GENETIC DIVERSITY AND POPULATION STRUCTURE OF THE ARABIAN HORSE POPULATIONS FROM SYRIA AND OTHER COUNTRIES

A Dissertation

by

ANAS MAHMOUD KHANSHOUR

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ernest Gus Cothran
Co-Chair of Committee,	Terje Raudsepp
Committee Members,	Jane Welsh
	James Derr
Head of Department,	Evelyn Tiffany-Castiglioni

August 2013

Major Subject: Biomedical Sciences

ABSTRACT

Humans and horses weaved together wonderful stories of adventure and generosity. As a part of human history and civilization, Arabian horses ignite imagination throughout the world. Populations of this breed exist in many countries. Here I explored different populations of Arabians representing Middle Eastern and Western populations. The main two aims of this study were to provide the genetic diversity description of Arabians from different origins and to examine the traditional classification system of the breed. A third aim was to tackle the distribution pattern of the genetic variability within the genome to show whether there are differences in relative variability of different types of markers.

First, I analyzed the genetic structure of 537Arabian horses from seven populations by using microsatellites. The results consistently showed higher levels of diversity within the Middle Eastern populations compared to the Western populations. All American-Arabians showed differentiation from Middle Eastern populations.

Second, I sequenced the whole mtDNA D-loop of 251 Arabian horses. The whole D-loop sequence was more informative than using just the HVR1. Native populations from the Middle East, such as Syrian, represented a hot spot of genetic diversity. Most importantly, there was no evidence that the Arabian horse breed has clear subdivisions depending on the traditional maternal based strain classification system.

Third, I tested the heterozygosity distribution pattern along the genome of 22 Peruvian Paso horses using 232 microsatellites and Single Nucleotide Polymorphisms (SNPs). The pattern of genetic diversity was completely different between these two markers where no correlation was found. Runs of homozygosity test of SNPs and associated microsatellites noticeably showed that all of associated microsatellites loci were homozygous in the matched case.

The findings of this study will help in understanding the evolutionary history and developing breeding and conservation programs of horses. This study provided databases including parentage testing system and maternal lineages that will help to recover the Syrian Arabian population after the armed conflict started in Syria in 2011. The results here can be applied not only to horses, but also to other animal species with similar criteria.

DEDICATION

To my father’s soul.

To my mother, my wife, my sisters, my brothers and friends.

To my best friend Dr. Nabeel Salam and his family.

To the great people in Syria who are standing up against oppression of tyrants.

ACKNOWLEDGEMENTS

Great thanks to my advisor Dr. Gus Cothran who is one of the great men in my professional and personal life. He stood by me every moment as leader, friend and big brother.

I would like to thank my committee co-chair and members, Dr. Raudsepp, Dr. Welsh and Dr. Derr for their great contributions and support during my dissertation preparation. I am so thankful to the Islamic Development Bank who supported my study by the IDB Merit Scholarship Programme.

Many thanks to our lab team Dr. Rytis, Dr. Eleanore, Marian and Smriti for their great support.

I would like to thank all the Syrian horse breeders who provided us their horse samples and everyone who helped to collect them, Edie Booth for providing Arabian samples with their pedigree information, Dr. G. Cholewinski who provided the Polish Arabian samples, F. Maloufi who provided the Persian Arabian samples and other who contributed samples.

I wish I can reach my father's hands to kiss them and tell him that I did it Haji! No words in this world can express my thankfulness that I owe to my father, mother, sisters and brothers.

Finally, I owe great big thanks to my soulmate, my wife Heba.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Since more than 6000 years ago, humans and horses weaved together wonderful stories of adventure, bravery and generosity. Horses were part of human history and civilization and participated in key events in human activities throughout the historical times. Horses attained a prominent role in transportation, welfare and warfare in human life. The ancient horse paintings on cave walls and the fascinating description in legends and poetries from the early and modern history of humans has shown horses as the best partners of human civilization.

1.2 The Arabian horse breed

Al-Asseel (purebred), *Al-Jwad* (generous), the horse of the desert, the drinker of the wind, the runner, these are some of the descriptions of the Arabian horse. The Arabian horse breed is one of the oldest and most influential breed throughout the world (Głażewska, 2010). It has been involved in the foundation of many other breeds such as the Thoroughbred (Bowling and Ruvinsky, 2000) and the Lipizzan (Zechner*, et al.*, 2002). From the historical point of view, the traditional breeders (*Bedouins*) have maintained the purity of the Arabian by avoiding any cross-breeding not only between Arabians and non-Arabians, but also by maintaining strictly separated strains (Pruski, 1983). This system of breed conservation has led to the formation of native populations which can be described as old populations at a candidate place of origin of the breed like the desert of the Middle East and the Arabian Peninsula. Consequently, regions like Syria and Saudi Arabia may have exceptional diversity within the breed, and may represent the original status of Arabian horses. On the other hand, Western Arabian populations, like the Polish Arabian, Shagya Arabian and American Arabian, were created in Europe and the USA using stocks originally imported from Middle Eastern Arabian populations from sources such as Syria and the Arabian peninsula no longer than 200 years ago (Bowling*, et al.*, 2000; Głażewska, 2010).

1.2.1 The Arabian horse populations in Syria

Officially, the Syrian horses are divided into two populations; the registered Syrian and the non-registered Syrian. The registered population consists of seven strains '*RASANs*' depending on their dam line: *Hadbaa, Hamadania, Dahmaa, Kahlila, Abian, Saklawia, and Muanakii*. These strains are considered as "pure" Arabian and expected to be completely separated from the non-registered horses and any other horse breeds. The non-registered horses, also are known as local horses, had not been considered as a pure Arabian. Neither of these populations have had in-depth efforts to discover their structure and diversity status. During the 1980s, the Ministry of Agriculture and Agrarian Reform in Syria started to determine the Syrian horse's lineage, and in 1989 they published the first studbook which contained 569 horses. After that, the Horses Office was created in order to register any new offspring. According to the data from the Horses Office in Syria in 2006, the total numbers of pure registered Arabian horses was 2574. Table 1 shows the number of horses in each of Syrian strains according to the data from the Horses Office in Syria 2006.

Table 1: The number of horses in each Syrian strains in 2006. (Arabian Horse Bureau 2006).

Hadbaa	*Hamadania*	*Dahmaa*	*Kahlila*	*Saklawia*	*Abian*	*Muanakii*	No-registered
26	263	6	1164	701	236	177	N/A

The horse breed in Syria faces difficult challenges such as the low total number of registered horses, an extinction and depletion danger especially in some strains (*Dahmaa and Hadbaa*), the traditional way of characterization, and the limited use of molecular biology methodologies in the identification and fingerprinting of local horses. Figure 1 shows some Syrian Arabian horses.

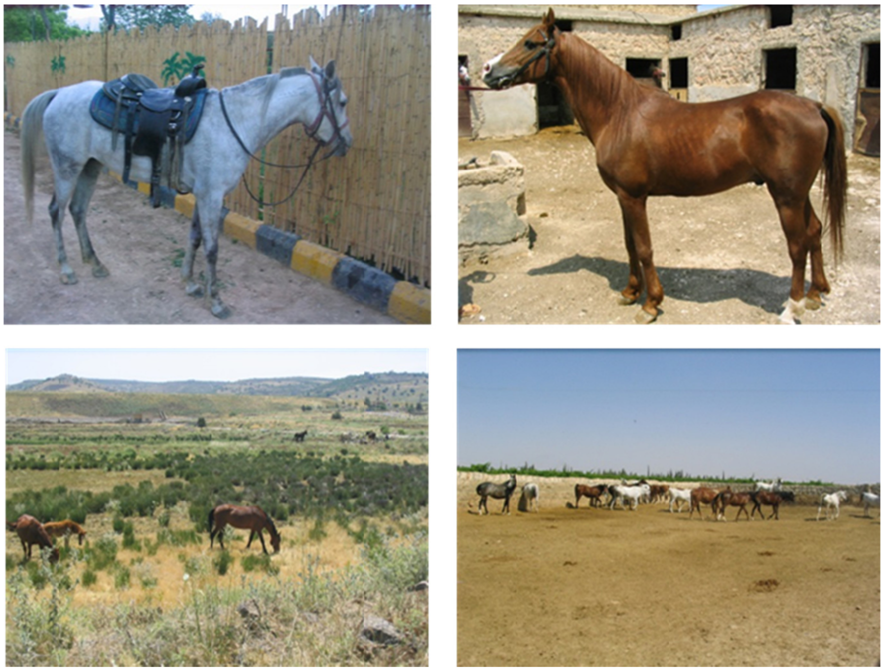


Figure 1: Syrian Arabian horses. These photos were taken by the author in 2009.

Furthermore, an armed conflict started in Syria in 2011. Horses under such circumstances, as well as all other animals, might be affected and Syrian horses might face serious negative consequences.

There are other Arabian horse populations in the Middle East such as the Saudi Arabian and the Iranian Arabian populations. These populations also consist of different registered strains, and all are considered as “pure” Arabians. As all were born and bred at the Middle East, Arabian Peninsula and Iran, respectively, they represent the Middle Eastern Arabian horse populations.

1.2.2 The Western-Arabian horse populations

The Western Arabian horse populations are represented by the American Arabian, Shagya Arabian and Polish Arabian. The first studbook of the American Arabian was established in 1908 and primarily consisted of mares that were exported from the Middle East in the mid to late 19th century (Bowling*, et al.*, 2000). The American Arabian includes horses originally from Egypt (USA-Egyptian) and horses originally from Saudi Arabia (USA-Saudi). These populations were mainly bred as separate breeds; there also was a cross-breed group (Egyptian-Saudi). In addition there is a group of horses known as the Davenport line that has been maintained as a closed population. The Shagya breed was originally developed in Hungary over 200 years ago, and its name came from the stallion Shagya which was probably imported from Syria 1836. The Polish Arabian breed was established in Poland in 1778 (Pruski, 1983), but most of its important studs were nearly destroyed during the World Wars I and II (Głażewska and Jezierski, 2004). The Polish studs were reconstructed after each war.

1.3 Genetic diversity and population structure

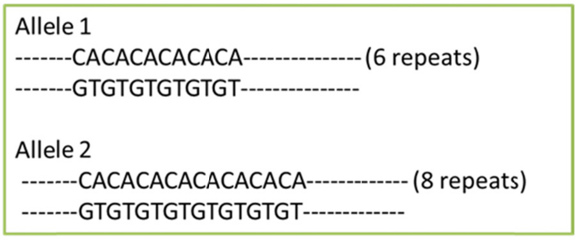
Comprehensive information about genetic diversity and population structure is highly important to draw the essential outlines for any appropriate conservation and sustainable management programs (Notter, 1999). Reducing the loss of genetic diversity is the main priority in management decisions (Weitzman, 1993). Concurrently, a high genetic diversity may indicate a genetic diversity hot spot which has been suggested as a tool for targeting conservation efforts of livestock spices (Bruford*, et al.*, 2003; Freeman*, et al.*, 2006). Also, genetic diversity and population structure studies are essential to understand the evolution, domestication and demographic history of populations as well as to support breeding programs and genome-wide association studies in plants and animals. Studying genetic diversity of any animal population will benefit not only these groups or species of animals, but results can be generalized and applied to other species and also can be used to achieve genetic improvements and medical discoveries in animals and humans. Horses have been successfully used in genetics and biomedical studies as model animals for many purposes (McIlwraith*, et al.*, 2010; Peffers*, et al.*, 2010).

1.4 Molecular biology for genetic diversity and population structure

Molecular markers have been used widely in genetics and biomedical studies and they have contributed very successfully in many discoveries and achievements in these fields (Vignal*, et al.*, 2002). In population genetics and conservation studies, the choice of molecular markers can be argued from two points of view. Biologists need simple and low cost genotyping procedures to generate as much data as possible. On the other hand, statisticians concerned about important characteristics such as information content, independent markers, neutrality, and sampling procedures. The population geneticist should fully understand the history and the nature of populations of interest in order to choose the best markers that may give an accurate interpretation of the results. There are many types of molecular markers such as allozymes, Restriction Fragment Length Polymorphisms (RFLP), Random Amplified Polymorphic DNA (RAPD), Microsatellites (STRs), Mitochondrial DNA (mtDNA) and Single Nucleotide Polymorphisms (SNPs). However, only STRs, mtDNA and SNPs will be reviewed in details in this study as the others are seldom used now.

1.4.1 Microsatellites

During the last fifteen years, DNA-based methodologies for genetic testing using polymerase chain reaction (PCR) technology, particularly microsatellites analysis, provide an obvious alternative to blood typing and become the basis of genetic analysis of populations. Microsatellites, or Short Tandem Repeats (STRs), are segments of repeated DNA with a short repeat unit, usually 1-6 nucleotides. Figure 2 shows an example of a dinucleotide repeat with two alleles**.**



```
Allele 1
-------CACACACACACA---------------(6 repeats)
-------GTGTGTGTGTGT---------------

Allele 2
 -------CACACACACACACACA-------------(8 repeats)
 -------GTGTGTGTGTGTGTGT-------------
```

Figure 2: An example of two alleles with two different numbers of repeats.

STR genotyping is based upon size determination of the amplified fragment containing a microsatellite and flanking regions. Figure 3 illustrate an example of STR genotyping in three animals A, B and AxB carrying allele1, allele 2 and both alleles, respectively.

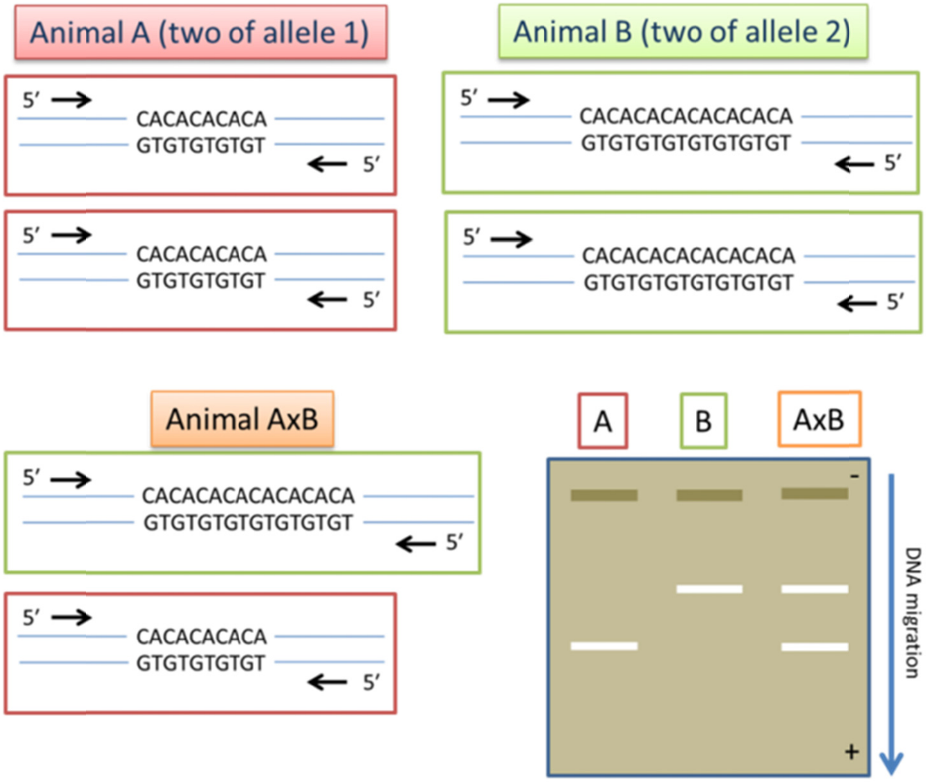


Figure 3: An illustration of STRs genotyping in three animals. A: homozygous for allele 1, B: homozygous for allele 2 and AxB: homozygous for alleles 1 and 2.

Stability, ease and accuracy of genotyping these co-dominant markers, together with their wide-spread distribution in the genome make microsatellite loci an attractive source of information for genetic diversity (Goldstein and Schlötterer, 1999). Usually, microsatellites have high mutation rates (on average 5 x 10^{-4}) that generate high levels of allelic diversity necessary for genetic studies (Schlotterer*, et al.*, 2004). In addition, STR typing can be automated with the ability of running multiplex amplification of several markers in a single PCR. Therefore, these selectively neutral genetic markers that follow Mendelian inheritance are extremely useful for the analysis of population structure for different species (Bruno-de-Sousa*, et al.*, 2011; Selkoe and Toonen, 2006), and can provide an indication of the levels of inter and intra-breed variability (Luís*, et al.*, 2007).

Microsatellite markers have been widely used to investigate genetic structure, population diversity estimation, individual genetic identification and pedigree analyses of different horse breeds (Achmann*, et al.*, 2004; Bigi and Perrotta, 2012; Koban*, et al.*, 2011; Luís*, et al.*, 2007; Prystupa*, et al.*, 2012b; Sereno*, et al.*, 2008). Also they have been successfully used in the analysis of small populations of closely bred animals (Kang*, et al.*, 2009). Many studies reported that microsatellite markers are useful in studying population structure and differentiation analysis better than allozymes (Barker*, et al.*, 1997; Estoup*, et al.*, 1998). Despite the fact that microsatellites are the most common markers for ecological and demographical applications with huge advantages, they have a few drawbacks. The main issues with microsatellite are: 1-Species-specific marker isolation: where specific primers are needed and a given primer often does not work across broad taxonomic groups, so primers design is usually required for each species (Glenn and Schable, 2005). 2- Problems with PCR amplification: that will cause the presence of null alleles (Paetkau and Strobeck, 1995). 3- Hidden allelic diversity: alleles of the identical size may have different evolutionary history, or in other words, the two alleles are identical in state but not in descent, a phenomenon known as microsatellite homoplasy. Homoplasy reduces the visible allelic diversity of populations and may inflate estimates of gene flow when mutation rate is high (Epperson, 2005; Rousset, 1996). Many of these challenges can be avoided by careful literature mining for previously tested loci for the organism of interest. The Food and Agriculture Organization (FAO) of the United Nations has published with the International Society for Animal Genetics (ISAG) a list of recommended microsatellite markers for different farm animal species including horses (FAO, 2011). Also, understanding the steps needed to evaluate the quality of a genetic data set is very important (Selkoe and Toonen, 2006) such as the testing of Hardy–Weinberg Equilibrium, Mendelian inheritance, linkage disequilibrium and the presence of null alleles. Then applying the suitable data analysis that may fit with the biological expectations and assumptions related to a study.

1.4.2 Mitochondrial DNA (mtDNA)

Since the mitochondrial genome discovery (Nass and Nass, 1963), its tiny fraction of organismal genome size has been one the most attractive subjects in animals and plants. Mitochondria are of major evolutionary and functional significance because they have their own small DNA genome (Chen and Butow, 2005) and accommodate some of the most critical functions of life (Chen*, et al.*, 2005). Mitochondria provide most of the cell's energy by oxidative phosphorylation that produces Adenosine Triphosphate ATP (Chen*, et al.*, 2005).

Therefore, mitochondria play a central role in metabolism and disease (Brand, 1997; Graeber and Muller, 1998). While the mitochondrial gene content is strongly conserved across animals (Boore, 1999; Gissi*, et al.*, 2008), the D-loop region is highly variable because of the elevated mutation rate in this region (Galtier*, et al.*, 2009). These unique structural characteristics combined with strictly maternal inheritance and lack of recombination, make mtDNA one of the most exploited markers in phylogenetic and genetic diversity studies (Moritz*, et al.*, 1987).

Horse was the 9th eutherian species after human, mouse, cow, rat, fin whale, harbor seal, blue whale and grey seal with the complete mtDNA sequenced in 1994 (Xu and Arnason, 1994). It appeared that mtDNA in horses was similar to other eutherian mtDNA with a total length of 16660 bp containing 13 peptide-coding genes, two rRNAs of the mitochondrial ribosome, 22 tRNAs and the control region or displacement loop (D-loop) region which is a highly variable region. The D-loop in horses contains two highly variable segments (HVR1 and HVR2), four conserved blocks (CSB), and variable repeats of 8 bp motifs (Ishida*, et al.*, 1994; Xu and Arnason, 1994). Figure 4 illustrates the basic structure of the D-loop region in horses using the description published in the literatures (Ishida*, et al.*, 1994; Kavar and Dovc, 2008; Xu and Arnason, 1994).

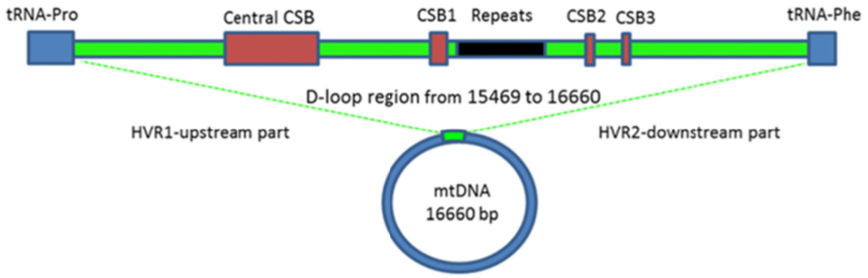


Figure 4: The basic structure of the D-loop region in horses. This figure was created using the description published in the literature (Ishida, et al., 1994; Kavar and Dovc, 2008; Xu and Arnason, 1994).

Sequencing of the whole equine mtDNA determined its structure and opened the door to applications available in other species, especially in humans, to discover more details about the equid mtDNA genomics.

Before the complete sequence of the mtDNA of horses, the first study involving mtDNA in evolutionary studies of the genus *Equus* was done by George and Ryder (1986). George and Ryder reported that the mitochondrial variation between the seven species of *Equus* supports a divergence of extant lineages from a common ancestor about 3.9 Million years before the present. Later on, the advent of mitochondrial DNA analysis in population genetics produced a revolutionary change regarding historical, biogeographic and phylogenetic perspectives on intra- and inter-specific genetic structure (Avise, 1994). Therefore, the use of mtDNA went farther than just the explanation the relationship between equine species, but also to understand intra-breed variation and infer the origin and domestication of horses. Therefore, the main common use of the mtDNA in horses can be found in the field of the evolutionary history and genetic diversity.

Thanks to the mtDNA control region, which represents a good model for studying the evolution of a non-coding region of mammalian DNA (Sbisa*, et al.*, 1997), more individuals can be sequenced and compared in order to follow the domestication history. Important questions related to this composite process were answered and valuable understandings have been provided (Patterson, 2001; Vila*, et al.*, 2001). Vila and co-authors (2001) evaluated the variation in the D-loop of the mtDNA from modern horse breeds and compared this with mtDNA from the remains of wild horses dating back 12,000–28,000 years. They used phylogenetic methods to make inferences about the evolutionary origins of modern breeds and reported that high mtDNA sequence diversity of horses suggests an unprecedented and widespread integration of matrilines and an extensive utilization and taming of wild horses (Vila*, et al.*, 2001). A similar study by Jansen and co-authors (2002) followed showing that the extensive genetic diversity revealed that several distinct horse populations were involved in the domestication of the horse. Furthermore, Cieslak and co-authors (2010) reported that the large diversity of mtDNA lineages is not a product of animal breeding, but represents ancestral variability (Cieslak*, et al.*, 2010). Recently, more haplogroups (group of similar haplotypes that share a common ancestor) that survived horse domestication have been defined by using the whole mtDNA sequence (Achilli*, et al.*, 2012).

Experimentally, mtDNA is relatively easy to extract, amplify and analyze. Add to that the specific characteristics of mtDNA that mentioned above. Therefore, this marker has been widely used to study the genetic diversity of many horse breeds (Achilli*, et al.*, 2012; Bowling*, et al.*, 2000; Cieslak*, et al.*, 2010; Cothran and Luis, 2005; Głażewska, 2010; Jansen*, et al.*, 2002; Kavar and Dovc, 2008; Lippold*, et al.*, 2011; Prystupa*, et al.*, 2012a) and developing equine mtDNA profiling for forensic application (Gurney*, et al.*, 2010).The outcomes of these genetic diversity studies were significantly important and each of them can be considered as the first step for the development of a conservation strategy for the breed tested. For example, the study by Prystupa et al. (2012a) provided the first real insight into the maternal gene flow and mitochondrial diversity within the native Canadian equine populations. In addition to the population studies above, the mtDNA insertions into the nuclear genome were also investigated in horses. Nergadze et al. (2010) reported that 82 percent of numts (nuclear sequences of mitochondrial origin) is represented in the nuclear genome.

Despite all these advantages, we need to consider some challenges about using mtDNA data sets. First, the mtDNA is the most widely used molecular tool in domestication studies, but it does not detect male mediated gene flow. This pattern of gene flow may have high influence on the evolution of livestock species in modern times (Diamond, 2002; MacHugh*, et al.*, 1997) where genetic variation can only be detected by Y chromosome markers (Wallner*, et al.*, 2013). Second, the mtDNA-nDNA interaction was not addressed very well in horses. Nuclear mitochondrial DNA (Numts) may give false polymorphic sites. Third, the whole field of phylogenetic analysis of the mtDNA heavily relies on the assumption of maternal inheritance of mtDNA. However, in some special cases such as in sheep (Zhao*, et al.*, 2001) and in *Drosophila* (Wolff*, et al.*, 2012) paternal contribution has been reported in the mtDNA inheritance.

1.4.3 Single Nucleotide Polymorphisms (SNPs)

A SNP is a single base change in a DNA sequence. This change happens as a result of a mutation in the DNA. Mutations can result by transitions, transversions and indels. Figure 5 shows examples of these mutations.

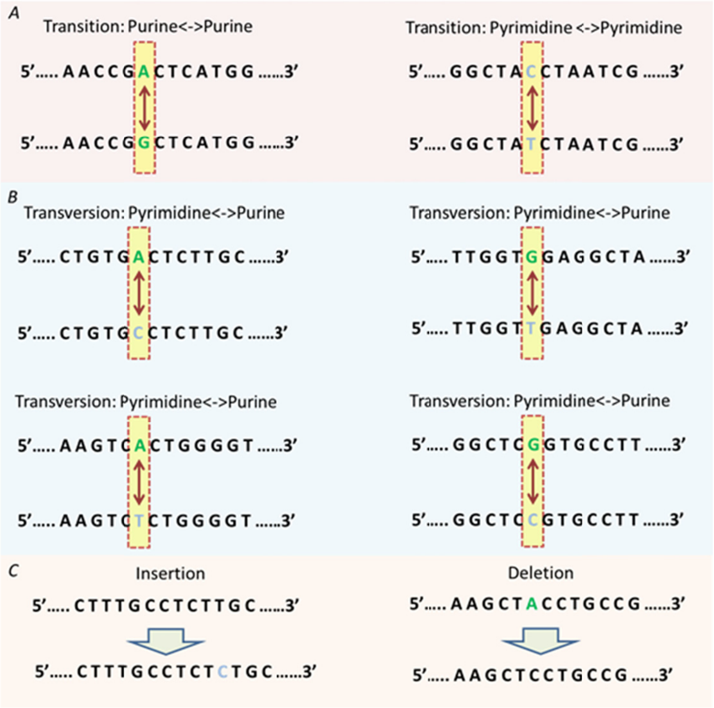

A
Transition: Purine<->Purine
5'..... AACCGACTCATGG ......3'
5'..... AACCGGCTCATGG ......3'
Transition: Pyrimidine <->Pyrimidine
5'..... GGCTACCTAATCG ......3'
5'..... GGCTATCTAATCG ......3'
B
Transversion: Pyrimidine<->Purine
5'..... CTGTGACTCTTGC ......3'
5'..... CTGTGCCTCTTGC ......3'
Transversion: Pyrimidine<->Purine
5'..... TTGGTGGAGGCTA ......3'
5'..... TTGGTTGAGGCTA ......3'
Transversion: Pyrimidine<->Purine
5'..... AAGTCACTGGGGT ......3'
5'..... AAGTCTCTGGGGT ......3'
Transversion: Pyrimidine<->Purine
5'..... GGCTCGGTGCCTT ......3'
5'..... GGCTCCGTGCCTT ......3'
C
Insertion
5'..... CTTTGCCTCTTGC ......3'
5'..... CTTTGCCTCTCTGC ......3'
Deletion
5'..... AAGCTACCTGCCG ......3'
5'..... AAGCTCCTGCCG ......3'

Figure 5: Examples of different mutations considered as SNPs. A: Transitions. B: Transversions. C: Indels. Only a single strand of DNA is shown.

During the last decade, SNPs were hypothesized to become the markers of the choice in ecological, evolutionary conservation studies (Seddon*, et al.*, 2005). Because of the rapid increase in characterization and availability of SNPs in non-model organisms (Hayes*, et al.*, 2007; Lindblad-Toh*, et al.*, 2005; Moen*, et al.*, 2008; Van Tassell*, et al.*, 2008; Wade*, et al.*, 2009; Wong*, et al.*, 2004a), a greater attention was given to this class of markers to address a broad range of evolutionary questions (Moen*, et al.*, 2008; Morin*, et al.*, 2004). SNP markers can help in understanding the recent evolutionary history of domestic animals (Goncalves*, et al.*, 2010; Pariset*, et al.*, 2009). Unlike microsatellite, SNPs have a lower mutation rate and very low false genotyping rate (Gärke*, et al.*, 2012) which makes it possible to automate and standardize SNPs analysis in high throughput technologies (Fries and Durstewitz, 2001; Xing*, et al.*, 2005). Furthermore, the recent technological advances have led to a decrease in both discovering and genotyping costs (Shen*, et al.*, 2005; Syvanen, 2005). Therefore, SNPs are likely to become the markers of choice for next generation population genetics data in the field of molecular ecology and conservation genetics (Pool*, et al.*, 2010). They have the highest density in genomes compared to other molecular markers.

Very recently, the significant SNPs discovery done by the National Human Genome Research project has produced the EquCab2.0 assembly that provided sufficient markers to construct a whole genome SNP panel for use in the domestic horse (McCue*, et al.*, 2012; Wade*, et al.*, 2009). Lately, some studies have been done using this SNPs array in horse genetics such as genetic diversity, association mapping, phylogeny study and inbreeding investigation (Binns*, et al.*, 2011; McCue*, et al.*, 2012; Petersen*, et al.*, 2013).

CHAPTER II

JUSTIFICATION AND OBJECTIVES OF THE STUDY

2.1 Justification and contributions of the study

There have been quite a few studies about the genetic diversity in the Arabian horse breed; all of which were about Western Arabian populations. Furthermore, all the previous mtDNA studies in horses were done using only a small part of the D-loop. In this study, I investigated the genetic structure of samples representing Middle Eastern and Western populations using microsatellite markers and whole mtDNA D-loop sequencing. I did a comparative analysis of the Arabian populations from different origins and provided an integrative description of the current status of genetic diversity using both nuclear and maternal inheritance approaches. This study will facilitate developing and implementing conservation programs for this important breed. The data from this study also provided new information for exploring the evolutionary history of domestication and breed origins which will contribute to international biodiversity programs. This work will contribute to both the scientific and economic aspects of horse breeding, and will guide the breeding process and support the population management of such important animals by integration of biotechnology methods (such as using molecular markers in parentage verification and genetic conservation in the Middle East). Also this study will provide a detailed comparison between the Arabian populations in the USA with other Arabian horse populations which will help breeders to maintain and improve the American-Arabian horses. A very unique importance of the current study is that it was done just before the Syrian revolution, that developed into an armed conflict, started in Syria two years ago. The outcomes of this study will help to recover the Syrian horse populations affected during the war. also the maternal inheritance results will help to track any horses that might be illegally taken out of the country during the war time. The results from this study could be applied not only in other horse populations but also in other animal species. The second part of this work is related to a whole genome scan analysis. This part is not related to the Arabian breed, but it used data from another breed (Peruvian Paso breed) to compare the variability and distribution of SNPs and microsatellites throughout the horse genome. This kind of comparison has not been done before in any horse breed.

2.2 Goals and objectives

My study aimed to:

1. Genetically survey Arabian horse populations and provide the genetic diversity and genetic structure database of samples representing Middle Eastern and Western populations to get an in depth description of the current status of the Arabian populations from different origins.

2. Determine genetic diversity and relationships between the Syrian Arabian horse populations and to optimize a suitable procedure for parentage testing for them.

3. Study the maternal diversity and phylogenetic relationships of Arabian populations and to examine the traditional classification system of the Arabian breed (RASANs system) that depends upon maternal family lines.

4. Provide data sets that may help in the recovery of the Syrian populations after the armed conflict happing in Syria.

5. Investigate the distribution pattern of the variability along different regions of the genome based upon microsatellite and SNPs markers to show whether there are differences in relative variability of the two types of markers within the same genomic region.

CHAPTER III

MICROSATELLITE ANALYSIS OF GENETIC DIVERSITY AND POPULATION STRUCTURE OF ARABIAN HORSE POPULATIONS[1]

3.1 Introduction

There have been quite a few studies of the nuclear genetic diversity in the Arabian horse breed (Bowling*, et al.*, 2000; Cervantes*, et al.*, 2008; Głażewska, 2010; Monies*, et al.*, 2011); all of them analyzed Western (USA and Europe) populations. In the present study, we investigated the genetic structure of samples representing Middle Eastern and Western populations to get an in depth description of the current status of the genetic diversity of Arabian populations from different origins.

3.2 Materials and methods

3.2.1 Population description

A total of (537) Arabian horses representing diverse set of Middle Eastern populations and Western were examined as shown in Table 2. The Middle Eastern Arabians are: Syrian Arabian (registered and non-registered), Saudi Arabian and Iranian Arabian. The Western Arabians are: Shagya Arabian, Polish Arabian and American Arabian (Davenport, Egyptian-Saudi mix, USA-Egyptian and USA-Saudi). In addition to the Arabian populations, also (128) non-Arabian horses were tested including Akhal Teke, Turkoman, and Caspian horses. The Przewalski horse was used as an out-group.

[1] Reprinted with permission from Khanshour A., Conant E., Juras R., Cothran G. (2013) Microsatellite analysis of genetic diversity and population structure of Arabian horse populations. Journal of Heredity. 104 (3): 386-398. Copyright 2013 The American Genetic Association.

Table 2: The description of the populations used in this study. The tested populations and their abbreviations (Pop., abb.), the population groups (Gr.), sample sizes (N) and sampling information.

Gr.	Pop., abb.	N	Sampling information
Middle Eastern Arabian	Saudi, SU2	33	Samples came from pure desert Arabians from Saudi Arabia. Samples were provided by breeders.
	Syrian registered, SY1	138	Samples were collected randomly from different places from Syria as following: South Syria (38), Middle Syria (60), North Syria (25), North east Syria (34), West Syria and the coastal mountains (23), National Center for horses breeding (72). All Syrian strains were represented in my samples collection. Breeders volunteered to give their horse samples. Samples from the governmental breeding station were collected under the permission from the Ministry of Agriculture and Agrarian Reform in Syria.
	Syrian non-registered, SY2	114	
	Iranian Arabian, KA	40	Samples came from Persian Arab *Asils* from Khuzestan in Iran. Samples were provided by breeders.
Western Arabian	Davenport, DV	23	Samples came from the Davenport registry in the USA. Samples came to the Animal Genetic lab at Texas A&M University for parentage testing.
	Egyptian-Saudi, mix SE	28	Samples came from the American Arabian Studbook registry. These horses were descended from a mixture of Saudi and Egyptian horses. Samples were provided by breeders.
	USA-Egyptian, EG	47	Samples came from the American Arabian Studbook registry. These horses were originally descended from Egyptian horses. Samples were provided by breeders.
	USA-Saudi, SU1	57	Samples came from the American Arabian Studbook registry. These horses were descended from Saudi horses. Samples were provided by breeders.
	Shagya Arabian, SA	21	Samples came from Performance Shagya Arabian Registry. Samples came to the Animal Genetic lab at Texas A&M University for parentage testing.
	Polish Arabian, PA	36	Samples came from the Polish Arabian horse breed in Poland. Samples were provided by Dr. G. Cholewinski from Agricultural University of Poznan.
Non-Arabian	Akhal Teke, AT	28	Samples came from the Akhal Teke horse breeders in the USA. Samples came to the Animal Genetic lab at Texas A&M University for parentage testing.
	Caspian, CS	35	Samples came from the Caspian Horse Society in the USA. Samples came to the Animal Genetic lab at Texas A&M University for parentage testing.
	Turkoman, TU	65	Samples came from the Turkoman Horse breeders in the USA. Samples came to the Animal Genetic lab at Texas A&M University for parentage testing.
	Przewaslski, PZ	17	Samples were provided by Wilds and Animal Genetic Lab

3.2.2 DNA extraction and microsatellite analysis

Total DNA was extracted from hair follicles using PUREGENE® DNA purification kit following the manufacturer's protocol.

A total of 15 microsatellite markers (ASB17, ASB2, AHT4, AHT5, HMS2, HMS7, HMS3, HMS6, ASB23, HTG10, HTG7, HTG4, HTG6, LEX33, and VHL20) specific to *Equus* caballus were used in this study. All these markers are included in the panel recommended by the International Society for Animal Genetics for diversity studies and parentage verification. Table 3 shows these markers and the chromosome number for each locus.

Table 3: The fifteen markers used in the study with the chromosome number of each locus.

Locus	Chromosome	Reference	Locus	Chromosome	Reference
ASB17	2	(Breen, *et al.*, 1997)	ASB23	3	(Irvin, *et al.*, 1998)
ASB2	15		HTG10	21	(Marklund, *et al.*, 1994)
AHT4	24	(Binns, *et al.*, 1995)	HTG7	4	
AHT5	8		HTG4	9	(Ellegren, *et al.*, 1992)
HMS2	10	(Guerin, *et al.*, 1994)	HTG6	15	
HMS7	1		LEX33	4	(Coogle, *et al.*, 1996)
HMS3	9		VHL20	30	(Van Haeringen, *et al.*, 1994)
HMS6	4				

The 15 microsatellites are amplified in three multiplex reactions using the method described by Juras et al. (2003). Fragment sizes of microsatellite alleles were determined using the STRand computer software (Locke, *et al.*, 2000). Alphabetical nomenclature was used for allele size designation in accordance with the International Society for Animal Genetics.

3.2.3 Statistical analysis

3.2.3.1 Molecular markers

Identification of possible genotyping errors due to null alleles, short allele dominance, typographic errors and the scoring of stutter peaks were detected and adjusted using Micro-checker software (Van Oosterhout*, et al.*, 2004) according to Brookfield's approach (Brookfield, 1996). Linkage disequilibrium (LD) between all pairs of loci was tested in the non-adjusted data by GENEPOP 3.4 (Raymond and Rousset, 2001) based on the exact test using the default parameters specified by the software. Hardy-Weinberg equilibrium (HWE) analyses by population and locus were carried out on the adjusted data using GENEPOP 3.4 based on the exact test. The exact p-values were obtained using MCMC simulation of 10,000 dememorization steps, 500 batches and 5,000 iterations. For the total markers together in each population, the Fisher's method implemented in GENEPOP 3.4 was used after Bonferroni correction to detect significant deviations of a population from HWE.

3.2.3.2 Gene diversity within and among populations

Gene diversity indices for each population were calculated from adjusted data using GENEALEX 6 (Paetkau*, et al.*, 1997). These included the average number of alleles per locus (Na), the effective number of alleles per locus (Ne), observed (H_O) and unbiased expected (H_E) heterozygosity or gene diversity. In addition, we calculated the number of rare alleles (Nr) (Allendorf, 1986). Allelic richness (AR), which could be considered as an alternative criterion to measure genetic diversity (Rodrigáñez*, et al.*, 2008), was used to estimate the diversity of the populations tested in this study. AR is standardized for variation in sample size and was calculated using FSTAT 2.9.3 (Goudet, 1995; Goudet, 2002) based on the minimal sample size of 12 diploid individuals. Wright's F statistics according to Weir and Cockerham (1984) were calculated using GENETIX 4.05 (Belkhir*, et al.*, 1996-2004) for the F_{IS}, and FSTAT 2.9.3 was used to calculate F_{ST}. Analysis of the Molecular Variance (AMOVA) was done for different partitions of the 13 populations (not including the out-group population) using GENEALEX 6 where the variation among populations was determined by Φ_{PT} using 999 permutations.

3.2.3.3 Relationships and genetic differentiation among populations

In order to study the relationships and genetic differentiation among tested populations, pairwise F_{ST} and R_{ST}, factorial correspondence analysis (FCA) and genetic distances were applied. Pairwise F_{ST} values were calculated using FSTAT 2.9.3.2, and P values were obtained after 10,000 permutations. The pairwise R_{ST} was done using MSAT software (http://genetics.stanford.edu/hpgl/projects/microsat/) using the standardized R_{ST} (Goodman, 1997). Representation of the genetic relationships among tested populations was done using the factorial correspondence analysis (Lebart*, et al.*, 1984) as implemented by GENETIX 4.05. Three different models for genetic distances were used. The first approach was to test the genetic relatedness among all individual horses depending on the simple matching dissimilarity indices of Jaccard's coefficient method (Perrier*, et al.*, 2003) using DARwin-5.0 (Perrier and Jacquemoud-Collet, 2006); the second approach, which included the Reynolds distance DR (Reynolds*, et al.*, 1983), Nei distance D (Nei, 1972) and Cavalli-Sforza Chord distance DC (Cavalli-Sforza and

Edwards, 1967) was estimated from 10,000 bootstrapped allele frequency datasets using PHYLIP package (Felsenstein, 1989-2006). For the third approach, genetic distances were calculated depending on the standardized RST method (Goodman, 1997) using MSAT software and PHYLIP package with 10,000 bootstraps. The dendrograms of phylogenetic trees were built from different distance matrices and were visualized by DARwin-5.0 and MEGA4 (Tamura*, et al.*, 2007) using the neighbour-joining method (Saitou and Nei, 1987).

3.2.3.4 Population structure and individuals assignment

We used the STRUCTURE 2.3.3 software (Pritchard*, et al.*, 2000) to study the relationships among the Arabian populations, and to assign samples into clusters using the Bayesian method under an admixture model. Different values of the length of the burn-in period (20,000 to 50,000) and MCMC repetitions (100,000 to 150,000). Different K values between K = 2 to K = 13, where K is the number of tested clusters, were applied. Runs for each K were repeated ten times. The software CLUMP (Jakobsson and Rosenberg, 2007) was used to align multiple replicates for each K in order to facilitate the interpretation of clustering results. The DISTRUCT application (Rosenberg, 2004) was used to graphically display the results. The best number of clusters was determined depending on ΔK value (Evanno*, et al.*, 2005) which was calculated and plotted using Structure Harvester application (Earl and vonHoldt, 2011).

3.3 Results

3.3.1 Microsatellite markers

All 15 loci tested in this study were found to be polymorphic in all populations except HTG6 and HTG7 which were not variable in the DV population. A total of 143 alleles were detected in 682 individuals of the 14 tested populations. The 15 loci in all tested populations showed no evidence of scoring errors due to stuttering or for large allele dropout. Furthermore, there was no evidence for null allele presence in any populations except for HTG7 in the non-registered Syrian population at 0.05 level.

The statistical significance of two-locus LD among 15 microsatellite loci was tested by the exact test; the LD P-values were obtained for 105 pairs of combinations in each population. At the level of $p<0.05$, there were 23 out of 105 pairs in linkage equilibrium (LE) in all tested populations. However, no pair was in constant LD in all populations. In addition, no population shows complete LE of all non-syntenic loci, while nine populations were in LE for three syntenic loci (HTG7, LEX33, and HMS6). Figure 6 shows all pairs of comparison among 15 markers with the number of populations out of 14 showing significant LD at $p<0.05$.

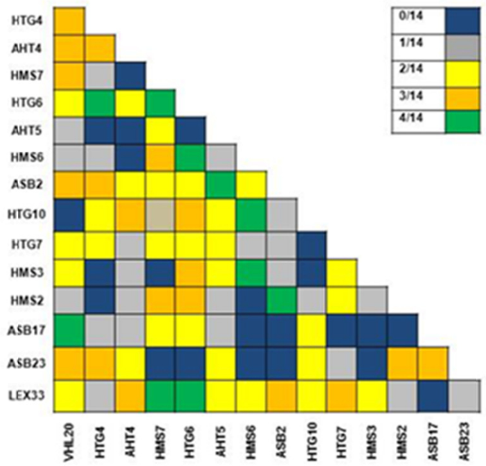


Figure 6: Pairs of comparison among 15 markers and the number of populations out of 14 for which any pair of markers shows significant LD at p<0.05

The exact test of deviation from HWE was done on adjusted data. Of the 210 population-locus combinations, 15 combinations deviated significantly (p<0.05) from HWE, five populations (PZ, PA, KA, CS, and SU2) were in complete equilibrium for all loci, and four populations (SU1, ES, SA, and TU) were in disequilibrium at only one locus (LEX33, ASB23, AHT4 and AHT5), respectively. The maximum number of the populations showing Hardy-Weinberg disequilibrium (HWD) for any one locus was three. The overall Fisher's test in each population with Bonferroni correction showed that only the SY1 and the DV populations were not in HWE at a significance level of ά = 0.0033 (ά = 0.05/15 = 0.0033). Figure 7 shows the HWE status in all combinations among tested populations and loci, and also shows the total probability of the deviation from HWE of each population over all loci.

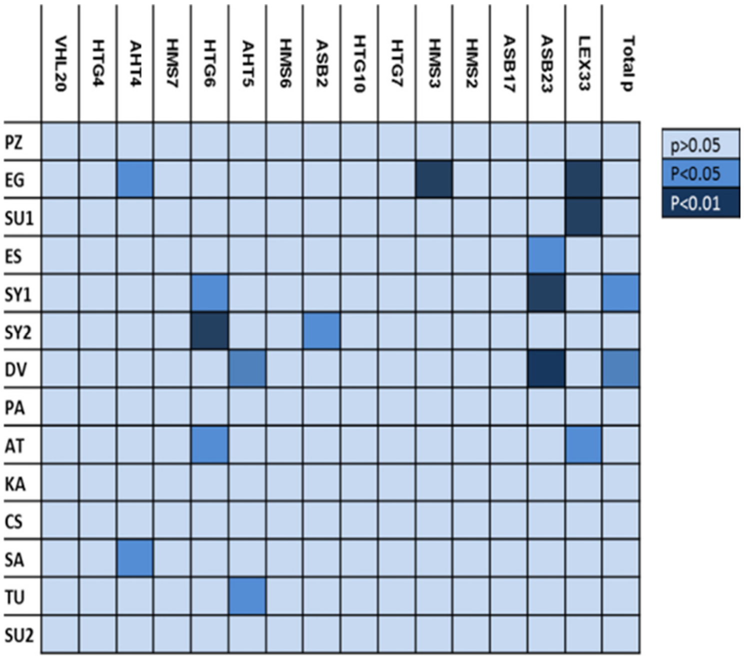


Figure 7: Hardy-Weinberg equilibrium status in 14 populations and 15 loci. For the total markers together in each population, the Fisher's method was used after Bonferroni correction to determine the significant deviation of a population from HWE. Population abbreviations could be seen in Table 2.

3.3.2 Genetic diversity within and among populations

The genetic diversity measures for each population are shown in Table 4.

Table 4: The genetic diversity measures for each population. Average number of alleles per locus per population (Na), average number of effective alleles per locus per population (Ne), observed heterozygosity (H_O), average number of rare alleles with frequency less than 0.1 per locus per population (Nr), unbiased expected heterozygosity (H_E), and allelic richness (AR). All means are combined with its standard error (SE), and FIS values are combined with the significant status. Population abbreviations could be seen in Table 2. *Values different from 0 at $p < 0.05$.

Populations	Na (SE)	Ne (SE)	Nr (SE)	H_O (SE)	H_E (SE)	F_{IS}	AR (SE)
SU2	5.13 (0.31)	3.30 (0.26)	1.87 (0.31)	0.68 (0.03)	0.68 (0.03)	0.008	4.51 (0.26)
SY1	6.47 (0.38)	3.51 (0.24)	2.87 (0.36)	0.70 (0.03)	0.69 (0.03)	-0.007	4.69 (0.22)
SY2	8.47 (0.59)	4.23 (0.27)	4.53 (0.45)	0.72 (0.02)	0.75 (0.02)	0.037*	5.62 (0.28)
KA	5.93 (0.37)	3.61 (0.24)	2.27 (0.29)	0.70 (0.02)	0.71 (0.02)	0.017	5.06 (0.27)
DV	3.00 (0.31)	2.12 (0.94)	0.40 (0.16)	0.40 (0.06)	0.46 (0.06)	0.132*	2.74 (0.24)
SE	3.53 (0.19)	2.35 (0.17)	0.60 (0.18)	0.58 (0.05)	0.55 (0.04)	-0.066*	3.20 (0.15)
EG	4.00 (0.26)	2.43 (0.17)	1.33 (0.28)	0.53 (0.04)	0.56 (0.04)	0.047	3.46 (0.20)
SU1	4.40 (0.31)	3.02 (0.22)	1.00 (0.28)	0.66 (0.02)	0.65 (0.03)	-0.015	3.78 (0.22)
SA	4.93 (0.30)	3.26 (0.21)	1.47 (0.35)	0.68 (0.03)	0.69 (0.03)	0.005	4.66 (0.24)
PA	5.67 (0.40)	3.41 (0.30)	2.67 (0.38)	0.69 (0.04)	0.68 (0.04)	-0.015	4.83 (0.32)
AT	5.27 (0.34)	3.31 (0.35)	1.93 (0.24)	0.72 (0.06)	0.65 (0.05)	-0.114*	4.60 (0.31)
CS	6.87 (0.48)	4.18 (0.36)	2.87 (0.35)	0.75 (0.03)	0.74 (0.03)	-0.009	5.70 (0.34)
TU	7.93 (0.65)	4.70 (0.41)	3.87 (0.62)	0.75 (0.02)	0.77 (0.02)	0.022	6.20 (0.41)
PZ	3.87 (0.22)	2.48 (0.19)	1.13 (0.21)	0.64 (0.04)	0.58 (0.03)	-0.094*	3.64 (0.19)

Na ranged from 3 in DV to 8.47 in SY2, and Ne ranged from 2.12 in DV to 4.7 in TU. Nr followed the Na pattern, and varied from 0.4 in the DV to 4.53 in SY2. H_O ranged between 0.4 in DV and 0.75 in TU, whereas H_E ranged between 0.46 in DV to 0.77 in TU. AR varied in the similar pattern of the other within-breed diversity measures and ranged from 2.74 in the DV to 6.20 in the TU. F_{IS} varied between -0.114 in the AT to 0.132 in the DV. While FIS in DV and SY2 was significantly positive, it was significantly negative in AT, SE, and PZ.

AMOVA was done for seven different combinations of the 13 populations which were partitioned according to prior knowledge about population origin as shown in Table 5.

Table 5: AMOVA table. Among (ΦPT) and within populations variation (WPV) under different combinations. ΦPT values were calculated at p = 0.001. Population abbreviations could be seen in Table 2.

no	Combinations	No. of populations	WPV	(Φ_{PT})
1	Arabian and non-Arabian (EG, SU1, SE, PA, KA, SA, DV, SY1, SY2, SU2, AT, CS, TU)	13	0.810	0.190
2	Arabian (EG, SU1, SE, PA, KA, SA, DV, SY1, SY2, SU2)	10	0.817	0.183
3	Middle Eastern Arabian (SY1, SY2, SU2, KA)	4	0.946	0.054
4	Western Arabian (EG, SU1, SE, PA, SA, DV)	6	0.670	0.330
5	Middle Eastern Arabian and non-Arabian (SY1, SY2, SU2, KA, AT, CS, TU)	7	0.903	0.097
6	Western Arabian and non-Arabian (EG, SU1, SE, PA, SA, DV)	9	0.751	0.250
7	Non-Arabian (AT, CS, TU)	3	0.861	0.139

Analyzing the Western and Middle Eastern Arabian in two separate structures (combination 3 and 4) showed a dramatic change in the genetic variation compared with analyzing all Arabian populations together (combination 2). Furthermore, excluding the Western Arabian (combination 5) from all populations (combination 1) caused a noticeable decrease in the ΦPT value. Such a big decrease of the Φ_{PT} value was not observed by excluding the non-Arabian populations (combination 2). In contrast, excluding the Middle Eastern Arabians (combination 6) caused a great increase in the Φ_{PT} value. This indicates the Western Arabian populations are the primary source of variation among populations in our study.

3.3.3 Genetic differentiation and relationships among populations

Pairwise F_{ST} and standardized R_{ST} are shown in Table 6.

Table 6: Analysis of pairwise population differentiation using F_{ST} (above diagonal) and standardized R_{ST} (below the diagonal). Patterns of Middle Eastern and Western population comparison for F_{ST} were represented by three colors (Middle Eastern Arabian vs. Western Arabian highlighted in yellow, Middle Eastern Arabian vs. Middle Eastern Arabian highlighted in green, Western Arabian vs. Western Arabian highlighted in pink, non-Arabian were not highlighted). Population abbreviations could be seen in Table 2.

Populations	PZ	EG	SU1	SE	SY1	SY2	DV	PA	AT	KA	CS	SA	TU	SU2
PZ		0.356	0.335	0.395	0.259	0.234	0.416	0.299	0.285	0.272	0.248	0.280	0.219	0.309
EG	0.599		0.176	0.161	0.113	0.107	0.319	0.152	0.238	0.077	0.170	0.141	0.136	0.118
SU1	0.578	0.174		0.151	0.094	0.066	0.243	0.092	0.171	0.075	0.133	0.096	0.088	0.042
SE	0.564	0.156	0.132		0.164	0.132	0.357	0.185	0.244	0.125	0.195	0.195	0.164	0.134
SY1	0.654	0.153	0.136	0.276		0.016	0.164	0.063	0.108	0.051	0.107	0.060	0.055	0.050
SY2	0.618	0.13	0.067	0.177	0.016		0.154	0.038	0.081	0.025	0.066	0.037	0.024	0.022
DV	0.637	0.334	0.210	0.314	0.156	0.150		0.228	0.271	0.208	0.267	0.214	0.179	0.199
PA	0.499	0.137	0.106	0.165	0.089	0.059	0.244		0.128	0.041	0.098	0.073	0.054	0.055
AT	0.488	0.286	0.159	0.24	0.277	0.182	0.303	0.129		0.102	0.126	0.104	0.074	0.122
KA	0.474	0.114	0.092	0.147	0.066	0.028	0.193	0.025	0.122		0.077	0.043	0.040	0.027
CS	0.456	0.196	0.101	0.148	0.151	0.103	0.273	0.102	0.131	0.093		0.097	0.050	0.108
SA	0.447	0.139	0.079	0.143	0.127	0.069	0.132	0.07	0.176	0.059	0.139		0.058	0.046
TU	0.495	0.172	0.07	0.164	0.058	0.023	0.207	0.067	0.077	0.047	0.041	0.124		0.054
SU2	0.536	0.167	0.042	0.141	0.13	0.059	0.147	0.101	0.195	0.068	0.124	0.026	0.102	

All of these values were significant at $p< 0.05$ after Bonferroni correction. The PZ, the out-group, showed very high F_{ST} values with all tested populations that ranged between 0.416 to 0.219 with the DV and TU, respectively. The lowest F_{ST} value, 0.016, was recorded between the two Syrian populations SY1 and SY2 as well as between the SY2 and SU2. Three of the Western Arabians populations PA, SA and SU1 were less differentiated from the Middle Eastern Arabian populations than the other two Western populations EG and SE which have relatively high F_{ST} with the SY1, SY2, and SU2. In addition, the six possible comparisons among the four Middle Eastern Arabian populations SY1, SY2, SU2 and KA showed low F_{ST} values. In most cases, the standardized R_{ST} showed the same pattern as F_{ST}. But R_{ST} only was greater than F_{ST} value in the comparisons of the out group PZ with all populations.

Figure 8 shows the result of the FCA among all populations except the PZ. Each population was represented by its center of gravity point. Figure 9 shows the result of the FCA using nine Arabian populations (EG, SU1, SE, SY1, SY2, PA, KA, SA and SU2) where each individual was plotted into the 3D plot.

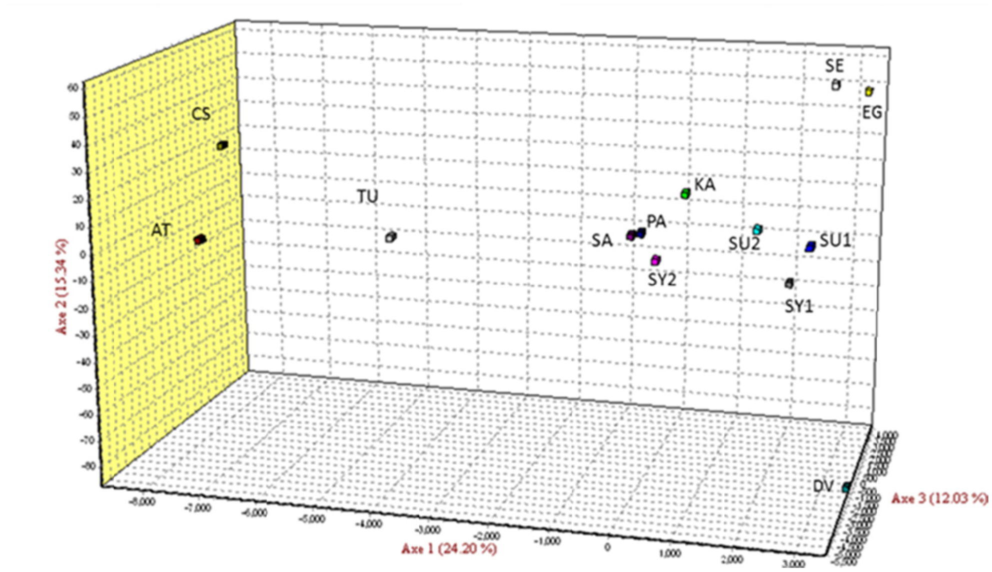


Figure 8: The FCA among all populations except the PZ. Each population was represented by its center of gravity point into 3D plot. Axis 1 accounts for 24.20% of the variation. Population abbreviations are found Table 2. The percentage of variance explained by each axis is found in the axes labels.

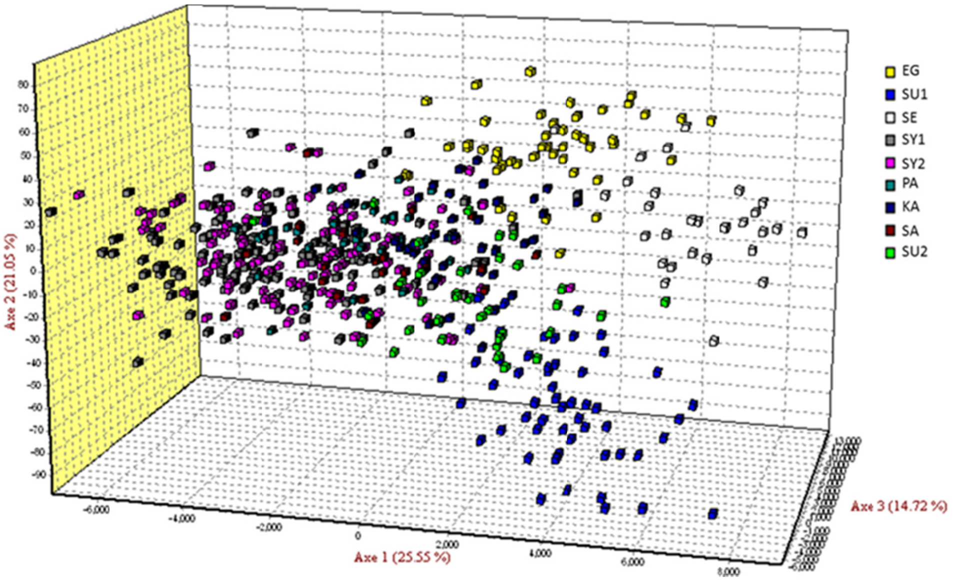


Figure 9: The FCA using only the Arabian populations. Each individual was plotted into 3D plot. Axis 1 accounts for 25.55% of the variation. Population abbreviations are as found in Table 2.

The FCA as shown in Figure 8 clearly differentiates the Arabian populations and the non-Arabian on the first axis which explains 24.2% of the genetic variance. However, the DV was clearly isolated from the rest of the Arabians and not used in this analysis. The FCA as shown in Figure 9 revealed some separation between the Middle Eastern Arabian populations and a few of the Western ones (SE, EG and SU1). Simultaneously, the PA and SA were clustered together with the Middle Eastern populations. Most Western populations were differentiated from each other.

The genetic relatedness based on the simple matching dissimilarity among all individual horses was represented in an individual-animal-based neighbor-joining dendogram, as shown in Figure 10.

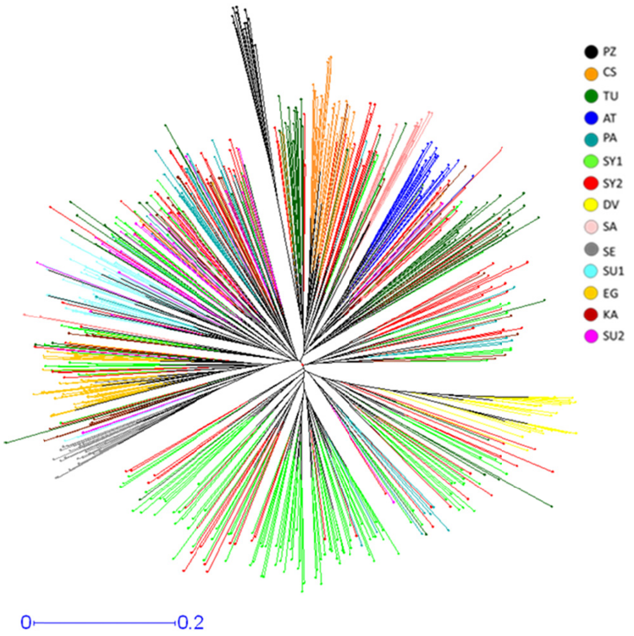


Figure 10: The individual-animal-based neighbor-joining dendogram depending on the simple matching dissimilarity indices of Jaccard’s coefficient among all individual horses. Population abbreviations are found in Table 2.

This dendogram shows that the majority of horses within each population were closely assembled in discrete branches, but there were some exceptions. Each of the non-Arabian populations (CS, AT and the PZ) and the Western Arabian populations (DV, PA, SA, SE, EG, SU1) were segregated clearly into a single branch. In contrast, all the Middle Eastern Arabian populations (SY1, SY2, SU2 and KA) did not show clear segregation in a single branch, with samples from each population distributed in more than one clade. Horses from the SY2 population were segregated into almost all branches including those branches of the non-Arabians. Similarly for the TU, horses were segregated into three distinct branches. Two of these branches were parts of the Arabian clade. Furthermore, the individual-animal based dendogram showed that the majority of the SY1 horses were not differentiated from the SY2 horses. Likewise, the SU2 shared some individuals with the SU1 and some with the KA, and the SE and the EG were very close to each other.

The estimates of genetic distances D, DC and DR revealed similar topologies for all populations tested here. While low bootstrap values were observed using D, the bootstrap values in both DC and DR were similar to each other and relatively high. Figure 11 shows the DR neighbor-joining dendogram including the PZ as the out-group.

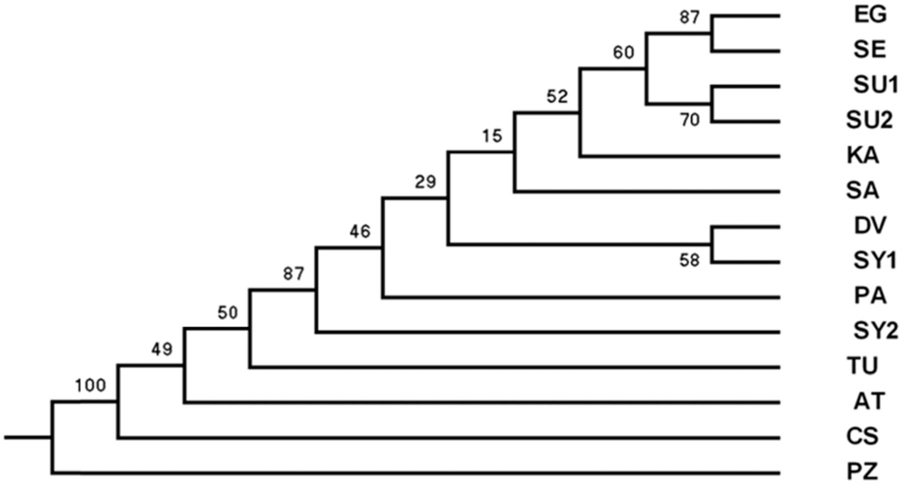


Figure 11: The chord DR neighbor-joining dendogram. It includes 13 horse populations and the Przewaslski as an out-group. The tree was estimated from 10,000 bootstrapped. Bootstrap values were listed as percentages. Population abbreviations could be seen in Table 2.

The non- Arabian populations (CS, AT and TU) cluster together and were separated from the Arabians. All Western and Middle Eastern Arabian populations fell into one big clade. Within the clade of the Arabian populations each of the (SY1 and DV), (EG and SE) and (SU1 and SU2) were paired together.

3.3.4 Population structure and individual assignment

Two STRUCTURE analyses were done in this study. The first was performed using all the Arabian populations (SY1, SY2, PA, KA, SA, SU1, SU2, SE, EG and DV) plus the out group. The DV population was excluded from the second analysis. Similar likelihood values were obtained using different values of the length of burn-in period (20,000 to 50,000) and MCMC repetitions (100,000 to 150,000), so we report only the results of using 20,000 burn in period and 100,000 of MCMC. For the all Arabian populations analysis we used 12 values of K (K = 2 to K = 13). The first STRUCTURE clustering did not give a clear mode for ΔK and the higher ΔK was associated with the lower value of K as shown in Figure 12. On the other hand, by excluding the DV population and using 11 values of K (K = 2 to K = 12) the highest ΔK was found at K = 5 as shown in Figure 13. That indicated the probable number of clusters (Evanno, *et al.*, 2005).

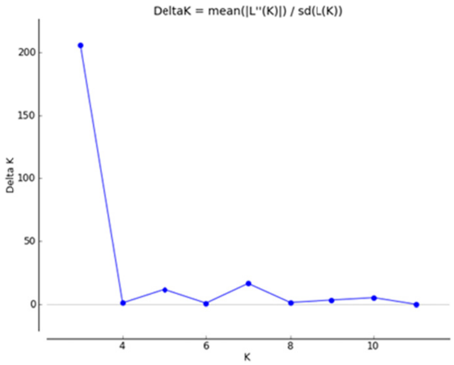


Figure 12: The ΔK distribution for different values of clusters (K) for 11 populations.

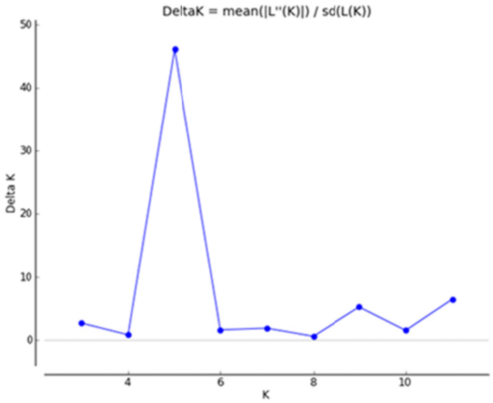


Figure 13: The ΔK distribution for different values of clusters (K) for ten populations.

Table 7 shows the proportion of individuals assigned into each of the five clusters depending on the Q value that resulted from the STRUCTURE analysis.

Table 7: The individuals' assignment into five clusters at K = 5. The highest value in each for each population is in bold. Population abbreviations could be seen in Table 2.

Populations	clusters				
	1	2	3	4	5
SU2	0.003	0.066	0.521	0.304	0.106
SY1	0.005	0.772	0.038	0.128	0.058
SY2	0.014	0.341	0.053	0.554	0.038
KA	0.007	0.061	0.072	0.671	0.189
SE	0.002	0.008	0.079	0.010	0.901
EG	0.004	0.017	0.012	0.029	0.939
SU1	0.002	0.011	0.964	0.009	0.015
SA	0.016	0.158	0.090	0.703	0.033
PA	0.004	0.084	0.050	0.831	0.030
PZ	0.984	0.007	0.002	0.005	0.002

The first cluster mainly consisted of the PZ where 98.4% of the PZ individuals were assigned into this cluster. The second cluster consisted of 77.2% of SY1 and 34.1% of SY2 in addition to 15.8% of the SA population. The third cluster contained 96.4% of SU1 and 52.1% of the SU2. Four different populations, 83.1% of the PA, 70.3% of the SA, 67.1% of the KA and 55.4% of the SY2, formed the fourth cluster which was the most admixed one. The fifth cluster mainly was formed by 93.9% of the EG and 90.1% of the SE.

As shown in Figure 14A, no clear separation of distinct Arabian populations was noticed at K = 2, but all the out-group individuals formed a separate cluster. At K = 3, three populations (SE, EG, SU1) plus most of the SU2 individuals were completely separated from the rest and the out-group can be recognized and completely separated. At K = 4, the SE and EG populations can be easily distinguished from all other populations but they still show some intermixture with the SU1. At K =5, which was the best value of the number of clusters that represent the structure of the data, the SE together with EG formed a distinct cluster with a very few individuals from KA and SY1 assigned into this cluster. Also, the SU1 plus most of SU2 individuals formed a second cluster. The SY2, KA, SA and PA together formed an admixed cluster with a few individuals from the SY1 population. The later and SY2 shared some other individuals. According to the results at K = 5, further analysis were done to determine if individuals of SE and EG (the fifth cluster) could be distinguished from each other. The PZ was used as an out-group. The highest ΔK was found at K =3 where the SE formed an independent cluster as did the EG, Figure 14B. Another subset of data was used to determine if the admixed cluster (the fourth cluster) which contains the SY2, KA, SA and PA could be separated into different substructures. The highest ΔK was found at K = 3. While only the PA formed a separate cluster, clear evidence of admixture was detected among SY2, KA and SA, Figure 14C.

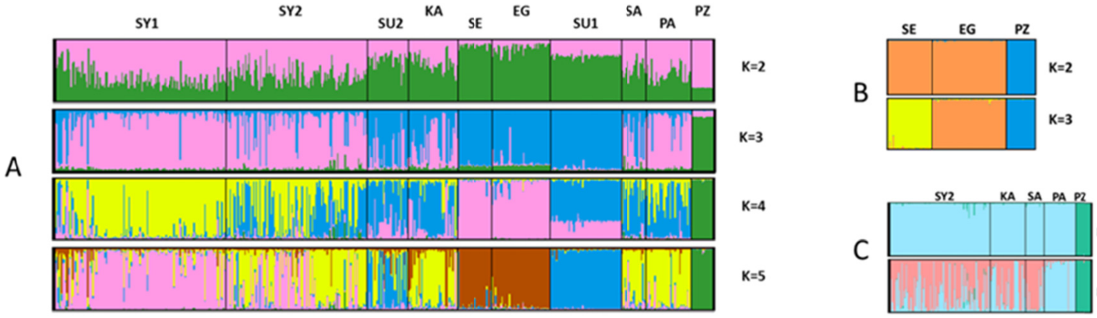


Figure 14: Clustering assignment depending on the Bayesian method under an admixture model obtained by STRUCTURE software. Each individual is represented by a single column which is divided into segments whose size and color correspond to the relative proportion of the animal genome corresponding to a particular cluster. Populations are separated by black lines. A: the analysis of nine populations (without Davenport) plus the out-group using K=2 to K=5. B: the analysis of two populations as new subset with the out-group using K=2 and K=3. C: analysis of 4 populations as new subset with the out-group using K=2 and K=3. Population abbreviations could be seen in Table 2.

3.4 Discussion

This work presents the first description of the genetic diversity and population structure in native Arabian horses sampled from Syria, Iran and Saudi Arabia and some Western Arabian populations bred in Europe and the USA.

3.4.1 Microsatellite markers

In order to identify the genotyping errors in our data, we tested it considering the possible excess of homozygotes and its distribution over all allele classes. That allows a visual inspection of a locus, and might reveal whether a locus contains null alleles or shows a deviation from HWE due to large allele dropout (Van Oosterhout*, et al.*, 2004). In our study, multi-locus genotyping were done, indeed we did can discriminate between inbreeding and Wahlund effects and the HWD caused by null alleles. Our results here crucially showed that there was not an overall homozygote excess distributed homogeneously across all homozygote classes in 13 out of 14 tested populations. However, this pattern was not stable only in the non-registered Syrian population for HTG7 locus. Such as finding indicates the possibility of the presence of null alleles for this locus in very few genotypes of the non-registered Syrian horses, and also confirms the genotyping accuracy in all other data we obtained. Therefore, to overcome these genotyping errors we used Brookfield's approach (Brookfield, 1996) to adjust the allele frequencies in a few genotypes of the non-registered Syrians. The adjusted allele frequencies can be used subsequently for further population genetic analysis (Van Oosterhout, *et al.*, 2004). Thus, in our study we used the adjusted allele frequencies in all following analyses except for the LD analysis.

The LD analysis did not give constant results for any pair of loci along all tested populations. Hence, none of these loci were excluded from further genetics analysis in our study. On the other hand, there was a big difference between populations in the number of pairs with significant LD. Relatively high numbers of pairs showing high LD were noticed in the Syrian registered, Davenport, and the Iranian Arabian populations, but fewer were seen for the Caspian population. It was well established from classical population genetics theory that genetic drift, migration, mutation and selection may generate LD (Bulmer, 1971; Ohta and Kimura, 1969; Stephens, *et al.*, 1994). Thus, it was not surprising to see such high LD in the DV population which is known to be small inbreed population with a founder effect. Also, the admixture and possible selection in the Syrian registered population may maintain and increase LD at least for some loci.

The exact test is considered as the most appropriate tool to check the deviation from HWE (Mukesh*, et al.*, 2009). It is the desirable test when there are multiple alleles and also when the number of some genotype categories are small (Hedrick, 2005). This test was done on the adjusted data to eliminate the possible effect of the null alleles which may cause significant deviation of the HWE. The overall Fisher's test in each population with Bonferroni correction showed that only the Syrian registered and the DV populations were not in HWE at a significance level of ά = 0.0033 (ά = 0.05/15 = 0.0033). Generally, deviation from the HWE can be a result of one or more of the following factors: selection against or favoring heterozygotes, inbreeding, gene flow, nonrandom mating, and Wahlund effect (Hedrick, 2005). While Wahlund effect and/or possible selection may cause this deviation in the Syrian registered, founder effect might be the reason of the HWD noticed in the DV population. Some populations (SU1, ES, SA, TU, EG, SY1, SY2, DV and AT) showed HWD in three or less loci. Deviation from HWE in some tested loci was already reported in many different horse breeds; in some European native horse breeds HWD were recorded for the HMS3 and HTG6 (Solis*, et al.*, 2005). VHL20 was not in HWE in two Portuguese breeds (Luís*, et al.*, 2002) unlike our result where the same locus was in HWE in all tested populations. Six loci at least were in HWD in some Pantaneiro horse breeds (Giacomoni*, et al.*, 2008) and five different Indian horse breeds were in HWD at many loci (Behl*, et al.*, 2007). Furthermore, the Brazilian Criollo breed showed significant HWD for HMS7, HMS6, AHT5, HMS3, HTG4, HTG10, AHT4 and VHL20 (Costa*, et al.*, 2010). A recent study by (van de Goor*, et al.*, 2011), using the same markers that we used, reported that five loci were deviated from the HWE in the Arabian horse breed in France and the HTG10 was the most frequent deviated locus a cross eight out of 35 different breeds.

3.4.2 Genetic diversity within and among populations

Within the Middle Eastern Arabian populations group, all four populations (Saudi, Syrian registered, Syrian non-registered and Iranian Arabian) had high heterozygosity values (0.68, 0.69, 0.75 and 0.71, respectively) but the Syrian non-registered was the highest. This value is among the highest heterozygosity values reported for other horse populations using the same or similar loci, (Leroy*, et al.*, 2009; Luís*, et al.*, 2007). Furthermore, the Syrian non-registered value of heterozygosity, according to our best knowledge, is the highest reported to date in the Arabian breed (Aberle*, et al.*, 2004; Conant*, et al.*, 2012; Glowatzki-Mullis*, et al.*, 2006; Iwanczyk*, et al.*, 2006; Ouragh, 2005; Plante*, et al.*, 2007; Solis*, et al.*, 2005; van de Goor*, et al.*, 2011). The high genetic diversity that was found in the Syrian non-registered likely reflects the wide and diverse base of this population, supported by the high Ne and may include introgression from non-Arabian horses. The later consideration is supported by a very high number of rare alleles in this population (Table 4) as compared other populations of this study. In addition, because the Syrian non-registered horses are not considered as pure Arabian in Syria, there is no liability to the Syrian breeders in outcrossing with any other horses, however, this outcrossing is limited and most breeders maintain and control their horse breeding. In contrast, outbreeding is prohibited in Syrian registered and Saudi and this maybe one of the reasons that Syrian registered and Saudi have lower values for heterozygosity, Na and Nr than Syrian non-registered.

The Western Arabian populations show lower values of variation compared to the Middle Eastern populations. The decrease of genetic diversity in populations being moved away from their possible center of origin was reported in different species including horses. Tozaki et al. (2003) reported that Japanese horses originated from Mongolian horses and the former had a lower genetic diversity than the later. Also Warmuth et al. (2012) mentioned the correlation between the loss diversity and east-to-west migrations of non-breed horses. Furthermore, it was also reported that the genetic diversity in human populations decreases with distance from Africa (Tishkoff*, et al.*, 2009) and that was consistent with the proposed serial founder effects resulting from the migration of modern humans out of Africa and across the globe (Jakobsson*, et al.*, 2008).

While the Davenport, Egyptian-Saudi mix, and USA-Egyptian have very low diversity, the Shagya Arabian and Polish Arabian have higher values. A very low heterozygosity in the Davenport is likely due to the founder effect, genetic drift and/or inbreeding. The records for this population indicated that only a few stallions and mares imported from the Middle East were used to start the line, and it has existed as a small, closed population since. The Shagya Arabian and Polish Arabian had the highest variability of the Western Arabian populations. For the Shagya Arabian, this population is known to be a mixture of Arabian and native Hungarian horses (Hendricks, 1995). In addition, new stallions from the Middle East were introduced into this population at a later time. For the Polish Arabian, the high heterozygosity seen here did not match that reported by Głażewska and Gralak (2006) where a very low diversity was found using protein markers. Also, another study, (Głażewska and Jezierski, 2004) reported a reduction in genetic diversity in the Polish Arabian due to inbreeding and founder effect. However, the two previous studies did not use microsatellites which have much higher variability and this may be the reason behind the dissimilar results. The high variation in the Polish Arabian is probably due to the reconstruction of the Polish Arab stock between 1918- 1946 (after each of the World Wars) using horses from the Near East and various European countries; most of these later imported European horses were of Polish origin or were the descendants of ancestors already present in the pedigree of horses in Polish studs (Głażewska and Jezierski, 2004). All non-Arabian breeds (Akhal Teke, Caspian and Turkoman) had high values of heterozygosity similar to what has been reported in different studies (Conant*, et al.*, 2012; van de Goor*, et al.*, 2011).

One of the powerful tools to support decisions that depend on heterozygosity in different populations is Na (Allendorf, 1986). It has been reported as the most relevant parameter in conservation programs (Barker, 2001; Petit*, et al.*, 1998). The preference of this parameter over heterozygosity is because in some cases heterozygosity provides an overly optimistic view when there are many alleles at a locus or when the population goes through a small or recent bottleneck (Allendorf, 1986; Luikart*, et al.*, 1998). Na ranged from 3 in the Davenport to 8.47 in the Syrian non-registered and showed low values in the most Western populations (Davenport, Egyptian-Saudi mix, USA-Egyptian, USA-Saudi and Shagya Arabian) which may indicate a recent bottleneck or founder effect in those populations. A drawback of the Na measure is that it is strongly influenced by sample size (Hedrick, 2005), for that reason, we also measured allelic richness (AR). AR showed the same pattern among all tested populations as Na which means that sample sizes for those populations had no noticeable effect on Na. A similar result was reported by Marletta et al. (2006).

F_{IS}, which reveals the degree of departure from random mating, varied between – 0.114 in the Akhal Teke to 0.132 in the Davenport. The negative significant F_{IS} seen in the Akhal Teke represent an excess of heterozygosity which may be a result of outbreeding. Similar FIS values have been reported in Akhal Teke (Conant*, et al.*, 2012). The excess of heterozygosity in the Egyptian-Saudi mix may be due to the mixing of some Egyptian and Saudi horses during the establishment this population based upon the pedigree record of this population. However, all samples tested here were North American (not directly from Egypt) and may not reflect an accurate picture of this population. The positive significant F_{IS} seen in the Davenport combined with low Na and Ne indicate a deficit of heterozygosity likely due to a high level of inbreeding in this small closed population. The significant positive F_{IS} found in the Syrian non-registered is most likely a result of the Wahlund effect considering the high Na and Ne found and that these samples represent a population that came from different geographic regions in Syria.

3.4.3 Relationships and genetic differentiation among populations

The AMOVA, which was done for seven different combinations of 13 populations, suggested that the Western Arabian populations were the main source of the among populations genetic variation found. This result likely was due to the low level of diversity within the Western Arabian populations caused by genetic drift and bottleneck effects in these populations. Also the non-Arabian breeds used here are genetically very close to the Arabian (Conant*, et al.*, 2012).

All pairwise comparisons of F_{ST} were significant. F_{ST} values are typically significant so it is not surprising to find such differences, and they may not necessarily be biologically meaningful (Hedrick, 1999; Waples, 1989). Regardless of the P-values, it has been suggested that a F_{ST} value lying in the range 0–0.05 indicates low genetic differentiation; a value between 0.05 and 0.15 indicates moderate differentiation; a value between 0.15 and 0.25 indicates great differentiation; and values above 0.25 indicates very great genetic differentiation (Hartl and Clark, 1997; Wright, 1978). For Syrian registered and Syrian non-registered, $F_{ST} = 0.016$ indicates low differentiation likely due to some admixture between these two populations. As well, in some cases registered stallions fathered offspring of the non-registered mares. The F_{ST} for each pair Syrian registered-Polish Arabian, Syrian non-registered-Polish Arabian and Saudi-Polish Arabian were relatively low which may reflect recent introduction of new horses into the Polish Arabian population from the Middle East. The low differentiation among the Syrian registered, Syrian non-registered, Iranian Arabian, Saudi Arabian and Turkoman may be due to a common ancestor for those populations. The Davenport showed high F_{ST} values which is due to the differentiation based upon loss of variation in this small population (Balloux and Lugon-Moulin, 2002). Finally, it was relevant to estimate and compare both F_{ST} and R_{ST} in our study, particularly, because we have Middle Eastern and Western population comparisons where important differences in levels of differentiation are expected (Balloux and Lugon-Moulin, 2002). Comparing F_{ST} and R_{ST}

values can provide insights into the main causes of population differentiation (Hardy*, et al.*, 2003). In most cases in our study, the R_{ST} showed the same pattern of differentiation as F_{ST}. R_{ST} values were greater than F_{ST} only in the comparisons of the out-group (Przewaslski) with all tested populations which is not unexpected. In all other comparisons R_{ST} values were close to F_{ST} values suggesting a common evolutionary pattern for these populations under domestication (Hardy*, et al.*, 2003), and reflect a short divergence time between those populations.

The out-group was not included in the first FCA, shown in Figure 8, in order to get better resolution of the relationship among tested populations. The general outcomes from the FCA matched the results explained by both the genetic relatedness among all individuals and the traditional genetic distances among populations tested here. The FCA using 13 populations (Akhal Teke, Caspian, Turkoman, Syrian registered, Syrian non-registered, Polish Arabian, Iranian Arabian, Shagya Arabian, USA-Saudi, Saudi, Egyptian-Saudi mix, Davenport, and USA-Egyptian), as well the traditional phylogenic trees, separated the Arabian populations (Syrian registered, Syrian non-registered, Polish Arabian, Iranian Arabian, Shagya Arabian, USA-Saudi, Saudi, Egyptian-Saudi mix, Davenport and USA-Egyptian) from the non-Arabians (Akhal Teke, Caspian, and Turkoman). This may indicate reproductive isolation in the last 100 or more years. Similar result for Arabian and some Italian populations using FCA was reported by Di Stasio*, et al.* (2008) based upon a similar set of microsatellite markers. Although the Turkoman population was separated from the Arabians, as shown in Figure 11 and Figures 8, it was closer to the Arabian than the Akhal Teke and Caspian as some

Turkoman individuals can be seen closely neighboring some Arabians Figure 10. This agrees with the report of Firouz (1998) about the origins of the Arabian and Turkoman horses. Also, the Davenport was isolated from the rest of the Arabians showing some differentiation likely because its low variability Figure 8, therefore, it was excluded from the second FCA shown in Figure 9. The FCA and the individual-animal-based dendogram showed close relationships among the Syrian populations Syrian registered and Syrian non-registered, where individuals hardly can be distinguished from each other. This is consistent with the common origin of these two populations, but with the additional diversity due to genetic introgression from non-Arabian horses into Syrian non-registered.

The outcomes from the FCA matched both the results of different phylogenic trees in our study, and the outcomes of the AMOVA, where it showed some separation between the Western and the Middle Eastern Arabian populations which was less evident in the phylogenic trees. The FCA showed a clear relationship between the Syrians horses and the Polish Arabian which had founders from the Middle East.

3.4.4 Population structure and individuals assignment

We did not include the non-Arabian populations in the STRUCTURE analysis because STRUCTURE works best with a small number of discrete populations (Pritchard*, et al.*, 2000). The STRUCTURE clustering using all Arabian populations did not give a clear mode for ΔK, Figure 12. Davenport population was excluded from the analysis because STRUCTURE algorithm assumes Hardy-Weinberg equilibrium within populations (Pritchard*, et al.*, 2000) and the use of a population with very low genetic variability like the Davenport may affect the STRUCTURE analysis and give no clear mode for ΔK (Vangestel*, et al.*, 2012). Thus, only nine Arabian populations (Syrian registered, Syrian non-registered, Polish Arabian, Iranian Arabian, Shagya Arabian, USA-Saudi, Saudi, Egyptian-Saudi mix, and USA-Egyptian) plus the out group were reported in our STRUCTURE analysis discussion.

The Bayesian clustering analysis at the optimal value of K confirmed the close relationship and the admixed structure in the Polish Arabian, Shagya Arabian, Iranian Arabian and Syrian non-registered that was suggested by both the FCA and pairwise F_{ST} test. A further STRUCTURE analysis, using only those four populations, showed isolation of the Polish Arabian from the rest Figure 14C, with a result of two different clusters; one formed by only the Polish Arabian horses and the second contained the Syrian non-registered, Shagya Arabian and Iranian Arabian. This suggests that those four populations have high levels of gene flow or share the same origin and have a recent divergence. Therefore, the Polish Arabian population is more differentiated from the Syrian non-registered than both the Shagya Arabian and Iranian Arabian populations. This outcome was not clear from the FCA or from other differentiation tests that were done; possibly because the clustering approach implemented in STRUCTRE can correctly infer the number of subpopulations in a dataset when genetic differentiation among groups is low (Latch*, et al.*, 2006).

The Bayesian clustering at K = 5 supported the result of the pairwise F_{ST} test, FCA, the individual-animal-based dendogram about the relationship between Syrian registered and Syrian non-registered, Figure 14A. This again confirms the similar origin and recent divergence of these two populations, as well as the high level of unidirectional gene flow (Syrian registered to Syrian non-registered) as a result of using some registered stallions in the reproduction of some non-registered animals.

The STRUCTURE analysis identified the American Western populations (Egyptian-Saudi mix, USA-Egyptian and USA-Saudi) as the most uniform. USA-Saudi was extremely homogeneous, probably due to the conservative breeding in this population which was descended from a limited number of founders. The STRUCTURE analysis showed similarity between USA-Saudi and most individuals from Saudi population. That confirmed the relationships between these two populations where both share a similar origin. Although Egyptian-Saudi mix together with USA-Egyptian formed another homogenous cluster at K = 5, further analysis using only Egyptian-Saudi mix and USA-Egyptian Figure 14B was able to discriminate between these two populations. The Egyptian-Saudi mix and USA-Egyptian share a similar pedigree background.

3.5 Conclusion

Overall, this work with the Middle Eastern and Western populations reveals a genetic structure of the Arabian horse breed not previously recognized and gives a comparative analysis of the Arabian populations from different origins. Genetic diversity was very high in Middle Eastern populations from Syria, Saudi Arabia and Iran. Some Western populations like the Polish-Arabian and Shaya-Arabian also have a high genetic diversity. In contrast, the Western American-Arabian showed less variability. Genetic differentiation was not strong among all Middle Eastern populations and the Polish-

Arabian and Shagya-Arabian populations, but the Western American-Arabians showed greater differentiation from these other groups and can be considered as uniform populations. The registered and non-registered Syrian populations were very close to each other but the later showed more diversity.

These results can facilitate conservation programs for this important breed, and enhance the effort to improve the management of Arabians to preserve the diversity found in the Middle Eastern Arabian populations. Furthermore, this study may encourage the Western Arabian horse breeders to expand the variability base of their lines, which has clearly been reduced, by introducing some new blood from the Middle Eastern populations. In addition to that, approaches used in this study can be applied to other domestic animals to discover their genetic diversity and population structure.

CHAPTER IV

MATERNAL PHYLOGENETIC RELATIONSHIPS AND GENETIC VARIATION AMONG ARABIAN HORSE POPULATIONS USING WHOLE MITOCHONDRIAL DNA D-LOOP SEQUENCING

4.1 Introduction

The traditional pattern of breeding Arabian horses affords special opportunities to evaluate variation in matrilineal markers, such as mitochondrial DNA. From a glance of historical records, the Arabian horse breed, in the desert, consists of five strains (*RASANs*) based upon dam lines: *Kahlila, Saklawia, Abiah, Shweemat, and Muanakii* (Hendricks, 1995) (some breeders and historians refer to an additional three *RASANs* which are *Hamadania, Dahmaa and Hadbaa*). The traditional breeders in the Middle East desert (*Bedouins*) have preserved the purity of the Arabian by avoiding any cross-breeding between the Arabians and non-Arabians and maintaining strictly separated *RASANs* (Pruski, 1983). Consequently, all individuals within a *RASAN* are expected to share the same maternal family line, and they should have similar mtDNA haplotype.

While many studies have been done in horses using mtDNA, only a few have included Arabians (Achilli*, et al.*, 2012; Bowling*, et al.*, 2000; Głażewska, 2010; Glazewska*, et al.*, 2007). Also, the Arabians used were mainly collected from Western populations. Most of the previous studies related to Arabian population genetics used only about 400 bp out of 1200 bp of the mtDNA D-loop. In the present study, we sequenced the whole mtDNA D-loop of Arabian horses collected from the Middle East as well as from Western populations. Our study was designed to investigate the maternal diversity and phylogenetic relationships of Arabian populations and to examine the traditional classification system of the Arabian breed (*RASANs* system) that depends upon maternal family lines.

4.2 Materials and methods

4.2.1 Sampling and DNA extraction

Hair samples were collected from 271 horses representing Middle Eastern Arabian, Western Arabian and non-Arabian populations. Tables 8 shows the number of animals used from each population.

Table 8: The tested populations and their abbreviations, the population groups and sample sizes.

Groups	Populations* (abbreviation)	Sample size
Middle Eastern Arabian	Syrian (SY)	114
	Saudi (SU2)	22
	Iranian (KA)	10
Western Arabian	USA-Egyptian (EG)	24
	USA- mix of Egyptian & Saudi (SE)	10
	USA-Saudi (SU1)	31
	Shagya Arabian (SA)	9
	Polish Arabian (PA)	13
	Davenport (DV)	19
non-Arabian	Mongolian (MON)	5
	Caspian (CS)	14

*descriptions about populations can be found in Table 2.

All tested horses were unrelated from the mothers' side for at least 3 generations at least based upon their pedigree records. Total genomic DNA was extracted from hair follicles using the PUREGENE® DNA purification kit following the manufacturer's instructions.

4.2.2 Whole D-loop sequencing and data analysis

We designed two pairs of primers based upon the horse mtDNA sequence reference X79547 (Xu and Arnason, 1994). We also considered the outcomes reported by Nergadze *et al*. (2010) to minimize the possible amplification of NUMTs that may overlap with D-loop. The designed primers were used to amplify the upstream part between sites 15440 and 16108 (Forward: 5′-AGCTCCACCATCAACACCCAAA-3′. Reverse 5’-CCATG GACTGAATAACACCTTATGGTTG-3′) and the downstream part between sites 16377 and 16642 (Forward 5′-ACCTACCCGCGCAGTAAGCAA-3′. Reverse 5′-AC GGGGGAAGAAGGGTTGACA-3′). Polymerase chain reactions were done for each part separately using the protocol described by Cothran*, et al.* (2005). A total of four sequencing reactions for each sample, including both strands in each part, were carried out using the BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, USA). Sequencing products were purified with the BigDye® XTerminator™ Purification Kit (Applied Biosystems, USA). DNA sequences were determined using the ABI 3130 xl Genetic Analyzer (Applied Biosystems, USA). Editing and aligning all sequences were carried out by MEGA 4 (Tamura*, et al.*, 2007) using the horse mtDNA sequence X79547 as a reference. Haplotype sequences included in this study were entered into the National Center for Biotechnology Information (NCBI) GenBank database available at http://www.ncbi.nlm.nih.gov/ with the accession numbers [NCBI: KC840701-KC840797]. The statistical quantities for the DNA sequences, including number of haplotypes and haplotype diversity and nucleotide diversity, were carried out using DnaSP 5.10.1 (Librado and Rozas, 2009). The statistical analysis was done for each population, as well as for each strain, using two sources of data HVR1, (450 sites) and whole D-loop sequences (951 sites).

Phylogenetic analysis of the haplotypes using a whole D-loop sequence was carried out with the PHYLIP software package (Felsenstein, 1989-2006) based upon the Kimura 2-parameter model to calculate genetic distances on the assumption of an equal substitution rate per site (Kimura, 1980). A consensus tree was also constructed with PHYLIP using the Neighbor-joining method (Saitou and Nei, 1987) with 1000 bootstrap repetitions. The donkey (*Equus* asinus) mtDNA sequence [NCBI: nc_001788] (Xu*, et al.*, 1996) was used as an out-group (Achilli*, et al.*, 2012; Vila*, et al.*, 2001).

Another approach for phylogenetic analysis was carried out by drawing the median-joining network (MJ network) (Bandelt*, et al.*, 1995) in accordance with the haplotype sequences of the whole D-loop using the NETWORK 4.6.1 software (available at http://fluxus-engineering.com). Default settings were applied ($r = 2$, $\varepsilon = 0$) (Jansen*, et al.*, 2002), and preliminary trials were done in order to determine the mutational hotspots. Four mutational hot spots were excluded and an additional four were down-weighted into 0.5 (Cieslak*, et al.*, 2010; Jansen*, et al.*, 2002). Each haplotype in the MJ network was shown by color codes representing the proportions of different strains (or populations) depending on the individual frequencies in each haplotype. Furthermore, the haplotype sequences were compared to the NCBI database using the BLAST search as implemented in MEGA 4, and haplogroups were named as defined by Achilli*, et al.* (2012).

To represent the genetic structure and differentiation of tested populations, principal coordinate analysis (PCoA), analysis of molecular variance (AMOVA) and pair-wise F_{ST} were applied. PCoA of the dissimilarity matrix according to Kimura (1980) based upon 951 bp of the 98 haplotypes sequences was carried out using DARwin 5.0 (Perrier*, et al.*, 2003; Perrier and Jacquemoud-Collet, 2006). AMOVA and pair-wise F_{ST} were done using the Kimura 2-parameter model (Kimura, 1980) with 1000 permutations and were carried out with Arlequin 3 (Excoffier*, et al.*, 2005). For the Pair-wise F_{ST} results, we followed the suggestion that refers that a value between 0–0.05 indicates little genetic differentiation; a value between 0.05 and 0.15, moderate differentiation; a value between 0.15 and 0.25, great differentiation; and values above 0.25, very great genetic differentiation (Hartl and Clark, 1997; Wright, 1978).

4.3 Results

Table 9 shows the diversity measures for populations including number of haplotypes (NHap), haplotype diversity (HapD), average number of nucleotide differences (k), the number of polymorphic sites (NPS) and nucleotide diversity (π) for each population. The results were shown for the HVR1 and the whole D-Loop separately.

Table 9: Diversity measures for populations tested in the study. N: number of individuals in each population. NHap: the number of haplotypes resulted in each population. HapD: haplotype diversity with its standard deviation. NPS: the number of polymorphic sites. π: nucleotide diversity with its standard deviation. k: average number of nucleotide differences. HVR1: part of the upstream D-loop (450 sites). W: the whole D-loop (951 sites).

Populations	N	NHap		HapD (SD)		NPS		π (SD)		k
		HVR1	W	HVR1	W	HVR1	W	HVR1	W	
SY	114	43	50	0.96(0.007)	0.97 (0.007)	44	69	0.0196(0.0006)	0.0142(0.0004)	8.6
SU2	22	10	10	0.84(0.06)	0.84 (0.06)	36	50	0.0192(0.0023)	0.0129(0.0015)	8.5
KA	10	8	8	0.96(0.06)	0.96 (0.06)	29	42	0.023(0.0019)	0.0153(0.0013)	10.2
EG	24	9	9	0.83(0.06)	0.83 (0.06)	26	38	0.019(0.0018)	0.0128(0.0012)	8.5
SE	10	4	5	0.79(0.09)	0.84 (0.08)	19	26	0.0199(0.0027)	0.0126(0.0015)	8.8
SU1	31	7	7	0.8(0.042)	0.8 (0.042)	34	51	0.0223(0.0016)	0.015(0.001)	9.9
SA	9	8	8	0.97(0.06)	0.97 (0.06)	30	41	0.0234(0.002)	0.0153(0.0018)	10.3
PA	13	6	6	0.82(0.08)	0.82 (0.08)	25	42	0.0213(0.002)	0.0163(0.0017)	9.4
DV	19	6	6	0.74(0.083)	0.74 (0.083)	26	36	0.020(0.0023)	0.01281(0.0016)	8.9
MON	5	5	5	1(0.12)	1 (0.12)	19	28	0.0195(0.0038)	0.013(0.0027)	8.6
CS	14	9	9	0.93(0.045)	0.93 (0.045)	35	54	0.023(0.0022)	0.017(0.0013	10.2

A total of 74 haplotypes from 60 polymorphic sites were found in 271 horses from 11 populations by using the HVR1. NHap increased to 97 using the whole D-loop sequences. Although π decreased from 0.022 to 0.015, NPS increased from 60 to 99 and k increased from 9.7 to 14.5 comparing to of the HVR1 to the whole D-loop, respectively Table 9. The highest HapD values among all tested Arabian populations were in SY, SA and KA (0.97, 0.97, 0.96), respectively. The non-Arabian populations also showed high values of HapD (1.0 in MON and 0.93 in CS). All American-Arabian populations (SU1, EG, SE and DV) showed relatively low HapD ranging between 0.74 and 0.83.

The tested samples were then grouped into strains according to pedigree records and regardless of their populations. We could assign 191 out of 271 samples into seven strains (*RASANs*). As shown in Table 10, a total of 44 haplotypes from 52 polymorphic sites were found in these 191 horses of the seven strains using the HVR1 part of the D-loop. The NHap increased to 55 using the whole D-loop sequences. Only the *Shweemat* strain had all individuals with a single haplotype. *Hadbaa* and *Dahmaa* also had low NHap (3 and 2, respectively). *Kahlila* was the most variable strain showing 26 haplotypes. The total NHap calculated from all individuals together (NHap = 55) was less than the sum of NHap calculated from each strain separately due to some shared haplotypes among strains.

Table 10: Diversity measures for strains (*RASANs*) tested in the study. N: number of individuals in each strain. NHap: the number of haplotypes resulted in each strain. HD: haplotype diversity with its standard deviation. NPS: the number of polymorphic sites. π: nucleotide diversity with its standard deviation. k: average number of nucleotide differences. HVR1: part of the upstream D-loop (450 sites). W: the whole D-loop (951 sites).

Strain (abbreviation)	N	NHap		HapD (SD)		NPS		π (SD)		π (SD)	k
		HVR1	W	HVR1	W	HVR1	W	HVR1	W		
Kahlila (K)	44	22	26	0.94 (0.022)	0.95 (0.022)	41	61	0.023(0.0006)	0.0149(0.0005)	9.9	14.2
Hamadania (H)	61	12	14	0.87 (0.019)	0.88 (0.02)	32	52	0.019(0.0009)	0.0148(0.0006)	8.4	14.1
Hadbaa (HD)	7	3	3	0.76 (0.115)	0.76 (0.115)	17	24	0.019(0.003)	0.0125(0.002)	8.5	12.0
Dahmaa (D)	7	2	2	0.57 (0.119)	0.57 (0.119)	11	19	0.014 (.0029)	0.0113(0.0023)	6.3	10.8
Saklawia (S)	43	14	15	0.92 (0.019)	0.92 (0.02)	33	48	0.019(0.001)	0.0127(0.0006)	8.4	12.1
Abiah (A)	24	10	10	0.85 (0.053)	0.85 (0.053)	28	39	0.019(0.0015)	0.012(0.001)	8.7	11.7
Shweemat (SH)	5	1	1	0	0	0	0	0	0	0	0
all	191	44	55	0.96 (0.004)	0.97 (0.004)	52	81	0.0218(0.0006)	0.0151(0.0004)	9.9	14.4

A consensus Neighbor-joining tree of the 97 haplotypes found from all tested populations is presented in Figure 15. No single population was found only in one cluster and different populations shared haplotypes. Fifteen haplotypes (24, 15, 29, 30, 44, 26, 11, 22, 33, 14, 23, 55, 16, 4, 28) were found in at least two Arabian populations; for example, haplotype 24 was found in two populations (SU1and EG); haplotype 14 appeared in five populations (SU1, SY, EG, SE and PA). In addition, haplotype 4 was found in Arabian and non-Arabian populations (SY and KA with CS). The dendrogram gave seven main clades plus the out-group. SY population was the most variable among all populations with individuals found in all clades. Figure 16 shows the consensus Neighbor-joining tree of the 55 haplotypes found in the individuals who were assigned to their strains. None of the tested strains, except SH, was represented by a single haplotype or phylogenetically close haplotypes. Each of the thirteen haplotypes (16, 23, 22, 12, 27, 14, 29, 15, 18, 75, 11, 74 and 17) was found in at least two strains. For example, haplotype 16 was present in two strains (H and D) and haplotype 23 in three strains (A, K and H). The most frequent mixing was noticed between S and K strains. The K strain was the most variable among all strains and its individuals were distributed among all clades.

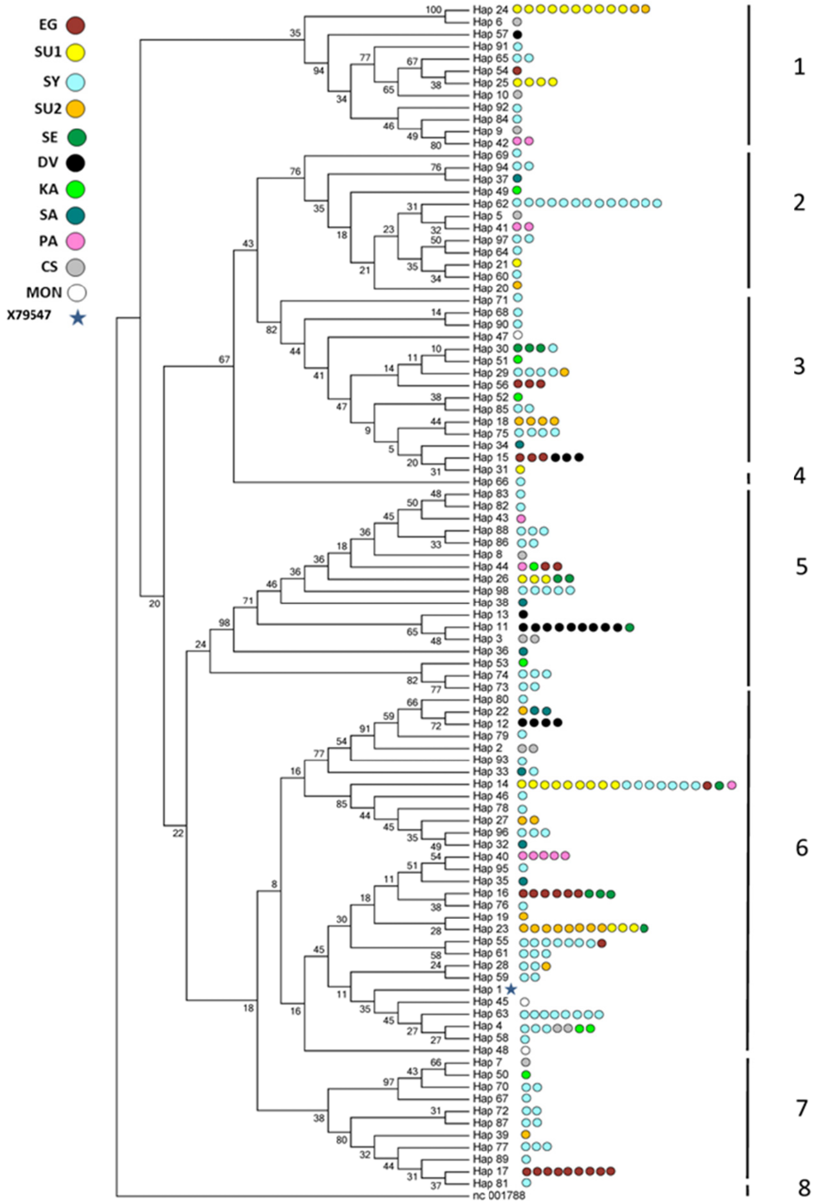


Figure 15: Consensus neighbor-joining tree of the 97 haplotypes found. The tree was drawn based upon 1000 bootstrap replicates. The reference donkey sequence nc 0017788 was used as an out-group. Bootstrap values are shown as percentages. The individuals with each haplotype are represented by colored circles depending on populations. Population abbreviations are as follow: Syrian (SY), Saudi (SU2), Iranian (KA), USA-Egyptian (EG), USA- mix of Egyptian & Saudi (SE), USA-Saudi (SU1), Shagya Arabian (SA), Polish Arabian (PA), Davenport (DV), Mongolian (MON), Caspian (CS). X79547 is the reference horse sequence

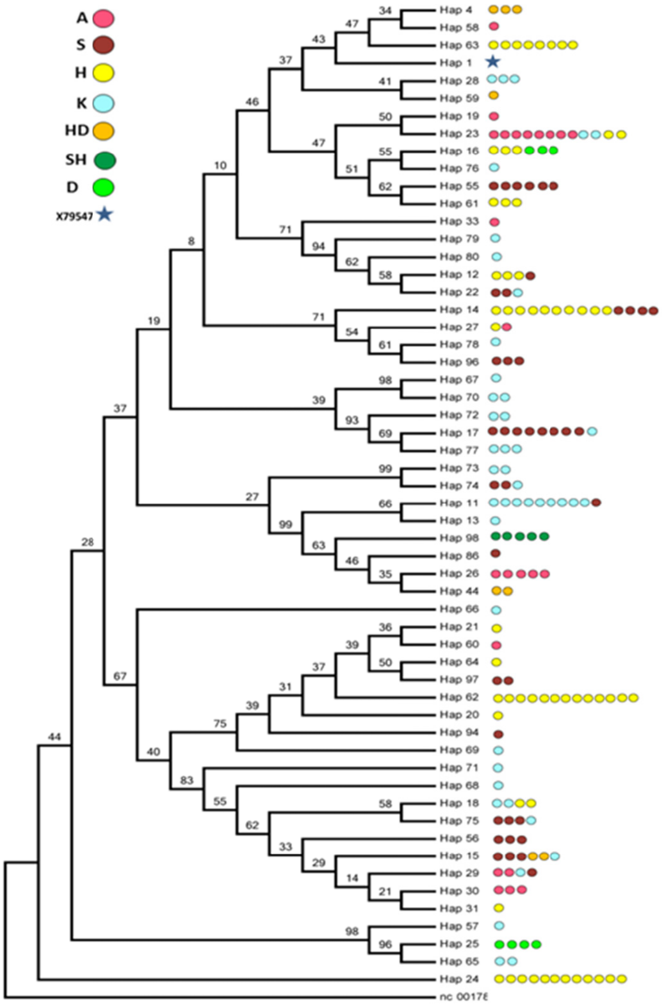


Figure 16: Consensus neighbor-joining tree of the 55 haplotypes found in strains. The tree was drawn based upon 1000 bootstrap replicates. The reference donkey sequence nc 0017788 was used as an out-group. Bootstrap values are shown as percentages. The individuals with each haplotype are represented by colored circles depending on strains. Strain abbreviations are as follow: *Kahlila* (K), *Hamadania* (H), *Hadbaa* (HD), *Dahmaa* (D), *Saklawia* (S), *Abiah* (A), *Shweemat* (SH). X79547 is the reference horse sequence

After taking into account mutational hot spots for the median-joining network (MJ network), the number of haplotypes dropped from 97 to 86. Figure 17 and Figure 18 show the MJ network based on 951 bp of the D-loop representing 271 samples by 86 haplotypes.

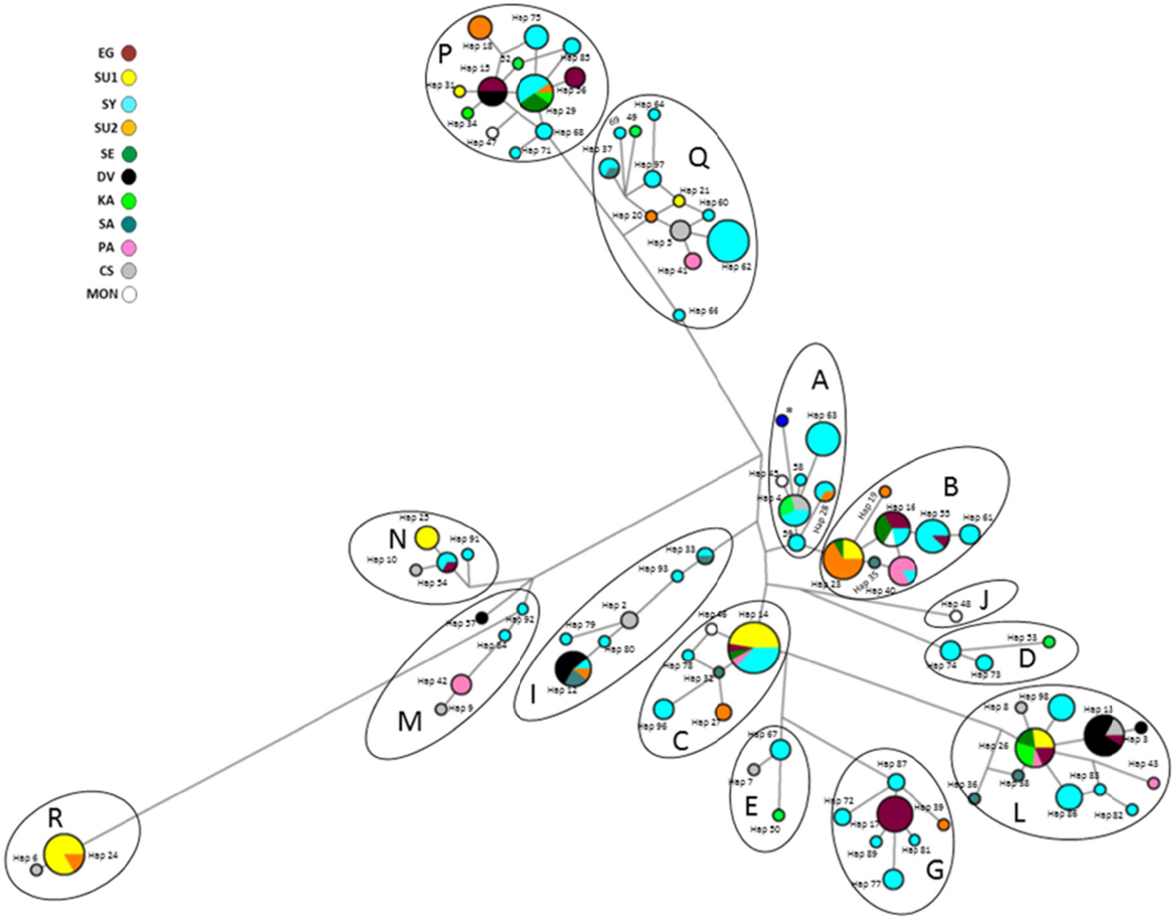


Figure 17: The median-joining network for populations. Analysis was done using 951 bp of the mitochondrial D-Loop and 272 horses within 87 haplotypes. Mutational hot spots were taken into account according to Cieslak et al 2010 and Jansen et al. 2002. The haplogroups were named as defined by Achilli et al 2012. Each population is shown by color and the proportions of different populations for each haplotype were shown. Population abbreviations are as follow: Syrian (SY), Saudi (SU2), Iranian (KA), USA-Egyptian (EG), USA- mix of Egyptian & Saudi (SE), USA-Saudi (SU1), Shagya Arabian (SA), Polish Arabian (PA), Davenport (DV), Mongolian (MON), Caspian (CS). The reference sample X79547 was labeled with a star.

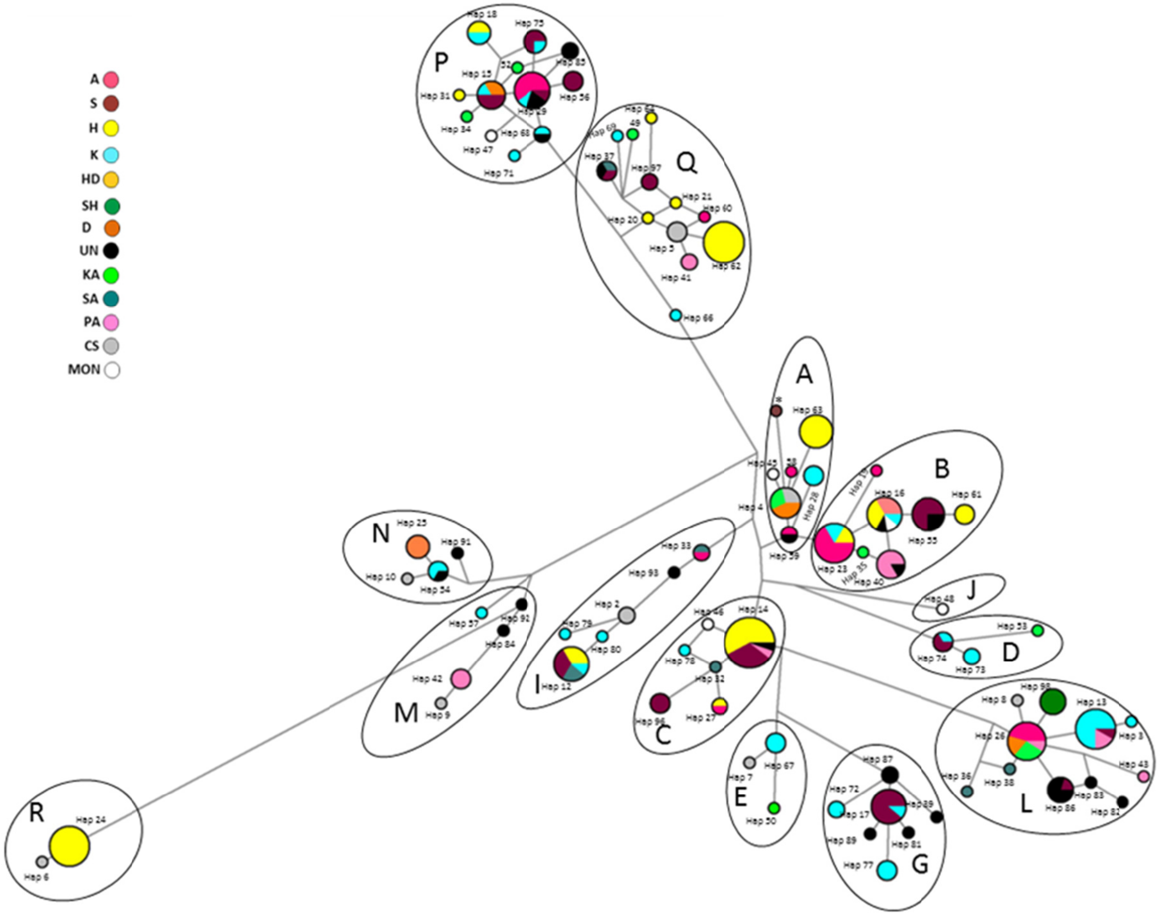


Figure 18: The median-joining network for strains. Analysis was done using 951 bp of the mitochondrial D-Loop and 272 horses within 87 haplotypes. Mutational hot spots were taken into account according to Cieslak et al 2010 and Jansen et al. 2002. The haplogroups were named as defined by Achilli et al. 2012 and mentioned as letters next to each haplogroup. Each strain was shown by color and the proportions of different strains for each haplotype were shown. Strain abbreviations are as follow: *Kahlila* (K), *Hamadania* (H), *Hadbaa* (HD), *Dahmaa* (D), *Saklawia* (S), *Abiah* (A), *Shweemat* (SH). Samples with unknown strain were represented by black. The reference sample X79547 was labeled with a star.

While in Figure 17 each haplotype is shown by the proportion of the different populations included in this haplotype, in Figure 18 each haplotype is shown by the proportion of different strains. The MJ networks showed 14 haplogroups (A, B, C, D, E, G, I, J, L, M, N, P, Q and R) as defined by Achilli*, et al.* (2012).

As shown in Figure 17, each of the 13 haplogroups (A, B, C, D, E, G, I, L, M, N, P, Q and R) contained identical or very close haplotypes from at least two populations. The highest number of populations was found in the haplogroup L. The Arabian populations were represented in all haplogroups except J. The non-Arabian samples were placed in the haplogroups (A, E, I, L, M, N, Q and R) and (A, B, C, J, and P) for the CS and the MON populations, respectively. SY population was the most variable with individuals distributed across all haplogroups except J and R. Individuals from SY had identical or very close haplotypes to individuals from all other Arabian and non-Arabian populations. The DV was the least variable Arabian population with only three haplogroups (I, L and P).

Figure 18 showed that individuals from different strains shared a single haplotype. Identical matching between two or more individuals from different strains was seen in 13 cases. Also, matching was found between known strains and other Arabian groups (PA, SA and KA) and non-Arabian populations (CS and MON). In addition, individuals from a single strain were found in distinctive haplogroups (for example: strain H in haplogroups P, C and R). The K strain was the most variable with individuals distributed across all haplogroups except J and R. All of the unknown-strain samples were identical or very close to samples of known strains. Although the SH strain was the least variable with only haplogroup L, it was very close to samples from some other strains.

Figure 19 shows the PCoA plot of the two first axes which explain 26.52 % and 18.77 % of the variability, respectively.

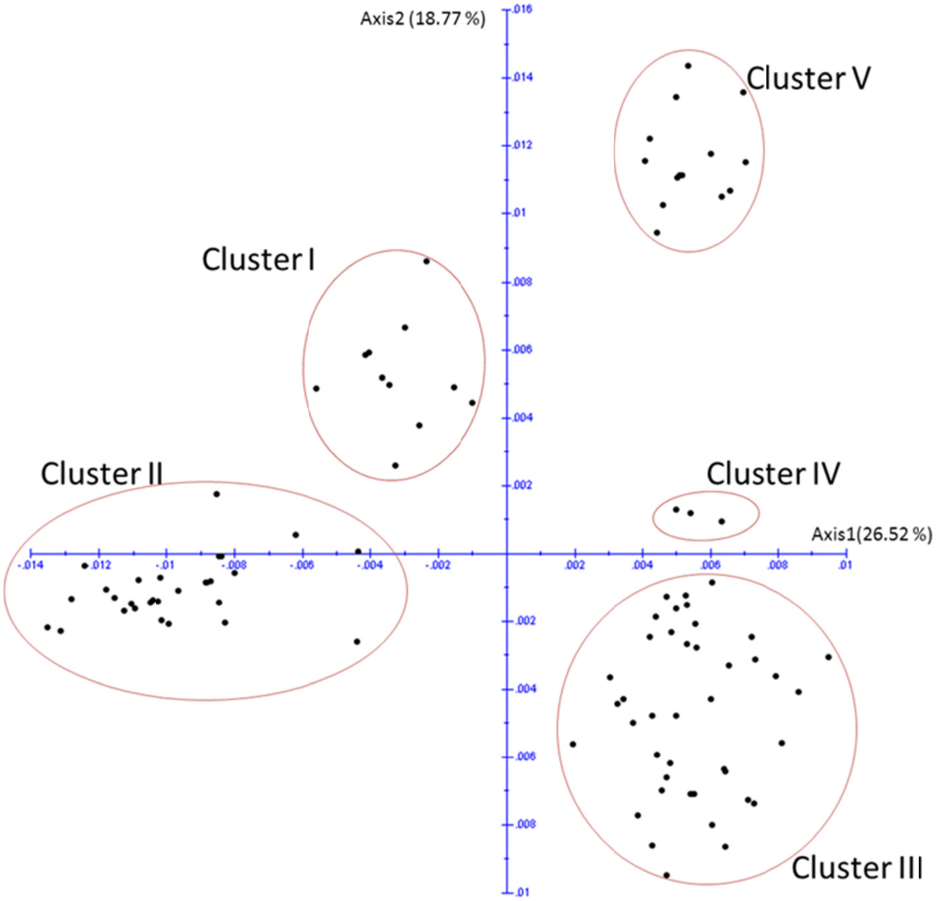


Figure 19: Plot of the two first axes of the principal coordinates analysis (PCoA). It was drown based upon the dissimilarity matrix according to Kimura (1980) based on 951 bp of the 98 haplotype sequences and carried out by using DARwin 5.0 software. Cluster I includes haplogroups M, N and R. Cluster II includes haplogroups P and Q. Cluster III includes A, B, C, E, G, I, and J. Cluster IV includes haplogroup D. Cluster V includes haplogroup L.

The PCoA plot grouped the 98 haplotypes into five clusters (Figure 19). Cluster I included a combination of three haplogroups (M, N and R). Cluster II consisted of two haplogroups (P and Q). Cluster III included seven haplogroups (A, B, C, E, G, I, and J). Cluster IV had only haplogroup D, and Cluster V included only haplogroup L. The clustering by PCoA did not show any differentiation among haplotypes that came from different populations (or different strains) but it showed that each cluster contained a mixture of individuals that represented different populations (or strains).

AMOVA showed that the proportion of the variation among populations was 8.25 % and the frequency of the variation within populations was 91.75 %. The fixation index was equal to 0.083.

The pairwise F_{ST} values are shown in Table 11.

Table 11: Pairwise F_{ST} values among populations. Negative values equate to zero.

Populations	CS	DV	EG	SE	SU2	SU1	SA	PA	MON	KA
DV	0.066									
EG	0.092	0.210								
SE	0.025	0.123	0.024							
SU2	0.076	0.243	0.058	0.033						
SU1	0.057	0.215	0.171	0.125	0.111					
SA	-0.015	0.069	0.066	-0.003	0.051	0.109				
PA	-0.004	0.155	0.092	0.007	0.072	0.104	0.032			
MON	0.027	0.209	0.014	-0.041	-0.040	0.093	-0.017	-0.011		
KA	-0.016	0.045	0.055	-0.029	0.092	0.113	-0.021	0.039	0.029	
SY	0.011	0.149	0.050	-0.001	0.050	0.125	0.015	0.034	-0.009	0.008

Out of 55 pairwise F_{ST} values 28 comparisons had F_{ST} values between 0 and 0.05 showing little genetic differentiation while 21 comparisons had Fst values between 0.05 and 0.15 showing moderate genetic differentiation, six comparisons had F_{ST} values between 0.15 and 0.25 showing great genetic differentiation. Negative F_{ST} values were recorded in some comparisons and these equate to zero F_{ST} values. While most of the lowest F_{ST} values were seen between SY and eight other populations (CS, EG, SE, SU2, SA, PA, MON and KA), the highest F_{ST} values were between the DV and five other populations (EG, SU2, SU1, PA and MON). None of the comparisons showed values corresponding to very great genetic differentiation.

4.4 Discussion

This study presents the first description of maternal genetic diversity based upon the whole mtDNA D-loop of native Arabian horses sampled from Syria, Iran and Saudi Arabia, as well as of Western Arabian populations. One of the unique aspects of this study is the inclusion of the traditional classification system (*RASANs* or strains system) of native Arabians.

4.4.1 HVR1 and the whole mtDNA D-loop comparison

Most previous maternal diversity studies of horses are based upon sequencing of the HVR1 (Bowling, *et al.*, 2000; Cieslak, *et al.*, 2010; Cothran, *et al.*, 2005; Cozzi, *et al.*, 2004; Georgescu, *et al.*, 2011; Glazewska, *et al.*, 2007; Guastella, *et al.*, 2011; Ivankovic, *et al.*, 2009; Jansen, *et al.*, 2002; Prystupa, *et al.*, 2012a). Our results of the comparison between HVR1 and the entire mtDNA D-loop showed that the variability in the upstream region of the D-loop revealed differences among 22 additional haplotypes which had identical sequences in the HVR1. This agrees with Kavar, *et al.* (1999) where such a pattern of variability had been found in the D-loop of the Lipizzan horse breed. Higher haplotype diversity (HapD = 0.98) and average number of nucleotide differences (k = 14.5) were found by using the whole mtDNA D-loop compared with the HVR1 (HapD = 0.98 and k = 9.5) (Table 9). Thus, using the whole mtDNA D-loop is more robust and powerful than using the HVR1 alone for analysis of genetic diversity of the mtDNA in horses. Similar results have been reported in goats (Kang*, et al.*, 2011).

4.4.2 Population genetic diversity

Maternal genetic diversity of the Arabian populations described in this study was similar to that reported in some other breeds (Cai*, et al.*, 2009; Guastella*, et al.*, 2011). Although SY, SA and KA populations had equally high HapD values (Table 9), the SY population was the most variable based on the consensus Neighbor-joining tree (Figure 15) where the SY individuals were found in eight clades compared to the KA and SA individuals found only in five and three clades, respectively. This result was also supported by the MJ-network (Figure 17) where the SY population was represented in 12 haplogroups compared to KA and SA with six and five haplogroups, respectively. According to Achilli*, et al.* (2012) there is a total of 18 major haplogroups of horses throughout Asia, Middle East, Europe and America; our results showed that SY population covers 12 of the 18 haplogroups showing extensive maternal genetic diversity. In our opinion, which is supported by results of (Cieslak*, et al.*, 2010), the huge diversity of SY population is not a consequence of recent animal breeding or outcrossing but instead a feature that was already present in this very old population. In addition, the huge diversity in the Arabian populations is consistent with the multiple origins in the maternal lineages of domestic horse breeds reported by other studies (Aberle*, et al.*, 2007; Cieslak*, et al.*, 2010; Georgescu*, et al.*, 2011; Jansen*, et al.*, 2002).

Some of the SY individuals were represented in haplogroup D (Figure 17), haplogroup E according to Jansen*, et al.* (2002), that was reported as a very rare and old haplogroup which may date back as far as Bronze age (Cieslak*, et al.*, 2010; Kakoi*, et al.*, 2007; Prystupa*, et al.*, 2012a).

The American-Arabian populations showed relatively low HapD values and were represented in a limited number of haplogroups. DV was the least variable with only three haplogroups (I, L and P). The low maternal diversity found in the American-Arabian populations is probably due to the founder effect. This result is supported by our previous work done by using microsatellite markers where American-Arabian populations showed less genetic variability compared with Middle Eastern populations (Khanshour*, et al.*, 2013). Also, PA did not show a very high genetic diversity with only 6 haplotypes distributed in four haplogroups. This result did not match with Glazewska*, et al.* (2007) where 14 distinct haplotypes were reported. This could be due to sample size or because the horses we used came from close maternal lines.

4.4.3 Population relationships and genetic structure

The low bootstrap values of the Neighbor-joining trees in Figure 15 and Figure 16 are primarily due to the overall high degree of relationship among horses (Cothran and Luis, 2005). Low bootstrap values have been reported in many mtDNA studies in horses (Cozzi*, et al.*, 2004; Georgescu*, et al.*, 2011; Kim*, et al.*, 1999; Lippold*, et al.*, 2011; Vila*, et al.*, 2001). Although bootstrap values were low, the populations consistently fell into the same groupings in the trees. The consensus Neighbor-joining tree (Figure 15) and the MJ-network (Figure 17) show that individuals from different populations share identical haplotypes. This indicates possible gene flow among those populations or common ancestry. Identical maternal lines were found between SY and PA populations revealing that Syrian mares were probably part of Polish Arabian founders, or some horses were recently introduced to this population. The identical maternal lines that were found between the American Arabian populations (SU1, SE, EG) and populations from the Middle East (SY, SU2 and KA) confirms that the current registered Arabian horses in America have been primarily founded by mares exported from the Middle East (Bowling*, et al.*, 2000). While SA population is thought to be descended from a Syrian stallion (Hendricks, 1995), our results show some shared maternal lines between SA and SY suggesting a maternal contribution of Syrian horses in SA population, or possibly recent gene flow between these two populations. Furthermore, the phylogenetic analysis revealed that different populations, including Arabian and non-Arabian, often had very close haplotypes, and none of these populations formed a distinct clade. These results together reveal the mixed origin and/or a likely common ancestor of these populations. The genetic clustering analysis using both phylogenic (Figure 15 and 17) and PCoA (Figure 19) did not show any clear pattern of differentiation among all populations. Haplotypes within a population were found in separate haplogroups. Similar results have been reported in other studies of horse mtDNA (Cothran*, et al.*, 2005; Jansen*, et al.*, 2002; Vila*, et al.*, 2001). F_{ST} analysis supports this unclear pattern of differentiation showing high rates of mtDNA sharing between populations. Negative F_{ST} values sometimes are produced by software which uses algorithms that include sampling error corrections, such as Arlequin, when the true

Fst values are close to zero (Musick, 2005), and usually appear when there are great differences between two random individuals from same population rather than between two random individuals from different populations (Arnason and Palsson, 1996). These negative values represent program idiosyncrasies and are effectively zero (Humphries and Winker, 2011) indicating no differentiation among the compared populations in the present study. AMOVA results also support within group variation with 91.75% of variability as within population variation.

4.4.4 Strain relationships and classification system

In the Middle East, strain breeding is still an important factor in the Arabian horse breed (Hendricks, 1995). According to Bedouin breeding traditions, Arabian horses were subdivided into strains depending on the maternal lineage. The phylogenetic and principle coordinate analyses in our study using 191 samples, of known strains, showed no evidence that the Arabian breed has clear divisions based upon traditional strain classification. There are four points that support this finding. First, 13 cases revealed that individuals from different strains shared a single haplotype. For example, haplotype 23 was found in individuals that came from three different strains (*Abiah, Kahlila* and *Hamadania*); haplotype 29 was in individuals from three strains (*Abiah, Kahlila* and *Saklawia*) (Figure 16). Second, individuals from different strains were found in a single haplogroup. For example, haplogroup P was seen in five strains (*Kahlila, Saklawia, Abiah, Hadbaa,* and *Hamadania*) (Figure 18). Third, each of the strains (*Kahlila, Saklawia, Abiah, Dahmaa, Hadbaa* and *Hamadania*) was represented in clearly separated haplogroups. For example, *Kahlila* was found in 12 haplogroups

(Figure 18). Finally, PCoA did not show any pattern of clustering that fits strains subdivision (Figure 19). Our results agree with the conclusion reported by Bowling et al. (2000) about American Arabian horses.

It is possible to have some minor mistakes in the pedigree records of any breed (Hill*, et al.*, 2002), but with our results we can confirm that these mistakes, if they existed in the records that we used, cannot be the reason behind having the huge admixture among tested strains. We do not suspect admixture into the Arabian horse breed, but it is clear that the pedigree records of the Arabian breed were not built using robust genetic tools that can recognize distinct maternal lines in the establishment of the pedigree.

Another important factor in the Bedouin breeding traditions is the sub-strain subdivisions (*MARBATT*) that subdivides each Arabian strain into related groups depending on the tribe's or owner's name (Hendricks, 1995). Although we did not test the sub-strain subdivisions of Arabians in our study because of a lack of information, we can say that the sub-strain system might be able to partially explain the third point mentioned above (related to the differences among individuals from same strain), but it does not answer the other questions.

5.5 Conclusion

The maternal phylogenetic analysis of native Arabian horses in our study revealed 1- That the analysis based upon the whole mtDNA D-loop sequence was more powerful to study the genetic diversity in Arabian horses than using just the HVR1. 2- That the maternal genetic diversity was wide in the Arabian horse populations especially in the Syrian population. 3- That there was no clear pattern of differentiation among all tested populations. 4-That the Syrian mares probably had maternal contributions to the Polish Arabian and Shagya Arabian populations. 5-That the current registered Arabian horses in America have been primarily founded by mares exported from the Middle East. 6. Most importantly, that there was no evidence, using mtDNA D-Lopp, that the Arabian breed has clear subdivisions depending on the traditional strain classification system.

CHAPTER V

MICROSATELLITE ANALYSIS FOR PARENTAGE TESTING OF THE ARABIAN HORSE BREED FROM SYRIA[2]

5.1 Introduction

In horses, parentage testing has been of particular importance for breed registration processes, studbook creation and validation. In general, parentage testing in animals is important for checking the genetic accuracy of progeny testing, in selection for traits (Jamieson and Taylor, 1997), while accurate pedigree information is important for a successful animal breeding program (Ozkan*, et al.*, 2009) and for conservation of animal populations (Sereno*, et al.*, 2008).

The Arabian breed might be expected to have a high level of homozygosity, because of the way in the (*Bedouins*) have conserved this breed by inbreeding and avoiding crossing to other breeds or horses of uncertain origins (Upton and Amirsadeghi, 1998). This manner of breeding becomes problematic in small populations, especially when the effects of natural selection are negated by inbreeding far away from the desert conditions under which the Arabian horses developed. In such as circumstances, the use of a set of highly polymorphic markers is required for reliable parentage testing.

[2] Reprinted from Khanshour A., Conant E., Juras R., Cothran G. (2013). Microsatellite analysis for parentage testing of the Arabian horse breed from Syria. Turkish Journal of Veterinary and Animal Sciences. 37: 9-14

Blood group and protein polymorphism tests were used for nearly three decades for horse pedigree records and successfully resolved queries of parentage in most cases (Bowling, *et al.*, 1997).

Recently DNA-based methodologies for genetic marker-testing using polymerase chain reaction (PCR) technology provided a more powerful alternative to blood typing, particularly the analysis of short tandem repeat loci (STRs or microsatellites) (Bowling, *et al.*, 1993). The purpose of this study was to determine if a panel of 16 STR markers was sufficient to validate parentage for Arabian horses collected directly from local breeders from Syria. This is the first in depth study of the Arabian horse breed originating from the Arabian Desert which may more closely reflect the original status of the genetic structure of the Arabian horse breed. I conducted this part of the study before the microsatellites and the maternal diversity studies of the Arabian populations mentioned in chapter III and IV. So when I did this part there was no information about the genetic diversity of the Syrian horses.

5.2 Materials and methods

5.2.1 Sampling and DNA extraction

Ninety-four hair samples were collected from different regions of Syria, including the government breeding center of the Arabian horses. Forty-nine samples were from non-registered horses, while the remaining 45 consisted of horses from all the registered groups of the Arabian horses in Syria. The animals from different *RASANs* were pooled for this analysis. Total DNA was extracted from the hair follicles using PUREGENE® DNA purification kit following the manufacturer's instructions.

5.2.2 Microsatellite analysis

Fifteen microsatellite markers (Table 3), specific to *Equus* caballus, were used in this study. All are recommended by the International Society for Animal Genetics, and one X chromosome marker, LEX3, was also typed. The 16 microsatellites were amplified in three multiplex reactions as follows: (8plex: AHT4, HT5, ASB17, ASB23, HMS6, HMS7, HTG4 and VHL20. 5plex: LEX3, HMS3, ASB2, HTG10 and LEX3. 3plex: HMS2, HTG6 and HTG7). Each reaction had a final volume of 12 µl, containing 50 ng of genomic DNA, from 0.07 to 0.8 pmol of primers, 1xPCR buffer, 2.5 mM MgCl2, 0.2 mM dNTPs, and 1 U AmpliTaq for the 8plex, while for the 3 and 5plex 1 U ChoiceTaq was used. For microsatellite amplification a hot start procedure was used, in which the genomic DNA and primers were combined and heated at 95 °C for 5 min. The temperature was then lowered and held at 85 °C for 10 min for the addition of the remaining reagents. Thirty five cycles were as follows: 95 °C for 1 minute, either 56 °C (5plex) or 60 °C (for 8plex) for 30 second and 72 °C for I minute annealing. The cycling was completed with a final extension at 72 °C for 15 minutes. The PCR products were separated by electrophoresis on a 6% polyacrylamide gel using the ABI PRISM 377 DNA Sequencer (Applied Biosystems, Foster City, CA, USA). Fragment sizes of microsatellite alleles were determined using the STRand computer software (Locke*, et al.*, 2000). Alphanumerical nomenclature was used for allele size designation in accordance with the International Society for Animal Genetics. All the tests were repeated at least three times, and both positive and negative controls were used in each reaction.

5.2.3 Statistical analyses

Standard diversity indices were calculated using Cervus 3.0 (Marshall*, et al.*, 1998). These include: the number of alleles (Na), number of effective alleles (Ne= 1/(1-He)), observed (Ho) and expected (He) heterozygosity (calculated from allele frequencies assuming Hardy-Weinberg equilibrium), polymorphic information content (PIC) which is a measure of informativeness related to expected heterozygosity (Botstein*, et al.*, 1980), frequency of the most common allele (FNA), probability of exclusion (PE) and combined probabilities of exclusion (CPE) (Jamieson and Taylor, 1997).

5.3 Results

PCR amplicons ranged between 93 base pair (bp) and 211 bp in size. Table 12 shows the standard diversity indices. The total number of alleles was 91 in the registered group, with a mean of 5.7 per locus, and 123 alleles in non- registered with a mean of 7.7. The number of alleles per locus ranged between 3 for HTG6 and HTG7 to 8 for ASB2 for the registered and in the non-registered group Na ranged between 4 for HTG7 to 14 for ASB17. Number of effective alleles (Ne) varied between 1.86 for HTG7 to 5.464 for ASB17 in the registered group, and between 2.141 for HTG7 to 5.988 for ASB17 in the non-registered. The mean Ne was 3.747 in the registered group and 4.476 in non-registered. Observed heterozygosity per locus in the registered group varied from 0.36 for HTG7 to 0.91 for LEX33 and from 0.47 for HTG7 to 0.88 for ASB17 for the non-registered with means of 0.69 and 0.71, respectively. The lowest value of PIC for both groups was for HTG7 (0.358 in the registered group and 0.469 in non-registered), while the highest value was for ASB17 (0.781 in the registered group and 0.803 in non-registered). The mean PIC was 0.657 in the registered group and 0.715 in non-registered. The individual probability of exclusion ranged from 32% in HTG7 locus to 80% in ASB17 for registered and 41% in HTG7 locus to 84% in ASB17 for the non-registered. The combined probability of exclusion (CPE) for all loci was more than 99.99% in each group. Figure 20 shows the CPE values for both groups as a function of the number of microsatellite loci.

Table 12: The standard diversity indices of tested loci. Number of alleles (Na), number of effective alleles (Ne), observed (Ho) and expected (He) heterozygosity, polymorphic information content (PIC) probability of exclusion (PE) and Combined probabilities of exclusion (CPE) for Registered (Reg) and Non-registered (Non-reg) Arabian horses.

Group	Registered						Non-registered					
locus	Na	Ne	Ho	He	PIC	PE	Na	Ne	Ho	He	PIC	PE
ASB17	7	5.4	0.87	0.817	0.781	0.800	14	5.9	0.88	0.833	0.803	0.839
HMS2	7	5.2	0.87	0.809	0.771	0.789	9	5.8	0.47	0.830	0.799	0.831
LEX3	7	4.7	0.43	0.790	0.748	0.755	9	5.7	0.84	0.825	0.791	0.814
ASB23	6	4.2	0.70	0.767	0.723	0.731	8	5.0	0.73	0.803	0.766	0.785
ASB2	8	4.0	0.77	0.754	0.707	0.713	9	5.0	0.84	0.800	0.763	0.784
HMS7	6	4.2	0.79	0.766	0.717	0.709	8	5.1	0.67	0.805	0.766	0.781
HMS3	6	4.0	0.81	0.751	0.701	0.697	7	5.0	0.71	0.803	0.762	0.769
HTG10	5	3.9	0.77	0.749	0.698	0.689	8	4.5	0.69	0.782	0.743	0.762
LEX33	5	3.8	0.91	0.742	0.690	0.675	9	4.5	0.8	0.780	0.737	0.747
VHL20	5	3.4	0.74	0.709	0.657	0.650	8	4.4	0.71	0.774	0.732	0.742
HMS6	5	3.2	0.70	0.696	0.639	0.626	6	4.2	0.78	0.765	0.720	0.726
AHT4	6	3.3	0.59	0.698	0.634	0.604	6	4.1	0.73	0.760	0.712	0.711
AHT5	6	3.0	0.66	0.675	0.610	0.594	7	3.5	0.67	0.720	0.666	0.657
HTG4	6	2.5	0.55	0.607	0.560	0.559	5	2.8	0.63	0.648	0.595	0.581
HTG6	3	2.4	0.57	0.594	0.499	0.434	6	3.2	0.65	0.690	0.621	0.587
HTG7	3	1.8	0.36	0.463	0.385	0.329	4	2.1	0.55	0.533	0.454	0.408
mean	5.7	3.7	0.69	0.712	0.657	CPE >0.999	7.7	4.4	0.71	0.759	0.715	CPE >0.999

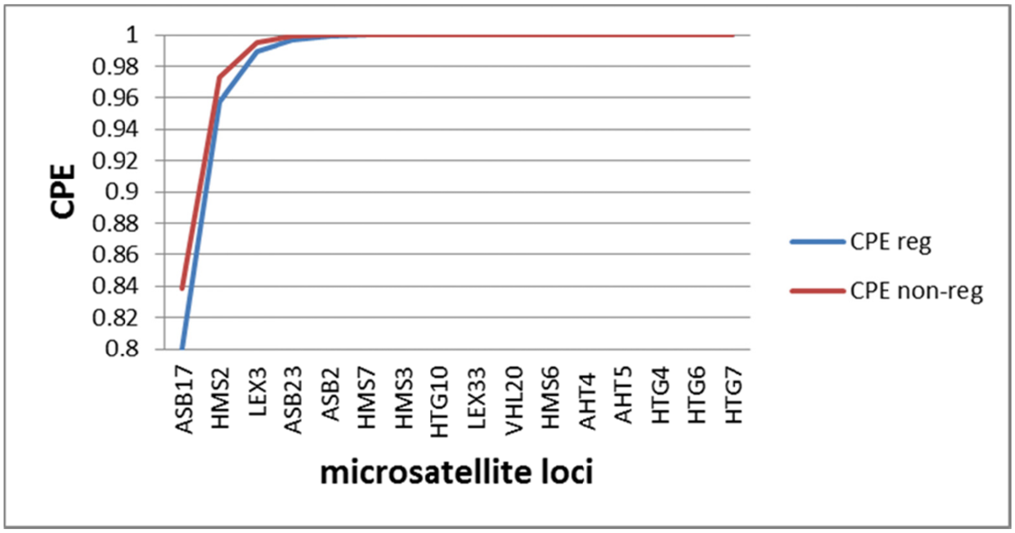


Figure 20: Combined probability of exclusion (CPE). CPE was shown as function of the number of 16 microsatellite loci in registered (reg) and non- registered (non-reg) Arabian horses.

5.4 Discussion

How informative a locus is depends upon the number of alleles exhibited by the locus and the frequency distribution of these alleles in a population (Ozkan*, et al.*, 2009). The mean Na in the present study was higher than that seen in some previous studies of horse breeds especially those which have a high level of inbreeding, such as the Sorraia horse breed (Luís*, et al.*, 2002). Values of diversity statistics similar to those observed here for the Arabian breed have been recorded for a number of equine populations, including the Thoroughbred (Lee and Cho, 2006), Lipizzaner (Achmann*, et al.*, 2004), Lithuanian native horse breeds (Juras and Cothran, 2004), and Pantaneiro horse (Sereno*, et al.*, 2008), however, the individual probability of exclusion (PE) for 13 out of 16 loci used in this study was higher than that reported in many other breeds, including the Thoroughbred (Lee and Cho, 2006). In this study, for the registered group, seven microsatellite markers (ASB17, HMS2, LEX3, ASB23, ASB2, HMS7, HMS3) had high PIC values (>0.7). A very high level of CPE (>0.99999) can be reached using only six of 16 loci (Figure 20), which makes these markers highly valuable for use in a parentage testing for these Arabian horses. (Keeping in mind that HMS2 and ASB23 are not included in the nine loci that make up the minimum standard of ISAG). Ellegren*, et al.* (1992) suggested that at least ten microsatellite loci should be used to achieve maximum exclusion in horses, but our results show that fewer can give a relatively high power, similar to results found by Sereno *et al*. (2008). Two markers, HTG6 and HTG7, were found to have PIC value less than 0.5 for registered group. As they are considered uninformative (Botstein*, et al.*, 1980), and they are in the less efficient 3-plex, these two loci plus HMS2 can easily be excluded from routine parentage testing for the Arabian horses.

In contrast, the non-registered group has PIC mean of 0.715 which was significantly higher than the PIC mean in the registered group (p <0.0001). This value reflects a higher level of variation in the non-registered group compared to the registered horses. The difference of variation between the registered and non-registered horses may be due to the restricted mating in the registered group, where registered horses must be mated within the same RASAN, while the non- registered horses can be crossed with any horse. Heterozygosity in both groups was within the range of heterozygosity in different horse breeds (Reis*, et al.*, 2008; Sereno*, et al.*, 2008). The heterozygosity levels are consistent with the high number of alleles per locus seen in the Arabian horses tested here, and indicate no serious loss of variability due to the breeding method employed by the local breeders in Syria, which is different than that practiced by most horse breeders. The International Stud Book Committee (ISBC) has required that the CPE value for parentage verification and an individual identification in horse be higher than 0.9995 (Tozaki*, et al.*, 2001). Here we showed that CPE using 12 autosomal loci was greater than the value required by the ISBC. Based on these results, we confirmed that loci of the 8 plex and 5 plex PCR can be used in parentage testing with high efficiency for the Arabian horses from Syria. The data presented here will help solve the problems related to registration issues and will provide the breeders with an effective tool for breeding. The unexpected results here about the high level of genetic diversity noticed in 94 Arabian Syrian horses opened the door to test more Syrian samples and to compare them to other Arabians.

CHAPTER VI

PATTERNS OF SINGLE NUCLEOTIDE POLYMORPHISMS AND MICROSATELLITE GENETIC HETEROZYGOSITY IN THE HORSE GENOME

6.1 Introduction

The description of the amount and distribution of genetic heterozygosity within a genome is essential to understand the history of species and the evolutionary forces such as selection, mutation, and recombination. Also it is important for the investigation of relatedness among individuals, genetic determinants of phenotypic variation and population demography including historical migration routes, population expansions and declines. (Payseur*, et al.*, 2011). Therefore, studying genetic diversity has important implications for organism evolution, forensics, and distribution of genetic diseases (Jorde*, et al.*, 2000). Genome wide effects of evolutionary forces, especially selection, are represented in different patterns of polymorphism distributions resulting from selective sweeps (Pool*, et al.*, 2010). Such genetic diversity patterns range from a deficit of variation around selected sites (Fu, 1997; Hudson and Kaplan, 1988) to an excess of high-frequency derived alleles in flanking regions (Fay and Wu, 2000). For example, negative selection reduces variation by elimination of some mutations, holding others in low frequency and also causing the loss of variants linked to deleterious alleles (Charlesworth*, et al.*, 1993). Positive selection leads to local reduction in genetic diversity through genetic hitchhiking effect (Smith and Haigh, 2007) where genes or group of sites will harbor fewer or more polymorphism than expected (Payseur*, et al.*, 2002). Genetic diversity is a complex function of different evolutionary and demographic factors not only selection. Thus, the signature of adaptation is expected to be smaller in the high recombination regions (Spencer*, et al.*, 2006). Demographic events such as founder effects, migration and consanguineous mating (mating between close relatives) may cause a reduction in genetic diversity (Khanshour*, et al.*, 2013; Kirin*, et al.*, 2010) and are highly common in horses. It is important to measure the amount of genetic diversity and its distribution throughout a genome to detect inbreeding and recognize any runs of homozygosity (ROH) or any loss of heterozygosity (LOH) which is the most common molecular genetic alteration observed in diseases such as cancers (Lindblad-Toh*, et al.*, 2000).

Different molecular markers have been used to measure genetic diversity. SNPs are the most common form of DNA sequence variation in a genome and were hypothesized to become the markers of the choice in ecological, evolutionary, conservation and medical studies (Sachidanandam*, et al.*, 2001; Seddon*, et al.*, 2005; Zheng*, et al.*, 2005). Also, STRs have been the markers of choice for different genomic studies such as genome-wide linkage studies, allelic imbalance studies, population genetic and evolution studies in many organisms over the past 20 years (Bruno-de-Sousa*, et al.*, 2011; Gulcher, 2012; Selkoe and Toonen, 2006). The key advantages of SNPs compared to STRs are a very low false genotyping rate, presence in coding and non-coding regions, a low mutation rate, the abundance in a genome and the most widespread form of DNA variation in a genome with a uniform distribution (Fries and Durstewitz, 2001; Gärke*, et al.*, 2012; Xing*, et al.*, 2005). On the other hand, the fact that

SNPs are biallelic and less informative than STRs is a disadvantage (Schaid*, et al.*, 2004). However, SNPs have replaced STRs in the recent years as the markers of choice for most large scale genomic studies in many organisms (Miller*, et al.*, 2005; Sabeti*, et al.*, 2007). In summary, both SNPs and STRs have advantages and disadvantages and different molecular backgrounds. SNPs have recently been successfully used in genome studies, and STRs have been widely used markers for genomic studies and remain so (Gulcher, 2012). Therefore, studying the interaction between SNP and STR distributions as two biological markers having different biological and molecular background, especially different mutation rates, might be an important tool to understand and explain the way in which genetic diversity is formed in a genome. Many such comparisons have been done in several organisms to study population genetics and genomic diversity (Ball*, et al.*, 2010; Coates*, et al.*, 2009; Forstmeier*, et al.*, 2012; Gärke*, et al.*, 2012; Glover*, et al.*, 2010; Hauser*, et al.*, 2011; Haynes and Latch, 2012; Morin*, et al.*, 2009; Narum*, et al.*, 2008; Rengmark*, et al.*, 2006; Smith*, et al.*, 2007; Thalamuthu*, et al.*, 2005; Varela and Amos, 2010) and genome linkages studies of diseases (Hoque*, et al.*, 2003; Schaid*, et al.*, 2004).None of these studies has been done in horses.

The aim of this study is to examine the pattern of genetic diversity provided by two different types of molecular markers, STRs and SNPs, at different levels of horse genome organization.

6.2 Materials and Methods

6.2.1 Samples

Horses have been successfully used in genetics and biomedical studies as model animals for many purposes. Therefore, we used 22 samples of the Peruvian Paso breed from the USA in this study.

6.2.2 Microsatellite STRs data collection

Microsatellite data came from the study done by Diane Strong in 2006 "The use of a whole genome scan to find a genetic marker for Degenerative Suspensory Ligament Desmitis in the Peruvian Paso horse" (MS thesis, University of Kentucky, under the direction of Dr. Cothran). All microsatellite markers used were from published sources where primer sequences and variability information was given. This thesis is available online including the STR genotypes (http://uknowledge.uky.edu/gradschool_theses/419). I assigned the 232 STRs markers to their chromosomal locations based upon the information from the horse genome project website at the NCBI (http://www.ncbi.nlm.nih.gov). For chromosomes 1 to 13, STRs were assigned onto the two arms (short arm p and long arm q). Linkage disequilibrium (LD) between all pairs of loci was tested for STRs data by GENEPOP 3.4 (Raymond and Rousset, 2001) based on the exact test using the default parameters specified by the software. Pairs of loci showing significant LD at the level of 0.05 were excluded. Table 13 shows 232 STRs markers on each autosomal chromosome arm with their positions in the genome.

Table 13: STR markers on each chromosome (Ch) with their positions in the genome. Shadowed areas refer to the short arm.

Ch	STR s	position	Chr	STRs	position	Chr	STRs	position	Chr	STRs	position	Chr	STRs	position
1	ASB41	18,265,689	4	ASB22	59,506,355	9	LEX019	75,595,151	16	AHT037	3,919,328	22	COR016	31,733,320
1	LEX020	20,590,361	4	COR089	59,843,221	10	COR020	9,995,159	16	TKY279	6,632,322	22	HMS47	39,952,965
1	COR100	50,781,058	4	HTG07	64,169,187	10	COR048	12,137,892	16	HTG03	8,149,566	22	HTG21	49,944,830
1	COR059	56,585,301	4	HTG22	98,452,420	10	ASB06	14,458,849	16	HMS20	27,660,000	23	COR055	3,251,291
1	TKY007	59,942,316	4	SGCV23	102,735,854	10	NVHEQ018	15,382,260	16	AHT038	30,274,261	23	LEX063	29,671,975
1	UCDEQ487	66,489,862	5	LEX004	7,818,164	10	NVHEQ007	36,378,123	16	L15-2	58,998,000	23	COR084	40,400,735
1	AHT021	89,894,655	5	AHT24	12,165,190	10	HMS002	52,713,200	16	LEX056	70,125,109	23	SGCV004	51,900,473
1	ASB08	99,984,303	5	VHL66	28,069,647	10	AHT86	76,695,520	16	I-18	74,985,165	24	LEX042	18,555,887
1	LEX058	102,644,095	5	HMS05	57,359,725	10	ASB09	54,960,72	16	AHT60	81,441,900	24	AHT32	20,909,877
1	TKY002	108,068,964	5	LEX069	63,737,858	10	COR085	73,487,975	16	AHT91	84,054,300	24	AHT4	23,415,673
1	1CA43	110,280,065	5	LEX034	76,173,038	11	UCDEQ439	8,460,189	17	COR007	6,608,667	24	COR061	33,238,445
1	1CA25	117,758,709	6	HTG31	4,340,328	11	SGCV24	19,537,692	17	LEX076	8,812,013	24	LEX074	34,015,752
1	TKY106	118,802,990	6	COR010	14,720,900	11	ASB35	25,599,244	17	NVHEQ79	20,690,900	24	COR024	41,000,139
1	UCDEQ493	119,389,495	6	NV82	15,512,496	11	SGCV13	26,147,201	17	COR032	41,428,036	25	COR080	8,780,724
1	HTG12	124,267,100	6	LEX065	20,784,689	11	TKY033	36,817,461	17	HMS25	61,873,600	25	COR018	15,686,913
1	UM004	129,608,032	6	UM015	34,558,289	11	NVHEQ90	37,311,816	18	TKY19	539,058	25	TKY018	18,489,234
1	UCDEQ440	130,126,191	6	TKY111	45,045,977	11	TKY648	38,782,434	18	LEX054	16,952,947	25	AHT007	28,114,731
1	HMS15	136,853,559	6	NVHEQ81	59,226,930	11	TKY010	39,679,362	18	UMNE50	23,060,120	25	AHT051	30,939,703
1	COR063	181,374,365	6	UCDEQ465	61,228,738	12	SGCV10	9,547,876	18	HMS46	25,125,738	25	NVHEQ043	31,051,894
2	COR065	1,737,180	6	COR070	65,850,909	12	SGCV08	21,559,085	18	SGCV07	26,364,970	26	COR071	19,052,877
2	ASB18	5,257,678	6	TKY412	70,589,233	12	COR030	24,880,227	18	TKY909	26,794,020	26	EB2E8	27,157,252
2	COR037	21,605,481	6	TKY284	73,768,431	12	COR058	27,946,790	18	TKY692	36,964,437	26	NVHEQ070	30,252,733
2	TKY024	23,738,085	7	TKY35	22,461,800	12	UCDEQ497	32,574,331	18	COR096	37,306,808	27	COR040	17,151,726
2	ASB17	30,600,996	7	TKY34	22,488,330	13	COR069	6,098,825	18	HTG28	38,640,668	27	HMS45	20,079,170
2	HMS051	32,978,100	7	TKY283	43,551,219	13	UM030	10,713,060	18	TKY017	66,813,831	27	COR017	35,275,682

Table 13 Continued

Chr	STR s	position	Chr	STRs	position	Chr	STRs	position	Chr	STRs	position	Chr	STRs	position
2	UCDEQ380	36,574,029	7	TKY005	43,597,947	13	ASB37	15,051,228	18	UCDEQ387	75,253,209	28	NVHEQ54	5,062,828
2	COR049	55,120,827	7	TKY272	51,294,156	13	VHL47	16,894,755	18	HLM3	74,489,252	28	UM003	10,560,340
2	COR094	70,292,709	7	COR004	51,416,907	13	AHT30	22,064,810	19	HTG23	9,108,342	28	HTG30	10,876,313
2	A-14	74,473,683	7	COR095	54,216,182	13	ASB01	31,743,456	19	LEX036	17,854,559	28	TKY319	25,543,673
2	ASB13	75,644,358	7	SGCV28	71,099,574	14	HTG29	12,821,063	19	LEX073	24,403,545	28	TKY515	30,492,665
2	UMNe076	87,061,070	7	AHT019	85,688,021	14	LEX043	16,144,786	19	COR092	45,782,951	28	UCDEQ425	43,085,558
2	TKY335	90,635,840	8	AHT005	740,641	14	UM010	25,466,225	19	AHT041	59,892,096	29	LEX018	3,007,558
2	TKY798	93,955,060	8	AHT025	2,570,692	14	VHL209	32,966,942	20	HTG5	10,511,401	29	COR082	4,277,206
2	TKY497	104,824,212	8	UM034	18,876,355	14	LEX047	34,561,952	20	LEX052	13,654,145	29	COR027	22,227,341
2	COR026	117,183,370	8	LEX023	25,944,092	14	TKY310	45,639,500	20	UM011	33,510,120	29	COR021	33,632,759
2	COR043	117,548,315	8	ASB14	41,189,276	14	TKY491	81,175,596	20	LEX071	61,179,336	30	LEX025	2,041,967
2	COR035	118,389,437	8	COR003	64,251,046	14	AHT83	81,754,500	20	HMS42	63,743,901	30	HTG27	7,292,962
2	TKY842	118,406,209	8	COR056	84,105,135	14	TKY749	86,872,839	21	SGCV16	3,013,600	30	HMS18	11,408,766
3	AHT036	2,948,130	8	SGCV32	57,499,947	14	LEX078	87,482,717	21	TKY021	448,857	30	VHL20	18,793,901
3	COR028	11,070,092	9	HTG4	1,497,890	14	COR002	90,003,124	21	SGCV14	1,604,070	31	AHT33	602,132
3	COR033	13,467,228	9	HMS03	16,895,898	14	TKY438	90,117,274	21	HTG10	17,139,092	31	COR038	632,737
3	AHT022	20,876,315	9	COR008	18,912,052	14	TKY636	91,846,301	21	COR073	20,250,418	31	TKY274	11,455,554
3	UCDEQ437	31,285,262	9	TKY627	20,348,236	15	B-8	21,787,962	21	CORO68	22,008,345	31	VIASH21	13,649,168
3	AHT097	99,036,446	9	COR013	23,219,062	15	LEX046	39,365,857	21	HTG32	32,476,035	31	AHT34	21,679,544
4	AHT043	2,915,350	9	HTG08	30,021,222	15	ASB02	55,316,355	21	LEX037	47,817,039			
4	HMS6	7,229,400	9	UM037	42,508,042	15	HTG06	73,962,712	22	TKY285	10,984,309			
4	LEX033	59,500,055	9	AHT53	51,322,320	15	COR014	86,778,890	22	COR022	22,898,531			

6.2.3 SNPs genotyping and quality control

The same 22 Peruvian Paso horses that we recently genotyped for microsatellites were used for SNPs genotyping. Genomic DNA was extracted from hair samples using PUREGENE® DNA purification kit following the manufacturer's instructions. SNPs were genotyped on Illumina EquineSNP50 BeadChip by Geneseek® and all genotype calls were extracted from the raw intensity data using GenomeStudio Genotyping Module with the minimum score cutoff of 0.15.

Data cleaning and filtering were performed using Plink (Purcell*, et al.*, 2007). Only autosomal SNPs were included in this study. The basic data cleaning were carried out according to (Petersen*, et al.*, 2013) where the missing rate per individual and per SNP were set to 0.1 (individuals with more than 10% missing genotypes have been excluded and only SNPs with a 90% call rate have been included, respectively). Hardy-Weinberg Equilibrium (HWE) exact test (Wigginton*, et al.*, 2005) filter was applied to exclude SNPs that deviated from HWE at $P < 0.001$ (Purcell*, et al.*, 2007). Further filters were applied by using three values of minor allele frequency MAF (0.01, 0.05 and 0.1) and four combinations of Linkage Disequilibrium (LD) pruning filter using pair-wise genotypic correlation in 50 and 100 SNPs, windows sliding by 5 and 25 SNPs along the genome with SNPs pruning at r2>0.5 and r2>0.2. Figure 21 shows the cleaning and filtering combinations tested in the SNPs data set, and the number of remaining SNPs in each data subset.

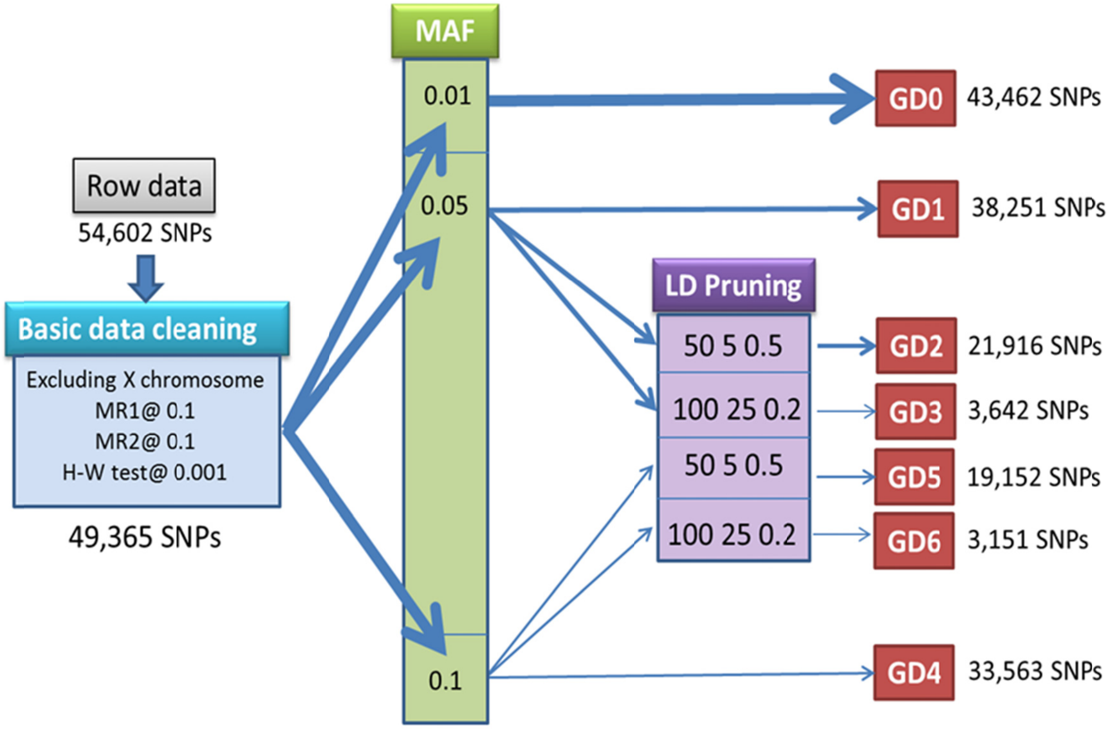


Figure 21: Cleaning and filtering the SNPs data set. MR1: missing rate per individual. MR2: missing rate per SNP. H-W test: Hardy-Weinberg test. MAF: minor allele frequency. LD pruning: linkage equilibrium pruning. GD0 to GD6: data subsets after filtering.

6.2.4 Data analysis

6.2.4.1 Genetic diversity calculation

Multi locus expected heterozygosity H_e is a suitable measure to capture information on genome-wide heterozygosity (Alho, *et al.*, 2010). STRH_e was calculated from STR data by Rhh version 1.0.1 (Alho, *et al.*, 2010) using R 3.0 extension package available through the Comprehensive R Archive Network (CRAN; http://cran.r-project.org). SNPH_e was calculated from SNPs data using PLINK (Purcell, *et al.*, 2007).

6.2.4.2 The general pattern of genetic diversity

General pattern of $STRH_e$ was analyzed with Each Pair Student’s test using JMP software (SAS Institute). The comparisons were done among 31 chromosomes, and among arms in the bi-arms chromosomes 1 to13. Chromosomes 14 to 31 are acrocentric with a single arm. Each Pair Student’s test was also applied for $SNPH_{e,}$. The comparisons were done among 31 chromosomes (for all subsets GD0 to GD6), and among segments in each chromosome (only GD1 and GD2). Chromosomes 1 to 13 have three segments (short arm, long arm and the centromeric region). Chromosomes 14 to 31 have only long arms and centromeric regions.

6.2.4.3 Comparisons between the genetic diversity of STRs and SNPs

To investigate the concordance between $STRH_e$ and $SNPH_e$, three levels of comparisons were done using JMP software. Level 1: overall heterozygosity (all 31 chromosomes together); level 2: chromosomal heterozygosity (by each of the 31 chromosomes); level 3: segmental heterozygosity (by each arm in each of chromosomes 1 to 13). In each level, two Each Pair tests were conducted: Student's Test and Nonparametric Test using the Wilcoxon method.

Correlation between $STRH_e$ and $SNPH_e$ was also studied at the three levels mentioned above. Two different approaches were applied: Pairwise correlation (R) and Nonparametric correlations of Spearman’s method (ρ). All correlation tests were done by JMP.

In addition to that, I looked for candidate segments for positive selection provided by SNPs data and associated STRs. I did the Runs Of Homozygosity test (ROH) along the genome using SNPs data by PLINK. The criteria applied in this test were as following: The sliding window 5000, segment length 500 kb, number of SNPs 50, the minimum of SNPs density 50 kb/SNP and the largest gap 1000 kb (Petersen*, et al.*, 2013). Then, STRs markers were assigned to the resulting homozygous segments based upon marker positions provided by Illumina EquineSNP50 BeadChip platform for SNPs and the information from the horse genome project website at the NCBI (http://www.ncbi.nlm.nih.gov) for STRs. Limited comparisons were available between the homozygosity provided by SNPs and Homozygosity status in the STRs marker in a same segment.

6.3 Results

6.3.1 SNPs quality control and filtering

As shown in Figure 21, the pre-analysis done on the different subsets from GD0 to GD6 showed that GD1 and GD2 were the best filter combinations that serve the aim of my study. GD1 was the subset resulting from the usual recommended filters used by other studies such as (Petersen*, et al.*, 2013). GD2 is the subset resulting from using the usual filters in addition to the LD filter which might be interesting to be compared with GD1 to see a possible effect of those SNPs having LD relationships on the heterozygosity distribution. GD0 is the subset that resulted from very basic filters that are not recommended in some studies such as studies of heterozygosity. GD3 and GD6 gave very low number of SNPs where more than 92% of total SNPs were excluded because of the high impact of the used filters. Also, GD5 had low number of SNPs simply because the MAF value (0.1) used here was high. Figure 22 shows SNPHe for all subsets (GD0 to GD6). Only GD1 and GD2 were included in the further analysis.

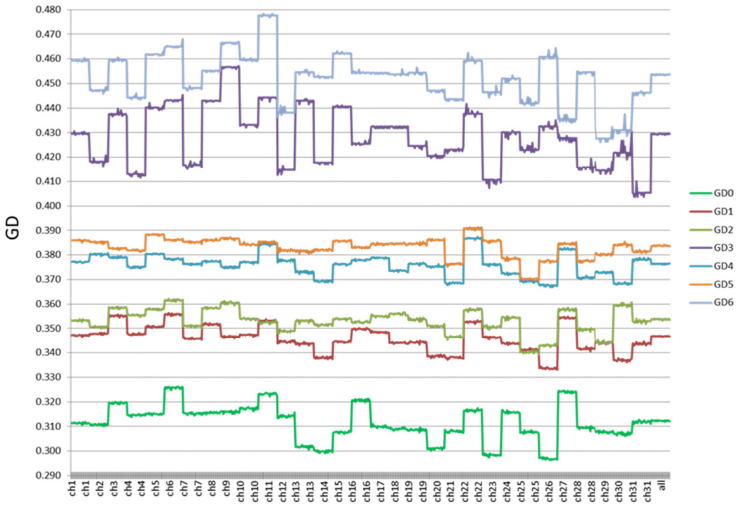


Figure 22: SNPHe for all subsets (GD0 to GD6). Information about GDs can be found in Figure 21.

6.3.2 The general pattern of genetic diversity

6.3.2.1 STRHe pattern

The box plots and means comparisons of STRH$_e$ among 31 chromosomes in Figure 23.A showed high level of variation within all chromosomes. The each pair means comparisons among chromosomes in Figure 3.B showed no significant differences among most of the 31 chromosomes. Only chromosome 7 was different from others (P<0.01). Limited cases were also found in other chromosome comparisons and most of them were significant only at 0.05 level as shown in Figure 23.B in blue.

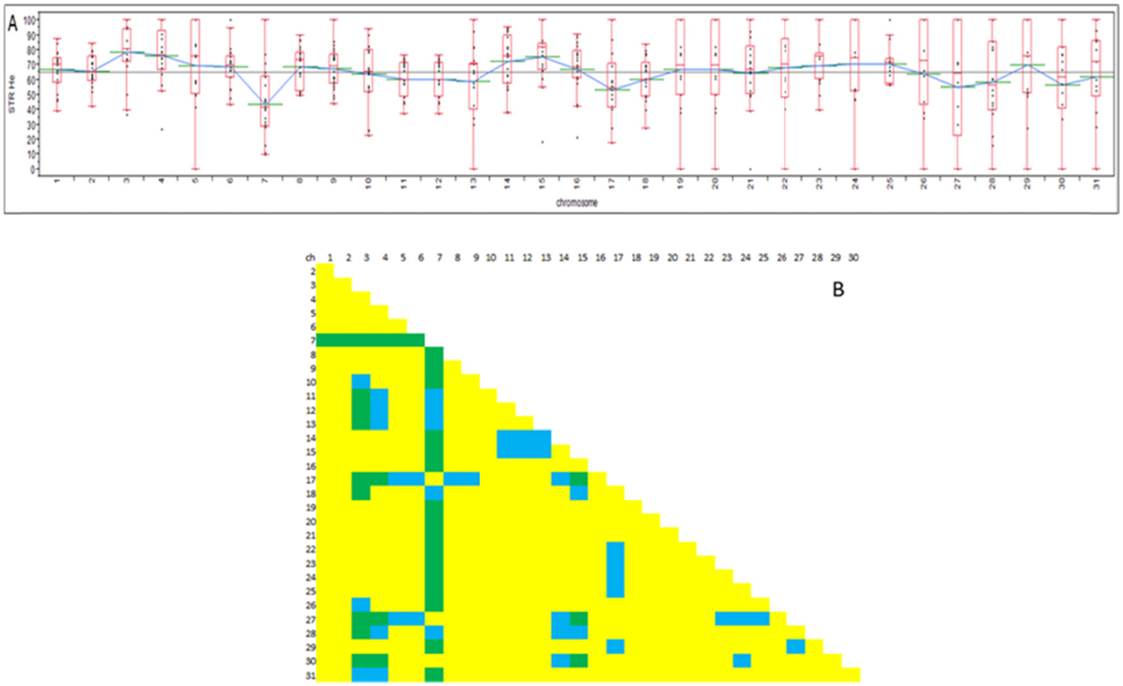


Figure 23: STRH$_e$ pattern by Each Pair Student's test among 31 chromosomes. A: Box plots and means comparisons among chromosomes. B: each pair comparisons. Green shows pairs of means that are significantly different at P<0.01 Blue shows pairs of means that are significantly different at P<0.05 Yellow: no difference.

The box plots and means comparisons of STRH$_e$ between long and short arms in each of chromosomes 1 to 13 (Figure 24) also showed high level of variation within each segment. There were no significant differences between short and long arms in 11 out of 13 chromosomes. Chromosome 8 showed a significant difference (P< 0.05) between segments where the mean STRH$_e$ was higher in 8q than 8p. Also, chromosome 12 showed a significant difference (P< 0.01) between segments where the mean STRH$_e$ was higher in 12p than 12q. The differences found in both chromosomes 8 and 12 might not have biological meaning due to the limited number of STRs tested in the short arms of these two chromosomes.

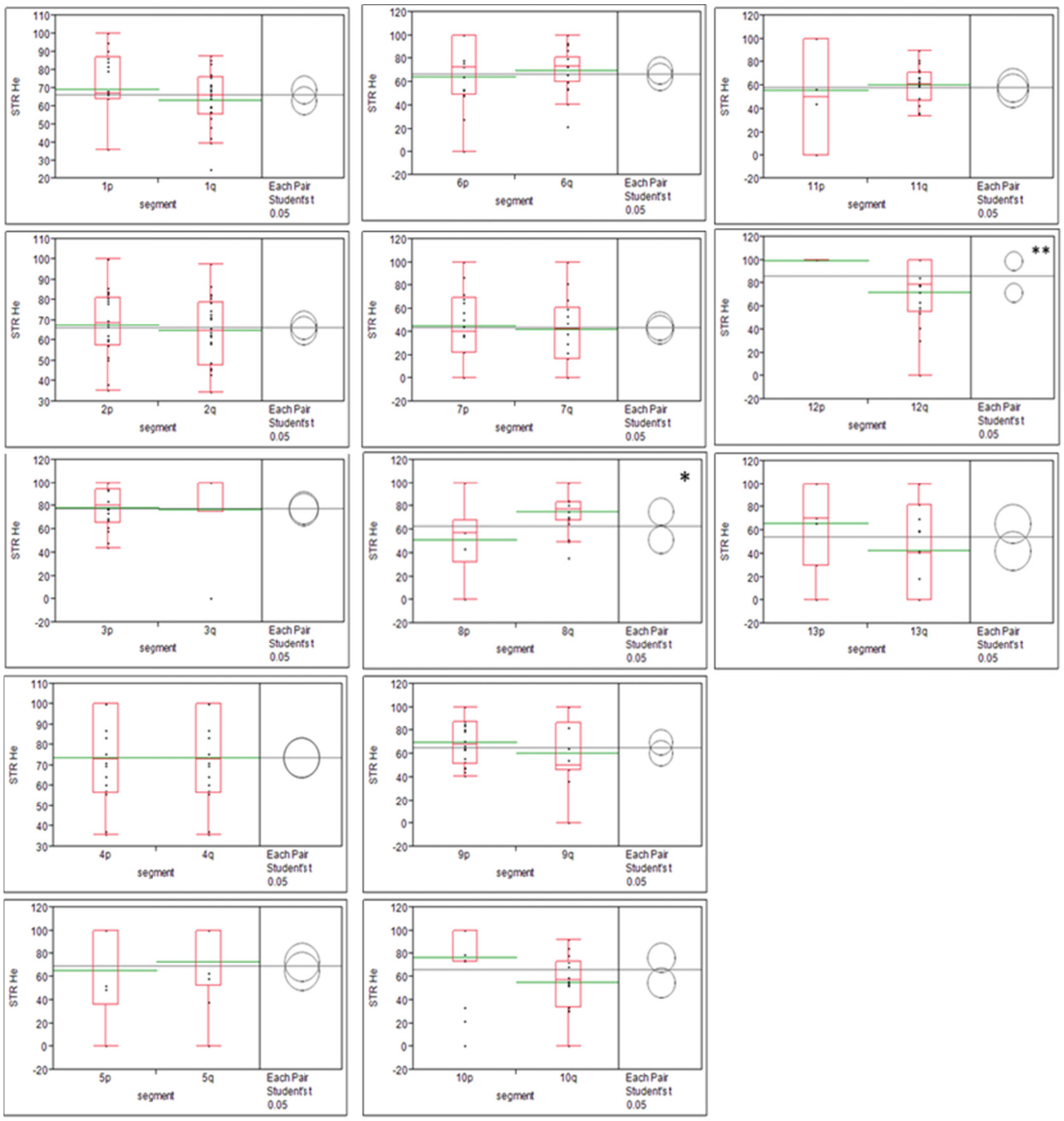


Figure 24: STRH$_e$ pattern by Each Pair Student's test comparisons of two arms (p and q). * P<0.05. ** P<0.01

6.3.2.2 SNPH$_e$ pattern

As shown in Figure 25.A and 26.A, the box plots and means comparisons of SNPH$_e$ (done using GD1 and GD2) among 31 chromosomes reveal low level of variation within each chromosome. In contrast to STRHe, the each pair means comparisons of SNPH$_e$ among chromosomes (Figure 25.B and 26.B) showed highly significant ($P<0.01$) differences among most of the 31 chromosomes (403 out of 420 possible comparisons) for both GD1 and GD2 subsets. In a limited number of cases (only 12 out of 420 possible comparisons) some chromosomes showed no significant differences, Figure 25.B and 26.B in yellow.

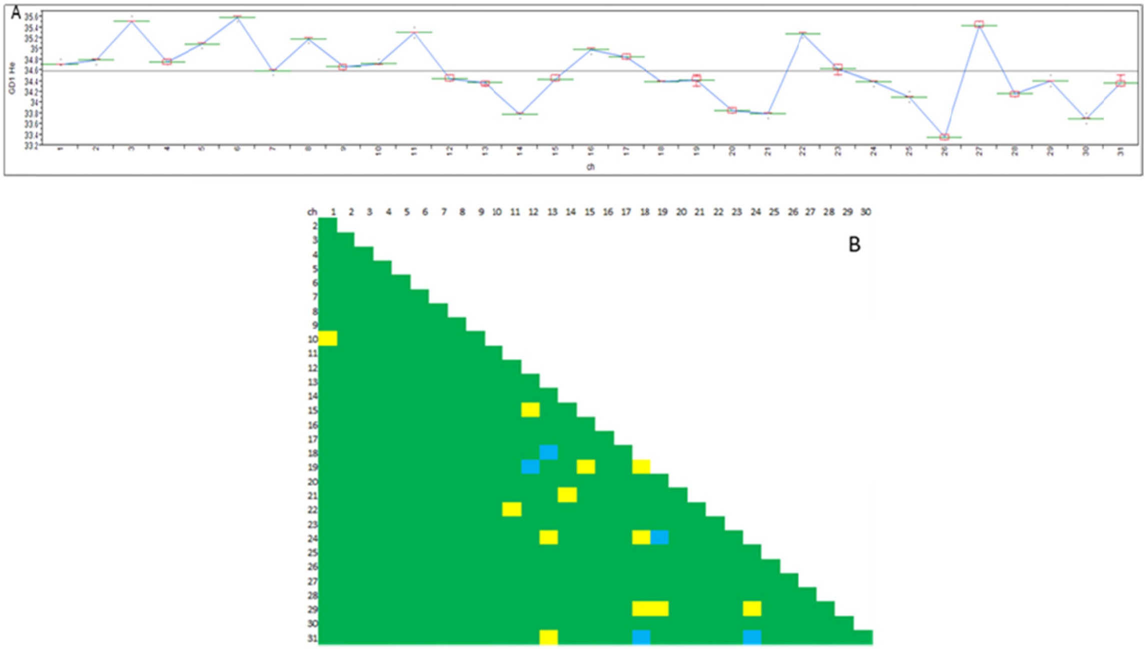


Figure 25: SNPH$_e$ pattern in GD1 by Each Pair Student's test among 31 chromosomes. A: Box plots and means comparisons among chromosomes. B: each pair comparisons. Green shows pairs of means that are significantly different at $P<0.01$ Blue shows pairs of means that are significantly different at $P<0.05$ Yellow: no different.

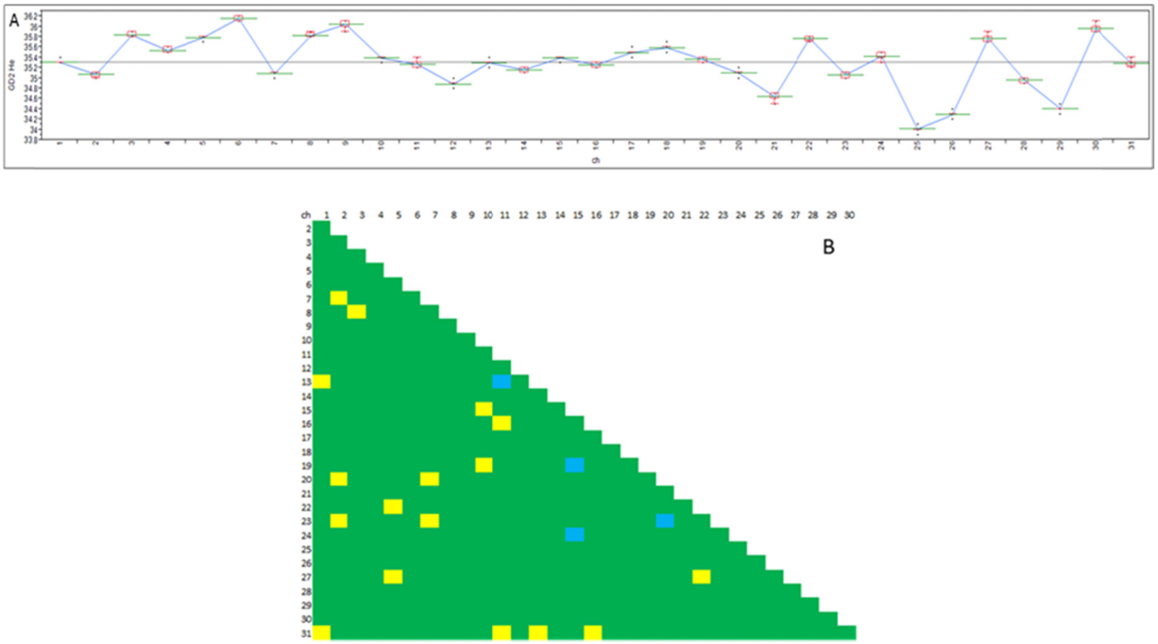


Figure 26: SNPH$_e$ pattern in GD2 by Each Pair Student's test among 31 chromosomes. A: Box plots and means comparisons among chromosomes. B: each pair comparisons. Green shows pairs of means that are significantly different at P<0.01 Blue shows pairs of means that are

The comparisons of SNPH$_e$ using GD1 between three segments (p, q and c) in each of chromosomes 1 to 13 (Figure 27.A) showed significant (P< 0.01) differences among all segments in all 13 chromosomes. Surprisingly, the centromeric regions did not always have the lowest values of heterozygosity compared with short and long arms. The comparisons of SNPH$_e$ done using GD2 in Figure 27.B between the three segments showed a similar pattern to that was found by using GD1 subset. Interestingly, in chromosome 6 and 7, the short and long arms in each of these two chromosomes were not significantly different.

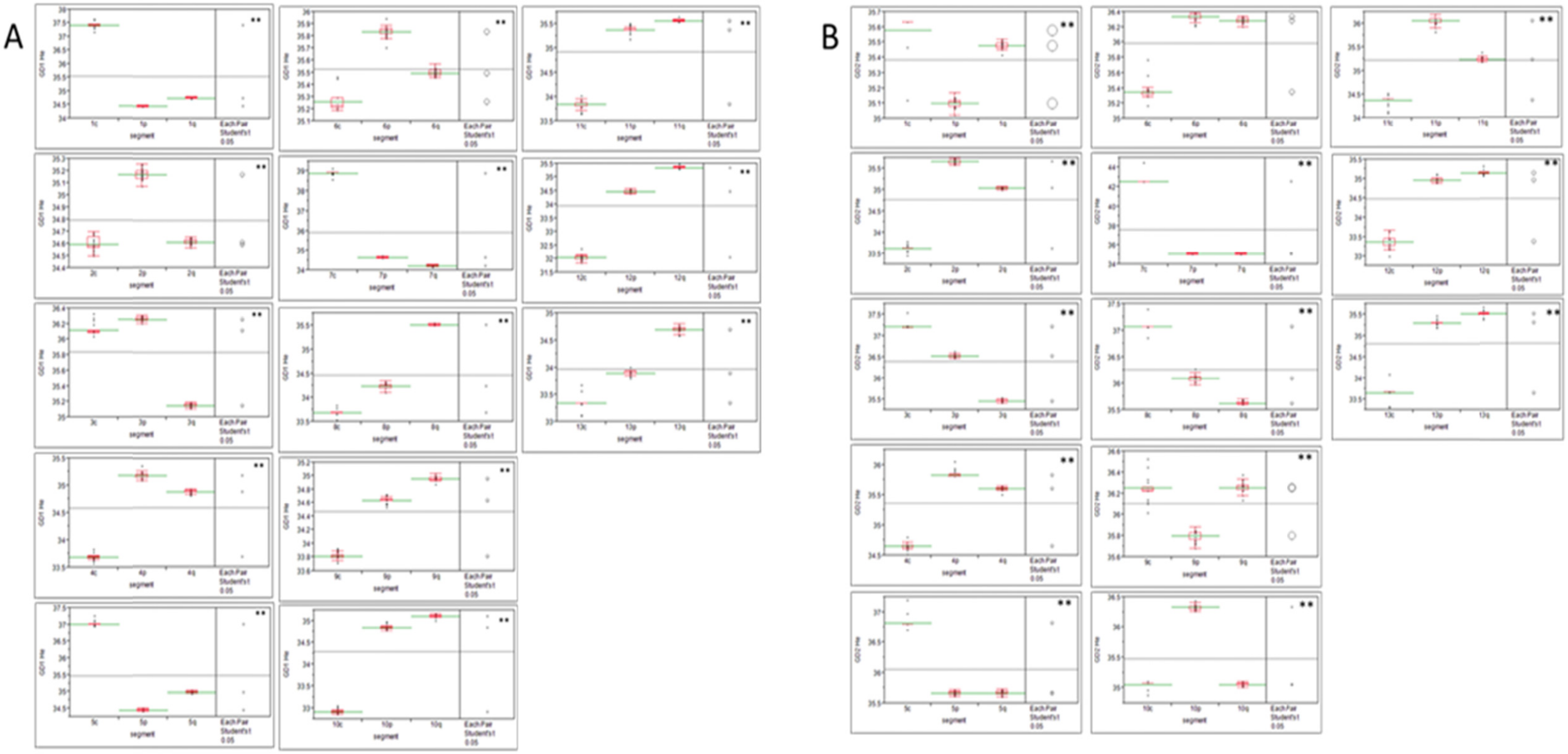


Figure 27: Each Pair Student's test comparisons of SNPH$_e$ pattern of three segments (p: short arm, q : long arm and c: centromeric region) in each of chromosomes (1 to 13). ** P<0.01 GD1 subset. A: GD1 subset. B: GD2 subset.

All comparisons of SNPH$_e$ done using GD1 between the two segments (c, q) in chromosomes 14 to 31 (Figure 28) showed significant (P<0.01) differences between c and q in all chromosomes except chromosomes18 and 31.

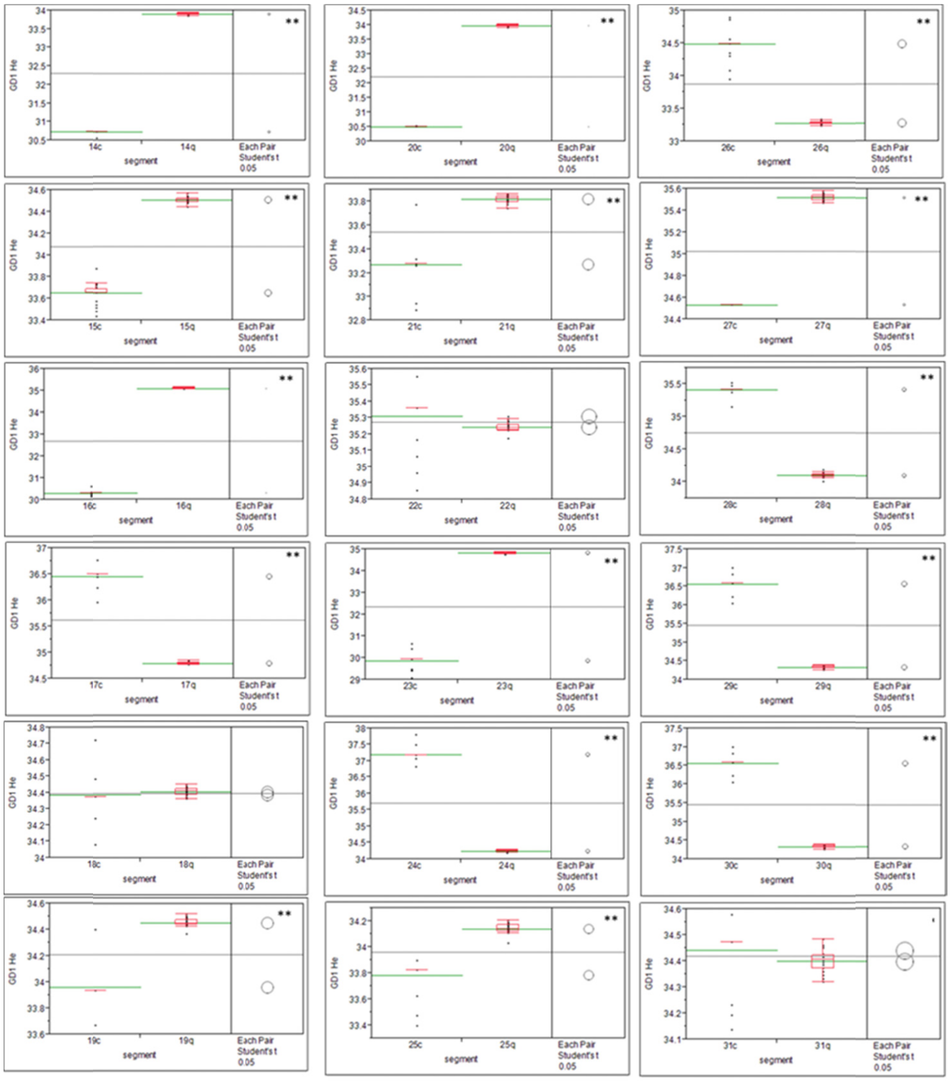


Figure 28: SNPHe pattern from GD1 subset by Each Pair Student's test comparisons of two segments (c: centromeric and q: long) in each of chromosomes (14 to 31). ** P<0.01.

Likewise, all comparisons of c and q were significant (P<0.01) with GD2 subset (Figure 29).

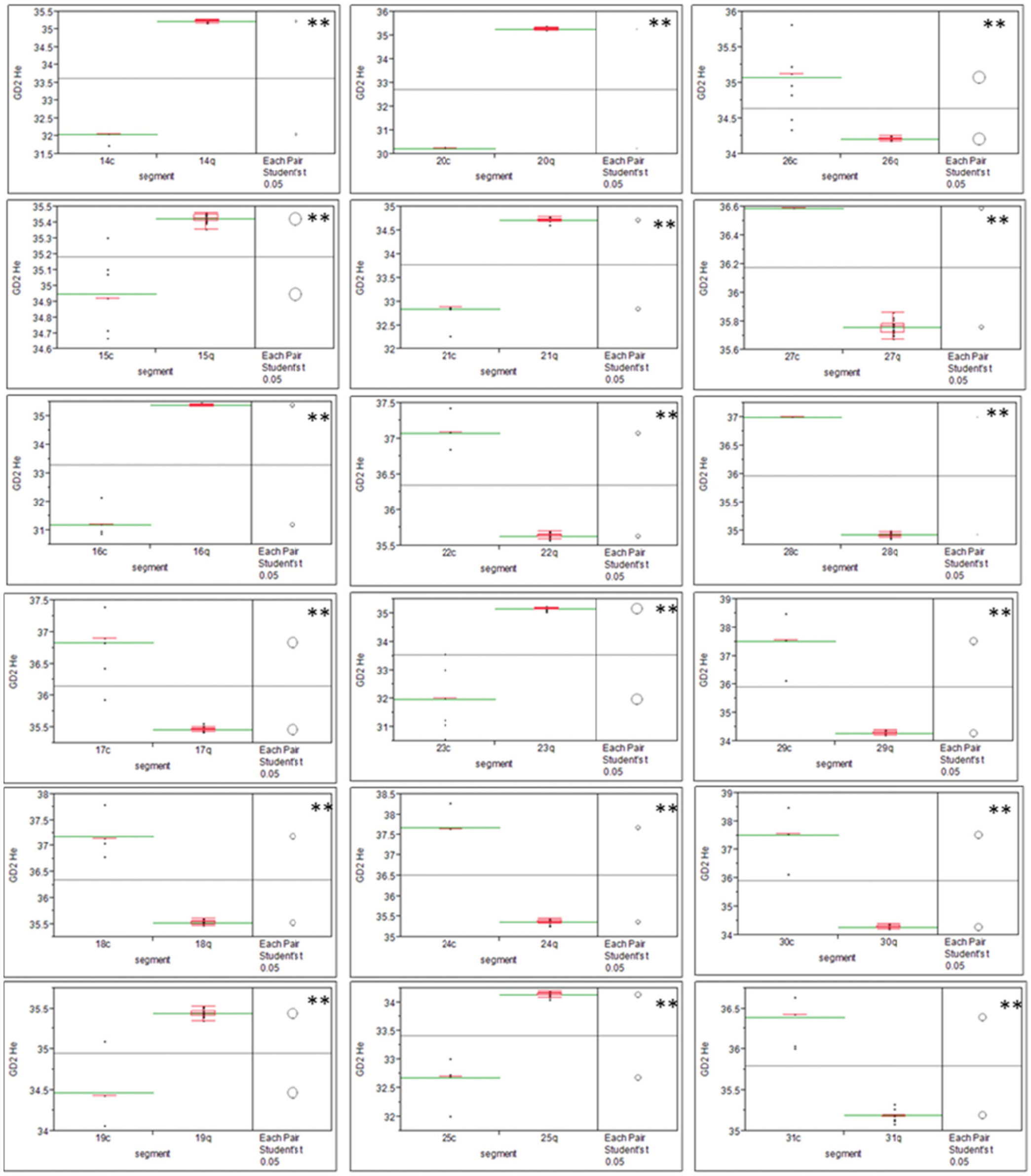


Figure 29: SNPH$_e$ pattern from GD2 subset by Each Pair Student's test comparisons of two segments (c: centromeric and q : long) in each of chromosomes (14 to 31). ** P<0.01.

6.3.3 Comparisons between the genetic diversity of STRs and SNPs

6.3.3.1 The concordance between STRH$_e$ and SNPH$_e$

6.3.3.1.1 Level 1

Figure 30 shows the Student's Test comparison between the overall heterozygosity of STRs and SNPs. The ranges were from 0 to 100 and 33.3 to 36.2 in STRH$_e$ and SNPH$_e$, respectively. The means were 65.4, 34.5 and 35.3 in STRHe, GD1-SNPH$_e$ and GD2-SNPH$_e$, respectively. Student's Test and Wilcoxon's test showed significant (P<0.01) differences between STRH$_e$ and SNPH$_e$. There was no significant difference between GD1 and GD2.

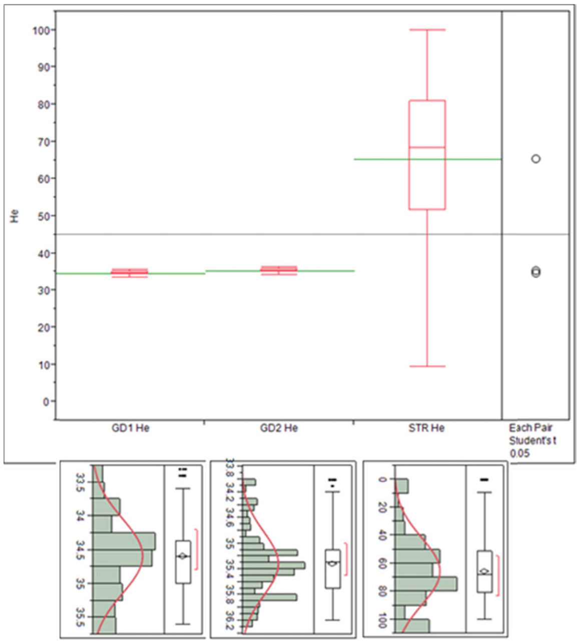
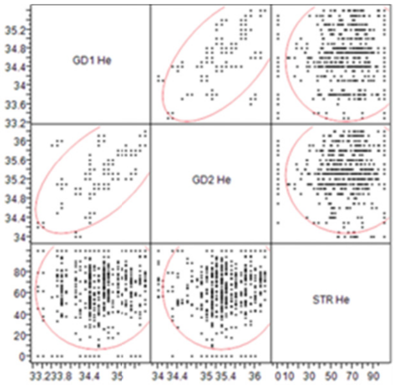


Figure 30: Each Pair Student's test comparisons between overall heterozygosity calculated from STRs and SNPs markers. GD1 and GD2 represented two subsets of SNPs as shown in materials and methods.

There was no significant correlation between overall heterozygosity of $STRH_e$ and $SNPH_e$ by using both the Pairwise correlation and Spearman correlation methods (Figure 31). GD1 and GD2 were significantly correlated using both statistical methods.



Variable	by Variable	Correlation	Count	Significantly
GD2 He	GD1 He	0.596364185	682	**
STR He	GD1 He	0.034265823	682	no
STR He	GD2 He	0.013143478	682	no

Figure 31: The pairwise correlation (R) between overall heterozygosity of $STRH_e$ and $SNPH_e$

6.3.3.1.2 Level 2

The chromosomal heterozygosity means comparisons between $STRH_e$ and $SNPH_e$ by the Student's and Wilcoxon’s tests showed significant ($p<0.01$) differences between $STRH_e$ and $SNPH_e$ in each of 31chromosmoes with two exceptions: differences in chromosome 27 were only significant at 0.05 Figure 32.A, and there was no significant difference in chromosome 7, Figure 32.B. There were no significant differences between GD1 and GD2 in all chromosomal heterozygosity comparisons.

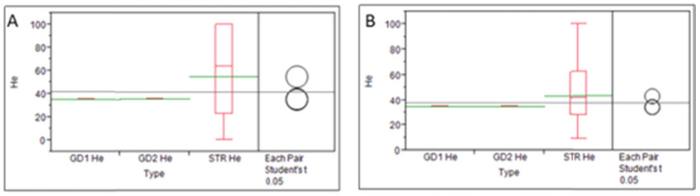


Figure 32: Chromosomal heterozygosity comparisons between $STRH_e$ and $SNPH_e$ by Each Pair Student's test. A: in chromosome 27. B: in chromosome 7.

Table 14 shows the correlations of the chromosomal heterozygosity between $SNPH_e$ and $STRH_e$. There were no significant correlations between $STRH_e$ and $SNPH_e$ using parametric and non-parametric methods for all chromosomes except for chromosome 4 where $STRH_e$ was negatively correlated with $SNPH_e$. Also chromosomes 10 and 20 had significant ($P<0.05$) positive correlations between GD2 and $STRH_e$ only by using R method, and chromosome 1 had significant ($P<0.05$) positive correlations between GD2 and $STRH_e$ only by using Spearman's method. For some chromosomes, correlations between GD1 and GD2 were significantly positive and in others there was no correlation at all. Both statistical methods give very similar results.

6.3.3.1.3 Level 3

The segmental heterozygosity means comparisons between $STRH_e$ and $SNPH_e$ by Student's Test in each of chromosomes 1 to 13 showed that each of chromosomes (1, 2, 3, 4, 5, 6, 9, 10, 12,13) had significant differences ($P< 0.01$) between $STRH_e$ and $SNPH_e$ in short arms. In long arms similar results were found except for chromosome 8 which also showed significant ($P<0.01$) differences. Wilcoxon's test gave a similar results came from the Student's Test.

Table 14: Correlations of chromosomal heterozygosity between SNPH$_e$ and STRH$_e$. Green: P<0.01. Blue: P<0.05.

chromosome	GD1 H_e by GD2 H_e correlations				GD1 H_e by STRH_e correlations				GD2 H_e by STRH_e correlations			
	Pairwise		Spearman		Pairwise		Spearman		Pairwise		Spearman	
	R	P	ρ	P	R	P	ρ	P	R	P	ρ	P
1	-0.1	no	-0.1	no	0.126	no	0.099	no	0.495	no	0.499	*
2	0.516	*	0.516	*	-0.04	no	0	no	0.123	no	0.129	no
3	0.525	*	0.525	*	0.011	no	0.0734	no	0.226	no	0.276	no
4	0.63	**	0.63	**	-0.52	*	-0.513	**	-0.59	**	-0.554	**
5	0.335	no	0.335	no	0.328	no	0.243	no	-0.077	no	-0.043	no
6	0.418	no	0.418	no	0.241	no	0.224	no	0.368	no	0.35	no
7	0.549	**	0.549	**	0.123	no	0.114	no	0.09	no	0.08	no
8	0.243	no	0.243	no	0.317	no	0.313	no	0.315	no	0.33	no
9	0.57	**	0.57	**	-0.221	no	-0.144	no	0.203	no	0.278	no
10	-0.402	no	-0.402	no	-0.095	no	-0.085	no	0.447	*	0.361	no
11	0.524	*	0.524	*	-0.17	no	-0.19	no	-0.289	no	-0.304	no
12	0.45	*	0.45	*	-0.096	no	-0.094	no	-0.258	no	-0.246	no
13	0.47	*	0.47	*	-0.08	no	-0.112	no	-0.053	no	-0.11	no
14	0.5817	**	0.5817	**	-0.259	no	-0.25	no	-0.048	no	-0.015	no
15	0.16	no	0.16	no	0.167	no	-0.23	no	-0.01	no	0.086	no
16	0.089	no	0.089	no	-0.087	no	-0.049	no	0.298	no	0.156	no
17	0.3	no	0.3	no	-0.197	no	-0.21	no	-0.193	no	-0.21	no
18	0	no	0	no	0	no	0	no	0.04	no	0.05	no
19	0.489	*	0.489	*	-0.128	no	-0.127	no	-0.019	no	-0.06	no
20	0.239	no	0.239	no	0.27	no	0.212	no	0.486	*	0.35	no
21	0.545	**	0.408	no	-0.353	no	-0.36	no	-0.366	no	-0.295	no
22	0.505	*	0.505	*	0.02	no	0.131	no	0.042	no	0.081	no
23	0.165	no	0.136	no	0.176	no	0.082	no	-0.02	no	0.015	no
24	0.455	*	0.382	no	0.19	no	0.19	no	-0.04	no	-0.078	no
25	-0.149	no	-0.149	no	0.26	no	0.22	no	-0.16	no	-0.14	no
26	0.058	no	0.058	no	0.07	no	0.007	no	-0.06	no	0	no
27	0.593	**	0.593	**	0.013	no	-0.007	no	-0.18	no	-0.151	no
28	0.498	*	0.498	*	-0.01	no	0.03	no	0.31	no	0.26	no
29	0.587	**	0.62	**	-0.14	no	-0.15	no	0.25	no	0.13	no
30	0.421	no	0.42	no	0.19	no	0.16	no	0.127	no	0.12	no
31	0.72	**	0.72	**	-0.156	no	-0.06	no	-0.175	no	-0.149	no

As shown in Table 15, all parametric correlations of the segmental heterozygosity between STRH_e and SNPH_e were not significant (P<0.01) except for chromosome 10 where significant correlation (P<0.05) between STRH_e and SNPH_e based on GD2 in the short arm. Spearman's method revealed similar results.

All results were similar for level3 comparisons and correlations between STRH_e and SNPH_e by using either GD1 or GD2 to calculate SNPHe. GD1 and GD2 showed significant correlations for all chromosomes except short arm of chromosome 10.

6.3.3.2 Runs of homozygosity test ROH

The ROH test using SNPs data of GD1 subset gave 36 cases of homozygous segments distributed in 15 chromosomes. Five STRs markers (TKY007, NVHEQ81, COR070, UCDEQ465, and LEX023) were found in only four segments. The other segments did not have any overlapping microsatellites in the tested data. When ROH test was applied to GD2 subset, no homozygous segments were found using the same criteria as with GD1.

Table 15: Correlations of the segmental heterozygosity between SNPH_e and STRH_e. Green: P<0.01. Blue: P<0.05.

segments	GD1 H_e by GD2 H_e correlations				GD1 He by STRs H_e correlations				GD2 H_e by STRs He correlations			
	Pairwise		Non-parametric		Pairwise		Non-parametric		Pairwise		Non-parametric	
	R	*P*	ρ	*P*	R	*P*	ρ	*P*	R	*P*	ρ	*P*
1p	0.806	**	0.71	**	0.215	no	0.24	no	0.11	no	0.077	no
2p	0.875	**	0.915	**	0.009	no	-0.007	no	-0.27	no	-0.08	no
3p	0.53	*	0.50	*	-0.13	no	-0.151	no	0.15	no	0.13	no
4p	0.76	**	0.635	**	0.04	no	-0.002	no	0.157	no	0.18	no
5p	0.608	**	0.55	**	-0.26	no	-0.269	no	0.037	no	0.057	no
6p	0.594	**	0.566	**	-0.357	no	-0.328	no	-0.158	no	-0.161	no
7p	0.891	**	0.745	**	0.256	no	0.231	no	0.204	no	0.136	no
8p	0.812	**	0.85	**	-0.124	no	-0.215	no	-0.088	no	-0.188	no
9p	0.721	**	0.57	**	-0.054	no	-0.111	no	0.15	no	0.097	no
10p	0.116	no	0.397	no	-0.28	no	-0.463	*	-0.436	*	-0.26	no
11p	0.878	**	0.827	**	-0.279	no	-0.226	no	-0.20	no	-0.135	no
12p	0.755	**	0.61	**	0	no	0	no	0	no	0	no
13p	0.76	**	0.675	**	0.148	no	0.07	no	-0.07	no	-0.231	no
1q	0.691	**	0.699	**	-0.186	no	-0.31	no	-0.002	no	-0.08	no
2q	0.642	**	0.59	**	0.046	no	0.065	no	-0.31	no	-0.342	no
3q	0.719	**	0.817	**	-0.401	no	-0.419	no	-0.316	no	-0.359	no
4q	0.703	**	0.715	**	-0.02	no	0.06	no	0.01	no	0.16	no
5q	0.439	*	0.43	*	-0.127	no	0.087	no	-0.234	no	-0.21	no
6q	0.597	**	0.574	**	0.392	no	0.341	no	0.141	no	0.305	no
7q	0.843	**	0.836	**	-0.143	no	-0.203	no	-0.223	no	-0.281	no
8q	0.757	**	0.773	**	-0.086	no	-0.077	no	-0.004	no	0.059	no
9q	0.806	**	0.831	**	-0.011	no	0.062	no	-0.021	no	0.112	no
10q	0.58	**	0.634	**	-0.246	no	-0.27	no	-0.288	no	-0.247	no
11q	0.805	**	0.60	**	0.009	no	-0.05	no	0.019	no	-0.03	no
12q	0.815	**	0.61	**	0.176	no	0.22	no	0.172	no	0.05	no
13q	0.693	**	0.446	*	0.155	no	0.36	no	-0.11	no	0.176	no

6.4 Discussion

6.4.1 The general pattern of genetic diversity using STRs and SNPs

Heterogeneity within a chromosome (different regions of the same chromosome could be compositionally different) and uniformity between chromosomes have been reported in eukaryote DNA sequences including plants, animals and yeasts (Bernardi, 1989; Li*, et al.*, 1998a; Tenaillon*, et al.*, 2001). In the present study, STRs markers did show uniformity among chromosomes, but could not reveal a clear pattern of heterogeneity between short and long arms along the genome. However, microsatellites are known to have high level of variation due to the multi-allelic polymorphism (Varela and Amos, 2010). Therefore, the unequal variance within a chromosome coming from the large range between the highest and the lowest STRH_e values for a chromosome might mask any possible differences among chromosomes in a genome. For example, in chromosome 5 the multi locus expected heterozygosity calculated from six polymorphic markers showed a range of 100%, and this chromosome does not show significant differences with most of the other chromosomes (Figure 23.A). The results here perfectly agree with Payseur*, et al.* (2011) “Statements about average microsatellite polymorphism mask remarkable heterogeneity in the levels of variation among loci”. The patterns of microsatellite polymorphism are intimately tied to the mutational process (Payseur*, et al.*, 2011). The STRs mutational process depends mainly on the mutation rate that is dependent on different complex factors such as recombination, GC rich and poor regions and the length and the number of tandem repeats (Brinkmann*, et al.*, 1998; Chakraborty*, et al.*, 1997; Ellegren, 2004; Molla*, et al.*, 2009; Payseur*, et al.*, 2011).

Therefore, another segmentation comparisons method considering those factors will be more accurate for the study of microsatellite heterozygosity pattern. Consequently, in order to make such different comparison a large number of STRs markers are needed and of course additional cost and time will be needed. In this case, SNPs might be a better tool providing higher density of markers than microsatellite along different regions of a chromosome.

SNP markers in the current study did discriminate heterogeneity between chromosomes as well as within a chromosome. Such finding agrees with Clark and colleagues 2007 where nonrandom distributions were found within and between all chromosomes in *Arabidopsis thaliana* (Clark*, et al.*, 2007). Also a similar finding was reported in human (Nekrutenko and Li, 2000). In contrast, uniformity among chromosomes has been reported in eukaryotic organisms (Li*, et al.*, 1998b) with an argument that inter-chromosomal uniformity might have happened through repeated polyploidization that occurred in many plant and animal genomes (Holland and GarciaFernandez, 1996; Spring, 1997). However, a recent study (Frenkel*, et al.*, 2012) using whole-genome sequences analyzed the heterogeneity of many vertebrate genomes and reported that genomes of higher eukaryotes are a mosaic of segments with various functions and evolutionary properties. Frenkel*, et al.* (2012) found wide variation among chromosomes in several taxonomic groups, including horses, where non-proportional distribution of variations was found among chromosomes. The pattern of SNP variation among chromosomes in the current study might be explained by variation of chromosomal features such as different GC content, repeated elements and gene density in each chromosome. Differences between chromosomes in human have been reported as results of GC and gene rich areas (Dunham*, et al.*, 1999; Grimwood*, et al.*, 2004; Hillier*, et al.*, 2005), mutation rate (Malcom*, et al.*, 2003) and repeated elements variation (Grimwood*, et al.*, 2004; Hillier*, et al.*, 2005; Zody*, et al.*, 2006).

The comparison within each chromosome showed that SNPH$_e$ was significantly different between arms and centromeric regions. In 14 out of 31 chromosomes SNPHe was higher in the centromeric regions than in the arms. Centromeric regions usually have reduced recombination rates and are expected to have low genetic variation, whereas arms exhibit more genetic diversity (Stephan and Langley, 1998). The reasons that some centromeric regions showed higher diversity than other arms probably is due to: first, there is no clear border between the centromeric region and chromosome arms. Second, the low number of SNPs represented from the centromeric regions in the used platform. Third, it is possible that some of these SNPs are not located in the correct physical position on a centromere or might be incorrectly assembled because of the centromere repositioning phenomena in horse (Carbone*, et al.*, 2006). Furthermore, the complex and repetitive structure of the centromeric regions makes studying this region highly difficult (Alkan*, et al.*, 2011; Neumann*, et al.*, 2012).

6.4.2 Comparisons between the genetic diversity of STRs and SNPs

From a statistical point of view, the heterozygosity values calculated from SNP data (SNPH$_e$) in the current study looked conserved and tended to be more normally distributed with a lower level of variance than STRH$_e$ as shown in Figures 6.10. The STR-based heterozygosity was significantly (P<0.0) higher than SNPs-based heterozygosity (usually by 2-folds) at all levels of comparison. Many studies have compared the heterozygosity between microsatellite and SNPs (Coates*, et al.*, 2009; Forstmeier*, et al.*, 2012; Varela and Amos, 2010; Wong*, et al.*, 2004b), and all reported that microsatellites have higher heterozygosity than SNP markers. This is likely because SNPs are bi-allelic and they have lower mutation rate compared with STRs. Also, microsatellite-based heterozygosity tends to be dominated by small number of markers that are usually used because of their high variability. Very few exceptions have been found where heterozygosity values calculated from both markers type were similar such as for chromosomes 7, 27, 8p and 11p in this study. For chromosomes 27, 8p, and 11p, STRs and SNPs comparison results should be taken with caution because a limited number of microsatellites were tested. In the case of chromosome 7, nine STR loci have been tested. However, five of these loci were reported as low variability microsatellites with only two alleles in horses (Hirota*, et al.*, 2001; Tozaki*, et al.*, 2000).

Our results also showed no correlation between STRs and SNPs based heterozygosity using both parametric and nonparametric statistical methods at all tested genomic levels. This result supports the description of the heterozygosity shown above where STRHe and SNPHe patterns were completely different. However, runs of homozygosity test of SNPs and associated STRs noticeably showed that all associated loci (TKY007, NVHEQ81, COR070, UCDEQ465, and LEX023) were homozygous in the matched case even though these markers were polymorphic in the other individuals. Association between SNPs and microsatellite markers within ROH through the human genome has been reported (Wong*, et al.*, 2004b).

6.5 Conclusion

The present study describes the distribution of heterozygosity in the horse genome using two types of polymorphic molecular markers: STRs and SNPs. The pattern of genetic diversity was completely different between these two markers and there was no correlation between these two patterns. Although limited number of tested STR loci associated with SNPs within runs of homozygosity segments were homozygous, the results are still interesting and need to be augmented by genotyping more loci within ROH segments. Finally, using molecular markers that have different mutation rate such as STRs and SNPs is useful to discover the complexity of a genome to understand the evolutionary history in organisms. More interestingly, having the whole genome sequencing of an organism gives the ability to perform unlimited comparisons by extraction of different markers along a genome using different segmentations and bioinformatics models where better view can be illustrated.

CHAPTER VII

CONCLUSION

There have been quite a few studies about the genetic diversity in the Arabian horse breed throughout the world. In this study I investigated the genetic structure of samples representing Middle Eastern and Western populations using microsatellite markers and whole mtDNA D-loop sequencing. The unique aspect of this study was that it is the first to look at different geographic populations of a single type of horse. Two important findings were that the populations from the Middle East were more genetically variable than those from Europe or North America. This result supports the idea that the Middle East is the place of origin of the important horse breed. The second finding was that North American Arabian horse populations have quite low variability and that some of these populations might be in danger of suffering the effects of inbreeding.

Another part of the research was an examination of the maternally based breeding system used by Arabian horse breeders, which is almost unique in domestic animal breeding. This analysis was based upon testing of the maternally inherited mitochondrial DNA. The research showed that there was no evidence that the Arabian breed has clear subdivisions depending on the traditional strain classification system.

This study will facilitate developing and implementing conservation programs for this important breed throughout the world. The data from this study also provided new information for exploring the evolution history of domestication and breed origins which will contribute to international biodiversity programs. This work will contribute to both the scientific and economic aspects of horse breeding, and will guide breeding process and support the population management of such important animals. The current study was done just before the Syrian armed conflict started in Syria two years ago. It is very well known that Syria is a candidate place of origin of many species including horses. The outcomes of this study will help to recover the Syrian horse populations affected during the war. As well as the maternal inheritance results will help to track any horses that might be illegally transformed out of the country during the war time. That is very important to preserve genetic diversity in hot spots areas of genetic diversity such as Syria.

There was one more part of this research that was separate from the Arabian horse work. This concerned a whole genome comparison of two different types of genetic variants with different mutations rates. A study of this type has never been done using the horse as a model. The results showed completely different patterns of variation between the variant types. However, there was a suggestion that regions of the genome that show high levels of homozygosity for single nucleotide variants also have homozygosity for microsatellite type variants. This could reflect strong selection for these areas of the genome.

Finally, the results from this study could be applied not only in horse populations but also in other animal species.

REFERENCES

Aberle KS, Hamann H, Drogemuller C, Distl O. 2004 Genetic diversity in German draught horse breeds compared with a group of primitive, riding and wild horses by means of microsatellite DNA markers. Animal Genetics. 35(4):270-277.

Aberle KS, Hamann H, Drogemuller C, Distl O. 2007 Phylogenetic relationships of German heavy draught horse breeds inferred from mitochondrial DNA D-loop variation. Journal of Animal Breeding and Genetics. 124(2):94-100.

Achilli A, Olivieri A, Soares P, Lancioni H, Kashani BH, *et al.* 2012 Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proceedings of the National Academy of Sciences of the United States of America. 109(7):2449-2454.

Achmann R, Curik I, Dovc P, Kavar T, Bodo I, *et al.* 2004 Microsatellite diversity, population subdivision and gene flow in the Lipizzan horse. Animal Genetics. 35(4):285-292.

Alho JS, Valimaki K, Merila J. 2010 Rhh: an R extension for estimating multilocus heterozygosity and heterozygosity-heterozygosity correlation. Molecular Ecology Resources. 10(4):720-722.

Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'brien SJ, *et al.* 2011 Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Research. 21(1):137-145.

Allendorf FW. 1986 Genetic drift and the loss of alleles versus heterozygosity. Zoo Biology. 5(2):181-190.

Arabian Horse Bureau. 2006 The Syrian Stud Book of Arabian Horses. Syrian Ministry of Agriculture and Agrarian Reform, Damascus, Syria.

Arnason E, Palsson S. 1996 Mitochondrial cytochrome b DNA sequence variation of Atlantic cod *Gadus morhua*, from Norway. Molecular Ecology. 5(6):715-724.

Avise JC. 1994 *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York, NY.

Ball AD, Stapley J, Dawson DA, Birkhead TR, Burke T*, et al.* 2010 A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). BMC Genomics. 11:218

Balloux F, Lugon-Moulin N. 2002 The estimation of population differentiation with microsatellite markers. Molecular Ecology. 11(2):155-165.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995 Mitochondrial portraits of human populations using median networks. Genetics. 141(2):743-753.

Barker JSF. 2001 Conservation and management of genetic diversity: a domestic animal perspective. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere. 31(4):588-595.

Barker JSF, Moore SS, Hetzel DJS, Evans D, Tan SG*, et al.* 1997 Genetic diversity of Asian water buffalo (*Bubalus bubalis*): microsatellite variation and a comparison with protein-coding loci. Animal Genetics. 28(2):103-115.

Behl R, Behl J, Gupta N, Gupta SC. 2007 Genetic relationships of five Indian horse breeds using microsatellite markers. Animal. 1:483-488.

Belkhir K, Borsa P, Chikhi L, Raufaste, Bonhomme F. 1996-2004. GENETIX 4.05, Windows^{TM} sofrtware for population genetics. Montpellier, France: Universite Montpellier II.

Bernardi G. 1989 The isochore organization of the human genome. Annual Review of Genetics. 23:637-661.

Bigi D, Perrotta G. 2012 Genetic structure and differentiation of the Italian Catria horse. Journal of Heredity. 103(1):134-139.

Binns MM, Boehler DA, Bailey E, Lear TL, Cardwell JM*, et al.* 2011 Inbreeding in the Thoroughbred horse. Animal Genetics. 43(3): 340–342.

Binns MM, Holmes NG, Holliman A, Scott AM. 1995 The identification of polymorphic microsatellites loci in the horse and their use in thoroughbred parentage testing. British Veterinary Journal. 151:9-15.

Boore JL. 1999 Animal mitochondrial genomes. Nucleic Acids Research. 27(8):1767-1780.

Botstein D, White RL, Skolnick M, Davis RW. 1980 Construction of a genetic-linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics. 32(3):314-331.

Bowling A, Ruvinsky A. 2000 *The Genetics of the Horse*. CAB International, Wallingford, UK.

Bowling AT, Del Valle A, Bowling M. 2000 A pedigree-based study of mitochondrial D-loop DNA sequence variation among Arabian horses. Animal Genetics. 31(1):1-7.

Bowling AT, Egglestonstott ML, Byrns G, Clark RS, Dileanis S*, et al.* 1997 Validation of microsatellite markers for routine horse parentage testing. Animal Genetics. 28(4):247-252.

Bowling AT, Penedo MCT, Stott ML, Malyj W. 1993. Introduction of DNA testing for horse and cattle parentage verification. The IVth International Symposium on Human Identification; Scottsdale, AZ. p. 75-80.

Brand MD. 1997 Regulation analysis of energy metabolism. Journal of Experimental Biology. 200(2):193-202.

Breen M, Lindgren G, Binns MM, Norman J, Irvin Z*, et al.* 1997 Genetical and physical assignments of equine microsatellites: first integration of anchored markers in horse genome mapping. Mammalian Genome. 8(4):267-273.

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B. 1998 Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. American Journal of Human Genetics. 62(6):1408-1415.

Brookfield JFY. 1996 A simple new method for estimating null allele frequency from heterozygote deficiency. Molecular Ecology. 5(3):453-455.

Bruford MW, Bradley DG, Luikart G. 2003 DNA markers reveal the complexity of livestock domestication. Nature Reviews Genetics. 4(11):900-910.

Bruno-De-Sousa C, Martinez AM, Ginja C, Santos-Silva F, Carolino MI*, et al.* 2011 Genetic diversity and population structure in Portuguese goat breeds. Livestock Science. 135(2-3):131-139.

Bulmer MG. 1971 The effect of selection on genetic variability. The American Naturalist. 105:201-211.

Cai DW, Tang ZW, Han L, Speller CF, Yang DYY*, et al.* 2009 Ancient DNA provides new insights into the origin of the Chinese domestic horse. Journal of Archaeological Science. 36(3):835-842.

Carbone L, Nergadze SG, Magnani E, Misceo D, Cardone MF*, et al.* 2006 Evolutionary movement of centromeres in horse, donkey, and zebra. Genomics. 87(6):777-782.

Cavalli-Sforza LL, Edwards AWF. 1967 Phylogenetic analysis models and estimation procedures. Evolution. 21(3):550-570.

Cervantes I, Molina A, Goyache F, Gutierrez JP, Valera M. 2008 Population history and genetic variability in the Spanish Arab horse assessed via pedigree analysis. Livestock Science. 113(1):24-33.

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. 1997 Relative mutation rates at di, tri, and tetranucleotide microsatellite loci. Proceedings of the National Academy of Sciences of the United States of America. 94(3):1041-1046.

Charlesworth B, Morgan MT, Charlesworth D. 1993 The effect of deleterious mutations on neutral molecular variation. Genetics. 134(4):1289-1303.

Chen XJ, Butow RA. 2005 The organization and inheritance of the mitochondrial genome. Nature Reviews Genetics. 6(11):815-825.

Chen XJ, Wang XW, Kaufman BA, Butow RA. 2005 Aconitase couples metabolic regulation to mitochondrial DNA maintenance. Science. 307(5710):714-717.

Cieslak M, Pruvost M, Benecke N, Hofreiter M, Morales A*, et al.* 2010 Origin and history of mitochondrial DNA lineages in domestic horses. PLoS ONE. 5(12): e15311.

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G*, et al.* 2007 Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science. 317(5836):338-342.

Coates BS, Sumerford DV, Miller NJ, Kim KS, Sappington TW*, et al.* 2009 Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. Journal of Heredity. 100(5):556-564.

Conant EK, Juras R, Cothran EG. 2012 A microsatellite analysis of five colonial Spanish horse populations of the southeastern United States. Animal Genetics. 43(1):53-62.

Coogle L, Bailey E, Reid R, Russ M. 1996 Equine dinucleotide repeat polymorphisms at loci LEX002, -003, -004, -005, -007, -008, -009, -10, -011, -013 and -014. Animal Genetics. 27(2):126-127.

Costa MaP, Bressel RMC, Almeida DB, Oliveira PA, Bassini LN*, et al.* 2010 Genotyping in the Brazilian Criollo horse stud book: resources and perspectives. Genetics and Molecular Research. 9(3):1645-1653.

Cothran EG, Juras R, Macijauskiene V. 2005 Mitochondrial DNA D-loop sequence variation among 5 maternal lines of the Zemaitukai horse breed. Genetics and Molecular Biology. 28(4):677-681.

Cothran EG, Luis C. 2005 Genetic distance as a tool in the conservation of rare horse breeds. Conservation Genetics of Endangered Horse Breeds.(116):55-71.

Cozzi MC, Strillacci MG, Valiati P, Bighignoli B, Cancedda M*, et al.* 2004 Mitochondrial D-loop sequence variation among Italian horse breeds. Genetics Selection Evolution. 36(6):663-672.

Di Stasio L, Perrotta G, Blasi M, Lisa C. 2008 Genetic characterization of the Bardigiano horse using microsatellite markers. Italian Journal of Animal Science. 7(2):243-250.

Diamond J. 2002 Evolution, consequences and future of plant and animal domestication. Nature. 418(6898):700-707.

Dunham I, Shimizu N, Roe BA, Chissoe S, Dunham I*, et al.* 1999 The DNA sequence of human chromosome 22. Nature. 402(6761):489-495.

Earl DA, Vonholdt BM. 2011 STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources. 4 (2) 359-361.

Ellegren H. 2004 Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics. 5(6):435-445.

Ellegren H, Johansson M, Sandberg K, Andersson L. 1992 Cloning of highly polymorphic microsatellites in the horse. Animal Genetics. 23(2):133-142.

Epperson BK. 2005 Mutation at high rates reduces spatial structure within populations. Molecular Ecology. 14(3):703-710.

Estoup A, Rousset F, Michalakis Y, Cornuet JM, Adriamanga M*, et al.* 1998 Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). Molecular Ecology. 7(3):339-353.

Evanno G, Regnaut S, Goudet J. 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology. 14(8):2611-2620.

Excoffier L, Laval G, Schneider S. 2005 Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics. 1:47-50.

FAO. 2011 *Molecular Genetic Characterization of Animal Genetic Resources*. FAO, Rome.

Fay JC, Wu CI. 2000 Hitchhiking under positive Darwinian selection. Genetics. 155(3):1405-1413.

Felsenstein J. 1989-2006. PHYLIP (phylogeny inference package). [Internet]. Seattle: Genomes Sciences, Department of Genetics, University of Washington; [accessed 2010 March 4]. Available from: URL http://evolution.genetics.washington.edu/phylip.html.

Firouz L. 1998. Original ancestors of the Turkoman and Caspian horses. First International Conference on Turkoman horses. Ashgabat, Turkmenistan. Avilable at: http://caspian.atomiclightwave.us/The-Original-Ancestors-of-the-Turkoman-Caspian-Horses.pdf. Page 1-7

Forstmeier W, Schielzeth H, Mueller JC, Ellegren H, Kempenaers B. 2012 Heterozygosity-fitness correlations in zebra finches: microsatellite markers can be better than their reputation. Molecular Ecology. 21(13):3237-3249.

Freeman AR, Bradley DG, Nagda S, Gibson JP, Hanotte O. 2006 Combination of multiple microsatellite data sets to investigate genetic diversity and admixture of domestic cattle. Animal Genetics. 37(1):1-9.

Frenkel S, Kirzhner V, Korol A. 2012 Organizational heterogeneity of vertebrate genomes. PLoS ONE. 7(2): e32076.

Fries R, Durstewitz G. 2001 Digital DNA signatures for animal tagging. Nature Biotechnology. 19(6):508-508.

Fu YX. 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 147(2):915-925.

Galtier N, Nabholz B, Glemin S, Hurst GDD. 2009 Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Molecular Ecology. 18(22):4541-4550.

Gärke C, Ytournel F, Bed'hom B, Gut I, Lathrop M*, et al.* 2012 Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. Animal Genetics. 43(4):419-428.

George M, Ryder OA. 1986 Mitochondrial-DNA evolution in the genus *Equus*. Molecular Biology and Evolution. 3(6):535-546.

Georgescu SE, Manea MA, Dudu A, Costache M. 2011 Phylogenetic relationships of the Hucul horse from Romania inferred from mitochondrial D-loop variation. Genetics and Molecular Research. 10(4):4104-4113.

Giacomoni EH, Fernandez-Stolz GP, Freitas TRO. 2008 Genetic diversity in the Pantaneiro horse breed assessed using microsatellite DNA markers. Genetics and Molecular Research. 7(1):261-270.

Gissi C, Iannelli F, Pesole G. 2008 Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity. 101(4):301-320.

Głażewska I. 2010 Speculations on the origin of the Arabian horse breed. Livestock Science. 129(1-3):49-55.

Głażewska I, Gralak B. 2006 Balancing selection in Polish Arabian horses. Livestock Science. 105(1-3):272-276.

Głażewska I, Jezierski T. 2004 Pedigree analysis of Polish Arabian horses based on founder contributions. Livestock Production Science. 90(2-3):293-298.

Glazewska I, Wysocka A, Gralak B, Sell J. 2007 A new view on dam lines in Polish Arabian horses based on mtDNA analysis. Genetics Selection Evolution. 39(5):609-619.

Glenn TC, Schable NA. 2005 Isolating microsatellite DNA loci. Molecular Evolution: Producing the Biochemical Data, Part B. 395:202-222.

Glover KA, Hansen MM, Lien S, Als TD, Hoyheim B*, et al.* 2010 A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. BMC Genetics, 11:2.

Glowatzki-Mullis ML, Muntwyler J, Pfister W, Marti E, Rieder S*, et al.* 2006 Genetic diversity among horse populations with a special focus on the Franches-Montagnes breed. Animal Genetics. 37(1):33-39.

Goldstein D, Schlötterer C. 1999 *Microsatellites: Evolution and Applications*. Oxford University Press, New York.

Goncalves GL, Moreira GRP, Freitas TRO, Hepp D, Passos DT*, et al.* 2010 Mitochondrial and nuclear DNA analyses reveal population differentiation in Brazilian Creole sheep. Animal Genetics. 41(3):308-310.

Goodman S. 1997 Rst Calc: a collection of computer programs for calculating estimates of genetic differentition from microsatellite data and a determining their significance. Molecular Ecology.(6):881-885.

Goudet J. 1995 FSTAT (Version 1.2): a computer program to calculate F-statistics. Journal of Heredity. 86(6):485-486.

Goudet J. 2002. FSTAT, a program to estimate and test gene diversities and fixation indices [Internet]; [accessed 2011 October 7]. Available from: URL http://www2.unil.ch/popgen/softwares/fstat.htm.

Graeber MB, Muller U. 1998 Recent developments in the molecular genetics of mitochondrial disorders. Journal of the Neurological Sciences. 153(2):251-263.

Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J*, et al.* 2004 The DNA sequence and biology of human chromosome 19. Nature. 428(6982):529-535.

Guastella AM, Zuccaro A, Criscione A, Marletta D, Bordonaro S. 2011 Genetic analysis of Sicilian autochthonous horse breeds using nuclear and mitochondrial DNA markers. Journal of Heredity. 102(6):753-758.

Guerin G, Bertaud M, Amigues Y. 1994 Characterization of seven new horse microsatellites: HMS1, HMS2, HMS3, HMS5, HMS6, HMS7 and HMS8. Animal Genetics. 25(1):62.

Gulcher J. 2012 Microsatellite markers for linkage and association studies. Cold Spring Harb Protoc. 2012(4):425-432.

Gurney SM, Schneider S, Pflugradt R, Barrett E, Forster AC*, et al.* 2010 Developing equine mtDNA profiling for forensic application. International Journal of Legal Medicine. 124(6):617-622.

Hardy OJ, Charbonnel N, Freville H, Heuertz M. 2003 Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. Genetics. 163(4):1467-1482.

Hartl D, Clark A. 1997 *Principles of Population Genetics*. Sinauer Associates Inc., Sunderland, MA.

Hauser L, Baird M, Hilborn R, Seeb LW, Seeb JE. 2011 An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. Molecular Ecology Resources. 11:150-161.

Hayes B, Laerdahl JK, Lien S, Moen T, Berg P*, et al.* 2007 An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. Aquaculture. 265(1-4):82-90.

Haynes GD, Latch EK. 2012 Identification of novel single nucleotide polymorphisms (SNPs) in deer (*Odocoileus* spp.) using the BovineSNP50 BeadChip. PLoS ONE. 7(5): e36536.

Hedrick PW. 1999 Perspective: highly variable loci and their interpretation in evolution and conservation. Evolution. 53(2):313-318.

Hedrick PW. 2005 *Genetics of Populations*. Jones and Bartlett, Boston, MA.

Hendricks BL. 1995 *International Encyclopedia of Horse Breeds*. University of Oklahoma Press, Norman.

Hill EW, Bradley DG, Al-Barody M, Ertugrul O, Splan RK*, et al.* 2002 History and integrity of thoroughbred dam lines revealed in equine mtDNA variation. Animal Genetics. 33(4):287-294.

Hillier LW, Graves TA, Fulton RS, Fulton LA, Pepin KH*, et al.* 2005 Generation and annotation of the DNA sequences of human chromosomes 2 and 4. Nature. 434(7034):724-731.

Hirota K, Tozaki T, Mashima S, Miura N. 2001 Cytogenetic assignment and genetic characterization of the horse microsatellites, TKY4-18, TKY20, TKY22-24, TKY30-41 derived from a cosmid library. Animal Genetics. 32(3):160-161.

Holland PWH, Garciafernandez J. 1996 Hox genes and chordate evolution. Developmental Biology. 173(2):382-395.

Hoque MO, Lee CC, Cairns P, Schoenberg M, Sidransky D. 2003 Genome-wide genetic characterization of bladder cancer: a comparison of high-density single-nucleotide polymorphism arrays and PCR-based microsatellite analysis. Cancer Res. 63(9):2216-2222.

Hudson RR, Kaplan NL. 1988 The coalescent process in models with selection and recombination. Genetics. 120(3):831-840.

Humphries EM, Winker K. 2011 Discord reigns among nuclear, mitochondrial and phenotypic estimates of divergence in nine lineages of trans-Beringian birds. Molecular Ecology. 20(3):573-583.

Irvin Z, Giffard J, Brandon R, Breen M, Bell K. 1998 Equine dinucleotide repeat polymorphisms at loci ASB 21, 23, 25 and 37-43. Animal Genetics. 29(1):67-67.

Ishida N, Hasegawa T, Takeda K, Sakagami M, Onishi A*, et al.* 1994 Polymorphic sequence in the D-Loop region of equine mitochondrial-DNA. Animal Genetics. 25(4):215-221.

Ivankovic A, Ramljak J, Konjacic M, Kelava N, Dovc P*, et al.* 2009 Mitochondrial D-loop sequence variation among autochthonous horse breeds in Croatia. Czech Journal of Animal Science. 54(3):101-111.

Iwanczyk E, Juras R, Cholewinski G, Cothran EG. 2006 Genetic structure and phylogenetic relationships of the Polish Heavy horse. Journal of Applied Genetics. 47(4):353-359.

Jakobsson M, Rosenberg NA. 2007 CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 23(14):1801-1806.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, Vanliere JM*, et al.* 2008 Genotype, haplotype and copy number variation in worldwide human populations. Nature. 451(7181):998-1003.

Jamieson A, Taylor SCS. 1997 Comparison of three probability formulae for parentage exclusion. Animal. Genetics. 28:397-400.

Jansen T, Forster P, Levine MA, Oelke H, Hurles M*, et al.* 2002 Mitochondrial DNA and the origins of the domestic horse. Proceedings of the National Academy of Sciences of the United States of America. 99(16):10905-10910.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE*, et al.* 2000 The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. American Journal of Human Genetics. 66(3):979-988.

Juras R, Cothran EG. 2004 Microsatellites in Lithuanian native horse breeds: usefulness for parentage testing. Biologija. 4:6–9.

Juras R, Cothran EG, Klimas R. 2003 Genetic analysis of three Lithuanian native horse breeds. Acta Agriculturae Scandinavica Section Animal Science. 53(4):180-185.

Kakoi H, Tozaki T, Gawahara H. 2007 Molecular analysis using mitochondrial DNA and microsatellites to infer the formation process of Japanese native horse populations. Biochemical Genetics. 45(3-4):375-395.

Kang BT, Kim KS, Min MS, Chae YJ, Kang JW*, et al.* 2009 Microsatellite loci analysis for the genetic variability and the parentage test of five dog breeds in South Korea. Genes & Genetic Systems. 84(3):245-251.

Kang J, Li X, Zhou R, Li L, Zheng G*, et al.* 2011 Genetic diversity and differentiation of four goat lineages based on analysis of complete mtDNA D-loop. Frontiers of Agriculture in China. 5(1):87-93.

Kavar T, Dovc P. 2008 Domestication of the horse: genetic relationships between domestic and wild horses. Livestock Science. 116(1-3):1-14.

Kavar T, Habe F, Brem G, Dovc P. 1999 Mitochondrial D-loop sequence variation among the 16 maternal lines of the Lipizzan horse breed. Animal Genetics. 30(6):423-430.

Khanshour A, Conant E, Juras R, Cothran EG. 2013 Microsatellite analysis of genetic diversity and population structure of arabian horse populations. Journal of Heredity. 104(3):386-398.

Kim KI, Yang YH, Lee SS, Park C, Ma R*, et al.* 1999 Phylogenetic relationships of Cheju horses to other horse breeds as determined by mtDNA D-loop sequence polymorphism. Animal Genetics. 30(2):102-108.

Kimura M. 1980 A simple methodfor estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 16:111-120.

Kirin M, Mcquillan R, Franklin CS, Campbell H, Mckeigue PM*, et al.* 2010 Genomic runs of homozygosity record population history and consanguinity. PLoS ONE. 5(11): e13996 .

Koban E, Denizci M, Aslan O, Aktoprakligil D, Aksu S*, et al.* 2011 High microsatellite and mitochondrial diversity in Anatolian native horse breeds shows Anatolia as a genetic conduit between Europe and Asia. Animal Genetics. 43(4):401–409.

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE. 2006 Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics. 7(2):295-302.

Lebart L, Morineau A, Warwick K. 1984 *Multivariate Descriptive Statistical Analysis*. J. Wiley, New York.

Lee SY, Cho GJ. 2006 Parentage testing of Thoroughbred horse in Korea using microsatellite DNA typing. Journal of Veterinary Science. 7(1):63-67.

Leroy G, Callede L, Verrier E, Meriaux JC, Ricard A*, et al.* 2009 Genetic diversity of a large set of horse breeds raised in France assessed by microsatellite polymorphism. Genetics Selection Evolution, 41:5.

Li W, Stolovitzky G, Bernaola-Galva´N P, Oliver JL. 1998a Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Genome Research. 8:916-928.

Li WT, Stolovitzky G, Bernaola-Galvan P, Oliver JL. 1998b Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Genome Research. 8(9):916-928.

Librado P, Rozas J. 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25(11):1451-1452.

Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO*, et al.* 2000 Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nature Biotechnology. 18(9):1001-1005.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB*, et al.* 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 438(7069):803-819.

Lippold S, Matzke NJ, Reissmann M, Hofreiter M. 2011 Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. BMC Evolutionary Biology. 11:328.

Locke M, Baack E, Toonen R. 2000. STRand software [Internet]: University of California at Davis; [accessed 2010 October 14]. Available from: URL http://www.vgl.ucdavis.edu/informatics/STRand/.

Luikart G, Allendorf FW, Cornuet JM, Sherwin WB. 1998 Distortion of allele frequency distributions provides a test for recent population bottlenecks. Journal of Heredity. 89(3):238-247.

Luís C, Cothran EG, Oom EE. 2002 Microsatellites in Portuguese autochthonous horse breed: usefulness for parentage testing. Genetics and Molecular Biology. 25:131-134.

Luís C, Cothran EG, Oom MDM. 2007 Inbreeding and genetic structure in the endangered Sorraia horse breed: implications for its conservation and management. Journal of Heredity. 98(3):232-237.

Machugh DE, Shriver MD, Loftus RT, Cunningham P, Bradley DG. 1997 Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and Zebu cattle (Bos taurus and Bos indicus). Genetics. 146(3):1071-1086.

Malcom CM, Wyckoff GJ, Lahn BT. 2003 Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. Molecular Biology and Evolution. 20(10):1633-1641.

Marklund S, Ellegren H, Eriksson S, Sandberg K, Andersson L. 1994 Parentage testing and linkage analysis in the horse using a set of highly polymorphic microsatellites. Animal Genetics. 25(1):19-23.

Marletta D, Tupac-Yupanqui I, Bordonaro S, Garcia D, Guastella AM*, et al.* 2006 Analysis of genetic diversity and the determination of relationships among western

Mediterranean horse breeds using microsatellite markers. Journal of Animal Breeding and Genetics. 123(5):315-325.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM. 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology. 7(5):639-655.

Mccue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E*, et al.* 2012 A high density SNP array for the domestic horse and extant perissodactyla: utility for association mapping, genetic diversity, and Phylogeny Studies. PLoS Genetics. 8(1):e1002451.

Mcilwraith CW, Frisbie DD, Kawcak CE, Fuller CJ, Hurtig M*, et al.* 2010 The OARSI histopathology initiative recommendations for histological assessments of osteoarthritis in the horse. Osteoarthritis and Cartilage. 18:S93-S105.

Miller RD, Phillips MS, Jo I, Donaldson MA, Studebaker JF*, et al.* 2005 High-density single-nucleotide polymorphism maps of the human genome. Genomics. 86(2):117-126.

Moen T, Hayes B, Baranski M, Berg PR, Kjoglum S*, et al.* 2008 A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. BMC Genomics. 9: 223.

Molla M, Delcher A, Sunyaev S, Cantor C, Kasif S. 2009 Triplet repeat length bias and variation in the human transcriptome. Proceedings of the National Academy of Sciences of the United States of America. 106(40):17095-17100.

Monies D, Abu Al Saud N, Sahar N, Meyer BF. 2011 Population studies and parentage testing for Arabian horses using 15 microsatellite markers. Animal Genetics. 42(2):225-226.

Morin PA, Luikart G, Wayne RK, Grp SW. 2004 SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution. 19(4):208-216.

Morin PA, Martien KK, Taylor BL. 2009 Assessing statistical power of SNPs for population structure and conservation studies. Molecular Ecology Resources. 9(1):66-73.

Moritz C, Dowling TE, Brown WM. 1987 Evolution of animal mitochondrial-DNA - relevance for population biology and systematics. Annual Review of Ecology and Systematics. 18:269-292.

Mukesh M, Sodhi M, Kataria RS, Mishra BP. 2009 Use of microsatellite multilocus genotypic data for individual assignment assay in six native cattle breeds from north-western region of India. Livestock Science. 121(1):72-77.

Musick JA. 2005 *Management Techniques for Elasmobranch Fisheries*. FAO, Rome, Italy.

Narum SR, Banks M, Beacham TD, Bellinger MR, Campbell MR*, et al.* 2008 Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. Molecular Ecology. 17(15):3464-3477.

Nass MMK, Nass SIJ. 1963 Intramitochondrial fibres with DNA characteristics. Cell Biology. 19:593-611.

Nei M. 1972 Genetic distance between populations. American Naturalist. 106(949):283-292.

Nekrutenko A, Li WH. 2000 Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Research. 10(12):1986-1995.

Nergadze SG, Lupotto M, Pellanda P, Santagostino M, Vitelli V*, et al.* 2010 Mitochondrial DNA insertions in the nuclear horse genome. Animal Genetics. 41:176-185.

Neumann P, Navratilova A, Schroeder-Reiter E, Koblizkova A, Steinbauerova V*, et al.* 2012 Stretching the rules: monocentric chromosomes with multiple centromere domains. PLoS Genetics. 8(6):e1002777.

Notter DR. 1999 The importance of genetic populations diversity in livestock populations of the future. Journal of Animal Science. 77(1):61-69.

Ohta T, Kimura M. 1969 Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics. 63(1):229-238.

Ouragh L. 2005 DNA polymorphism of Arabian, Thoroughbred and Anglo-Arab horses in Morocco - Application to identification and parentage verification of individual horses. Applications of Gene-Based Technologies for Improving Animal Production and Health in Developing Countries.621-629.

Ozkan E, Soysal MI, Ozder M, Koban E, Sahin O*, et al.* 2009 Evaluation of parentage testing in the Turkish Holstein population based on 12 microsatellite loci. Livestock Science. 124(1-3):101-106.

Paetkau D, Strobeck C. 1995 The molecular-basis and evolutionary history of a microsatellite null allele in bears. Molecular Ecology. 4(4):519-520.

Paetkau D, Waits LP, Clarkson PL, Craighead L, Strobeck C. 1997 An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. Genetics. 147(4):1943-1957.

Pariset L, Cuteri A, Ligda C, Ajmone-Marsan P, Valentini A. 2009 Geographical patterning of sixteen goat breeds from Italy, Albania and Greece assessed by single nucleotide polymorphisms. BMC Ecology. 9:20.

Patterson M. 2001 *Equus* - how it all began. Nature Reviews Genetics. 2(3):163-163.

Payseur BA, Cutter AD, Nachman MW. 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Molecular Biology and Evolution. 19(7):1143-1153.

Payseur BA, Jing PC, Haasl RJ. 2011 A genomic portrait of human microsatellite variation. Molecular Biology and Evolution. 28(1):303-312.

Peffers MJ, Milner PI, Tew SR, Clegg PD. 2010 Regulation of SOX9 in normal and osteoarthritic equine articular chondrocytes by hyperosmotic loading. Osteoarthritis and Cartilage. 18(11):1502-1508.

Perrier X, Flori A, Bonnot F. 2003 Data analysis methods. In: Hamon, P., Seguin, M., Perrier, X. ,Glaszmann, J. C. Ed., *Genetic Diversity of Cultivated Tropical Plants*. Enfield Science Publishers, Montpellier, France.

Perrier X, Jacquemoud-Collet JP. 2006. DARwin software [Internet]; [accessed 2011 April 22]. Available from: URL http://darwin.cirad.fr/darwin.

Petersen JL, Mickelson JR, Cothran EG, Andersson LS, Axelsson J*, et al.* 2013 Genetic diversity in the modern horse illustrated from genome-wide SNP data. PLoS ONE. 8(1):e54997.

Petit RJ, El Mousadik A, Pons O. 1998 Identifying populations for conservation on the basis of genetic markers. Conservation Biology. 12(4):844-855.

Plante Y, Vega-Pla JL, Lucas Z, Colling D, De March B*, et al.* 2007 Genetic diversity in a feral horse population from Sable Island, Canada. Journal of Heredity. 98(6):594-602.

Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010 Population genetic inference from genomic sequence variation. Genome Research. 20(3):291-300.

Pritchard JK, Stephens M, Donnelly P. 2000 Inference of population structure using multilocus genotype data. Genetics. 155(2):945-959.

Pruski W. 1983 *Dwa Wieki polskiej Hodowli Koni Arabskich (1778-1978) Ijej Sukcesy Na Swiecie*. PWRiL, Warszawa.

Prystupa JM, Hind P, Cothran EG, Plante Y. 2012a Maternal lineages in native Canadian equine populations and their relationship to the Nordic and Mountain and Moorland pony breeds. Journal of Heredity. 103(3):380-390.

Prystupa JM, Juras R, Cothran EG, Buchanan FC, Plante Y. 2012b Genetic diversity and admixture among Canadian, Mountain and Moorland and Nordic pony populations. Animal. 6(1):19-30.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MaR*, et al.* 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 81(3):559-575.

Raymond M, Rousset F. 2001. GENEPOP update of the version described in Raymond, M. Rousset, F (1995)-GENEPOP: population genetics software for exact tests and ecumenicism. Version 3.4.

Reis SP, Goncalves EC, Silva A, Schneider MPC. 2008 Genetic variability and efficiency of DNA microsatellite markers for paternity testing in horse breeds from the Brazilian Marajo archipelago. Genetics and Molecular Biology. 31(1):68-72.

Rengmark AH, Slettan A, Skaala O, Lie O, Lingaas F. 2006 Genetic variability in wild and farmed Atlantic salmon (*Salmo salar*) strains estimated by SNP and microsatellites. Aquaculture. 253(1-4):229-237.

Reynolds J, Weir BS, Cokerham C. 1983 Estimation of the coancestry coefficients: basis for a short-term genetic distance. Genetics. 105:767-779.

Rodrigáñez J, Barragán C, Alves E, Cortázar C, Toro MA*, et al.* 2008 Genetic diversity and allelic richness in Spanish wild and domestic pig population estimated from microsatellite markers. Spanish Journal of Agricultural Research. 6:107-115.

Rosenberg NA. 2004 DISTRUCT: a program for the graphical display of population structure. Molecular Ecology Notes. 4(1):137-138.

Rousset F. 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics. 142(4):1357-1362.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E*, et al.* 2007 Genome-wide detection and characterization of positive selection in human populations. Nature. 449(7164):913-918.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD*, et al.* 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 409(6822):928-933.

Saitou N, Nei M. 1987 The neighbor-joining method a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 4(4):406-425.

Sbisa E, Tanzariello F, Reyes A, Pesole G, Saccone C. 1997 Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. Gene. 205(1-2):125-140.

Schaid DJ, Guenther JC, Christensen GB, Hebbring S, Rosenow C*, et al.* 2004 Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. American Journal of Human Genetics. 75(6):948-965.

Schlotterer C, Kauer M, Dieringer D. 2004 Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality. Proceedings of the Royal Society B-Biological Sciences. 271(1541):869-874.

Seddon JM, Parker HG, Ostrander EA, Ellegren H. 2005 SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. Molecular Ecology. 14(2):503-511.

Selkoe KA, Toonen RJ. 2006 Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecology Letters. 9(5):615-629.

Sereno FTPD, Sereno JRB, Vega-Pla JL, Delado JV. 2008 DNA testing for parentage verification in a conservation nucleus of Pantaneiro horse. Genetics and Molecular Biology. 31(1):64-67.

Shen R, Fan JB, Campbell D, Chang WH, Chen J*, et al.* 2005 High-throughput SNP genotyping on universal bead arrays. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis. 573(1-2):70-82.

Smith CT, Antonovich A, Templin WD, Elfstrom CM, Narum SR*, et al.* 2007 Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon: a comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. Transactions of the American Fisheries Society. 136(6):1674-1687.

Smith JM, Haigh J. 2007 The hitch-hiking effect of a favourable gene. Genetics Research. 89(5-6):391-403.

Solis A, Jugo BM, Meriaux JC, Iriondo M, Mazon LI*, et al.* 2005 Genetic diversity within and among four south European native horse breeds based on microsatellite DNA analysis: implications for conservation. Journal of Heredity. 96(6):670-678.

Spencer CCA, Deloukas P, Hunt S, Mullikin J, Myers S*, et al.* 2006 The influence of recombination on human genetic diversity. PLoS Genetics. 2(9):1375-1385.

Spring J. 1997 Hypothesis vertebrate evolution by interspecific hybridisation - are we polyploid? FEBS Letters. 400(1):2-8.

Stephan W, Langley CH. 1998 DNA polymorphism in Lycopersicon and crossing-over per physical length. Genetics. 150(4):1585-1593.

Stephens JC, Briscoe D, Obrien SJ. 1994 Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. American Journal of Human Genetics. 55(4):809-824.

Syvanen AC. 2005 Toward genome-wide SNP genotyping. Nature Genetics. 37:S5-S10.

Tamura K, Dudley J, Nei M, Kumar S. 2007 MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution. 24(8):1596-1599.

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF*, et al.* 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Natl Acad Sci U S A. 98(16):9161-9166.

Thalamuthu A, Mukhopadhyay I, Ray A, Weeks DE. 2005 A comparison between microsatellite and single-nucleotide polymorphism markers with respect to two measures of information content. BMC Genetics. 6(Suppl 1):S27.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A*, et al.* 2009 The genetic structure and history of Africans and African Americans. Science. 324(5930):1035-1044.

Tozaki T, Kakoi H, Mashima S, Hirota K, Hasegawa T*, et al.* 2001 Population study and validation of paternity testing for Thoroughbred horses by 15 microsatellite loci. Journal of Veterinary Medical Science. 63(11):1191-1197.

Tozaki T, Kakoi H, Mashima S, Hirota K, Hasegawa T*, et al.* 2000 The isolation and characterization of 18 equine microsatellite loci, TKY272-TKY289. Animal Genetics. 31(2):149-150.

Tozaki T, Takezaki N, Hasegawa T, Ishida N, Kurosawa M*, et al.* 2003 Microsatellite variation in Japanese and Asian horses and their phylogenetic relationship using a European horse outgroup. Journal of Heredity. 94(5):374-380.

Upton P, Amirsadeghi H. 1998 *Drinkers of the Wind*. Thames and Hudson, London, UK.

Van De Goor LHP, Van Haeringen WA, Lenstra JA. 2011 Population studies of 17 equine STR for forensic and phylogenetic analysis. Animal Genetics. 42(6):627-633.

Van Haeringen H, Bowling AT, Stott ML, Lenstra JA, Zwaagstra KA. 1994 A highly polymorphic horse microsatellite locus - VHL20. Animal Genetics. 25(3):207-207.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P. 2004 MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes. 4(3):535-538.

Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD*, et al.* 2008 SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods. 5(3):247-252.

Vangestel C, Mergeay J, Dawson DA, Callens T, Vandomme V*, et al.* 2012 Genetic diversity and population structure in contemporary house sparrow populations along an urbanization gradient. Heredity. 109(3):163-172.

Varela MA, Amos W. 2010 Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence. Genomics. 95(3):151-159.

Vignal A, Milan D, Sancristobal M, Eggen A. 2002 A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution. 34(3):275-305.

Vila C, Leonard JA, Gotherstrom A, Marklund S, Sandberg K*, et al.* 2001 Widespread origins of domestic horse lineages. Science. 291(5503):474-477.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S*, et al.* 2009 Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 326(5954):865-867.

Wallner B, Vogl C, Shukla P, Burgstaller JP, Druml T*, et al.* 2013 Identification of genetic variation on the horse Y chromosome and the tracing of male founder lineages in modern breeds. PLoS ONE. 8(4):e60015.

Waples RS. 1989 A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics. 121(2):379-391.

Warmuth V, Manica A, Eriksson A, Barker G, Bower M. 2012 Autosomal genetic diversity in non-breed horses from eastern Eurasia provides insights into historical population movements. Animal Genetics. 44(1):53–61.

Weir BS, Cockerham CC. 1984 Estimating F-statistics for the analysis of population structure. Evolution. 38(6):1358-1370.

Weitzman M. 1993 What to preserve? An application of diversity theory to crane conservation. The Quarterly Journal of Economics. 107:363-405.

Wigginton JE, Cutler DJ, Abecasis GR. 2005 A note on exact tests of Hardy-Weinberg equilibrium. American Journal of Human Genetics. 76(5):887-893.

Wolff JN, Nafisinia M, Sutovsky P, Ballard JWO. 2012 Paternal transmission of mitochondrial DNA as an integral part of mitochondrial inheritance in metapopulations of Drosophila simulans. Heredity. 110:57–62.

Wong GKS, Liu B, Wang J, Zhang Y, Yang X*, et al.* 2004a A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature. 432(7018):717-722.

Wong KK, Tsang YT, Shen J, Cheng RS, Chang YM*, et al.* 2004b Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. Nucleic Acids Research. 32(9):e69.

Wright S. 1978 *Evolution and the Genetics of Population, Variability within and among Natural Populations*. The University of Chicago Press, Chicago.

Xing C, Schumacher FR, Xing G, Lu Q, Wang T*, et al.* 2005 Comparison of microsatellites, single-nucleotide polymorphisms (SNPs) and composite markers derived from SNPs in linkage analysis. BMC Genetics. 6(Suppl 1):S29.

Xu XF, Arnason U. 1994 The complete mitochondrial-DNA sequence of the horse, *Equus* caballus: extensive heteroplasmy of the control region. Gene. 148(2):357-362.

Xu XF, Janke A, Arnason U. 1996 The complete mitochondrial DNA sequence of the greater Indian Rhinoceros, Rhinoceros unicornis, and the phylogenetic relationship among Carnivora, Perissodactyla, and Artiodactyla (+ Cetacea). Molecular Biology and Evolution. 13(9):1167-1173.

Zechner P, Solkner J, Bodo I, Druml T, Baumung R*, et al.* 2002 Analysis of diversity and population structure in the Lipizzan horse breed based on pedigree information. Livestock Production Science. 77(2-3):137-146.

Zhao XB, Chu MX, Li N, Wu CX. 2001 Paternal inheritance of mitochondrial DNA in the sheep (*Ovine aries*). Science in China Series C-Life Sciences. 44(3):321-326.

Zheng HT, Peng ZH, Li S, He L. 2005 Loss of heterozygosity analyzed by single nucleotide polymorphism array in cancer. World Journal of Gastroenterology. 11(43):6740-6744.

Zody MC, Garber M, Adams DJ, Sharpe T, Harrow J*, et al.* 2006 DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature. 440(7087):1045-1049.