgression Between Caballine and Non-Caballine Equids**, **Plant and Animal Genome XXVI (PAG) Conference** held in San Diego, CA, USA from 13-17, January 2018.
- **Kalpani de Silva**, Ernest Bailey, Joel Claiborne Stephens, Theodore S. Kalbfleisch (2017). **Analysis in modern horses for non-caballine introgression. 16th Annual UT-ORNL-KBRIN Bioinformatics Summit 2017**, held in Burns, TN, USA from 21-23, April 2017.

- **Kalpani de Silva**, Theodore S. Kalbfleisch & Charles T. Robbins (2017). **Rudimentary Genomic Reference for the Grizzly Bear derived using Short Read data from several Animals. Plant and Animal Genome XXIV (PAG) Conference** held in San Diego, CA, USA from 08 - 13, January 2017.

- **Kalpani de Silva,** Ernest Bailey, Theodore S. Kalbfleisch (2017). **EquCab3 Viewed through the Lens of Comparative Genomics for Non-Caballene Species**, **Plant and Animal Genome XXIV (PAG) Conference** held in San Diego, CA, USA from 08-13, January 2017.

- **Kalpani de Silva**, Michael P. Heaton, Theodore S. Kalbfleisch (2016). **Identification of conserved genomic regions and variation therein amongst sixteen Cetartiodactyla species using Next Generation Sequencing. 15th Annual UT-KBRIN Bioinformatics Summit 2016,** held in Cadiz, KY, USA from 8-10 April 2016.

- **Kalpani de Silva,** Poornima S. Dodangoda, Sampath S. Senevirathne (2014). **A Peculiar biogeographic history for a flameback woodpecker (*Dinopium benghalense*) revealed through high throughput sequencing. 51st Annual Meeting of the Association for Tropical Biology and Conservation (ATBC) 2014**, held in Cairns, Australia from 20-24 July 2014.

AWARDS

- **Recognition for student involvement**, Dean's Reception **2020, 2019**
  University of Louisville

- **ISAG/IFAG Travel Award**, *International Society for Animal Genetics* **2019**

- **Nichols Professional Development Award**, *Women's Center,* **2018**
  University of Louisville

- **Graduate Student Council Travel Award**, **2019, 2018, 2017**
  University of Louisville

- **Horse Genome Coordinator Fund Travel Award**, **2019, 2018, 2017**
  University of Kentucky

- **Graduate School Fellowship**, *University of Louisville* **2015 to 2019**

UNIVERSITY SERVICES

- **Vice President**, Sri Lankan Student Association **2020-2021**
  University of Louisville

- **GS Graduate Student Ambassador,** *University of Louisville* **2019-2021**
  Represent and assist Graduate School with graduate student orientations, graduate diversity welcome reception, visitation day, doctoral hooding ceremony and admissions

- **Graduate Student Council Representative for Interdisciplinary Studies,** **2019-2021**
  University of Louisville
  Communicate with students to serve their needs by offering travel grants, research grants and planning events (Welcome breakfast, Halloween party, Graduate student trivia night, Graduate student regional conference) for graduate students.

- **Organizing committee member** **2020**
  GSRRC (Graduate Student Regional Research Conference), *University of Louisville*

- **Executive board member/Parliamentarian** **2019-2020**
  Multicultural Association for Graduate Students, *University of Louisville*