이학박사학위논문

Identification of adaptive signatures and genomic features in response to selective pressure in mammals

포유류 유전체 내 선택압에 의한 적응 흔적 및 특성 발굴

2019 년 8 월

서울대학교 대학원

협동과정 생물정보학과

김 권 도

Identification of adaptive signatures and genomic features in response to selective pressure in mammals

By

Kwondo Kim

Supervisor: Professor Heebal Kim

Aug, 2019

Interdisciplinary Program in Bioinformatics

Seoul National University

포유류 유전체 내 선택압에 의한 적응 흔적 및 특성 발굴

지도교수 김 희 발

이 논문을 이학박사 학위논문으로 제출함

2019 년 6 월

서울대학교 대학원

협동과정 생물정보학과

김 권 도

김권도의 이학박사 학위논문을 인준함

2019 년 6 월

위 원 장 김 선 **(인)**

부위원장 김 희 발 **(인)**

위 원 한 재 용 **(인)**

위 원 조 서 애 **(인)**

위 원 유 재 웅 **(인)**

Abstract

Identification of adaptive signatures and genomic features in response to selective pressure in mammals

Kwondo Kim
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

The central goal of evolutionary biology is to understand the genetic basis of evolutionary processes and adaptive traits. In this regard, the recent advances in sequencing technologies and the explosion of sequence data provide a better opportunity to reach this goal. Various genomic variations are now easily and precisely obtained for large-scale of samples. They are expanding the scope of typical genomic studies, allowing us to take into account diverse evolutionary processes. The aim of this thesis is to demonstrate the applications of such genomic variations while taking into account diverse evolutionary scenarios and time scales. As such, this thesis will fill in the gaps in the knowledge of mammalian genetic background underlying adaptive traits through genome-wide scan and comparative genome analysis.

This thesis consists of five chapters and includes results of genome analysis for detecting evolutionary signatures in three mammal species; pinnipeds, primates, and cattle. The basic background, terminologies and recent example studies related to this thesis were introduced in chapter 1. Chapter 2 and 3 focused on divergence between species (macroevolution), while chapter 4 and 5 focused on polymorphism within species (microevolution).

Pinnipeds are a remarkable group of marine animals with unique adaptations to semi-aquatic life. However, their genomes are poorly characterized. In chapter 2, evolutionary signatures of pinnipeds have been investigated using amino acid substitutions. Novel genome assemblies of 3 pinniped species; *Phoca largha*, *Callorhinus ursinus*, and *Eumetopias jubatus* have been generated. These genome assemblies have been used to detect rapidly evolving genes and substitutions unique to pinnipeds associated with their specificities. As a result, unique substitutions were found within the *TECTA* gene and are likely related to the adaptation to amphibious sound perception in pinnipeds. In addition, several genes (*FASN*, *KCNA5*, and *IL17RA*) containing substitutions specific to pinnipeds were found to be potential candidates of phenotypic convergence in all marine mammals. It indicates the weak link between molecular and phenotypic convergence, and confirms the results of previous studies. This study provides candidate targets for future studies of gene function, as well as backgrounds for convergent evolution of marine mammals.

Humans have the largest brain among extant primates with specialized neuronal connections. However, how the human brain rapidly evolved compared to that of closely related primates is not fully understood. In chapter 3, a genome-wide survey has been performed to find an explanation for the rapid evolution of human brain. Based on the hypothesis that tandem repeats could play a key role in introducing genetic variations due to their unstable nature, a genome-wide survey detected 152 human-specific TRs (HSTR) that have emerged only in the human lineage. The HSTRs are associated with biological functions in brain development and synapse function, and the expression level of HSTR-associated genes in brain tissues was significantly higher in human than in other primates. These results suggest a possibility that *de novo* emergence of TRs might have contributed to the rapid evolution of human brain.

The genetic history of cattle is complex, but contains plentiful information to comprehend mammalian evolutionary process such as domestication, and environmental adaptations. In chapter 4, the genomic influence of recent artificial selection has been examined in the case of Korean native cattle, Hanwoo. Using runs of homozygosity (ROH), an increase of inbreeding for decades has been shown, and at the same time, it has been demonstrated that inbreeding has been of little influence on body weight trait. In chapter 5, admixture between two cattle populations; *Bos taurus*, and *Bos indicus* has been examined in Indigenous African cattle populations., Several evidences based on single nucleotide polymorphism (SNP) support that adaptive admixture is at the root of the success of African cattle’s rapid dispersion across African continent.

The findings in this thesis demonstrated applications of various genomic variations under diverse evolutionary scenarios and time scales, and thus may contribute to the understanding of evolutionary processes in mammals.

Key words: genomic variation, evolution, adaptation, positive selection, introgression, mammal

Student number: 2016-30124

Contents

List of Tables

List of Figures

General Introduction

Evolution is a continuous process, which leaves distinctive footprints on the genome. Using these footprints, evolutionary parameters and histories can be inferred. However, the footprints could be diverse as much as the complexity of evolutionary history. Especially, different molecular markers should be used to detect signatures under different time scales of evolution.

Macroevolution refers to changes on a scale at or above the species level, while microevolution refers to changes in allele frequencies that occurs within a species or population (Reznick and Ricklefs 2009). Macroevolution and microevolution depend on fundamentally identical processes, but occur on different time scales (Dietrich 2010). For this reason, separate approaches are generally employed into the two categories of evolution. For instance, reference genome sequences are directly compared, and then functional changes are inferred based on amino acid changes at a species level. In contrast, allele frequency difference or length of haplotypes are usually considered at a population level using genotype data.

The rates of evolution are not uniform across genome of an organism, as different genomic regions experience different evolutionary process. For example, mutation rates of tandem repeats are 10 to 100,000 times higher than that in other parts of the genome (Legendre, Pochet et al. 2007). In addition, gene flow among populations could significantly increase genetic variation and rates of evolution at a particular genomic region (Verhoeven, Macel et al. 2010).

Although such non-uniform evolutionary rates definitely increase the complexity of evolutionary process, they could be useful in some cases such accelerated evolution or adaptation of a species. For instance, many recent studies attempted to explain rapid adaptation of a species that has been exposed to new environmental challenges by genetic variation from gene flow (Hedrick 2013).

This thesis is largely divided into three categories (Terminologies and recent example studies, macroevolution studies, and microevolution studies). The basic background and recent example studies related to this thesis were introduced in the first category (chapter 1). The second category consist of two macroevolution studies with novel analysis using tandem repeat variation (chapter 3) as well as conventional evolutionary analysis (chapter 2). The third category consist of two microevolution studies with novel analysis for very recent evolutionary event (chapter 4) as well as analysis for gene flow (chapter 5). Each study of latter two categories has been performed based on different genomic variations under different time scales. Given these results, this thesis suggests relevance of each genomic variations corresponding to different time scales of evolution.

Chapter 1. Literature Review

1.1 Genomic variations to detect evolutionary signatures

1.1.1 Overview of genomic variations

Genomic variations or genetic polymorphisms are generally referred to as heritable DNA sequence differences among individuals or populations. They arise through several types of mutations; a change from one type of nucleotide to another, an insertion or a deletion, or a rearrangement of nucleotides (Ismail and Essawi 2012).

Due to its polymorphic nature, genomic variation has been used in a variety of research fields including population genetics, phylogenetics, and forensic science. As it also has close relationship with phenotypic difference or disease, one of the major scientific challenges in human genetics field is decoding the genomic basis of human health and disease (Sharp, Cheng et al. 2006, Frazer, Murray et al. 2009).

Genomic variations can be classified into two broad categories; single nucleotide variants (e.g. SNP) and structural variants (e.g. copy number variation, and translocation). They can be analyzed by conventional methods such as polymerase chain reaction (PCR), gel electrophoresis, and fluorescence in situ hybridization (FISH). However, the recent development of new sequencing technologies has provided an opportunity to directly examine the sequences of genomic variations (Frazer, Murray et al. 2009).

1.1.2 Single nucleotide polymorphism (SNP)

Single nucleotide polymorphism (SNP) is a single nucleotide variant that is caused by a point mutation at a particular nucleotide site (Sharp, Cheng et al. 2006). SNPs are the most abundant type of genomic variation in the human genome. On the basis of sequencing results, SNP has been estimated to occur at 1 out of every 1,000 bases in the human genome (Syvänen 2001).

SNP information of an individual can be produced by genotyping or using sequence-based methods. Generally, genotype-based methods are cost-effective, but only previously defined loci can be assayed. Sequencing-based methods, in the other hand, cost more than genotyping methods but they can expand the range of loci to whole genome scale.

The vast majority of SNPs are located outside of genic regions with minor effects for phenotypic variations. However, sometimes, SNPs are located within coding regions of a gene, which can lead to several types of changes for its protein products. In detail, a synonymous SNP, or silent mutation does not result in a change of amino acid due to the degeneracy of the genetic code, whereas a substitution from one amino acid to the other occurs for a nonsynonymous SNP, or missense mutation. If the nonsynonymous SNP result in a premature termination of polypeptide by introducing an internal stop codon, it is called nonsense mutation (Frazer, Murray et al. 2009).

SNP is a major type of genomic variation that is widely used in a variety of current research fields including population genetics, phylogenetics, and evolutionary genetics. In population genetics, it is often used to infer population structure or history, and to detect positive selection signatures based on linkage disequilibrium (Sabeti, Varilly et al. 2007) or allele frequency spectrum (Chen, Patterson et al. 2010). Evolutionary genetics, in particular, used it as a marker for molecular clock (Hasegawa, Kishino et al. 1985) or detecting positive selection signatures based on a ratio of non-synonymous and synonymous SNPs (Nielsen 2005).

1.1.3 Tandem repeats (TR)

Tandem repeats (TR) refers to a repetitive pattern of one or more nucleotides that are connected directly each other (Consortium 2001). Since they were initially detected as a form of satellite bands in density-gradient centrifugal DNA separation, TRs are also known as satellite DNA, and can be divided into microsatellite and minisatellite according to their unit length. Microsatellites is usually used to define short tandem repeats with a unit length shorter than 10 nucleotides, while the unit length of minisatellite is longer their size being up to 100bp (Vergnaud and Denoeud 2000, Gemayel, Vinces et al. 2010).

Traditionally, TRs could be detected by DNA amplification as in human identity testing (Butler 2006), while sequencing technologies now enables to detect TRs across whole genome sequence. Therefore, many TR detection algorithms and software have been developed so far to identify TRs comprehensively for a haploid sequence of whole genome (e.g. Tandem repeat finder (Benson 1999)). Moreover, the recent explosion of sequence data led by next generation sequencing (NGS) facilitates the production of a catalog of TRs at a population level.

TRs are ubiquitously distributed in eukaryotic genomes. Since they are often found in genic regions as well as intergenic regions, TRs could have diverse effects on cellular processes, depending on their locations (Usdin 2008). For example, the expansion of repeats in *HTT* (huntingtin) gene leads to a modification of protein product, which result in Huntington's disease (Hannan 2018). TRs also have higher mutation rates than that in other parts of the genome, as mutations around them is caused by strand slippage or homologous recombination errors rather than point mutations (Gemayel, Vinces et al. 2010).

TRs together with other repeat sequences, have been regarded as junk DNA (Gemayel, Cho et al. 2012). However, the unstable nature of TRs along with their possible biological effects is recently attracting the researchers' attention. According to recent studies, TRs might have more diverse and important roles in gene expression and regulation than expected (Sonay, Carvalho et al. 2015, Gymrek, Willems et al. 2016, Quilez, Guilmatre et al. 2016). Furthermore, TRs give an opportunity to provide insights into evolutionary processes, especially when evolution occurred rapidly as it is the case in human.

1.1.4 Runs of homozygosity (ROH)

Runs of homozygosity (ROH) refers to a continuous DNA segment where all loci are homozygous. Although ROH could be created by chance, ‘autozygous segment’, created by inbreeding is generally a target to study (McQuillan, Leutenegger et al. 2008).

The autozygous segment is generally shortened over generations due to meiotic recombinations (Kirin, McQuillan et al. 2010), and can present a degree of inbreeding in one population. Hence, the frequency and length of ROH could be different across different populations according to their genetic histories (McQuillan, Leutenegger et al. 2008).

Several recent studies have highlighted the link between ROH and phenotypes (Yang, Guo et al. 2010, Keller, Simonson et al. 2012, Ghani, Reitz et al. 2015), although their causal relationship is still not clear. For instance, the deleterious effect of inbreeding on a reproductive trait has been examined by ROH in a pig population (Saura, Fernández et al. 2015). In addition, ROH correlates relatively well with pedigree-based methods to measure inbreeding, and provides inbreeding information at both population and individual-level (Saura, Fernández et al. 2015). In livestock population, controlling the level of inbreeding and maintaining genetic diversity is of major concern to maximize livestock’s productivity. Therefore, ROH could be a promising marker for monitoring inbreeding in livestock population.

1.2 Signatures of positive selection

1.2.1 Overview of positive selection

Positive selection, a type of natural selection, is the force that drives the increase in prevalence of a favored phenotype, causing adaptive evolution of a species (Biswas and Akey 2006). Although the contribution of positive selection could be diverse in each evolutionary process, it is clear that it has played a major role in the evolutionary history of all extant species.

The evidences of positive selection can be found in extant genomes in many different forms. For example, it might induce an excess of derived allele at a locus, and reduce the level of genetic variation compared to a neutral locus (Nielsen 2005, Biswas and Akey 2006). The putative targets of positive selection can be obtained when it is possible to identify such genomic features. A common way of identifying the target is to study the genomic features of a few loci that are hypothesized to have been under positive selection. However, whole genome sequences produced by rapid development of sequencing technologies have enabled obtaining genomic features and the identification of targets under positive selection in a genomic scale.

Although the functional significance is still challenging to verify, the ‘genome scan’ searching whole genome sequence has given numerous candidates worth investigating further. Recently, the field and laboratory experiments combined with genomics directly demonstrated the functional connections between candidate loci, phenotype, and fitness (Barrett, Laurent et al. 2019). Moreover, genome-wide survey facilitate the examination of non-coding genomic regions that have not been extensively studied compared to coding regions (Kelley and Swanson 2008).

1.2.2 Measures to detect positive selection

Many methods have been developed to detect positive selection, however, they could be classified into two fundamental categories; polymorphism-based methods, and divergence-based methods.

Polymorphism-based methods are used to identify recent positive selection within a species. Most of them are focus on identifying the signatures of 'selective sweeps' that are caused by the rapid fixation of an advantageous mutation and by the hitchhiking of beneficial mutations on linked neutral loci. The signatures of selection sweep include 1) an excess of rare alleles, 2) reduced levels of genetic variation, and 3) elevated frequency of long haplotypes relative to neutral expectations.

Tajima's D is a classic example of methods used to detect a skew in the allele frequency distribution caused by the excess of rare alleles. It measures the difference between two measures of genetic diversity; the average number of pairwise differences and the number of segregating sites (Tajima 1989). Significant deviations from zero indicate a skew in the allele frequency distribution relative to neutral expectations, while the expected value is zero under the neutral null model (Benson 1999). The similar tests to Tajima's D include Fu and Li’s D (Fu and Li 1993) and F, and Fay and Wu’s H test (Fay and Wu 2000).

Fst is often used to measure levels of differentiation between subpopulations, whereas, it can also present a signature of positive selection. The differentiation between populations is largely determined by genetic drift. However, when a locus is subjected to positive selection in one of the subpopulations, the allele frequencies around the locus could change rapidly. This leads to an elevated level of population differentiation around the locus, which can be interpreted as a signature of positive selection (Kelley and Swanson 2008). Many *Fst*-based methods have been developed to detect signatures of positive selection. One of them, population branch statistics (PBS) has been devised to detect genetic adaptation to high altitude in Tibetan population. It is computed as population divergence time since the divergence from closely-related population quantified using pairwise *Fst* in three populations. It could be, therefore, interpreted as the amount of allele frequency changes at a given locus in the history of target population (Yi, Liang et al. 2010). The recent extension of *Fst*, FLK and hapFLK, enables the detection of positive selection from multi-population samples, accounting for hierarchical population structure (Fariello, Boitard et al. 2013).

The hitchhiking of linked neutral loci due to recent positive selection also results in an excess of long haplotypes on the selected loci. It is because insufficient time has elapsed to allow recombination to break down long haplotypes. The pattern of long haplotypes around the selected loci can be measured by determining the extended haplotype homozygosity (EHH) (Sabeti, Reich et al. 2002). The iHS (integrated haplotype score) is inferred from EHH, which measures the amount of EHH at a given SNP along the ancestral allele relative to the derived allele (Voight, Kudaravalli et al. 2006). XP-EHH is computed identically to iHS except that it compares EHH between each allele of two populations (Sabeti, Varilly et al. 2007).

Divergence-based methods are used to identify ancient positive selection between species. For protein-coding regions, the discrepancy between the number of nonsynonymous substitutions and the number of synonymous substitutions could give an indicative of positive selection. Therefore, dN/dS-based methods compare the ratio of non-synonymous (dn) to synonymous (ds) substitutions in a protein coding region. This ratio provides information about the evolutionary forces operating on a particular gene; dN/dS = 1 indicates neutrality, dN/dS < 1 indicates functional constraint for non-synonymous substitutions, and dN/dS > 1 indicates positive selection on the gene (Biswas and Akey 2006). Although dN/dS test has generally been performed for single genes, the recent emergence of whole genome sequences from multiple species allows a genome-wide survey of protein evolution. Hudson-Kreitman-Aguade (HKA) test (Hudson, Kreitman et al. 1987) uses both polymorphism and divergence. This tests compares the polymorphism within each species and the divergence between two species at two or more loci. McDonald Kreitman (MK) test (McDonald and Kreitman 1991) is also based on polymorphism and divergence, but accounting for synonymous and non-synonymous sites.

1.2.3 Recent examples of positive selection in mammals

With the advent of next-generation sequencing, numerous genome sequences across multiple species have been generated, leading extending the limit of evolutionary studies. Genome-wide surveys of positive selection already became common across all species, and it is now crucial to enhance the reliability of its outcome and to validate it.

Human population has been most intensively studied to detect signatures of positive selection, as humans have been exposed to different environmental conditions, leading to many adaptations in different environmental conditions. Skin pigmentation is a prominent example of adaptation to environmental conditions in humans. Several genes associated with skin pigmentation have been shown to be subject to selective pressure in different human populations (Lao, De Gruijter et al. 2007, Sturm and Duffy 2012, Martínez-Cadenas, López et al. 2013). Another example of human adaptation can be found in Tibetan, Andean, and Ethiopian populations. In a series of studies, different sets of genes have been detected to be associated with high altitude adaptation in each population (Bigham, Mao et al. 2009, Yi, Liang et al. 2010, Scheinfeldt, Soi et al. 2012). In addition to diversities within human species, distinctive features of human have led to the comparison of human with its closely-related species. For example, positive selection during human brain evolution has been actively studied using different genomic variations (Popesco, MacLaren et al. 2006, Beiter, Khramtsova et al. 2017)

The generation of whole genome sequences from multiple species also allow the detection of positive selection in mammals other than humans. For example, recent studies have attempted to detect positive selection related to domestication in pig (Frantz, Schraiber et al. 2015, Moon, Kim et al. 2015) in order to understand the domestication process that is a main concern in livestock animals. Similar to human population, adaptive signatures to diverse environmental conditions have been identified in various mammal species. For example, positive selection in marine mammals have been detected as they have distinct characteristics such as tolerance to hypoxia, sound perception, and limbs adaptation, that might originate from the second adaptation to aquatic environments (Foote, Liu et al. 2015, Chikina, Robinson et al. 2016). In particular, marine mammals have been also highlighted due to their independent adaptations in each clade, which is called convergent evolution. The convergent evolution, as a form of positive selection, has provided an insight into mammalian evolution in various mammals, as represented by echolocation in bats and dolphins (Parker, Tsagkogeorga et al. 2013), and adaptive pseudothumb in different panda species (Hu, Wu et al. 2017).

1.3 Signatures of introgression

1.3.1 Overview of introgression

Population admixture refers to the process or consequence of hybridization between two or more populations or species (Verhoeven, Macel et al. 2010). If the admixture has a direction by a movement of genetic materials from one specie into the gene pool of another specie via hybridization and backcrossing, it is called ‘introgression or introgressive hybridization’ (Harrison and Larson 2014).

Introgression plays an important role in evolutionary processes, as it enables the introduction of genetic variation faster than through mutation alone. Such introgressed genetic materials are often maintained by natural selection, contributing to the adaptation of a species (Hedrick 2013, Suarez-Gonzalez, Lexer et al. 2018). For this reason, introgression has been consistently used as a tool in breeding crop species (Zhang, Percy et al. 2014). Besides, numerous studies recently highlighted adaptive introgression in natural populations (Frantz, Schraiber et al. 2015).

1.3.2 Measures to detect introgression

The detection of introgression on a genomic sequence is generally straightforward. However, it also requires multiple evidences to exclude other alternative demographic processes such as incomplete lineage sorting, and prove it being adaptive.

Patterson's D or ABBABABA statistic measures an excess of shared alleles between sympatric populations relative to allopatric populations (Durand, Patterson et al. 2011). It could constitute an indicative of introgression, but does not provide information on the amount of admixture. On the other hand, Martin's f statistic, a modified version of the D statistic, has been developed to estimate the genome-wide fraction of admixture, and has been shown to perform better in identitying introgressed loci compared to the D statistic (Martin, Davey et al. 2014).

Ancestry inference has been generally used to examine admixture or introgression among human populations. While numerous methods have been developed to infer ancestry in population data, the majority of them reconstruct high-resolution recombination maps in admixed populations and detect segments introduced by introgression based on probabilistic models such as hidden Markov model (HMM) (Geza, Mugo et al. 2018). As these methods requires biological parameters such as genetic map and admixture generations, it is challenging to employ these methods to non-model species except humans. However, a recent non-probabilistic algorithm based on optimization problem allows inferring ancestry without any biological parameters (Dias-Alves, Mairal et al. 2018).

Identical by descent (IBD) refers to genomic segments that have a same ancestral origin in two or more individuals (Browning 2008). If a particular locus has been introgressed in a target population, this region would share more IBD segments with the sympatric population rather than with the allopatric population. Therefore, IBD can be also used to detect introgression by calculating relative IBD sharing (rIBD) of target population with two reference populations (sympatric and allopatric population) (Bosse, Megens et al. 2014).

1.3.3 Recent examples of introgression in mammals

The relationships between modern and archaic humans have been one of the major concerns in retracing the human history. Especially, a number of recent studies for archaic genomes have provided strong evidences for the introgression of archaic humans genes such as Neanderthal and Denisovan into the modern human genomes (Racimo, Sankararaman et al. 2015). However, the extent and influence of the archaic introgressions on modern humans are still on debate.

In the other mammal species, introgression has been primarily studied as a source of adaptive genetic variation. A series of studies in the butterfly genus *Heliconius* is one important example to observe adaptive introgression related to the mimicry of butterfly (Dasmahapatra, Walters et al. 2012, Martin, Dasmahapatra et al. 2013, Jay, Whibley et al. 2018). The range of species that is targeted in the introgression studies is currently being expanded, giving insights into understanding of the role of introgression in evolution across species. For example, adaptive introgression has been investigated in *Bos* (Medugorac, Graf et al. 2017, Chen, Cai et al. 2018, Wu, Ding et al. 2018) and *Canis* (Miao, Wang et al. 2016) species, suggesting possible contribution of introgression to adaptive traits.

This chapter was published in *Scientific Reports*
as a partial fulfillment of Kwondo Kim's Ph.D program.

Chapter 2. Deciphering the evolutionary signatures of pinnipeds using novel genome sequences: The first genomes of *Phoca largha*, *Callorhinus ursinus*, and *Eumetopias jubatus*

2.1 Abstract

The pinnipeds, which comprise seals, sea lions, and walruses, are a remarkable group of marine animals with unique adaptations to semi-aquatic life. However, their genomes are poorly characterized. In this study, we sequenced and characterized the genomes of three pinnipeds (*Phoca largha*, *Callorhinus ursinus*, and *Eumetopias jubatus*), focusing on site-wise sequence changes. We detected rapidly evolving genes in pinniped lineages and substitutions unique to pinnipeds associated with amphibious sound perception. Phenotypic convergence-related sequence convergences are not common in marine mammals. For example, *FASN*, *KCNA5*, and *IL17RA* contain substitutions specific to pinnipeds, yet are potential candidates of phenotypic convergence (blubber, response to hypoxia, and immunity to pathogens) in all marine mammals. The outcomes of this study will provide insight into targets for future studies of convergent evolution or gene function.

2.2 Introduction

Marine mammals are a classic example of convergent evolution in terms of adaptation of terrestrial mammals to the marine environment. During secondary adaptation to the marine environment, marine mammals experienced similar environmental challenges, which have resulted in shared morphological or physiological features across distant taxa. For instance, they have experienced similar changes in skin and limbs, and subsequently became streamlined (Fish, Howle et al. 2008, Chikina, Robinson et al. 2016). Adaptive traits related to hypoxia are shared features of marine mammals (Andersen 1966, Chikina, Robinson et al. 2016).

Marine mammals include three orders: cetaceans (whales, dolphins, and porpoises), pinnipeds (seals, sea lions, and walruses), and sirenians (manatees and dugongs) (Jefferson, Leatherwood et al. 1993). They have evolved to inhabit the ocean in multiple lineages. Cetaceans and sirenians emerged around 40–50 million years ago (mya) from *Cetartiodactyla* and *Afrotheria*, respectively (Berta, Sumich et al. 2005). Pinnipeds emerged within the *Carnivora* approximately 20 million years later (Berta, Sumich et al. 2005). This implies that different molecular changes occurred across separate lineages, possibly resulting in divergent phenotypic changes. However, most studies related to marine mammals have focused on convergent evolution, although some of the adaptations of marine mammals to an aquatic lifestyle vary among species (Berta, Sumich et al. 2005).

Pinnipeds, which consist of three families (*Phocidae*, *Otariidae*, and *Odobenidae*) are distinguishable from other marine mammals (Berta 2002). Most pinnipeds are semi-aquatic, unlike other marine mammals that spend their entire lives in the water (Jefferson, Leatherwood et al. 1993), and have modified limbs as flippers that propel them both in the water and on land (Rybczynski, Dawson et al. 2009). In addition, with the exception of the walrus, which is the only extant species of the family *Odobenidae*, all pinnipeds have fur coats (Riedman 1990). These distinct characteristics have not been sufficiently characterized at the molecular level. Although a draft fur seal genome has recently been assembled (Humble, Martinez-Barrio et al. 2016), the evolutionary and biological aspects of pinnipeds have not been investigated. Indeed, the genome of the Weddell seal (family *Phocidae*) has not been completed (http://software.broadinstitute.org/allpaths-lg/blog/?p=647). In addition, most phylogenetic studies of pinnipeds have used limited marker sequences, such as that of the mitochondrial genome (Slade, Moritz et al. 1994, Davis, Delisle et al. 2004, Fulton and Strobeck 2010).

Comparative genomics enables investigation of the convergent evolution of distant species. For example, convergent amino acid changes for vocal learning were identified by sequencing 48 avian genomes (Zhang, Li et al. 2014). Similarly, Parker *et al.* (Parker, Tsagkogeorga et al. 2013) reported nearly 200 convergent loci in the genomes of echolocating mammals. Although there are more studies to demonstrate to phenotypic convergence-linked sequence convergence, molecular convergence toward phenotypic convergence, at least in marine mammals, seems to be uncommon. By analyzing 22 mammalian genomes, including those of three marine mammals, Foote *et al.* (Foote, Liu et al. 2015) suggested that different molecular pathways could be used to reach the same phenotype. In a study of the *Hox* gene family in mammals, only a fraction of sites had positive selection signatures shared by three independent marine mammal lineages (Nery, Borges et al. 2016). Rather than sequence-level, gene-level convergence was presented as widespread signatures when evolutionary rates were used (Chikina, Robinson et al. 2016). Therefore, there is convergence at the functional level or higher in separate mammalian lineages, and different marine mammal lineages have used different molecular pathways to achieve phenotypic convergence.

Here, we constructed draft genomes of three species of two pinniped families: *Phoca largha* (*Phocidae*) and *Callorhinus ursinus* and *Eumetopias jubatus* (*Otariidae*) (Figure 2.1). We identified genes with a positive selection signature that were common to the three pinnipeds but absent from other mammals, which are likely related to the unique traits of pinnipeds. In addition, divergent molecular changes likely to occur only in the pinniped lineage during phenotypic convergence of marine mammals were investigated.

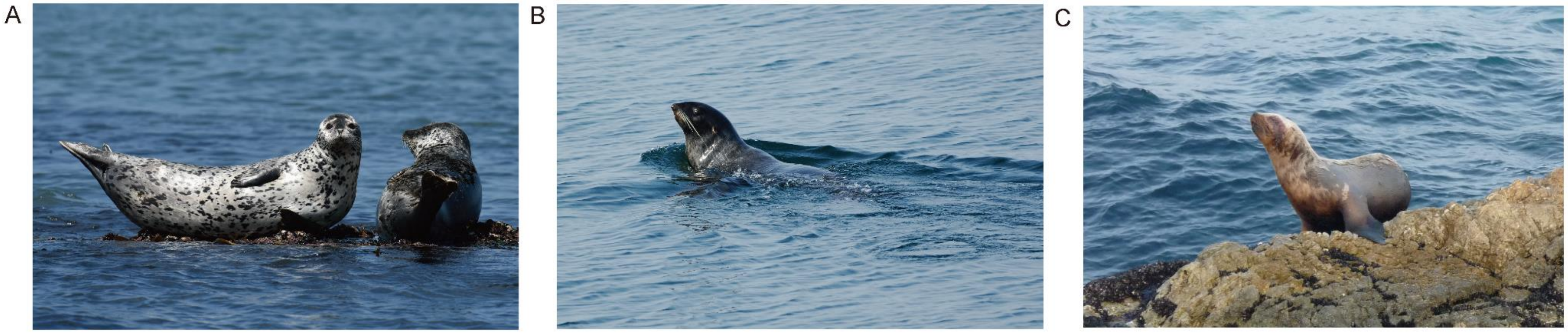


Figure 2.1 Examples of (a) *Phoca largha*; Spotted seal, (b) *Callorhinus ursinus*; Northern fur seal, and (c) *Eumetopias jubatus*; Steller sea lion

2.3 Materials and Methods

2.3.1 Ethics statement

No ethics approval was required for the collection of DNA from blood samples of bycaught carcasses.

2.3.2 Sample information and collection

We collected five pinniped samples from Korean waters. Three male Northern fur seals (*Callorhinus ursinus*) were bycaught in set nets and collected during January and February 2016 (one was used to produce sequence data). A bycaught female Steller sea lion (*Eumetopias jubatus*) was collected in April 2008. A female spotted seal (*Phoca largha*) was collected on a beach in August 2015. All of the above were found in the waters off Gangwon-do, northeastern South Korea.

2.3.3 DNA sequencing and genome assembly

For whole-genome shotgun sequencing and draft genome assembly, we constructed two paired-end libraries with insert sizes of 350 and 700 bp using the Illumina TruSeq DNA Sample Preparation Kit (Illumina, San Diego, CA, USA). For the Steller sea lion genome, mate-pair libraries with insert sizes of 3, 9, and 40 kb were constructed as scaffolds using the Illumina Nextera mate-pair library construction protocol (Illumina). Sequence reads were generated using the Illumina Nextseq500 platform. Information on the constructed libraries and sequencing data is provided in Table 2.1

The 19-mer distribution of the paired-end library with an insert size of 350 bp was calculated using Jellyfish (Marçais and Kingsford 2011), and the sizes of three genomes were estimated (Figure 2.2). To retrieve high-quality sequence reads, the quality of the raw data was controlled using FASTQC (Andrews 2010). Artifact sequences were removed via Trimmomatic (Bolger, Lohse et al. 2014) for paired-end libraries, and Nxtrim (O'Connell, Schulz-Trieglaff et al. 2015) for mate-pair libraries. Sequencing errors within each read were estimated and discarded using the error-correction module of Allpaths-LG (Gnerre, MacCallum et al. 2011). We assembled error-corrected paired-end reads using IDBA_UD (Peng, Leung et al. 2012) with the option of pre-correction and kmin = 40. Scaffolding on initial contigs was conducted using the paired-end reads with a 700 bp insert size, and mate-pair reads sequentially by SSPACE (Boetzer, Henkel et al. 2010) and ScaffMatch (Mandric and Zelikovsky 2015). After scaffolding, we iteratively filled gaps using Gapcloser (Luo, Liu et al. 2012) with the -l 155 and -p 31 parameters.

RepeatModeler (Smit and Hubley 2010), which includes RECON (Bao and Eddy 2002), RepeatScout (Price, Jones et al. 2005), and TRF (Benson 1999), was used to create a custom database for each species. A custom library was constructed by integrating the custom databases into the Repbase (Jurka, Kapitonov et al. 2005) database of mammals. Repeat elements were identified and masked using RepeatMasker (Tarailo-Graovac and Chen 2009) with the custom library and '-q, no_is' options.

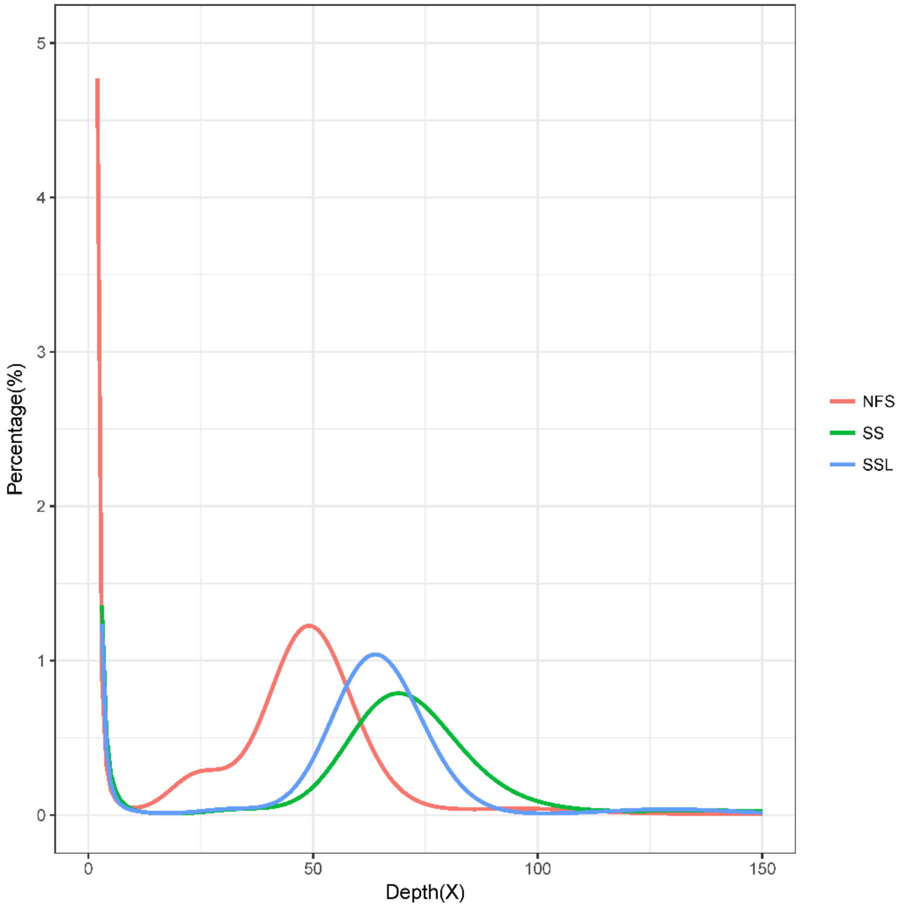


Figure 2.2 Estimation of three pinnipeds' genome sizes based on 19-mer

Table 2.1 Summary of sequencing data

Species	Library name	Library type	Insert size	Platform	Read length	No. reads	Total bp
Steller sea lion	350bp	Paired-end	350	Nextseq500	150	1,278,234,364	191,735,154,600
	700bp	Paired-end	700	Nextseq500	150	899,461,340	134,919,201,000
	3k	Mate-pair	3000	Nextseq500	151	421,988,284	63,720,230,884
	9k	Mate-pair	9000	Nextseq500	151	352,293,972	53,196,389,772
	40k	Mate-pair	40000	Nextseq500	151	278,537,902	42,059,223,202
Total						3,230,515,862	485,630,199,458
Spotted Seal	350bp	Paired-end	350	Nextseq500	150	1,365,397,528	204,809,629,200
	700bp	Paired-end	700	Nextseq500	150	1,009,079,390	151,361,908,500
Total						2,374,476,918	356,171,537,700
Northern Fur Seal	350bp	Paired-end	350	Nextseq500	150	1,011,695,208	151,754,281,200
	700bp	Paired-end	700	Nextseq500	150	1,807,129,054	271,069,358,100
Total						2,818,824,262	422,823,639,300

2.3.4 Genome annotation

Two approaches were used to predict protein-coding genes. First, manually curated protein sequences of Mammalia were retrieved from Swiss-Prot (Consortium 2014) and aligned to the pinniped genomes using tBLASTn (Altschul, Madden et al. 1997). The homologous genome sequences with E-values ≤ 1E-5 were extracted and realigned to the matched proteins using Exonerate (Slater and Birney 2005) to predict splice sites. Ab initio gene prediction was conducted using Augustus (Stanke, Keller et al. 2006), Geneid (Blanco, Parra et al. 2007), and GlimmerHMM (Majoros, Pertea et al. 2004) software with the default options. Predicted genes using each approach were combined using EvidenceModeler (Haas, Salzberg et al. 2008) into a consensus gene set.

For assessment of the quality of the draft genome, we remapped paired-end reads with a 350 bp insert size and investigated completeness of core-orthologs using BUSCO (Simão, Waterhouse et al. 2015).

For the three gene sets, the best match of a BLASTP (Altschul, Gish et al. 1990) search against the SwissProt and TrEMBL databases (Boeckmann, Bairoch et al. 2003) was assigned to putative functions. Gene motifs and domains were determined using InterProScan v. 5.19 (Zdobnov and Apweiler 2001). The GO IDs for each gene were obtained from the corresponding InterPro entries.

2.3.5 Ortholog identification

The complete proteome datasets were downloaded from UCSC Genome Browser (Tyner, Barber et al. 2016) for the following nine mammals: human (hg19), mouse (mm10), dog (canFam3), cow (bosTau8), manatee (triMan1), dolphin (turTru2), Minke whale (balAcu1), opossum (monDom5), and elephant (loxAfr3). Gene clusters for these nine mammals and three pinnipeds were identified using OrthoMCL v. 2.0.9 (Li, Stoeckert et al. 2003) with the default settings. A custom python script was used to generate a dataset comprising strict one-to-one orthologs (core-orthologs) from the 12 mammals.

2.3.6 Phylogenomic analyses using a genome-wide set of one-to-one orthologs

Amino acid sequences of 12 mammals corresponding to the one-to-one orthologs were individually aligned using ClustalW v. 2.1 (Larkin, Blackshields et al. 2007). A concatenated alignment was then prepared by merging individual alignments. The concatenated alignment was trimmed using Gblocks v. 0.91b (Castresana 2000) with auto settings.

The best-fit substitution model for the alignment was determined using ModelGenerator (Keane, Naughton et al. 2004). For phylogenetic analyses, RAxML v. 7.2.8 (Stamatakis 2006) was used to generate ML trees. Rapid bootstrap analysis and identification of the best-scoring ML tree (-f a option) were performed using RAxML v. 7.2.8 (Stamatakis 2006). Bootstrap support values/percentages were determined using 100 replicates. A Jones-Taylor-

Thornton amino acid substitution model (Jones, Taylor et al. 1992) (with the PROTCATIJTTF option) as recommended by ModelGenerator (Keane, Creevey et al. 2006) was used to construct the ML trees.

2.3.7 Detection of lineage-specific gene losses and gains

Using the gene clusters defined by Orthomcl v. 2.0.9 (Li, Stoeckert et al. 2003), the genes in each gene family group were enumerated and converted to input data for CAFÉ software v. 3.1 (De Bie, Cristianini et al. 2006). Expansion or contraction of the gene families was defined by comparing the cluster size of the ancestor to that of each of the current species using CAFÉ (De Bie, Cristianini et al. 2006).

2.3.8 Detection of positively selected genes and substitutions

To detect positively selected genes, coding sequence alignments were prepared by pal2nal v. 14 (Suyama, Torrents et al. 2006) using the amino acid alignments of the one-to-one orthologs. After trimming of the poorly aligned regions, alignments that are shorter than 100 bp or contain an internal stop codon were excluded.

To detect positive selection affecting a few sites in particular lineages (foreground branches, pinniped lineage in this study), we employed a branch-site model, which allows the ω ratio to vary both among lineages and among sites. We used the ML method of codeml in PAML v. 4.9 (Yang 2007), which estimates the rate of non-synonymous substitutions (dN), the rate of synonymous substitutions (dS), and the ratio of the non-synonymous to synonymous substitution rates (ω) values using the F3X4 codon frequencies. An alternative codon substitution model was specified using model = 2, NSsites = 2 (model A (Yang, Wong et al. 2005, Zhang, Nielsen et al. 2005), number of parameters k = 4), which was compared with the corresponding null model ω2 = 1 (ω ratio of foreground branches) fixed (fix_omega = 1 and omega = 1) using a likelihood-ratio test (LRT). From the alternative model, two different ω ratios of site class 2b (proportion: $(1 - p_0 - p_1)$ p1/$(p_0 + p_1)$, $\omega_1 = 1$, $\omega_2 \geq 1$) for pinniped branches (foreground branches) and other branches (background branches) were estimated (Figure 2.3a) to detect positive selection.

To identify fast-evolving genes in marine mammals (pinnipeds, cetaceans, and sirenians), we employed a branch model, which allows the ω ratio to vary among branches (Yang 1998). In codeml, an alternative codon substitution model was specified using model = 2 and NSistes = 0, which was compared with the basic null model (model = 0, NSsites = 0) by LRT. From the alternative model, two different ω ratios for marine mammal branches (foreground branches) and other branches (background branches) were estimated (Figure 2.2b).

Genes with maximum dS of > 3 or maximum dN/dS of > 5 in all branches or a log-likelihood ratio of < 0 were filtered from the output of each analysis. The Bonferroni method (Dunn 1961) was used to correct for multiple testing, and a value of $p < 0.05$ was taken to indicate statistical significance.

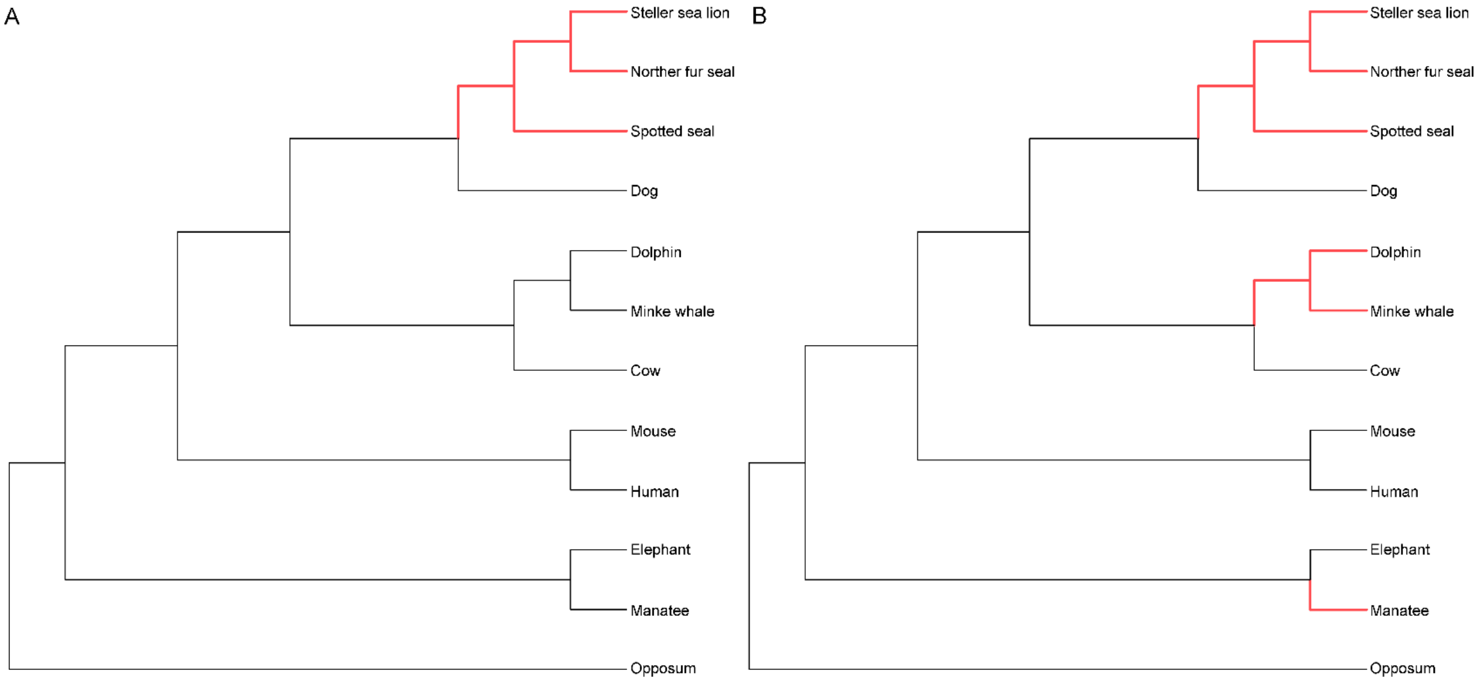
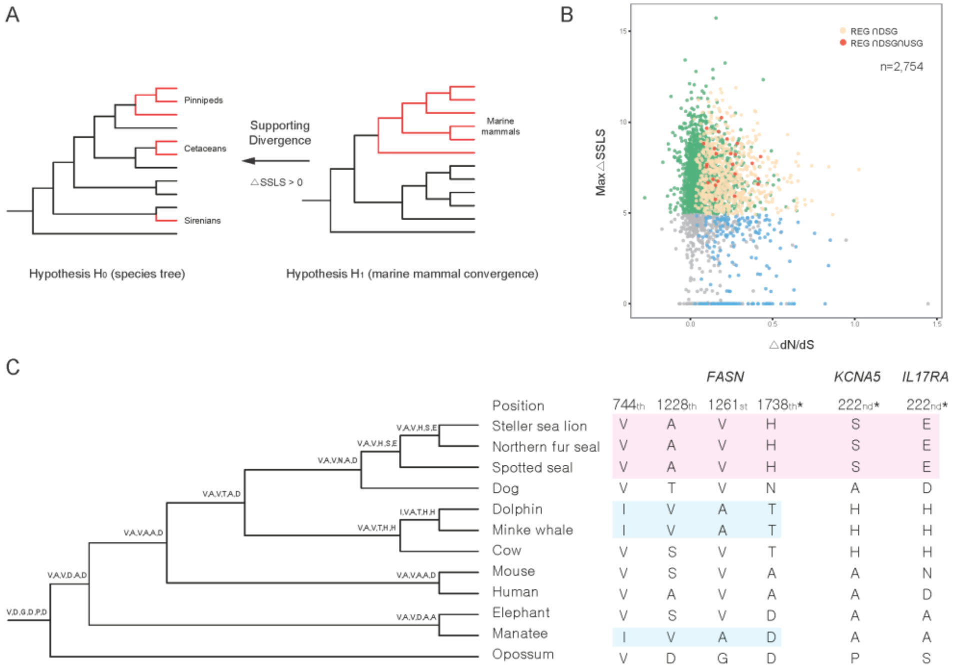


Figure 2.3 Foreground branches used in (a) branch-site model, and (b) branch model. Red line indicates foreground branches in each analyses

2.3.9 Calculation of site-wise likelihood support

To detect sites with molecular divergence that supported the monophyly of pinnipeds, we fitted the amino acid sequence alignment of one-to-one orthologs to a null model (H0, species tree) and an alternative model (H1, monophyly of marine mammals) (Figure. 2.4a). The goodness-of-fit of each site to a pair of phylogenetic trees under a given model was calculated as the SSLS value and directly compared as ΔSSLS = lnL (H0) - lnL (H1). Positive ΔSSLS values indicate a better fit of the model to the species tree, H0 (supporting divergence), whereas negative ΔSSLS values indicate a better fit to H1 (supporting convergence). The substitution model for each gene was determined by ModelGenerator (Keane, Creevey et al. 2006). The SSLS value for each site of alignment was estimated by RAxML v. 7.2.8 (Stamatakis 2006)

.



	FASN				*KCNA5*	*IL17RA*
Position	744th	1228th	1261st	1738th*	222nd*	222nd*
Steller sea lion	V	A	V	H	S	E
Northern fur seal	V	A	V	H	S	E
Spotted seal	V	A	V	H	S	E
Dog	V	T	V	N	A	D
Dolphin	I	V	A	T	H	H
Minke whale	I	V	A	T	H	H
Cow	V	S	V	T	H	H
Mouse	V	S	V	A	A	N
Human	V	A	V	A	A	D
Elephant	V	S	V	D	A	A
Manatee	I	V	A	D	A	A
Opossum	V	D	G	D	P	S

Figure 2.4 Analysis of rapidly evolving genes, divergent substitution genes, and unique substitution genes. (a) Hypotheses used to calculate ΔSSLS. (b) ΔdN/dS and ΔSSLS distribution in 2,754 orthologs. (c) Unique substitutions of *FASN*, *KCNA5*, and *IL17RA*. Asterisks, substitutions unique to pinnipeds. Other positions represent substitutions unique to cetaceans + sirenians

2.3.10 Identification of parallel and unique substitutions

We defined parallel substitutions as any amino acid change at the same position in marine mammals different from that of the ancestral node of each marine group, but identical in the three marine groups. To identify parallel amino acid changes in marine mammals, the species tree constructed in this study was used to reconstruct the ancestral sequences. The ancestral sequences for each node were reconstructed by Joint method using FastML v. 3.1 (Ashkenazy, Penn et al. 2012). We allowed FastML 3.1 to estimate the branch length of the phylogenetic tree for each gene when the ancestral sequences were reconstructed using the set of 12 mammals. For the sites with parallel and unique substitutions, the amino acid sequences of 100 vertebrates were investigated by 100-way multi-alignment (Blanchette, Kent et al. 2004) with the UCSC genome browser.

2.3.11 Conserved domain search

To determine whether positively selected sites are located in gene functional domains, we searched for conserved domains within positively selected genes using the CD-Search tool in the NCBI (Marchler-Bauer, Derbyshire et al. 2014). The amino acid sequences of human orthologs were used as a query set with the following settings: data source, CDD v. 3.16; expected value threshold, 0.01; composition-based statistical adjustment, applied; low-complexity filter, not applied.

2.3.12 Gene ontology analysis

We mapped the identified genes to GO categories in Ensembl (Flicek, Amode et al. 2011) to identify those putatively associated with a specific function, such as adipose tissue development. Gene set enrichment tests were performed by DAVID functional annotation (Dennis, Sherman et al. 2003) using a cutoff P-value of < 0.05.

2.4 Results

2.4.1 Genome assembly and annotation

Before assembling the genomes of the three pinnipeds, we estimated the genome sizes using the 19-mer distribution of paired-end reads. The estimated genome sizes were 2.61, 2.71, and 2.64 Gbp for the spotted seal (SS), northern fur seal (NFS), and Steller sea lion (SSL), respectively. The genomic DNA of the three pinnipeds was assembled to a size of approximately 2.5 Gbp, which is similar to that of previously assembled genomes (Antarctic fur seal (Humble, Martinez-Barrio et al. 2016), Hawaiian monk seal [**https://www.ncbi.nlm.nih.gov/assembly/GCF_002201575.1**], and Weddell seal[**https://www.ncbi.nlm.nih.gov/assembly/GCF_000349705.1**]).

Summary statistics of the final assembly are provided in Table 2.2. To assess the quality of the draft genomes, we remapped paired-end reads with a 350 bp insert size, which yielded alignment rates of > 98% for the three genomes (98.24, 98.74, and 98.73% for SS, NFS, and SSL, respectively). The completeness of core-orthologs was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO). Each of the three genomes contained more than 90% core-orthologs from the class Mammalia, in the form of either complete or fragmented sequences (Table 2.3). The GC contents of the three genomes were investigated using 500 bp bins, and were similar to those of the draft genomes of related species (Figure 2.5).

Repeat elements accounted for 35.83, 40.40, and 35.78% of the SS, NFS, and SSL genomes, respectively. Of the repeat regions, long interspersed nuclear element (LINE) was the most extended element in terms of base pairs (Table 2.4). After masking the identified repeat elements, 33,988, 32,740, and 28,081 protein-coding genes were predicted for SS, NFS, and SSL, respectively (Table 2.5). Of the predicted genes, ~ 95% were functionally annotated to at least one of the InterPro, SwissProt, and TrEMBL databases (Table 2.6).

Therefore, the SS, NFS, and SSL genomes were not significantly different from one another in terms of various statistics related to genome assembly. Because the three species are related, this similarity suggests that the three genomes have similar levels of completeness.

Table 2.2 Summary of assembly statistics (>2000bp)

	Spotted Seal	**Northern Fur seal**	**Steller Sea Lion**
Size (Haploid)	2.26 Gb	2.37 Gb	2.36 Gb
GC level	41.36%	41.39%	40.98%
No. scaffolds	50053	69688	17807
N50 of scaffolds (bp)	87534	67633	331138
N bases in scaffolds (%)	0.2 Mb (0.01%)	0.6 Mb (0.03%)	21 Mb (0.90%)
Longest (shortest) scaffolds (bp)	732428	606905 (2000)	2503745 (2000)
Average scaffold length (bp)	45467.84	33959.64	132798.99

Table 2.3 Summary of genome assessment results

	Spotted Seal	Norther Fur Seal	Stella Sea Lion
Overall remapping rate	98.73%	98.74%	98.24%
Complete single-copy BUSCOs	3652 (89.0%)	3101 (75.6%)	3169 (77.2%)
Complete duplicate BUSCOs	49 (1.2%)	39 (1.0%)	28 (0.7%)
Fragmented BUSCOs	241 (5.9%)	641 (15.6%)	613 (14.9%)
Missing BUSCOs	162 (3.9%)	323 (7.8%)	294 (7.2%)

Table 2.4 Summary statistics for repeat elements

	Spotted Seal		Norther Fur Seal		Stella Sea Lion	
Repeat element	**No. element**	**Length (%)**	**No. element**	**Length bp (%)**	**No. element**	**Length (%)**
SINE	1,036,137	141,553,944 (5.58%)	1,129,582	161,422,763 (5.71%)	851,418	133,941,433 (5.07%)
LINE	2,214,162	585,943,176 (23.08%)	3,011,201	743,144,050 (26.28%)	2,377,806	660,209,680 (25.01%)
LTR element	197,943	82,674,816 (3.26%)	354,083	99,715,643 (3.53%)	158,944	69,491,321 (2.63%)
DNA element	125,810	35,396,811 (1.39%)	142,116	36,834,655 (1.30%)	97,051	29,844,098 (1.13%)
Small RNA	975,134	134,338,873 (5.29%)	1,066,608	153,715,199 (5.44%)	801,275	128,404,897 (4.86%)
Satellites	2,177	697,132 (0.03%)	12,501	2,277,985 (0.08%)	1,750	223,507 (0.01%)
Simple repeat	767,846	35,008,073 (1.38%)	886,899	41,726,927 (1.48%)	819,929	39,293,888 (1.49%)
Low complexity	106,255	5,545,681 (0.22%)	133,029	6,805,611 (0.24%)	108,031	5,560,112 (0.21%)
Unclassified	145,816	24,567,042 (0.97%)	371,817	52,945,392 (1.87%)	33,385	6,259,843 (0.24%)

Table 2.5 Summary statistics of gene prediction results

Species	Element	No. of elements	Average length	Count per gene	Total length	Genome coverage
Spotted Seal	**Gene**	33,988	13,012.64	-	442,273,447	0.17
	Exon	177,734	194.80	5.23	34,622,532	0.01
	Intron	143,746	2,835.91	4.23	407,650,915	0.16
Northern Fur Seal	**Gene**	32,740	12,642.54	-	413,916,607	0.15
	Exon	172,034	192.00	5.25	33,030,109	0.01
	Intron	139,294	2,734.41	4.25	380,886,498	0.13
Steller Sea Lion	**Gene**	28,081	21,049.79	-	591,099,025	0.22
	Exon	180,207	209.79	6.42	37,805,425	0.01
	Intron	152,126	3,637.07	5.42	553,293,600	0.21

Table 2.6 Summary of functional annotation results

		Spotted Seal		Northern Fur Seal		Stella Sea Lion	
Annotated	**InterPro**	23,890	70.29%	23,829	72.78%	21,310	75.89%
	GO	18,368	54.04%	18,590	56.78%	17,045	60.70%
	SwissProt	31,713	93.31%	30,807	94.10%	25,905	92.25%
	TrEMBL	31,462	92.57%	30,401	92.86%	25,664	91.39%
Not annotated		1,922	5.65%	1,682	5.14%	1,962	6.99%
Total		33,988		32,740		28,081	

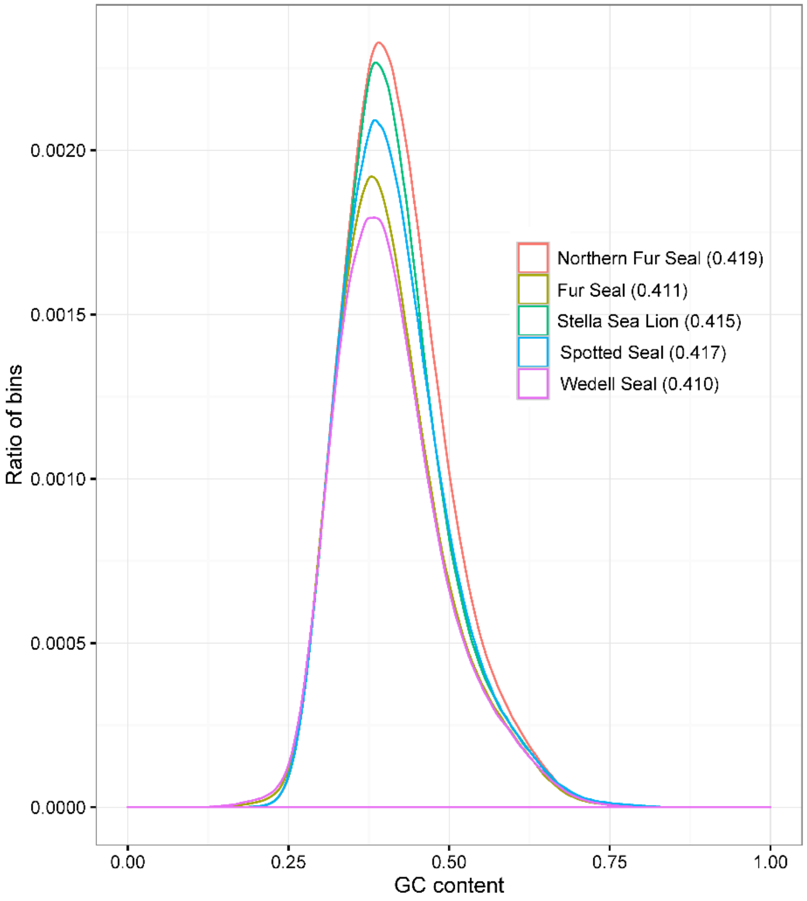


Figure 2.5 Distribution of GC content for the pinnipeds genomes

2.4.2 Phylogenomics and protein-coding gene families

To identify the relationships among SS, NFS, and SSL and other related species, we constructed a maximum-likelihood (ML) tree using the amino acid sequence of one-to-one orthologs generated using a dataset of the proteomes of nine species available in public databases. In total, there were 2,907 one-to-one orthologs, the combined length of which was 982,250 amino acid residues. The newly constructed tree provided robust support for the known phylogenetic tree of marine mammals (**http://www.timetree.org/**) (Figure 2.6a), and the phylogenetic tree is used in the downstream analysis for positively selected genes and substitutions.

We constructed orthologous gene clusters using the genomes of six marine mammals to identify gene clusters and their functions unique to pinnipeds (Figure 2.7). The pinniped genomes contained 13,919 (NFS), 13,441 (SS), and 14,165 (SSL) orthologous gene families, respectively, 9,639 of which were shared by all three pinnipeds (Figure 2.6b). Of these gene families, 1,874 were present in all pinnipeds, but not in three other mammals. By Gene Ontology (GO) enrichment analysis, we found these gene families to be enriched in 31 terms (p-value < 0.05), several of which were related to an aquatic lifestyle, such as ‘aorta development’, ‘sterol biosynthetic process’, ‘cardiac septum development’, ‘coronary vasculature development’, and ‘cellular response to oxidative stress’ (Table 2.7).

To investigate gene-family expansion and contraction, a computational analysis of gene-family sizes using the orthologous gene clusters was performed in CAFÉ (De Bie, Cristianini et al. 2006). By comparing six marine mammals, we found that 874 gene families were expanded, while 1,925 gene families were contracted in the pinniped lineage. Of these gene families, a subset of the Protocadherin (Pcdh) family (herein named family 34) was significantly expanded in the pinniped lineage (p = 0.000346). The genomes of the pinnipeds contained a larger number of Pcdh genes than those of the other marine mammals (Figure 2.6c). Pcdhs are the largest mammalian subgroup of the cadherin superfamily (Hulpiau and Van Roy 2009), and have functions associated with the nervous system (Wang, Weiner et al. 2002, Chen, Alvarez et al. 2012) such as in olfactory sensory neurons (Hasegawa, Hamada et al. 2008). The number of Pcdhs varies among vertebrate lineages (Yagi 2008).

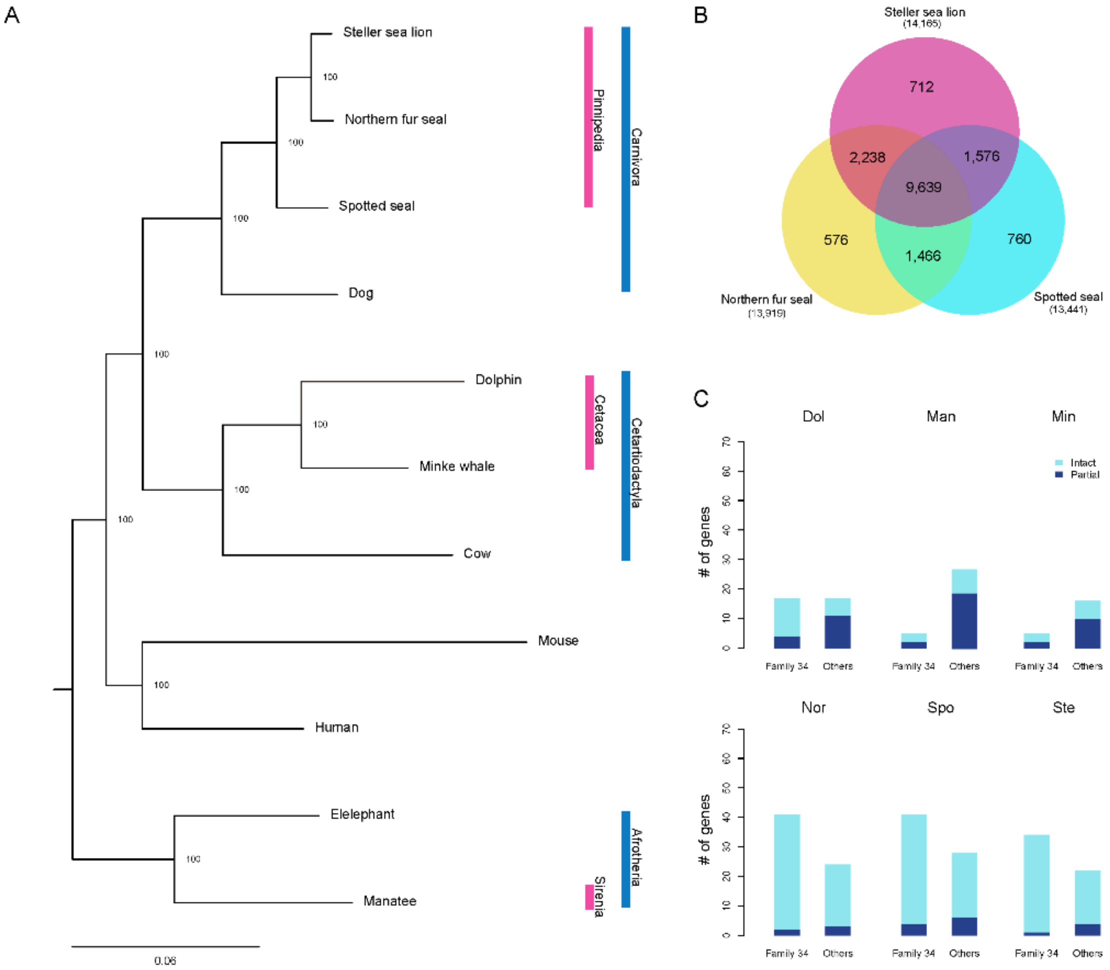


Figure 2.6 Phylogenomics and protein-coding gene families of pinnipeds. (a) Species tree of 12 terrestrial and marine mammals constructed by the maximum-likelihood method. (b) Orthologous gene clusters in three pinnipeds. (c) Number of intact (coverage ≥ 90%) and partial (coverage < 90%) genes that belong to Protocadherin gene families, named family 34 in our dataset (Dol, dolphin; Man, manatee; Min, Minke whale; Nor, northern fur seal; Spo, spotted seal; Ste, Steller sea lion)

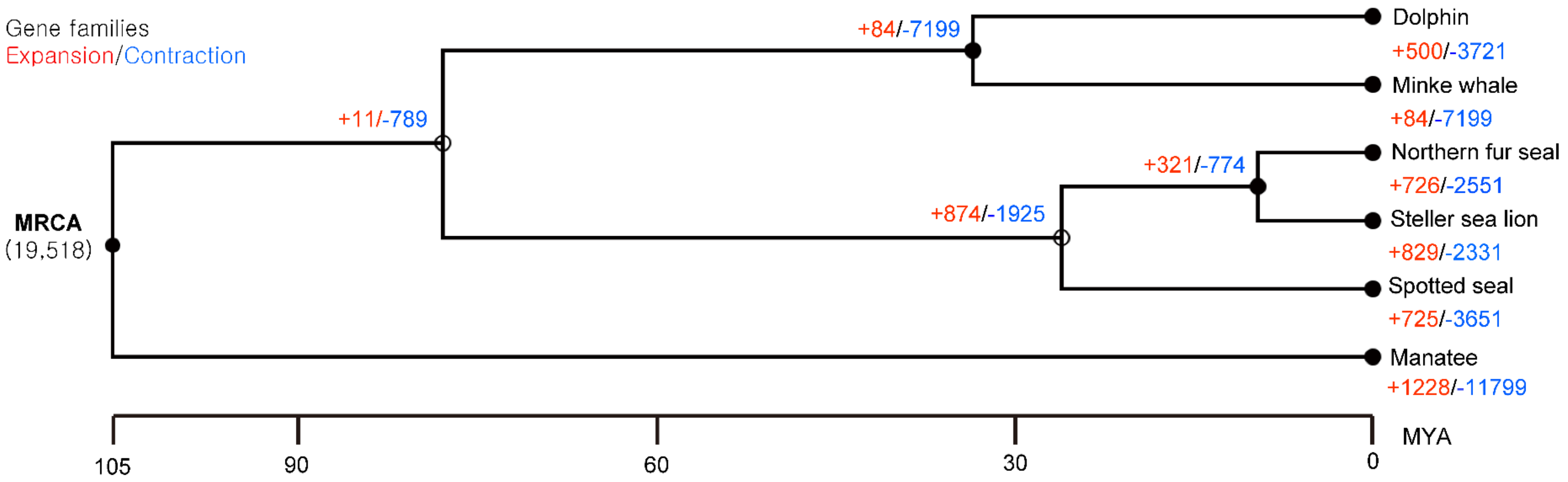


Figure 2.7 Gene family expansion or contraction across 6 marine mammals

Table 2.7 Gene ontology (GO) enrichment analysis of pinnipeds specific gene families (P-value < 0.05)

ID	GO Term	Count	%	PValue	Pop Hits	Fold Enrichment
GO:0032024	positive regulation of insulin secretion	8	0.448	0.004	41	3.929
GO:0006366	transcription from RNA polymerase II promoter	40	2.241	0.005	513	1.570
GO:0019985	translesion synthesis	7	0.392	0.008	36	3.915
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	66	3.697	0.010	981	1.355
GO:0035904	aorta development	5	0.280	0.011	18	5.593
GO:0016126	sterol biosynthetic process	4	0.224	0.011	10	8.054
GO:0070987	error-free translesion synthesis	5	0.280	0.013	19	5.298
GO:0042276	error-prone translesion synthesis	5	0.280	0.013	19	5.298
GO:0007005	mitochondrion organization	10	0.560	0.014	77	2.615
GO:0002244	hematopoietic progenitor cell differentiation	9	0.504	0.016	66	2.746
GO:0048538	thymus development	7	0.392	0.019	43	3.278
GO:0003279	cardiac septum development	4	0.224	0.019	12	6.711
GO:0032228	regulation of synaptic transmission, GABAergic	4	0.224	0.019	12	6.711
GO:0000722	telomere maintenance via recombination	6	0.336	0.020	32	3.775
GO:1901796	regulation of signal transduction by p53 class mediator	13	0.728	0.020	124	2.111

GO:0009792	embryo development ending in birth or egg hatching	3	0.168	0.022	5	12.081
GO:1903232	melanosome assembly	3	0.168	0.022	5	12.081
GO:0006486	protein glycosylation	12	0.672	0.025	113	2.138
GO:0006297	nucleotide-excision repair, DNA gap filling	5	0.280	0.029	24	4.195
GO:0006974	cellular response to DNA damage stimulus	18	1.008	0.030	208	1.742
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	48	2.689	0.032	720	1.342
GO:0006273	lagging strand elongation	3	0.168	0.032	6	10.067
GO:0060976	coronary vasculature development	5	0.280	0.033	25	4.027
GO:0033683	nucleotide-excision repair, DNA incision	6	0.336	0.038	38	3.179
GO:0034599	cellular response to oxidative stress	8	0.448	0.039	64	2.517
GO:0051593	response to folic acid	3	0.168	0.044	7	8.629
GO:1900264	positive regulation of DNA-directed DNA polymerase activity	3	0.168	0.044	7	8.629
GO:0006260	DNA replication	14	0.784	0.045	155	1.819
GO:0030326	embryonic limb morphogenesis	6	0.336	0.046	40	3.020
GO:0006470	protein dephosphorylation	12	0.672	0.049	126	1.918
GO:0048663	neuron fate commitment	4	0.224	0.049	17	4.737

2.4.3 Genes with accelerated evolution in the pinniped lineage

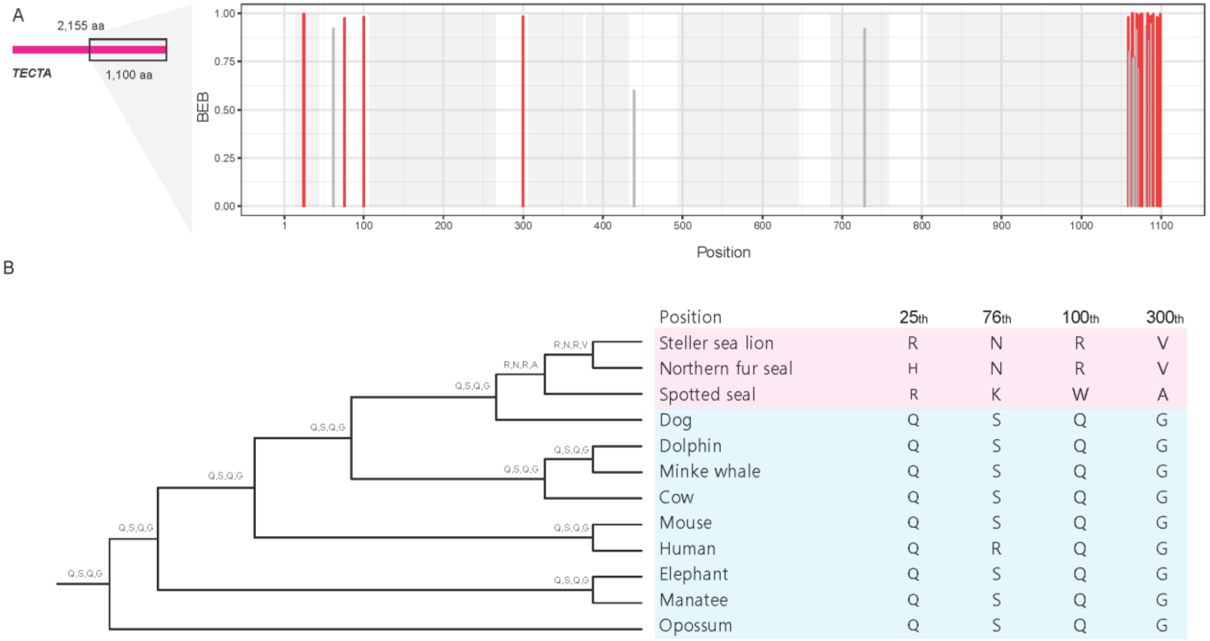
To detect positive selection in the pinniped lineage, a dN/dS analysis using the branch-site model was performed. The branch-site model allows dN/dS (ω) to vary both among sites in the protein and across branches on the tree (Yang, Wong et al. 2005). Therefore, we hypothesized a few sites in the pinniped branches to have different ω ratios compared to other branches and that the genes containing these sites might be related to the unique features of pinnipeds. After the filtering step (see Materials and Methods), we analyzed 2,754 one-to-one orthologs identified in the proteomes of 12 mammals, of which seven genes with 145 sites were under positive selection (Bonferroni-corrected $p < 0.05$, posterior probability based on Bayes empirical Bayes inference [BEB] > 0.95; Table 2.8). Of these genes, transmembrane protein 132B (*TMEM132B*) contained the largest number of positively selected sites (52 sites). Of the seven genes, six contained 29 conserved domains with 74 sites (51%) under positive selection. GO terms were assigned to each gene, and the following functional associations with pinniped lifestyle were found: *TECTA*, sensory perception of sound (GO:0007605), *SPEG*, muscle organ development (GO:0007517), and *ADAMTS5*, defense response to bacterium (GO:0042742) and tooth eruption (GO:0044691). *TECTA* encodes alpha-tectorin, a major non-collagenous glycoprotein of the tectorial membrane, an extracellular matrix in the inner ear (Hulpiau and Van Roy 2009). Mutations in *TECTA* result in hearing loss (Meyer, Alasti et al. 2007, Alasti, Sanati et al. 2008, Collin, de Heer et al. 2008) (OMIM: 602574). *SPEG* is required for cardiac development and is associated with cardiac myopathy (Liu, Ramjiganesh et al. 2009, Agrawal, Pierson et al. 2014) (OMIM: 615950). *ADAMTS5*, which encodes an extracellular matrix-degrading enzyme, plays an important role in the T-cell immune response to viral infection (McMahon, McCulloch et al. 2016, Stambas, Ye et al. 2017).

To assess their uniqueness, the amino acid residues positively selected in the pinniped lineage were compared to other species in our analysis as well as in publicly available databases. For example, we investigated 4 of the 18 sites within *TECTA* after manually filtering out amino acid residues with spurious alignment (Figure 2.8a). The four sites were pinniped-specific compared to the other nine species (Figure 2.8b). Moreover, a 100-way multi-alignment showed that two pinnipeds (Pacific walrus and Weddell seal) had residues identical to those in the three pinnipeds in this study (Figure 2.9). We could only find a small number of residues matching those in 100 vertebrates at these sites (Figure 2.9). Consequently, the four sites within *TECTA* might be unique to pinnipeds and generated during their adaptation to a semi-aquatic environment.

Table 2.8 Genes with accelerated evolution in the pinniped lineage. H1_fg_omega: dN/dS value (ω) on foreground given H1 (ω varies across the branches); H0_lnl: log likelihood given H0 (ω does not vary across the branches); H1_lnl: log likelihood given H1; H0_lnl: log likelihood given H0

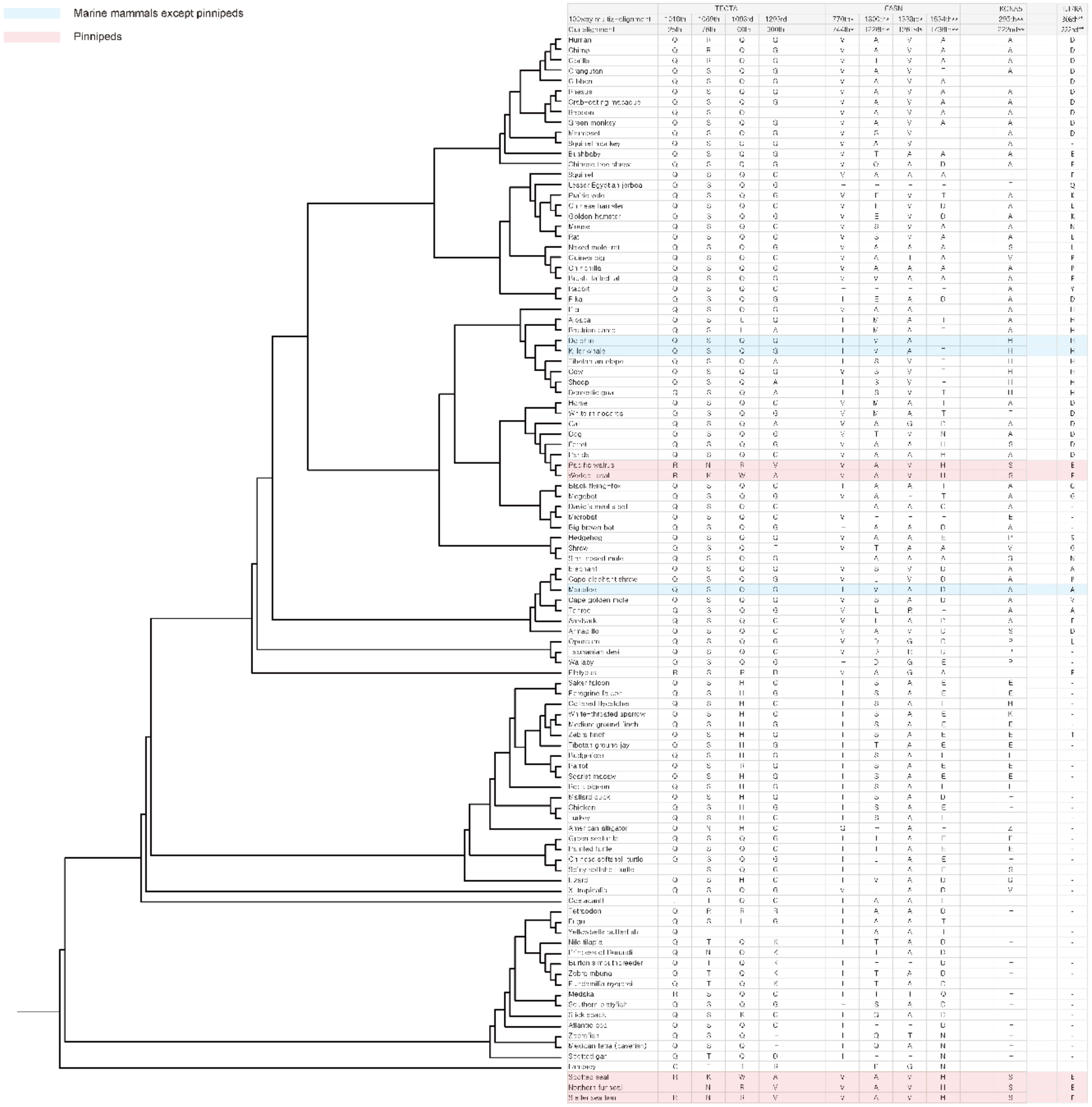
Gene	H1_fg_omega (ω_2)	Proportion (H_1) $(1 – p_0 – p_1)p_1/(p0 + p_1)$	H0_lnl	H1_lnl	Likelihood ratio	p-value	Adjusted p-value	# of positively selected sites*
TMEM132B	3.81581	0.01666	-6438.78	-6419.68	38.20475	6.37E-10	1.18E-06	52 (22)
PARP1	4.76894	0.00604	-5357.53	-5341.29	32.48145	1.20E-08	2.22E-05	23 (22)
TECTA	3.67139	0.00194	-12076.1	-12060.4	31.42787	2.07E-08	3.83E-05	18 (14)
FUBP3	4.89809	0.01916	-4880.95	-4869.76	22.38143	2.24E-06	0.004144	12 (1)
IGF2BP1	4.96893	0.00201	-4448.2	-4438.13	20.13898	7.20E-06	0.01332	19 (2)
SPEG	4.81594	0.00254	-11218.8	-11209.4	18.85029	1.41E-05	0.026085	13 (13)
ADAMTS5	4.38148	0.00124	-4320.48	-4311.43	18.1014	2.09E-05	0.038665	8 (0)

*Number of positively selected sites with a BEB of > 0.95. The numbers of positively selected sites within domain regions are shown in parentheses.



Position	25th	76th	100th	300th
Steller sea lion	R	N	R	V
Northern fur seal	H	N	R	V
Spotted seal	R	K	W	A
Dog	Q	S	Q	G
Dolphin	Q	S	Q	G
Minke whale	Q	S	Q	G
Cow	Q	S	Q	G
Mouse	Q	S	Q	G
Human	Q	R	Q	G
Elephant	Q	S	Q	G
Manatee	Q	S	Q	G
Opossum	Q	S	Q	G

Figure 2.8 Results of a branch-site model analysis of *TECTA*. (a) Bayes empirical Bayes (BEB) posterior probability in *TECTA*. Shaded area, conserved domain regions. (b) Sequence of sites with significant BEB (> 0.95). Red and blue shaded areas, pinnipeds and other mammals, respectively



	TECTA				FASN				KCNA5	IL17RA
100way multiz-alignment	1010th	1069th	1093rd	1293rd	770th*	1800th**	1333rd*	1634th**	295th**	306th**
Our alignment	25th	78th	90th	300th	744th*	1228th**	1261st*	1738th**	222nd**	222nd**
Human	Q	R	Q	G	V	A	V	A	A	D
Chimp	Q	R	Q	G	V	A	V	A	A	D
Gorilla	Q	R	Q	G	V	I	V	A	A	D
Orangutan	Q	S	Q	G	V	A	V	-	A	D
Gibbon	Q	S	Q	G	V	A	V	A		D
Rhesus	Q	S	Q	G	V	A	V	A	A	D
Crab-eating macaque	Q	S	Q	G	V	A	V	A	A	D
Baboon	Q	S	Q		V	A	V	A	A	D
Green monkey	Q	S	Q	G	V	A	V	A	A	D
Marmoset	Q	S	Q	G	V	S	V		A	D
Squirrel monkey	Q	S	Q	G	V	A	V		A	-
Bushbaby	Q	S	Q	G	V	T	A	A	A	E
Chinese tree shrew	Q	S	Q	G	V	Q	A	D	A	F
Squirrel	Q	S	Q	C	V	A	A	A	-	F
Lesser Egyptian jerboa	Q	S	Q	G	-	-	-	-	-	Q
Prairie vole	Q	S	Q	G	V	F	V	T	A	K
Chinese hamster	Q	S	Q	C	V	I	V	D	A	L
Golden hamster	Q	S	Q	G	V	E	V	D	A	K
Mouse	Q	S	Q	C	V	S	V	A	A	N
Rat	Q	S	Q	C	V	S	V	A	A	L
Naked mole-rat	Q	S	Q	G	V	A	A	A	S	I
Guinea pig	Q	S	Q	C	V	A	I	A	V	F
Chinchilla	Q	S	Q	G	V	A	A	A	A	F
Brush-tailed rat	Q	S	Q	G	V	V	A	A	A	F
Rabbit	Q	S	Q	C	-	-	-	-	A	Y
Pika	Q	S	Q	G	I	E	A	D	A	D
Pig	Q	S	Q	G	V	A	A		A	H
Alpaca	Q	S	L	G	I	M	A	I	A	H
Bactrian camel	Q	S	I	A	I	M	A	-	A	H
Dolphin	Q	S	Q	G	I	V	A		H	H
Killer whale	Q	S	Q	G	I	V	A	-	H	H
Tibetan antelope	Q	S	Q	A	I	S	V	-	H	H
Cow	Q	S	Q	G	V	S	V	-	H	H
Sheep	Q	S	Q	A	I	S	V	-	H	H
Domestic goat	Q	S	Q	A	I	S	V	T	H	H
Horse	Q	S	Q	C	V	M	A	I	A	D
White rhinoceros	Q	S	Q	G	V	M	A	T	-	D
Cat	Q	S	Q	A	V	A	G	D	A	D
Dog	Q	S	Q	G	V	T	V	N	A	D
Ferret	Q	S	Q	G	V	A	A	H	S	D
Panda	Q	S	Q	C	V	A	A	H	A	D
Pacific walrus	R	N	R	V	V	A	V	H	S	E
Weddell seal	R	K	W	A	V	A	V	H	S	F
Black flying-fox	Q	S	Q	C	I	A	A	I	A	C
Megabat	Q	S	Q	G	V	A	-	T	A	G
David's myotis bat	Q	S	Q	C		A	A	C	A	-
Microbat	Q	S	Q	C	V	-	-	-	E	-
Big brown bat	Q	S	Q	G	-	A	A	D	A	-
Hedgehog	Q	S	Q	G	V	A	A	E	P	S
Shrew	Q	S	Q	-	V	T	A	A	V	G
Star-nosed mole	Q	S	Q	G		A	A	A	S	N
Elephant	Q	S	Q	G	V	S	V	D	A	A
Cape elephant shrew	Q	S	Q	G	V	_	V	D	A	F
Manatee	Q	S	Q	G	I	V	A	D	A	A
Cape golden mole	Q	S	Q	G	V	S	A	D	A	V
Tenrec	Q	S	Q	G	V	L	P	-	A	A
Aardvark	Q	S	Q	C	V	I	A	D	A	F
Armadillo	Q	S	Q	C	V	A	V	D	S	D
Opossum	Q	S	Q	G	V	D	G	D	P	L
Tasmanian devil	Q	S	Q	C	V	D	E	D	'	-
Wallaby	Q	S	Q	G	-	D	G	E	P	-
Platypus	R	S	R	D	V	A	G	A		F
Saker falcon	Q	S	H	C	I	S	A	E	E	-
Peregrine falcon	Q	S	H	G	I	S	A	E	E	-
Collared flycatcher	Q	S	H	C	I	S	A	I	H	-
White-throated sparrow	Q	S	H	C	I	S	A	E	K	-
Medium ground finch	Q	S	H	G	I	S	A	E	E	-
Zebra finch	Q	S	H	G	I	S	A	E	E	T
Tibetan ground jay	Q	S	H	G	I	T	A	E	E	-
Budgerigar	Q	S	H	G	I	S	A	I	I	
Parrot	Q	S	R	G	I	S	A	E	E	-
Scarlet macaw	Q	S	H	G	I	S	A	E	E	-
Rock pigeon	Q	S	H	G	I	S	A	I	I	
Mallard duck	Q	S	H	G	I	S	A	D	-	-
Chicken	Q	S	H	G	I	S	A	E	-	-
Turkey	Q	S	H	C	I	S	A	I		-
American alligator	Q	N	H	C	G	-	A	-	Z	-
Green seaturtle	Q	S	Q	G	I	I	A	F	F	-
Painted turtle	Q	S	Q	C	I	I	A	E	E	-
Chinese softshell turtle	Q	S	Q	G	I	_	A	E	-	-
Spiny softshell turtle		S	Q	G	I		A	F	S	
Lizard	Q	S	H	C	I	V	A	D	G	-
X. tropicalis	Q	S	Q	G	V	_	A	D	V	-
Coelacanth	.	I	Q	C	I	A	A	I		
Tetraodon	Q	R	R	R	I	A	A	D	-	-
Fugu	Q	S	I	G	I	A	A	T		
Yellowbelly pufferfish	Q				I	A	A	I		-
Nile tilapia	Q	T	Q	K	I	T	A	D	-	-
Princess of Burundi	Q	N	Q	K		I	A	D		
Burton's mouthbreeder	Q	I	Q	K	I	-	-	D	-	-
Zebra mbuna	Q	T	Q	K	I	T	A	D	-	-
Pundamilia nyererei	Q	T	Q	K	I	T	A	D		
Medaka	R	S	Q	C	I	I	I	Q	-	-
Southern platyfish	Q	S	Q	G	-	S	A	D	-	-
Stickleback	Q	S	K	C	I	G	A	D		-
Atlantic cod	Q	S	Q	C	I	-	-	D	-	-
Zebrafish	Q	S	Q	-	I	Q	T	N	-	-
Mexican tetra (cavefish)	Q	S	Q	-	I	Q	A	N	-	-
Spotted gar	Q	T	Q	D	I	-	-	N	-	-
Lamprey	C	-	T	R		F	G	N		
Spotted seal	R	K	W	A	V	A	V	H	S	E
Northern fur seal		N	R	V	V	A	V	H	S	E
Steller sea lion	R	N	R	V	V	A	V	H	S	F

Figure 2.9 Amino acid sequences across 100 vertebrates at the site with rapid evolution or substitution unique to three pinnipeds (Spotted seal, Northern fur seal, and Steller sea lion). Sites at *TECTA, FASN, KCNA5 and IL17RA* are shown. ** indicates substitutions unique to pinnipeds. * indicates substitution unique to cetaceans + sirenians

.

2.4.4 Unique substitutions of pinnipeds contributed to the phenotypic convergence of marine mammals

Parallel substitutions are widespread in marine mammals; however, most are not unique to marine mammals (Foote, Liu et al. 2015, Zhou, Seim et al. 2015). Moreover, molecular convergences are rarely linked to phenotypic convergences in marine mammals (Foote, Liu et al. 2015, Chikina, Robinson et al. 2016, Nery, Borges et al. 2016). In this study, about half of the parallel substitutions shared by marine mammals were also found in terrestrial mammals, and a considerable number of unique substitutions was found between species with no obvious phenotypic convergence (Figure 2.10–2.12). Therefore, we hypothesized the existence of pinniped-specific substitutions that contributed to aquatic adaptation and are shared by marine mammals.

First, we focused on gene-level convergence (Figure 2.13) and conducted a dN/dS analysis of one-to-one orthologs using the branch model. The branch model allows the dN/dS (ω) ratio to vary among branches in the phylogeny and is useful for detecting positive selection acting on particular lineages (Yang 1998). In this way we aimed to detect candidate genes with different ω ratios among the marine mammal lineages rather than candidate sites, which may contribute to phenotypic convergence among marine mammals. Of the 2,754 filtered one-to-one orthologs, the branch model-based dN/dS analysis detected 853 positively selected genes in marine mammal lineages (Figure 2.3b, cetaceans, pinnipeds, and sirenians, Bonferroni corrected p-value < 0.05). These are hereafter referred to as rapidly evolving genes (REGs). A subset of

853 REGs covered the following functional categories possibly associated with marine mammals' adaptation to the ocean: muscle physiology (GO:0007015, GO:0035914, GO:0007519, and GO:0035914), lipid metabolism (GO:0006629, GO:0006869, GO:0006631, and GO:0016042), sensory system (GO:0007605, GO:0042472, and GO:0021772), skin and connective tissue (GO:0008544, GO:0043588, and GO:0030216), cardiovascular system (GO:0086091, GO:0060976, and GO:0007507), and resistance to oxidative stress (GO:0001666).

We also calculated the site-wise log likelihood support (SSLS) values for the amino acid sequences of 2,754 genes (996,522 residues in total) and calculated the ΔSSLS values to detect site-wise signatures of divergent evolution. The ΔSSLS value is indicative of the goodness-of-fit of each site to a pair of phylogenetic trees. We aimed to detect genes positively selected in three marine mammal lineages with substitutions unique to pinnipeds. Therefore, we calculated the SSLS for two hypotheses: H0, divergence among marine mammal clades and H1, convergence among marine mammal clades. Therefore, a ΔSSLS (log likelihood of H0 − log likelihood of H1) value > 0 means that the site in question supports divergence among marine mammal clades. We used the ΔSSLS value as a filtering criterion to exclude sites supporting convergence among marine mammals. By excluding those with low ΔSSLS values, we identified pinniped-specific sites that support the separation clades of marine mammals. We expected that this analysis would generate more reliable sites than directly extracting unique substitutions over REGs, as it considers the overall phylogeny not just the sequence itself.

We regarded the 9,965 residues with the top 1% ΔSSLS values as being supported by divergent substitutions (support for H0) rather than convergent substitutions among three marine mammal clades (support for H1) (Figure 2.4a). We termed the 2,159 genes containing at least one of these residues as divergent substitution genes (DSGs). DSGs covered most of the 2,754 one-to-one orthologs (78%), and 85% of total residues had positive ΔSSLS values. Therefore, the majority of the sequences supported the commonly accepted phylogeny.

Of the 853 REGs, 658 (3,277 residues) had a least one top 1% ΔSSLS site (Figure 2.4b). Although these genes covered the functional categories associated with marine mammals' adaptation, a single residue divergence supported by likelihood divergence (ΔSSLS) could be vulnerable to sequencing error. We also focused on sequence changes common to the pinniped clade; that is, changes from the ancestral node sequence shared by dog to that of the node of each pinniped. Therefore, we investigated unique substitutions (any amino acid residue at the same position in all three pinnipeds that was found in neither the ancestral nodes with their respective terrestrial taxa [dog] nor in other mammals) to rule out sequence divergences other than ancestral substitutions unique to the pinniped clade. There were 1,731 genes with at least one unique substitution (7,878 residues); these were termed unique substitution genes (USGs), 63 of which contained top 1% ΔSSLS residues at the same positions as unique substitutions. Finally, we obtained 24 REGs containing top 1% ΔSSLS residues and unique substitutions at the same positions (Figure 2.4 and Table 2.9).

Although the 24 REGs are supported by rapid evolutionary rates (dN/dS) and fixation of amino acid residues within the pinniped clade, the precise phenotypic effects of the unique substitutions cannot currently be ascertained. However, several of the 24 REGs have known functional associations that suggest a role in the convergent phenotypic evolution of marine mammal lineages. For instance, *FASN* encodes fatty-acid synthase, which catalyzes the conversion of acetyl-CoA and malonyl-CoA to long-chain saturated fatty acids (Wakil 1989) and is related to obesity (Loftus, Jaworsky et al. 2000). *KCNA5* (potassium voltage-gated channel subfamily A member 5) encodes voltage-gated potassium channels in pulmonary artery smooth muscle cells and mediates the response to hypoxia (Platoshyn, Brevnova et al. 2006, Firth, Platoshyn et al. 2009). *IL17RA* encodes the interleukin 17A receptor, a ubiquitous type I membrane glycoprotein that binds to interleukin 17A. Interleukin 17A and its receptor play a key role in the immune response to pathogen infection (Cypowyj, Picard et al. 2012, Bär, Whitney et al. 2014).

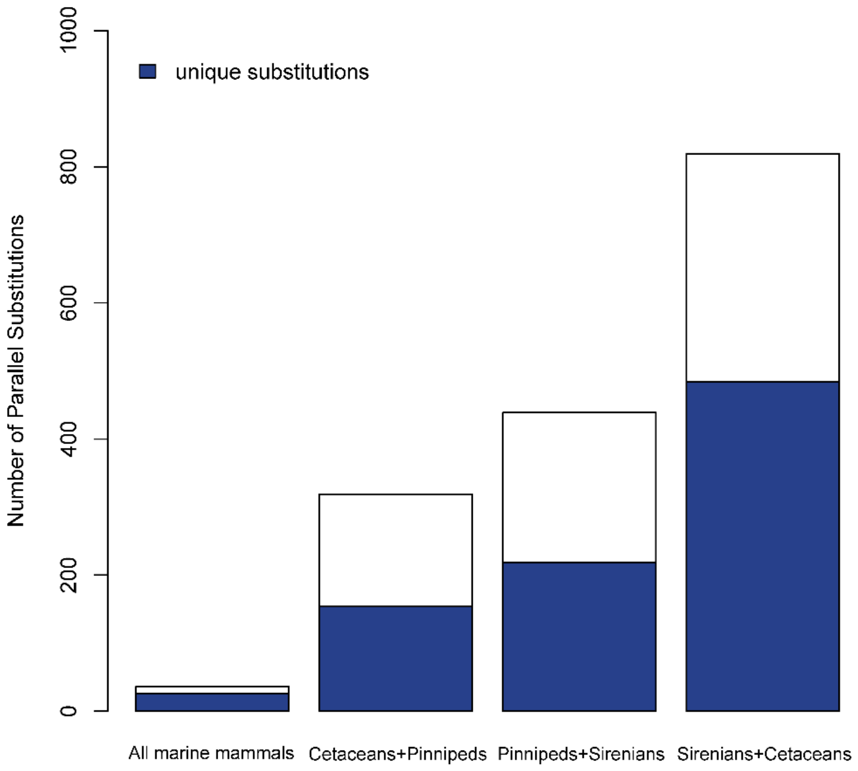


Figure 2.10 Number of parallel and unique substitutions across marine mammal clades

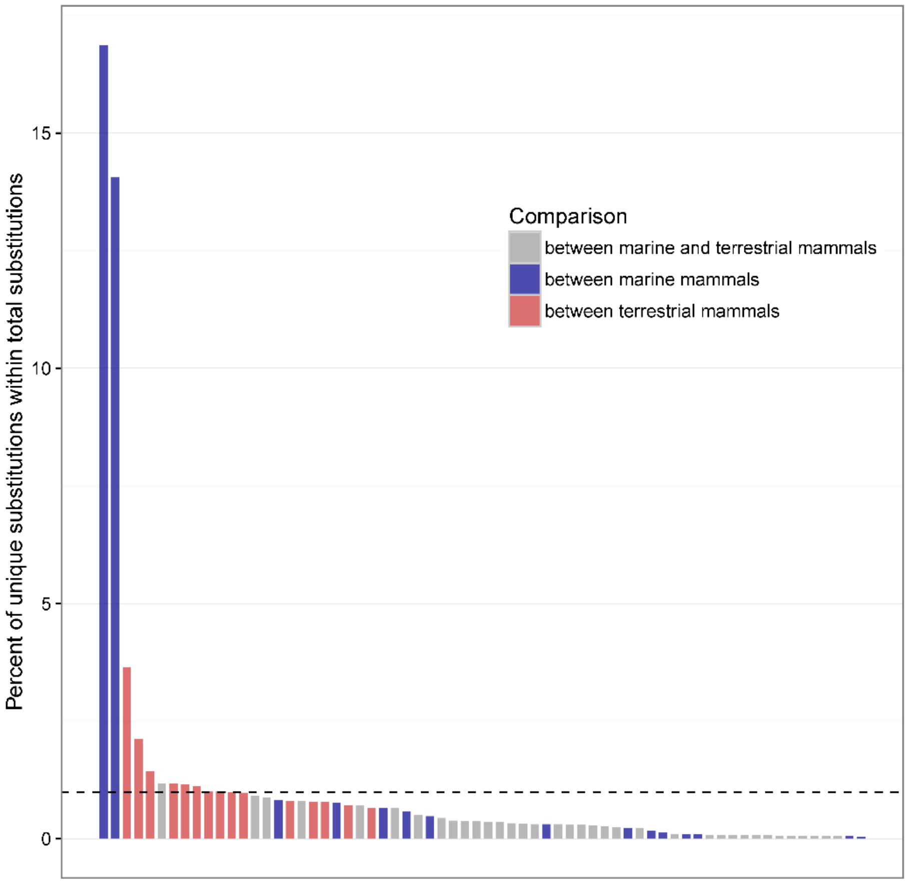


Figure 2.11 Percentage of unique substitutions in total substitutions for all pairwise comparisons among all 12 species in the phylogeny. The dashed line indicates the average

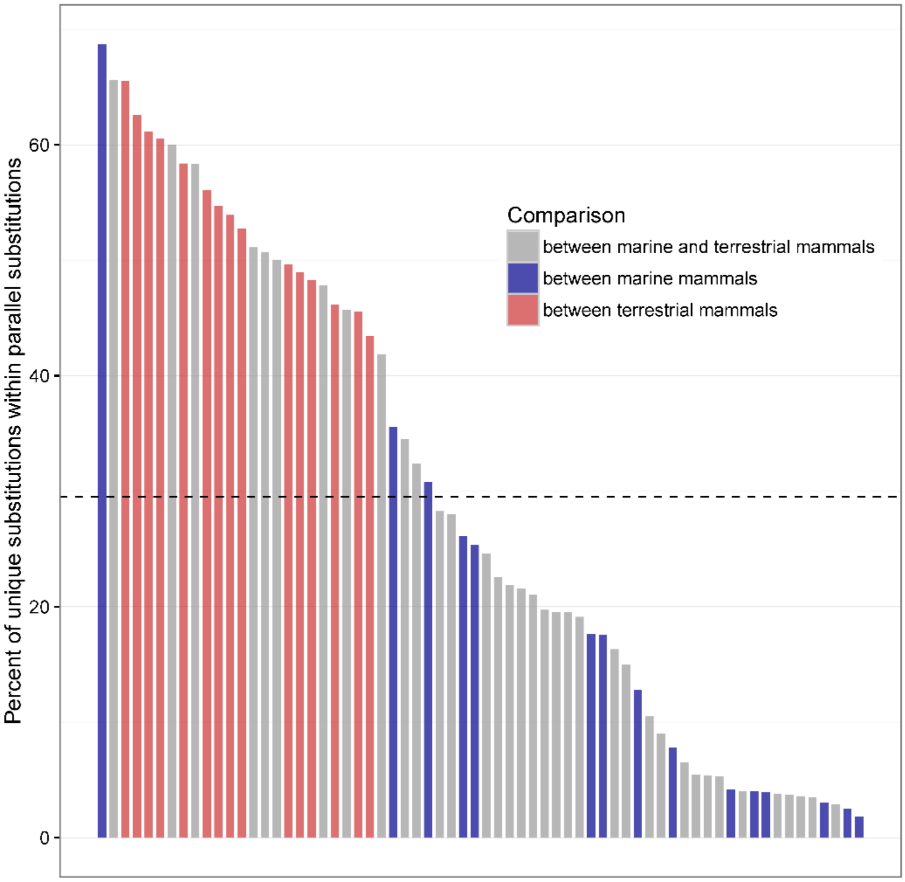


Figure 2.12 Percentage of unique substitutions in parallel substitutions for all pairwise comparisons among all 12 species in the phylogeny. The dashed line indicates the average

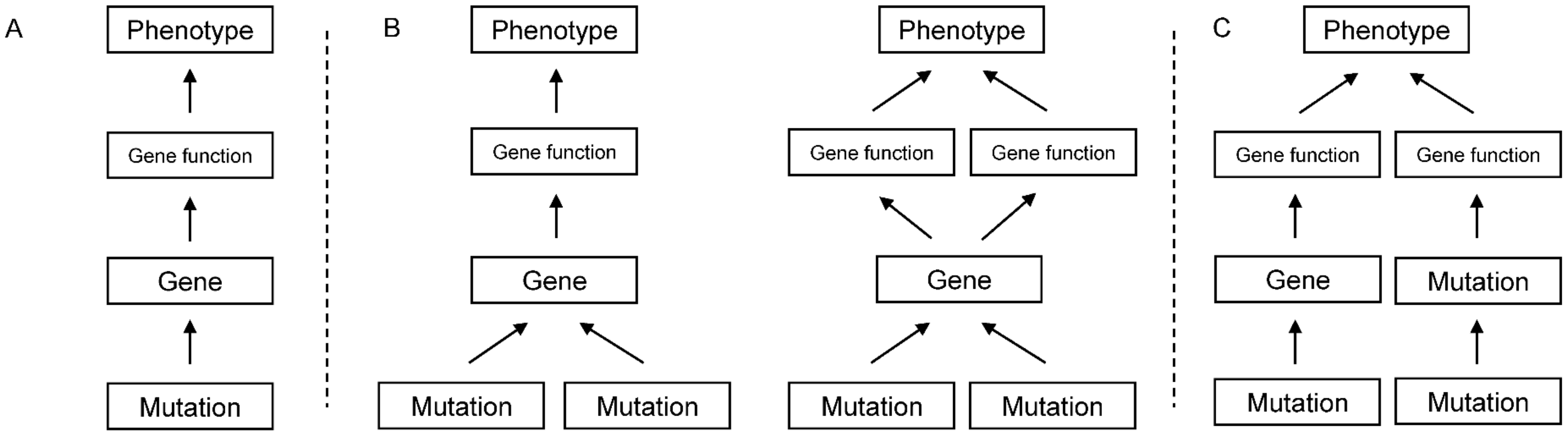


Figure 2.13 Convergent evolution at multiple levels. Similar phenotypes can evolve at: (a) molecular; (b) gene; (c) phenotypic level (Manceau, Domingues et al. 2010). We have focused e on the gene-level convergence (b) in this study

Table 2.9 Genes with sequence changes likely to occur in only the pinniped lineage when gene-level convergence took place in marine mammals. H1_fg_omega: dN/dS value (ω) on foreground branches given H1 (ω varies across the branches); H0_lnl: log likelihood given H0 (ω does not vary across the branches); H1_lnl: log likelihood given H1.

Gene	H1_fg_omega	H0_lnl	H1_lnl	p-value	Adjusted p-value	Max ΔSSLS	# of unique substitutions
VPS45	0.40038	-4037.348759	-3956.066722	3.11E-37	8.3037E-34	7.730292	1
ABCC10	0.44339	-18216.8972	-18153.02735	1.28E-29	3.4176E-26	8.828157	20
FASN	0.19743	-40595.23443	-40538.33849	1.45E-26	3.8715E-23	9.661292	54
DUS3L	0.34814	-8646.525061	-8591.484612	9.41E-26	2.51247E-22	8.224026	5
DDAH2	0.45032	-2977.374327	-2929.688683	1.58E-22	4.2186E-19	6.582644	3
SASH1	0.19451	-6515.324513	-6472.937933	3.35E-20	8.9445E-17	6.525431	6
GPR155	0.57001	-5926.033389	-5888.825926	6.33E-18	1.69011E-14	6.877673	6
DUSP27	0.28847	-13197.13408	-13162.41783	7.91E-17	2.11197E-13	6.749015	79
EMILIN3	0.26942	-9364.829765	-9346.305662	1.15E-09	3.0705E-06	8.265838	11
DCLRE1A	0.70785	-6950.035862	-6931.672364	1.36E-09	3.6312E-06	8.100957	6
DGKQ	0.1879	-12994.3629	-12976.13466	1.56E-09	4.1652E-06	7.537842	13

VWF	0.21584	-26711.98598	-26695.2834	7.48E-09	1.99716E-05	8.820315	34
GUCY2C	0.45014	-6447.80369	-6431.307957	9.26E-09	2.47242E-05	5.926296	4
ABCD4	0.23237	-6428.830395	-6414.05014	5.42E-08	0.000144714	6.854527	7
TACC3	0.48752	-7423.970534	-7410.509676	0.000000212	0.00056604	8.265423	6
LMTK2	0.34446	-19363.6547	-19351.05935	0.000000519	0.00138573	8.54542	12
RIN3	0.27223	-6861.555137	-6849.333436	0.000000765	0.00204255	5.750655	6
KCNA5	0.18524	-6876.444906	-6864.677432	0.00000123	0.0032841	6.664093	6
TRMT12	0.48457	-6214.487554	-6203.00578	0.00000165	0.0044055	7.141302	7
POLL	0.41568	-7373.848752	-7362.382731	0.00000168	0.0044856	9.073179	9
ANKRD5	0.3288	-9488.47528	-9477.525821	0.00000287	0.0076629	9.485991	10
LAMB2	0.23606	-17812.14129	-17801.84156	0.00000566	0.0151122	7.663766	8
IL17RA	0.40977	-10647.044	-10636.87528	0.00000649	0.0173283	10.242048	12
TRIML1	0.45219	-4898.490822	-4888.744854	0.0000101	0.026967	7.674387	4

2.5 Discussion

In this study, we presented three genomes of pinnipeds (*Phoca largha*, *Callorhinus ursinus* and *Eumetopias jubatus*) that belong to *Phocidae*, and *Otariidae* family for the first time. *Pinnipedia* is a monophyletic group distinct from other marine mammals in many respects, such as its semi-aquatic lifestyle and well-developed flippers (Berta, Sumich et al. 2005). Our findings provide insight into the common features of pinniped genomes, which is less clear than the convergent evolution of pinnipeds.

Pinnipeds are the most amphibious mammalian species. Possibly, for that reason, their auditory systems are challenged by the need to function efficiently underwater and in air, unlike the solely underwater hearing of cetaceans and sirenians (Wartzok and Ketten 1999, Reichmuth, Holt et al. 2013). *TECTA*, which is related to sound perception (Alasti, Sanati et al. 2008) was identified as positively selected in the pinniped lineage. *TECTA* encodes α-tectorin, a non-collagenous component of the tectorial membrane in the cochlea (Verhoeven, Van Laer et al. 1998). The tectorial membrane is an extracellular matrix that covers the surface of the sensory epithelium in the cochlea and plays a vital role in transmitting sound to the stereocilia of hair cells, where the sound is transduced into neural signals (Michalski and Petit 2015). Therefore, mutations in *TECTA* might be involved in the semi-aquatic adaptation of pinnipeds by tuning their hearing ranges. Indeed, mutations in *TECTA* are responsible for loss of hearing at particular frequencies (Collin, de Heer et al. 2008, Moteki, Nishio et al. 2012, Ishikawa, Naito et al. 2014). Interestingly, the four positively selected sites in *TECTA* were very rare among 100 vertebrates (Figure 2.9). Although its relationship with amphibious sound perception is unclear, *TECTA* should be investigated in future studies of amphibious sound perception in pinnipeds. The pinniped lifestyle might influence the function of other candidate genes, such as *SPEG* and *ADAMTS5*. Comparative analysis of amphibious mammals may reveal their adaptations at the molecular level and show that an amphibious lifestyle results in selection pressure.

We found that a considerable number of parallel substitutions are not unique to marine mammals, consistent with two recent reports (Foote, Liu et al. 2015, Zhou, Seim et al. 2015). This implies that molecular convergence is not a driving force of phenotypic convergence among marine mammals, and that different clades of marine mammals used different molecular pathways to reach similar phenotypes. Although this phenomenon has been observed several times in marine mammals, whether it also applies to other clades is unclear. More evidence in other clades is needed to generalize this phenomenon to other forms of phenotypic convergence.

Because sequence convergences leading to phenotypic convergences are not common, we assumed that unique substitutions contributed to the aquatic adaptation of pinnipeds. In our analyses, three genes, *FASN*, *KCNA5*, and *IL17RA*, were identified as candidates. The well-defined roles of these genes (blubber (Dunn 1961), resistance to hypoxia (Davis 2014), and the immune response to pathogens (Foote, Liu et al. 2015), respectively) support their contributions to phenotypic convergences of marine mammals. *FASN* and *KCNA5* were not found to be positively selected in the branch-site model analysis using all marine mammal branches as foreground branches. In addition, only ~ 17% of the REGs were found to be positively selected genes by the branch-site model analysis (Fig. S11). Such results suggest that rapid evolution occurred at different sites of the candidate genes between marine mammal clades, an example of gene-level convergent evolution.

Convergent evolution can occur at molecule, gene, and function levels (Parker, Tsagkogeorga et al. 2013, Zhou, Seim et al. 2015). We focused on convergence at the gene level. However, the functions of the majority of the putative convergent genes were unrelated to apparent phenotypic convergence, such as lipid metabolism and resistance to oxidative stress. This may be due to the missing link between convergent genes and phenotypic convergences. In this case, the results can be complemented by studying the gene functions and convergence at a higher-level.

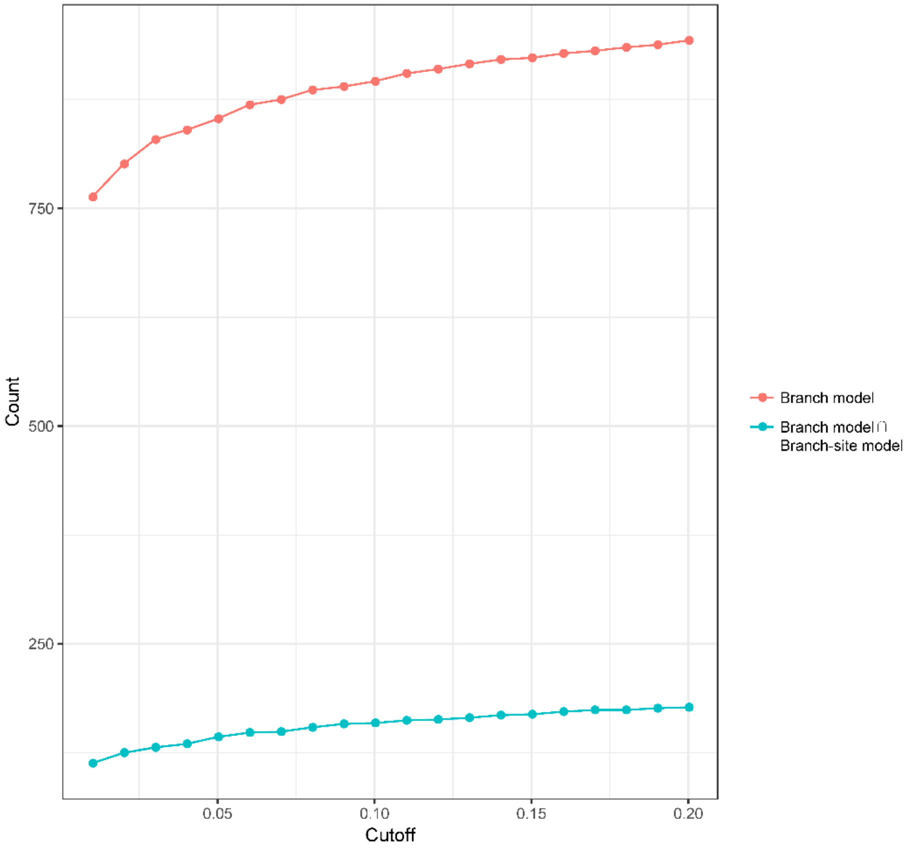


Figure 2.14 Intersect between the positively selected genes from branch model analysis and branch-site model analysis for all marine mammal's branches with continuous p-value cutoffs. Red line indicates the number of positively selected genes from branch model analysis and turquoise line indicates the number of positively selected genes detected in both of branch model and branch-site model analysis

This chapter was published in *Human Genetics*
as a partial fulfillment of Kwondo Kim's Ph.D program.

Chapter 3. *De novo* emergence and potential function of human-specific tandem repeats in brain-related loci

3.1 Abstract

Tandem repeats (TRs) are widespread in the genomes of all living organisms. In eukaryotes, they are found in both coding and noncoding regions and have potential roles in the regulation of cellular processes such as transcription, translation and in the modification of protein structure. Recent studies have highlighted TRs as a key regulator of gene expression and a potential contributor to human evolution. Thus, TRs are emerging as an important source of variation that can result in differential gene expression at intra- and inter-species levels.

In this study, we performed a genome-wide survey to identify TRs that have emerged in the human lineage. We further examined these loci to explore their potential functional significance for human evolution. We identified 152 human-specific TR (HSTR) loci containing a repeat unit of more than 10 bases, with most of them showing a repeat count of two. Gene set enrichment analysis showed that HSTR-associated genes were associated with biological functions in brain development and synapse function. In addition, we compared gene expression of human HSTR loci with orthologues from non-human primates (NHP) in seven different tissues. Strikingly, the expression level of HSTR-associated genes in brain tissues was significantly higher in human than in NHP.

These results suggest the possibility that de novo emergence of TRs could have resulted in altered gene expression in humans within a short time frame and contributed to the rapid evolution of human brain function.

3.2 Introduction

Tandem repeats (TR) are DNA segments where a pattern of one or more nucleotides is repeated in the directly adjacent sequence (Lander, Linton et al. 2001). TRs are also known as satellite DNA, since they were initially detected as satellite bands in density-gradient centrifugal DNA separation. Although there is no clear definition, TRs with a unit ranging from 1 to 9 nucleotides in length are generally referred to as microsatellites while those with a longer repeat unit (≥10 nucleotides) are known as minisatellites (Gemayel, Vinces et al. 2010). Typical minisatelites range in size up to 100bp (Vergnaud and Denoeud 2000), however, some minisatelites up to 500bp have been identified in the human genome (Lander, Linton et al. 2001).

TRs are ubiquitous in eukaryotic genomes. In the human genome, TRs comprise 3% of its length (Lander, Linton et al. 2001). TRs are found in both coding and noncoding regions and are often associated with alterations to cellular processes including transcription, translation as well as altering protein structure, depending on their locations (Gemayel, Vinces et al. 2010). Huntington's disease is a classic example, where affected individuals have a larger number of CAG sequence repeats in the first exon of their *HTT* (huntingtin) gene (Walker 2007). The expansion of TRs in the protein coding region *HTT* gives rise to abnormal protein, which gradually damages cells in the brain (Usdin 2008, Gymrek, Willems et al. 2017). TRs located in the noncoding region are also able to modify the binding of transcription factors. TRs in the promoter region of the *PIG3* gene directly interact with the P53 transcription factor to mediates its induction (Contente, Dittmer et al. 2002). Expression of the dopamine transporter gene (*DAT1*) is similarly regulated by the number of TRs located in the 3′ UTR (Mill, Asherson et al. 2002).

Recent studies have highlighted the relationship between TR variation and gene expression (Sonay, Carvalho et al. 2015, Gymrek, Willems et al. 2016). Gymrek et al. surveyed genome-wide short tandem repeats (STRs) and identified 2,060 STRs within promoter regions, which were significantly associated with altered expression of the neighboring genes. By analyzing genome-wide association studies (GWAS), they predicted that TRs were associated with various clinical conditions (Gymrek, Willems et al. 2016). Another study by Sonay et al. showed that polymorphic TRs in promoters alters gene expression across human and primates and in several different tissues. This study also suggested that TRs are associated with biological processes such as stimulus detection, sensory perception, and skin development, which are related to the evolution of human cognitive traits and adaptation to new environments (Sonay, Carvalho et al. 2015).

Much research has focused on genomic alterations contributing to the unique evolutionary phenotypes in human (O'bleness, Searles et al. 2012). These studies suggest that alterations to the DNA sequence are likely to have contributed to the evolution of human-specific phenotypes, prompting us to explore the potential relevance of human-specific TRs to the evolution of human-specific traits (Enard, Gehre et al. 2009, Dumas, O’Bleness et al. 2012, Suzuki, Miyabe et al. 2018).

Mutation rates of TRs are 10 to 100,000 times higher than that in other parts of the genome. Variation in these regions is generated by strand slippage and homologous recombination errors rather than point mutations (Legendre, Pochet et al. 2007). We hypothesized that under neutral condition, TRs are unlikely to be fixed in length and sequence composition. Therefore, *de novo* TRs that are fixed could indicate a positive selection signature and loci that are related to advantageous traits.

In this study, we extracted the loci where putative de novo TRs emerged and are fixed, specifically in the human lineage by comparing the genome sequences of humans with three non-human primates (NHP) (Chimpanzee, Orangutan, and Gorilla). Then, we analyzed their potential association with gene expression and their functional role in the evolution of human-specific traits.

3.3 Materials and Methods

3.3.1 Detection of human-specific tandem repeat

A list of whole tandem repeats for the human genome (GRCh38) generated by TRF (Benson 1999) was downloaded from UCSC genome browser (https://genome.ucsc.edu/). Among the total 1,014,212 TR loci, 1,014,188 non-redundant TR loci were used for downstream analyses.

To identify TRs that have emerged only in the human lineage, we employed a method from a study examining Alu elements (Sen, Han et al. 2006) (Figure 3.1B). For each tandem repeat loci identified in the human genome, 400bp of upstream and downstream were extracted from unmasked genome sequences. RBH (Reciprocal best hits) were detected to find orthologous sequences in masked NHP genomes using Blast with default settings (ver. 2.2.30) (Moreno-Hagelsieb and Latimer 2008). We used bitscore to find the best hits, and filtered out those that did not cover whole query sequence in both the upstream and downstream regions.

The intervening sequence (IS) between the two flanking sequences on NHP genomes was examined to determine HSTRs, and the loci with IS that include "N" base or do not have any sequence were removed. For the remaining loci, we only considered the loci where NHP have repeat count of one to reduce false positives derived from variation within each NHP population (Figure 3.1D). The criteria for this process are as follows; 1) Human TR (HTR) length ≥ IS length*2, 2) HTR length – IS length ≥ unit length of HTR, and 3) At least one match with percent identity > 95% between unit sequence of HTR and IS. We repeated the above procedure for three comparisons (human vs chimpanzee, human vs gorilla and human vs orangutan) and intersected three TR lists. The TRs in the list were considered as human-specific TRs that are expanded only in human genome.

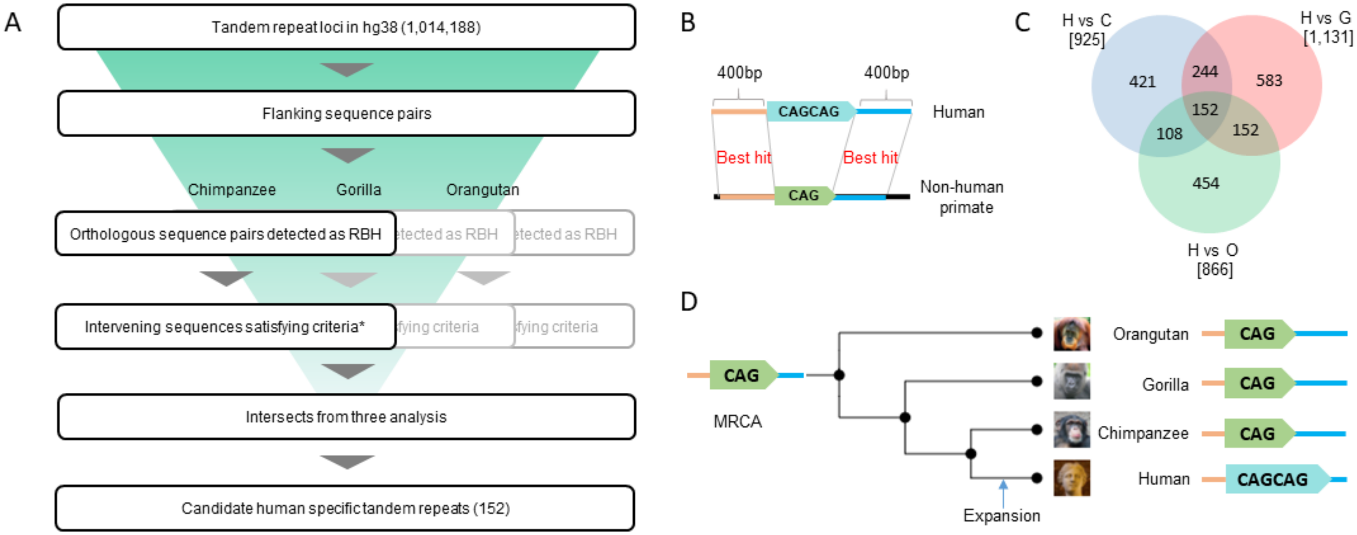


Figure 3.1 Workflow for detecting human-specific tandem repeat. (a) A diagram representing whole analysis steps to detect HSTR. TRs were filtered by the criterion of each steps (see Materials and Methods). (b) A schematic of method for detecting HSTRs. The concept of method was modified from a previous study (Sen, Han et al. 2006). (c) The number of TRs detected in each comparison (human vs chimpanzee, human vs gorilla and human vs orangutan). (d) A type of TRs (HSTR) we expect to identify in this study

3.3.2 Basic statistics and genomic location of HSTR

For the sequences of total TRs and HSTRs, five characteristics (total length, unit length, number of repeats, and percentage of match between repeats) were independently investigated. The percentage of match between repeat units of HSTRs was compared to total TRs using a random sample. The random sampling was performed with the subset of the total TRs that have same number of repeats to HSTRs. After generating 152 random samples (same as HSTRs), we calculated the mean percentage of match, and repeated this process 1000 times. The location of TR was classified into exon, intron, boundaries of exon and intron, CDS, and intergenic using Ensembl annotations (Hubbard, Barker et al. 2002).

3.3.3 *In silico* validation of HSTRs in 1000 genome data set

To investigate HSTR variation at the population level, we examined them in 24 ethnically diverse, high coverage genomes from the 1000 genomes project (Consortium 2015) (Table 3.1). BLAST (Altschul, Gish et al. 1990) databases were built from the whole genome sequence reads for each individual. We used HSTR sequences with 20 nucleotides of flanking sequence as queries. Since the read length of the data was only 250bp, queries longer than 250bp were excluded from the BLAST search. A hit was counted if the high-scoring segment pair (HSP) contained at least 98% of the query sequence.

Table 3.1 Information of 24 high coverage data from 1000 genome project (Consortium 2015)

Sample Name	SRS ID	Read length	Read Count (pair)	Population	SRR ID
HG01112	SRS010781	250	418,550,608	CLM	SRR1291024/SRR1291070
HG00096	SRS006837	250	386,212,832	GBR	SRR1291026/SRR1291035
HG01583	SRS368308	250	340,123,297	PJL	SRR1291030/SRR1291036
NA19648	SRS003635	250	301,246,465	MXL	SRR1291041/SRR1291138
HG01051	SRS010747	250	391,232,599	PUR	SRR1291141/SRR1291157
HG00268	SRS008538	250	394,472,884	FIN	SRR1293236/SRR1293262
HG03742	SRS352914	250	399,265,890	ITU	SRR1293251/SRR1293283
HG00759	SRS179211	250	396,523,555	CDX	SRR1293295/SRR1293326
HG01500	SRS074310	250	405,476,927	IBS	SRR1295423/SRR1298981
NA20502	SRS001670	250	318,668,689	TSI	SRR1295424/SRR1298988
HG02568	SRS368094	250	208,279,272	GWD	SRR1295425/SRR1298980
HG01565	SRS178998	250	421,765,162	PEL	SRR1295426/SRR1298989
HG03052	SRS344046	250	362,366,156	MSL	SRR1295432/SRR1295535
HG00419	SRS008605	250	365,091,105	CHS	SRR1295433/SRR1295554
NA20845	SRS003814	250	411,438,756	GIH	SRR1295465/SRR1295542
HG03642	SRS350861	250	326,040,894	STU	SRR1295466/SRR1295515
NA18525	SRS001364	250	386,942,089	CHB	SRR1295532/SRR1295539
HG01879	SRS212451	250	408,157,062	ACB	SRR1295533/SRR1295534
HG01595	SRS179136	250	404,889,598	KHV	SRR1295536/SRR1295552
NA18939	SRS000715	250	403,121,310	JPT	SRR1295537/SRR1295540
NA19625	SRS003634	250	335,694,770	ASW	SRR1295538/SRR1295545
HG02922	SRS344063	250	394,384,474	ESN	SRR1295543/SRR1295553
NA19017	SRS001488	250	393,888,768	LWK	SRR1295544/SRR1295546
HG03006	SRS350823	250	418,697,479	BEB	SRR1295568/SRR1295570

3.3.4 Investigation of common variants in HSTR loci

A list of common variants was downloaded from dbSNP (build 151) (Sherry, Ward et al. 2001). In this list, common variants are defined as follows; variants with a minor allele of frequency ≥ 1% in at least one population of the 1000 genomes project and for which two or more founders contribute to that minor allele frequency. The common variant was counted if it was located in any HSTR regions. The list of common variants does not include unplaced scaffolds, thus two HSTRs located in unplaced scaffolds were excluded. For comparison with non-HSTRs, 10,000 control sets were generated by random sampling TR sets that have the same number of TRs and minimum/maximum TR length the same as that of HSTRs.

3.3.5 Expression quantification of orthologous genes in human, chimpanzee, gorilla, orangutan and brain organoids

RNA-seq data of human, chimpanzee, gorilla, and orangutan were obtained from the Sequence read archive (SRA) database with accession number SRP007412 (Brawand, Soumillon et al. 2011). The data set consists of 48 tissue samples across frontal cortex, cerebellum, liver, heart, kidney, and testis (Table 3.2). The raw reads were quality checked and trimmed for low-quality regions and adaptor sequences using Trimmomatic v0.36 (Bolger, Lohse et al. 2014). The clean reads were aligned to the reference genomes (GRCh38, panTro4, gorGor4, and ponAbe2) using HISAT v2-2.1.0 (Kim, Langmead et al. 2015). We then quantified gene expression by counting the aligned reads using FeatureCounts with annotation file (.GTF) from Ensembl (Hubbard, Barker et al. 2002). To compare gene expression between species, 15,508 orthologous groups for human and other primates were determined by intersecting pairwise 1-to-1 orthologous genes between human and three NHP using Ensembl bioMart. For each orthologous group, RPKM (Reads per kilobase per million) and quantile normalization were performed in each tissues. We also used the gene expression dataset of cerebral organoids (52 human and 344 chimpanzee samples), which were obtained from GEO (Gene expression omnibus) with accession number GSE86207 (Mora-Bermúdez, Badsha et al. 2016) (Table 3.3). Because this dataset has been already normalized by RPKM method and includes orthologous group information, RPKM normalization and orthologous group identification were not performed for this dataset.

For each two RPKM normalized data, we independently performed re-normalization in each tissues using quantile normalization (Brawand, Soumillon et al. 2011, Barbash and Sakmar 2017). In each tissue, we calculated the logFC of HSTR genes by dividing the mean expression level of human by that of NHP (log2(Human/Other primates)). For comparing the logFC of HSTR genes with that of all genes, we resampled the same number of genes as HSTR associated genes, and calculated logFC. By repeating this process one million times, we constructed a null distribution and calculated empirical P-values by determining the position of HSTR logFC in that distribution.

Table 3.2 Information of RNA-sequencing data from SRP007412 (Brawand, Soumillon et al. 2011)

Sample Name	Species	Age	Sex	Tissue
SRS214076	Homo sapiens	NA	NA	Brain, prefrontal cortex
SRS214077	Homo sapiens	NA	Male	Brain, prefrontal cortex
SRS214078	Homo sapiens	NA	Male	Brain, prefrontal cortex
SRS214074	Homo sapiens	NA	NA	Brain, frontal cortex
SRS214075	Homo sapiens	NA	NA	Brain, frontal cortex
SRS214047	Pan troglodytes	44 years	Female	Brain, prefrontal cortex
SRS214048	Pan troglodytes	12.3 years	Male	Brain, prefrontal cortex
SRS214049	Pan troglodytes	35 years	Male	Brain, prefrontal cortex
SRS214050	Pan troglodytes	12.1 years	Male	Brain, prefrontal cortex
SRS214051	Pan troglodytes	6.7 years	Male	Brain, prefrontal cortex
SRS214052	Pan troglodytes	12 years	Male	Brain, prefrontal cortex
SRS214036	Gorilla gorilla	50 years	Female	Brain, prefrontal cortex
SRS214037	Gorilla gorilla	51 years	Female	Brain, prefrontal cortex
SRS214027	Pongo pygmaeus	56 years	Female	Brain, prefrontal cortex
SRS214028	Pongo pygmaeus	16 years	Male	Brain, prefrontal cortex
SRS214080	Homo sapiens	NA	Female	Cerebellum
SRS214081	Homo sapiens	NA	Male	Cerebellum
SRS214081	Homo sapiens	NA	Male	Cerebellum
SRS214038	Gorilla gorilla	50 years	Female	Cerebellum
SRS214039	Gorilla gorilla	51 years	Male	Cerebellum
SRS214029	Pongo pygmaeus	56 years	Female	Cerebellum
SRS214053	Pan troglodytes	44 years	Female	Cerebellum
SRS214054	Pan troglodytes	12 years	Male	Cerebellum
SRS214082	Homo sapiens	NA	Female	Heart
SRS214083	Homo sapiens	NA	Male	Heart
SRS214083	Homo sapiens	NA	Male	Heart
SRS214084	Homo sapiens	NA	Male	Heart
SRS214055	Pan troglodytes	44 years	Female	Heart

SRS214056	Pan troglodytes	12 years	Male	Heart
SRS214040	Gorilla gorilla	50 years	Female	Heart
SRS214041	Gorilla gorilla	51 years	Male	Heart
SRS214030	Pongo pygmaeus	56 years	Female	Heart
SRS214031	Pongo pygmaeus	21 years	Male	Heart
SRS214085	Homo sapiens	NA	Female	Kidney
SRS214086	Homo sapiens	NA	Male	Kidney
SRS214087	Homo sapiens	NA	Male	Kidney
SRS214057	Pan troglodytes	44 years	Female	Kidney
SRS214058	Pan troglodytes	12 years	Male	Kidney
SRS214042	Gorilla gorilla	50 years	Female	Kidney
SRS214043	Gorilla gorilla	51 years	Male	Kidney
SRS214032	Pongo pygmaeus	56 years	Female	Kidney
SRS214033	Pongo pygmaeus	21 years	Male	Kidney
SRS214088	Homo sapiens	NA	NA	Liver
SRS214088	Homo sapiens	NA	NA	Liver
SRS214089	Homo sapiens	NA	NA	Liver
SRS214059	Pan troglodytes	44 years	Female	Liver
SRS214060	Pan troglodytes	12 years	Male	Liver
SRS214044	Gorilla gorilla	50 years	Female	Liver
SRS214045	Gorilla gorilla	51 years	Male	Liver
SRS214034	Pongo pygmaeus	56 years	Female	Liver
SRS214035	Pongo pygmaeus	21 years	Male	Liver
SRS214090	Homo sapiens	NA	NA	Testis
SRS214091	Homo sapiens	NA	NA	Testis
SRS214061	Pan troglodytes	12 years	Male	Testis
SRS214046	Gorilla gorilla	51 years	Male	Testis

Table 3.3 Information of RNA-sequencing data from GSE86207 (Mora-Bermúdez, Badsha et al. 2016).

	Homo sapiens	Pan troglodytes
# of samples	52	344
Stage	60 days	56.59 days ± 10.12
Source	iPSC SC102-A1	iPSC SandraA (18.02%) iPSC PR818-5 (81.97%)
Tissue	ventricle	cerebral organoid (66%) ventricle (33%)

3.3.6 Gene set enrichment analysis

For gene sets containing or nearby a HSTR, enrichment tests were performed using DAVID functional annotation (Dennis, Sherman et al. 2003). P-value < 0.05 was used as the cutoff value of enrichment tests.

3.3.7 Multiple alignments data of 99 vertebrate genomes with human genome

Multiple alignments of 100 vertebrates (GRCh38) were downloaded from UCSC genome browser (Casper, Zweig et al. 2017).

3.3.8 ChIP-seq data for histone modification overlap with HSTRs

In order to identify histone modifications that are likely to be responsible for regulating gene expression in brain tissues, we used 49 bed files, from the ENCODE (Consortium 2007) database. Four histone marks were considered; H3K4me3 (n=15) and H3K9ac (n=7) for promoter activity, and H3K4me1 (n=15) and H3K27ac (n=12) for enhancer activity (Table 3.4). For each histone marks, consensus peaks across samples were generated by DiffBind (Stark and Brown 2011) with following options; minOverlap=2, summits=250.

Table 3.4 Information of 49 ChIP-seq data from ENCODE database (Consortium 2007)

File accession	Experiment accession	Region	Sex	Age	Histone mark
ENCFF073BUA	ENCSR608XIG	temporal lobe	female	75	H3K27ac
ENCFF236KXA	ENCSR123HEE	layer of hippocampus	female	75	H3K27ac
ENCFF464AKF	ENCSR195CFR	middle frontal area 46	male	81	H3K27ac
ENCFF470WBY	ENCSR355UYP	cingulate gyrus	male	81	H3K27ac
ENCFF495HNT	ENCSR494MDB	caudate nucleus	male	81	H3K27ac
ENCFF512SFD	ENCSR798RTU	caudate nucleus	female	75	H3K27ac
ENCFF719XFH	ENCSR380KOO	angular gyrus	male	81	H3K27ac
ENCFF740HPX	ENCSR014TDK	temporal lobe	male	81	H3K27ac
ENCFF756LUQ	ENCSR604JDV	cingulate gyrus	female	75	H3K27ac
ENCFF823HIP	ENCSR321LKT	layer of hippocampus	male	73	H3K27ac
ENCFF874BBA	ENCSR912TVO	layer of hippocampus	male	81	H3K27ac
ENCFF894UON	ENCSR016XBE	middle frontal area 46	female	75	H3K27ac
ENCFF437RXH	ENCSR724TOA	angular gyrus	female	75	H3K4me1
ENCFF430MTV	ENCSR253JFN	middle frontal area 46	female	75	H3K4me1
ENCFF452SPQ	ENCSR901SCD	temporal lobe	male	81	H3K4me1
ENCFF266IMI	ENCSR150GWJ	substantia nigra	female	75	H3K4me1
ENCFF462EUH	ENCSR303OIB	caudate nucleus	male	81	H3K4me1
ENCFF260KOR	ENCSR638HSG	layer of hippocampus	female	75	H3K4me1
ENCFF280IMI	ENCSR264YGM	caudate nucleus	female	75	H3K4me1
ENCFF356UEG	ENCSR559CSJ	substantia nigra	male	81	H3K4me1
ENCFF304IPZ	ENCSR479GJI	temporal lobe	female	75	H3K4me1
ENCFF712GAI	ENCSR117IJB	cingulate gyrus	female	75	H3K4me1
ENCFF629VBT	ENCSR443NEP	cingulate gyrus	male	81	H3K4me1

ENCFF813JPB	ENCSR263BGI	angular gyrus	male	81	H3K4me1
ENCFF860LKA	ENCSR051FXN	middle frontal area 46	male	81	H3K4me1
ENCFF859KZN	ENCSR797IXN	layer of hippocampus	male	73	H3K4me1
ENCFF747LBC	ENCSR040TUS	layer of hippocampus	male	81	H3K4me1
ENCFF379RQU	ENCSR840KVX	caudate nucleus	male	81	H3K4me3
ENCFF146OJJ	ENCSR551QXE	substantia nigra	female	75	H3K4me3
ENCFF159KZF	ENCSR418JIS	layer of hippocampus	female	75	H3K4me3
ENCFF446VYW	ENCSR693GVU	cingulate gyrus	female	75	H3K4me3
ENCFF733MXP	ENCSR477BHF	temporal lobe	female	75	H3K4me3
ENCFF386SAW	ENCSR883QMZ	substantia nigra	male	81	H3K4me3
ENCFF583LHL	ENCSR717AJD	temporal lobe	male	81	H3K4me3
ENCFF979YXX	ENCSR486QMV	caudate nucleus	female	75	H3K4me3
ENCFF408LCT	ENCSR032BMQ	cingulate gyrus	male	81	H3K4me3
ENCFF756UKX	ENCSR956CFX	layer of hippocampus	male	81	H3K4me3
ENCFF884KFF	ENCSR057RET	angular gyrus	female	75	H3K4me3
ENCFF535LNZ	ENCSR401VZL	middle frontal area 46	male	81	H3K4me3
ENCFF069XCJ	ENCSR535XRY	angular gyrus	male	81	H3K4me3
ENCFF084KQP	ENCSR157EML	middle frontal area 46	female	75	H3K4me3
ENCFF071XLQ	ENCSR383AEO	layer of hippocampus	male	73	H3K4me3
ENCFF413LSS	ENCSR058AUB	substantia nigra	female	75	H3K9ac
ENCFF647IZD	ENCSR591YVZ	angular gyrus	female	75	H3K9ac
ENCFF950UGM	ENCSR768DMK	temporal lobe	female	75	H3K9ac
ENCFF878QJG	ENCSR892RZN	layer of hippocampus	female	75	H3K9ac
ENCFF195JWO	ENCSR672FMZ	caudate nucleus	female	75	H3K9ac
ENCFF359FMD	ENCSR860PEV	cingulate gyrus	female	75	H3K9ac
ENCFF251GMK	ENCSR448MML	middle frontal area 46	female	75	H3K9ac

3.3.9 Scan for binding motif of transcription factor within HSTR sequences

To scan for transcription factor binding motifs in our HSTRs, we used position weighted matrices from the JASPAR 2016 database (Mathelier, Fornes et al. 2015), which consist of 386 transcription factor motif profiles implemented in TFBSTools (Tan and Lenhard 2016). Multiple testing problems were corrected with the Bonferroni method (Dunn 1961) and adjusted p-value < 0.05 was used as the cutoff value.

3.4 Results

3.4.1 Tandem repeats present only in the human lineage

To retrieve as many candidates as possible, we started with all possible TRs loci, detected by tandem repeat finder (TRF) on the human reference genome (GRCh38), which resulted in approximately one million loci (1,014,188) (Figure 3.1a). We next filtered for TR loci that have flanking sequences with orthologous blast hits in each NHP genome. This resulted in the elimination of 255,465 / 238,644 / 284,520 in the chimpanzee, gorilla and orangutan respectively (loci were eliminated since they had e-value above 10; Table 3.5, 1. Blast hit). The remaining loci (600,132 / 606,841 / 598,864) were then filtered for orthologous queries that covered the entire target sequence (Table 3.5, 2. Coverage). A further, 136,167 / 144,338 / 116,802 loci were removed using reciprocal best hits (RBH) analysis between flanking sequences of human and NHP TR loci (Table 3.5, 3. RBH). Only a few loci (386 / 375 / 1,220) had non-matching flanking sequences in both 3' and 5' directions and were removed (Table 3.5, 4. Pair). Consequently, 22,038, 23,999 and 13,472 loci in the chimpanzee, gorilla and orangutan respectively with shared orthologues between human and NHPs were retained.

After removing the loci that either have repeat count less than 2 (e. g. 1.5) or ambiguous base compositions (e. g. "N"), 17,840 / 19,287 / 10,326 orthologous loci remained. Among them, 23.8% / 26.2% / 30.5% were not detected as HSTRs after filtering for length and identity (Table 2), 69.8% / 66.2% / 56.4% due to Length criterion (Table 2, Length), and 2.7% / 3.9% / 8.2% due to Identity criterion (Table 2, Identity). Finally, 925 / 1,131 / 866 loci remained in each human vs NHPs comparisons (Table 2, HSTR), and 152 loci that were shared across all three comparisons were considered as HSTR loci (Figure 1C).

We then compared the HSTRs with the total TRs detected by TRF, in terms of unit length, number of repeats, total length, and percentage of match (Figure 3.2). The mean unit length of HSTRs (30.3bp) was longer than that of total TRs (28.4bp), and showed a length distribution from 11bp to 219bp. This shows that HSTRs did not include TRs that have a unit length shorter than 10bp. In the process of determining HSTR, 98.2% of microsatellites and 64.1% of all TRs (in average of three NHPs) were removed due to length criterion (Table 3.6 and Table 3.7). This resulted in the complete removal of microsatellites from the final list of HSTRs shared by all three NHPs. Most HSTRs were repeated 2 times (mean: 2.2, median : 2), whereas total TRs were repeated more than 2 times (mean: 11.7, median : 3). The total TRs contained much longer sequences (mean: 330.7 bp, median : 47 bp) than HSTRs (mean : 70.8 bp, median : 53 bp), as expected since the each total TRs contain more repeats than each HSTRs. The percentage match between repeat unit sequences was higher in HSTRs (average 97.3%) than in total TRs (average 88.7%), which was consistent in a comparison between HSTRs and control sets (see Materials and Methods, Figure 3.3).

Table 3.5 The number of loci satisfying the criteria for orthologous flanking sequences. The figures indicate the number of loci satisfying or not satisfying criteria in each filtering step (see Materials and Methods). RBH: reciprocal best hits

	1.Blast hit		**2.Coverage**		**3.RBH**		**4.Pair**	
	not satisfying	**satisfying**	**not satisfying**	**satisfying**	**not satisfying**	**satisfying**	**not satisfying**	**satisfying**
Chimpanzee	255,465	758,723	600,132	158,591	136,167	22,424	386	22,038
Gorilla	238,644	775,544	606,841	168,712	144,338	24,374	375	23,999
Orangutan	284,520	729,668	598,864	130,804	116,802	14,002	1,220	13,472

Table 3.6 The number of TR loci that did not pass criteria for determining HSTR. The figures in each column indicate the number of loci that did not pass each criteria. Length criteria are: 1) Human TR (HTR) length ≥ intervening sequence (IS) length in NHP *2 and 2) HTR length – IS length in NHP ≥ unit length of HTR. Identity selects loci with percent identity > 95% between sequence units of the HTR and IS in NHP. Table indicates the number of HSTRs that remain after filtering TRs through the previous three criteria ("Length & Identity", "Length" and "Identity")

	Length & Identity	**Length**	**Identity**	**HSTR**	**Total**
Chimpanzee	4,252(23.8%)	13,112(69.8%)	476(2.7%)	925(4.9%)	18,765(100%)
Gorilla	5,063(26.2%)	13,464(66.2%)	760(3.9%)	1,131(5.5%)	20,418(100%)
Orangutan	3,157(30.5%)	6,320(56.4%)	849(8.2%)	866(7.7%)	11,192(100%)

Table 3.7 The number of microsatellites (unit length ≤ 10bp) that did not pass criteria for determining HSTR

	Length & Identity	**Length**	**Identity**	**HSTR**	**Total**
Chimpanzee	32 (0.56%)	5,610 (98.8%)	6 (0.1%)	26 (0.45%)	5,674 (100%)
Gorilla	30 (0.51%)	5,784 (98.7%)	12 (0.2%)	29 (0.49%)	5,855 (100%)
Orangutan	51 (1.5%)	3,242 (97.2%)	8 (0.24%)	31 (0.93%)	3,332 (100%)

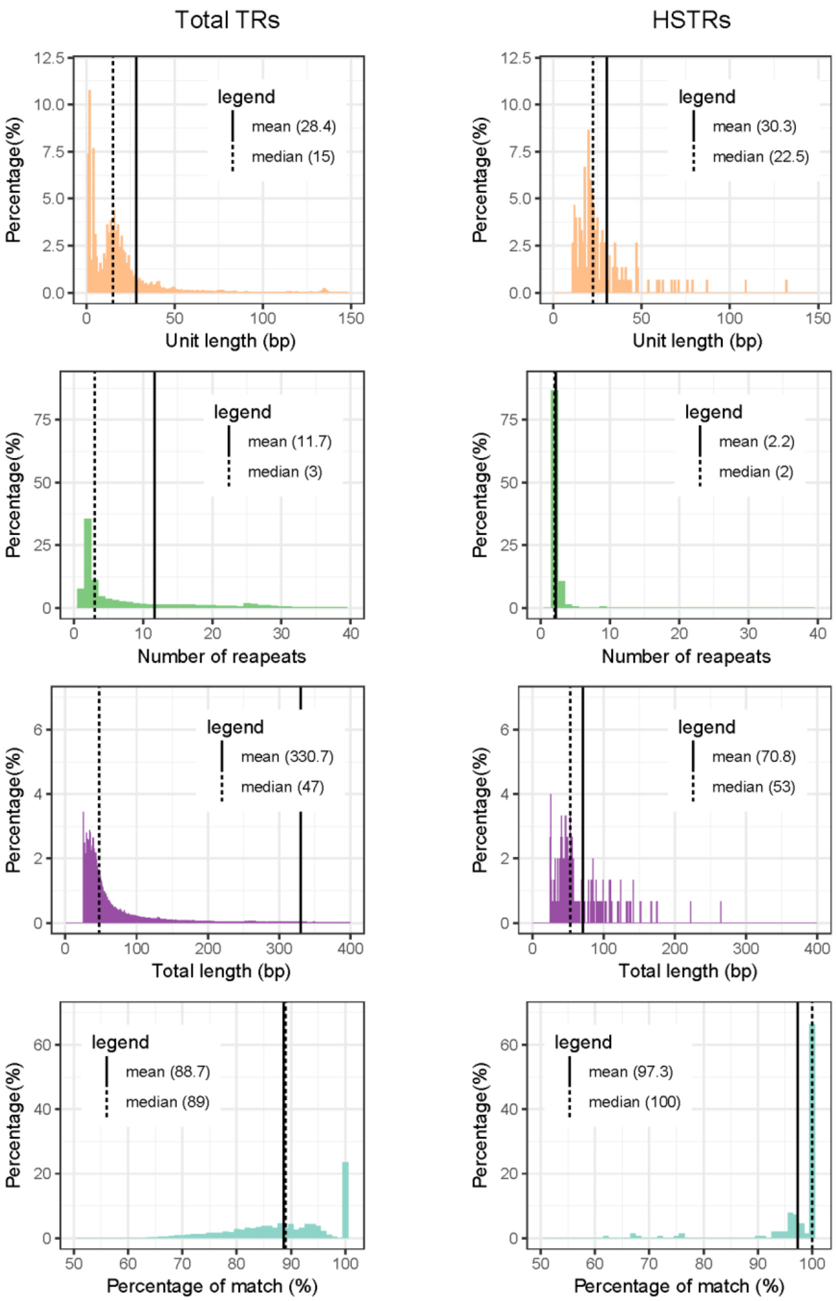


Figure 3.2 Characteristics of total TRs and HSTRs. Characteristics (Unit length, number of repeats, total length and percentage of match) of total TR and HSTR

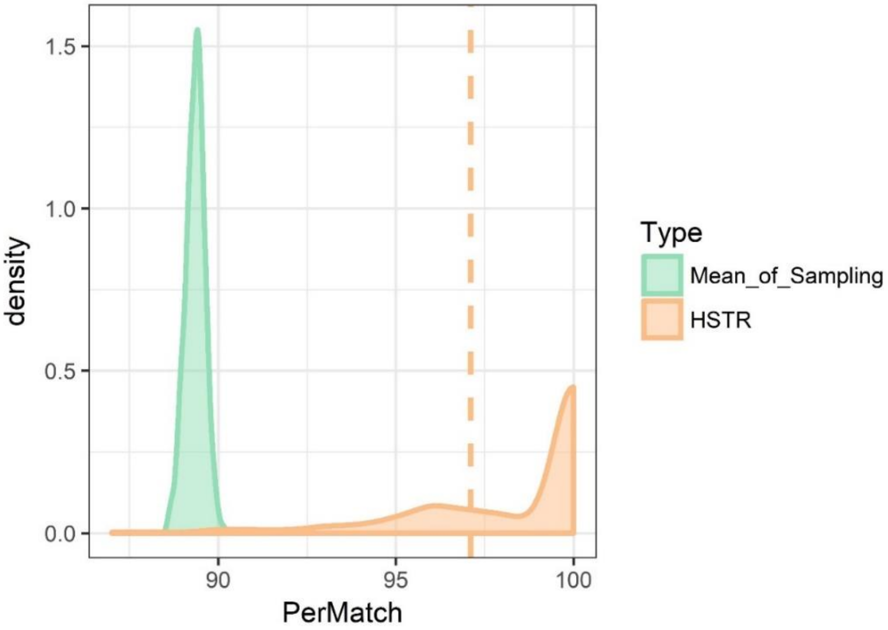


Figure 3.3 Distribution of bootstrapped percentages of matches. Dashed line indicates the mean percentage of match of HSTRs

3.4.2 Validation and fixation of HSTRs

If the HSTRs that we have detected de novo emerged in the human lineage, they should be absent in other species as well as in the three NHPs. In order to assess this, we examined the presence/absence of sequences at the HSTR loci in an additional 99 species. In a multiple alignment of 100 species from the UCSC database, the HSTRs showed extremely high frequency of ‘deletion’ (a nucleotide that is absent in one of species’ sequence in the alignment) compared to flanking sequences (Figure 3.4a, Wilcoxon rank sum test, p-value < 2.2e-16). As expected, all NHPs had a deletion at the exact orthologous position (Figure 3.4b and 3.4c).

It is highly likely that TRs associated with important genomic functions are fixed in the human population. We examined variations at the HSTR loci in human population data from the whole genome sequences of the 1000 genomes project (Consortium 2015). The sequence data from 24 individuals with high sequence coverage (153X on average) were used to reduce false negatives due to the lack of sequencing depth. Approximately 64% of the HSTRs sequences were identified in all 24 individuals (95 out of total 148 HSTRs, 4 loci that were longer than the read length (>250bp) were excluded from the 152 HSTRs, Figure 3.5a). Total 86.3% of HSTRs was identified in more than 10 individuals (sum of 64.1% found in all samples, an additional 4.7% were found in 23 samples, an additional 2% in 22 samples, an additional 2.7% in 21 individuals and 12.1% in 11-20 individuals). Only 9% of HSTRs (14 loci) were not detected in all individuals. The 24 individuals in the 1000 genomes project come from diverse genetic backgrounds and represent 24 different ethnic groups.

Nevertheless, 95 HSTR sequences were identified across all individuals. It is also worth noting that there may be more than 95 fixed HSTRs out of total 148 candidates but these have been removed because of our strict criteria applied for identification of intact HSTR sequences in the 1000 genomes project data.

We next investigated the incidence of common variants retrieved from dbSNP database within HSTR regions. In general, TRs along with other types of repeat sequences have higher mutation rates than other genomic regions (Vergnaud and Denoeud 2000). We generated control sets consisting of only TR sequences with various lengths using bootstrapping and compared them with the HSTRs. The control sets showed a clear trend that longer TRs contain more variants. In contrast, HSTRs showed much lower variation than expected for their mean length compared with TRs (Figure 3.5b). In addition, separate comparisons for genic and intergenic regions showed the same tendency (Figure 3.6). This shows that the low incidence of variation in HSTRs is not simply attributed to their genomic location.

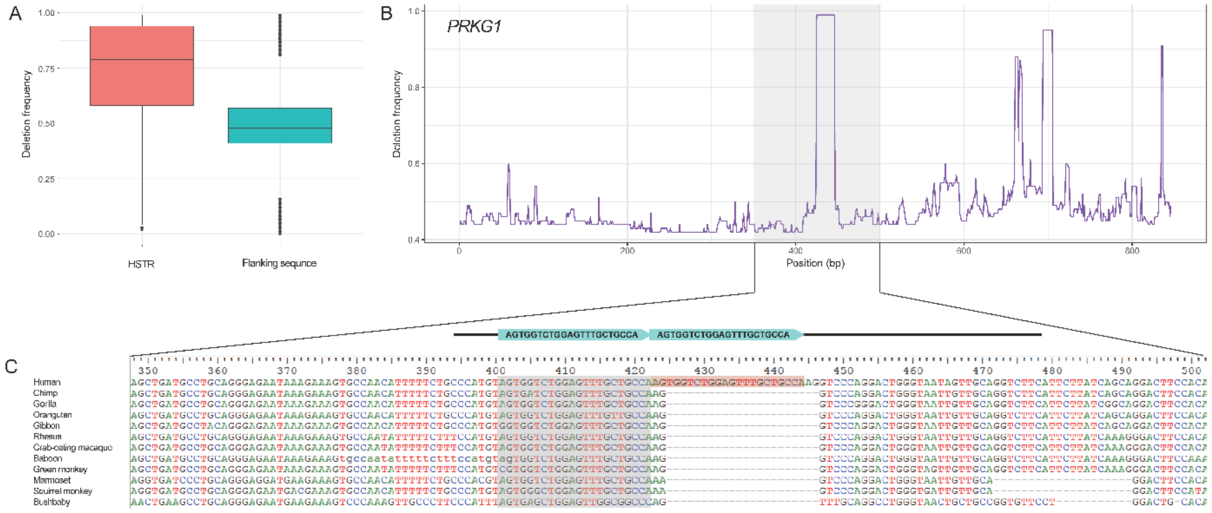


Figure 3.4 Deletion frequency of identified HSTRs and flanking sequences. (a) Per base deletion frequencies across 100 vertebrate's alignments between sequence of HSTRs and Flanking sequences. The 152 HSTRs have extremely high frequency of deletions compared to flanking sequences (Wilcoxon rank sum test, p-value < 2.2e-16). (b) An example of deletion frequency trend for the HSTR in PRKG1 gene along with flanking sequences. Almost all species have deletion at the HSTR region except human. (c) Multiple sequence alignment of shaded area in (b). Of 100 vertebrates, only the alignment of 12 primates were shown

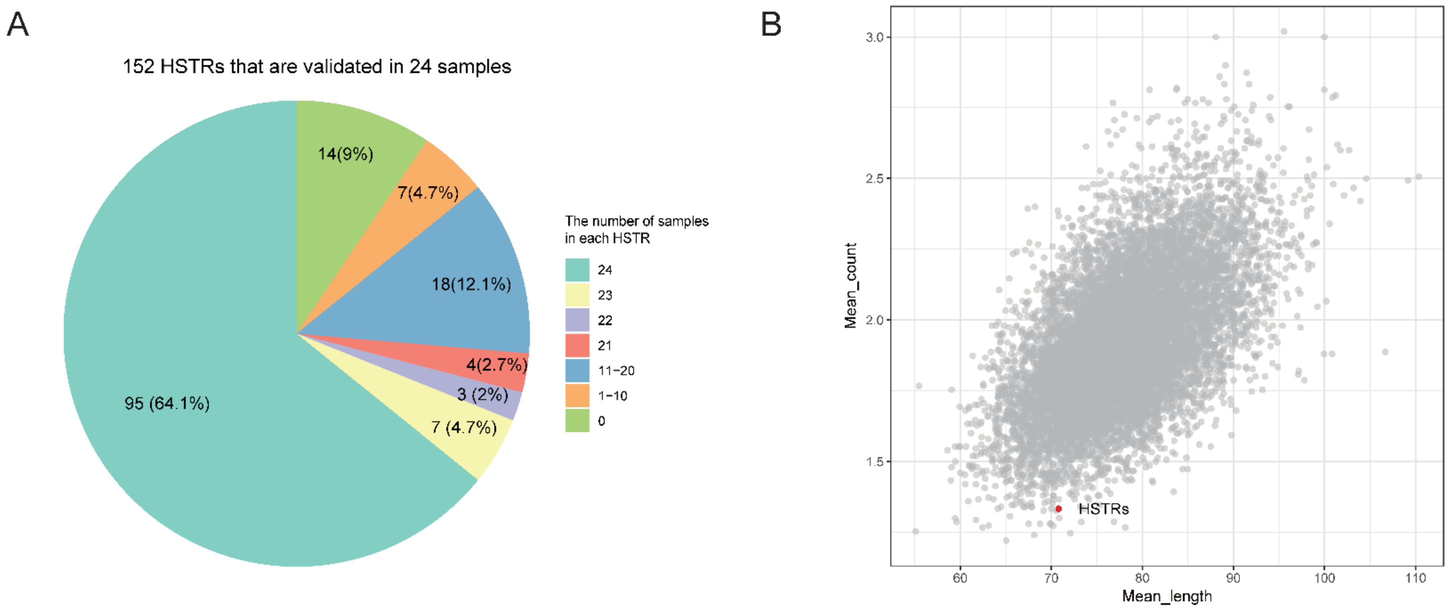


Figure 3.5 Validation and fixation of HSTRs. (a) Distribution of samples that have intact HSTR sequences. (b) Mean count of common variants within HSTRs. Control sets generated by resampling 10,000 times were plotted as grey dots, and the HSTR set was plotted as a red dot. X and Y-axis indicate mean length of TR and mean count of common variants for each sets, respectively

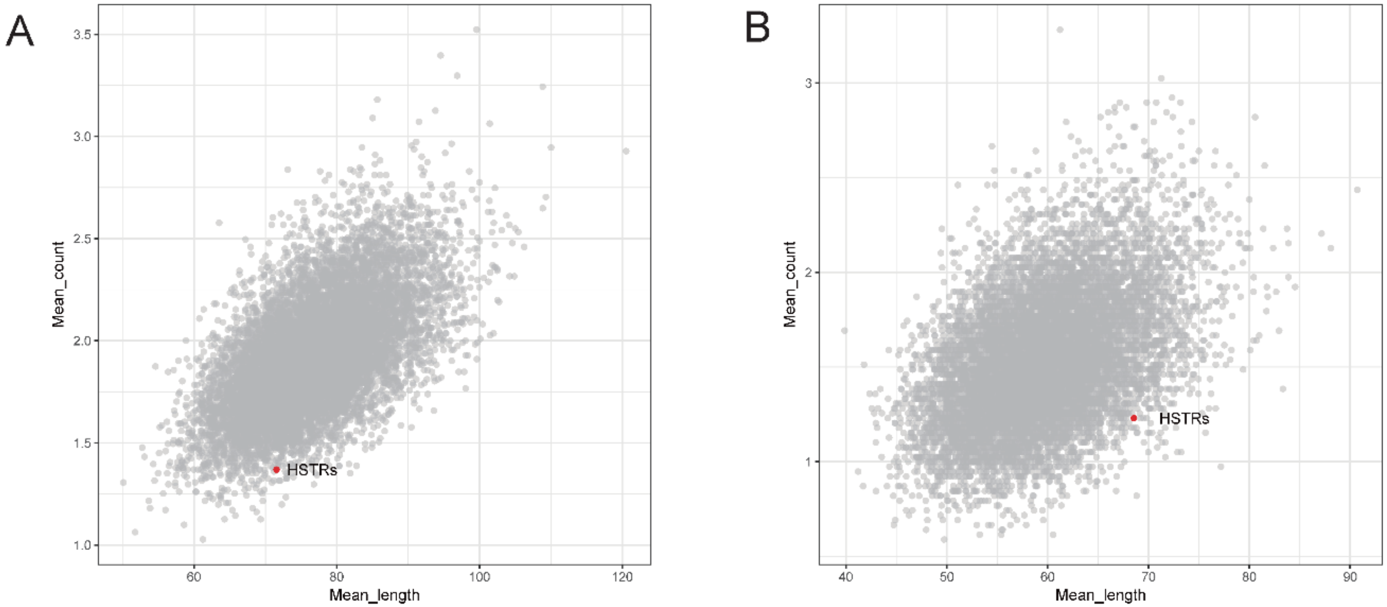


Figure 3.6 Mean count of common variants within HSTRs in (a) genic region and (b) intergenic region. Control sets generated by resampling 10,000 times were plotted as grey dots, and the HSTR set was plotted as a red dot. X and Y-axis indicate mean length of TR and mean count of common variants for each sets, respectively

3.4.3 Enrichment of HSTR in brain-related functions

To assess potential functionality of HSTRs, we first investigated their position within human genome (Figure 3.7a). Similar to a previous study (Legendre, Pochet et al. 2007), we the found majority of our HSTRs mapped within genic regions (72.4%). Out of the HSTRs within genic regions, 59.9 (91), 9.9 (15), and 2.6 (4) % were located in intron, exon, or intron-exon boundary respectively. The proportion of genic HSTRs located in introns was greater than that for total TRs (42 %). Of 15 HSTRs in exon, seven were located within the coding sequences.

Using gene annotation information for HSTRs, we performed gene set enrichment tests to identify potential functions using our full set of HSTRs. Out of the 152 HSTRs, 110 were assigned to at least one gene annotation, and 118 genes contained at least one HSTR. Out of the remaining 118 genes, 81 genes were assigned to Gene Ontology (GO) terms (Biological Process) and were used in gene set enrichment analysis (Figure 3.7b). The most significant GO term was "receptor localization to synapse (GO:0097120)". Also, among non-statistically significant results were several synapse-related terms; "innervation (GO:0060384)", "positive regulation of excitatory postsynaptic potential (GO:2000463)", "synaptic vesicle exocytosis (GO:0016079)". Intriguingly, three genes (*DYNC2H1*, *PRKG1*, and *SSTR1*) belonging to the statistically significant term, "forebrain development (GO:0030900)" were found. Moreover, as the HSTRs might have long distance effects on gene regulation, we also expanded the gene list to include those within 100kb of any gene TSS.

If multiple TSS are located within 100kb of any HSTR, the closest gene was selected. The expanded gene list consisting of 117 genes was still enriched in synapse-related or brain development, biological processes (Figure 3.8). 22 genes including 26 HSTRs were assigned to one of the genetic disorders in the OMIM (Online Mendelian Inheritance in Man) database, of which six genes were related to neurodegenerative disorders (*VPS53*), mental retardation (*TNIK*, *DPP6* and *KCNH1*), and brain-related diseases (*TRPC3* and *TDGF1*).

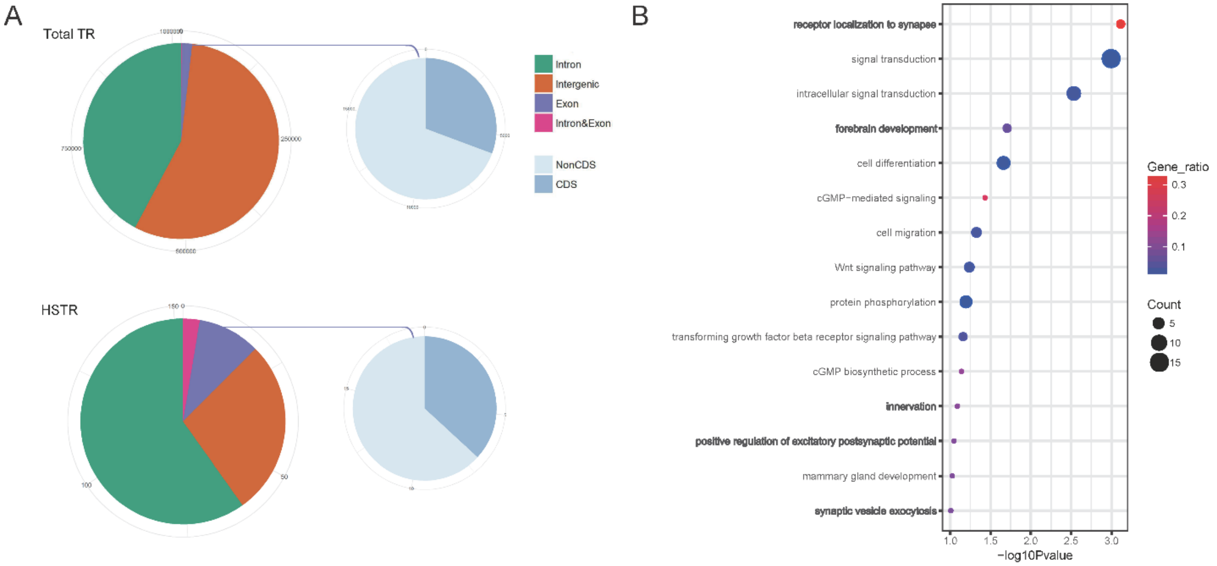


Figure 3.7 Genomic location and functional enrichment of HSTRs. (a) Genomic location of total TR and HSTR. The ratio for the location of TRs (Intron, Intergenic, Exon and boundaries of Intron and Exon) was illustrated with pie chart. Among TRs located in exon, the ratio of TR in CDS (coding sequence) and untranslated region (NonCDS) were also shown. (b) Result of gene set enrichment analysis for the gene set which contain HSTRs

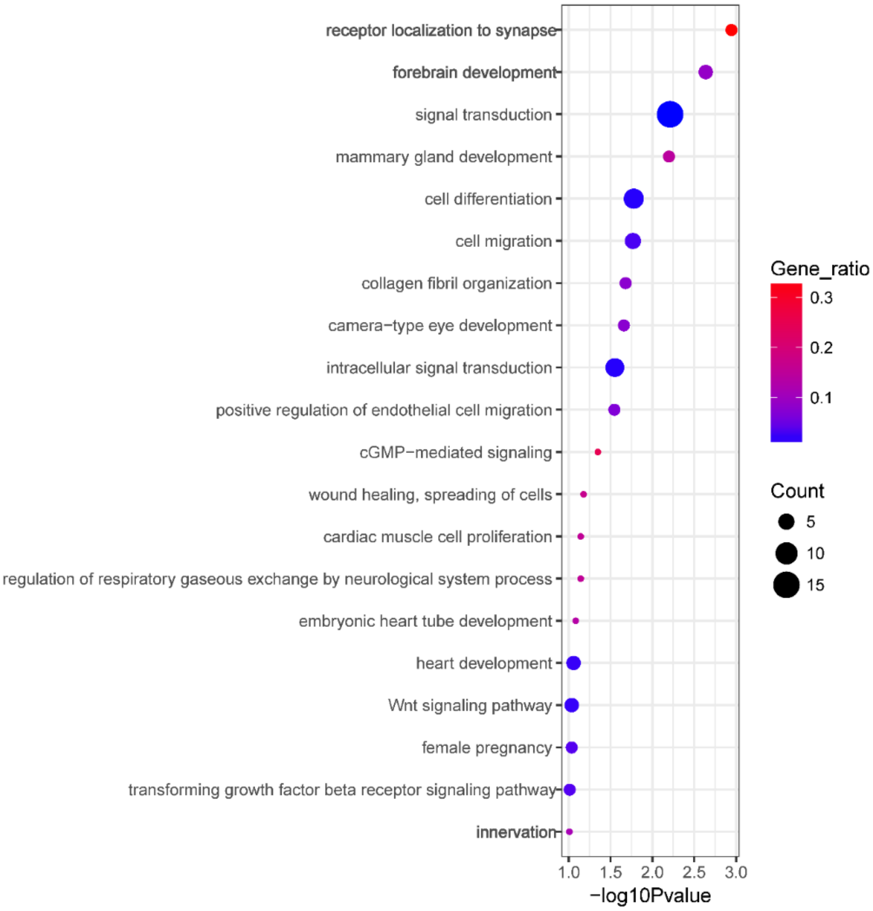


Figure 3.8 Gene set enrichment analysis for a gene set nearby HSTRs (within 100kb). Gene ratio indicates ratio for the number of genes to total number of genes in a gene ontology category

3.4.4 Brain-specific expression of HSTR associated genes

Given that TRs can affect the expression of nearby genes (Hamada, Seidman et al. 1984, Bennett, Lucassen et al. 1995, Warpeha, Xu et al. 1999, Streelman and Kocher 2002), we hypothesized that HSTR associated genes would show a greater degree of differential expression levels between human and NHP compared to non-HSTR associated genes. In order to examine this hypothesis, we used “log2 fold change” (logFC) as a measure of differential expression between human and NHP, and compared logFC of the HSTR associated gene set with that of the total orthologous gene set (see Materials and Methods). For investigating the expression of orthologues between humans and NHP, two sets of publicly available RNA-seq data were used. The first data set consists of 6 tissues from 4 species (human, chimpanzee, gorilla, and orangutan). Out of 15,508 orthologous gene sets defined in the 4 species, those genes which are located within 100kb from HSTR were selected; 82 orthologues were selected and the mean of logFC was calculated. In the entire tissues, most of mean logFCs were close to zero with small deviations. The mean values of logFC distributions in each tissues were -0.033 (Frontal cortex), -0.128 (Cerebellum), -0.166 (Liver), -0.116 (Heart), -0.082 (Kidney) and -0.072 (Testis). Strikingly, enhancement of HSTR gene expression was only observed in two brain tissues. The frontal cortex had the highest mean logFC of HSTR associated genes (Mean logFC: 0.204, Empirical p-value: 0.059), and Cerebellum had the second highest mean logFC (Mean logFC: 0.028, Empirical p-value: 0.15) (Figure 3.9). Increased gene expression was not observed in any other tissues with other tissues showing a decrease or similar expression (Liver: -0.378, Heart: -0.070, Kidney: -0.089, and Testis: -0.156).

Out of 16,377 orthologous gene sets defined in the second data set, those genes which are located within 100kb from HSTR were selected same as the first data set; 64 orthologues were selected and the mean of logFC was calculated. Similar to the first data set, the overall expression of the orthologues was higher in brain organoids from humans than chimpanzees (Mean logFC: -2.827, empirical p-value: 0.12, Figure 3.9). There were 68, differentially expressed HSTR associated genes (DEHG) between human and NHP (29 in the frontal cortex, 18 in the cerebellum, and 21 in brain organoids, p-value < 0.05). The mean logFCs of the DEHGs were 1.052, 1.062, and 1.474, respectively. Note that all the logFCs for the DEHGs were in the positive direction or zero, that is, the expressions of DEHGs are consistently higher in human than those of NHP in three brain tissues.

The results from the brain transcriptome and brain organoid data consistently show the increased expression of HSTR associated genes specifically in brain tissues. Comparisons with non-HSTR associated loci also supports the conclusion that altered expression is likely associated with HSTRs.

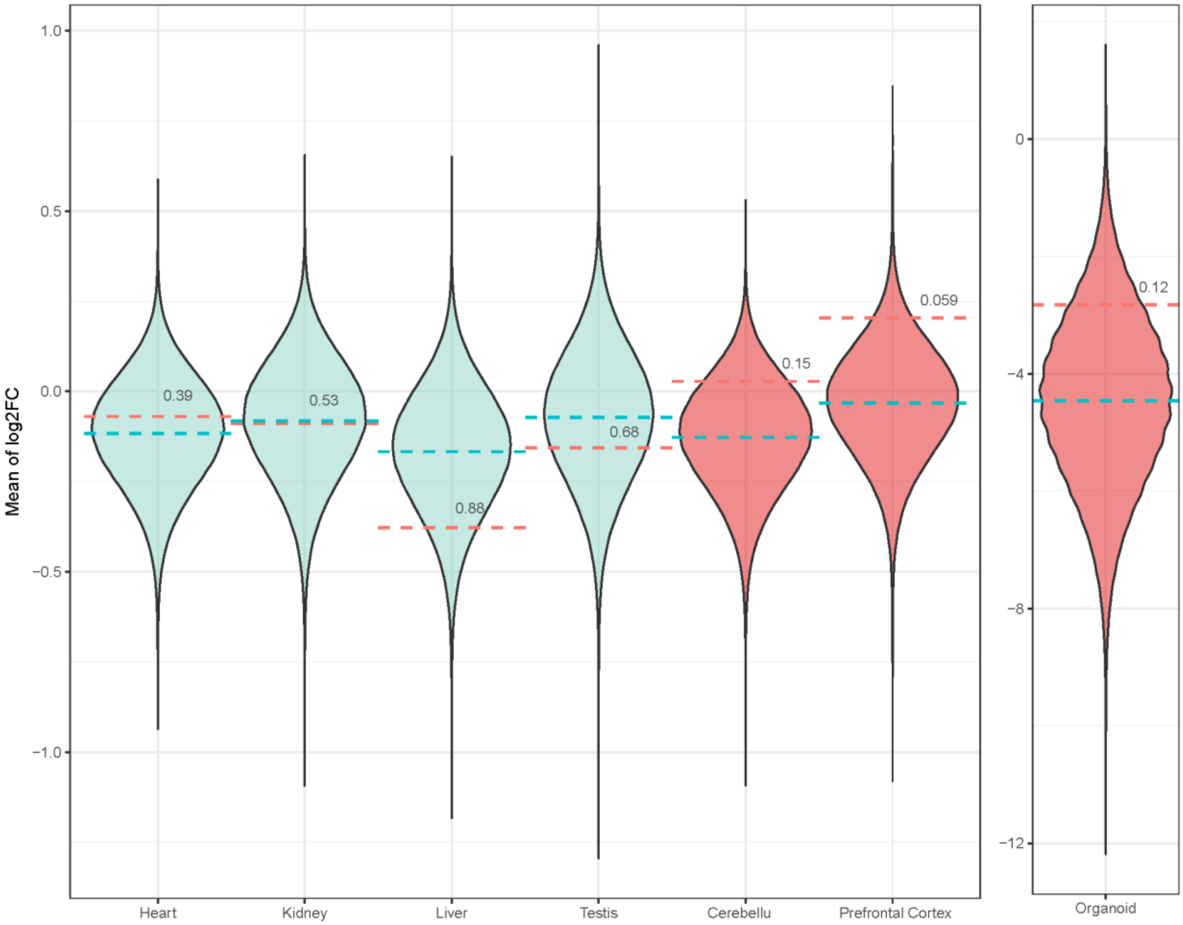


Figure 3.9 Expression of HSTR genes across seven tissues. Distributions of mean logFC (log2 fold change) for random samples in seven tissues. Dashed line colored by red indicates mean logFC for total orthologous gene sets. Dashed line colored by blue indicates mean logFC for HSTR gene set

3.4.5 Potential roles of HSTRs in regulation of gene expression

The possibility of HSTRs causing altered gene expression led us to investigate their potential regulatory role. We first investigated signals of histone modification within the HSTR regions. 49 ChIP-seq data-sets produced from human brain tissues were retrieved from the ENCODE database (Consortium 2007) and used for checking the potential promoter (H3K4me3 and H3K9ac) and enhancer (H3K4me1 and H3K27ac) activity of HSTR regions (Table 3.4). For each histone mark, we found 4 (H3K4me3), 15 (H3K9ac), 22 (H3K4me1) and 22 (H3K27ac) peaks, which covered ~28% of all HSTRs. We then investigated HSTR sequences to identify potential binding motifs for transcription factors. Although the types of binding motifs were varied, ~35% of all HSTRs contained at least one binding motif.

3.5 Discussion

The brain is one of the most active evolutionary sites in humans. Humans have the largest brain among extant primates with specialized neuronal connections (Hill and Walsh 2005, Sousa, Meyer et al. 2017), which lead to superior abilities related to cognition and behavior such as long-term planning and the capacity to create art (Sousa, Meyer et al. 2017). As such, the brain is regarded as the core component of the human identity. However, how the human brain evolved from that of closely related primates is not fully understood. Several studies suggested gene expression changes as one of the driving forces for the distinctive attributes of human brain function (Cáceres, Lachuer et al. 2003, Gu and Gu 2003, Carroll 2005, Khaitovich, Lockstone et al. 2008). However, the reason for difference in gene expression across different species is still an open question.

Recent studies have examined TR as a potential driver of altered gene expression. Their unstable nature facilitates rapid evolutionary events and can cause variations in gene expression and function within or between species (Gemayel, Vinces et al. 2010). Divergence in gene expression is observed in the presence or absence of TR using six tissues across four primates (Sonay, Carvalho et al. 2015). In this study, we examined the distribution and potential functional consequences of TRs that emerged de novo in the human lineage (HSTRs). The HSTRs we found showed a low number of repeats compared to total TRs in the human genome. As we did not limit the number of repeats in any of our steps for detecting HSTR, two unit repetitions in our HSTR is not likely to be caused by a systematic bias. Rather, it may be a unique characteristic of TRs that recently emerged. A previous study (Ahmed and Liang 2012) suggested that older TRs have more repeats than younger ones. In addition, the higher percentage match between repeats in the HSTRs provides further support for their recent emergence (Kumar and Subramanian 2002). Additionally, the rareness of HSTR across 100 species suggests that the HSTRs not only recently emerged, but also *de novo* emerged after split from chimpanzee.

HSTRs were enriched within or nearby genes related to brain function. This is consistent with previous studies (Legendre, Pochet et al. 2007) examining unfiltered TRs . In addition, we observed an enrichment in synapse related functions and showed that the presence of HSTRs was associated with increased gene expression, exclusively in human brain tissues.

Approximately 28% of HSTRs overlapped with histone marks associated with promoter or enhancer functions in brain tissues and ~35% of HSTR sequences contained at least one transcription factor binding motif (Bonferroni corrected p-value < 0.05). These results were consistent with a role for TR's in cis-regulatory elements such as promoters or enhancers (Usdin 2008, Gemayel, Vinces et al. 2010). Considering the genomic location of HSTRs (Most of HSTRs were located in intron region), there might be other underlying mechanisms. Several previous studies suggested intron size as one factor related to splicing (Wieringa, Hofer et al. 1984, Bell, Cowper et al. 1998, Pai, Henriques et al. 2017). Because most of our HSTRs have a relatively long unit length, the influence of TR expansion on the splicing of messenger RNAs is worth examining in evolutionary context. Moreover, among HSTRs located in exonic regions, the majority were contained in either the first or last exon. This may be due to their role in regulating gene expression as the first and last exons are related to the recognition of regulatory elements and polyadenylation processes respectively (Bieberstein, Carrillo Oesterreich et al. 2012, Matoulkova, Michalova et al. 2012).

Although more functional investigations are required, we found a large number of HSTRs which might play a role in human brain evolution (Figure 3.10). However, the HSTRs described in this study do not account for entire set in the human genome. We used flanking sequence similarity to detect orthologous regions in NHP thus, we could only detect HSTRs in highly conserved regions. This caveat can explain the enrichment of HSTRs in genic regions. In addition, we only examined TRs with a fixed number of repeats, there might be HSTRs where the number of repeats are not fixed in length, but which show a defined range of repeat units which are unique to humans. For example, the DUF1220 protein domain, which is related to brain size (Dumas, O'Bleness et al. 2012), has a human-specific copy number range (human have ~300 copies, while chimp and gorilla have 95~140 copies) (O'Bleness, Dickens et al. 2012, Zimmer and Montgomery 2015). Hence, the list of HSTR we analyzed should be considered as a subset of the entire HSTRs. Finally, the imbalance of annotation information between human and NHP along with incomplete reference genomes limits our expression analysis, especially in defining the orthologous gene sets. More refined reference genomes and annotations might reduce the missing orthologues.

To the best of our present knowledge, this is the first study to suggest a relationship between de novo emerged HSTRs and gene expression. Moreover, our findings provide novel insights into the role of TR emergence in human brain evolution. Our computational approach can provide a useful mechanism for discovering additional candidate TRs responsible for phenotypic variation and help us further understand their contribution to the regulatory landscape.

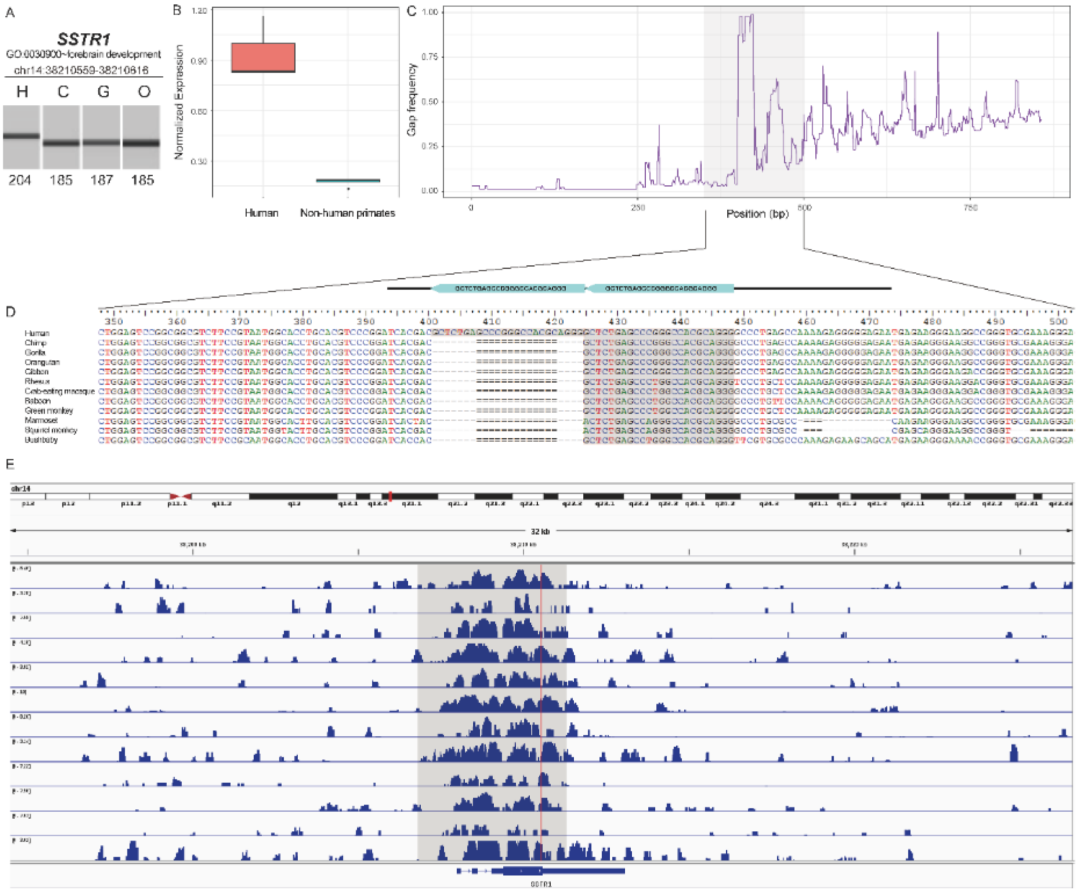


Figure 3.10 An example of candidate HSTR that might be related to human brain function. (a) *in vitro* validation of HSTR in *SSTR1* gene. We amplified the HSTR loci within *SSTR1* gene by PCR using human, chimpanzee, gorilla and orangutan genomic DNA samples and confirmed that the PCR products were longer only in the human sample. The numbers below indicate fragment length. (b) Boxplot of normalized expressions in cerebellum tissue (logFC : 2.4328, Wilcoxon rank sum test p-value : 0.0339). (c) Gap frequency distribution for the HSTR in *SSTR1* gene along with flanking sequences. Almost all species have gap at the HSTR region except human. (d) Multiple sequence alignment of shaded area in (c). Of 100 vertebrates, only the alignment of 12 primates were shown. (e) Fold enrichment over control at the HSTR region in *SSTR1* gene for Histone modification H3K27ac. Shaded area indicates known enhancer region (GH14I038206) in GeneHancer database (Fishilevich, Nudel et al. 2017). Red line indicates HSTR location in *SSTR1* gene

This chapter was published in *PLoS ONE*
as a partial fulfillment of Kwondo Kim's Ph.D program.

Chapter 4. Artificial selection increased body weight but induced increase of runs of homozygosity in Hanwoo cattle

4.1 Abstract

Artificial selection has been demonstrated to have a rapid and significant effect on the phenotype and genome of an organism. However, most previous studies on artificial selection have focused solely on genomic sequences modified by artificial selection or genomic sequences associated with a specific trait. In this study, we generated whole genome sequencing data of 126 cattle under artificial selection, and 24,973,862 single nucleotide variants to investigate the relationship among artificial selection, genomic sequences and trait. Using runs of homozygosity detected by the variants, we showed increase of inbreeding for decades, and at the same time demonstrated a little influence of recent inbreeding on body weight. Also, we could identify ~0.2 Mb runs of homozygosity segment which may be created by recent artificial selection. This approach may aid in development of genetic markers directly influenced by artificial selection, and provide insight into the process of artificial selection.

4.2 Introduction

Artificial selection creates genetic signatures on a genome as well as alteration of phenotypes (Innan and Kim 2004, Flori, Fritz et al. 2009, Gouveia, Silva et al. 2014). These genetic signatures refer to any types of sequence alteration which can be generated from selection process. For instance, extended linkage disequilibrium (LD), and reduced nucleotide diversity are examples of the genetic signatures (Wright, Bi et al. 2005, Consortium 2009). In the artificial selection process, one of the factors that give rise to such genetic signatures is inbreeding, which is the production of offspring from mating or breeding of individuals that are genetically close (Allard, Jain et al. 1968). Breeders allow only certain individuals with desirable characteristics to reproduce over the course of decades resulting in intensive inbreeding (Robertson 1961).

Runs of homozygosity (ROH) refers to the contiguous homozygous segment in the genome of an individual (McQuillan, Leutenegger et al. 2008). During inbreeding, identical-by-descent tracts could be formed in the genome of an individual, which refers to 'autozygous segment'. The homozygosity of the tract refers to 'autozygosity'. Although ROH segment could include randomly created segments, when the length is long enough it has been commonly regarded as autozygous segment. One of the most important features of autozygous segments is that their length shortens over generations due to meiotic recombination (Kirin, McQuillan et al. 2010). For this reason, by controlling the length threshold, it is possible to obtain the autozygous segment which was created during the period we are interested in. This was recently demonstrated using simulated data (Howrigan, Simonson et al. 2011). Several studies have analyzed ROHs including the autozygous segments, to detect recent artificial selection signatures (Pryce, Haile-Mariam et al. 2014, Kim, Sonstegard et al. 2015). The general process of artificial selection is accompanied by inbreeding and the degree of inbreeding increases as artificial selection progresses (Kim, Sonstegard et al. 2015). Therefore, it was also possible to investigate the change of autozygosity level as signatures of recent artificial selection by using genotype data which was accumulated for decades (Kim, Sonstegard et al. 2015). In addition to detecting the signature of artificial selection, ROH have been used as genetic markers for complex traits as well as Mendelian traits. For instance, ROH was suggested as the risk factor for schizophrenia in the human population (Keller, Simonson et al. 2012) and used to detect inbreeding depression for two reproductive traits in the pig population (Saura, Fernández et al. 2015).

Hanwoo cattle originated from natural crossbreeding between taurine and zebu cattle in Korea (Lee, Chung et al. 2013). In earlier times of its 5000 years history, Hawoo was used as draft animal and was not artificially selected (Lee, Park et al. 2014). In the early 1900's, Hanwoo was selected by appearance traits including body type and hair color to determine the breed, following the guideline established by the government agency (Park, Choi et al. 2013). Since the establishment of selection purpose for beef cattle in 1963 by Livestock Industry Act (LIA), Artificial selection of Hanwoo commenced along with the development of AI (Park, Choi et al. 2013). In 1974, performance tests was developed to select cattle with superior characteristics as beef cattle (Park, Choi et al. 2013). In 1980s, progeny test, which is a method that tests cattle by examining the economic traits observed from their offspring, was introduced by Hanwoo Improvement Center (HIC) (Park, Choi et al. 2013). Since the introduction of the two tests, Hanwoo breeding program has expanded and bulls with the proven quality could be produced. In the current performance test, young bulls from 6 to 12 month of age are selected through genetic evaluations for yearling weight (YW) and marbling score (MS) (Park, Choi et al. 2013). Semen collected from the selected young bulls is then inseminated in cows whose lineage is known (Park, Choi et al. 2013). Progenies of the cows are weighed at 6, 12, 18 and 24 months of age, and harvested at 24 months of age to investigate carcass traits including carcass weight (CW), eye muscle area (EMA), back fat thickness (BF) and MS. Selection of proven bulls is then conducted, based on a selection index which is measured by using weighted breeding values for BF, MS and EMA (Lee, Park et al. 2014). Although the current breeding program of Hanwoo cattle has been established recently, their economic traits has improved overall. Especially, YW remarkably increased from 315.54 kg in 1998 to 355.06 kg in 2011, resulting in about 40 kg of improvement over 13 years (Park, Choi et al. 2013).

For decades, Hanwoo breeding program has selected a subset of Hanwoo population, and the selected bulls have been used as breeding population through AI. Therefore, it is likely that there has been increase of inbreeding, considering similar cases of cattle population (Andersen 1966, Cleveland, Blackburn et al. 2005). In case of Hanwoo population, there was a report that the averaged inbreeding coefficients increased by 0.3% from 1997 to 2007, when a total of 1,123,162 cattle was investigated (Hwang, Park et al. 2009). While the inbreeding is inevitable during artificial selection or breeding, it has also deleterious effect on reproductive or production traits of individual (Sanchez, Toro et al. 1999) due to increased homozygosity at loci carrying rare recessive deleterious alleles or exhibiting overdominance (Charlesworth and Willis 2009, Huisman, Kruuk et al. 2016). In Hawoo, there was a report that inbreeding affected body weights at birth or at weaning in a negative way (Park YI 1969). However, the degree of inbreeding has been known to be lower than other commercial cattle breeds so far (Porto-Neto, Lee et al. 2014). For this reason, in Hanwoo population, it is important to control genetic diversity of the population to manage deleterious effect of inbreeding resulted from intensive artificial selection.

In this study, we aimed to investigate the changes of inbreeding in Hanwoo population during the decades of selection program, and identify genomic regions related to inbreeding induced by artificial selection and a trait (body weight). Especially, Hanwoo population used in this study has been intensively selected for decades to improve their economical traits including the body weight of a calf. Therefore, identifying the genomic regions affecting body weight during breeding program is important. In order to accomplish this, whole genome sequencing data of 126 cattle which were selected by a breeding program during ~20 years were generated. After defining autozygous segments on the whole genome data, we demonstrated the change of autozygous level in the population. Also, two statistical analyses were performed between three elements (artificial selection, autozygosity level, and body weight) using two regression models. Using the combined results of the two regression analysis, we attempted to demonstrate the relationship between these three elements.

4.3 Materials and Methods

4.3.1 Ethics statement

No ethics statement was required for the collection of DNA samples. DNA was extracted either from artificial insemination bull semen straws or from blood samples obtained from the Hanwoo Improvement Center of the National Agricultural Cooperative Federation (HICNACF) with the permission of the owners. The protocol was approved by the Committee on the Ethics of Animal Experiments of the National Institute of Animal Science (Permit Number: NIAS2015-774).

4.3.2 Sample information and whole genome sequencing

Blood samples for whole genome sequencing were obtained from 136 Korean beef cattle, (Hanwoo), 126 of which were bulls that were selected from HICNACF. The breeding program consisted of two steps; Performance and Progeny test of candidate bulls. In the first step, 66 of 900 bulls were selected by their weighted breeding values of economic traits, weight at 12 months, and marbling score. In the second step, 4,500 cows were inseminated with the semen from the selected bulls and 30 bulls were selected based on selection index which used weighted breeding values for BF, MS, and EMA of 800 male calves (Lee, Kim et al. 2013, Park, Choi et al. 2013, Lee, Park et al. 2014). Our 126 bulls have been selected through this breeding program from 1998 to 2015.

The unselected cattle (n=10) reared in Hanwoo Research Institute in National Institute of Animal Science had never been involved and selected in Hanwoo breeding program.

Using DNA from the blood samples, we produced indexed shotgun paired-end (PE) libraries with approximately 500bp inserts using TruSeq Nano DNA Library Prep Kit (Illumina, USA) following standard Illumina sample-preparation protocol. Briefly, 200 ng of gDNA was fragmented by Covaris M220 (Woburn, MA, USA) resulting in a median fragment size of ~500 bp followed by end repair, A-tailing, and indexed adapter ligation (~125bp adapter). For the next step, the gel-based size selection was done in the range of 550 to 650 bp for the adapter ligated DNA and PCR amplification was performed for 8 cycles in the case of library. The size-selected libraries were analyzed by an Agilent 2100 Bioanalyzer (Agilent Technologies) to determine the size distribution and to check for adapter contamination. The resulting libraries were sequenced in Illumina HiSeq 2500 (2x125bp paired-end sequences) and NextSeq500 (2x150bp paired-end sequences) sequencer.

4.3.3 Sequence read mapping and variant calling

A quality control for per-base quality of reads and removal of potential adaptor sequences was performed using fastQC (Andrews 2010) and Trimmomatic (Bolger, Lohse et al. 2014) software, respectively. Then, high quality sequence reads were mapped to the Bos taurus reference genome (UMD 3.1.78) using

Bowtie2 (Langmead and Salzberg 2012) with default settings. A series of downstream processes were performed to improve the quality of called variants as well as sequence alignment: Picard tools (http://picard.sourceforge.net) was used to sort reads, remove potential PCR duplicates, and ensure the mate pair information of paired-end reads. SAMtools (Li, Handsaker et al. 2009) was used to create index files for the reference and bam files. Genome Analysis Toolkit (GATK) (McKenna, Hanna et al. 2010) was used to conduct local realignment of sequence reads to correct misalignments aroused by small insertion and deletion. Also, base quality scores were recalibrated to obtain more accurate quality scores and correct the variation in quality with machine cycle and sequence context. Lastly, for variant calling and filtering step, "UnifiedGenotyper" and "SelectVariants" arguments implemented in GATK were used. High quality variants were retrieved by employing following criteria: The variants with 1) a Phred-scaled quality score < 30, 2) read depth < 5, 3) MQ0 (total count across all samples of mapping quality zero reads) > 4; or a 4) Phred-scaled P-value using Fisher's exact test > 200 were filtered out to reduce false positive calls.

4.3.4 Detection of ROH and autozygous segments

Using the information of genotypes provided by variant calls, ROH for each sample was identified using vcftools (Auton, Bryc et al. 2009, Danecek, Auton et al. 2011). As all of the genotypes were obtained from bulls, the homozygosity segments within only autosomal regions were considered. To ensure that the detected ROHs had not been generated by random events but by recent inbreeding, the proportion of ROHs in the population was investigated according to change of ROH length threshold. From this, the threshold of ROH length was set to 500kb, and ROHs shorter than 500kb were filtered out. We regarded the remaining ROHs as autozygous segments for the following downstream analysis. The proportion of these autozygous segments in UMD 3.1 reference genome (F_{roh}) was calculated and comparison was made between groups by Wilcoxon's rank sum test.

4.3.5 Detection of inbreeding and selection signatures

Inbreeding coefficients of each individual based on SNP (F_{snp}) were estimated as follows. First, whole SNP data set was pruned using the option --indep-pairwise implemented in plink 1.9 (Purcell, Neale et al. 2007) with three parameters (window size=50, step size=5, and r2=0.5). Then, individual Fsnp was estimated from the data set using the option –het. LD between pairs of markers were assessed using plink 1.9 (Purcell, Neale et al. 2007). The r2 value was calculated between all pairs of SNPs within 20 kb (--r2 and --ld-window-kb parameters). Moving averages of the pairwise r2 were then carried out with 5-kb steps.

To detect recent selection signatures, we used SHAPEIT v2 (Delaneau, Marchini et al. 2012) to infer the haplotype phase and impute missing alleles for the SNP data set generated after filtering out SNPs based on genotype missing rate > 0.05, and minor allele frequency < 0.01. Then, integrated haplotype score (iHS) was calculated with the Selscan (Szpiech and Hernandez 2014) using the default settings except --maf 0.01 option. The raw iHS was standardized by 'norm' module implemented in Selscan with 100 frequency bins.

4.3.6 Statistical analysis to identify candidate genomic region

To identify genomic regions associated with inbreeding induced by artificial selection, traits or both, an association test with regression analysis was performed. The unit for the association test was determined to be 10Mb bins, and *Bos taurus* autosome (BTA) were divided into 269 bins of 10Mb. After filtering out the bins where all samples had an equal ROH length, 264 bins were finally used for statistical analysis.

In the first step of regression analysis, an association test between artificial selection and ROH was performed for each bin (Analysis 1). The progress of artificial selection was represented by 'KPN (Korean Proven Bull) number'. KPN number is a registration number given to a bull finally selected through the breeding program. The breeding program has been performed approximately once a year. Accordingly, KPN numbers are highly correlated to the birth years of the selected bulls (Spearman's correlation test, $\rho = 0.9992$, p-value < 2.2e-16). The data including birth year records were not available for some of the samples and the exact time of birth was not necessary for the measure of relative progress of artificial selection. Thus, KPN number which is available in all samples was used instead of birth year. Consequently, we performed statistical analysis between ROH status and KPN number with the logistic regression model (Kim, Sonstegard et al. 2015). In the model, indicator of the ROH status (0: no ROH, 1: at least one ROH) and KPN number were considered as response and explanatory variables, respectively.

In the second analysis, the mixed effect model was used to analyze an association between ROH length and body weight for each bin (Analysis 2). In the model, ROH length and facility where the sample was raised from, were considered as fixed effects. To adjust background genetic effect on the model, genomic relationship matrix (GRM) generated by GCTA tool (Yang, Lee et al. 2011) was considered as random effect. Statistical test was performed using asreml and wald function implemented in ASReml-R package (Butler, Cullis et al. 2009), where Body weight at 12 month was used as response variable. In these two analyses, we determined significance of each bin with p-value < 0.01.

4.3.7 Validation of ROH segments using additional dataset

To confirm the existence of ROH segments in candidate regions, we additionally generated whole genome sequencing data of 77 Hanwoo cattle whose KPN numbers range from 634 to 1017, by using the same sequencing procedure. The ROH segments of the additional data were detected by the identical process which was used for the original dataset.

4.4 Results

4.4.1 The increase of genome-wide autozygosities in 126 cattle under selection

As the length of ROH which is autozygous, decreases over generations, it is necessary to set a suitable threshold of ROH length in order to investigate the effects of recent inbreeding on a genome. Although previous studies on ROH of cattle population suggested several criteria for defining autozygous segments (Pryce, Haile-Mariam et al. 2014, Kim, Sonstegard et al. 2015), these were not applicable in our case due to the characteristics of Hanwoo population as well as differences in the data generation platform.

Instead, we first investigated the average count or length of ROH with controlling the threshold of ROH length in our data (Figure 4.1). Average count and length of ROH at a 500kb threshold were ~46.13 and ~33Mb, respectively. Although it is difficult to directly compare the detected ROHs to a previous study (Lee, Chung et al. 2013), these count and length of ROHs were relatively higher. This can be partially explained by the difference of method used to detect ROH, which include smoothing of homozygosity. In addition, at a 500kb threshold, 125 of 126 cattle had at least one ROH above the threshold, and the number dropped drastically around 500kb. Considering these results, ROH length of at least 500kb was chosen for the downstream analysis; the ROH frequency in our population was large enough (~99%) to cover almost all individuals, and, at the same time, the length was maximum to enable statistical analysis using ROHs.

Although ROHs longer than 500kb occurred at least once in almost all individuals in our population, this alone does not suggest that 500kb ROHs were generated by recent inbreeding. Thus, it was necessary to show the difference between selected population that has gone through the inbreeding induced by artificial selection, and unselected population when the threshold was applied. For this, we used the genome data of unselected individuals which were processed in the same way as those of 126 selected individuals, and calculated individual genome-wide autozygosities (Froh) using ROHs following previously published protocol (McQuillan, Leutenegger et al. 2008). Froh values of the unselected group were compared to those of 126 selected individuals. As expected, selected populations generally had a higher inbreeding coefficient than unselected (Figure 4.2a, Wilcoxon rank sum test, p-value = 9.704e-05). Unselected population had an average of 0.0034 F_{roh}; on the other hand, average F_{roh} of the selected population was 0.0131. This difference was also observed when we generated the data by random sampling thousand times from 126 selected individuals to reduce bias due to sample size (Figure 4.3). F_{snp} values between unselected and selected population also showed similar pattern (Figure 4.2b). From this result, we inferred that recent inbreeding had an influence on genome and $F_{roh,}$ and found that a threshold of 500kb could be sufficient to represent the recent inbreeding which our samples had undergone.

In the same context of comparison between selected and unselected cattle, the effect of breeding program over years was also investigated in selected cattle. During ~ 20 years, F_{roh} steadily increased with several outliers (Figure 4.2c). Additionally, the F_{roh} of cattle born earlier were close to zero, whereas most of cattle born later had F_{roh} close to 0.02. This demonstrates throughout ~20 years long breeding program, there was a ~2% increase of ROH at a genome-level. This gradual increase of inbreeding level has also been observed in a study using pedigree-based method (Dang, Lee et al. 2011).

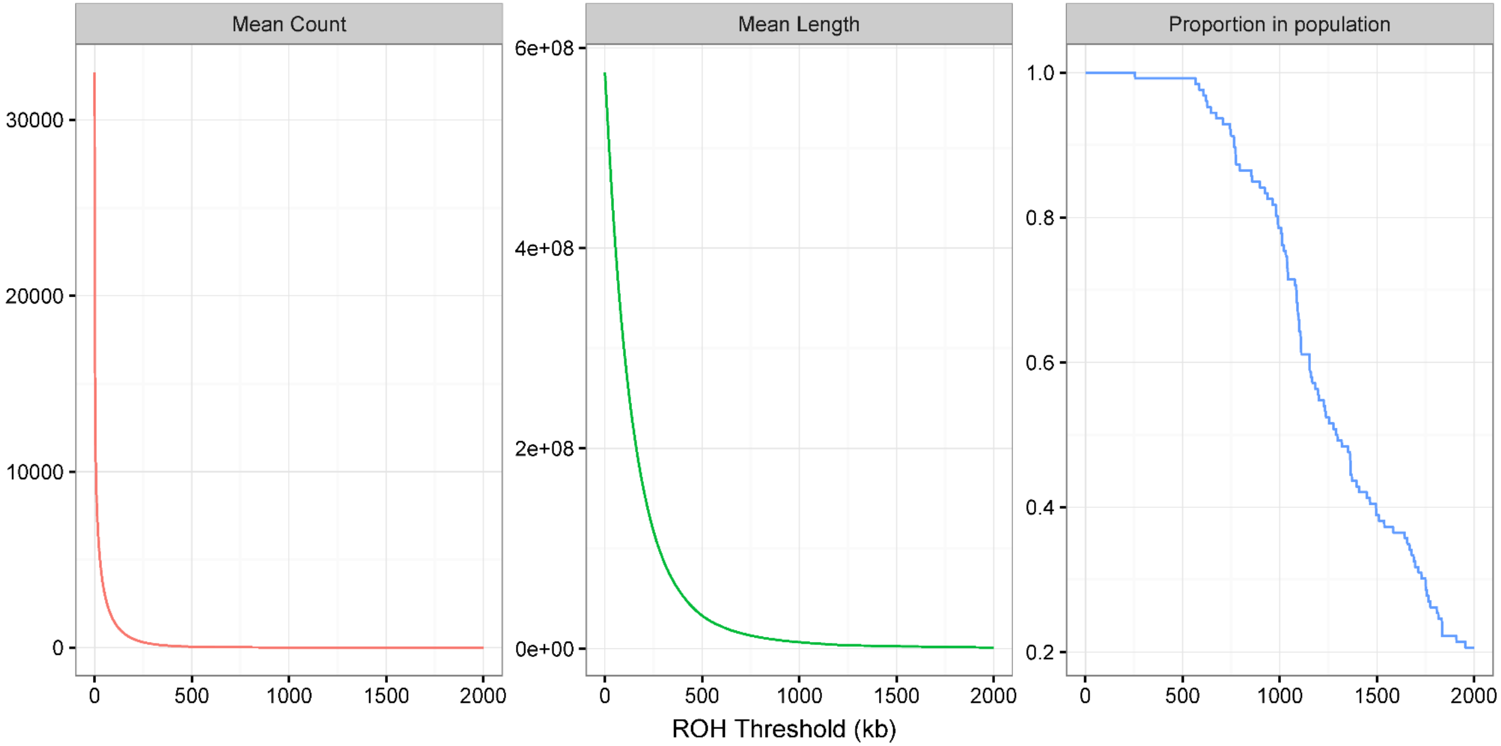


Figure 4.1 Mean count, length, and frequency of ROH in 126 cattle according to the change of ROH threshold. ROH threshold was controlled from 0 to 2000kb with 1000kb as a unit

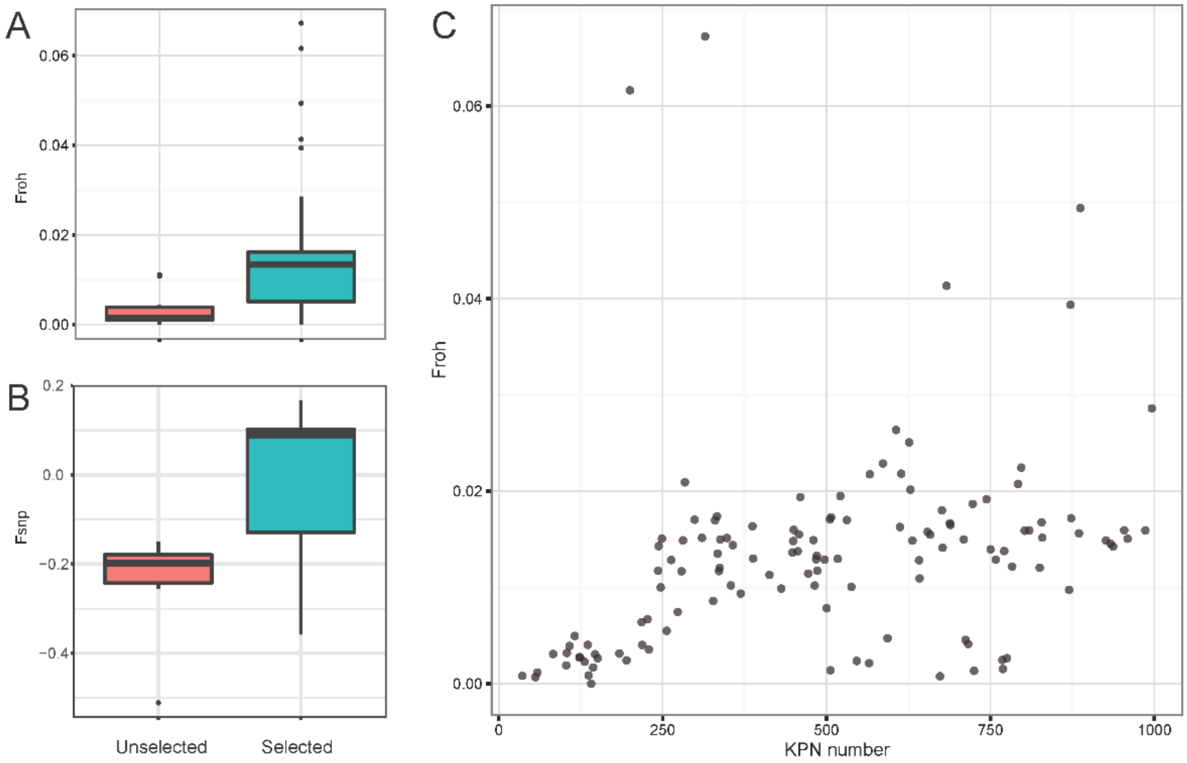


Figure 4.2 Individual genome-wide autozygosities (F_{roh}). Comparison of (a) F_{roh} and (b) F_{snp} between selected (n = 126) and unselected (n = 10) cattle populations. Selected and unselected cattle populations were significantly different in both F_{roh} and F_{snp} (Wilcoxon rank sum test, p-value = 9.704e-05 and 2.979e-04, respectively). (c) Change of F_{roh} during the past ~20 years. KPN number was used instead of cattle birth year

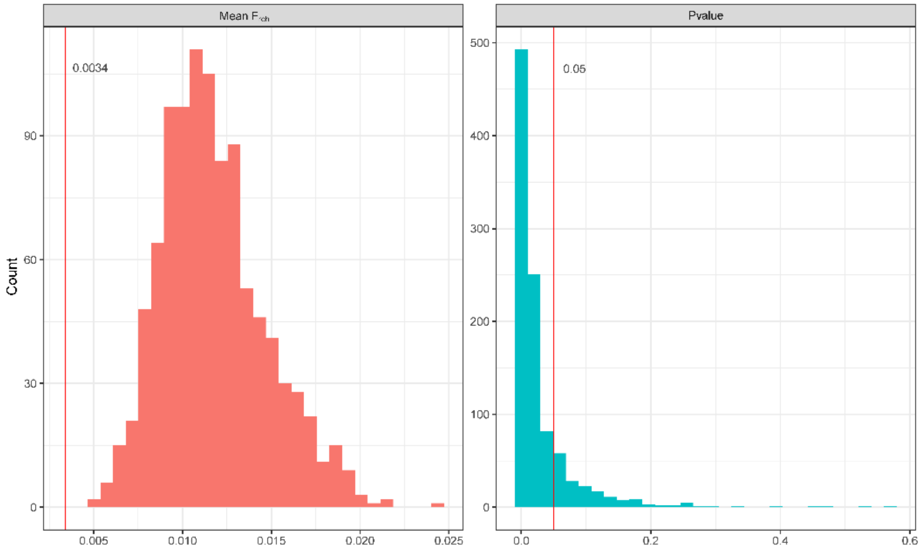


Figure 4.3 Distribution of Mean Froh and p-values of 1000 data sets generated by sampling 10 individuals from selected population iteratively. Froh values for each data set were averaged, and p-values were calculated by Wilcoxon's rank sum test between 10 selected individuals and unselected individuals (n=10). Note that the vertical red line indicates Mean Froh and p-value of unselected population, respectively

4.4.2 Genomic regions with increase of autozygosity during artificial selection

When genome-wide distribution of ROHs was investigated, there was variation of ROH length as well as change of ROH length among genomic regions (Figure 4.4); changes of ROH length were calculated by subtracting the mean length of one population from the other divided by KPN number. Mean length of ROHs ranged from 0.0Mb to 0.44Mb, and the change of ROH mean length ranged from -0.31Mb to 0.52Mb. This genome-wide heterogeneity of ROH allowed genome-wide mapping analysis, which need the prerequisite that ROHs are not evenly distributed on a genome. In addition to this, most of bins gained ROHs rather than losing them (Figure 4.4), which is consistent with the change of individual autozygosity according to KPN number (Figure 4.2c). Consequently, there were discrepancies in rates of inbreeding across a genome, however, the overall tendency was toward an increment of ROHs during artificial selection.

On the basis of ROH heterogeneity, we first tried to identify genomic regions whose ROH significantly increased or decreased throughout ~20 years of the breeding program. When ROHs within each bin were fitted by the model (Analysis 1), in 225 of total 264 bins, coefficients were positive (Figure 4.5). This again demonstrated that in most of regions in genome, ROH increased even though the rates of increase were heterogeneous among the regions. Similar tendency was observed in nominally significant bins (p-value < 0.05). For the result of statistical test, there were 8 bins (p-value < 0.01) which were statistically significant, with coefficients ranging from 0.00234 to 0.00542 (Table 4.1). Of 8 bins, only one bin overlapped with 16 ROHs in a previous study (Lee, Chung et al. 2013) using whole genome sequencing of Hanwoo bull. This ROH segment were specific to Hanwoo breed compared to Black Angus and Holstein. The ROH might have been created during recent Hanwoo breeding program where the population was not outbred with other breeds (Figure 4.6).

To gain deeper insight into the 8 regions identified above, we investigated the distribution of ROH segments (Figure 4.7). We found that one of them (BTA 25, 30~40Mb) showed increasing pattern at a narrow region. 15 individuals contained at least one ROH segment in this region, and 12 segments among them shared ~0.2Mb region (BTA 25, 30,931,767~31,129,826) (Figure 4.8a). When we separated total population into two groups (Group A: KPN≤486, and Group B: KPN>486), Group B had more ROH segments than Group A, and Average LD and F coefficients of Group B was higher than Group A (Figure 4.8b and c). These measures of inbreeding suggest that in this particular region, there was an increase of inbreeding induced by artificial selection. Moreover, strong selection signature (|iHS| > 3.623, as highest 1% of all |iHS| values at genome-wide level) was detected in this region, especially in the overlapped ROH segment (Figure 4.8d). In a recent study, significant correlation between the EHH-based selection signature and actual trend in haplotype frequencies was demonstrated (Glick, Shirak et al. 2012). Similarly, we inferred that strong selection signature based on EHH could be the evidence of recent artificial selection contributing to change of ROH. The artificial selection causes sequence alteration, which in turn creates the inbreeding signature presented in this region. The existence of ROH segments in this region was also confirmed using additional dataset (Figure 4.7)

.

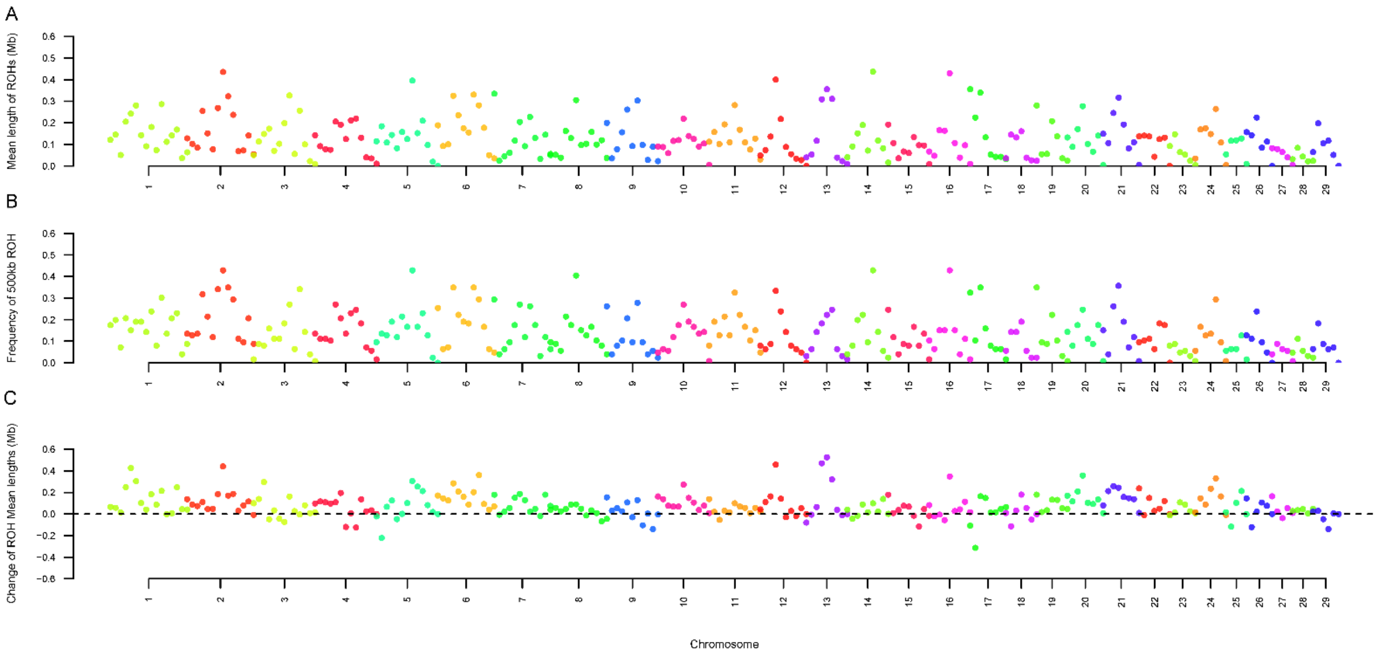


Figure 4.4 Genome-wide distribution of ROH in 126 cattle. (a) Distribution of ROH mean length in 10Mb bin. (b) Frequency of ROH longer than 500kb in 10Mb bin. (c) Change of ROH mean length when comparing ROH mean length of two groups (Group A: KPN≤486, and Group B: KPN>486)

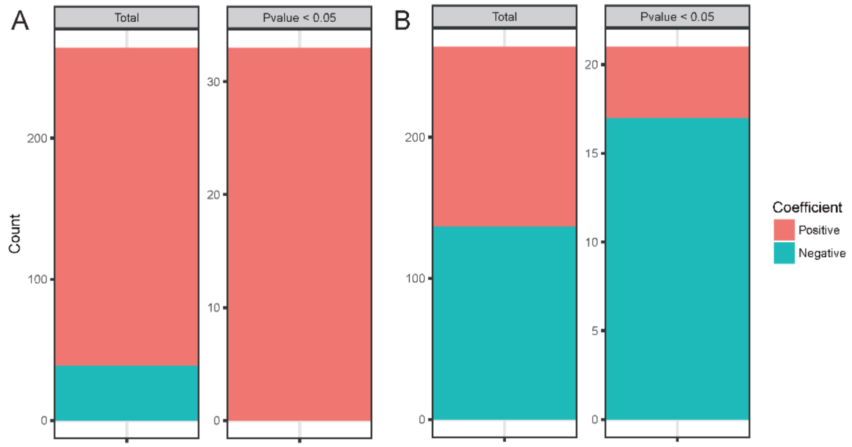


Figure 4.5 Direction of regression coefficients in two association test (Analysis 1, and Analysis 2). (a) Bin counts according to their direction of regression coefficients in association test between artificial selection and ROH (Analysis 1). (b) Bin counts according to the direction of regression coefficients in association test between ROH and body weight (Analysis 2)

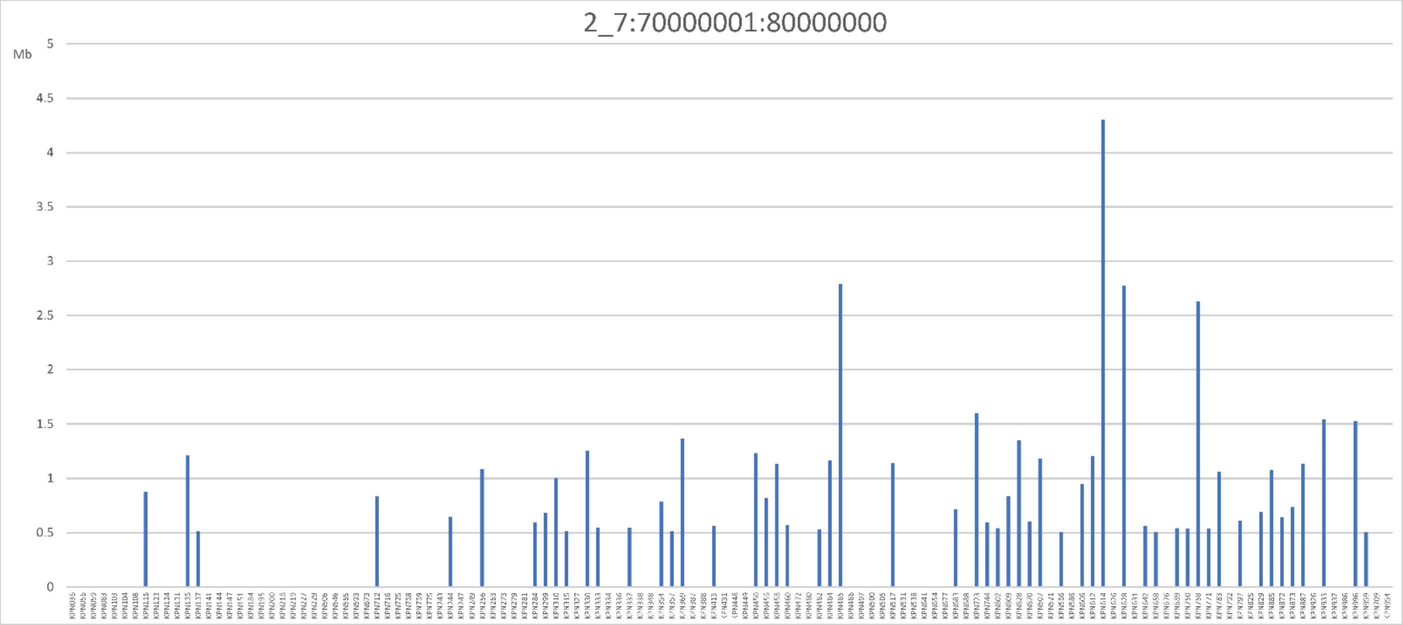


Figure 4.6 ROH length of each individuals in a bin (BTA2:70,000,001~80,000,000) overlapped with a previous study (Lee, Chung et al. 2013). X axis indicates the individual ID sorted by their KPN number, and Y axis indicates ROH length in Mb..

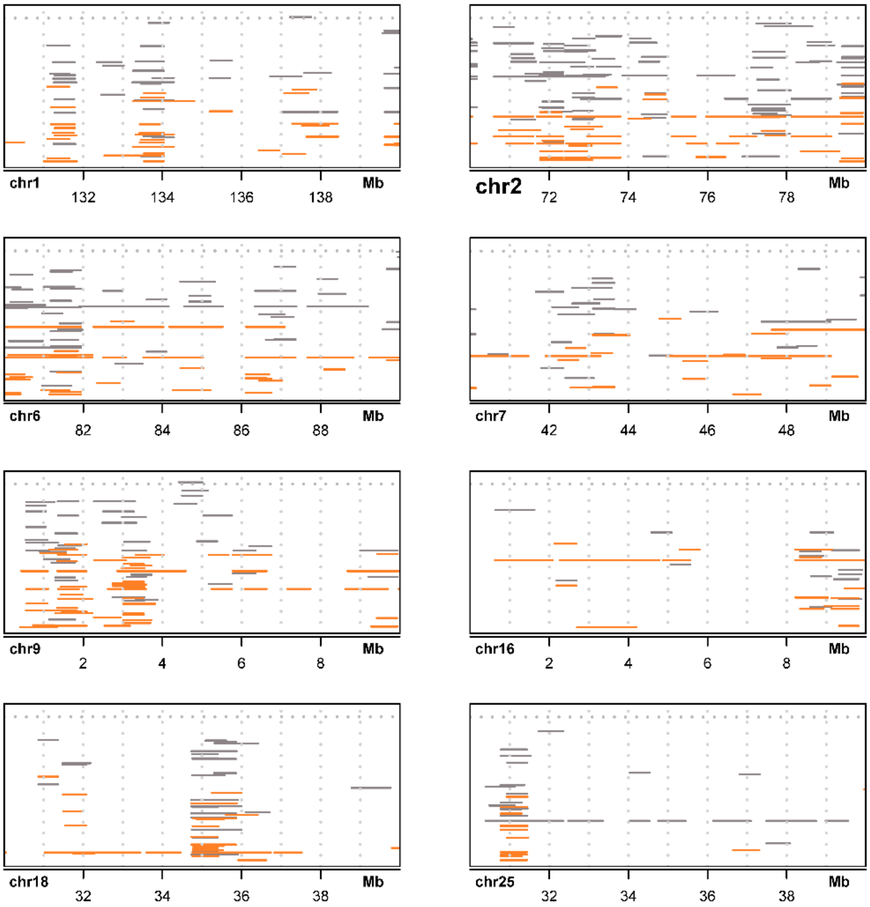


Figure 4.7 Distribution of ROH segments in candidate regions with validation dataset. Y axis indicates the individual ID sorted by their KPN number with increasing order, and X axis indicates coordinates on UMD 3.1 reference genome. ROH segments in original dataset (n=136), and in validation dataset (n=77) are marked by grey and orange color, respectively. Note that the first 10 individuals are “unselected population” without KPN numbers

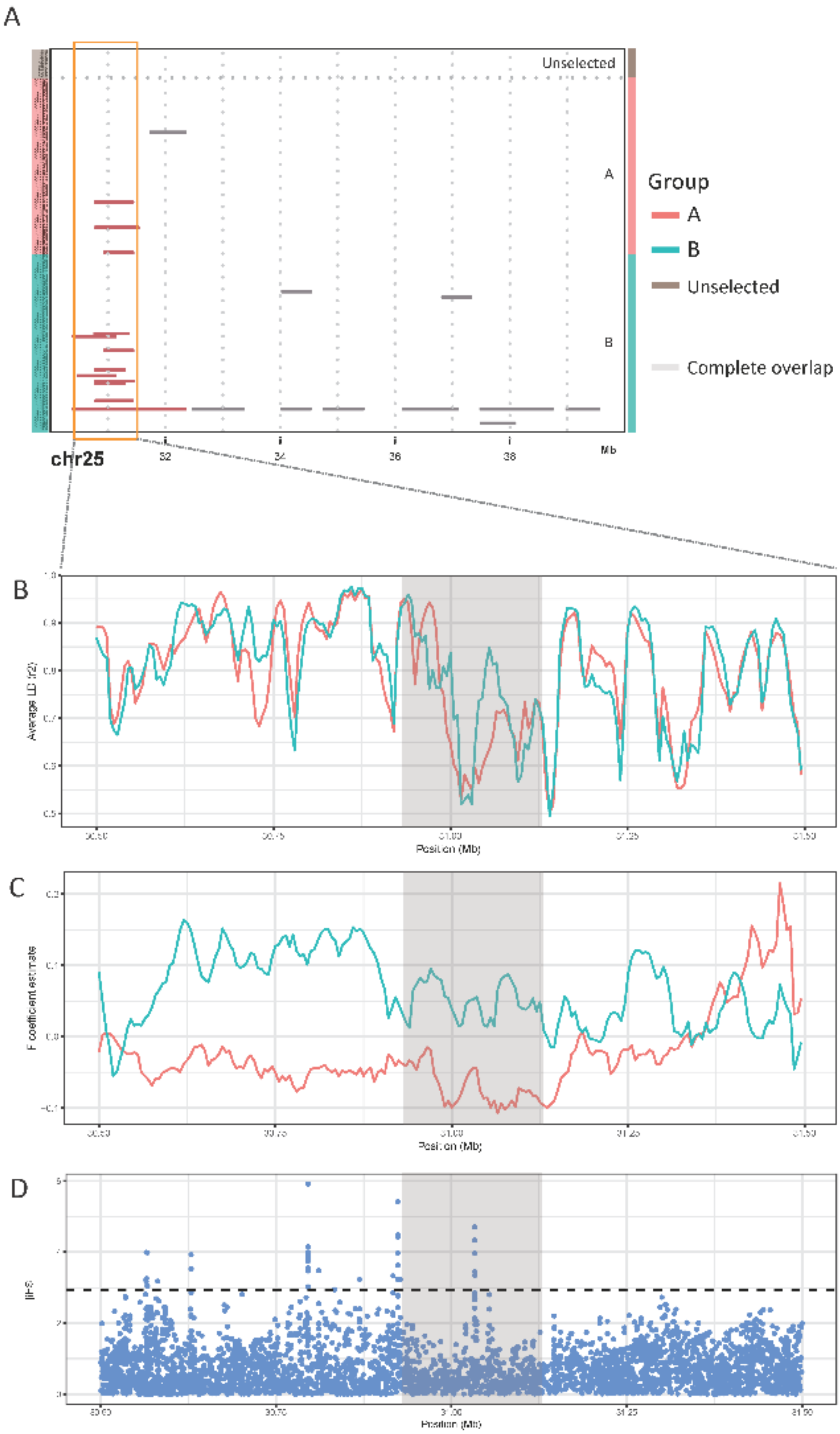

A
Unselected
A
B
Group
A
B
Unselected
Complete overlap
chr25
32
34
36
38
Mb
B
Average LD (r2)
1.0
0.9
0.8
0.7
0.6
0.5
30.50
30.75
31.00
31.25
31.50
Position (Mb)
C
F coefficient estimate
0.2
0.1
0.0
-0.1
30.50
30.75
31.00
31.25
31.50
Position (Mb)
D
|iHS|
6
4
2
0
30.50
30.75
31.00
31.25
31.50
Position (Mb)

Figure 4.8 Signatures of inbreeding at the candidate region in BTA 25. (a) Distribution of ROH segments in the candidate region. “Complete overlap region” refers to the genomic regions that have the maximum number of samples which have at least one ROH segment. (b) Inbreeding signatures of candidate region are presented by Average LD and F coefficient. “Complete overlap region” are shaded in grey. Unselected individuals, Group A (Individuals with KPN≤486), and Group B (Individuals with KPN>486) are represented by dark brown, red and green color, respectively

Table 4.1 Candidate regions associated with years. Statistical test using Analysis 1 was performed. Only the regions with Pvalue less than 0.01 are shown

BTA	Start	End	Coefficient	r^2	Pvalue
1	130,000,001	140,000,000	0.00250	0.06192	0.00551
2	70,000,001	80,000,000	0.00235	0.05965	0.00209
6	80,000,001	90,000,000	0.00245	0.06217	0.00318
7	40,000,001	50,000,000	0.00294	0.07861	0.00436
9	1	10,000,000	0.00234	0.05587	0.00620
16	1	10,000,000	0.00542	0.18820	0.00092
18	30,000,001	40,000,000	0.00312	0.08871	0.00205
25	30,000,001	40,000,000	0.00327	0.08881	0.00674

4.4.3 Lack of significant influence of recent inbreeding on body weight

At a genome-wide level, there was no significant relationship between body weight and genome-wide F_{roh} (Figure 4.9, Spearman's correlation test, ρ = -0.0393, p-value = 0.699). Instead, the body weight steadily increased during ~20 years, although F_{roh} steadily increased for the period (Figure 4.2c). Consequently, there was little influence of F_{roh} on body weight at an individual level. According to records for Hanwoo breeding program, there was indeed a gradual increase of weight for the decades (Park, Choi et al. 2013) similar to our data (Figure 4.2). If we assume that F_{roh} represents the inbreeding level of an individual (McQuillan, Leutenegger et al. 2008, Bjelland, Weigel et al. 2013), this suggests that although Hanwoo has experienced inbreeding pressure induced by artificial selection in ~20 years, the negative effect of inbreeding had less influence on an individual's body weight.

When we focused on particular regions of 10Mb length, the regions with fitted models showing positive coefficient (Analysis 2) was almost equivalent to that of negative coefficient unlike the previous result (Figure 4.5). However, the regions with negative coefficients are dominant if we look at the most nominally significant bins (p-value < 0.05). This discrepancy indicates that there are regions where ROHs have negative correlation with body weight. The 7 candidate regions associated with body weight are shown in Table 2 (p-value < 0.01). All of these regions were found to have a negative correlation between ROH length and body weight. Among these regions, 4 contain at least one body weight QTLs (body weight, or body weight gain) according to animal QTLdb (Hu, Park et al. 2016). Increase of 1 base pair of the ROHs in these regions corresponded to only 0.00002 body weight loss on average. Although ROHs located in these regions have negative correlation with body weight, it does not imply that the actual body weight of an individual will decrease due to these ROH. The actual trend of body weight of Hanwoo is towards an increase, and effect of the ROHs was one of many factors affecting body weight.

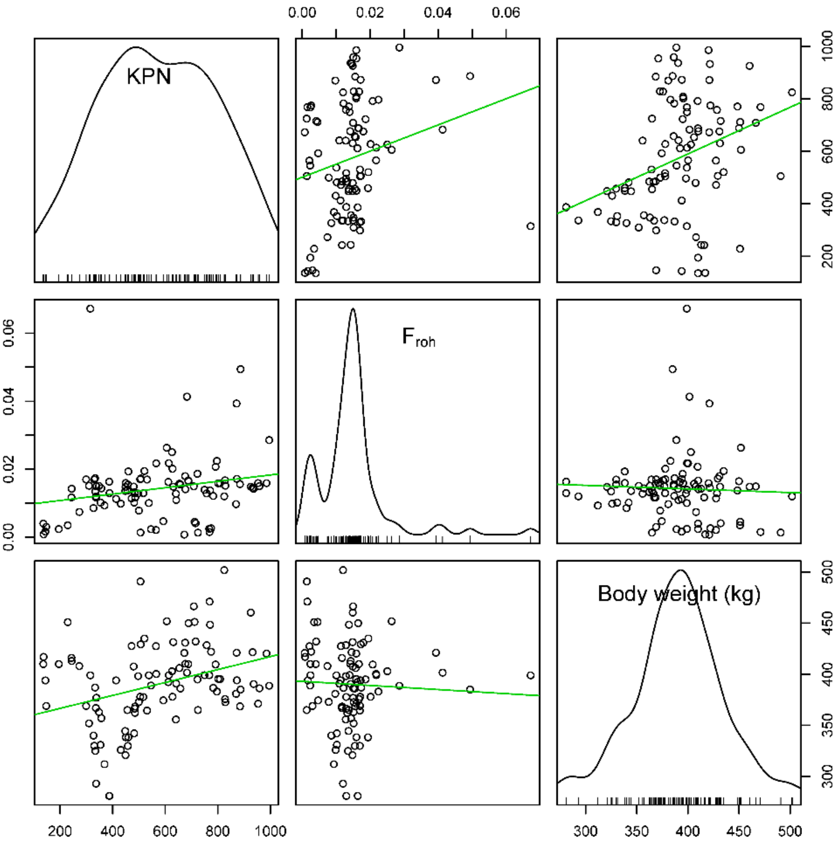


Figure 4.9 Scatterplots for KPN, Froh, Body weight. Correlation between each elements was tested by spearman’s method. KPN vs Froh : ρ = 0.46697, p-value = 5.203e-08; Froh vs Body weight : ρ = -0.03930, p-value = 0.69900; KPN vs Body weight : ρ = 0.33921, p-value = 0.00059

Table 4.2 Candidate regions associated with cattle body weight. Statistical test using Analysis 2 was performed. Only the regions with Pvalue less than 0.01 are shown

BTA	Start	End	Coefficient	Pvalue	Body weight QTL
2	100,000,001	110,000,000	-0.00005	0.00076	0
3	90,000,001	100,000,000	-0.00002	0.00366	2
5	50,000,001	60,000,000	-0.00003	0.00037	1
7	70,000,001	80,000,000	-0.00002	0.00034	14
15	50,000,001	60,000,000	-0.00002	0.00451	0
16	40,000,001	50,000,000	-0.00001	0.00634	12
27	20,000,001	30,000,000	-0.00002	0.00441	0

4.4.4 Signatures of selection detected by integrated haplotype score (iHS)

To detect loci under recent selection, we independently calculated iHS as a measure of selection. In each non-overlapping windows of 100kb, a proportion of SNPs with |iHS| > 2 was calculated (In this step, windows containing less than 10 SNPs were dropped). Then, we considered the windows with empirically highest 1% of the proportion (0.3623) to be candidates for containing selective sweep (Voight, Kudaravalli et al. 2006). As a result, we could identify 250 windows with selective sweeps (Figure 4.10). Of 8 regions in which ROH has increased significantly, 4 regions showed selective sweep signature, and 5 of 7 regions which were associated with body weight displayed selective sweep signature.

The candidate region in BTA 25 were not significantly affected by selective sweep. They showed relatively moderate signals (The proportion of SNPs with |iHS| > 2: 0.0432, and 0.0417 respectively). However, when we locally investigated the regions, we could find strong selection signals, especially in the overlapped ROH segment (Figure 4.8d).

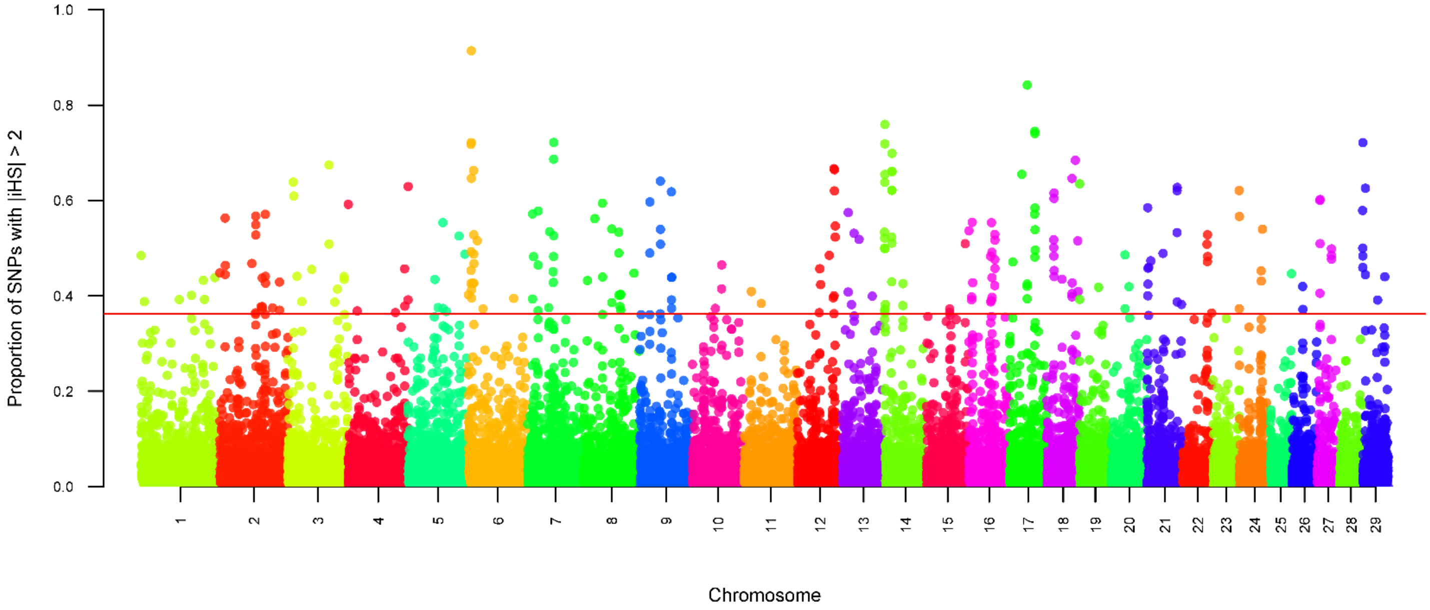


Figure 4.10 Selective sweep regions identified by integrated haplotype score (iHS). The horizontal red line indicates top 1% proportion of SNPs with |iHS| > 2 in a 100kb window

4.5 Discussion

In this study, we traced the change of inbreeding in cattle population under artificial selection, and observed increase of inbreeding for the decades of period. Also, we could suggest candidate regions which showed significant increase of inbreeding. However, we could not suggest strong evidences for the relation of the candidate regions produced by artificial selection to the increment of body weight. Indeed, there was no statistically significant regions shared by both of regression analyses (Analysis 1, and Analysis 2). Individual body weight has increased, even though the degree of inbreeding, which lowers fitness-related characters in many species (Charlesworth and Willis 2009) has also increased for decades of artificial selection, and the inbreeding of most candidate regions showed weak negative correlation with body weight. This might mean that the increment of inbreeding at genome-wide level or specific region did not have large impact on body weight.

Hanwoo breed has a short history of artificial selection compared to well-known commercial breeds such as Angus and Holstein (Park, Choi et al. 2013, Lee, Park et al. 2014). There has also been several evidences that the Hanwoo population has a lower degree of inbreeding than other commercial breeds (Lee, Cho et al. 2011, Lee, Kim et al. 2014). Therefore, it might be possible that the negative effect of inbreeding is not as high as the positive effect of the breeding program to cause deleterious effects on body weight in the Hanwoo population. Moreover, for weight/growth traits including body weight, it has been shown that there is less influence of inbreeding than other traits related to reproduction or production (Leroy 2014). As a result, we inferred that either insufficient inbreeding load or characteristics of body weight trait could be the reason why increment of inbreeding did not have large effects on body weight.

In addition to investigation of inbreeding increment, we independently calculated iHS as a measure of recent selection. As a result, we could identify 250 windows with selective sweeps. However, the 250 windows could not include all of the previously identified regions with significant increase of autozygosity. Since inbreeding affects all loci equally and genetic drift changes frequency of loci randomly, inbreeding may not induce LD between neighboring loci, whereas selection will drive linked alleles to high frequency (MacEachern, Hayes et al. 2009). Selection signatures based on LD could not therefore capture all the actual change of ROH. Moreover, we investigated selection signatures for the entire population, which might result in more difference in the two analyses.

In previous studies, the patterns of ROH was investigated, in which ROHs of particular length in Hanwoo were compared to those in other commercial breeds (Lee, Chung et al. 2013, Porto-Neto, Lee et al. 2014). However, the purpose of this study was to demonstrate the change of ROH within particular population. The threshold for autozygous segments, therefore, needs to be set according to the characteristics of our population. Furthermore, the difference of marker density produced the difference of power to detect ROH that are IBD (Marras, Gaspa et al. 2015). It is, therefore, expected that there are considerable differences in ROH detected from SNP chip data and sequencing data. Consequently, we set the threshold length of ROH adjusted to our data, unlike that of most previous studies, in which SNP chip data of commercial breeds were employed.

Here, we attempted to identify the direct connection between a genetic marker, breeding program and a trait. In past efforts to develop genetic markers in livestock, the main focus has been to find genetic markers significantly associated with economic traits. This approach could be promising in livestock industry if we have access to the tools that can directly manipulate genomic sequences with high resolution. However, there has been no such tool in livestock industry. Instead, artificial selection or breeding program have been widely used to indirectly modify genome sequences. Therefore, the chances of developing useful applications of the genetic markers are expected to be higher if breeding strategy as well as the trait are considered to identify those markers.

Artificial selection often leads to inbreeding (Sanchez, Toro et al. 1999), thus there have been many efforts to manage rates of inbreeding (Toro and Perez-Enciso 1990, Brotherstone and Goddard 2005). The Hanwoo population used in this study has been known to have a relatively low inbreeding coefficient and larger effective population size than other commercial breeds such as Holstein (Dang, Lee et al. 2011, Lee, Park et al. 2014). However, the degree of inbreeding in Hanwoo population rapidly increased in recent years, which has created an urgent need for the control of inbreeding rates (Dang, Lee et al. 2011). Moreover, most Hanwoo calves are usually produced by artificial insemination which uses semen from a few selected bulls. For this reason, inbreeding of few bulls could affect the whole Hanwoo population. The approaches used in this study are advantageous for monitoring the change of inbreeding through breeding process along with inbreeding depression related to specific traits. This will be especially useful when investigating population of certain breed as Hanwoo.

This chapter will be published elsewhere
as a partial fulfillment of Kwondo Kim's Ph.D program.

Chapter 5. The mosaic genome architecture of indigenous African cattle as a unique genetic resource for adaptation to local environments

5.1 Abstract

Across African human societies, cattle play a crucial role as a source of animal proteins (milk and meat), draught power and wealth. They are present across all agro-ecologies, from the driest to the most humid ones, with their adaptations to various local environmental conditions directly linked to their survival and productivity. The majority of African cattle are now classified as zebu, sanga or zenga. They are admixed populations with high genome diversities, a legacy of their crossing with African taurine, the most ancient African cattle populations, following multiple indicine (zebu) introductions. Indicine are of larger size and they are more adapted to the arid environment than their taurine counterparts, while African taurine are uniquely tolerant to African tropical disease challenge, most notably trypanosomosis. We present here evidences that such admixed diversity is at the root of the success of African cattle pastoralism. First, we analyzed whole-genome sequences of 162 indigenous African cattle, representing 15 indigenous populations, and demonstrated high degree of their genome admixture compared to pure taurine and indicine cattle. We then reconstructed the local ancestries of the genomes of African indicine populations. Against a genome-wide background predominantly indicine, we found an exceptionally long gene rich region around 6 Mb at chromosome 23 with a significant excess of taurine ancestry. This region includes 121 annotated genes and it is enriched in immune response and stimulus perception locus. Then, we analyzed two African taurines with

evidence of indicine introgression. We identified a conserved haplotype commonly shared within African taurine breeds, but distinct from that of European taurine as well as Asian indicine cattle. This specific haplotype was located on upstream of *CARD11*, previously reported to be linked to trypanotolerance, which supports its role as a regulatory element. It represents an example of a barrier locus against introgression. Our results support that the rapid dispersion across African agro-ecologies of cattle was driven by adaptive admixture.

5.2 Introduction

Cattle play an important role across African economy and society. They provide a source of nutrition, manure, and draught power, and are often used to pay bride and blood fines as a primary wealth of each household (Schneider 1964, Esse, Bürkert et al. 2001, Di Lernia, Tafuri et al. 2013). To date, at least 150 indigenous cattle breeds have been recognized across the continent (Mwai, Hanotte et al. 2015). They are widespread in diverse agro-ecologies, displaying unique characteristics distinguishing them from each other. One of the well-known examples in taurine breeds, e.g. N'Dama, is genetic tolerance to tsete-fly transmitted trypanosomosis (Berthier, Brenière et al. 2016), while African zebu cattle are recognized for their heat tolerance adaptation (Hansen 2004).

According to previous studies (Hanotte, Tawah et al. 2000, Hanotte, Bradley et al. 2002), the dispersion and diversity of African cattle followed the history and development of African cattle pastoralism that shapes the basis of life for millions across the continent. It is now well established that the humpless *Bos taurus* and the humped *B. indicus* originated from separate domestication events ~10,000 years ago (Pitt, Sevane et al. 2019). The humpless *Bos taurus* has been introduced first into the African continent (Epstein 1971). Archeological evidences support that these first African cattle have reached the western and eastern parts of the continent (Hanotte, Tawah et al. 2000, Hanotte, Bradley et al. 2002). After the taurine dispersion, the first wave of arrival and dispersion of humped cattle, *Bos indicus* started around 700

AD along the Red Sea and Indian Ocean coastal area of the continent following the development of the Swahili civilization with subsequent rapid dispersion across the continent via Horn of Africa, the Sahelien belt, and the African Rift Valley in several waves. (Epstein 1971).

Such multiple introductions of cattle types led to extensive crossbreeding between indicine and taurine cattle populations across the continent. Today, the great majority of African cattle are *B. taurus* x *B. indicus* admixed populations, and they are classified as sanga, zenga and African zebu according to their phenotypes (Mwai, Hanotte et al. 2015). The pure taurine populations are found today in only West Africa. Molecular studies on mitochondrial DNA (mtDNA) diversity indicate the presence of only taurine mtDNA haplotypes within African cattle (Bradley, MacHugh et al. 1996, Bonfiglio, Ginja et al. 2012), which suggest a predominantly male mediated pattern of indicine introgression. Autosomal genome analysis indicate that African humped cattle do have a taurine background of different proportions among breeds or populations but with little variation within a population (Mbole-Kariuki, Sonstegard et al. 2014).

Population admixture has significant effects on the genetic composition of a population, as it provides new genetic resource and diversity to a population (Verhoeven, Macel et al. 2010). It may facilitate dispersion and colonization of new habitats (Arnold and Kunte 2017). However, it may also have a cost, reducing reproductive fitness of hybrid populations following genome incompatibility (Abbott, Albach et al. 2013). Several recent studies have addressed the effects of population admixture and introgression among cattle and related-species at genome-wide level. They have identified loci that derived from donor populations, which may have contributed to the adaptation of recipient populations (Medugorac, Graf et al. 2017, Chen, Cai et al. 2018, Wu, Ding et al. 2018). Despite the root of the today success of African pastoralism, the effect and consequence of *B. taurus* x *B. indicus* admixture on the genome of African cattle have not been reported yet.

In this study, we newly generated whole genome sequences of 114 individuals across 11 indigenous African cattle breeds, which have diverse genetic background from taurine-indicine ancestry. Incorporating previously generated data of 5 indigenous African cattle breeds, we performed comparative genome-wide analysis with 5 European taurine and 3 Asian indicine cattle breeds (Figure 5.1b). This allows us to demonstrate the prevalent taurine-indicine admixture in African cattle populations at population-wide and genome-wide. We also infer an influence of such admixture on their local environmental adaptations since they arrived in the continent.

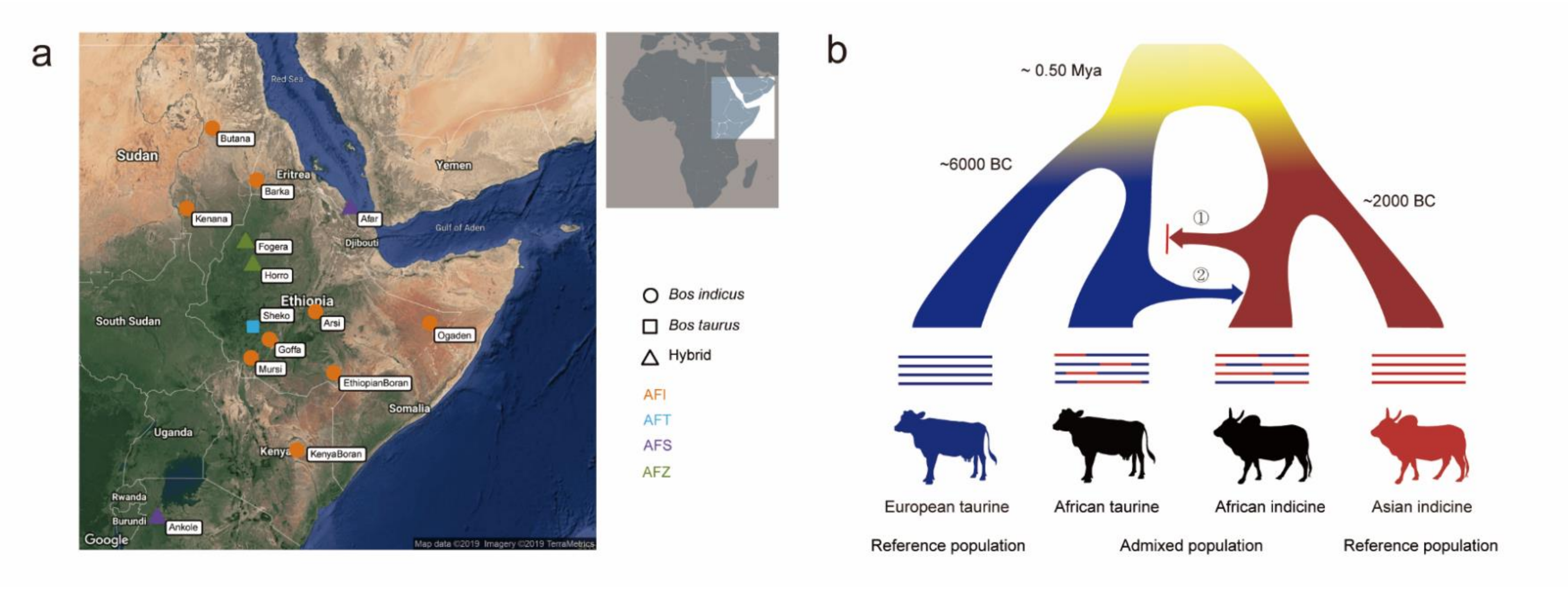


Figure 5.1 (a) Distribution of indigenous African cattle breeds used in this study. Only Eastern African breeds were shown, excluding N’Dama breed. (b) Scheme of comparative analysis in this study. Two analysis have been performed to identify adaptive signature in indigenous African cattle; ① To identify introgression loci in African indicine cattle ② To identify barrier loci against introgression in African taurine cattle

5.3 Materials and Methods

5.3.1 Sample information

The current study used 282 individuals in total, including 45 additional individuals for validating the candidate regions. These samples were grouped according to their origin and types; EUT (European taurine cattle as a reference population), ASI (Asian indicine cattle as a reference population), AFI (African indicine cattle), AFS (African sanga cattle that is a crossbreed between taurine and indicine cattle), AFZ (African zenga cattle that is a crossbreed between sanga and indicine cattle), and AFB (African buffalo, as an outgroup). The detailed information for each groups and corresponding breeds is provided in Table 5.1.

Table 5.1 Summary information of cattle breeds used in this study

Breeds	Origin	Species/Subspecies	Animal count	Study
N'Dama	North-eastern Africa	Bos taurus	13	Kim et al. 2017 (10); This study (3)
Sheko	South-western Ethiopia	Bos taurus	9	This study
Arsi	Ethiopia	Bos indicus	10	This study
Barka	North-western Ethiopia	Bos indicus	9	This study
Butana	Sudan	Bos indicus	20	This study
EthiopianBoran	Ethiopia	Bos indicus	10	This study
Goffa	South-western Ethiopia	Bos indicus	10	This study
Kenana	Northern Sudan	Bos indicus	13	Kim et al. 2017 (9); This study (4)
KenyaBoran	Northern Kenya	Bos indicus	10	This study
Mursi	South-western Ethiopia	zBos indicus	10	Kim et al. 2017
Ogaden	South-eastern Ethiopia	Bos indicus	9	Kim et al. 2017
Afar	North-eastern Ethiopia	Bos taurus x Bos indicus (Sanga)	9	This study

Ankole	Uganda	Bos taurus x Bos indicus (Sanga)	10	Kim et al. 2017
Fogera	North-western Ethiopia	Bos taurus x Bos indicus (Zenga)	9	This study
Horro	North-western Ethiopia	Bos taurus x Bos indicus (Zenga)	11	This study
Angus	Great Britain	Bos taurus	10	Kim et al. 2017
Hanwoo	Korea	Bos taurus	23	Lee et al. 2014 (11); Shin et al. (12)
Holstein	Netherlands	Bos taurus	10	Lee et al. 2014
Jersey	Channel Islands, France	Bos taurus	10	Kim et al. 2017
Brahman	India	Bos indicus	10	Heaton et al. 2016 (5); Taylor et al. 2016 (5)
Gir	India	Bos indicus	4	Taylor et al. 2016
Nelore	India	Bos indicus	6	Taylor et al. 2016
AfricanBuffalo	South-eastern Africa	Syncerus caffer	2	This study
Total			237	

5.3.2 Sequencing and variant calling

Our previously published data of 53 commercial taurine (Lee, Kim et al. 2014, Shin, Lee et al. 2014, Kim, Hanotte et al. 2017) and 48 African (Kim, Hanotte et al. 2017) cattle, and publicly available data of 20 Asian zebu (Heaton, Smith et al. 2016, Taylor, Whitacre et al. 2016) were incorporated into our newly generated raw sequence data.

We then examined a per-base sequence quality for the raw sequence reads using fastQC software (Andrews 2010), and removed low quality bases and artifact sequences using Trimommatic (Bolger, Lohse et al. 2014). The high quality sequence reads were then mapped against the reference bovine genome (UMD 3.1) using Bowtie2 (Langmead and Salzberg 2012) with default parameters. The overall alignment rate of reads to the reference sequence was 98.56% with an average read depth of 10.96×. On average across the whole samples, the reads covered 98.44% of the reference genome.

For the mapped reads, potential PCR duplicates were filtered using the "REMOVE_DUPLICATES = true" option in "MarkDuplicates" command-line tool of Picard (http://broadinstitute.github.io/picard), and mate-pair information was checked and fixed using the "FixMateInformation" command-line tools of Picard. We then used SAMtools (Li, Handsaker et al. 2009) to create index files for reference and bam files. Genome analysis toolkit 3.4 (GATK) (McKenna, Hanna et al. 2010) was used to perform local realignment of reads to correct misalignments due to the presence of indels ("RealignerTargetCreator" and "IndelRealigner" arguments). The

"BaseReclibrator" and "PrintReads" arguments of GATK were used to perform base quality recalibration.

For calling candidate SNPs and InDels from the bam files, the "UnifiedGenotyper" and "SelectVariants" arguments of GATK were used. To avoid possible false positives in these SNPs, we first masked SNPs that are located at the same position with InDels and used argument "VariantFiltration" of the same software with the following options: (1) SNP clusters were filtered with "--clusterSize 3" and "--clusterWindowSize 10" options; (2) SNPs with MQ0 (mapping quality zero; total count across all samples of mapping quality zero reads) > 4 and proportion of MQ0 reads > 0.1 were filtered; (3) SNPs with a phred-scaled quality score (QUAL) < 30 were filtered; (4) SNPs with FS (phred-scaled P value using Fisher's exact test to detect strand bias) > 200 were filtered since FS represents variation on either the forward or the reverse strand, which are indicative of false-positive calls. The remaining high-quality SNPs were annotated according to their positions using SnpEff v4.3 (Cingolani, Platts et al. 2012) and used in all of the downstream analysis of this study (Table 5.2 and 5.3).

To check the confidence of variant calls from resequencing analysis, we additionally genotyped 72 cattle samples (of which blood samples were available) using BovineSNP50 Genotyping BeadChip (Illumina, Inc.). After filtering out SNPs based on GeneCall score < 0.7, common loci of SNP chip and DNA resequencing data were extracted and examined to assess concordance between genotypes from two different platfroms.

5.3.3 Population differentiation and structure

For principle component analysis (PCA), we used genome-wide complex trait analysis (GCTA) (Yang, Lee et al. 2011) tool to estimate the eigenvalue and eigenvectors, incorporating genotype data from 235 individuals except 2 AFB. For admixture analysis, we restricted the genotype data to SNPs with MAF (Minor allele frequency) ≥ 0.01 and proportion of missing genotypes ≤ 0.01 to reduce computational load. Admixture v1.3.0 (Alexander, Novembre et al. 2009) was run for this genotype data, while increasing K from 2 to 22, where K is the assumed number of ancestral populations. The cross validation (--cv=10) was also performed to choose optimal K. Genetic distances between cattle breeds were estimated using *Fst* estimators as described in Weir and Cockerham (Weir and Cockerham 1984) that is implemented in PLINK 1.9 (Purcell, Neale et al. 2007).

Table 5.2 Variant calling statistics

	Whole	Autosome
# of variants	60,776,473	58,444,833
mean missing rate	0.01614688	0.0151154
mean minor allele frequency	0.1204713	0.1205122
mean SNP density	22.76457/Kbp	23.26563/Kbp

Table 5.3 Variant annotation statistics

Variant type	SNP count
Intergenic	43,850,529
Intronic	18,362,942
Downstream	2,755,568
Upstream	2,743,386
Exonic	482,254
Splice site acceptor	1,242
Splice site donor	1,160
Splice site region	47,826
UTR 3 prime	143,054
UTR 5 prime	31,889

5.3.4 Phylogenetic reconstruction and genetic diversity

A maximum likelihood tree was constructed with SNPhylo (Lee, Guo et al. 2014) on the basis of SNPs of 235 individuals. To assess the confidence of each node in the tree, bootstrapping was conducted with 1000 replicates (-B 1000). The maximum likelihood tree was visualized by interactive tree of life (iTOL) (Letunic and Bork 2019). We also reconstructed a population-level phylogeny of the 22 cattle populations using TREEMIX (Pickrell and Pritchard 2012) to address population relationships and to identify pairs of populations that are related to each other independent of that captured by this tree. The window size of 1000 was used to account for linkage disequilibrium (-k) and "-global" to generate the ML tree. Migration events (-m) were added into the initial tree until 99.8% of the variance in ancestry between populations was explained by the tree (Pickrell and Pritchard 2012, Decker, McKay et al. 2014).

To infer the genetic diversity and inbreeding of African cattle, observed heterozygosity and runs of homozygosity (ROH) were analyzed using VCFtools (Danecek, Auton et al. 2011). The ROH segments < 50kb were filtered to remain reliable identical by descent segments.

5.3.5 Identification of introgression signatures in African indicine/hybrid cattle

ABBABABA statistics

We used ABBABABA statistics (Durand, Patterson et al. 2011, Martin, Davey et al. 2014) (also called D statistics) to test for introgression and admixture pattern at the genome-scale. We assumed that EUT and ASI are not admixed from each counterpart and maintain pure taurine and indicine ancestry, respectively. The allele sharing pattern (ABBA and BABA sites) with these two pure population in African cattle breeds was measured by computing ABBABABA statistics based on the phylogeny as follows;

$$(((\mathrm{ASI},\mathrm{X}),\mathrm{EUT}),\mathrm{AFB}) \qquad (1)$$

$$\mathrm{D}(P_1, P_2, P_3, P_4) = \mathrm{D}(\mathrm{ASI}, \mathrm{X}, \mathrm{EUT}, \mathrm{AFB}) = \frac{\sum C_{ABBA}(i) - C_{BABA}(i)}{\sum C_{ABBA}(i) - C_{BABA}(i)} \qquad (2)$$

where *X* is the target African breed that we are interested in, AFB was used as an outgroup, P_i is each population, and $C_{ABBA}(i)$ and $C_{BABA}(i)$ are counts of either 1 or 0, depending on whether or not the specified pattern (ABBA or BABA) is observed at site i in the genome. For straightforward calculation for multiple sample case, we used the frequency of the derived allele at each site rather than binary counts of fixed ABBA and BABA sites to compute $C_{ABBA}(i)$ and $C_{BABA}(i)$ as follows;

$$C_{ABBA}(i) = (1 - \hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1 - \hat{p}_{i4}) \qquad (3)$$

$$C_{BABA}(i) = \hat{p}_{i1}(1 - \hat{p}_{i2})\hat{p}_{i3}(1 - \hat{p}_{i4}) \qquad (4)$$

where $\hat{p}_{ij}$ is the observed frequency of the derived allele at site *i* in population *j*.

To estimate the admixture proportion at genome-scale, we used *f* statistics (Durand, Patterson et al. 2011, Martin, Davey et al. 2014) computed as follows;

$$\mathrm{S}(P_1, P_2, P_3, P_4) = \sum C_{ABBA}(i) - C_{BABA}(i) \qquad (5)$$

$$f(P_1, P_2, P_3, P_4) = \frac{S(P_1, P_2, P_3, P_4)}{S(P_1, P_{3a}, P_{3b}, P_4)} \qquad (6)$$

where S indicates the difference between sums of $C_{ABBA}(i)$ and $C_{BABA}(i)$. f statistics compares the excess of ABBA over BABA sites, to that which would be expected under complete admixture by splitting P_3 into P_{3a} and P_{3b}, and replacing P_2 with P_{3a}. Therefore, we arbitrary split each breeds of EUT into two groups and set the P_{3a} and P_{3b} populations.

Standard errors of D and f statistics are computed by block jackknife method (Efron 1981) with block size 1Mb. Z score of D statistics and confidence interval of *f* statistics are computed based on the standard errors. All of these analysis for ABBABABA statistics were performed using the python scripts from https://github.com/simonhmartin.

Local ancestry inference

To minimize ambiguous information, we filtered SNP loci with the proportion of missing genotypes > 0.1 on whole genotype data before analyzing for regional signature. We then used BEAGLE v4.1 (Browning and Browning 2007) to infer the haplotype phase and impute missing alleles for the entire set of cattle populations (except AFB) simultaneously. We performed local ancestry inference implemented in LOTER package (Dias-Alves, Mairal et al. 2018) to infer taurine-indicine ancestry along the genome. LOTER has been shown to outperform other methods when admixture occurred earlier than 150 generations ago in human population. This time frame (around 4500 ya) is relevant for Bos indicus cattle history of reaching to African continent for the first time (Mwai, Hanotte et al. 2015). Moreover, LOTER is a parameter-free software that we were not required to specify uncertain biological parameters for bovine genome (e.g. genetic map).

We specified 53 individuals of EUT and 20 individuals of ASI as reference populations. Also, we assumed that a haplotype of an admixed African indicine or hybrid cattle consists of a mosaic of existing haplotypes from the reference populations. The resulting ancestral haplotypes were converted to ancestral allele frequency within African indicine/hybrid population. The allele frequencies were then averaged over each non-overlapping window with 20 SNPs. The windows showing strong deviation of indicine ancestry (3 SDs above or below the genome-wide average) were considered as potential candidates for decreased or increased taurine ancestry (Bryc, Auton et al. 2010, Jin, Xu et al. 2012). For checking bias derived from sample size difference between two reference populations, we generated 100 random subsets of EUT with same sample size as ASI (n=20), and performed local ancestry inference for each subset.

Relative frequency of IBD (rIBD) sharing

The genotype data filtered by proportion of missing genotypes > 0.1 was used as input for IBD detection program, IBDseq v2.0 (Browning and Browning 2013). To calculate the frequency of shared IBD in different regions of the genome, the genome was divided into 20kb windows with 10kb step (Wu, Ding et al. 2018). The frequencies of IBD segments shared between all pair-wise individuals were then counted for each window. The relative frequency of IBD (rIBD) for each window was calculated as follows (Bosse, Megens et al. 2014);

$$\text{rIBD} = \frac{cIBD_{EUT}}{tIBD_{EUT}} - \frac{cIBD_{ASI}}{tIBD_{ASI}} \qquad (7)$$

where $cIBD_i$ indicates the count of all shared IBD between *i* population and African indicine/hybrid population, $tIBD_i$ the count of total pairwise comparisons between *i* population and African indicine/hybrid population.

5.3.6 Identification of selection regions in African taurine cattle

Population branch statistics (PBS)

To detect selection signatures in AFT after divergence from EUT, we employed a java script (https://github.com/rronen/selection_stats) (Udpa, Ronen et al. 2014) implementing Populations branch statistics (PBS) developed by Yi et al. (Yi, Liang et al. 2010). For each window with 50kb size and 2kb step, we calculated this statistic as follows;

$$\text{T} = -\log(1 - F_{st}) \qquad (8)$$

$$\text{PBS} = \frac{T^{AE} + T^{AO} - T^{EO}}{2} \qquad (9)$$

where T^{ij} represent estimated branch length between i and j populations based on pair-wise Weir and Cockerham (Weir and Cockerham 1984) *Fst*. *A* represents a study population, AFT, *E* represents control population, EUT and *O* represents an outgroup, ASI. A population's PBS value conceptually represents the amount of allele frequency change at a given locus since its divergence from the other two populations. From this statistic, we intended to discover selection regions created after reaching of AFT to African continent.

Haplotype cluster frequency calculation

We used Hapflk method (Fariello, Boitard et al. 2013) to calculate haplotype cluster frequencies at a candidate region. The Reynold's genetic distance matrix between pairs of breeds was generated for a random subset of one percent of the total variants using Hapflk (Fariello, Boitard et al. 2013, Alberto, Boyer et al. 2018) and converted to the kinship matrix by a R script that the developer provided. fastPHASE (Scheet and Stephens 2006) and R scripts under https://github.com/inzilico/kselection were used to select the number of haplotype clusters, K in Hapflk program.

5.3.7 Annotation and functional enrichment analysis

The annotation for candidate regions was based on UMD 3.1 Gene Transfer Format file (.gtf) from Ensembl release 94 (Zerbino, Achuthan et al. 2017) and Cattle QTL db (Hu, Park et al. 2018). For functional enrichment analysis of candidate genes, functional annotation tool in DAVID v.6.8 (Huang, Sherman et al. 2009) was used based on GO Biological Process Direct terms with default settings. A FDR-corrected p-value of 0.05 was used as the threshold for statistical significance.

5.4 Results

5.4.1 Sequencing, mapping and identification of SNPs

Individual genomes of 116 indigenous African cattle and two African buffalo, *Syncerus caffer caffer* (AFB) were newly generated to ~11 X depth coverage each and were jointly genotyped with publicly available genomes of commercial taurine cattle breeds (Angus, Jersey, Holstein, and Hanwoo), Asian indicine cattle breeds, and indigenous African cattle breeds from our previous study (Kim, Hanotte et al. 2017). The total of 22 cattle breeds (except AFB) comprise cattle classified based on their phenotypes and geographic distribution as African Indicus (AFI: Afar, Arsi, Barka, Butana, Ethiopian Boran, Goffa, Kenana, and Mursi), African Sanga (AFS: Ankole, and Afar), and Zenga (AFZ: Fogera, and Horro), African Taurine (AFT: N'Dama, and Sheko), European Taurine (EUT: Angus, Holstein, Jersey), and Asian Taurine (Hanwoo) (Table 5.1).

In total, ~35 billion reads or ~3.50 Tbp of sequences were generated or retrieved. Sequence reads of each sample have aligned against the reference genome with average alignment rate of 98.71% and covered 98.57% of the reference genome. Concordant with previous analysis of zebu Nellore (Canavez, Luche et al. 2012), overall alignment rate of the African *B. indicus* and crossbreed samples (98.65% in average) to the taurine reference genome (UMD 3.1) was found comparable to the one obtained for the African taurine samples

(98.94% in average). In addition to this, AFB also has overall alignment rate (BF001: 89.02%, and BF002: 91.24%, respectively).

After variant calling and filtering steps, a total of ~60 million SNPs were finally retained. An average missing rate and minor allele frequency of the SNPs were 0.016 and 0.120, respectively. The density of SNP loci was 22.764/kbp along the genome (Table 5.2). Overall genotype concordance of 72 samples was 95.13%, between the additional genotype data and the re-sequencing results across the samples, providing confidence on the accuracy of SNP calling. Of the 72 samples, there were three samples that showed low genotype concordance. However, the DNA samples of these samples are not expected to produce reliable genotypes, considering the low call rates.

5.4.2 Population structure and genetic diversity of African cattle

Population structure and relationships

We performed principal component analysis, ignoring group and breed membership (Figure 5.2b). All of the African humped cattle samples (AFI, AFS, and AFZ) were located between EUT and ASI samples according to eigenvector 1, which explained ~10% of total variation, and they are also clearly separated from AFT. Therefore, There is no evidence for recent admixture between EUT and/or ASI and African cattle. Rather, these support ancient admixture for the African humped cattle with unique genetic material as explained by eigenvector 2 (~2.5% of total variation). Of African taurine cattle breeds, N'Dama was clearly separated from other breeds. However, Sheko that has been referred as AFT breed clustered with other African humped breeds as previously reported (Hanotte, Tawah et al. 2000, Mbole-Kariuki, Sonstegard et al. 2014), and it is even closer to them than the Sanga breed, Ankole. African humped breeds are not separated by their breed membership, and they didn't form distinct clusters except the Ankole breed. An identical pattern to PCA results was also observed in an individual-level phylogenetic tree reconstructed based on genome-wide SNPs (Figure 5.2a); We could not distinguish individual classified phenotypically as AFI, AFS and AFZ. The phylogenetic tree showed that AFI, AFS, and AFZ were clustered with ASI.

We then reconstructed the maximum likelihood tree of the 19 cattle populations using Treemix (Pickrell and Pritchard 2012) to address population-level relationships and to identify pairs of populations that are related to each other independent of that captured by this tree. The population-level phylogeny without any migration edges explained 95% of total variance and showed that most of African populations are clustered with ASI and are far closer to ASI than EUT except N'Dama. Migration events were added into initial tree until 99.8% of the variance in ancestry between populations was explained by the tree. The 13 migration edges added into the tree where the variance explained reached 99.8% (Figure 5.3), were statistically significant. They are mostly observed among AFI, AFS, AFZ and ASI breeds. (Figure 5.4).

In addition to the relationship of cattle breeds, we performed Admixture (Alexander, Novembre et al. 2009) analysis to address admixture patterns between groups as well as breeds. A subset of the total SNPs (~14 million SNP loci) was used with increasing K from 2 to 22. The cross validation of this analysis suggested K = 6 as the most likely number of genetically distinct groups for our data (cross validation error = 0.33). At K = 2, while both of EUT and ASI populations showed intact genetic background, all of the African cattle breeds apparently shared genome ancestry with taurine as well as indicine. The taurine ancestry of African hybrids (AFS, and AFZ) showed no difference compared to that of AFI except for Ankole. Also, Sheko also displays a similar amount of taurine ancestry to that of other African humped breeds. Approximately, 18% of each African humped cattle genomes (except Ankole) came from taurine ancestry. Of them, the Mursi population especially showed a higher proportion than admixed population. At K=3, most of taurine ancestry at K=2 has been replaced by putative African taurine ancestry (light blue). It indicates that the existing taurine ancestry across African humped cattle genomes originate from AFT. The admixture plot above K=3 showed intercontinental divergence and admixture; Asian taurine cattle, Hanwoo breed showed distinct ancestry, and most of African humped breed showed two ancestries different from ASI ancestry. North-western AFI including Barka, Butana and Kenena (Figure 5.1a), especially showed a different pattern to the other breeds (orange).

Genetic distance and diversity

To present genetic distances between the 22 cattle populations, pairwise *Fst* values was estimated (Figure 5.5). Based on these *Fst* values, the genetic background of EUT was observed to be definitely different from all other cattle breeds except N'Dama. The EUT showed ~0.2 and ~0.3 genetic distance in *Fst* values against African cattle breeds and ASI, respectively. As already shown in PCA and phylogenetic analysis, all of African breeds regardless of their classification were close to each other, except N'Dama, and they showed pairwise *Fst* close to zero.

The genetic diversity for the whole autosomal SNPs showed reduced levels of heterozygosity compared to all other breeds in the taurine cattle, except for the Sheko. Heterozygosity values of African humped cattle were similarly higher across breeds, which reflect that genetic diversity was consistently conserved across the Horn of Africa. Note that ASI breeds showed similar level of heterozygosity compared to African humped breeds, and showed higher level of heterogeneity within each breeds (Figure 5.6). The degree of inbreeding measured by runs of homozygosity (ROH) showed that taurine breeds including N'Dama have higher level of inbreeding compared to the other breeds. ASI breeds showed a similar pattern of ROH distribution to African humped breeds (Figure 5.7).

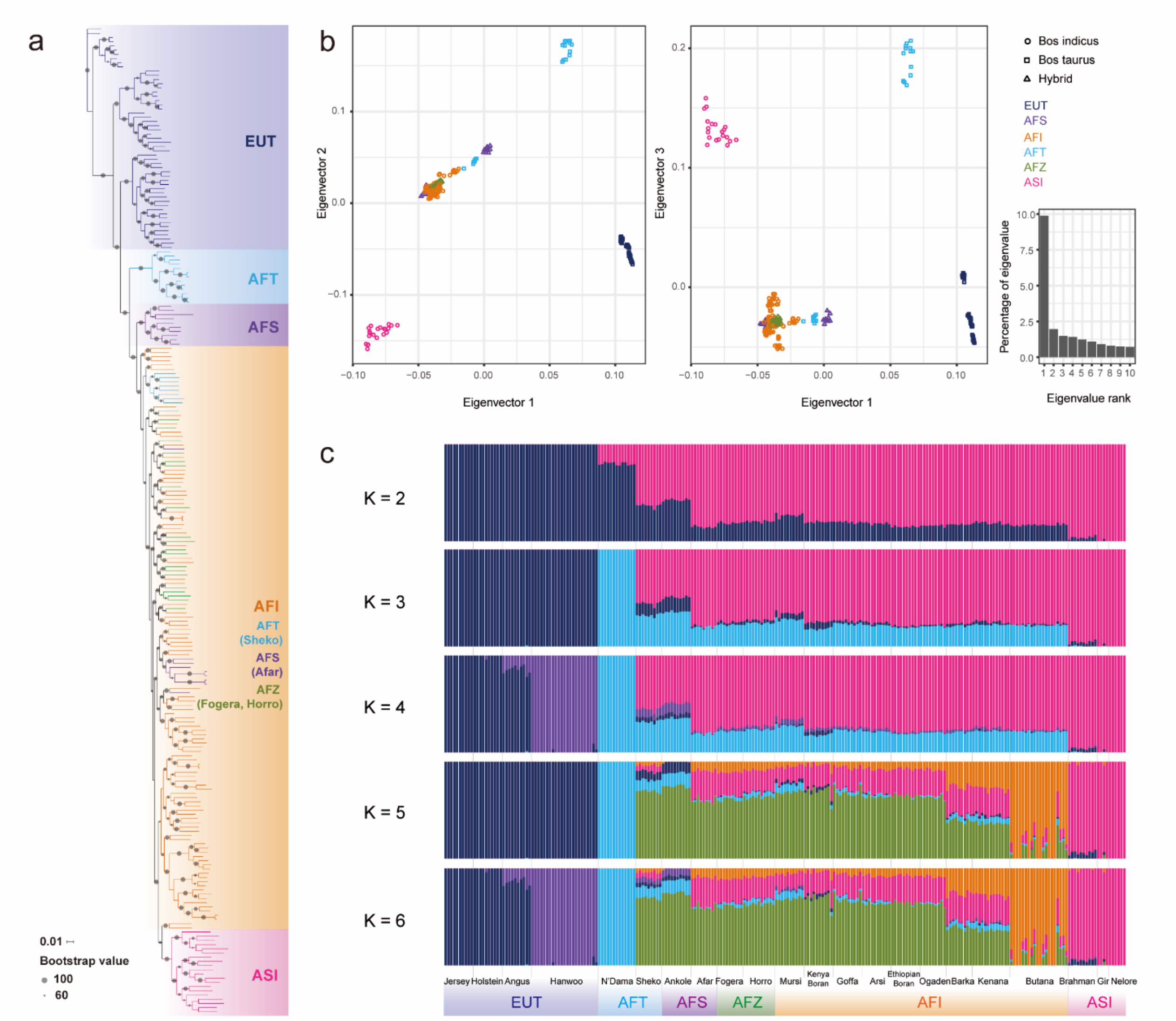


Figure 5.2 Populations structure of indigenous African cattle. (a) Maximum likelihood tree reconstructed for the 235 cattle samples. The size of dot at each node indicates bootstrap value. (b) PCA plot of 235 cattle samples. The shape and color of points indicate type and group information, respectively. (c) The results of admixture analysis by using cluster from K=2 to K=6

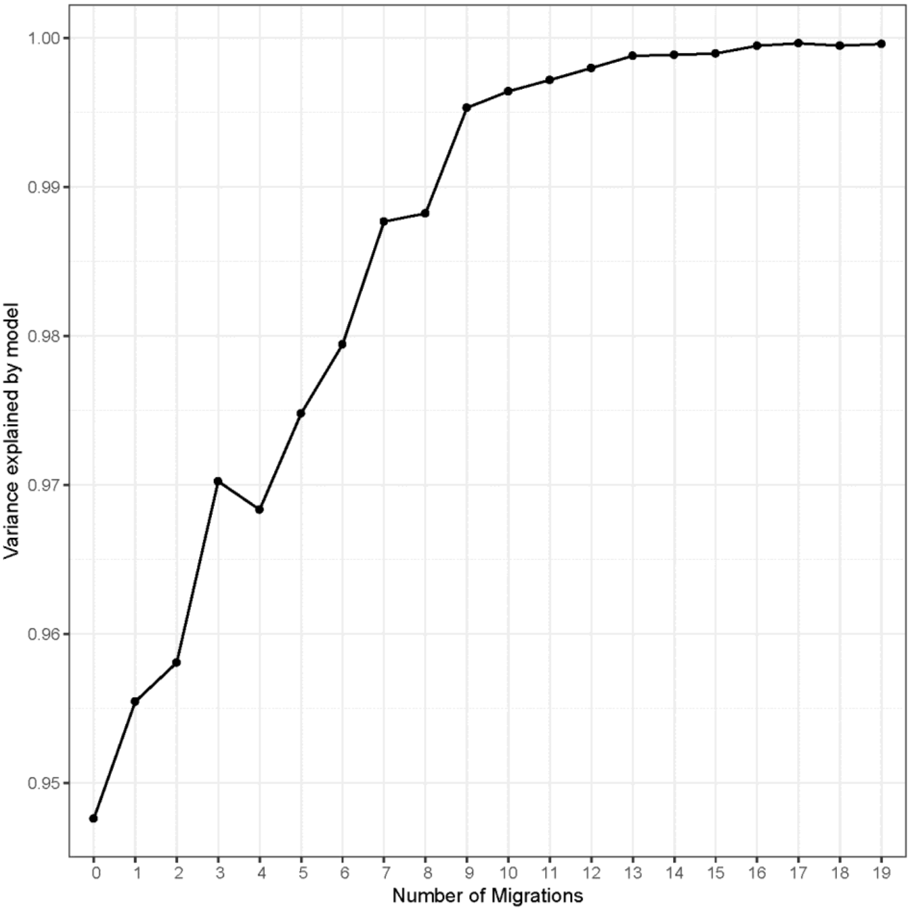


Figure 5.3 Variance explained by model in Treemix analysis

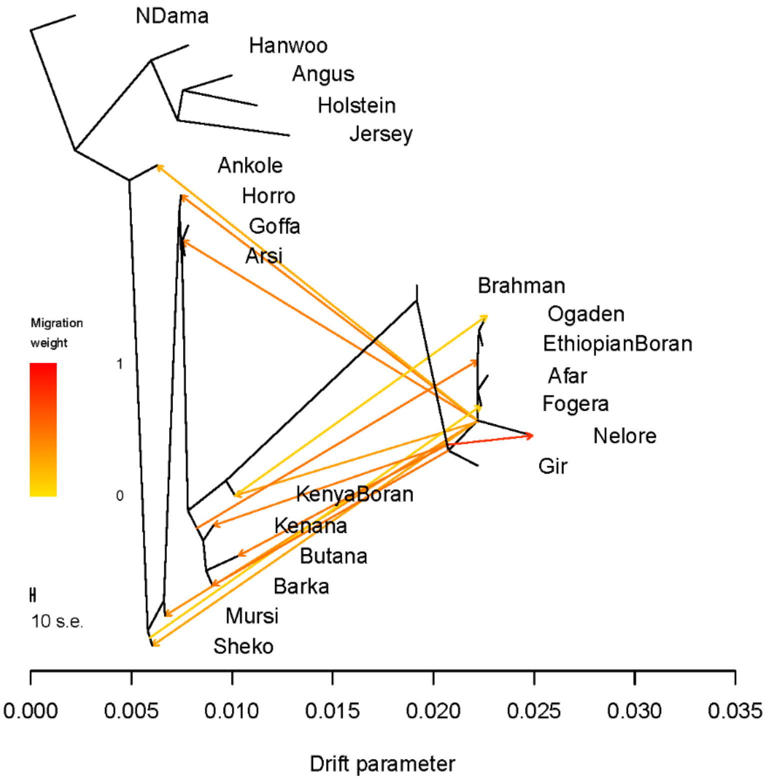


Figure 5.4 Population-based phylogeny with 13 migration edges

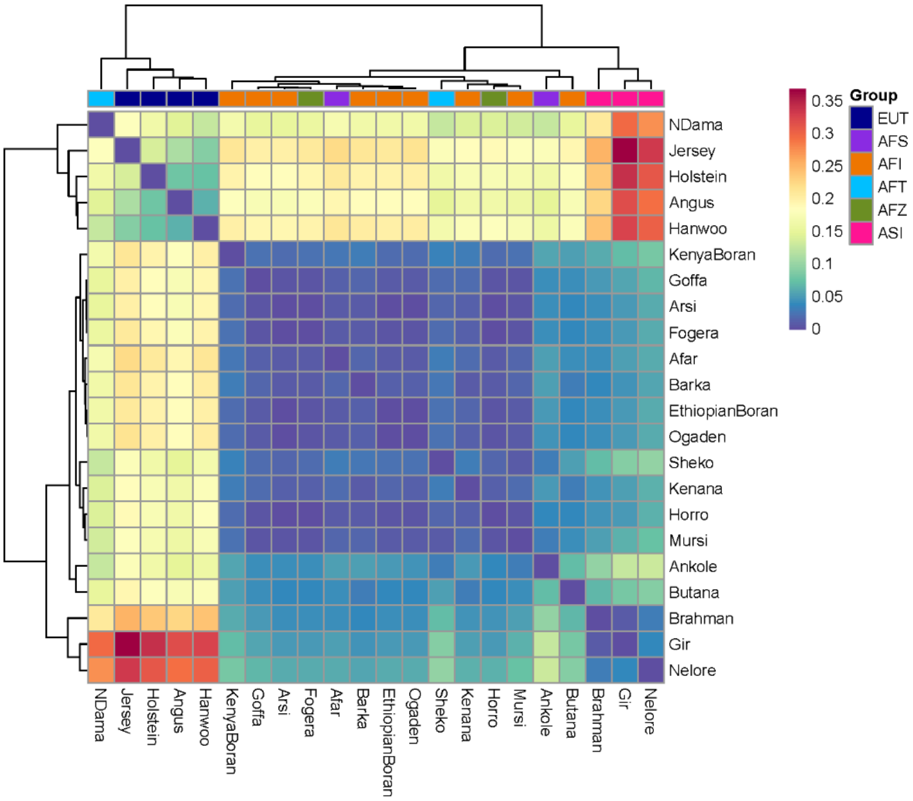


Figure 5.5 Mean pairwise *Fst* value between cattle breeds

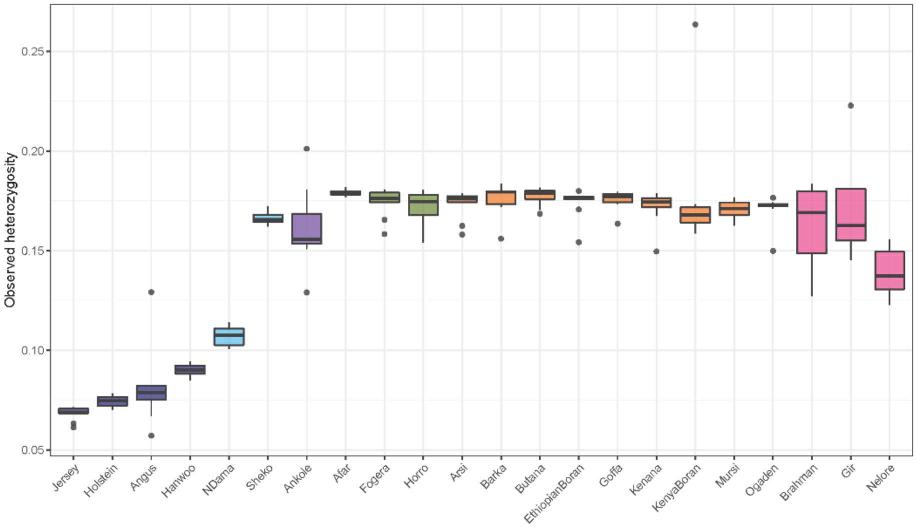


Figure 5.6 Observed heterozygosity of all cattle breeds in this study

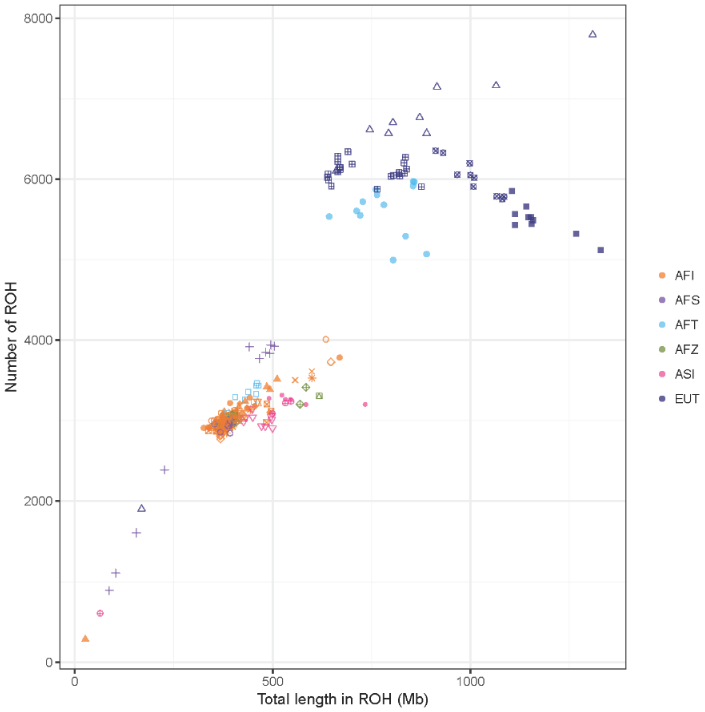


Figure 5.7 Runs of homozygosity patterns of all cattle breeds in this study

5.4.3 Admixed signatures in African indicine genomes

Signatures of intensive admixture across African cattle genomes

To further understand the significance and degree of admixture across African cattle breeds, allele sharing patterns were examined using ABBABABA statistics. Under assumption that EUT and ASI are pure populations that have intact genetic background from taurine and indicine, all African breeds showed an evidence of admixture (Figure 5.8a), consistent with the result of Admixture analysis. The average D-statistics and *f*-statistics across all breeds were 0.2742 and 0.3158, respectively, and Z-transformed D-statistics are much higher than zero in all breeds. Also, we could not observe a particular pattern according to group membership of each breed. For instance, Afar breed has a lowest proportion of taurine admixture across African indicine/hybrids (21.79%), although it is known as a crossbreed, sanga. Mursi breed rather has a higher proportion of taurine admixture than hybrids such as Horro (33.55%, Figure 5.8b). Interestingly, KenyaBoran and EthiopianBoran have different taurine admixture proportion, in spite of their short history since a split of populations.

Potential introgression loci identified by local taurine-indicine ancestry

Based on the observation that all African cattle breeds have genome ancestry from both of taurine and indicine, we estimated local ancestry along the genome to identify potential introgression loci from taurine to AFI, AFS and AFZ (140 individuals in total). In the estimation, each individual shared 39.9% haplotype ancestry with taurine on average, and for each locus, 110.6 of total 280 chromosomes were determined to have taurine haplotype ancestry on average. On a genome-wide window scale of 20 SNPs, we found 7,443 significant windows that have lower or higher indicine ancestry than the significance threshold ($\mathrm{mean\ ancestry} \pm 3 * \mathrm{standard\ deviation}$). The significant windows were annotated with overlapped protein-coding genes, which showed that the gene functions are significantly enriched in sensory perception and immune responses (G-protein coupled receptor signaling pathway, sensory perception of smell, detection of chemical stimulus involved in sensory perception, antigen processing and presentation, and antigen processing and presentation of peptide antigen via MHC class I, FDR-corrected p-value < 0.05, Figure 5.9c). As 20 SNPs for each window is an arbitrary criterion, the significant windows were then merged into a region to examine continuous signals with excess of taurine ancestry. This process resulted in 47 regions where all close windows are located within 1Mbp.

The longest region with significant excess of taurine ancestry was ~6Mb long at BTA 23. This region included 3,751 significant windows with ~81% of average taurine ancestry, much higher than the overall taurine ancestry across the whole genome (Figure 5.9a). Moreover, whole BTA 23 showed a lower level of indicine ancestry compared to that of whole genome (Figure 5.9b). A further examination using rIBD and pair-wise *Fst* values supported the excess of taurine ancestry, and showed the current status of the sequence at this region.

In rIBD analysis, AFI, AFS, and AFZ shared more IBD segments with EUT than ASI exactly at the same region which is also only one distinct and wide region with excess of IBD shared with taurine across the whole genome (Figure 5.10). The pair-wise *Fst* values showed that at some parts of the region (especially at each ends of the region), African population has higher level of genetic differentiation against ASI than EUT (Figure 5.9d). For the *Fst* distribution, the opposite trend was observed in the target region compared to that of the whole genome; the genome-wide *Fst* distribution of ASI *versus* African population was higher than that of EUT *versus* African population, while the opposite was observed in the target region (Figure 5.9e).

We then used gene and QTL information around the candidate region to understand its functional implication. The 121 protein-coding genes were found within the region, and all of the immune-related QTL at BTA 23 (10 Antibody-mediated immune response QTLs, and 3 Cell-mediated immune response QTLs) were found within or around the candidate region. Especially, the gene density of this region is exceptionally higher compared to the other regions in BTA 23. e.g. 10 protein-coding genes within 50kb. Moreover, the enrichment of gene functions across all significant windows with excess of taurine ancestry mostly derived from these 121 protein-coding genes.

We also found other candidate regions with excess of taurine ancestry. The strongest signal was found in an intergenic region between *FAM207A* and *ADARB1* gene at BTA1 with 96.7% taurine ancestry. In addition, other candidate regions including genes that are related to inflammation (*NLRC4*

(Zhao, Yang et al. 2011, Canna, de Jesus et al. 2014, Kitamura, Sasaki et al. 2014), and *SLC30A6* (Adelino, Addobbati et al. 2016)), immune system (*GP2* (Hase, Kawano et al. 2009), and *IBTK* (Liu, Quinto et al. 2001)), and spermatogenesis (*SRD5A2* (Elzanaty, Giwercman et al. 2006, Zhao, Wu et al. 2012)) were found. For the excess of indicine ancestry, we found only one region on *ITSN1* gene at BTA1. This region has nearly pure indicine ancestry (95.7% in average) compared to the whole genome level.

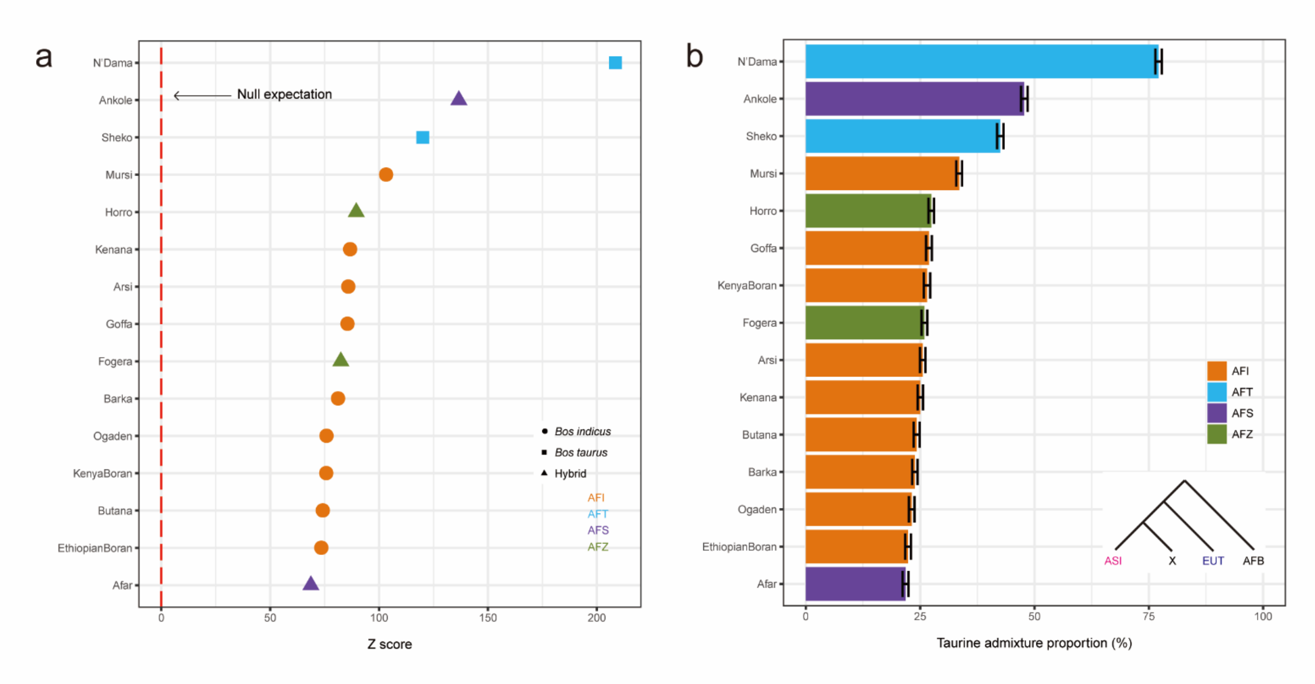


Figure 5.8 Admixed signatures in African indicine genomes. (a) Z transformed ABBABABA statistics. The dotted red line indicates expected statistic at a neutral locus (b) Admixture proportions measured by *f*-statistics. The inset indicates a phylogeny assumed in calculating both of ABBABABA and *f*-statistics

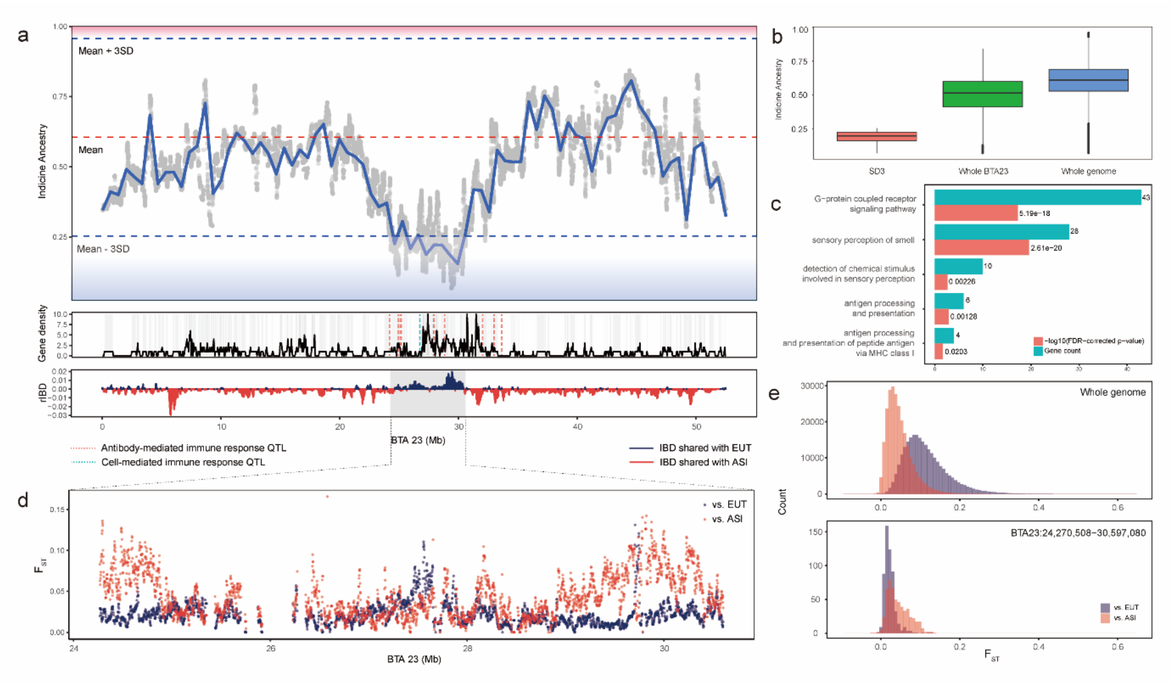
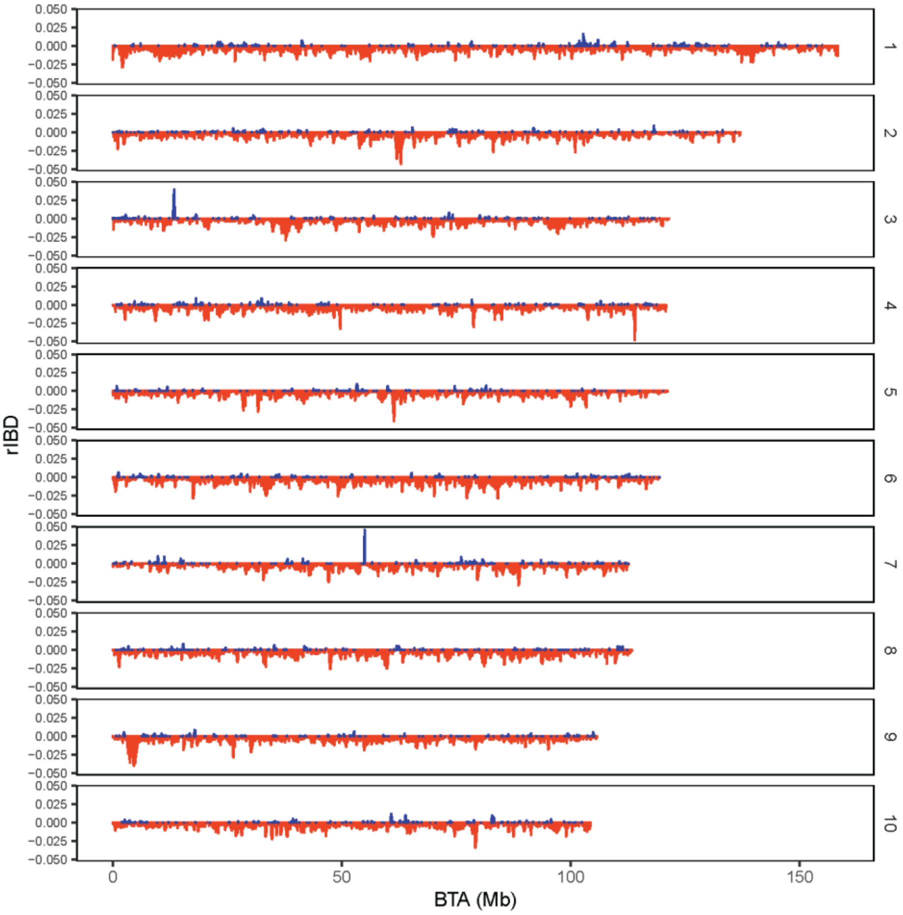
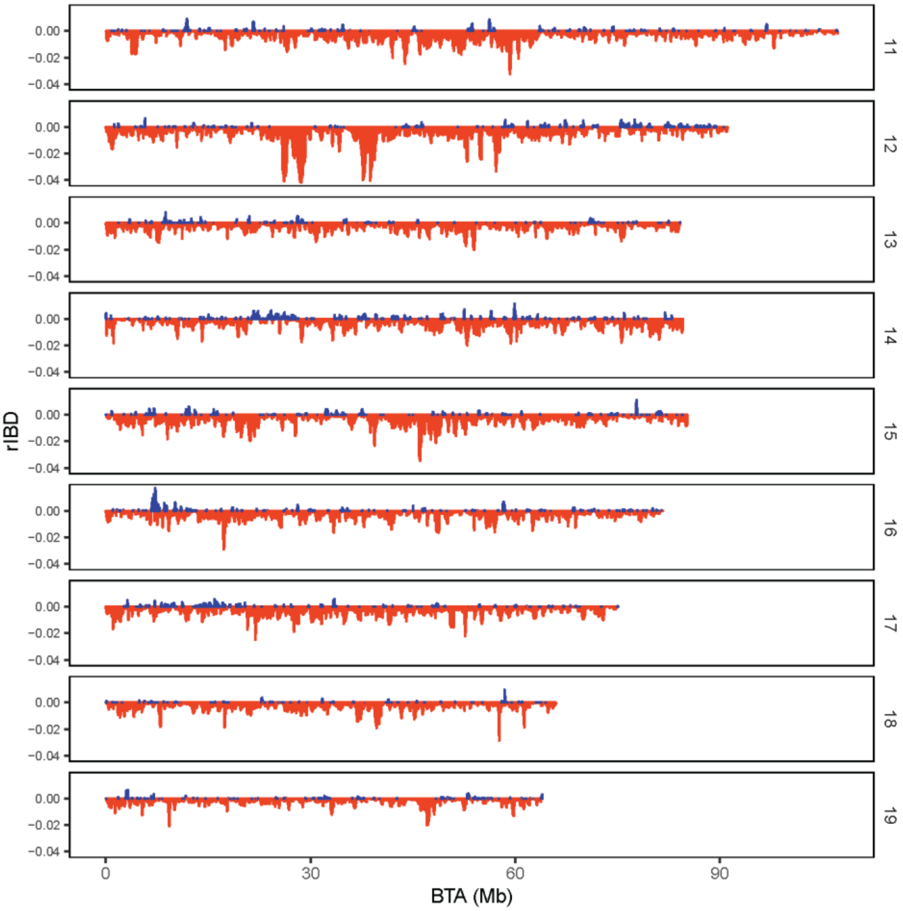


Figure 5.9 An example of introgression loci in African indicine genomes (a) Indicine ancestry inferred by LOTER. (b) Indicine ancestry according to genomic regions. (c) Result of gene set enrichment analysis. (d) Pairwise *Fst* value in the candidate locus. (e) Genome-wide distribution of pairwise *Fst* values compared to that of the candidate locus


rIBD
0.050
0.025
0.000
−0.025
−0.050
1
2
3
4
5
6
7
8
9
10
0
50
100
150
BTA (Mb)


0.00
−0.02
−0.04
rIBD
11
12
13
14
15
16
17
18
19
0
30
60
90
BTA (Mb)

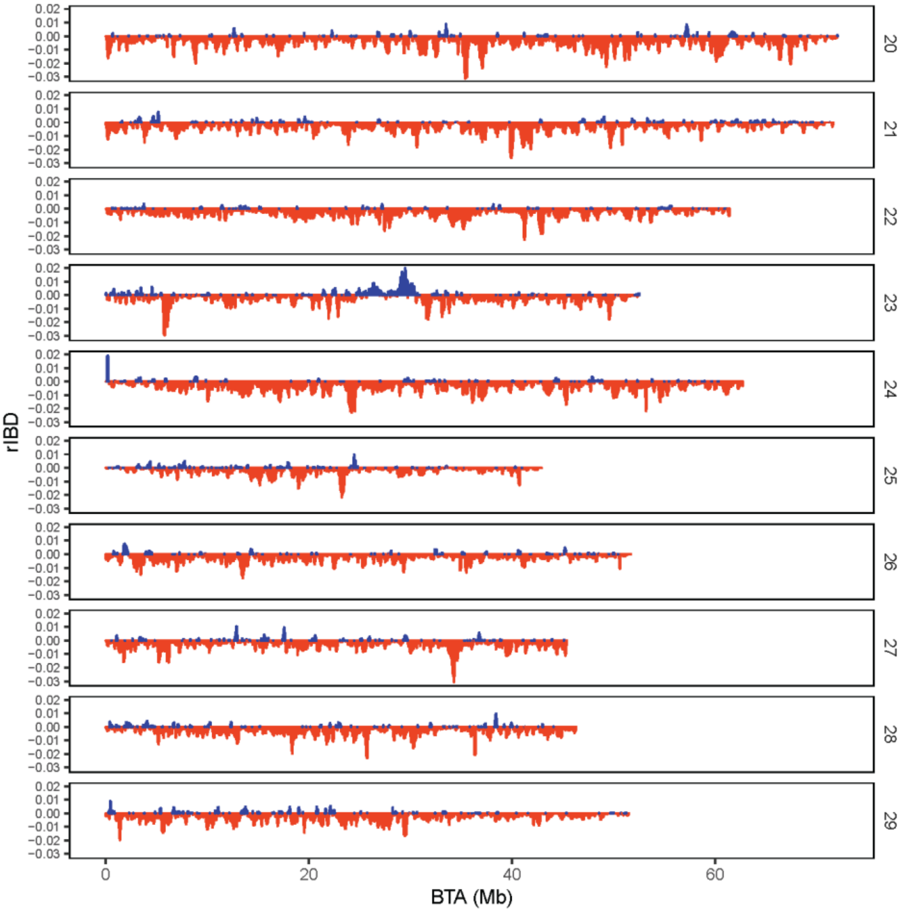


Figure 5.10 Distribution of rIBD values at a genome-wide level. rIBD has been calculated with 20kb window and 10kb step. Red: IBD sharing toward ASI, Blue: IBD sharing toward EUT

5.4.4 Selection signatures of African taurine cattle after split from European taurine cattle

Intensive indicine introgression into Eastern African taurine genomes

In the population structure analysis, we identified that the two AFT populations have some degree of indicine admixture. Especially, we repeatedly observed that the Sheko genomes has a significant indicine admixture even higher than for other hybrids such as Ankole (Sheko: 57.5%, and Ankole: 52.3%, Figure 5.8b), and the breed is genetically closer to those of AFI than AFT (Figure 5.5). As the Sheko breed was historically referred as an intact AFT population (Rege, Ayalew et al. 2006, Bahbahani, Afana et al. 2018), it shows that it has experienced intensive admixture with surrounding AFI populations at the opposite of the Western African breed, N'Dama. This intensive admixture in Sheko population has been previously shown (Hanotte, Tawah et al. 2000, Mbole-Kariuki, Sonstegard et al. 2014, Edea, Bhuiyan et al. 2015)

Potential barrier loci inferred by selection signatures of African taurine genomes

AFT is the most ancient African cattle population, and it would have had more time to adapt to the local environmental challenges of the African continent than African humped cattle. Therefore, we hypothesized that a selective sweep for environmental adaptation in African humped cattle might be of taurine ancestry. We also expected that Sheko admixed genomes preserved such selective sweeps.

To identify these loci, we used PBS analysis using both N'Dama and Sheko genomes. There were 1,255,150 windows with -0.05012 PBS in average (min = -0.64071, and max = 0.57369), and we found a region with a strongest signal upstream of *CARD11* gene at BTA 25 (Figure 5.11a). Pairwise *Fst* among the AFT, EUT, and ASI, showed high genetic distances at this region compared to the genome-wide genetic distances among these three groups (Figure 5.11b and Figure 5.12).

When we expanded the comparisons to the other African populations, pairwise *Fst* values were relatively high in any comparisons exactly at the 50kb window region with top PBS (BTA 25 40,603,333-40,653,333) (Figure 5.13a). In the haplotype clustering frequencies across breeds, N'Dama and Sheko breeds have distinct haplotypes compared to other breeds including control populations in PBS analysis (EUT, and ASI) (Figure 5.13b). Note that although most of Sheko breed's haplotypes were similar to that of N'Dama, which was a stronger signal than any other AFI breeds, Sheko breed has also some haplotypes similar to that of AFI, and the other AFI breeds such as Mursi have some degree of haplotype similar to N'Dama. We also validated the haplotype found at the region with additional populations including 1 West African taurine breed (Muturu), and 3 East African indicine breeds (Bagaria, Bale, and Semien). As the original dataset, Muturu that is also AFT breed has very similar haplotypes as N’Dama in contrary to other indicine breeds (Bagaria, Bale, and Semien).

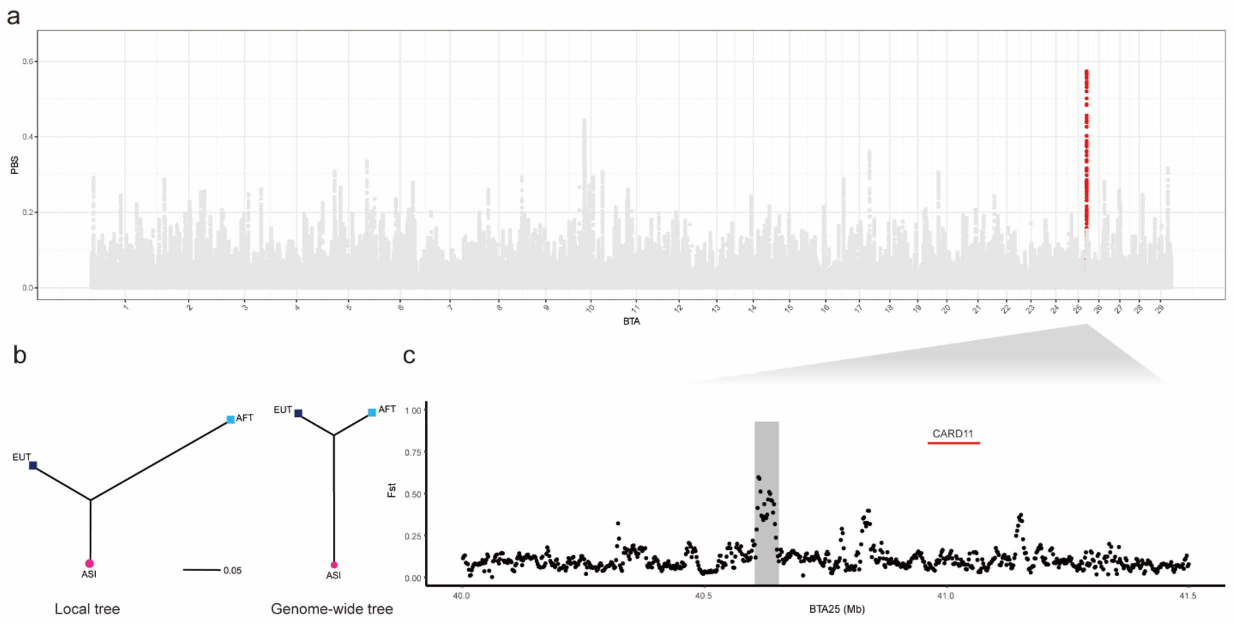


Figure 5.11 Selection signatures of African taurine cattle after split from European taurine cattle. (a) Genome-wide distribution of PBS values with 50kb window and 2kb step. (b) *Fst*-based phylogeny among AFT, EUT, and ASI. The branch length indicates *Fst* value. (c) Pairwise *Fst* value between AFT and EUT around the peak with highest PBS

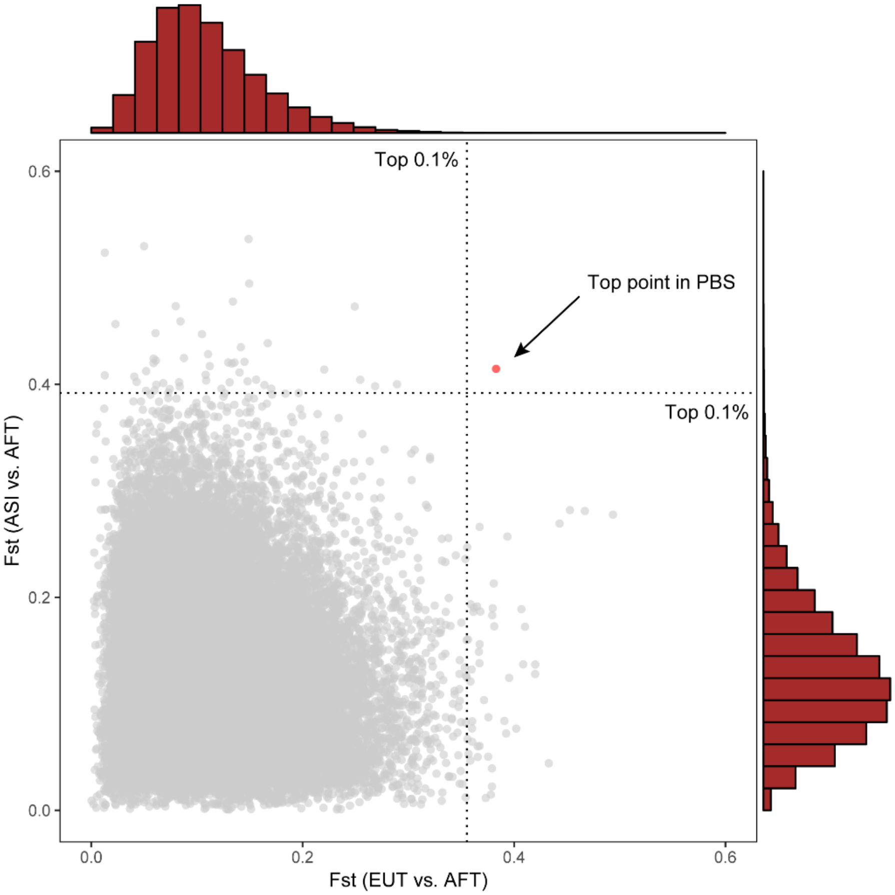


Figure 5.12 Distribution of Pairwise *Fst* values for two comparisons; EUT vs. AFT, and ASI vs. AFT

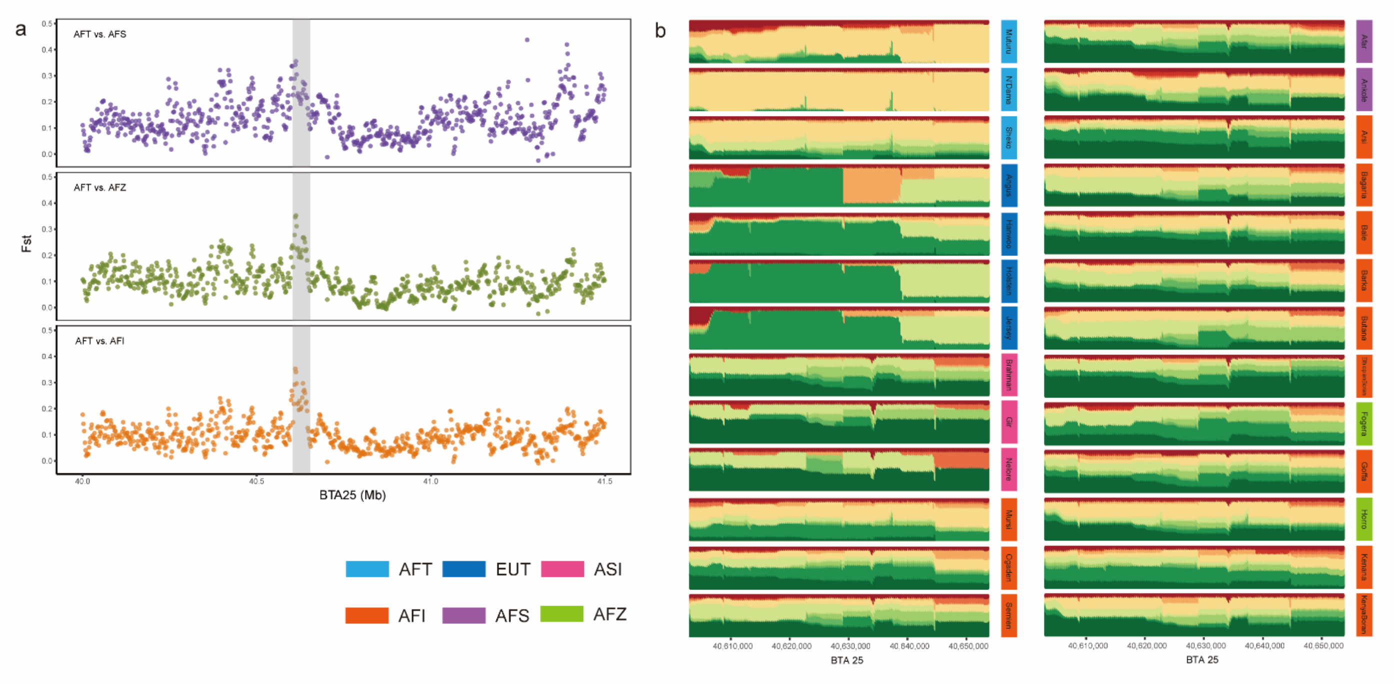


Figure 5.13 Comparisons with other breeds for the candidate locus at BTA25. (a) Pairwise *Fst* values around the peak with highest PBS. (b) Haplotype clustering frequencies at the 50kb window with highest PBS. Four breeds have been added into this analysis. (Bagaria, Bale, Muturu, and Semien)

5.5 Discussion

After its domestication, cattle have become the most widespread animal across the world within less than 10,000 years. One of its striking examples is the rapid dispersion of cattle populations across African continent, African cattle pastoralism. Especially, the admixture events caused by initial taurine introduction with following multiple indicine (zebu) introductions has been the basis of the present African cattle distribution and diversity.

In this study, we have highlighted the mosaic-like characteristics of present African cattle genomes, and suggested that such admixed diversity is at the root of the success of African cattle pastoralism. First, we demonstrated that all of the 15 indigenous populations are admixed, sharing genome ancestry with both taurine and indicine at a whole-genome scale. The overall status of population structure suggested an ancient admixture rather than recent admixture from EUT and ASI, which is consistent with historical records (Epstein 1971, Hanotte, Tawah et al. 2000) where zebu introductions has mainly occurred more than thousands years ago.

Interestingly, the populations that have been defined as crossbreeds, have no difference in their admixture proportion compared to other indicine breeds. Rather, Mursi breed that is known as indicine breeds, has the highest taurine admixture if we exclude Ankole breed. This is due to the genetic homogeneity as shown in the low levels of genetic differentiation between indicine populations. Several previous studies (Dadi, Tibbo et al. 2008, Edea, Dadi et al.

2013, Edea, Bhuiyan et al. 2015) have also reported the low levels of genetic differentiation in East African cattle. In pastoral systems, migrations of animals across different countries and uncontrolled mating between populations could frequently occur. The low level of genetic differentiation might be caused by an intensive gene flow between the populations in this system, as supported by the result of treemix analysis, where most of migration event were found between African indicine/hybrids populations (Figure 5.4). Also, we can infer that Mursi population might have experienced unique admixture history, considering relatively higher taurine ancestry that has been maintained despite of the intensive gene flow.

The mosaic of AFI/AFS/AFZ genomes showed extremely different taurine ancestry across genomic regions, which led us to find several regions with excess of taurine ancestry in 13 indicine populations. As shown in Figure 1b, the habitats of 13 populations are distributed around 'Horn of Africa' which is believed to be a main migration route of zebu introductions (Mwai, Hanotte et al. 2015). Therefore, we expected here to identify signatures of admixture that could represent the legacy of taurine introgression across the whole African continent.

Most of the significant regions were located in intergenic regions. However, the longest region at BTA23 contains more than 100 protein-coding genes. Moreover, those were enriched in olfactory receptor and immune functions. For mammals, olfaction is essential to adapt different environmental conditions by avoiding dangers and searching for food. Their genomes, therefore, contain more than hundreds of olfactory receptor genes in a form of gene clusters. For cattle genome, it has been estimated that there are nearly 1,000 functional olfactory receptor genes that are clustered in each 49 regions across 26 chromosomes (Niimura and Nei 2007, Lee, Nguyen et al. 2013). The candidate region we found in the local ancestry analysis was overlapped with the largest cluster in the cattle genome, which contained 41 olfactory receptor genes (~5% of total olfactory receptor genes).

The migration of zebu cattle from Asia is believed to be a big challenge in terms of environmental conditions such as food, weather, and local diseases. Therefore, we could hypothesize that the genetic variations from AFT that has been already adapted to local environments might provide a novel genetic resource to subsets of olfactory receptor genes, and help the introduced populations adapt to local environments within a relatively short time. Similarly, we can understand the changes of immune-related genes that might play a crucial role in local disease challenges. The link between olfactory receptor and immune genes and environmental conditions has been continuously suggested in previous studies (Niimura and Nei 2005, Niimura and Nei 2007, Hayden, Bekaert et al. 2010). For instance, Hayden *et. al.* suggested that adaptation to aquatic lifestyle have resulted in a loss of functional olfactory receptor genes in marine mammals (Hayden, Bekaert et al. 2010). In our result, it was difficult to expect such a big alteration, considering the extent of divergence between taurine and indicine cattle since their common ancestor. Instead, we could detect a broad pattern of genetic differentiation and haplotype sharing across the candidate region that is significantly different to that of whole-genome. The difference is, therefore, expected to mostly affect the functions of gene in different ways (e.g. cis-regulation of gene expression) rather than non-synonymous changes or duplications of protein-coding genes.

African trypanosomosis is caused by tsetse fly-transmitted parasite of the genus Trypanosoma and is a severe obstacle on productivity of livestock in Africa (Courtin, Berthier et al. 2008). AFT is believed to genetically have some level of tolerance to control the disease development, which is called trypanotolerance (Murray, Trail et al. 1984). N'Dama breeds is especially well-characterized in that character (Mattioli, Pandey et al. 2000), whereas Sheko breed is believed to have relatively reasonable level of trypanotolerance (Lemecha, Mulatu et al. 2006). In our result of PBS analysis, a selection signature common to these two breeds was located in upstream of *CARD11* gene. The protein encoded by this gene is a member of membrane-associated guanylate kinase (MAGUK) family, and carries a characteristic caspase-associated recruitment domain (CARD) (Safran, Dalah et al. 2010). Also, *CARD11* is essential for signaling of T- and B-cells in both the innate and adaptive immune system (Pomerantz, Denny et al. 2002, Hara, Wada et al. 2003), and more importantly, it is previously reported to be related to trypanotolerance as a positive selection gene (Kim, Ka et al. 2017) and a differentially expressed gene (Noyes, Brass et al. 2011). Therefore, we suggest that the candidate region could play a role in a regulation of *CARD11* expression, and contribute the adaptation of AFT under selection pressure caused by trypanosomiosis. If the hypothesis is true, we might be able to suggest one possible explanation for the reason why this haplotype has been maintained in spite of the severe introgression from indicine cattle into AFT breeds. However, it is difficult to insist that the potential regulatory element is critical to control the disease development, as the trypanotolerance is likely to be a complex quantitative trait according to previous studies (Naessens, Teale et al. 2002, Hanotte, Ronin et al. 2003, Courtin, Berthier et al. 2008). Therefore, the potential regulatory element should be understood as one of genetic factors to achieve trypanotolerance.

Interestingly, the weaker signal of Sheko haplotypes than that of N'Dama or Muturu's, is consistent with the known level of trypanotolerance in each breeds, and some of indicine populations presented weak signal, however, similar haplotype as that of AFT breeds. Especially, Mursi breeds that have been reported to have relatively reasonable levels of trypanotolerance (Terefe, Haile et al. 2015) is shown to have relatively high frequency of haplotypes similar to that of AFT. This could support the link between the potential regulatory element and trypanotolerance. Also, the existence of haplotypes with low frequencies across some breeds suggests a possibility of unknown breeds with low level of trypanotolerance.

General discussion

Evolutionary processes can leave distinctive footprints on the genome that can be detected by various molecular markers as shown in chapter 2-5. As each molecular marker can represent different time scale of evolution, the selection of markers should take into account for the time scale when we are interested in. For example, amino acid changes that are directly related to functional changes have been considered in a comparative analysis between species diverged more than 9 million years ago in chapter 2, while allele frequency changes have been used to compare populations diverged less than 1 million years ago in chapter 5. In addition to time scale, various evolutionary conditions of each species were considered. In chapter 3, tandem repeats that have much higher mutation rates than other genomic regions have been used to infer the reason of accelerated evolution in human lineage. The chapter 5 also dealt with acceleration of adaptation that might be caused by gene flow between populations. In chapter 4, very recent trend of inbreeding has been estimated by using ROH that is a distinct genomic structure made by mating between closely related individuals. In this regard, this thesis has demonstrated the usage of diverse markers to infer the origin of evolutionary footprints in various conditions as well as different time scales. Especially, the novel trials using structural variations such as tandem repeats and ROH suggest that relevant molecular markers can expand our research scope.

Signatures of evolutionary process are not always to be interpreted in one way. For example, other evolutionary forces (e.g. genetic drift) than positive selection could result in false positive selection signatures. For this reason, evolutionary signatures should be interpreted with caution and supporting evidences. The findings of this thesis thus emphasizes rather on hypothesis generating than hypothesis testing.

Sequencing technologies have been moving from their second to third generation. One key achievement of this third generation, which has already been reached to some extent, would be the production of much longer reads. Longer reads are expected to solve many issues arising in the current short read-based studies. For example, the assembly quality in chapter 2 could be significantly improved, and repetitive variations such as tandem repeats in chapter 3 could be obtained more precisely. This is likely to generate more reliable genomic variations and evolutionary signatures.

The genetic diversity is an important concern in animal breeding both in reducing inbreeding depression and enhancing productivity. However, the genetic diversity within cattle populations, especially in commercial taurine populations, is being threatened by inbreeding. In this context, the approaches used in chapter 4 are advantageous for monitoring the genetic diversity and for controlling the deleterious effects of inbreeding. The latter will be especially useful when investigating cattle populations managed under breeding programs such as the Hanwoo. On the other hand, indigenous African cattle is placed on a different situation. At present, around 150 African cattle breeds or populations are recognized in the African continent. They maintain high genome diversity that may have contributed to their adaptation towards various environmental conditions. However, uncontrolled mating between cattle breeds including exotic breeds is rapidly reducing the genetic diversity across the African continent as shown in chapter 5. The opportunity to explore the genetic diversity, resource for animal breeding and fundamental research, may not last for very much longer. It constitutes another reason to perform genomic studies such as the ones in chapters 4 and 5.

References

Abbott, R., et al. (2013). "Hybridization and speciation." Journal of evolutionary biology **26**(2): 229-246.

Adelino, J., et al. (2016). "A genetic variant within SLC30A6 has a protective role in the severity of rheumatoid arthritis." Scandinavian journal of rheumatology: 1-2.

Agrawal, P. B., et al. (2014). "SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy." The American Journal of Human Genetics **95**(2): 218-226.

Ahmed, M. and P. Liang (2012). "Transposable elements are a significant contributor to tandem repeats in the human genome." Comparative and functional genomics **2012**.

Alasti, F., et al. (2008). "A novel TECTA mutation confirms the recognizable phenotype among autosomal recessive hearing impairment families." International journal of pediatric otorhinolaryngology **72**(2): 249-255.

Alberto, F. J., et al. (2018). "Convergent genomic signatures of domestication in sheep and goats." Nature communications **9**(1): 813.

Alexander, D. H., et al. (2009). "Fast model-based estimation of ancestry in unrelated individuals." Genome research **19**(9): 1655-1664.

Allard, R., et al. (1968). "The genetics of inbreeding populations." Advances in genetics **14**: 55-131.

Altschul, S. F., et al. (1990). "Basic local alignment search tool." Journal of molecular biology **215**(3): 403-410.

Altschul, S. F., et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research **25**(17): 3389-3402.

Andersen, H. T. (1966). "Physiological adaptations in diving vertebrates." Physiological Reviews **46**(2): 212-243.

Andrews, S. (2010). "FastQC: A quality control tool for high throughput sequence data." Reference Source.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arnold, M. L. and K. Kunte (2017). "Adaptive genetic exchange: a tangled history of admixture and evolutionary innovation." Trends in ecology & evolution **32**(8): 601-611.

Ashkenazy, H., et al. (2012). "FastML: a web server for probabilistic reconstruction of ancestral sequences." Nucleic acids research **40**(W1): W580-W584.

Auton, A., et al. (2009). "Global distribution of genomic diversity underscores rich complex history of continental human populations." Genome research **19**(5): 795-803.

Bär, E., et al. (2014). "IL-17 regulates systemic fungal immunity by controlling the functional competence of NK cells." Immunity **40**(1): 117-127.

Bahbahani, H., et al. (2018). "Genomic signatures of adaptive introgression and environmental adaptation in the Sheko cattle of southwest Ethiopia." PloS one **13**(8): e0202479.

Bao, Z. and S. R. Eddy (2002). "Automated de novo identification of repeat sequence families in sequenced genomes." Genome research **12**(8): 1269-1276.

Barbash, S. and T. P. Sakmar (2017). "Brain gene expression signature on primate genomic sequence evolution." Scientific reports **7**(1): 17329.

Barrett, R. D., et al. (2019). "Linking a mutation to survival in wild mice." Science **363**(6426): 499-504.

Beiter, E. R., et al. (2017). "Polygenic selection underlies evolution of human brain structure and behavioral traits." BioRxiv: 164707.

Bell, M. V., et al. (1998). "Influence of intron length on alternative splicing of CD44." Molecular and Cellular Biology **18**(10): 5930-5941.

Bennett, S., et al. (1995). "Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus." Nature genetics **9**(3): 284.

Benson, G. (1999). "Tandem repeats finder: a program to analyze DNA sequences." Nucleic acids research **27**(2): 573-580.

Berta, A. (2002). "Pinnipedia, overview." J. Zool **83**: 1525-1531.

Berta, A., et al. (2005). Marine mammals: evolutionary biology, Academic Press.

Berthier, D., et al. (2016). "Tolerance to trypanosomatids: a threat, or a key for disease elimination?" Trends in parasitology **32**(2): 157-168.

Bieberstein, N. I., et al. (2012). "First exon length controls active chromatin signatures and transcription." Cell Rep **2**(1): 62-68.

Bigham, A. W., et al. (2009). "Identifying positive selection candidate loci for high-altitude adaptation in Andean populations." Human genomics **4**(2): 79.

Biswas, S. and J. M. Akey (2006). "Genomic insights into positive selection." TRENDS in Genetics **22**(8): 437-446.

Bjelland, D., et al. (2013). "Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding." journal of dairy Science **96**(7): 4697-4706.

Blanchette, M., et al. (2004). "Aligning multiple genomic sequences with the threaded blockset aligner." Genome research **14**(4): 708-715.

Blanco, E., et al. (2007). "Using geneid to identify genes." Current protocols in bioinformatics: 4.3. 1-4.3. 28.

Boeckmann, B., et al. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic acids research **31**(1): 365-370.

Boetzer, M., et al. (2010). "Scaffolding pre-assembled contigs using SSPACE." Bioinformatics **27**(4): 578-579.

Bolger, A. M., et al. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics: btu170.

Bolger, A. M., et al. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Bonfiglio, S., et al. (2012). "Origin and spread of Bos taurus: new clues from mitochondrial genomes belonging to haplogroup T1." PloS one **7**(6): e38601.

Bosse, M., et al. (2014). "Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression." Nature communications **5**: 4392.

Bradley, D. G., et al. (1996). "Mitochondrial diversity and the origins of African and European cattle." Proceedings of the National Academy of Sciences **93**(10): 5131-5135.

Brawand, D., et al. (2011). "The evolution of gene expression levels in mammalian organs." nature **478**(7369): 343.

Brotherstone, S. and M. Goddard (2005). "Artificial selection and maintenance of genetic variance in the global dairy cow population." Philosophical Transactions of the Royal Society of London B: Biological Sciences **360**(1459): 1479-1488.

Browning, B. L. and S. R. Browning (2013). "Detecting identity by descent and estimating genotype error rates in sequence data." The American Journal of Human Genetics **93**(5): 840-851.

Browning, S. R. (2008). "Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes." Genetics **178**(4): 2123-2132.

Browning, S. R. and B. L. Browning (2007). "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 **81**(5): 1084-1097.

Bryc, K., et al. (2010). "Genome-wide patterns of population structure and admixture in West Africans and African Americans." Proceedings of the National Academy of Sciences **107**(2): 786-791.

Butler, D., et al. (2009). "ASReml-R reference manual." The State of Queensland, Department of Primary Industries and Fisheries, Brisbane.

Butler, J. M. (2006). "Genetics and genomics of core short tandem repeat loci used in human identity testing." Journal of forensic sciences **51**(2): 253-265.

Cáceres, M., et al. (2003). "Elevated gene expression levels distinguish human from non-human primate brains." Proceedings of the National Academy of Sciences **100**(22): 13030-13035.

Canavez, F. C., et al. (2012). "Genome sequence and assembly of Bos indicus." Journal of Heredity **103**(3): 342-348.

Canna, S. W., et al. (2014). "An activating NLRC4 inflammasome mutation causes autoinflammation with recurrent macrophage activation syndrome." Nature genetics **46**(10): 1140.

Carroll, S. B. (2005). "Evolution at two levels: on genes and form." PLoS biology **3**(7): e245.

Casper, J., et al. (2017). "The UCSC Genome Browser database: 2018 update." Nucleic acids research **46**(D1): D762-D769.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Molecular biology and evolution **17**(4): 540-552.

Charlesworth, D. and J. H. Willis (2009). "The genetics of inbreeding depression." Nature reviews. Genetics **10**(11): 783.

Chen, H., et al. (2010). "Population differentiation as a test for selective sweeps." Genome research **20**(3): 393-402.

Chen, N., et al. (2018). "Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia." Nature communications **9**(1): 2337.

Chen, W. V., et al. (2012). "Functional significance of isoform diversification in the protocadherin gamma gene cluster." Neuron **75**(3): 402-409.

Chikina, M., et al. (2016). "Hundreds of genes experienced convergent shifts in selective pressure in marine mammals." Molecular biology and evolution **33**(9): 2182-2192.

Cingolani, P., et al. (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly **6**(2): 80-92.

Cleveland, M., et al. (2005). "Changes in inbreeding of US Herefords during the twentieth century." Journal of animal science **83**(5): 992-1001.

Collin, R. W., et al. (2008). "Mid-frequency DFNA8/12 hearing loss caused by a synonymous TECTA mutation that affects an exonic splice enhancer." European journal of human genetics **16**(12): 1430-1436.

Collin, R. W., et al. (2008). "Mid-frequency DFNA8/12 hearing loss caused by a synonymous TECTA mutation that affects an exonic splice enhancer." European journal of human genetics **16**(12).

Consortium, B. H. (2009). "Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds." Science **324**(5926): 528-532.

Consortium, E. P. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." nature **447**(7146): 799.

Consortium, G. P. (2015). "A global reference for human genetic variation." nature **526**(7571): 68.

Consortium, I. H. G. S. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860.

Consortium, U. (2014). "UniProt: a hub for protein information." Nucleic acids research: gku989.

Contente, A., et al. (2002). "A polymorphic microsatellite that mediates induction of PIG3 by p53." Nature genetics **30**(3): 315.

Courtin, D., et al. (2008). "Host genetics in African trypanosomiasis." Infection, Genetics and Evolution **8**(3): 229-238.

Cypowyj, S., et al. (2012). "Immunity to infection in IL-17-deficient mice and humans." European journal of immunology **42**(9): 2246-2254.

Dadi, H., et al. (2008). "Microsatellite analysis reveals high genetic diversity but low genetic structure in Ethiopian indigenous cattle populations." Animal Genetics **39**(4): 425-431.

Danecek, P., et al. (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.

Dang, C.-G., et al. (2011). "Estimation of inbreeding coefficients and effective population size in breeding bulls of Hanwoo (Korean cattle)." Journal of animal science and technology.

Dang, C.-G., et al. (2011). "Estimation of inbreeding coefficients and effective population size in breeding bulls of Hanwoo (Korean cattle)." Journal of animal science and technology **53**(4): 297-302.

Dasmahapatra, K. K., et al. (2012). "Butterfly genome reveals promiscuous exchange of mimicry adaptations among species." Nature **487**(7405): 94.

Davis, C. S., et al. (2004). "A phylogeny of the extant Phocidae inferred from complete mitochondrial DNA coding regions." Molecular phylogenetics and evolution **33**(2): 363-377.

Davis, R. W. (2014). "A review of the multi-level adaptations for maximizing aerobic dive duration in marine mammals: from biochemistry to behavior." Journal of Comparative Physiology B **184**(1): 23-53.

De Bie, T., et al. (2006). "CAFE: a computational tool for the study of gene family evolution." Bioinformatics **22**(10): 1269-1271.

Decker, J. E., et al. (2014). "Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle." PLoS genetics **10**(3): e1004254.

Delaneau, O., et al. (2012). "A linear complexity phasing method for thousands of genomes." Nature methods **9**(2): 179-181.

Dennis, G., et al. (2003). "DAVID: database for annotation, visualization, and integrated discovery." Genome biology **4**(9): R60.

Di Lernia, S., et al. (2013). "Inside the “African cattle complex”: Animal burials in the Holocene central Sahara." PloS one **8**(2): e56879.

Dias-Alves, T., et al. (2018). "Loter: A software package to infer local ancestry for a wide range of species." Molecular biology and evolution **35**(9): 2318-2326.

Dietrich, M. R. (2010). "Microevolution and macroevolution are governed by the same processes." Francisco J. Ayala and Robert Arp: 169.

Dumas, L. J., et al. (2012). "DUF1220-domain copy number implicated in human brain-size pathology and evolution." The American Journal of Human Genetics **91**(3): 444-454.

Dunn, O. J. (1961). "Multiple comparisons among means." Journal of the American Statistical Association **56**(293): 52-64.

Durand, E. Y., et al. (2011). "Testing for ancient admixture between closely related populations." Molecular biology and evolution **28**(8): 2239-2252.

Edea, Z., et al. (2015). "Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds." Animal **9**(2): 218-226.

Edea, Z., et al. (2013). "Genetic diversity, population structure and relationships in indigenous cattle populations of Ethiopia and Korean Hanwoo breeds using SNP markers." Frontiers in genetics **4**: 35.

Efron, B. (1981). "Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods." Biometrika **68**(3): 589-599.

Elzanaty, S., et al. (2006). "Significant impact of 5α-reductase type 2 polymorphisms on sperm concentration and motility." International journal of andrology **29**(3): 414-420.

Enard, W., et al. (2009). "A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice." Cell **137**(5): 961-971.

Epstein, H. (1971). The origin of the domestic animals of Africa, Africana publishing corporation.

Esse, P., et al. (2001). "Decomposition of and nutrient release from ruminant manure on acid sandy soils in the Sahelian zone of Niger, West Africa." Agriculture, ecosystems & environment **83**(1-2): 55-63.

Fariello, M. I., et al. (2013). "Detecting signatures of selection through haplotype differentiation among hierarchically structured populations." Genetics **193**(3): 929-941.

Fay, J. C. and C.-I. Wu (2000). "Hitchhiking under positive Darwinian selection." Genetics **155**(3): 1405-1413.

Firth, A. L., et al. (2009). "Hypoxia selectively inhibits KCNA5 channels in pulmonary artery smooth muscle cells." Annals of the New York Academy of Sciences **1177**(1): 101-111.

Fish, F. E., et al. (2008). "Hydrodynamic flow control in marine mammals." Integrative and Comparative Biology **48**(6): 788-800.

Fishilevich, S., et al. (2017). "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards." Database **2017**.

Flicek, P., et al. (2011). "Ensembl 2012." Nucleic acids research **40**(D1): D84-D90.

Flori, L., et al. (2009). "The genome response to artificial selection: a case study in dairy cattle." PLOS ONE **4**(8): e6595.

Foote, A. D., et al. (2015). "Convergent evolution of the genomes of marine mammals." Nature genetics **47**(3): 272.

Frantz, L. A., et al. (2015). "Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes." Nature genetics **47**(10): 1141.

Frazer, K. A., et al. (2009). "Human genetic variation and its contribution to complex traits." Nature Reviews Genetics **10**(4): 241.

Fu, Y.-X. and W.-H. Li (1993). "Statistical tests of neutrality of mutations." Genetics **133**(3): 693-709.

Fulton, T. L. and C. Strobeck (2010). "Multiple markers and multiple individuals refine true seal phylogeny and bring molecules and morphology back in line." Proceedings of the Royal Society of London B: Biological Sciences **277**(1684): 1065-1070.

Gemayel, R., et al. (2012). "Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences." Genes **3**(3): 461-480.

Gemayel, R., et al. (2010). "Variable tandem repeats accelerate evolution of coding and regulatory sequences." Annual review of genetics **44**: 445-477.

Geza, E., et al. (2018). "A comprehensive survey of models for dissecting local ancestry deconvolution in human genome." Briefings in bioinformatics.

Ghani, M., et al. (2015). "Association of long runs of homozygosity with Alzheimer disease among African American individuals." JAMA neurology **72**(11): 1313-1323.

Glick, G., et al. (2012). "Signatures of contemporary selection in the Israeli Holstein dairy cattle." Animal genetics **43**(s1): 45-55.

Gnerre, S., et al. (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." Proceedings of the National Academy of Sciences **108**(4): 1513-1518.

Gouveia, J. J. d. S., et al. (2014). "Identification of selection signatures in livestock species." Genetics and molecular biology **37**(2): 330-342.

Gu, J. and X. Gu (2003). "Induced gene expression in human brain after the split from chimpanzee." Trends in Genetics **19**(2): 63-65.

Gymrek, M., et al. (2016). "Abundant contribution of short tandem repeats to gene expression variation in humans." Nature genetics **48**(1): 22.

Gymrek, M., et al. (2017). "Interpreting short tandem repeat variations in humans using mutational constraint." Nat Genet **49**(10): 1495-1501.

Haas, B. J., et al. (2008). "Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments." Genome biology **9**(1): R7.

Hamada, H., et al. (1984). "Enhanced gene expression by the poly (dT-dG). poly (dC-dA) sequence." Molecular and cellular biology **4**(12): 2622-2630.

Hannan, A. J. (2018). "Tandem repeats mediating genetic plasticity in health and disease." Nature Reviews Genetics **19**(5): 286.

Hanotte, O., et al. (2002). "African pastoralism: genetic imprints of origins and migrations." Science **296**(5566): 336-339.

Hanotte, O., et al. (2003). "Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle." Proceedings of the National Academy of Sciences **100**(13): 7443-7448.

Hanotte, O., et al. (2000). "Geographic distribution and frequency of a taurine Bos taurus and an indicine Bos indicus Y specific allele amongst sub-Saharan African cattle breeds." Molecular ecology **9**(4): 387-396.

Hansen, P. (2004). "Physiological and cellular adaptations of zebu cattle to thermal stress." Animal reproduction science **82**: 349-360.

Hara, H., et al. (2003). "The MAGUK family protein CARD11 is essential for lymphocyte activation." Immunity **18**(6): 763-775.

Harrison, R. G. and E. L. Larson (2014). "Hybridization, introgression, and the nature of species boundaries." Journal of Heredity **105**(S1): 795-809.

Hase, K., et al. (2009). "Uptake through glycoprotein 2 of FimH+ bacteria by M cells initiates mucosal immune response." Nature **462**(7270): 226.

Hasegawa, M., et al. (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." Journal of molecular evolution **22**(2): 160-174.

Hasegawa, S., et al. (2008). "The protocadherin-α family is involved in axonal coalescence of olfactory sensory neurons into glomeruli of the olfactory bulb in mouse." Molecular and Cellular Neuroscience **38**(1): 66-79.

Hayden, S., et al. (2010). "Ecological adaptation determines functional mammalian olfactory subgenomes." Genome research **20**(1): 1-9.

Heaton, M. P., et al. (2016). "Using diverse US beef cattle genomes to identify missense mutations in EPAS1, a gene associated with pulmonary hypertension." F1000Research **5**.

Hedrick, P. W. (2013). "Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation." Molecular ecology **22**(18): 4606-4618.

Hill, R. S. and C. A. Walsh (2005). "Molecular insights into human brain evolution." nature **437**(7055): 64.

Howrigan, D. P., et al. (2011). "Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms." BMC genomics **12**(1): 1.

Hu, Y., et al. (2017). "Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas." Proceedings of the National Academy of Sciences **114**(5): 1081-1086.

Hu, Z.-L., et al. (2016). "Developmental progress and current status of the Animal QTLdb." Nucleic acids research **44**(D1): D827-D833.

Hu, Z.-L., et al. (2018). "Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB." Nucleic acids research **47**(D1): D701-D710.

Huang, D. W., et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature protocols **4**(1): 44.

Hubbard, T., et al. (2002). "The Ensembl genome database project." Nucleic acids research **30**(1): 38-41.

Hudson, R. R., et al. (1987). "A test of neutral molecular evolution based on nucleotide data." Genetics **116**(1): 153-159.

Huisman, J., et al. (2016). "Inbreeding depression across the lifespan in a wild mammal population." Proceedings of the National Academy of Sciences **113**(13): 3585-3590.

Hulpiau, P. and F. Van Roy (2009). "Molecular evolution of the cadherin superfamily." The international journal of biochemistry & cell biology **41**(2): 349-369.

Humble, E., et al. (2016). "A draft fur seal genome provides insights into factors affecting SNP validation and how to mitigate them." Molecular ecology resources **16**(4): 909-921.

Hwang, J. M., et al. (2009). "The Inbreeding Trend of Hanwoo Cow Population." Annals of Animal Resources Sciences **20**: 1-5.

Innan, H. and Y. Kim (2004). "Pattern of polymorphism after strong artificial selection in a domestication event." Proceedings of the National Academy of Sciences of the United States of America **101**(29): 10667-10672.

Ishikawa, K., et al. (2014). "A Japanese family showing high-frequency hearing loss with KCNQ4 and TECTA mutations." Acta oto-laryngologica **134**(6): 557-563.

Ismail, S. and M. Essawi (2012). "Genetic polymorphism studies in humans." Middle East Journal of Medical Genetics **1**(2): 57-63.

Jay, P., et al. (2018). "Supergene evolution triggered by the introgression of a chromosomal inversion." Current Biology **28**(11): 1839-1845. e1833.

Jefferson, T. A., et al. (1993). Marine mammals of the world, Food & Agriculture Org.

Jin, W., et al. (2012). "Genome-wide detection of natural selection in African Americans pre-and post-admixture." Genome research **22**(3): 519-527.

Jones, D. T., et al. (1992). "The rapid generation of mutation data matrices from protein sequences." Computer applications in the biosciences: CABIOS **8**(3): 275-282.

Jurka, J., et al. (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenetic and genome research **110**(1-4): 462-467.

Keane, T., et al. (2004). "ModelGenerator: amino acid and nucleotide substitution model selection." National University of Ireland, Maynooth, Ireland: 34.

Keane, T. M., et al. (2006). "Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified." BMC evolutionary biology **6**(1): 29.

Keller, M. C., et al. (2012). "Runs of homozygosity implicate autozygosity as a schizophrenia risk factor." PLoS genetics **8**(4): e1002656.

Keller, M. C., et al. (2012). "Runs of homozygosity implicate autozygosity as a schizophrenia risk factor." PLoS Genet **8**(4): e1002656.

Kelley, J. L. and W. J. Swanson (2008). "Positive selection in the human genome: from genome scans to biological significance." Annu. Rev. Genomics Hum. Genet. **9**: 143-160.

Khaitovich, P., et al. (2008). "Metabolic changes in schizophrenia and human brain evolution." Genome biology **9**(8): R124.

Kim, D., et al. (2015). "HISAT: a fast spliced aligner with low memory requirements." Nature methods **12**(4): 357.

Kim, E.-S., et al. (2015). "Recent artificial selection in US Jersey cattle impacts autozygosity levels of specific genomic regions." BMC genomics **16**(1): 1.

Kim, E.-S., et al. (2015). "The relationship between runs of homozygosity and inbreeding in Jersey cattle under selection." PLOS ONE **10**(7): e0129967.

Kim, J., et al. (2017). "The genome landscape of indigenous African cattle." Genome biology **18**(1): 34.

Kim, S.-J., et al. (2017). "Cattle genome-wide analysis reveals genetic signatures in trypanotolerant N'Dama." BMC genomics **18**(1): 371.

Kirin, M., et al. (2010). "Genomic runs of homozygosity record population history and consanguinity." PloS one **5**(11): e13996.

Kitamura, A., et al. (2014). "An inherited mutation in NLRC4 causes autoinflammation in human and mice." Journal of Experimental Medicine **211**(12): 2385-2396.

Kumar, S. and S. Subramanian (2002). "Mutation rates in mammalian genomes." Proceedings of the National Academy of Sciences **99**(2): 803-808.

Lander, E. S., et al. (2001). "Initial sequencing and analysis of the human genome." nature **409**(6822): 860-921.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4): 357.

Lao, O., et al. (2007). "Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms." Annals of human genetics **71**(3): 354-369.

Larkin, M. A., et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Lee, H.-J., et al. (2014). "Deciphering the genetic blueprint behind Holstein milk proteins and production." Genome biology and evolution **6**(6): 1366-1374.

Lee, K.-T., et al. (2013). "Whole-genome resequencing of Hanwoo (Korean cattle) and insight into regions of homozygosity." BMC genomics **14**(1): 519.

Lee, K., et al. (2013). "Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant." BMC genomics **14**(1): 596.

Lee, S.-H., et al. (2014). "Hanwoo cattle: origin, domestication, breeding strategies and genomic selection." Journal of animal science and technology **56**(1): 2.

Lee, S.-H., et al. (2014). "Hanwoo cattle: origin, domestication, breeding strategies and genomic selection." Journal of animal science and technology **56**(1): 1.

Lee, S., et al. (2011). "Linkage disequilibrium and effective population size in hanwoo korean cattle." Asian-Australasian journal of animal sciences **24**(12): 1660-1665.

Lee, S. H., et al. (2013). "Strategies to multiply elite cow in Hanwoo small farm." Journal of Embryo Transfer **28**(2): 79-85.

Lee, T.-H., et al. (2014). "SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data." BMC genomics **15**(1): 162.

Legendre, M., et al. (2007). "Sequence-based estimation of minisatellite and microsatellite repeat variability." Genome research **17**(12): 1787-1796.

Lemecha, H., et al. (2006). "Response of four indigenous cattle breeds to natural tsetse and trypanosomosis challenge in the Ghibe valley of Ethiopia." Veterinary parasitology **141**(1-2): 165-176.

Leroy, G. (2014). "Inbreeding depression in livestock species: review and meta-analysis." Animal genetics **45**(5): 618-628.

Letunic, I. and P. Bork (2019). "Interactive Tree Of Life (iTOL) v4: recent updates and new developments." Nucleic acids research.

Li, H., et al. (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, L., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome research **13**(9): 2178-2189.

Liu, W., et al. (2001). "Direct inhibition of Bruton's tyrosine kinase by IBtk, a Btk-binding protein." Nature immunology **2**(10): 939.

Liu, X., et al. (2009). "Disruption of striated preferentially expressed gene locus leads to dilated cardiomyopathy in mice." Circulation **119**(2): 261-268.

Loftus, T. M., et al. (2000). "Reduced food intake and body weight in mice treated with fatty acid synthase inhibitors." Science **288**(5475): 2379-2381.

Luo, R., et al. (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." Gigascience **1**(1): 18.

MacEachern, S., et al. (2009). "An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle." BMC genomics **10**(1): 181.

Majoros, W. H., et al. (2004). "TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders." Bioinformatics **20**(16): 2878-2879.

Manceau, M., et al. (2010). "Convergence in pigmentation at multiple levels: mutations, genes and function." Philosophical Transactions of the Royal Society B: Biological Sciences **365**(1552): 2439-2450.

Mandric, I. and A. Zelikovsky (2015). "ScaffMatch: scaffolding algorithm based on maximum weight matching." Bioinformatics **31**(16): 2632-2638.

Marchler-Bauer, A., et al. (2014). "CDD: NCBI's conserved domain database." Nucleic acids research **43**(D1): D222-D226.

Marçais, G. and C. Kingsford (2011). "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." Bioinformatics **27**(6): 764-770.

Marras, G., et al. (2015). "Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy." Animal genetics **46**(2): 110-121.

Martin, S. H., et al. (2013). "Genome-wide evidence for speciation with gene flow in Heliconius butterflies." Genome research **23**(11): 1817-1828.

Martin, S. H., et al. (2014). "Evaluating the use of ABBA–BABA statistics to locate introgressed loci." Molecular biology and evolution **32**(1): 244-257.

Martínez-Cadenas, C., et al. (2013). "Simultaneous purifying selection on the ancestral MC1R allele and positive selection on the melanoma-risk allele V60L in south Europeans." Molecular biology and evolution **30**(12): 2654-2665.

Mathelier, A., et al. (2015). "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles." Nucleic acids research **44**(D1): D110-D115.

Matoulkova, E., et al. (2012). "The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells." RNA Biol **9**(5): 563-576.

Mattioli, R. C., et al. (2000). "Immunogenetic influences on tick resistance in African cattle with particular reference to trypanotolerant N'Dama (Bos taurus) and trypanosusceptible Gobra zebu (Bos indicus) cattle." Acta Tropica **75**(3): 263-277.

Mbole-Kariuki, M. N., et al. (2014). "Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya." Heredity **113**(4): 297.

McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature **351**(6328): 652.

McKenna, A., et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research **20**(9): 1297-1303.

McMahon, M. K., et al. (2016). ADAMTS5: A critical enzyme for the maintenance of influenza-specific CD8+ T cell memory, Am Assoc Immnol.

McQuillan, R., et al. (2008). "Runs of homozygosity in European populations." The American Journal of Human Genetics **83**(3): 359-372.

Medugorac, I., et al. (2017). "Whole-genome analysis of introgressive hybridization and characterization of the bovine legacy of Mongolian yaks." Nature genetics **49**(3): 470.

Meyer, N. C., et al. (2007). "Identification of three novel TECTA mutations in Iranian families with autosomal recessive nonsyndromic hearing impairment at the DFNB21 locus." American Journal of Medical Genetics Part A **143**(14): 1623-1629.

Miao, B., et al. (2016). "Genomic analysis reveals hypoxia adaptation in the Tibetan mastiff by introgression of the gray wolf from the Tibetan plateau." Molecular biology and evolution **34**(3): 734-743.

Michalski, N. and C. Petit (2015). "Genetics of auditory mechano-electrical transduction." Pflügers Archiv-European Journal of Physiology **467**(1): 49-72.

Mill, J., et al. (2002). "Expression of the dopamine transporter gene is regulated by the 3′ UTR VNTR: Evidence from brain and lymphocytes using quantitative RT-PCR." American Journal of Medical Genetics Part A **114**(8): 975-979.

Moon, S., et al. (2015). "A genome-wide scan for signatures of directional selection in domesticated pigs." BMC genomics **16**(1): 130.

Mora-Bermúdez, F., et al. (2016). "Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development." Elife **5**: e18683.

Moreno-Hagelsieb, G. and K. Latimer (2008). "Choosing BLAST options for better detection of orthologs as reciprocal best hits." Bioinformatics **24**(3): 319-324.

Moteki, H., et al. (2012). "TECTA mutations in Japanese with mid-frequency hearing loss affected by zona pellucida domain protein secretion." Journal of human genetics **57**(9): 587-592.

Murray, M., et al. (1984). "Genetic resistance to African Trypanosomiasis." The Journal of infectious diseases **149**(3): 311-319.

Mwai, O., et al. (2015). "African indigenous cattle: unique genetic resources in a rapidly changing world." Asian-Australasian journal of animal sciences **28**(7): 911.

Naessens, J., et al. (2002). "Identification of mechanisms of natural resistance to African trypanosomiasis in cattle." Veterinary Immunology and Immunopathology **87**(3-4): 187-194.

Nery, M. F., et al. (2016). "Selection on different genes with equivalent functions: the convergence story told by Hox genes along the evolution of aquatic mammalian lineages." BMC evolutionary biology **16**(1): 113.

Nielsen, R. (2005). "Molecular signatures of natural selection." Annu. Rev. Genet. **39**: 197-218.

Niimura, Y. and M. Nei (2005). "Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods." Proceedings of the National Academy of Sciences **102**(17): 6039-6044.

Niimura, Y. and M. Nei (2007). "Extensive gains and losses of olfactory receptor genes in mammalian evolution." PloS one **2**(8): e708.

Noyes, H., et al. (2011). "Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection." Proceedings of the National Academy of Sciences **108**(22): 9304-9309.

O'bleness, M., et al. (2012). "Evolution of genetic and genomic features unique to the human lineage." Nature Reviews Genetics **13**(12): 853.

O’Bleness, M. S., et al. (2012). "Evolutionary history and genome organization of DUF1220 protein domains." G3: Genes, Genomes, Genetics **2**(9): 977-986.

O’Connell, J., et al. (2015). "NxTrim: optimized trimming of Illumina mate pair reads." Bioinformatics **31**(12): 2035-2037.

Pai, A. A., et al. (2017). "Intron Length and Recursive Sites are Major Determinants of Splicing Efficiency in Flies." BioRxiv: 107995.

Park, B., et al. (2013). "National genetic evaluation (system) of Hanwoo (Korean native cattle)." Asian-Australasian journal of animal sciences **26**(2): 151.

Park YI, L. J., Cho YY (1969). "Effect of inbreeding on birth and weaning weights in Korean Native Cattle." Korean J Anim Sci **11**: 36-39.

Parker, J., et al. (2013). "Genome-wide signatures of convergent evolution in echolocating mammals." nature **502**(7470).

Parker, J., et al. (2013). "Genome-wide signatures of convergent evolution in echolocating mammals." Nature **502**(7470): 228.

Peng, Y., et al. (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." Bioinformatics **28**(11): 1420-1428.

Pickrell, J. K. and J. K. Pritchard (2012). "Inference of population splits and mixtures from genome-wide allele frequency data." PLoS genetics **8**(11): e1002967.

Pitt, D., et al. (2019). "Domestication of cattle: Two or three events?" Evolutionary applications **12**(1): 123-136.

Platoshyn, O., et al. (2006). "Acute hypoxia selectively inhibits KCNA5 channels in pulmonary artery smooth muscle cells." American Journal of Physiology-Cell Physiology **290**(3): C907-C916.

Pomerantz, J. L., et al. (2002). "CARD11 mediates factor-specific activation of NF-κB by the T cell receptor complex." The EMBO journal **21**(19): 5184-5194.

Popesco, M. C., et al. (2006). "Human lineage–specific amplification, selection, and neuronal expression of DUF1220 domains." Science **313**(5791): 1304-1307.

Porto-Neto, L., et al. (2014). "Genome-wide detection of signatures of selection in Korean Hanwoo cattle." Animal genetics **45**(2): 180-190.

Porto-Neto, L., et al. (2014). "Genome-wide detection of signatures of selection in Korean Hanwoo cattle." Animal genetics **45**(2): 180-190.

Price, A. L., et al. (2005). "De novo identification of repeat families in large genomes." Bioinformatics **21**(suppl_1): i351-i358.

Pryce, J. E., et al. (2014). "Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle." Genetics Selection Evolution **46**(1): 1.

Purcell, S., et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American Journal of Human Genetics **81**(3): 559-575.

Quilez, J., et al. (2016). "Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans." Nucleic acids research **44**(8): 3750-3762.

Racimo, F., et al. (2015). "Evidence for archaic adaptive introgression in humans." Nature Reviews Genetics **16**(6): 359.

Rege, J., et al. (2006). DAGRIS (Domestic Animal Genetic Resources Information System). International Livestock Research Institute, Addis Ababa, Ethiopia.

Reichmuth, C., et al. (2013). "Comparative assessment of amphibious hearing in pinnipeds." Journal of Comparative Physiology A **199**(6): 491-507.

Reznick, D. N. and R. E. Ricklefs (2009). "Darwin's bridge between microevolution and macroevolution." Nature **457**(7231): 837.

Riedman, M. (1990). The pinnipeds: seals, sea lions, and walruses, Univ of California Press.

Robertson, A. (1961). "Inbreeding in artificial selection programmes." Genetics Research **2**(2): 189-194.

Rybczynski, N., et al. (2009). "A semi-aquatic Arctic mammalian carnivore from the Miocene epoch and origin of Pinnipedia." nature **458(**7241): 1021.

Sabeti, P. C., et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832.

Sabeti, P. C., et al. (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **449**(7164): 913.

Safran, M., et al. (2010). "GeneCards Version 3: the human gene integrator." Database **2010**.

Sanchez, L., et al. (1999). "Improving the efficiency of artificial selection: more selection pressure with less inbreeding." Genetics **151**(3): 1103-1114.

Saura, M., et al. (2015). "Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data." Genetics Selection Evolution **47**(1): 1.

Scheet, P. and M. Stephens (2006). "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase." The American Journal of Human Genetics **78**(4): 629-644.

Scheinfeldt, L. B., et al. (2012). "Genetic adaptation to high altitude in the Ethiopian highlands." Genome biology **13**(1): R1.

Schneider, H. K. (1964). "A model of African indigenous economy and society." Comparative Studies in Society and History **7**(1): 37-55.

Sen, S. K., et al. (2006). "Human genomic deletions mediated by recombination between Alu elements." The American Journal of Human Genetics **79**(1): 41-53.

Sharp, A. J., et al. (2006). "Structural variation of the human genome." Annu. Rev. Genomics Hum. Genet. **7**: 407-442.

Sherry, S. T., et al. (2001). "dbSNP: the NCBI database of genetic variation." Nucleic acids research **29**(1): 308-311.

Shin, D.-H., et al. (2014). "Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level." BMC genomics **15**(1): 240.

Simão, F. A., et al. (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." Bioinformatics **31**(19): 3210-3212.

Slade, R. W., et al. (1994). "Multiple nuclear-gene phylogenies: application to pinnipeds and comparison with a mitochondrial DNA gene phylogeny." Molecular biology and evolution **11**(3): 341-356.

Slater, G. S. C. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison." BMC bioinformatics **6**(1): 31.

Smit, A. and R. Hubley (2010). "RepeatModeler Open-1.0." Repeat Masker Website.

Sonay, T. B., et al. (2015). "Tandem repeat variation in human and great ape populations and its impact on gene expression divergence." Genome research **25**(11): 1591-1599.

Sousa, A. M., et al. (2017). "Evolution of the human nervous system function, structure, and development." Cell **170**(2): 226-247.

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics **22**(21): 2688-2690.

Stambas, J., et al. (2017). ADAMTS5 and its substrate versican play a critical role in influenza virus immunity, Am Assoc Immnol.

Stanke, M., et al. (2006). "AUGUSTUS: ab initio prediction of alternative transcripts." Nucleic acids research **34**(suppl_2): W435-W439.

Stark, R. and G. Brown (2011). "DiffBind: differential binding analysis of ChIP-Seq peak data." R package version **100**: 4.3.

Streelman, J. T. and T. D. Kocher (2002). "Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia." Physiological genomics **9**(1): 1-4.

Sturm, R. A. and D. L. Duffy (2012). "Human pigmentation genes under environmental selection." Genome biology **13**(9): 248.

Suarez-Gonzalez, A., et al. (2018). "Adaptive introgression: a plant perspective." Biology Letters **14**(3): 20170688.

Suyama, M., et al. (2006). "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments." Nucleic acids research **34**(suppl_2): W609-W612.

Suzuki, S., et al. (2018). "Novel brain-expressed noncoding RNA, HSTR1, identified at a human-specific variable number tandem repeat locus with a human accelerated region." Biochemical and biophysical research communications **503**(3): 1478-1483.

Syvänen, A.-C. (2001). "Accessing genetic variation: genotyping single nucleotide polymorphisms." Nature Reviews Genetics **2**(12): 930.

Szpiech, Z. A. and R. D. Hernandez (2014). "Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection." Molecular biology and evolution: msu211.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-595.

Tan, G. and B. Lenhard (2016). "TFBSTools: an R/bioconductor package for transcription factor binding site analysis." Bioinformatics **32**(10): 1555-1556.

Tarailo-Graovac, M. and N. Chen (2009). "Using RepeatMasker to identify repetitive elements in genomic sequences." Current protocols in bioinformatics: 4.10. 11-14.10. 14.

Taylor, J. F., et al. (2016). "Lessons for livestock genomics from genome and transcriptome sequencing in cattle and other mammals." Genetics Selection Evolution **48**(1): 59.

Terefe, E., et al. (2015). "Phenotypic characteristics and trypanosome prevalence of Mursi cattle breed in the Bodi and Mursi districts of South Omo Zone, southwest Ethiopia." Tropical animal health and production **47**(3): 485-493.

Toro, M. and M. Perez-Enciso (1990). "Optimization of selection response under restricted inbreeding." Genetics Selection Evolution **22**(1): 1.

Tyner, C., et al. (2016). "The UCSC Genome Browser database: 2017 update." Nucleic acids research **45**(D1): D626-D634.

Udpa, N., et al. (2014). "Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes." Genome biology **15**(2): R36.

Usdin, K. (2008). "The biological effects of simple tandem repeats: lessons from the repeat expansion diseases." Genome Res **18**(7): 1011-1019.

Usdin, K. (2008). "The biological effects of simple tandem repeats: lessons from the repeat expansion diseases." Genome research **18**(7): 1011-1019.

Vergnaud, G. and F. Denoeud (2000). "Minisatellites: mutability and genome architecture." Genome research **10**(7): 899-907.

Verhoeven, K., et al. (1998). "Mutations in the human α-tectorin gene cause autosomal dominant non-syndromic hearing impairment." Nature genetics **19**(1): 60-62.

Verhoeven, K. J., et al. (2010). "Population admixture, biological invasions and the balance between local adaptation and inbreeding depression." Proceedings of the Royal Society B: Biological Sciences **278**(1702): 2-8.

Voight, B. F., et al. (2006). "A map of recent positive selection in the human genome." PLoS biology **4**(3): e72.

Wakil, S. J. (1989). "Fatty acid synthase, a proficient multifunctional enzyme." Biochemistry **28**(11): 4523-4530.

Walker, F. O. (2007). "Huntington's disease." The Lancet **369**(9557): 218-228.

Wang, X., et al. (2002). "Gamma protocadherins are required for survival of spinal interneurons." Neuron **36**(5): 843-854.

Warpeha, K., et al. (1999). "Genotyping and functional analysis of a polymorphic (CCTTT) n repeat of NOS2A in diabetic retinopathy." The FASEB journal **13**(13): 1825-1832.

Wartzok, D. and D. R. Ketten (1999). "Marine mammal sensory systems." Biology of marine mammals **1**: 117.

Weir, B. S. and C. C. Cockerham (1984). "Estimating F-statistics for the analysis of population structure." evolution **38**(6): 1358-1370.

Wieringa, B., et al. (1984). "A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron." Cell **37**(3): 915-925.

Wright, S. I., et al. (2005). "The effects of artificial selection on the maize genome." Science **308**(5726): 1310-1314.

Wu, D.-D., et al. (2018). "Pervasive introgression facilitated domestication and adaptation in the Bos species complex." Nat Ecol Evol **2**(7): 1139-1145.

Yagi, T. (2008). "Clustered protocadherin family." Development, growth & differentiation **50**(s1).

Yang, J., et al. (2011). "GCTA: a tool for genome-wide complex trait analysis." The American Journal of Human Genetics **88**(1): 76-82.

Yang, T.-L., et al. (2010). "Runs of homozygosity identify a recessive locus 12q21. 31 for human adult height." The Journal of Clinical Endocrinology & Metabolism **95**(8): 3777-3782.

Yang, Z. (1998). "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution." Molecular biology and evolution **15**(5): 568-573.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Molecular biology and evolution **24**(8): 1586-1591.

Yang, Z., et al. (2005). "Bayes empirical Bayes inference of amino acid sites under positive selection." Molecular biology and evolution **22**(4): 1107-1118.

Yi, X., et al. (2010). "Sequencing of 50 human exomes reveals adaptation to high altitude." Science **329**(5987): 75-78.

Zdobnov, E. M. and R. Apweiler (2001). "InterProScan–an integration platform for the signature-recognition methods in InterPro." Bioinformatics **17**(9): 847-848.

Zerbino, D. R., et al. (2017). "Ensembl 2018." Nucleic acids research **46**(D1): D754-D761.

Zhang, G., et al. (2014). "Comparative genomics reveals insights into avian genome evolution and adaptation." Science **346**(6215): 1311-1320.

Zhang, J., et al. (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." Molecular biology and evolution **22**(12): 2472-2479.

Zhang, J., et al. (2014). "Introgression genetics and breeding between Upland and Pima cotton: a review." Euphytica **198**(1): 1-12.

Zhao, D., et al. (2012). "Variants in the SRD5A2 gene are associated with quality of semen." Molecular medicine reports **6**(3): 639-644.

Zhao, Y., et al. (2011). "The NLRC4 inflammasome receptors for bacterial flagellin and type III secretion apparatus." Nature **477**(7366): 596.

Zhou, X., et al. (2015). "Convergent evolution of marine mammals is associated with distinct substitutions in common genes." Scientific reports **5**.

Zimmer, F. and S. H. Montgomery (2015). "Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion." Genome biology and evolution **7**(8): 2083-2088.

요약(국문초록)

포유류 유전체 내 선택압에 의한 적응 흔적 및 특성 발굴

김권도

협동과정 생물정보학

서울대학교 대학원 자연과학대학

진화 생물학의 핵심 목표는 진화 과정과 적응 형질의 유전적 기초를 이해하는 것이다. 이와 관련하여 최근 시퀀싱 기술의 진보와 서열 데이터의 폭발적 증가는 이러한 목표를 달성 할 수 있는 더 좋은 기회를 제공하고 있다. 실제로 시퀀싱 기술의 발달로 대규모 시료에 대하여 좀 더 쉽고 정확하게 다양한 유전변이를 얻을 수 있게 되었으며, 이 덕분에 일반적인 유전체 연구의 범위가 확장되었고 다양한 진화 과정을 고려할 수도 있게 되었다. 이에 이 논문의 목적은 다양한 진화과정 하에서 여러 유전변이의 유용성을 보여주고, 전장 유전체 및 비교 유전체 분석을 통해 포유류 적응 형질의 유전적 배경을 밝히는 것이다.

이 논문은 세 가지 포유 동물 종(기각류, 영장류, 소)의 유전체 분석 결과를 포함한 5 개의 장으로 구성되어있다. 제 1 장에서는 이 논문과 관련된 배경 지식과 최근의 연구 사례를 소개하고 있다.

전반부 (제 2, 3 장)는 종간 비교분석에 중점을 두었고, 후반부(제 4, 5 장)는 종내의 다형성에 초점을 두고 있다.

기각류는 반 수생 환경에 적응한 특징적인 해양 동물이다. 그러나 그 유전체는 특성이 잘 알려져 있지 않다. 제 2 장에서는 아미노산 치환 정보를 이용하여 기각류의 진화 및 적응 흔적을 조사하였다. 구체적으로 기각류 3 종의 새로운 유전체를 이용하여 기각류의 생활환경과 관련된 양성선택 유전자 및 아미노산 치환 흔적을 발굴하였다. 특히 *TECTA* 유전자 내의 고유한 아미노산 치환 흔적은 기각류의 청각과 밀접한 관련이 있을 것으로 예상된다. 또한, 이전 연구결과와 같이 해양 포유류에서 표현형의 수렴진화와 직접적으로 연관되어 있는 서열 수렴은 흔하지 않다는 것을 확인하였다. 예를 들어, *FASN*, *KCNA5* 및 *IL17RA* 는 기각류에 특이적인 아미노산 치환을 포함하지만 모든 해양 포유 동물에서 공통적으로 표현형 수렴진화 (두꺼운 지방조직, 저산소 적응 및 병원체에 대한 면역 반응)가 일어났을 것으로 예상된다. 이러한 연구 결과들은 해양 포유류의 수렴 진화 특성에 대한 지식을 제공함과 동시에 유전자 기능 연구에 대한 후보 표적을 제공할 것으로 기대된다.

인간은 현존하는 영장류 중에서 가장 큰 두뇌를 가지고 있다. 그러나 인간의 뇌가 어떻게 영장류 중에서 특히 빠르게 진화했는지는 완전히 밝혀지지 않았다. 제 3 장에서는 인간 두뇌의 급속한 진화에 대한 가설을 찾기 위해 유전체 분석을 수행하였다. 반복서열이 그 불안정한 성질 때문에 급속한 유전적 변이를

일으키는 데 핵심적인 역할을 할 수 있다는 가설에 근거하여, 유전체 비교분석에서 인간 특이적인 152 개의 반복서열을 검출하였다. 특이하게도, 이러한 반복서열들은 뇌 발달 및 시냅스 기능과 관련이 있었으며, 뇌 조직에서 해당 반복서열과 관련된 유전자의 발현 수준은 다른 영장류보다 인간에서 유의하게 높았다. 이러한 결과는 반복서열이 인간 두뇌의 급속한 진화에 기여하였을 수도 있다는 하나의 가능성을 제시한다.

소의 유전적 역사는 복잡하지만 가축화 및 환경 적응과 같은 포유 동물의 진화 과정을 이해할 수 있는 풍부한 정보를 담고있다. 제 4 장에서는 한국 토종 소품종인 한우의 유전체 선발이 한우 집단에 미친 유전적 영향을 조사 하였다. Runs of humozygosity 를 이용하여, 최근에 일어난 근친 교배의 증가를 보여 주었고 동시에, 근교약세가 체중에 영향을 미칠 만큼 크지 않았다는 것을 유전정보를 통해 보여주었다. 제 5 장에서는 소의 두 아종 (*Bos taurus*, *Bos indicus*)사이의 유전적 혼합을 아프리카 토착 소의 단일 염기 다형성 자료를 통해 분석하였다. 이를 통해 아프리카 소의 환경에 대한 빠른 적응의 원인은 유전적 혼합에 있다는 여러 증거를 제시하였다.

이 논문은 다양한 진화과정 하에서 다양한 유전변이의 적용사례를 보여주고, 또한 이를 통해 포유 동물의 다양한 진화 과정을 이해하는 데에 기여할 수 있을 것으로 기대된다.

주요어: 유전변이, 진화, 적응, 양성선택, 유전자 이입, 포유류

학번: 2016−30124