Inaugural-Dissertation zur Erlangung der Doktorwürde
der Tierärztlichen Fakultät der Ludwig-Maximilians-Universität
München

Genetic characterization of paternal bloodlines in modern horse breeds based on Y-chromosomal sequence information

von Viktoria Remer

aus Wien

München, 2022

Aus dem Veterinärwissenschaftlichen Department der Tierärztlichen Fakultätder Ludwig-Maximilians-Universität München

Lehrstuhl für Molekulare Tierzucht und Biotechnologie

Arbeit angefertigt unter der Leitung von: Priv.-Doz. Dr. Ivica Međugorac

Angefertigt am: Institut für Tierzucht und Genetik der Veterinärmedizinischen Universität Wien

Mentor: O.Univ. Prof. DI DDr. Habil. Dr.h.c. Gottfried Brem

Gedruckt mit Genehmigung der Tierärztlichen Fakultät

der Ludwig-Maximilians-Universität München

Dekan:	Univ.-Prof. Dr. Reinhard K. Straubinger, Ph.D.
Berichterstatter:	Priv.-Doz. Dr. Ivica Međugorac
Korreferent:	Priv.-Doz. Dr. Elisabeth G. Kemter

Tag der Promotion: 30. Juli 2022

Die vorliegende Arbeit wurde gemäß § 6 Abs. 2 der Promotionsordnung für die Tierärztliche Fakultät der Ludwig-Maximilians-Universität München in kumulativer Form verfasst.

Folgende wissenschaftliche Arbeiten sind in dieser Dissertationsschrift enthalten:

1. Felkel, S., Vogl, C., Rigler, D., **Dobretsberger, V.**, Chowdhary, B. P., Distl, O., Fries, R., Jagannathan, V., Janečka, J. E., Leeb, T., Lindgren, G., McCue, M., Metzger, J., Neuditschko, M., Rattei, T., Raudsepp, T., Rieder, S., Rubin, C. J., Schaefer, R., … Wallner, B.. **The horse Y chromosome as an informative marker for tracing sire lines**, erschienen in *Scientific Reports* 2019, online verfügbar unter doi: 10.1038/s41598-019-42640-w

   (dazwischen Namensänderung durch Heirat von Viktoria Dobretsberger zu Viktoria Remer)

2. **Remer, V.**, Bozlak, E., Felkel, S., Radovic, L., Rigler, D., Grilz-seger, G., Stefaniuk-szmukier, M., Bugno-poniewierska, M., Brooks, S., Miller, D. C., Antczak, D. F., Sadeghi, R., Cothran, G., Juras, R., & Khanshour, A. M.. **Y-Chromosomal Insights into Breeding History and Sire Line Genealogies of Arabian Horses,** erschienen in Genes 2022, online verfügbar unter doi: 10.3390/genes13020229

My children, who gifted me the time for this thesis,

and my father for his great will to live.

TABLE OF CONTENTS

INDEX OF ABBREVIATIONS

AQHA	American Quarter Horse Association
bp	base pair
CE	common era
CH	crown haplogroup
eMSY	equine male-specific region of the Y chromosome
ESSA	European State Studs Association
FAO	Food and Agriculture Organization
FAOSTAT	Food and Agriculture Organization Corporate Statistical Database
HG	haplogroup
HT	haplotype
indels	single base insertions/deletions
kb	kilobases
kya	thousand years ago
LIF	Lipizzan International Federation
Mb	megabase
MRCA	most recent common ancestor
MSY	male-specific, non-recombining region of the Y chromosome
mtDNA	mitochondrial DNA
mya	million years ago
Ne	effective population size
NGS	next-generation sequencing
PAR	pseudoautosomal region
sHG	subhaplogroup
SNP	single nucleotide polymorphism
STRs	Short tandem repeats
TB	Thoroughbred

I. INTRODUCTION

The domesticated horse (*Equus caballus*) is a species inextricably linked to human societies and exhibits a pronounced cultural heritage and economic value (Kelekna, 2009; Librado et al., 2021). Anthropogenic selection gradually shaped horses over millennia and led to the formation of a variety of specialized horse breeds (Nissen, 1997). Horses embody a species with particular importance on paternal lines in breeding. Stallions and sire lines are especially celebrated in modern horse breeds that were created within the recent 200 to 400 years. The spatial and temporal descent of paternal bloodlines contributing to the formation of modern breeds is of great public interest (Nissen, 1998). The Y chromosome, in particular the male-specific, non-recombining region of the Y chromosome, is the genetic compartment that mirrors these male-driven population dynamics and paternal relationships best; since it is inherited uniparentally as a single linkage group from the father to his sons (Batini & Jobling, 2017; Calafell & Larmuseau, 2016).

Studying male demography in populations based on Y chromosome information is well accepted in humans, while in horses it is still at the stage of development (Wallner et al., 2017). With human studies as a role model, this thesis aims to establish the horse Y chromosome as a genetic marker to address population dynamics, male demography, and forensic questions in modern horse breeds.

II. LITERATURE REVIEW

1. Horses: From early domesticates to modern breeds

1.1. Horses and humans

The horse is one of seven species in the equid family. Together with the domesticated donkey (*Equus asinus*), horses are the only equines that are domesticated today (Librado & Orlando, 2021; Wang et al., 2020).

The first evidence for horse husbandry dates back to around 5.5 thousand years ago (kya) and is given by the archaeological site of Botai in the Pontic Caspian steppe, modern-day Kazakhstan (Outram et al., 2009). However, as recently shown, modern horses do not descend from these early domesticates and also not from horses in other hypothetical domestication regions, like Iberia (Fages et al., 2019) and Anatolia (Guimaraes et al., 2020). None of these domestication centers contributed significantly to the genetic pool of modern horses. The cradle of extant domestic lineages has rather been located on the Western Eurasian steppes (Librado et al., 2021). From there, anthropogenic dispersal facilitated, some four kya, the rapid spread of domestic horses throughout Eurasia (Orlando, 2020).

All present horses originate from the Eurasian wild horse population; including those on the American continent which are either domesticated or feral (Luís et al., 2006). The wild horse populations on this continent got extinct due to a mass reduction of large mammalian species about 12 kya (Burney & Flannery, 2005; Faith & Surovell, 2009). It is notable that no truly wild ancestor of domestic horses is roaming any continent of the planet today. Recent genetic work revealed that even Przewalski's horses (*Equus przewalskii*), a today wild living sister taxon of domestic horses, are genetically quite similar to the horses herded by the Botai culture; thus the authors claimed them as feral descendants of these early domesticates (Gaunitz et al., 2018).

In the course of history, horses were initially used as a source of meat and milk similar to other ungulates (Outram et al., 2009). In addition, conjointly with donkeys, horses were soon used for transportation and human mobility (Wang et al., 2020). Horses' stamina and speed revolutionized warfare, transportation, and agriculture. By facilitating cultural exchange, horses contributed to the spread of languages, diseases, religions, science, and art (Kelekna, 2009; Librado et al., 2016; Nissen, 1997). With the introduction of harnesses and horseshoes, horses were increasingly used for agriculture. Horsepower was supported and replaced by machinery in the industrialized countries only in the last hundred years (Librado et al., 2016; Petersen et al., 2014).

Since the mid-20th century horses in developed countries are mainly used for sports (horse racing, equestrian sports, leisure sports, games such as polo), work (therapeutic horseback riding, draft horses in forestry work or tourism), and resources (meat, milk, serum) or for pure pleasure (hobby) (Kräußlich, 1997). Overall, horses are passionately loved as companion animals and the emphasized interest of breeders and owners in their horses' ancestries is remarkable (Nissen, 1997).

1.2. Anthropogenic selection

The onset of domestication reflects the commencement of artificial selection. Only a limited number of horses from the wild seemed to have formed the basis of the effective domestic breeding stock (Warmuth et al., 2012). Breeding animals were continuously selected according to whether they showed particular characteristics of interest (Petersen et al., 2013). Favoring locomotion and behavioral traits was already evident in early equestrianism (Librado et al., 2021). Anthropogenic selection gradually shaped horses over millennia in terms of socio-economic needs. Desired characteristics have included performances, such as speed, endurance, and strength, as well as gait, temperament, and appearances, like size, color, and conformation (Bowling, 1996a; Kräußlich, 1997). The human-mediated breeding process drastically influenced the evolutionary trajectories of the species and led to the creation of the domestic horse population as we know it today (Librado et al., 2016; Librado & Orlando, 2021; Orlando & Librado, 2019).

Horse breeding is characterized by a sex bias toward males and this is consistent from ancient breeding to modern breeding practices where selected stallions are used to cover multiple mares (Vilà et al., 2001). Recent significant turning points in horse husbandry techniques were the commencement of organized horse breeding and the popularity of linebreeding concepts, see chapter II. 1.4 ‘Historical review on the formation of European horse breeds’. These management shifts led to the foundation of closed studs and the fragmentation of some breeding populations into demarcated reproductive isolates. A development that launched the formation of horse breeds as we know them today (Nissen, 1997; Orlando, 2020).

1.3. Modern horse breeds

Horse breeds represent managed populations. Members of a breed share selected characteristics and are related to a variable extent. Explicit attributes such as gait, coat color, standing height, racing speed, or endurance may be distinctive for a specific horse breed. According to the production statistics of the Food and Agricultural Organization of the United Nations (FAOSTAT), the extant global horse population counts almost 58 million individuals (FAO, 2021b), which are divided into hundreds of specialized horse breeds. However, breed definitions, as well as numeric data on horse breeds, vary among literature sources (Hendricks, 2007; Kräußlich, 1997; Nissen, 1997). Breed counts depend on the breed categorizations applied. The FAO for example classifies mammalian livestock breeds in only three categories, ‘local breeds’ and ‘transboundary breeds’, with the latter subclassified into ‘regional transboundary breeds’ and ‘international transboundary breeds’. According to the FAOSTAT, the highest number of reported horse breeds is recorded in Europe with 269 local and 38 regional transboundary horse breeds listed in 2006 (Khadka et al., 2010).

The FAO system is a pragmatic classification; however, it is often too simplified given the variation of horse breeds. Hence more commonly, horses are classified into more general defined breed types which underlie local gradings and ideals rather than a global reference norm. Such categorizations are based on temperament, such as ‘coldblooded’, ‘warmblooded’, and ‘hotblooded’ horses; nature of work like ‘riding’ or ‘draft’ horses; as well as conformation, like ‘light’ and ‘heavy’ horses or ponies. Also, studbook keeping classifies horses, dependent on the applied breeding

strategy, into ‘purebred’ or ‘crossbred’ horses (Khadka et al., 2010). Some horse breeds are even further sub-categorized into breeding sections; for example, the ‘Finnish horse’ is divided into ‘riding horses’, ‘pony-sized horses’, ‘draught horses’, and ‘harness trotters’ (Kvist et al., 2019).

Overall, the denomination of breeds is a quite recent, European-conceptualized phenomenon in the history of domestic horses. Before breed classifications, horses were defined rather by their geographic origin, for example, ‘North African’, ‘Spanish, ‘Shire’ or ‘Arabian’ horse; but also by their specific type of usage, such as ‘Courser’ (warhorse) or just ‘Carriage horse’ (Bowling, 1996a; Druml, 2011). The emergence of horse breeds is coupled with the formation of breed associations.

1.4. Historical review on the formation of European horse breeds

This brief historical discourse aims to summarize important circumstances and facts over the past 1,500 years that have influenced the emergence of European horse breeds.

Several migration waves and military invasions influenced European horse populations within this timeframe. The migration period from the 4th to the 8th century common era facilitated the distribution of Germanic horses. Later on, Hungarian invasions, in the 9th and 10th centuries, and the Mongol invasion into Europe in the 13th century promoted the distribution of horses from the East (Kräußlich, 1997). In parallel, in the West, the Moors occupied the Iberian Peninsula from the 8th to the 15th centuries entailing genetic exchange between North African and Iberian stock (Luís et al., 2006). More recently, the Ottoman wars, lasting from the 15th century up through the early 20th century, promoted a long-lasting distribution of horses from the Middle East (Nissen, 1997).

In the Middle Ages, from the 5th to the 15th century, the selection of European horses was mainly guided by the demands of warfare. Large horses were required as the weight of a knight’s armor increased from 80 kg in the 12th century to 130 kg in the 15th century (Grilz-Seger & Druml, 2017b). From the late 15th century onwards, driven by the introduction of firearms and by the formation of cavalry forces, the criteria altered towards faster and more agile war horses. This initiated the introduction of Arabs, Barbs, and Turkman horses as refining forces in horse breeding (James, 1986; Kräußlich, 1997). Within the time frame from the 15th up to the 19th century, the increasing demand for highly relevant military horses induced the centralization of horse stock and breeding became more organized; numerous military studs were founded in central Europe, for example, ‘Mezöhegyes’ in 1785, ‘Bábolna’ in 1789, ‘Radautz’ in 1792, and ‘Piber’ in 1798 (Druml, 2011).

Horses were not only used for war but were also considered as an animal of prestige by excellence and thus status symbols. From the 16th century onwards royal and clerical stud farms arose to produce noble, representative horses, such as ‘Kladrub’ in 1579, ‘Lipizza’ in 1585, and ‘Hillerødsholm’ later ‘Frederiksborg’ in 1560. Imported stallions, mostly of Spanish and Neapolitan origin were used in those studs. Historically, an exchange of stud and regional breeding stocks reflects an exceptional open breeding practice within the following century; archbishops permitted private citizens to breed with imported stallions (Grilz-Seger & Druml, 2011, 2017b). Horse types for representational

purposes were especially hyped during the Baroque era, from the 17th until the 18th century (Druml, 2011; Fages et al., 2020). This era promoted classical riding as performed art and as advantageous for cavalrymen (Giacomini, 2014).

Crossbreeding, with no demand of known ancestries of breeding animals, was commonly used in Europe until the early modern period, 15th to 18th century. In the 18th century, applied breeding schemes in European studs changed from cross- to inbreeding concepts. The idea of purebred horses was initially co-imported with Arabian horses as these horses were already bred in purity under the principle of line-breeding in their homelands of origin (Nissen, 1998). The English racehorse community adopted the newly imported inbreeding practices and could thereupon impressively improve the racing performances of their horses. In the late 18th century, the very first 'general stud book' in Europe was introduced for the 'English Thoroughbred horse' (Weatherbys, 1791). Once the notion of pure stock was accredited to successful animal breeding, the practice became popular across many horse populations. The concept of breeds was further promoted by the European aristocracy by applying their 'blue-blooded' theory to domestic animal populations. As a consequence of this development, the formation of breed associations commenced on a large scale within the recent 200 years and depict a Western World phenomenon (Bowling, 1996a; Druml, 2011; Grilz-Seger & Druml, 2011; Kräußlich, 1997; Nissen, 1998; Scharnhölz, 2017).

Besides warfare and representation purposes, horses were also bred for work. From the onset of the industrial revolution, in the late 18th century, the breeding of heavy horses was intensified. Paradoxically, the demand for horsepower exploded in the first phase of the industrialization process and collapsed only at its end (Grilz-Seger & Druml, 2017a). In the 20th century, horsepower experienced serious deprecation. With the progressive industrialization and after the conclusion of World War II in 1945, horses lost their monopoly position as a power source (Orlando & Librado, 2019). This resulted in severe population reductions and quite recent genetic bottlenecks, especially in most draft breeds and heavy horse types (Petersen et al., 2013).

1.4.1. Significant introgression waves

In Europe, the breeding success in the past 500 years was mainly driven by stallions that were introduced from foreign studs to improve the existent breeding stock. Notably, three waves of introgression can be described (Druml, 2011; Grilz-Seger & Druml, 2011; Nissen, 1997) and those influential waves shaped the basis of modern breeds.

i) From the 16th to the 18th century horse breeding in central Europe was impacted by the **'Neapolitan and Spanish'** wave. Organized stallion-mediated refinement promoted the import of Spanish and, now extinct, Neapolitan stallions (Druml, 2011; Grilz-Seger & Druml, 2011). In parallel to their influence in Europe, the transmission of horses to the American continent took place. This was undertaken by colonialists and conquerors, who brought horses from the Iberian Peninsula to the New World. The first translocation was in the year 1493 during the second voyage of Christopher Columbus (Cortés et al., 2017).

ii) The **'Oriental'** wave followed from the 18th to the 19th century and is marked by the introduction of Oriental stallions. In this period, imports were largely restricted to horses from Turkoman (from the steppes of central

Asia) and Arabian (from the Arabian Peninsula) origin (Grilz-Seger & Druml, 2011; Nissen, 1997). The impact of the Oriental wave was intensified by the co-imported new breeding schemes, which lead to the focus on pure-blooded horses, closed populations, line-breeding, and inbreeding (Nissen, 1998). The foundation of the Thoroughbred horse breed falls into the timeframe of this wave (Weatherbys, 2021). The popularity of Arabian horses in particular and their influence on European horse populations can be addressed as the **'Arabian'** wave. Throughout the 19th and into the first decades of the 20th century several Arabian horse lineages were established in many countries.

iii) The third, the **'English'** wave, started in the 19th century and is consistent until today. Stallion-mediated improvement peaked in this wave in a way that Thoroughbred stallions were, and still are, used as a refining force on a global scale. Various sport horse and warmblood breeds derive from Thoroughbred ancestry (Nissen, 1997).

1.4.2. Substantial founder populations of modern horse breeds

Four horse breeds respectively horse types, that still exist today, can be associated with the above-mentioned introgression waves. They are the legacy of historic populations that have contributed extraordinarily to the genetic pool of modern breeds (Grilz-Seger & Druml, 2011; Nissen, 1997) and are now briefly introduced.

i) The Thoroughbred horse breed

The Thoroughbred horse breed, initially founded in England, has been introduced all over the world, where it is bred for racing or used as a refiner for local populations (Hendricks, 2007; Nissen, 1997; Petersen et al., 2014). It is administered by one of the oldest closed studbooks, which harbors presumably the most comprehensive pedigree records among domestic animal breeds. An 'Introduction to a General Stud book' dedicated to the English Thoroughbred horse was published in 1791 (Weatherbys, 1791) and compiled earlier sources of pedigree information for a small number of horses (Bower et al., 2012; Cunningham et al., 2001).

The breed itself bases on refinement breeding as local English mares were bred with Oriental stallions (Bower et al., 2012). All extant breed members trace their male ancestry to one of three imported stallions, namely Darley Arabian, born in the year 1700, Byerley Turk, born in the year 1680, and Godolphin Arabian, born in the year 1724 (Bower et al., 2012). Whereby Darley Arabian has had a disproportionate influence on living male lineages, 95% of extant paternal bloodlines are attributed to him (Cunningham et al., 2001).

ii) Arabian horses

The cradle of Arabian horses is the Middle East, from where they spread around the world. Arabian horses are bred in many countries outside their homelands of origin, among the leading ones are Poland, Egypt,

France, Russia, Great Britain, and the US. In total 63 Arabian pedigree registries exist today (WAHO, n.d.). The pedigree records of extant populations in Europe, Russia, Egypt, and the US comprise 200 years of breeding history tracing back to imported Arabian horses. The popularity of Arabian horses grew in the Western World from the 18th century onwards. In their homelands, Arabian horses were bred in purity over centuries and their import heavily influenced European horse breeding throughout the 19th and into the first decade of the 20th century. The foundation stocks of many modern breeds comprise Arabian horses (Głazewska, 2010; Nissen, 1997).

iii) North African horses

Initially bred by nomadic people, Barb horses are the most renowned North African breed today (Hendricks, 2007; Krischke, 2011). Pedigree documentation for Barbs, to the extent it exists at all, began only recently at the end of the 20th century. In the late 19th century, North African Barb horses were crossbred with Arabian horses to combine the hardiness, endurance, and stamina of the Barb with the elegance and speed of the Arabians. This refinement breeding resulted in the Arab-Barb breed (Berber et al., 2014). Apart from the Maghreb countries, Barbs and Arab-Barbs are actively bred in Central Europe (mainly France and Germany), where breeding studs are built upon horses recently imported from the countries of origin. Historically, the cultural and economic ties between the Maghreb and the Iberian Peninsula induced a kinship between North African and Iberian horses (Druml, 2011; Nissen, 1997). In Europe, North African horses were already mentioned in medieval written records and are suggested to have influenced many modern breeds, especially Baroque horse breeds (Krischke, 2011).

iv) Iberian horses

A variety of modern breeds are native to the Iberian Peninsula, among those are Coldblood breeds, Pony breeds, and Warmblood breeds (Nissen, 1997). To what extent Iberian horses evolved under the influence of the local landrace ‘Sorraia’, or via the influences of North African, Arabian, and European horses, specifically Celtic ponies, is still open for speculation (Druml, 2011; Royo et al., 2005). Notably, modern Baroque Type breeds retain the distinctive characteristics of an Iberian horse type that rose to prominence in Europe during the Baroque era (Nissen, 1997). Historically, this famous Iberian horse type developed in the Middle Ages from horses used in bullfights and the Reconquista. From the 16th to the 18th century this Iberian horse type was used as a refining force across Europe. The dispersal of horses from the Iberian Peninsula was driven by a baroque ideal and the supremacy of the Spanish imperium (Druml, 2011; Nissen, 1997). Due to a brief annexing period of Portugal to Spain at the height of the Baroque period, from 1578 to 1640, renowned Iberian horses were generally known as ‘Spanish horses’ (Giacomini, 2014). Founder stocks of modern Baroque Type breeds, such as the Lusitano breed, the Pura Raza Española breed, the

Lipizzaner breed, the Kladruber breed, the Frederiksborger breed, and the Friesian breed descent in large parts from Spanish horses (Druml, 2011; Giacomini, 2014).

1.5. Modern breed management

Modern horse breeds are managed by their breed associations. Such organizations formulate the breeding program and keep records of breeding animals. Breed associations are administrated either privately or governmentally and multiple associations may exist for the same horse breed. Management aspects vary among them, in terms of the quality and effectiveness of the organization, period of foundation, and length of service as well as breeding goals (Encyclopaedia Britannica, 1998; ESSA, 2018; Nissen, 1997).

The official records of horses approved for reproduction are called 'Breed registries', nominal studbooks, or herdbooks (Encyclopaedia Britannica, 2011). Studbooks comprise accurate biographical data, such as gender, date of birth, or reproductive success, performance results, identifying characteristics, like brands and markings, as well as the pedigrees of registered horses. Overall, they contain the demographic history of the entire managed population. Studbooks are used to correlate traits with ancestry, making pedigree records a valuable basis for predicting breeding success. Horses must meet either a conformation or a performance standard, or both, to be registered (Bowling, 1996a). There are two types of studbooks in regards to the descent of the breeding stock.

i) Studbooks may be kept closed. Under this condition, the use of foreign blood for breeding is prohibited. Thus, all horses registered trace back to a defined foundation stock of the specific breed ensuring their purebred status. The gene pool of the breed is strictly defined and the only source of novel genetic variation is undetected crossbreeding or mutation. Closed populations promote inbreeding, a term that refers to the mating of related animals (Bowling, 1996a). The most representative example of a closed studbook is the Thoroughbred horse breed (Weatherbys, 2021).
ii) Alternatively, studbooks may be kept open. Registries with open studbooks allow the immigration of breeding animals into a given population to variable extents (Brem et al., 1997). This means that horses approved for breeding may have an unknown ancestry or may derive from another breed. For example, the American Quarter Horse Association (AQHA) accepts Thoroughbred horses for breeding (AQHA, 2000). A common application of this breeding scheme is refinement crossing, mostly conducted via stallions since they can produce effectively more offspring than mares. Arabian and Thoroughbred horses have been widely used as enhancement sources (Bowling, 1996b).

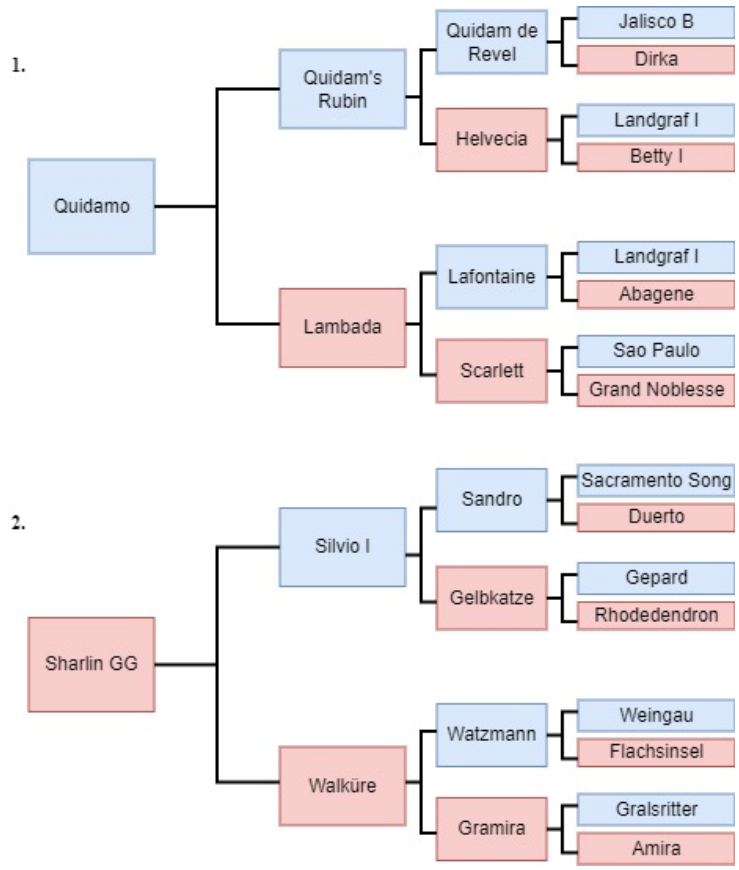
1.5.1. Pedigrees and the importance of sire lines and mare lines

The pedigree of a horse portrays its ancestry, like a family tree. The paternal line of descent in the pedigree is addressed as 'tail-male lineage', while the maternal line of descent as 'tail-female lineage'. Pedigrees of extant breeds trace ancestries back to horses of unknown parentage, which are mostly referred to as the founders (or 'foundation animals') of the designated horse breed.

The cumulative impact of a stallion's male offspring leads to the establishment of a 'sire line', those of a mare's female offspring to the establishment of a 'dam line'. Sire and dam lines are a strategically impactful breeding tool to fix inherited traits in a breed; but they are also helpful, to maintain an overview of breed diversity and to avoid high-level inbreeding (Bowling, 1996a; Brem et al., 1997). Since stallions can produce a higher number of offspring, their breeding impact is comparatively greater than that of a mare. Thus, a few well-supported sire lines in contrast to a variety of mare lines are rather the rule than the exception in modern breeds (Nissen, 1997). In modern horses a sex bias towards males in breeding practice is evident, sire lines that are inherited from a single renowned stallion are celebrated. Today, the impact of specific sire lines is further promoted through the feasibility of artificial insemination, which is not allowed in the Thoroughbred horse breed (Bowling, 1996b).

When talking about the ancestry of a horse, sires are rather addressed than mares, because mostly more famous. A breeder who promotes the stallion 'Quidamo' would adduce the ancestral sires 'Quidam's Rubin', 'Lafontaine', and 'Sao Paulo'. Talking about the mare 'Sharlin GG, promoting would sound somehow like: "The mare 'Sharlin GG', sired by 'Silvio I' out of the mare 'Walküre' from 'Watzmann' by 'Gralsritter, is an extraordinarily beautiful horse."; exemplified pedigrees are shown in Figure 1.

Figure 1: Pedigree examples of two horses of a European Warmblood breed. 1.) shows the pedigree of a stallion, number two 2.) shows the pedigree of a mare. In both pedigrees the sires are bluish, mares are pinkish colored. The top line of a pedigree, the alignment of blue boxes, illustrates the tail-male lineage; the low line, the alignment of pink boxes, illustrates the tail-female lineage. © Viktoria Remer



Pedigree records of modern breeds differ tremendously in time depth. In some breeds, it is feasible to trace ancestries up to 250 years into the past, for example in the Thoroughbred horse breed (Weatherbys, 2021) or the Lipizzan breed (Grilz-Seger & Druml, 2011). Overall, studbooks bear witness to how domestic horse breeds emerged. To explain the origin and development of breeds, information from stud books and anecdotal historical statements can be supplemented with information encoded on the horses' genome.

2. Molecular genetic research in horses

The horse reference genome assembly was first generated in 2009 from the Thoroughbred mare 'Twilight' and incorporates the 31 horse autosomes, the X chromosome, and the mitochondrial genome (Kalbfleisch et al., 2018; Wade et al., 2009).

The provision of a reference sequence together with the advent of high throughput DNA sequencing technology (Dijk et al., 2014), which considerably increased the sensitivity and reduced the costs related to the analysis of genomic variation, promoted a wealth of studies of the horse genome in the past decade. The development of genome-wide screening tools, including the development of DNA microarrays targeting thousands of single nucleotide polymorphisms (SNPs) scattered across the entire genome (Schaefer et al., 2017), enable the study of the horses' genomes from many aspects. Several variants causal for various congenital diseases, performance, and other phenotypic traits were already mapped to particular genomic regions (Bellone et al., 2010; Bower et al., 2012; Makvandi-Nejad et al., 2012; Raudsepp et al., 2019).

Genetic approaches are also applied to elucidate the demographic history of horses, ranging from domestication scenarios to modern breed formation. In such studies the genetic composition of horse breeds is in the focus, to illustrate their relatedness, uniqueness, and/or their ancestry (Cortés et al., 2017; Pérez-Gutiérrez et al., 2008; Petersen, et al., 2013; Wade et al., 2009). Technical wise, either numerous polymorphic markers, like single nucleotide variants (SNVs), highly polymorphic short tandem repeat markers (STRs or microsatellites), short insertions/deletions (indels), or copy number variants (CNVs), scattered along many chromosomes are screened (Petersen et al., 2013; Raudsepp et al., 2019; Wade et al., 2009), or trajectories of a single or few particular, sometimes trait-specific, loci are investigated (Ludwig et al., 2009; McCoy et al., 2014; Wutke et al., 2016). The latter approaches often incorporate data from ancient and historic samples which enable to redraw past to present allele frequencies (Fages et al., 2019; Librado et al., 2017, 2021). Results enlightened the changes the horse population underwent through time and space since the commencement of anthropogenic selection. The emergence and distribution of particular traits such as coat color (Ludwig et al., 2009; Pruvost et al., 2011), gait or racing performance (Bower et al., 2012; Librado et al., 2021; Wutke et al., 2016), as well as metabolic and energy metabolism (McCoy et al., 2014) are a few examples.

Two unique segments in the genome are by far the most studied 'single loci' for inferring the demographic history of populations: the mitochondrial DNA (mtDNA) and the male-specific region of the Y chromosome (Underhill &

Kivisild, 2007). While the rest of the genome in sexually reproducing organisms gets reshuffled whenever sperm and eggs are produced, these two compartments are passed from parents to their offspring unilinear and almost unmodified (Hurles & Jobling, 2001). This turns the mtDNA and the Y chromosome into powerful systems harboring the potency to elucidate the sex-specific demographic history of a population. In horses, mtDNA variation has been studied to a great extent, while the information content from Y chromosome studies was still lacking behind (Raudsepp et al., 2019).

2.1. The maternally inherited mtDNA

The mtDNA is found in the cell's mitochondria and is a circular-shaped molecule. It is the only portion of genomic DNA outside of the cell's nucleus. In horses, as in other mammals, the mtDNA is a small, around 16 kilobases (kb), molecule. Every cell contains many mtDNA copies whereas copy numbers differ extremely between cell types. During fertilization, mtDNA contained in the sperm cell, notably in low copy numbers, get degraded. Thus, the mtDNA is transmitted, except in rare cases, from the mother to her offspring. Consequently, the mtDNA is inherited as a complete linkage group (a haplotype) in the matrilineage. Accordingly, mtDNA sequence variation is suitable to trace maternal lineages in populations. Due to the small genome size and the high copy numbers, sequencing the mtDNA is quite straightforward. The complete mtDNA sequence is available in most domestic animals (Jiang et al., 2010).

In horses, studies were initially based on the hypervariable mtDNA control region (D-loop) (Cieslak et al., 2017; Devi & Ghosh, 2013; Jansen et al., 2002), but more recently complete mitochondrial genomes are investigated as well. The outcomes show, that the phylogenetic origin of mtDNA haplotypes in present horses predates the domestication process by far. Horse mtDNA haplotypes coalesce some 93 to 160 kya (Achilli et al., 2012; Lippold et al., 2011). The major mtDNA haplogroups are distributed evenly across horse populations worldwide and show no clear correspondence to geographic areas or breeds. This pattern was explained by irreducible complex interactions including broad restocking from the wild, long-distance dispersal rates of mares, and recent human management (Warmuth et al., 2012; Warmuth et al., 2013). The evolutionary origins of mtDNA lineages as well as their exact trajectories during the ancient and historic development of the world's horse population are extensively discussed (Achilli et al., 2012; Librado et al., 2016, 2017; Lippold et al., 2011; Schubert et al., 2014; Vilà et al., 2001). In addition, tracing recently established mare lines with mtDNA analysis was conducted in modern breeds to verify pedigree correctness and to investigate founder mares' affinities (for example Bowling et al., 2000; Khanshour & Cothran, 2013b; Kavar et al., 2002).

Besides the application in horses, mtDNA studies have provided insights into the domestication process and historical development of several domestic species such as cattle, yak, water buffalo, sheep, goats, pigs, chicken, camelids (Almathen et al., 2016; Groeneveld et al., 2010), cats (Ottoni et al., 2017) and dogs (Savolainen et al., 2002).

2.2. The paternally inherited Y chromosome

While most mammalian chromosomes are inherited from both parents to their offspring, the male sex chromosome, the Y chromosome, is passed only from father to sons (Liu, 2010).

2.2.1. Y chromosome evolution and sequence structure

Most mammals have an 'XY' sex-determination system and the two sex chromosomes differ significantly in size, structure, and gene content. The X (female) and Y (male) chromosomes evolved from a pair of homologous autosomes about 180 million years ago (mya) (Cortez et al., 2014; Graves, 2006).

The 'proto-Y chromosome' acquired a sex-determining locus and subsequently male-specific genes accumulated on this chromosome. Repression of recombination evolved to shelter the male-specific genes on the Y chromosome, ensuring that those genes are retained in the correct (male) organism (Graves, 2006; Lahn & Page, 1999). Repression of recombination led to the absence of genetic exchange between the X and Y chromosomes over most of their length (Cortez et al., 2014; Graves, 2006; Li et al., 2013; Waters et al., 2007).

Only in small regions at the tips of the chromosome arms still occurs meiotic crossing over between the X and Y chromosomes. These parts, called 'pseudoautosomal regions' (PAR), are homologous between the X and the Y chromosome. The substantially larger part of a mammalian Y chromosome escapes the reshuffling processes and meiotic recombination with the X chromosome (Graves, 2006; Lahn & Page, 1999; Skaletsky et al., 2003). This region is either called the 'male-specific region of the Y chromosome' (MSY) or the 'non-recombining region of the Y chromosome' (NRY) (Tomaszkiewicz et al., 2017; Tyler-Smith, 1995). In humans and horses, the MSY comprises 95% or 54 megabases (Mb) (EMBL-EBI, 2021; Skaletsky et al., 2003) and 96% or 45 – 50 Mb of the total chromosome respectively (Janečka et al., 2018; Raudsepp et al., 2004).

The MSY is best described in humans and based on human observations the sequence content of the MSY was categorized into certain sequence classes (Skaletsky et al., 2003). Those classes apply more or less also for other mammalian Y chromosomes. Hence; sequence classes described in humans (reviewed in Skaletsky et al., 2003) are briefly introduced taking the equine Y chromosome (reviewed in Janečka et al., 2018) into account. The MSY contains a large block of heterochromatin which represents highly repetitive, tightly packed, and transcriptionally silenced heterochromatic DNA (Skaletsky et al., 2003). Two-thirds of the equine Y chromosome is heterochromatic while the small distant euchromatic segment is approximately 12 Mb (Janečka et al., 2018). Generally, the euchromatic regions contain the units of transcription (Skaletsky et al., 2003). The current state of knowledge is that the horses' Y chromosome contains at least 52 genes/transcripts. About half of the euchromatic region (54%) of the horse Y is composed of various interspersed repeats while only the remaining half are non-repetitive (Janečka et al., 2018). The human euchromatic MSY is divided into three classes, the 'X-degenerate', the 'ampliconic', and the 'X-transposed' regions. X-degenerate sequences descend from the proto-sex chromosome and still share nucleotide sequence identity with their X-linked homologs to various extents. The second class comprises the 'ampliconic' or 'multicopy' segments. Those parts are characterized by highly identical sequence blocks (up to 99,9% homology). Genes in this region have multiple copies. The marked similarities among the copies are preserved by frequent intrachromosomal gene conversion within this area (Skaletsky et al., 2003). In horses, MSY gene density comes close to that of autosomes and some multi-copy Y genes show a relatively broad expression pattern. This is noteworthy since Y genes in most

mammals are predominantly if not exclusively expressed in testes (Janečka et al., 2018). The human-specific X-transposed region resulted from a single X-to-Y transposition, 3 to 4 mya, and this region shows 99% identity to those of the X chromosome. In horses, a recent PAR to Y transposition around 3.3 – 4.1 mya was identified comprising two genes (Janečka et al., 2018).

2.3. The Y chromosome as a genetic marker

The MSY is inherited as a complete linkage group (a haplotype) in a patrilineal manner from the father to his male offspring. This makes the MSY an ideal tool to trace male-driven population dynamics. To conduct MSY lineage tracing, informative polymorphic variants need to be ascertained. Two types of variation on the MSY are generally applied. Either slowly-evolving single-base DNA mutations leading to single nucleotide variants (SNVs) and single-base insertions/deletions (indels), or faster-evolving variations in tandemly repeated sequences leading to microsatellites (or short tandem repeats (STRs)) (Calafell & Larmuseau, 2016; Claerhout et al., 2019; Jobling & Tyler-Smith, 2017; Poznik et al., 2016; Tao et al., 2019).

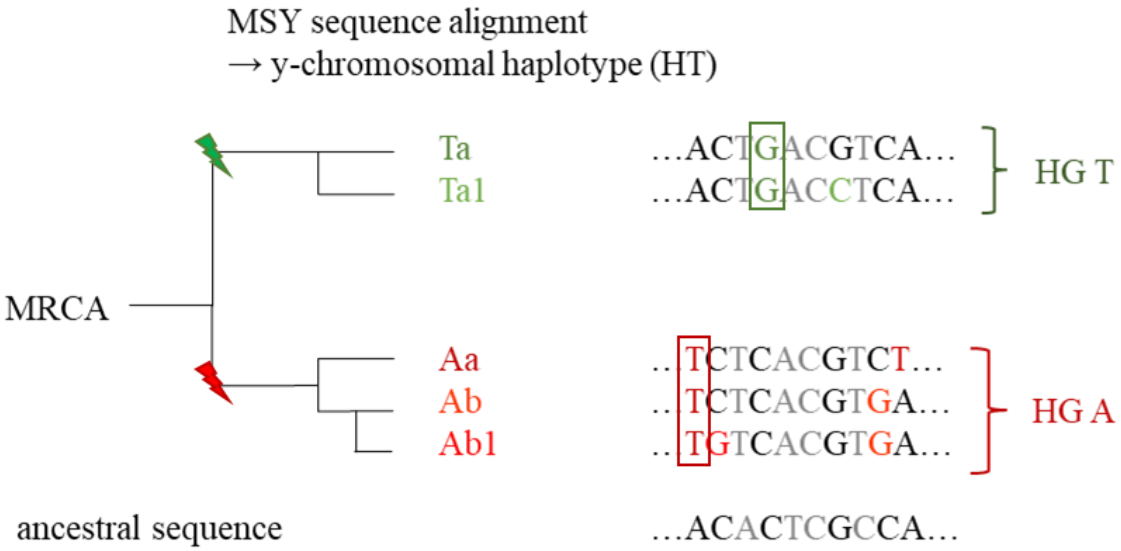
The uniparental inheritance and the lack of recombination imply that the MSY undergoes a simple history of sequence changes. MSY haplotypes show identity by descent and retain mutations that have occurred during evolution in strict hierarchical order (Jobling, 2012); as illustrated in Figure 2, new haplotypes evolve stepwise due to the accumulation of sequence variations. Whereas, SNVs are base substitutes that occur rarely enough to be regarded as unique events and can thus be applied as stable evolutionary markers (Jobling & Tyler-Smith, 2017).

The phylogenetic relationship among MSY haplotypes can be inferred under the principle of parsimony and illustrated as trees. The observed haplotypes coalesce in a single ancestral sequence, their most recent common ancestor (MRCA). Haplotypes forming a monophyletic group can be termed together as a haplogroup, and haplogroups are conditioned by haplogroup-determining variants (Batini & Jobling, 2017), see Figure 2. The nomenclature of haplotypes and haplogroups should follow a unified hierarchical system and reflects their phylogenetic clustering (Hammer, 2002).

Figure 2: The emergence of MSY haplotypes and nomenclature.

Right panel: The alignment of sequence variants on the MSY determines y-chromosomal haplotypes (HTs). Haplotypes forming a monophyletic group are termed together as a haplogroup (HG). Unaltered bases are written in grey, polymorphic haplotype-determining variants are colored, haplogroup-determining variants are framed.

Left panel: The MRCA of the haplotypes refers to the ancestral sequence from which all MSY haplotypes derived. A simplified haplotype tree is assembled under the parsimony principle with arrows indicating the emergence of haplogroup-determining allele 'G' in green and allele 'T' in red. The haplotypes 'Aa', 'Ab', and 'Ab1' cluster in the haplogroup 'A' which is defined by the allele 'T'; while the haplotypes 'Ta' and 'Ta1' cluster in haplogroup 'T' which is defined by the allele 'G'. Phylogenetic clustering is also reflected in haplotype/haplogroup nomenclature; © Viktoria Remer



Based on the assumption of the molecular clock, DNA sequences evolve at a rate that is relatively constant over time (Zuckerkandl, 1987). Thus, sequence differences between two DNA segments are expected to be proportional to the time since they last shared a common ancestor. The molecular clock is an often-applied method for estimating evolutionary timescales of the Y chromosome (Wei et al., 2013). Hence, branching points of phylogenetic trees can be dated based on the number of accumulated variants and a given mutation rate (Batini & Jobling, 2017; Tyler-Smith, 1995).

To ascertain MSY variants, next-generation sequenced (NGS) data are mapped to a Y reference sequence. NGS data are either generated from whole-genome sequencing data, comprising the full genome, or from targeted resequencing data. In this approach, specific regions of the genome, such as the Y chromosome are enriched before sequencing (Jobling & Tyler-Smith, 2017). NGS platforms made resequencing cost-effective (Jobling & Tyler-Smith, 2017). However, the Y is, due to its repetitive landscape, notoriously difficult to sequence (Hughes & Rozen, 2012). Only particular parts of the MSY represent uniquely mappable regions; approximately 10 Mb in humans (Jobling & Tyler-Smith, 2017; Poznik et al., 2013).

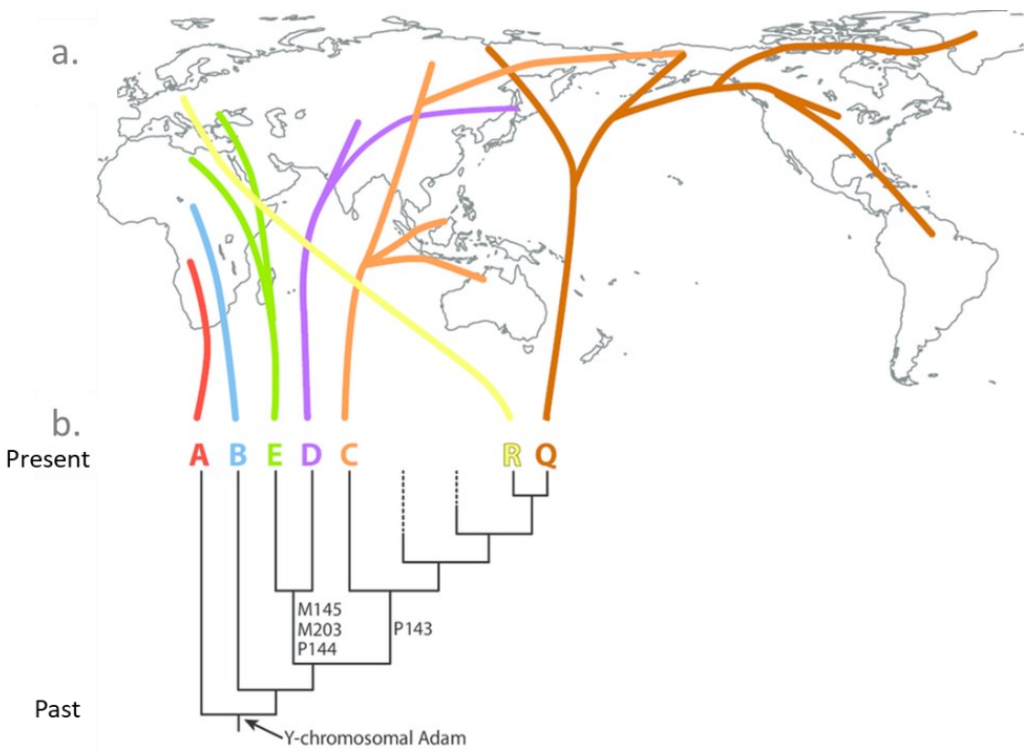
An alternative way to assess the distribution of haplotypes respectively haplogroups in a large number of cohorts is to genotype NGS-ascertained variants. MSY genotyping represents a time and cost-effective alternative to sequencing

methods. The allelic state of numerous, previously NGS-ascertained variants can be evaluated even in a high sample size. However, genotyping is limited to already identified MSY variants (Ballantyne et al., 2014; Gršković et al., 2010; Kayser, 2017).

Y-chromosomal lineage tracing is best established in humans to study population dynamics, male demography, and forensic genetics applications. Two decades of technical progress in variant calling and haplotype analyses in men shed light on many issues (Hallast et al., 2014; Jobling & Tyler-Smith, 2017; King & Jobling, 2009; Poznik et al., 2014; Teitz et al., 2018). The Y chromosome bears, for example, information on the ancestry of populations revealing the origin and distribution of patrilineages. To infer the origin of extant haplotypes respectively haplogroups ancient samples may be incorporated in Y studies. Haplogroups can be correlated with specificities, such as geographic areas (Batini et al., 2015; Hughes & Rozen, 2012).

In Europe, for example, four y-chromosomal haplogroups are predominant and the analysis of European Y lineages revealed the genetic trace of European colonisations (Navarro-López et al., 2021). In human studies, molecular phylogenies were also superimposed upon maps to infer and illustrate migration history dynamics (Hughes & Rozen, 2012; Jobling, 2012).

Figure 3: Abridged genealogical tree of human MSY haplogroups superimposed on a geographical map, adopted from *Hughes & Rozen, 2012*. a) the geographical distribution and specificity of MSY haplogroups. b) MSY haplotype tree. Present haplotypes cluster in defined haplogroups (A, B, E, D, C, R, Q) and derived from one ancestral haplotype in the past (the MRCA or, in humans, 'y-chromosomal Adam') *(Hughes & Rozen, 2012)*.



In humans, the MSY is also extensively used in genetic genealogy research (Calafell & Larmuseau, 2016) and here for example a relationship between MSY and patrilineal surnames has been supported. Heritable patrilineal surnames were linked with MSY haplotypes and their (co)ancestry was illustrated (King & Jobling, 2009b, 2009a; Martinez-Cadenas et al., 2016). A very prominent example of family history research is the case of US President Thomas Jefferson (1743 – 1826). The president was the alleged father of at least one son of Sally Heming`s, a slave at the president's Virginia estate. His declared son shared his MSY haplotype with a male-line descendant of Jefferson's paternal uncle. The MSY thus supported the paternity case (Foster, 1998).

MSY analysis is interesting for academic scientists and the huge and lively community of amateur genealogists. For humans, direct-to-consumer NGS services are already available and people can get their Y chromosome sequenced (Calafell & Larmuseau, 2016; Jobling & Tyler-Smith, 2017).

In other mammalian species, MSY haplotype analysis was used to illuminate the male side of the domestication process as well as the process of breed formation. Such studies were performed on swamp buffaloes (Wang et al., 2018), sheep (Meadows et al., 2006; Zhang, 2014), goats (Kul et al., 2015; Vidal et al., 2017), pigs (Guirao-Rico et al., 2018; Cliffe, 2010; Ramírez et al., 2009), dogs (Brown et al., 2011; Larson et al., 2012; Oetjens et al., 2018) and cattle (Edwards et al., 2011; Götherström et al., 2005). For cattle also animal exchange between past cultures has been addressed (Ginja et al., 2019; Pérez-Pardal et al., 2018).

2.3.1. Equine MSY diversities

MSY sequencing and the assembly of a reference sequence in horses were hampered for a long time by its repeat-rich nature. Additionally, modern breeding practices conditioned the monomorphic appearance of Y-chromosomal lineages in modern breeds and thus scientists could not detect significant Y-chromosomal sequence diversity (Lindgren et al., 2004; Wallner et al., 2004; Wallner et al., 2013). The achievements necessary to perform comprehensive Y-chromosomal studies in horses were accomplished only within the last decade, namely the assembly of a reference sequence (Janečka et al., 2018; Wallner et al., 2013) and data generation based on NGS technologies (Felkel et al., 2018; Wallner et al., 2017).

The first MSY markers informative in domestic horses were published in the year 2013. NGS sequencing technology was applied to screen 186 kbs of the MSY in 17 domestic horses and one Przewalski horse. Only six haplotypes could be distinguished among several domestic horse breeds, and those were distinct from the Przewalski horse haplotypes. While two haplotypes were widely distributed at high frequencies among modern European horse breeds, private haplotypes were detected in Northern European breeds. The Northern European haplotypes were fixed within but not shared among breeds. The influence of Thoroughbred stallions on modern sport horse breeds was demonstrated for the first time (Wallner et al., 2013).

In 2017, the first Y chromosome genealogy of modern horses was resolved by screening 1.46 Mb of the MSY. For this purpose, a draft horse MSY reference sequence based on short-read data was generated and 52 whole-genome sequenced individuals representing 21 breeds were screened for variants. The low sequence diversity on the horse MSY was confirmed with only 53 variants based on 50 SNPs and three indels detectable. Together with the formerly published MSY variants, 24 individual haplotypes were distinguishable (Wallner et al., 2017). Additionally, a *de novo* mutation rate for horse MSY was estimated from deep pedigree data available in horses. The outcome showed that modern horses' MSY haplotypes coalesce much later than the domestication of the species. MSY haplotyping was carried out to assess haplotype frequencies in 363 purebred horses from 57 breeds. Almost all modern horse breeds clustered together in one predominant haplogroup of oriental origin, the so-called 'crown haplogroup' (CH). Again, the high impact of English Thoroughbred sires on modern horses was displayed and the few private northern European haplotypes were confirmed. In conclusion, it was shown that breeding history has led to a situation in which only a few, extremely influential stallions exerted disproportionate influence on modern breeds (Wallner et al., 2017).

By 2018, a study revealed a much broader MSY spectrum in Asian horses than in European breeds. Asian haplotypes introduce a deep split to the phylogeny. The authors conclude that Asian horses retained signatures of the far more diverse ancient horse population, a signal that is lost in cosmopolitan breeds (Felkel et al., 2018). The signature of native MSY lineages was also evident in Chinese Mongolian Horses (Han et al., 2019) and Estonian native horses (Castaneda et al., 2019).

III. RESEARCH OUTLINE

This doctoral thesis aims to develop meaningful MSY analyses in modern horse breeds and to show their information content using practical questions. To reach this aim, strategies and workflows were established that include improving the ascertainment of variants from NGS data and comprehensive screening of MSY haplotype distribution. An overarching objective was to determine diagnostic MSY markers for sire lines active in modern breeds. Two renowned breeds, namely Thoroughbred horses and Arabian horses, were the main focus. Both breeds have been extensively used as refining forces in breeding within the recent 200 years and the utilization of the MSY as a genetic marker makes their influence on modern breeds genetically traceable.

Two peer-reviewed manuscripts are included in the thesis and are presented in the results section. In addition to the published scientific papers, results for a publication in preparation are presented. The lists of literature of the peer-reviewed manuscripts are held in the quoting style of the respective journals. Enumeration of figures and tables refers to the published version of the manuscripts. References quoted in the published manuscripts are not listed at the end of the doctoral thesis. References quoted in the third manuscript, that presents results for a publication in preparation, are given in the literature section (chapter VIII. List of literature).

- **Publication I**

The horse Y chromosome as an informative marker for tracing sire lines

In Felkel et al. (2019) a draft reference of the horse MSY was assembled, and strategies for efficient variant ascertainment from NGS data were established. This enables fine-scaled Y-chromosomal lineage tracing on the sire line level. By incorporating pedigree information, the haplotypes of the three English Thoroughbred founder stallions were distinguished. Consequently, this publication carved out MSY markers that are diagnostic for Thoroughbred ancestry. Additionally, a genealogical question in the Thoroughbred horse breed was successfully addressed. Based on MSY haplotype frequency data assessed by genotyping haplotype-determining variants in a comprehensive dataset, the predominance of the CH in modern breeds was confirmed. The contribution to this article was in determining haplotype distribution in a range of breeds. These data were analyzed for dating internal branching points and served as a basis for selecting samples for NGS target-enriched resequencing to refine the haplotype phylogeny.

- **Publication II**

Y-chromosomal insights into the breeding history and sire line genealogies of Arabian horses

In the second manuscript, Remer et al. 2022, the CH was resolved in more detail by ascertaining variants from NGS target-enriched resequencing data. The Y signatures of another important breed, the Arabian horse, were identified and diagnostic MSY markers for Arabian ancestry were determined. Additionally, MSY analyses were successfully applied on a genealogical scale to answer paternity questions in Arabian horses.

- **Publication III in preparation**

Equine Y-chromosomal variants develop into a diagnostic tool for paternal ancestry studies and enlighten the legacy of Spanish horses

Pedigree records of modern horse breeds illustrate that besides Arabian and Thoroughbred stallions, numerous other historic populations, especially Spanish studs, have had a remarkable influence on breed formation. Hence, the next logical step was to infer MSY haplotypes in Iberian breeds and breeds that developed to a certain extent under the influence of Spanish horses, like New World breeds, Baroque Type breeds, and European Coldblood breeds. MSY spectra in those breeds should lead to a more complete picture of the dissemination of CH haplotypes that derived from Spanish ancestry.

IV. RESULTS

1. Publication I

SCIENTIFIC REPORTS

OPEN

The horse Y chromosome as an informative marker for tracing sire lines

Received: 4 September 2018
Accepted: 4 April 2019
Published online: 15 April 2019

Sabine Felkel[1,2], Claus Vogl[1], Doris Rigler[1], Viktoria Dobretsberger[1], Bhanu P. Chowdhary[3], Ottmar Distl[4], Ruedi Fries[5], Vidhya Jagannathan[6], Jan E. Janečka[7], Tosso Leeb[6], Gabriella Lindgren[8,9], Molly McCue[10], Julia Metzger[4], Markus Neuditschko[11], Thomas Rattei[12], Terje Raudsepp[13], Stefan Rieder[11], Carl-Johan Rubin[14], Robert Schaefer[11], Christian Schlötterer[15], Georg Thaller[16], Jens Tetens[16,17], Brandon Velie[8,18], Gottfried Brem[1] & Barbara Wallner[1]

Analysis of the Y chromosome is the best-established way to reconstruct paternal family history in humans. Here, we applied fine-scaled Y-chromosomal haplotyping in horses with biallelic markers and demonstrate the potential of our approach to address the ancestry of sire lines. We *de novo* assembled a draft reference of the male-specific region of the Y chromosome from Illumina short reads and then screened 5.8 million basepairs for variants in 130 specimens from intensively selected and rural breeds and nine Przewalski's horses. Among domestic horses we confirmed the predominance of a young'crown haplogroup' in Central European and North American breeds. Within the crown, we distinguished 58 haplotypes based on 211 variants, forming three major haplogroups. In addition to two previously characterised haplogroups, one observed in Arabian/Coldblooded and the other in Turkoman/Thoroughbred horses, we uncovered a third haplogroup containing Iberian lines and a North African Barb Horse. In a genealogical showcase, we distinguished the patrilines of the three English Thoroughbred founder stallions and resolved a historic controversy over the parentage of the horse 'Galopin', born in 1872. We observed two nearly instantaneous radiations in the history of Central and Northern European Y-chromosomal lineages that both occurred after domestication 5,500 years ago.

The horse (*Equus caballus*) has accompanied humans ever since its domestication more than five millennia ago. While initially serving as a food source, the horse soon revolutionised agriculture, transportation and warfare. Vast empires were ruled from the back of the horse[1]. Today we count 58 million horses worldwide[2] and

[1]Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna, Vienna, 1210, Austria. [2]Vienna Graduate School of Population Genetics, Vienna, Austria. [3]United Arab Emirates University, 15551, Al Ain, UAE. [4]Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Hannover, 30559, Germany. [5]Lehrstuhl fuer Tierzucht, Technische Universitaet Muenchen, Freising, 85354, Germany. [6]Institute of Genetics, Vetsuisse Faculty, University of Bern, Bern, 3001, Switzerland. [7]Department of Biological Sciences, Duquesne University, Pittsburgh, 15282, USA. [8]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, 75007, Sweden. [9]Department of Biosystems, KU Leuven, Leuven, 3001, Belgium. [10]Veterinary Population Medicine Department, University of Minnesota, St. Paul, MN, 55108, USA. [11]Agroscope, Swiss National Stud Farm, Avenches, 1580, Switzerland. [12]Department of Microbiology and Ecosystem Science, Division of Computational Systems Biology, University of Vienna, Althanstrasse 14, 1090, Vienna, Austria. [13]Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX, 77843-4458, USA. [14]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala, 75123, Sweden. [15]Institut fuer Populationsgenetik, University of Veterinary Medicine Vienna, Vienna, 1210, Austria. [16]Institute of Animal Breeding and Husbandry, University of Kiel, Kiel, 24098, Germany. [17]Functional Breeding Group, Department of Animal Sciences, Georg-August-University Göttingen, Göttingen, 37077, Germany. [18]School of Life and Environmental Sciences, University of Sydney, Sydney, 2006, Australia. Correspondence and requests for materials should be addressed to B.W. (email: barbara.wallner@vetmeduni.ac.at)

distinguish about 700 breeds[3]. Extant horses are genetically clearly distinct from early domesticates[4]. Substantial turnovers in the horse genome coincide with the expansion and spread of human cultures starting from the Early Bronze Age over Scythian steppe riders and Roman times into the Middle Ages[5–7]. As such shifts were mainly achieved by breeding from limited numbers of stallions, they have led to a much-reduced male compared to female effective population size, explaining the difference in Y chromosome and mtDNA diversity[8].

Strategic breeding in the past 300 years caused the most dramatic change in the genetic make-up of horses[9]. The formation of European and American breeds is marked by an enormous impact of imported stallions. According to written records, this process, intended as a refinement of local stocks, started in Europe with the popularity of Iberian sires from the 15th to the 18th century. It was followed by the so-called "Oriental wave" from the late 18th to the late 19th century. During this period "Original Arabian" stallions have been imported from Syria to Egypt to achieve speed and elegance. From the early 19th century onwards, systematic upgrading of horse populations was mainly done through English Thoroughbred stallions. The establishment of most modern horse breeds attributes to this period[10]. Given the impact stallions had and still have to achieve breeding goals, a pedigree-independent, precise genetic analysis of sire lines is important.

The male-specific region of the mammalian Y chromosome (MSY) perfectly mirrors the patrilineage and MSY haplotype (HT) data have been widely used to infer the paternal ancestry of populations[11–16]. In humans, for example, the most used genetic markers to combine genetic data with family history are located on the MSY[17,18]. Tracing back the history of sire lines based on such genetic information, however, has been hampered by the low variability on the domestic horse MSY[19–21]. We recently demonstrated that MSY haplotyping can supplement pedigree information[22,23]. Based on a male genealogy inferred from an MSY reference spanning 1.46 million basepairs (Mbp) we showed that apart from an early branching Asian clade and a few other Asian and Northern European lineages, all Western domestic horses cluster in a recently established 'crown haplogroup'[23,24]. We interpreted the predominance of the crown haplogroup in Western European and North American breeds as a consequence of the extreme preference of stallions of Oriental origin during the past few hundred years. Given the dominance of crown lineages in intensively bred horse populations, full understanding of the recent use of stallions requires a clear resolution of the crown haplotypes.

In this article, we generated a 6.46 Mbp sized reference of the horse MSY that goes far beyond the 1.46 Mbp non-repetitive MSY (nonrepMSY) used in previous studies. We further developed a probabilistic method to accurately define the regions on the repeat rich mammalian MSY[12] that are suitable for unambiguous variant calling. By increasing the region screened for polymorphic markers and augmenting our previous dataset with more crown group breeds, we dissected the crown HT structure with a resolution only reached in human MSY studies so far. Our data permit new insights into the recent history of horse breeding.

Materials and Methods

A detailed description of methods including program parameters and program versions used is available in the Supplementary Information.

Ethics statement. The study was discussed and approved by the institutional ethics and welfare committee of the University of Veterinary Medicine Vienna in concordance with GSP guidelines and national legislation (ETK-10/05/2016). The research was performed in accordance with relevant guidelines reported in the above-mentioned document. All samples are coded and an informed consent was obtained from horse owners.

Samples and raw data processing. *Next generation sequencing (NGS) data.* Whole-genome Illumina data from 130 male domestic horses, nine Przewalski's horses, one donkey and five female horses used in the study were either obtained form previously published studies[23–35] and downloaded from publicly available sources or sequenced as part of this study. Details on samples are given in Supplementary Table S1 and Supplementary Fig. S1. Removal of adapter sequences and a quality-based trimming was performed for all samples using ReadTools[36].

DNA samples used for genotyping. Hair root samples from male horses were derived from breeding associations and private horse owners. Genomic DNA was isolated using DNeasy Blood and Tissue Kit (Qiagen®).

Y chromosome de novo assembly. Prior to the MSY *de novo* assembly we reduced the complexity of the input data. Therefore, we used a bioinformatics-based approach to enrich paired end Illumina short reads of three whole-genome sequenced Lipizzan stallions (sample IDs Lip111, Lip113 and Lip169 in Supplementary Table S1) for Y-specific sequences. We used published Y-chromosomal sequences – the 1.6 Mbp nonrepMSY contigs[23], six BAC-clones[22], human and mouse Y-chromosomal sequences (NCBI GenBank), and equine Y-chromosomal and XY-homologous GenBank entries[27,37] as baits to identify putatively Y-specific reads. A detailed description of the Y read enrichment step is given in the Supplementary Information and bait details are listed in Supplementary Table S2. We mapped each Lipizzan male with bwa aln[38] to a file with multiple FASTA-formatted sequences containing all the bait sequences. We then extracted mapped read-pairs using samtools[39] and used them as input for the generation of a *de novo* assembly with SPAdes[40]. Contigs shorter than 200 bp were discarded. We used REAPR[41] to correct for assembly errors. Details are shown in the Supplementary Information.

Classification of scY, mcY and nonMSY windows. Using a probabilistic model, we classified the assembly into single-copy (scY), multi-copy (mcY) Y and not MSY (nonMSY) regions (X-chromosomal, autosomal, pseudoautosomal or XY-homologous regions) based on coverage differences between females and males in 50 bp windows.

Preparing input data for the classification. Whole-genome Illumina data from five females and ten males (samples indicated in Supplementary Table S1) were mapped to the REAPR-corrected assembly and the equine X chromosome with bwa aln[38]. Per-site mean coverage of each 50 bp window and horse was determined using Unix commands, Python[42] and bedtools[43]. As sequencing-depth differs among individuals (see Supplementary Table S1, Supplementary Fig. S1), we normalised each window's mean coverage: an individual's diploid coverage was inferred as its mean coverage of windows in the pseudoautosomal region (PAR) of the X chromosome. For each horse, the mode of the distribution of the mean coverages of the PAR windows was calculated in R[44] and used to normalise the mean coverage per window in the Y chromosome assembly, such that a relative coverage of one corresponds to a diploid state. Details are given in the Supplementary Information.

During visual inspection of alignments, we noted that reads from females spuriously map to single-copy MSY regions at a background level. Each female's background coverage was estimated as its mean normalised mapping coverage in confirmed single-copy MSY windows. These single-copy MSY windows were regions to which the single-copy MSY contigs from the previously published nonrepMSY reference (GenBank accession MPVR00000000)[23] mapped using bwa mem[38] with the default settings.

Probabilistic model. The normalised coverages in 50 bp windows were modelled as Poisson distributed. Using R[44], each Y assembly window was assigned to one of two classes: i) MSY or ii) nonMSY. The observed mean coverages per window k and horse (i for males and j for females; provided to get the normalised mapping coverages y), each horse's mean mapping coverage of the PAR c and the female background coverages b, all obtained as described above, are provided as input for the R script. Assuming equal prior probabilities, the probability of assignment of a window to class i) or ii) can be calculated from this ratio (the R script and derivation of the formula are given in the Supplementary Information):

$$\frac{Pr\left(y_{i,k}, y_{j,k} | \hat{\nu}_k, b, c\right)}{Pr\left(y_{i,k}, y_{j,k} | \hat{\mu}_k, c\right)} = \frac{\prod_{i=1}^{I} (\hat{\nu}_k c_i/2)^{y_{i,k}} e^{-\hat{\nu}_k c_i/2} \prod_{j=1}^{J} b_k^{y_{j,k}} e^{-b_k}}{\prod_{i=1}^{I} (\hat{\mu}_k c_i)^{y_{i,k}} e^{-\hat{\mu}_k c_i} \prod_{j=1}^{J} \left(\hat{\mu}_k c_j\right)^{y_{j,k}} e^{-\hat{\mu}_k c_j}}$$

According to the observed distribution of windows assigned to class i), a relative coverage cut-off value less than one was selected to distinguish scY from mcY windows in the MSY fraction (see Fig. 1).

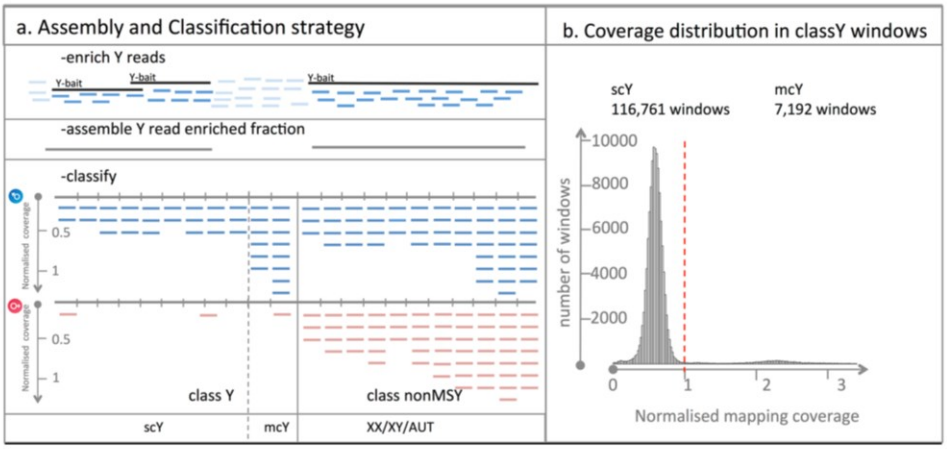
LipY764 predefinition. Contigs with a Y-specific content less than 45% were discarded from the assembly (see Supplementary Fig. S2). After filtering for contig length ≥ 300, the remaining classified 764 Y-specific reference contigs were further characterised (see below) and released (LipY764, PRJNA428358). The scY and mcY coordinates are provided in Supplementary Table S3 and other contig details are given in Supplementary Table S5. Summary statistics of the final assembly were computed with Python[42].

LipY764 comparisons and gene content. BLAST[45] was used to align the nonrepMSY reference[23] to the assembled contigs. BLAST[45] was also used to compare the LipY764 contigs with eMSYv3[46] by setting the thresholds for a match to nident > 299 and pident $> 95\%$.

Horse Y phylogeny. A horse Y phylogeny was generated using variants ascertained from whole-genome NGS data of 139 males (Supplementary Table S1).

Variant ascertainment and validation. The FastQC-checked, adapter free and trimmed data were mapped to the assembly using bwa aln[38]. PCR duplicates and low quality mappings were filtered with samtools[39]. Variant calling was performed using HaplotypeCaller and CombineGVCFs from GenomeAnalysisTK[47] (details in the Supplementary Information). As CombineGVCFs does not detect the full diversity when using a sample set with uneven haplotype distribution (for example many modern domestic and few Przewalski's horses) we ran CombineGVCFs for several sample compositions (for example only the Przewalski's horses, the deep-branching Asian horses or the crown group) to ensure valid calling also for HTs underrepresented in our dataset. Several filtering steps were performed on the outputs using Python[42] to predict a final list of high quality biallelic variants. First, only variants found in scY windows were considered and then phased variants, variants with multiple alternatives and reference errors were excluded. In the third step a read depth of at least three in one individual and a genotype quality $\geq$ nine were set as limit to keep the variant in the list. In the last step variants with heterozygous or no-calls in more than 10% of the samples were excluded. A final CombineGVCFs run with all samples was performed to implement the identified variants for all samples. Still ambiguous or missing variants for low-coverage samples were subsequently corrected according to their phylogenetic clustering. All positive single nucleotide variants (SNVs) and insertions or deletions (indels) detected in the Dom-West horses were checked visually in IGV[48], by comparing the site in multiple samples in parallel. 166 LipY764 variants (indicated in Supplementary Table S8) are already lab validated using LGC KASP technology, as described previously[23].

Phylogeny. Allelic states of 2,193 variants (2,035 described herein the first time; Supplementary Table S8) were catenated to construct MSY haplotypes. The ancestral variant state was inferred from the donkey or the Przewalski's horse for variants polymorphic only in domestic horses. A maximum parsimony (MP) tree rooted with the donkey was generated with PAUP[49] and bootstrapping with 1,000 replicates was performed. Further, a maximum likelihood (ML) tree was reconstructed with RAxML[50] (-m gtrgamma) with 1,000 bootstrap replicates. Trees were finalised using FigTree[51]. The topology and frequencies of the MSY haplotypes were additionally visualised with Network[52].



C.	total length in Mbp	number of contigs	N50 in bp	scY in Mbp	mcY in Mbp	nonMSY in Mbp
Raw assembly	11.73	1,935	21,876			
After REAPR and length filter	11.45	1,432	15,830			
After Classification	11.45	1,432	15,830	5.87	0.36	5.17
LipY764	6.46	764	15,552	5.84	0.36	0.26

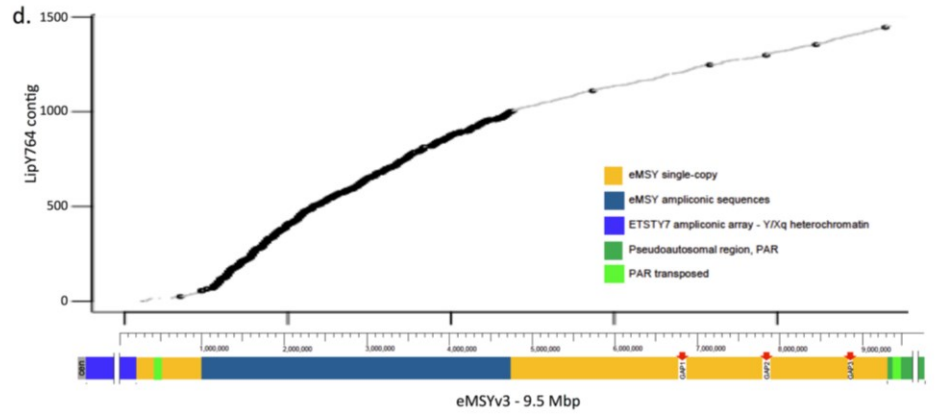


Figure 1. LipY764 Assembly. (**a**) Assembly step: whole genome NGS reads from males (blue) are mapped to Y-specific bait sequences (black). Mapped reads (dark blue) are then extracted and assembled (grey). Classification step: the assembly is shown in grey with hatchmarks representing 50 bp windows. Male (blue) and female (red) reads are mapped to the assembly and mapping coverages normalised to autosomal coverages per window were estimated. The probability of Y- or nonMSY-specificity per window is obtained by comparing normalised coverages in males and females. class nonMSY: XX/XY/AUT ≈ 1 in males and ≈ 1 in females; class Y: scY ≤ 1 in males and ≈ 0 in females, mcY > 1 in males and ≈ 0 in females. (**b**) Frequency distribution of normalised mapping coverage in classY windows. The cut-off scY to seperate mcY is set to 1 (red dashed line). (**c**) Resulting statistics for the assembly and classification approach. (**d**) Position of LipY764 contigs on eMSYv3[46] (data in Supplementary Table S5). Contigs having a single unique position on eMSYv3 are shown in grey, contigs with multiple hits in black/bold.

eMSYv3 variant coordinates. With BLAST[45] and Python[42] the variant coordinates were lifted over to eMSYv3[46] based on 100% identity of the flanking sequences (Supplementary Table S8). For variants on contigs with a single BLAST hit on eMSYv3[46] (Supplementary Table S5), whose positions were not detected with Python[42], the positions on eMSYv3[46] were inferred manually by allowing mismatches in the flanking region using CLC Genomics Workbench[53].

Microsatellite analysis. 109 individuals were genotyped for the tetranucleotide microsatellite fBVB (Supplementary Table S9, for details see Supplementary Information). Genomic DNA was isolated from hair roots using nexttec®. One primer was labelled with FAM fluorescent dye to allow analysis of fragments (fwd_FAM: ACAACCTAAGTGTCTGTGAATGA; rev: CCCAATAATATTCCACTGCGTGT, expected amplicon length 204 bp) on an ABI 3130xl Genetic Analyzer. PCR was performed in a 20 µl volume containing 0,4 µM of each primer. The DNA was initially denatured at 95 °C for 5 min, followed by 35 cycles of 30 s at 95 °C, 40 s at 58 °C

annealing temperature and 40 s at 72 °C, and a final extension of 30 min at 72 °C. Alleles were sized relative to the internal size standard using GeneMarker®.

Pedigree reconstruction and generation time intervals. We inferred the paternal genealogy for 333 males using pedigree information provided by breeding associations or matched information from several web-based databases (listed in Supplementary Information). Horses with inconsistent genealogical records across different databases were discarded. The reconstruction revealed 1,722 father-son pairs; the years of birth of sons range from 1680 to 2000. MSY haplogroups (HGs) of these 333 individuals were inferred by either whole genome sequencing or KASP genotyping (see Supplementary Information).

***De novo* mutation rate and branch length estimates.** We generated a maximum parsimony tree using only samples with a mean single-copy Y coverage ≥ six, considered only SNVs (Supplementary Table S8) and counted the number of mutations on branches. We tested pairs of branches of MSY haplotypes descending from the same split for equality of their numbers of mutations assuming a Poisson distribution with the same expectation using a chi-square test of equal frequencies and determined p-values by 10,000 simulations in R[44]. Correlation between branch length and mean generation interval was tested with a chi-square test for given probabilities with simulated p-value (based on 10,000 replicates, see Supplementary Information). We inferred a mutation rate with Darley Arabian, born in 1700 as calibration point. From pedigrees of 592 father-son pairs tracing back to Darley Arabian, on average 28.3 generations were counted in 320 years, resulting in a mean generation interval of 11.36 years.

We estimated a *de novo* mutation rate from our data by dividing the mean observed *de novo* mutations (2.75) per 28.3 generations by the length of the single-copy MSY sequence (5.83 Mb).

Molecular dating. We dated the most important nodes with BEAST[54] assuming a mutation rate of $1.69*10^{-9}$ mutations/site/year, which is based on the rate/site/generation calculated from our data (see results section) and a ten years generation interval. We generated a confidence interval allowing for plus/minus two years generation interval resulting in $2.11*10^{-9}$ mutations/site/year for eight and $1.41*10^{-9}$ mutations/site/year for twelve years generation time. The substitution model to best fit the data was chosen according to datamonkey[55] (HKY and gamma site heterogeneity). We used a constant-sized coalescent tree prior and a strict clock. A prior with a normal distribution based on the 95% CI of the substitution rate was applied. Only the variant sites were used and the number and composition of invariant sites was defined in the BEAST xml file. Markov chain Monte Carlo (MCMC) samples were based on 20,000,000 generations, logging every 1,000 steps. Two runs were combined using BEAST's[54] LogCombiner and TreeAnnotator with the first 10% discarded as burn-in. The final tree was visualised using FigTree[51].

Results and Discussion

Generation of a Y chromosome draft assembly. Y-chromosomal regions are often underrepresented in whole-genome shotgun assemblies[56]. Tomaszkiewicz *et al.*[57] efficiently assembled gorilla Y-chromosomal regions by flow-cell sorting the Y chromosome prior to Illumina sequencing. We mimicked the flow-cell enrichment computationally and generated a Y-enriched short read dataset by mapping whole-genome NGS data from three Lipizzan males (Supplementary Table S1) to a collection of publicly available Y-specific sequences (details in Supplementary Table S2). In total, 2,549,458 read-pairs mapped to the captured sequences and these reads were used for generating the raw assembly. The raw assembly had 1,935 contigs and a total size of 11,727,306 bp (with contig lengths between 56 and 112,142 bp and an N50 of 21,876 bp).

Next, we developed and applied a probabilistic method to classify the assembly into i) single-copy (scY) and multi-copy (mcY) MSY regions and ii) not Y-specific (nonMSY) regions based on different mapping coverages in males and females (see Material and Methods). An overview of the assembly and classification strategy including assembly statistics is shown in Fig. 1. After the classification and filtering steps, 764 non-overlapping contigs with a total size of 6.46 Mbp remained (LipY764; SAMN08288327). We classified 5.84 Mbp of LipY764 as scY, 0.36 Mbp as mcY and 0.26 Mbp as nonMSY (region details in Supplementary Table S3). The assignment of windows to class i) or ii) is strongly influenced by the mapping settings and complexity of the reference. Most scY windows are expected to be located in X-degenerated regions in old evolutionary strata[58] with high sequence divergence to the X gametolog.

The low repetitive content in our 6.46 Mbp MSY assembly can be explained by difficulties of the assembler to bridge highly repetitive regions based on short read information only[59]. A similar method, the 'chromosome quotient method', to classify single-copy regions on the hemizygous sex chromosome has been proposed earlier[60,61]. As our probabilistic model inferred posterior probabilities of assignment to scY, mcY, and nonMSY in 50 bp windows, we achieved a refined classification with little loss of information caused by false negatives. At first glance, LipY764 corresponds to 54% of the 12 Mbp euchromatic region proposed for the horse MSY[46]. Tomaszkiewicz *et al.*[57] doubled the size of their 13.3 Mbp short read Illumina assembly of the gorilla Y chromosome with mate-pair and PacBio data. However, as the ampliconic regions added by such efforts would not significantly increase the power for variant ascertainment and haplotype analysis, we did not attempt to further extend the assembly.

The finding that 2,765 of the 2,794 nonrepMSY contigs (GenBank accession MPVR00000000) used in previous studies[4,23] were embedded in LipY764 (Supplementary Table S4) confirms that LipY764 is an improvement of the nonrepMSY assembly. Very recently, Janečka *et al.*[46] published the first comprehensive assembly of the horse MSY with a total length of 9.5 Mbp, by sequencing 192 BACs from a BAC tiling path map (eMSYv3, GenBank accession MH341179).

This offered us the unique opportunity to assess the quality and complexity of LipY764 in a direct comparison. We found perfect eMSYv3 homologies for 585 LipY764 contigs (4.9 Mbp in total; Supplementary Table S5). When

we allowed for multiple matches on eMSYv3, LipY764 covered even 7.05 Mbp (74%) of eMSYv3 (Fig. 1d). Most LipY764 contigs with multiple matches were already classified as mcY over most of their length (Supplementary Table S5). The 1,122 gaps were evenly distributed along eMSYv3 (Supplementary Table S6) and gap lengths in the X-degenerate and the multicopy region of eMSYv3 ranged from 1 to 150,228 bp (median = 503 bp, 1st qu. = 163 bp, 3rd qu. = 1,408 bp).

Interestingly, 179 contigs (1.48 Mbp) were not found on eMSYv3 (which is based on a Thoroughbred), but were unique to LipY764 (Supplementary Table S5). These contigs were evenly covered by NGS reads in all males analysed, including Thoroughbreds. Hence, they are not large insertions in the Lipizzans used for the LipY764 assembly, but rather represent regions not yet assembled and potentially informative for closing the remaining gaps in eMSYv3.

The 9.5 Mbp eMSYv3 harbours about 4 Mbp that are not accessible for variant detection using short read data[46], such as the ampliconic region, the ETSY7 array, parts of the PAR and the PAR transposed region. For the ascertainment of variants to trace male lineages, LipY764 is advantageous because it contains 1.48 Mbp unique sequences and has therefore a larger total amount of scY regions. In contrast, only 853 kbp of the X-degenerate region of eMSYv3 are not covered by any LipY764 contig (Fig. 1, Supplementary Table S6). Altogether, the high concordance between the two references proves the reliability and the achievements we made with the more cost-effective and rapid approach we used to generate LipY764.

Of the 174 transcripts annotated on eMSYv3, 85 (48%) were covered to at least 95% on LipY764. Of the rest, 35 (20%) transcripts were partially and 54 (31%) not at all represented on LipY764 (Supplementary Table S7). Among the missing genes were those identified as recently transposed from the PAR to the MSY[46] (Fig. 1, light green) and many other MSY genes with high sequence similarity to X gametologs or autosomal paralogs (Supplementary Table S7). With our algorithm, such regions would be classified as X-chromosomal or autosomal and removed during the classification step from downstream analyses.

Fine-scaled Y-chromosomal haplotype structure in domestic horse breeds. We ascertained SNVs and small indels only in unambiguous scY windows of LipY764 predicted by the probabilistic model. By doing so we stringently excluded multi-copy and X-chromosomal/autosomal homologous regions, where unambiguously mapped reads could lead to inaccurate variant prediction (Supplementary Fig. S3). In total, we screened 139 whole-genome NGS sequenced males (see Methods) and ascertained 2,187 variants (2,027 SNVs, 159 indels and one microsatellite; see Supplementary Table S8), of which 152 have been already described previously[22–24]. Our dataset covers most Central European breeds. In addition to the 104 horses from Felkel *et al.*[24] we analysed eleven Thoroughbreds, 16 horses from Western European breeds and eight Przewalski's horses (sample details in Supplementary Table S1, Supplementary Fig. S1). Based on the 2,187 variants detected on LipY764 and implementing allelic states from six additional variants described previously[24] (of which one is invariant in our dataset), we distinguished 76 haplotypes. Within domestic horses, we observed 71 haplotypes determined by 740 variants (735 on LipY764 plus the five additional variants); the remaining 1,452 variants separated the Przewalski's horses from domestic horses. On the horse MSY, coalescence times are so recent that no double mutations at any site were detected and MP can be used to infer the genealogy (Fig. 2; corresponding network in Supplementary Fig. S4).

The phylogenetic trees (Fig. 2, Supplementary Fig. S5) correspond to previous results[23,24] based on the shorter nonrepMSY: the Przewalski's haplogroups (HG) Pz-a and Pz-b are clearly separated from the domestic horses. Within the domestic horses a few Asian samples branch early (HG O) and are connected to the crown via few Northern European (HGs N and I) and other Asian lines (HGs J, M and Y).

As expected, most modern horse breeds (115 horses or 83% of all samples) clustered in the previously described crown haplogroup (Fig. 2). This held true even for the Dülmen Pony, an ancient autochthonous Central European horse[10]. Within the crown, we distinguished 58 HTs defined by 211 variants. As the number of crown HTs almost doubled, we switched to human guidelines for MSY haplotype nomenclature[62], but kept the first four letters consistent with the previous studies[23,24]. We now resolved three haplogroups forming a trichotomy at the base of the crown group. Apart from the previously described clades A (green in Fig. 2) and T (yellow in Fig. 2)[22,24], we now resolved a new HG H (pink in Fig. 2) characterised by four variants (fRO, fRO, fYR, fBZU; Supplementary Fig. S4). All HG H samples except the Barb horse (subgroup Hs) were analysed previously[24] and formed independent groups (S - Sorraia horse, L - Lipizzan horses and C - Chinese Chakouyi horse). The history of breeds representing subgroup Hs so far (Barb horse, Spanish breeds, Sorraia horse, Baroque breeds) suggests that this HG entered Europe from the West via Spain by the introduction of North African horses[1,10,63].

The microcosmos of the Tb-clade and English Thoroughbred sire lines. More than half of the domestic horses in our dataset (76 of 130) carried a haplotype of HG Tb. These included English Thoroughbreds, Standardbreds, many Thoroughbred-influenced breeds (Warmbloods, American Quarter horses, Franches-Montagnes), a Lipizzan stallion, and the Akhal-Tekes. Previously, we identified HG Tb as a signature of the Turkoman horse, an ancient horse population from the steppes of central Asia[23]. During the past 300 years HG Tb was extensively spread by the English Thoroughbred. The Thoroughbred sire lines trace back to three founder stallions that were imported to England at the end of the 17th century[10,64]. Here, we fully resolved the heritage of the Thoroughbred sire lines with MSY haplotyping. The haplotype structure of 65 males was highly consistent with their paternal genealogy inferred from the pedigree (Fig. 3). We now clearly discriminate discrete sublines of Darley Arabian, born in 1700 (Tb-d) and Godolphin Arabian, born in 1724 (formerly Tb-g, now Tb-oB3b). The third founder, Byerley Turk, born in 1680, was newly characterised by an allelic variation of the tetranucleotide microsatellite fBVB ($GATA_{14}$/$GATA_{15}$; Supplementary Table S8) that defines the Tb-oB1 clade. According to pedigree information only few of the tested males trace back paternally to Byerley Turk. We thus screened a representative dataset of 109 purebred males by genotyping fBVB and detected allelic variation of fBVB only in Tb

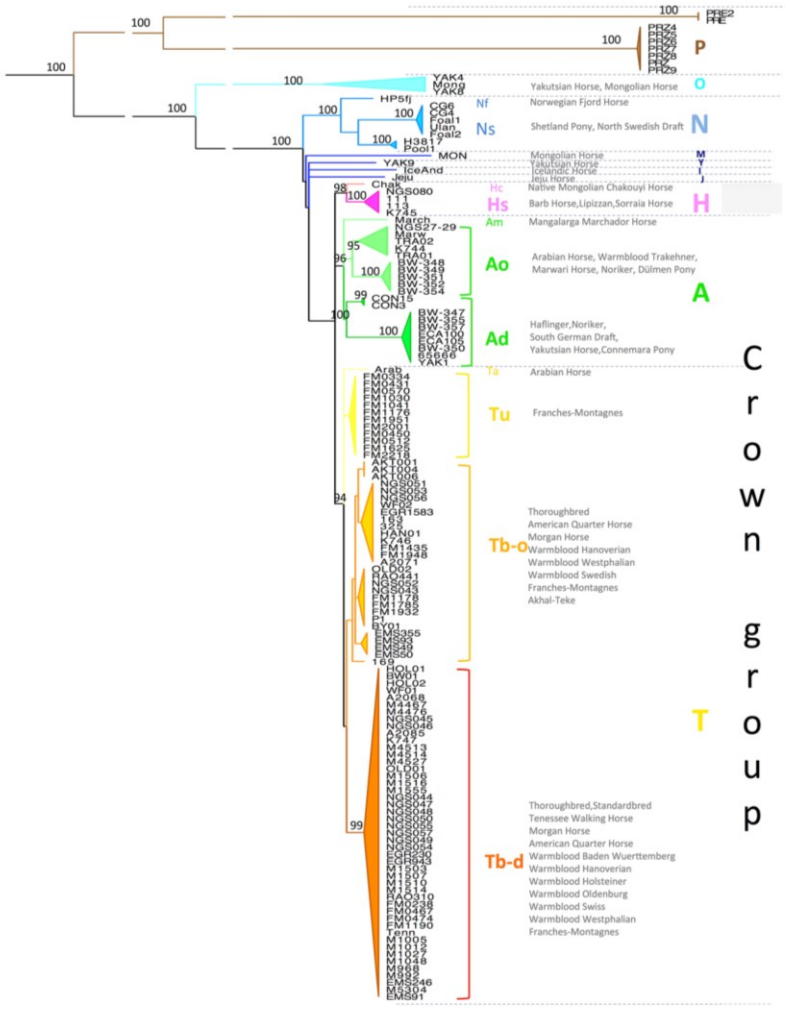


Figure 2. Horse MSY tree. A maximum parsimony tree showing the horse MSY phylogeny based on 2,192 scY variants detected in 139 males. The tree is rooted with the donkey and bootstrap values of 90% or higher are shown. The Przewalski's horses are shown in brown. Blueish clades correspond to early splitting Asian samples (O), Northern European breeds (N and I) and other autochthonous Asian samples (M, Y and J). The three clearly separated crown group clades are represented in pink (H), green (A) and orange (T) shades. Assigned haplogroups are shown on the right. A detailed haplotype network with variants is shown in Supplementary Fig. S4, variant details are given in Supplementary Table S8.

clade horses (Supplementary Table S9). All 30 patrilineal descendants that coalesce in Herod, born in 1758, whose ancestry in turn traces back to Byerley Turk, carried the Tb-oB1 specific allele 208 (Supplementary Table S9). We also confirmed Tb-oB1 in eleven horses tracing back to St. Simon, born in 1881. According to stud records, St. Simon was the son of Galopin, born in 1872, who was sired by Vedette, born in 1854. The line should trace back to Eclipse, born in 1764 (Fig. 3) according to stud books. All descendants of St. Simon carry Tb-oB1, undoubtedly the HT of Herod (Fig. 3), and not that of Eclipse (Tb-dW*). Thus, an incorrect paternity assignment must have occurred in this lineage. In a discussion recorded in the early to mid 19th century, one party claimed that instead of Vedette a moderate performer named Delight, born in 1863, a Byerley Turk descendant (Fig. 3b), fathered Galopin[65–67]; our molecular data support this view.

In addition to the patrilines of the English Thoroughbred, we distinguished Tb-oB3a, Tb-oB2a, and Tb-oL represented by the Akhal-Tekes, the Morgan horses and a Lipizzan horse in the Tb clade (Fig. 3a). The private clustering of these lines suggests an origin from a similar source population but independent from the Thoroughbreds.

Potential and perspectives for individual MSY haplotyping in horses. In humans analysis of the MSY is well-established in forensics[68]. Due to the Y chromosome being a single locus and mirroring the genealogy of only the male sex, its analysis allows only limited conclusions about the rest of the genome. However, as the male side plays such an important role in horse breeding, MSY genealogies reveal not only paternal ancestry of horses but also breeding history in general. Our fine-scaled resolution of the individual Thoroughbred lines

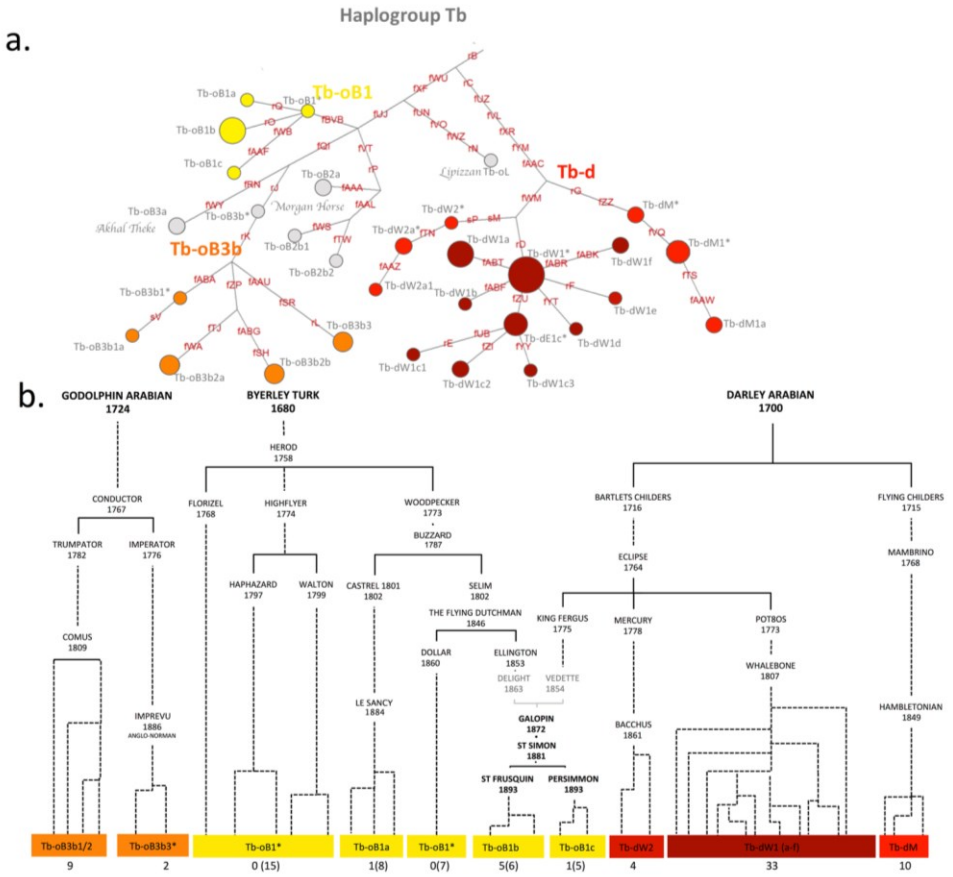


Figure 3. Detailed view on haplogroup Tb. (**a**) Haplotype network of group Tb. Circle sizes correspond to the number of samples. Nomenclature of HTs is based on Wallner *et al.*[23] and subbranches according to human guidelines[62]. Determining variants are given on branches (details in Supplementary Table S8). HTs derived from Darley Arabian are shown in red, Godolphin Arabian in orange and Byerley Turk in yellow. (**b**) Pedigree reconstruction of English Thoroughbred descendants and the respective HTs. Dotted lines connect relatives where at least one ancestor is omitted. For each HT the number of samples in the NGS dataset is given and the number of genotyped individuals, if available (Table S9), shown in parentheses.

underlines that horse MSY haplotyping is a practically valuable and accurate method to assess male ancestry. By further augmenting the haplotype network with sire line representatives from other classical refiner breeds (e.g., the imported Arabians and the Iberian/North African horses), our knowledge of the development and history of present-day horse breeds can be further improved. Horses with doubtful ancestry can now be assigned to sire lines and recent male-driven introgressions can be documented in detail[69].

Currently we have no indication of fitness effects of the variants reported here. Four of the 24 variants in transcribed regions (Supplementary Table S8) were in UTRs and thus not coding. Of the remaining 20, only five were polymorphic in domestic horses, the other 15 separate Przewalski's and domestic horses. Due to the limited number of candidates and the lack of phenotypic data, we refrained from further functional investigation. So far, the major phenotypic effects (mainly reduced fertility) described for the Y chromosome are due to structural and copy number variations[12,70–72], which we did not examine in the context of this work.

Dating historic horse radiations. The high mtDNA diversity observed in modern domestic horses reflects the preservation of diverse wild ancestral maternal lineages[73,74]. In contrast, most ancient MSY variation[5,8] was lost and the low MSY variation in modern domestic horses mirrors a very recent origin of extant sire lines. Knowledge of the exact timing and location of the emergence of the predominant MSY lineages would allow identification of the corresponding human cultures and improve our understanding of the human-horse history.

We detected two early splits in the MSY phylogeny, one separating the Przewalski's from modern domestic horses (Cab-Prz) and the other between Western and Asian haplotypes (Dom-All). Additionally, we observed two recent consecutive radiations: 'Dom-West', encompassing all non-Asian domestic horses with the deepest branch leading to the Shetland Pony and the North Swedish Draft Horses (N), followed by the more recent crown radiation (Fig. 4a). We postulate that the dispersal of the crown started from a population already harbouring the basal HTs of the crown's sublines (grey rings in Fig. 4a). The ancestors of the classical refiner populations - North African and Iberian, Arabian, and Turkoman horses - likely originate from this ancestral pool. Additionally, a Chinese lineage falls into this group.

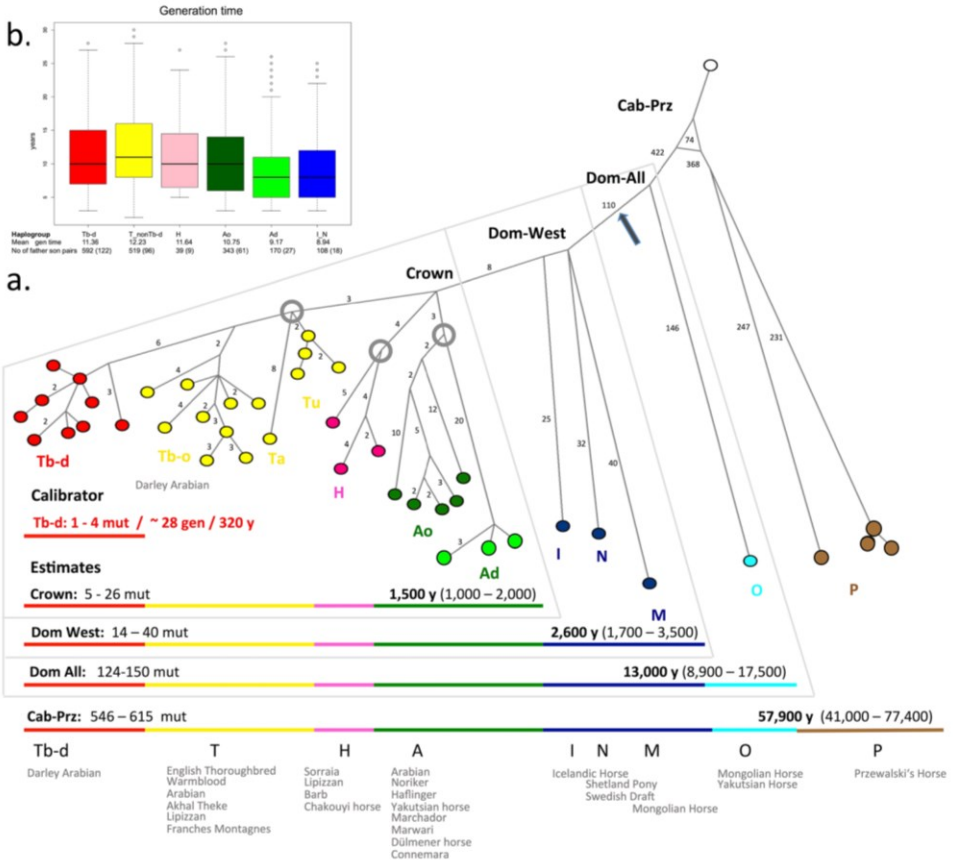


Figure 4. Divergence time estimates. (**a**) Maximum parsimony tree rooted with the donkey. Coloured circles represent contemporary haplogroups; grey rings indicate basal HTs of the crowns' sublines. The number of mutations on a branch is given on its left, unless it is one. In the lower panel the full range of mutations observed after respective coalescence points (mut) and years back to the MRCA (y) under the assumption of a mutation rate of 1.69×10^{-8} site/generation and assuming a generation interval from eight to twelve years (**b**) is shown. 95% highest posterior density intervals are given in brackets. The position of variant fBOI (indicative for Y-HT1 in Wutke *et al.*[77]) is marked by an arrow. Details on variants are given in Supplementary Table S8. (**b**) Mean generation intervals calculated from deep pedigrees from males genotyped for the respective haplogroup. The number of father-son pairs is given with genotyped individuals in parentheses (data in Supplementary Table S10).

The unbiased variant ascertainment in our approach would allow for exact dating of coalescent times under the assumption of a molecular clock: with a constant mutation rate and identical generation intervals, branches descending from one split are expected to accumulate new mutations at the same constant rate.

For dating, we omitted low-coverage samples, did not consider indels (see Methods) and inferred the genealogy from 42 HTs based on 1,856 variants assuming parsimony (Fig. 4a). To test for the expected equality in branch lengths, we assumed a Poisson distribution (see Methods). We found significant deviations from an equal distribution of mutations on individual branches within recently emerged haplogroups, e.g., between branches in Dom-West (14 to 40, $X^2 = 34.532$, p-value < 0.01) and branches in the crown (five to 26 mutations; $X^2 = 34.55$, p-value < 0.01). No such deviation was observed when contrasting number of mutations accumulating from older branching points (dom_all, dom_prz). Most obvious was the contrast between the exceptionally long Ad-h branch (Noriker Coldblood horse) and the shorter Tb branches (Thoroughbred).

Overestimation of branch lengths could have been induced by several mutations that occurred simultaneously due to X-Y gene conversion[22,75] but were by chance not marked as phased variants by GATK. These variants would have been retained in our final variant list and by mistake assumed as independent. Additionally, it could be that the X chromosomes of our five females are too dissimilar from potential gene conversion regions such that our method could not identify them as nonMSY regions. We ruled out X-Y gene conversion as causative for the long branches, since we did not observe a signature of closely clustering variants occurring on the same contig (Supplementary Fig. S6, Supplementary Table S8).

As stated above, the excess of variation could also be due to varying generation times. Mean generation intervals in father-son pairs inferred from deep rooting pedigrees, genotyped for respective HGs (Supplementary Table S10), followed the trend observed in branch lengths (Fig. 4b). Coldblooded males, often carrying Ad, tended to have shorter mean generation intervals (8.94 years) than horses carrying Tb (mean 11.3 to 12.23 years;

Fig. 4b) in the past 300 years. We tested the two most extreme branches, namely Ad-h and T carriers; differences in generation intervals were insufficient to account for the different numbers of mutations ($X^2 = 8.6541$, p-value < 0.01).

The violation of the assumption of a constant mutation rate per unit time compromised dating. Nevertheless, we estimated the time to most recent common ancestors (MRCA; Fig. 4b). We calibrated the mutation rate using mutations after Darley Arabian (Tb-d). The resulting mutation rate (1.69×10^{-8} mutations/site/generation or one *de novo* mutation every 10.2 generations in the MSY region under investigation) was lower than the previous estimate (2.91×10^{-8} mutations/site/generation) based on four *de novo* mutations in paternal families[23], but agreed much better with genome-wide estimates in humans[76]. Despite the wide range of our estimates (Fig. 4b, Supplementary Fig. S7), we confirmed that all patrilineal Western horse lineages (Dom-West) arose after domestication (MRCA 1,700–3,500 years before present) with the crown embedded therein (MRCA 1,000–2,000 years before present). The analysis of dated archaeological samples will make it possible to more precisely determine the cultural and biogeographic origin of Dom-West and the crown, and to more thoroughly resolve paternal ancestry of horses. In a recent study based on a collection of ancient samples capturing nearly the entire history of horse domestication[77] it was revealed that a single Y-HT, first observed in a 4,200 year old sample, invaded and spread in the population until it reached fixation. The SNV fBOI, defining the 'fixation Y-HT1' in Wutke *et al.*[77], is one of the 110 variants separating Dom-West from the Przewalski's horse and the Asian haplogroup (indicated by an arrow in Fig. 4b, details in Supplementary Table S8). The signal of Y-HT1 in Wutke *et al.*[77] therefore comprises all Dom-West HTs and its emergence is consistent with our dating estimate.

Conclusion

Until recently, the only means to trace sire lines were often incomplete or even erroneous pedigree data. Here, we resolved the MSY sequence variation of horses at a resolution comparable to that in humans. We have enabled Y-chromosomal barcoding of individual sire lines and paved the way for forensic applications. Our robust MSY phylogeny based on biallelic markers will serve as backbone for studying the paternal ancestry of horses on a worldwide scale. Moreover, the incorporation of ancient DNA data should further elucidate the origin of extant lineages in the near future.

Data Availability

The LipY764 contigs can be downloaded from NCBI PRJNA428358 (SUB3434928). Mapped NGS reads of all samples in this paper have been submitted to SRA archive PRJNA430351 (SUB3548921). The newly identified variants, their coordinates and flanking sequences can be found in Supplementary Table S8. Variants being polymorphic in domestic horses and their position detected on eMSYv3 were submitted to ENA (ID will be included in Supplementary Table S8 as soon as received).

References

1. Kelekna, P. *The Horse in Human History*. (Cambridge University Press 2009).
2. FAOSTAT www.fao.org/faostat. Accessed 2019 March 24 (2017).
3. Rischkowsky, B. & Pilling, D. *The State of the World's Animal Genetic Resources for Food and Agriculture*. (FAO 2007).
4. Gaunitz, C. *et al.* Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science.* **360**, 111–114 (2018).
5. Librado, P. *et al.* Ancient genomic changes associated with domestication of the horse. *Science.* **445**, 442–445 (2017).
6. Peters, J. *Römische Tierhaltung und Tierzucht: eine Synthese aus archäozoologischer Untersuchung und schriftlich-bildlicher Überlieferung*. (M. Leidorf 1998).
7. Wutke, S. *et al.* Spotted phenotypes in horses lost attractiveness in the Middle Ages. *Sci Rep.* **7**(6), 38548 (2016).
8. Lippold, S. *et al.* Discovery of lost diversity of paternal horse lineages using ancient DNA. *Nat Commun* **2**, 450 (2011).
9. Librado, P. *et al.* The Evolutionary Origin and Genetic Makeup of Domestic Horses. *Genetics* **204**, 423–434 (2016).
10. Hendricks, B. L. *International Encyclopedia of Horse Breeds*. (University of Oklahoma Press 2007).
11. Poznik, G. D. *et al.* Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.* **12**, 809–809 (2016).
12. Jobling, M. A. & Tyler-Smith, C. Human Y-chromosome variation in the genome-sequencing era. *Nat. Rev. Genet.* **18**, 485–497 (2017).
13. Edwards, C. J. *et al.* Dual origins of dairy cattle farming–evidence from a comprehensive survey of European Y-chromosomal variation. *PLoS One* **6**, e15922 (2011).
14. Guirao-Rico, S. *et al.* Porcine Y-chromosome variation is consistent with the occurrence of paternal gene flow from non-Asian to Asian populations. *Heredity (Edinb).* **120**, 1–14 (2018).
15. Vidal, O. *et al.* Differential distribution of Y-chromosome haplotypes in Swiss and Southern European goat breeds. *Sci. Rep.* **7**, 1–6 (2017).
16. Ding, Z.-L. L. *et al.* Origins of domestic dog in Southern East Asia is supported by analysis of Y-chromosome DNA. *Heredity (Edinb).* **108**, 507–514 (2012).
17. King, T. E. & Jobling, M. A. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends Genet* **25**, 351–360 (2009).
18. Kayser, M. Forensic use of Y-chromosome DNA: a general overview. *Hum. Genet.* **136**, 621–635 (2017).
19. Lindgren, G. *et al.* Limited number of patrilines in horse domestication. *Nat Genet* **36**, 335–336 (2004).
20. Kreutzmann, N., Brem, G. & Wallner, B. The domestic horse harbours Y-chromosomal microsatellite polymorphism only on two widely distributed male lineages. *Anim Genet* **45**, 460 (2014).
21. Wallner, B., Piumi, F., Brem, G., Müller, M. & Achmann, R. Isolation of Y chromosome-specific microsatellites in the horse and cross-species amplification in the genus Equus. *J Hered* **95**, 158–164 (2004).
22. Wallner, B. *et al.* Identification of Genetic Variation on the Horse Y Chromosome and the Tracing of Male Founder Lineages in Modern Breeds. *PLoS One* **8** (2013).
23. Wallner, B. *et al.* Y Chromosome Uncovers the Recent Oriental Origin of Modern Stallions. *Curr Biol. Jul 10* **27**, 2029–2035 (2017).
24. Felkel, S. *et al.* Asian horses deepen the MSY phylogeny. *Anim. Genet.* **49**, 90–93 (2018).
25. Frischknecht, M. *et al.* Imputation of sequence level genotypes in the Franches-Montagnes horse breed. *Genet. Sel. Evol.* **46**, 1–8 (2014).
26. Jun, J. *et al.* Whole genome sequence and analysis of the Marwari horse breed and its genetic origin. *BMC Genomics* **15**, S4 (2014).

27. Huang, J. *et al.* Analysis of horse genomes provides insight into the diversification and adaptive evolution of karyotype. *Sci. Rep.* **4**, 1–8 (2014).
28. Schaefer, R. J. *et al.* Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics* **18**, 1–18 (2017).
29. Metzger, J. *et al.* Next generation sequencing gives an insight into the characteristics of highly selected breeds versus non-breed horses in the course of domestication. *BMC Genomics* **15**, 562 (2014).
30. Rafati, N. *et al.* Large Deletions at the SHOX Locus in the Pseudoautosomal Region Are Associated with Skeletal Atavism in Shetland Ponies. *G3* **6**, 2213–2223 (2016).
31. Andersson, L. S. L. *et al.* Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature* **488**, 642–646 (2012).
32. Finno, C. J. *et al.* SERPINB11 Frameshift Variant Associated with Novel Hoof Specific Phenotype in Connemara Ponies. *PLoS Genet.* **11**, 1–17 (2015).
33. Der Sarkissian, C. *et al.* Evolutionary genomics and conservation of the endangered Przewalski's horse. *Curr. Biol.* **25**, 2577–2583 (2015).
34. Doan, R. *et al.* Whole-Genome sequencing and genetic variant analysis of a quarter Horse mare. *BMC Genomics* **13**, 78 (2012).
35. Bauer, A. *et al.* A Nonsense Variant in the ST14 Gene in Akhal-Teke Horses with Naked Foal Syndrome. *G3 (Bethesda).* **7**, 1315–1321 (2017).
36. Gómez-Sánchez, D. & Schlötterer, C. ReadTools: A universal toolkit for handling sequence data from different sequencing platforms. *Mol. Ecol. Resour.* **18**, 676–680 (2018).
37. Paria, N. *et al.* A gene catalogue of the euchromatic male-specific region of the horse y chromosome: comparison with human and other mammals. *PLoS One* **6**, e21374 (2011).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows – Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
40. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
41. Hunt, M. *et al.* REAPR: A universal tool for genome assembly evaluation. *Genome Biol.* **14** (2013).
42. Python Software Foundation. Python v3.5.1 *Available at* http://www.python.org.
43. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
44. Team, R. C. R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria.* (2014).
45. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **15**, 521 (2009).
46. Janečka, J. E. *et al.* Horse Y chromosome assembly displays unique evolutionary features, putative stallion fertility genes and horizontal transfer. *Nat. Commun.* **9**, 2945 (2018).
47. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).
48. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
49. DL, S. Phylogenetic analysis using parsimony and other methods version 4.0 beta version (2002).
50. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
51. Rambaut, A. FigTree v1.4.2. http://tree.bio.ed.ac.uk/software/figtree/.
52. Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
53. © QIAGEN 2018. CLC Genomics Workbench 7.7.1 https://www.qiagenbioinformatics.com/.
54. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, 1–5 (2018).
55. Delport, W., Poon, A. F. Y., Frost, S. D. W. & Kosakovsky Pond, S. L. Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455–2457 (2010).
56. Renaud, G. *et al.* Improved de novo genomic assembly for the domestic donkey. *Sci. Adv.* **4**, 1–11 (2018).
57. Tomaszkiewicz, M. *et al.* A time- and cost-effective strategy to sequence mammalian Y chromosomes: An application to the de novo assembly of gorilla Y. *Genome Res.* **26**, 530–540 (2016).
58. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
59. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* **13**, 36–46 (2013).
60. Smeds, L. *et al.* Evolutionary analysis of the female-specific avian W chromosome. *Nat. Commun.* **6** (2015).
61. Hall, A. B. *et al.* Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. *BMC Genomics* **14** (2013).
62. Hammer, M. A nomenclature system for the tree of human Y-Chromosomal binary haplogroups. *Genome Res.* **12**, 339–348 (2002).
63. Royo, L. J. *et al.* The origins of Iberian horses assessed via mitochondrial DNA. *J. Hered.* **96**, 663–669 (2005).
64. Cunningham, E. P., Dooley, J. J., Splan, R. K. & Bradley, D. G. Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Anim Genet* **32**, 360–364 (2001).
65. Richardson, C. In *The New Book of the Horse, Volume 1* 628 (Funk & Wagnalls Company 1911).
66. Wall, J. F. In *Famous Running Horses: Their Forbears and Descendants Sportsman's Press book* 313 (Infantry Journal Press 1949).
67. Moorhouse, E. In *The history and romance of the Derby, Vol 2* (Biogr. Press 1911).
68. Larmuseau, M. H. D., Van Geystelen, A., Kayser, M., van Oven, M. & Decorte, R. Towards a consensus Y-chromosomal phylogeny and Y-SNP set in forensics in the next-generation sequencing era. *Forensic Sci. Int. Genet.* **15**, 39–42 (2015).
69. Librado, P. *et al.* Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proc. Natl. Acad. Sci.* **112**, 201513696 (2015).
70. Peterlin, B., Kunej, T. & Hristovski, D. Diagnostic test for Y chromosome microdeletion screening in male infertility. *Genet. Test.* **8**, 45–49 (2004).
71. Toure, A. *et al.* A New Deletion of the Mouse Y Chromosome Long Arm Associated with the Loss of Ssty Expression, Abnormal Sperm Development and Sterility. *Genetics* **166**, 901–912 (2004).
72. Raudsepp, T. *et al.* Molecular heterogeneity of XY sex reversal in horses. *Anim. Genet.* **41**, 41–52 (2010).
73. Achilli, A. *et al.* Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl. Acad. Sci.* **109**, 2449–2454 (2012).
74. Lippold, S., Matzke, N. J., Reissmann, M. & Hofreiter, M. Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. *BMC Evol Biol* **11**, 328 (2011).
75. Trombetta, B., Cruciani, F., Underhill, Pa, Sellitto, D. & Scozzari, R. Footprints of X-to-Y Gene Conversion in Recent Human Evolution. *Mol. Biol. Evol.* **27**, 714–725 (2010).
76. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
77. Wutke, S. *et al.* Decline of genetic diversity in ancient domestic stallions in Europe. *Sci. Adv.* **4** (2018).

Acknowledgements
Funding was provided by the Innovation Fund of the Austrian Academy of Sciences (ÖAW) and the Austrian Federal Ministry of Agriculture, Forestry, Environment and Water Management (DAFNE, 101184). We thank Leif Andersson for providing NGS data; Anthony Byles and Judy Baugh from bloodlines.net for helping with the literature on English Thoroughbred genealogies; Ines Butler-Wemken, Suzi Pritchard Jones, Max Dobretsberger for organising samples and Bogarka Balint for microsatellite analysis.

Author Contributions
S.F., C.V. and B.W. conceived and designed the experiments. S.F., C.V. and B.W. analysed the data. D.R. and V.D. performed genotyping experiments. B.P.C., O.D., R.F., V.J., J.J.J., T.L., G.L., M.M., J.M., M.N., T.R., T.R., S.R., C.J.R., R.S, C.S., G.T., J.T., B.V., and G.B. contributed data and resources. S.F., C.V., B.C. and B.W. wrote the paper. All authors read and approved the manuscript.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-42640-w.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2. Publication II

genes

MDPI

Article

Y-Chromosomal Insights into Breeding History and Sire Line Genealogies of Arabian Horses

Viktoria Remer [1], Elif Bozlak [1,2], Sabine Felkel [1,2,3], Lara Radovic [1,2], Doris Rigler [1], Gertrud Grilz-Seger [1], Monika Stefaniuk-Szmukier [4,5], Monika Bugno-Poniewierska [5], Samantha Brooks [6], Donald C. Miller [7], Douglas F. Antczak [7], Raheleh Sadeghi [7,8], Gus Cothran [9], Rytis Juras [9], Anas M. Khanshour [9,10], Stefan Rieder [11], Maria C. Penedo [12], Gudrun Waiditschka [1], Liliya Kalinkova [13], Valery V. Kalashnikov [13], Alexander M. Zaitsev [13], Saria Almarzook [14], Monika Reißmann [14], Gudrun A. Brockmann [14], Gottfried Brem [1] and Barbara Wallner [1,*]

1 Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna, 1210 Vienna, Austria; vet.remer@gmail.com (V.R.); elif.bozlak@vetmeduni.ac.at (E.B.); lara.radovic@vetmeduni.ac.at (L.R.); doris.rigler@vetmeduni.ac.at (D.R.); gertrud.grilz@vetmeduni.ac.at (G.G.-S.); GWaiditschka@in-the-focus.com (G.W.); gottfried.brem@bremkg.de (G.B.); sabine.felkel@boku.ac.at (S.F.)
2 Vienna Graduate School of Population Genetics, University of Veterinary Medicine Vienna, 1210 Vienna, Austria
3 Institute of Computational Biology, Department of Biotechnology, BOKU University of Life Sciences and Natural Resources, 1190 Vienna, Austria
4 Department of Animal Molecular Biology, National Research Institute of Animal Production, 32-083 Balice, Poland; monika.stefaniuk@iz.edu.pl
5 Department of Animal Reproduction, Anatomy and Genomics, University of Agriculture in Kraków, 30-059 Kraków, Poland; monika.bugno-poniewierska@urk.edu.pl
6 Department of Animal Science, UF Genetics Institute, University of Florida, Gainesville, FL 32610, USA; samantha.brooks@ufl.edu
7 Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA; dm96@cornell.edu (D.C.M.); dfa1@cornell.edu (D.F.A.); raheleh.sadeghi2@gmail.com (R.S.)
8 Equestrian Federation of The Islamic Republic of Iran, Tehran 193954754, Iran
9 College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA; GCothran@cvm.tamu.edu (G.C.); rjuras@cvm.tamu.edu (R.J.); anas.khanshour@gmail.com (A.M.K.)
10 Center for Pediatric Bone Biology and Translational Research, Texas Scottish Rite Hospital for Children, Dallas, TX 75219, USA
11 Identitas AG, R&D, 3014 Bern, Switzerland; stefan.rieder@identitas.ch
12 Veterinary Genetics Laboratory, University of California, Davis, CA 95617-1102, USA; mctorrespenedo@ucdavis.edu
13 All-Russian Research Institute for Horse Breeding, 391105 Ryazan, Russia; l.v.kalinkova@gmail.com (L.K.); vniik08@mail.ru (V.V.K.); amzaitceff@mail.ru (A.M.Z.)
14 Albrecht Daniel Thaer-Institut, Humboldt-Universität zu Berlin, 10115 Berlin, Germany; saria.marzook.de@gmail.com (S.A.); monika.reissmann@agrar.hu-berlin.de (M.R.); gudrun.brockmann@agrar.hu-berlin.de (G.A.B.)
* Correspondence: barbara.wallner@vetmeduni.ac.at; Tel.: +43-1-25077-5625

check for updates

Citation: Remer, V.; Bozlak, E.; Felkel, S.; Radovic, L.; Rigler, D.; Grilz-Seger, G.; Stefaniuk-Szmukier, M.; Bugno-Poniewierska, M.; Brooks, S.; Miller, D.C.; et al. Y-Chromosomal Insights into Breeding History and Sire Line Genealogies of Arabian Horses. *Genes* **2022**, *13*, 229. https://doi.org/10.3390/genes13020229

Academic Editor: Anna V. Kukekova

Received: 21 December 2021
Accepted: 21 January 2022
Published: 26 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: The Y chromosome is a valuable genetic marker for studying the origin and influence of paternal lineages in populations. In this study, we conducted Y-chromosomal lineage-tracing in Arabian horses. First, we resolved a Y haplotype phylogeny based on the next generation sequencing data of 157 males from several breeds. Y-chromosomal haplotypes specific for Arabian horses were inferred by genotyping a collection of 145 males representing most Arabian sire lines that are active around the globe. These lines formed three discrete haplogroups, and the same haplogroups were detected in Arabian populations native to the Middle East. The Arabian haplotypes were clearly distinct from the ones detected in Akhal Tekes, Turkoman horses, and the progeny of two Thoroughbred foundation sires. However, a haplotype introduced into the English Thoroughbred by the stallion Byerley Turk (1680), was shared among Arabians, Turkomans, and Akhal Tekes, which opens a discussion about the historic connections between Oriental horse types. Furthermore, we genetically traced Arabian sire line breeding in the Western World over the past 200 years.

This confirmed a strong selection for relatively few male lineages and uncovered incongruences to written pedigree records. Overall, we demonstrate how fine-scaled Y-analysis contributes to a better understanding of the historical development of horse breeds.

Keywords: horse breeding; foundation sire; Y chromosome; paternal lineage tracing; haplotype; Arabian horse; pedigree; male genealogy; genotyping

1. Introduction

Since their first domestication approximately 5000 years ago, horses have been inextricably linked to human societies [1,2]. Apart from their economic value and significant impact as working animals, horses have a pronounced cultural heritage value and exert a strong emotional attraction on people. The evolution of horse breeds [3], the relationship among breeds [4], or even the ancestry and influence of single individuals [5] are of particular interest to the scientific and public communities.

Most of our insights into the origins and the genetic compositions of horse breeds rely on studbook records and on genomic studies which are predominantly driven by investigations of autosomal loci or the maternally inherited mitochondrial DNA (mtDNA) (reviewed in [6]). These studies delineated the fact that the domestic horse populations continuously changed in the course of historic development [7] with the most pronounced changes occurring in the past 200 years as a result of intensive selective breeding.

In this recent period, the prevalent strategy to introduce new phenotypes or consolidate traits was and remains the increased use of carefully selected stallions [8,9]. To date, insights into stallion-mediated refinement derived from studbook information indicates that stallions used for breed improvement were often imported from distant regions, that they were extensively shifted between studs, and that their heritage was amplified by their sons and grandsons. This cumulative impact of a stallion led to the establishment of a 'sire line'. By definition, members of a sire line trace back to a single, often renowned, foundation individual in their paternal line of inheritance. Due to intensive selection of stallions, the predominance of a few sire lines is the rule rather than the exception in intensively managed horse breeds [10,11]. The English Thoroughbreds, with a well-documented closed studbook first published in 1791 [12], provides a compelling example of this sex bias in horse breeding with the tail-male lineages of only three foundation sires being retained today [13,14].

However, the depth and quality of breeding records vary significantly among breeds. Besides the English Thoroughbred studbook, the deepest breed registries, dating back to the 18th century (up to 30 generations back), exist for some Central European breeds (for example [15,16]). The pedigree records illustrate the remarkable influence of Spanish, Arabian, and Thoroughbred stallions on the formation of many modern breeds. Given the eminence of stallions in breeding, a pedigree-independent genetic characterization of patrilines in a breed would be a milestone for the horse community.

The genetic compartment that perfectly matches the paternal lineage in mammals is the male-specific, nonrecombining part of the Y chromosome (MSY). The MSY is inherited as a single linkage group, defined as a haplotype (HT), from a father to his sons. Due to the close association with the patriline, the MSY became the most popular marker in human genetic genealogy [17,18]. In horses, MSY analysis could reveal the influence and origin of breeding stallions, similar to human family history research, or even be useful for forensic applications [19].

Recently, a stable horse MSY HT topology based on slow evolving biallelic markers was defined using next generation sequencing (NGS) technology [20–22]. Domestic horse MSY HTs are clearly distinct from those in the Przewalski's Horse. The most pronounced MSY signature in modern horse breeds is the predominance of a very recently expanded haplogroup (HG; in this context a defined group of closely related HTs). All Central and

South European, American, and most East Asian horses investigated to date carried this so-called 'crown' HG. The most recent common ancestor (MRCA) of the crown was dated to 1000–2000 years before present [22]. The crown was postulated to be the footprint of Oriental horses [20], but the exact ways and timeframes by which the Oriental HTs were disseminated in the past are not yet fully resolved. 'Non-crown' HTs have so far been detected only in Asian horses [21,23,24] and in some northern European breeds [20,25].

The feasibility of MSY patriline tracing with biallelic markers, even among the recently established crown HTs, was demonstrated in the three English Thoroughbred sire lines [22], where deep pedigree information exists back to the 18th century. The HTs carried by the foundation sires of the Thoroughbred were inferred by combining sire line genealogies from pedigrees with MSY HTs ascertained from NGS sequencing data in numerous extant descendants. The three Thoroughbred foundation sires carried an HT in the crown HG 'Tb'. Moreover, several 'subline-HTs', which arose from de novo mutations within the pedigree-supported timeframe, were detected. Those 'subline-HTs' could be unambiguously linked to a specific offspring and thus serve as specific markers for the breeding influence of particular stallions on the genealogical scale. The most prominent HT 'Tb-dW1', specific to the progeny of the stallion Whalebone (1807), unequivocally delineates the influence of this bloodline on many breeds [20,26,27]. However, apart from the Thoroughbred lines, the MSY HTs for the many other influential stallions and refining breeds are not yet defined.

In this study, we introduce fine-scaled MSY haplotyping in Arabian horses. Selectively bred over centuries by Arab people, the Arabian horse can be considered the oldest horse breed in the world. Arabian horses were used for enhancement in many of today's breeds, but there are few areas in the world that have not used Arabian bloodlines to refine native horses [8,28]. As shown in a recent study based on ancient DNA [7], a significant influence of Persian horse lines on European and Central Asian stocks is evident from the 7th–9th century onwards. In more recent periods, Arabian horses grew in popularity in the Western world from the 18th century onwards. European nobles and breeders purchased breeding animals in the Middle East directly from Arabian owners or traders and imported them to Europe with the intention to either improve their local herds or breed Arabians outside their region of origin. The popularity of Arabian horses throughout the 19th and into the first decades of the 20th century marks the 'Arabian wave' in European horse breeding. In this period, several Arabian horse lineages were established in many countries and today the Polish, Egyptian, French, Russian, British, and US Arabians are among the leading ones. Nowadays, Arabian horses are bred for several types around the world through 63 Arabian pedigree registries [29].

Whereas for centuries Arabian horses were bred in their native based on maternal strains, selective breeding outside the Middle East became focused on sire lines. Consequently, few tail-male lines are retained in Arabian populations today, with the foundation sires originating from different Bedouin tribes, strains, geographical regions, and time periods. Many of the Arabian sire lines are globally active today, and some lines have been re-introduced into their regions of origin [29].

Here we apply fine-scaled Y-chromosomal HT analysis in Arabians for the first time and present what this powerful lineage tracer informs about in terms of the origin, purity, genealogies, dispersal, and influence of Arabian stallions.

2. Materials and Methods

A glossary with the definitions for the terms used in the following sections is given in Appendix A Table A1.

2.1. Creating a Refined Y-Chromosomal Haplotype Phylogeny

2.1.1. Next Generation Sequencing (NGS) Data

- Whole Genome NGS Data

Whole genome NGS (WGS) data from 118 male horses (117 domestic horses and a single Przewalski's Horse) mapped to the LipY764 (GCA_002166905.2) Y chromosome

assembly were used from a previous study [22]. Among these samples, 114 belong to the crown haplogroup. An Icelandic Horse, a Shetland Pony, and a Przewalski's Horse were included as outgroups. Details about WGS data are given in Table S1.

- Target Enriched Sequencing Data

We performed Y-chromosome-target enriched sequencing (TES) of 39 males. The horses selected for TES included Arabians, Arabian-influenced breeds, Baroque Type breeds, and Coldbloods (Table S1). Genomic DNA was isolated from the Biosample given in Table S1 for each individual using DNeasy Blood and Tissue Kit from Qiagen, Vienna, Austria. For NGS library preparation and target enrichment, the Custom SureSelectXT low input target enrichment from Agilent, Vienna, Austria, was used. As previously described [22], the LipY764 assembly harbors 5.8 Mb of so-called single-copy Y (scY) regions that were shown to be suitable for unambiguous variant calling. Baits were generated which covered 5.063 Mb of scY regions of the LipY764 assembly, the mtDNA, and some autosomal loci. The 4032 Y-chromosomal bait segments that cover 5.06 Mb (86.7%) of the LipY764 scY regions are shown in Table S2. Indexed libraries were generated, and the enrichment was performed according to the protocol supplied by Agilent. Libraries were pooled and sequenced on two Illumina NextSeq550Medium PE150 runs (San Diego, CA, USA) (Table S1) at Vienna Biocenter Core Facilities, Vienna. Only reads mapping to the Lip764 assembly (see below) were analyzed in this study (Figure A1).

2.1.2. Data Analysis

- Variant Ascertainment

TES data were demultiplexed, adapters were removed with AdapterRemoval [30], and reads were trimmed with ReadTools [31]. Versions for programs used are provided in Appendix B Table A2. Adapter-free and trimmed reads were mapped to the LipY764 assembly (GCA_002166905.2) using bwa [32] with the parameters bwa aln −n 0.02 −l 200. Unmapped reads, PCR duplicates, and low-quality mappings (−q 20) were filtered out with samtools [33]. Variant calling was performed on the 39 TES mappings using GenomeAnalysisTK's [34]. First, gvcf files were produced by HaplotypeCaller, and then these files were merged with GenotypeGVCFs. Only the variants in the scY windows (these regions in LipY764 were defined in [22]) were considered for further analysis using bedtools [35]. We filtered out reference errors, sites called heterozygous only and variants with multiple alternatives. Variants with a read depth <3 and genotype quality <5 in more than 10% of the samples were excluded from further analyses using a custom python script (available from the authors on request). Furthermore, we just retained variants that were called in 75 % of the samples. We then merged the final variant list with the Y-chromosomal variant panel from [22]. All variants that were first ascertained in the TES data were visually checked in the IGV browser [36].

- MSY Haplotype Tree

The allelic state of 2276 MSY variants (2199 variants published [22] and 77 variants newly ascertained from the target enriched data) were genotyped in the 118 WGS sequenced and the 39 target-enriched samples sequenced using freebayes [37]. In total, 1639 MSY variants were polymorphic in our genotyping panel. In order to impute sites with missing calls in NGS data, which occurred either because of low coverage in WGS data or because the site is not covered in TES data, the samples were first clustered into haplogroups based on previously described haplogroup-determining variants [22]. Missing calls were manually added by introducing the allelic state observed in samples of the same group (strategy described in [38]). After gap filling, the states of the 1639 variants were concatenated to construct MSY HTs. Variant ascertainment and haplotype reconstruction strategy is outlined in Appendix B Figure A2. The Przewalski's Horse was used to infer the ancestral state for the variants that are polymorphic in the domestic horse. The topology and frequencies of the MSY HTs were visualized with the program Network [39]. For HT nomenclature, the first four letters were kept consistent with our previous studies [20,22],

followed by a successive letter/number according to human guidelines for MSY HTs [40]. The first four-letter code is deliberately not in alphabetical order but rather informs in which breeds/populations the HG/HT was first described. For details see Appendix C Table A3.

2.2. MSY Haplotyping

2.2.1. Creating the Backbone Structure

Out of the 290 crown variants, we defined 118 markers (110 Single Nucleotide Variants (SNVs), seven short insertions/deletions (Indels), and one short tandem repeat (STR)). The selected markers were checked visually in the IGV browser [36], by comparing the site in multiple samples in parallel. The markers are underlined in Figure 1 and listed in Table S3. These were catenated to generate a simplified crown HT tree (Figure S1, HTs in Table S4). This tree served as a backbone for MSY HT screening.

2.2.2. Haplotype Determination

For MSY haplotyping, genomic DNA from male horses was isolated using nexttec, Hilgertshausen, Germany. DNA Isolation Kits, using the biological material given in Table S5. To genotype 117 variants (SNVs and Indels), PCR-based KASP™ (Kompetitive Allele-Specific PCR) genotyping technology (KASP™, lgcgroup.com) was used. KASP assays were designed and KASP™ screening was performed on a CFX96 Touch™ Real-Time PCR Detection System using the standard protocol provided by the supplier (LGC, Berlin Germany). Cluster plots were analyzed using Bio-Rad CFX Manager 3.1 (Biorad, Vienna, Austria). Samples with known allelic states (DNA from sequenced individuals, and DNA from a Przewalski's Horse, a Shetland Pony, and an Icelandic Horse) were included as positive controls in each run. DNA from females and water controls were used as negative controls.

To genotype the amplicon length of the STR fBVB on an ABI 3130xl Genetic Analyzer, one primer was labeled with FAM fluorescent dye (fwd_FAM: ACAACCTAAGTGTCT-GTGAATGA; rev: CCCAATAATATTCCACTGCGTGT). PCR was performed on a 20 µL volume containing 0.4 µM of each primer. The DNA was initially denatured at 95 °C for 5 min, followed by 35 cycles of 30 s at 95 °C, 40 s at 58 °C annealing temperature, 40 s at 72 °C, and a final extension of 30 min at 72 °C. Alleles were sized relative to the internal size standard using GeneMarker®(Softgenetics, State College, USA).

Genotyping was performed in a consecutive manner. First, we confirmed that each sample belonged to the crown haplogroup by testing for the key variants rAY and rAX. Second, we clustered them into one of the major crown clades T, A, or H by testing for rA, rW, and fYR (see Figure 1). We then genotyped each sample for variants that determine the haplotype in the respective clade and imputed the ancestral allele in variants informative in the other clades. For the 38 individuals included in the MSY haplotyping approach that had NGS data available (Table S5), the allelic states of the 118 defined variants were extracted from the freebayes genotyping results (Table S6).

Haplotype networks were generated using the program Network [39]. The HT network was redrawn with the program draw.io. Not yet fully NGS-ascertained (most possibly internally branching off) HTs are marked with an asterisk and are outlined in the network with a dashed line branching off at the respective internal node (see Figure S1).

2.2.3. Datasets

- Globally Active Arabian Sire Lines

A dataset was assembled with sire lines that are presently active in the occidental Arabian horse population. Tail-male line evaluation from pedigrees of present-day breeding stallions formed the basis for sample collection. We first reconstructed the paternal genealogies of hundreds of horses using pedigree information provided by breeding associations, studbooks, or by reconciling results from multiple databases (listed in Table S5) and stored the tail-male line in a string format. Based on these analyses, we chose two to three

present male descendants of each sire active in the 1970s and collected a sample (biological material given in Table S5) for genotyping. We further included sire lines that evidently originated from imported Arabians but survived only in other breeds, like the Shagya Arabian, the Trakehner, and the Lipizzaner. With this strategy, we attempted to obtain the progeny of as many imported foundation sires identified and also represent their sublines, while minimizing oversampling of certain lines. Furthermore, we selected the samples from different breeding registries if a line was universally represented. However, for some lines, only a few samples or sometimes even a single sample were available. In total, our final sample set of 145 males comprised 81 registered Arabian Horses, 31 Shagya Arabians, 5 Partbred Arabians, 4 Fredriksborg Horses, 1 Kladruber, 2 Knabstrupper, 1 Lewitzer, 11 Lipizzaner, 1 Pintabian, 5 Trakehner, and 3 Warmbloods. Tail-male line reconstruction as a string format is outlined in Table S6, where the most recent ancestor given for the sampled horse was born no later than 1995, in order to protect the anonymity of the sample analyzed.

We need to state here that seven Arabian horses and four Akhal Teke were observed to carry the Tb-dW1 haplotype. This HT was previously attributed to a recent admixture with Thoroughbreds [20,22,27]. These horses were omitted from further analysis as the goal of this work was to focus solely on the Arabian Y-chromosomal lineages.

- Sampling Local Arabian Lines and Other Populations

A dataset of 35 Arabian horses, 8 from Iran and 27 from Syria, representing local lines was collected. The samples derive from three independent sampling projects, and pedigree information was available only for a few animals. If the pedigree information of a sample existed, it confirmed that the horse did not trace back to any globally active line. Furthermore, twelve Turkoman horses from Iran and 28 Akhal Teke horses from several countries were used. Finally, 24 Thoroughbreds and 15 Trotters, representing the three foundation sires of the Thoroughbred were included in the dataset. A full description of this dataset including sampling information is given in Table S5.

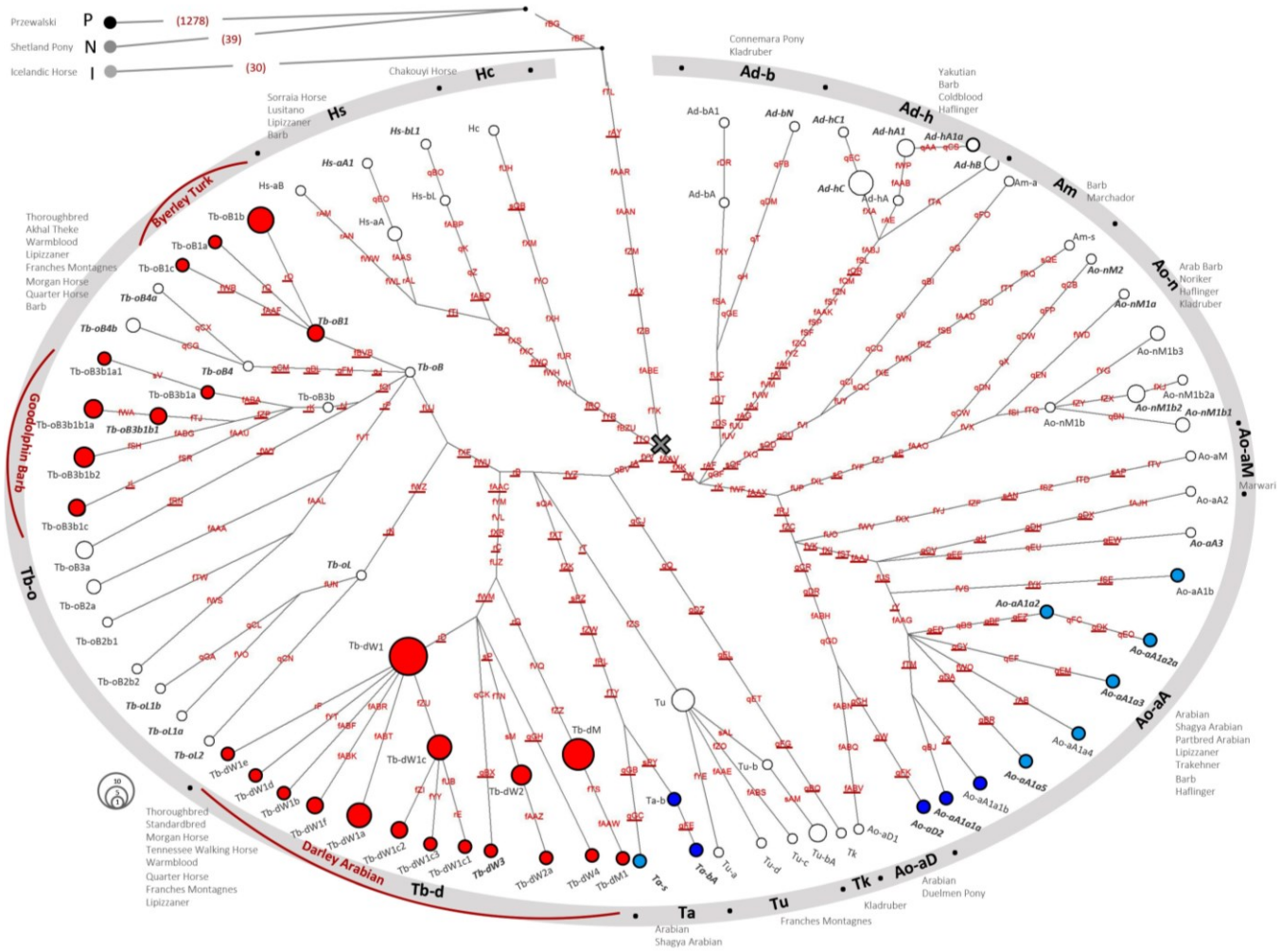


Figure 1. NGS HT Network. The MSY HT network from 118 WGS sequenced males and 39 TES males, based on a total of 1639 variants (281 crown, 1358 outside the crown). HTs are indicated as circles, with circle size being proportional to frequency. HT-IDs and sample information are provided in Table S1. HTs first described in this study based on TES data are shown in bold. Variants are indicated on branches and underlined when selected for genotyping. The position of the crown MRCA is marked with a cross. The 14 crown HGs are indicated in the outer circle, with the breeds listed beside them. Blue HTs were detected in Arabians, and light blue HTs were detected in a horse that traced back to an imported Arabian in the paternal lineage. The signatures of the three founders of the English Thoroughbred (Tb-oB1, Tb-oB31, and Tb-d) are marked with red lines.

3. Results and Discussion

3.1. A refinement of the Crown HT Structure from NGS Data

In order to perform MSY tracing of sire lines in modern breeds, one needs to distinguish among the very recently established crown HTs with informative markers. Therefore, we first generated a refined topology of the HTs within the crown HG based on NGS sequencing data. In total, 115 HTs from whole-genome sequenced males which were confirmed as crown haplogroup carriers in a previous study [22], were analyzed together with TES data from 39 males generated in the course of this study. The TES data covered the 5 Mb scY region of the MSY with a mean coverage of 41.95 × (Table S1 and Figure A1). While the whole genome sequenced sample collection from the previous study consisted mainly of Sport horses and Thoroughbred-influenced breeds, the horses selected for TES represented Arabians and Arabian-influenced breeds as well as some Baroque Type breeds (such as Lipizzaner, Kladruber, Friesian), Coldblood, Iberian breeds, and some Barb horses.

A Shetland Pony, an Icelandic Horse, and a Przewalski's Horse, previously determined as 'non-crown' HT carriers [22] were included as outgroups. The 157 males with NGS data analyzed in this study, including breed, sire line, and raw data information are listed in Table S1. The LipY764 shotgun assembly, which allows unambiguous MSY variant calling on a total length of 5.8 Mb scY regions, was used as reference. We first ascertained variants in the 39 TES samples and retained 177 variants after the filtering steps described in the Material and Methods section. Out of those, 18 did not pass the IGV quality check. From the 159 remaining variants, 82 (53.4 %) were previously described in the WGS dataset and 77 variants (67 SNVs and 10 Indels) were newly ascertained in this study (q-variants). The results from variant ascertainment in TES data in .vcf format are shown in Table S6.

We then combined the newly ascertained variants with the 2192 MSY specific variants ascertained in [22] and genotyped those in the full dataset (WGS and TES data; see Material and Methods and Figure A2). We detected 1639 variable sites in the 157 males, and those variants are listed in Table S6. The majority of variants (1278 (78%)) separated the Przewalski's Horse HT from the domestic horses. The two 'non-crown' domestic horse HTs in our analysis panel, N (the Shetland Pony) and I (the Icelandic Horse), had 39 and 30 private variants, respectively, and were separated from the crown lineages by nine shared variants. A total of 281 variants (259 SNVs, 20 Indels, 2 STRs) was detected among crown haplogroup carriers. As mentioned above, 77 variants (26.6%) were newly ascertained in this study (q-variants), whereas the remaining 212 crown variants (73.4%) were previously described (r-, s- and f-variants in [20–22,26]). The 281 crown variants distinguished 86 HTs. Reconstructed HTs are shown in Table S6. The sequential order of mutations on the nonrecombining MSY allowed a straightforward reconstruction of the HT phylogeny under the principle of parsimony. The resulting topology drawn as a network with variants given on branches is shown in Figure 1. The strictly hierarchic evolution of the MSY HTs is also reflected in the HT nomenclature (for details see Material and Methods, Table A1).

The refined phylogeny confirmed the polytomy of the crown HTs proposed previously [22] and substantiated the three major crown clades H, A, and T (Figure 1). The extended dataset further revealed that the three major clades clearly split into subclades. Hence, the crown HT phylogeny forms a pronounced star-like structure. This observation substantiates the previous hypothesis that the crown HTs are the signature of a remarkable expansion of a population harboring the basal HTs of the crown's sublines. Within the crown, we define 14 HGs. In Figure 1, we mark the MSY footprint of the three Thoroughbred lineages, which were defined previously [22] as HG Tb-d for Darley Arabian, subhaplogroup (sHG) Tb-oB3b1 for Godolphin Barb, and sHG Tb-oB1 for Byerley Turk.

The Coldblood horses in our dataset grouped primarily into HG Ad-h, while Iberian, Baroque Type, and North African horses, as well as British Ponies, were distributed across several HGs. The five Arabian horses (one each from the Saklawi, Bairactar, and Krzyzyk sire lines, two without pedigree information) clustered together with the seven horses from other breeds (Shagya Arabian, Partbred Arabian, Lipizzaner, Trakehner) that trace back to an Arabian foundation patriline (Siglavy senior (three), Siglavy Bagdady, Ibrahim, Ilderim, and Shagya) into HGs Ao-aA, Ao-aD, and Ta. Each of the twelve horses with Arabian ancestry carried a unique HT. Most Arabian HTs clustered into Ao-aA (eight), followed by Ta (three), and Ao-aD (one). The clustering is shown in Figure 1 and details about samples including breed, sire line information, and information on sequencing data are provided in Table S1. The refined crown topology is a prerequisite for informative sire line investigation in Arabians and other modern breeds. In Felkel et al. 2019 [22], the MRCA of the crown was dated to approximately 1500 years before present. Hence, to accumulate variation, only approximately 190 sequential generations are between the MRCA and present crown carriers when assuming a mean generation interval of eight years [41]. In view of this short evolutionary period, the revealed resolution of crown HTs based on biallelic markers is a remarkable achievement.

3.2. Allocation of the MSY Signature in Arabian Sire Lines

In previous studies, Arabian horses were roughly grouped into the crown HGs Ao, Ta, and some racing Arabians into Tb [20,22,27]. The HTs now detected in sequenced males of Arabian origin indicate three HGs (Ao-aA, Ao-aD, and Ta) as being characteristic of Arabian ancestry. However, this finding builds on only twelve sequenced individuals which represent only a few sire lines. Hence, we performed further investigation by genotyping HT-determining variants in a more comprehensive Arabian dataset. For this approach, we selected 118 variants (116 in the crown and the two crown-determining variants, rAY and rAX) to build a complexity-reduced 'crown-backbone'. In the variant selection, we ensured that the proposed Arabian lineages as well as the Thoroughbred HTs were fully traceable. For HGs that were detected only in breeds other than Arabians and Thoroughbreds (for example Ad-h, Ad-b, Am, Ao-n, Hs, and Hc), we chose fewer variants that would allow the inference of the major clusters but not a fine-scaled HT analysis yet. Selected variants are underlined in Figure 1 and listed in Table S3; the resulting backbone structure is given in Figure S1.

In order to delineate the MSY HT signatures of Arabian horses, we started with Arabian sire lines that became established in Europe based on imported Original Arabians. Many of these lines remain active on a global scale. For horses of these lines, high-quality pedigree information exists, which allowed a systematic representative sampling. Hence, a reconstruction of the tail-male line from the pedigree of registered horses was integrated into the sample selection process (see Material and Methods). Finally, the Arabian HT screening panel consisted of 145 males; 81 registered Arabians from leading breeding populations around the globe, and 64 registered horses from European breeds which trace back to an Arabian stallion imported into Europe in their paternal lineage. In total, the dataset (termed as 'occidental Arabian lines') comprised the legacy of 26 Arabian sires that were either imported to Europe from the Middle East during the 19th and 20th centuries or Egyptian foundation stallions. These 26 stallions originated from various Bedouin tribes, represented several strains, and covered more than a 100-year timespan. The first stallion transferred to Europe was Siglavy db, imported in 1814 to Austria, and the latest was Kuhailan Afas, imported in 1931 to Poland (Table 1, Full dataset Table S5).

Table 1. Y chromosome haplotypes in occidental Arabian horse lines. Full information about the 145 males sampled (81 registered Arabians and 67 from other breeds), including the tail-male line, is given in Table S5. See Figure 2 for genetic relationships among the haplotypes.

Foundation Sire [1]	Imported	Line Represented in Dataset via	Registered Arabians Sampled From [2]	Breed Sampled [3]	Y-Chromosomal Haplotype	Remarks
Koheilan Adjuze db Sbaa Anazeh	1885 Hungary	Piolun, 1934, Poland Jaszmak, 1928, Poland	Austria(1), Russia(5)	Trak(2) ShA(3)	Ao-aA1a* Ao-aA1a*	
Mahmoud Mirza db *Mukhalladiyah* Asad (Iraq)	India, Great Britain, Hungary	Jussuf I, 1962, Shagya Arabian		ShA(2)	Ao-aA1a*	
Old Jellabi Speckled (Bahrain)	Bahrain Foundation horse	Dhahmaan Alawwal, 1938, Bahrain	Austria(2), Poland(1)		Ao-aA1a*	
Zarif *Muniqi* Shammar Bedouins	1840 Germany	Rex II 372, 1941, Fredriksborg Horse Hermolin, 1937, Knabstrupper		FH(4) Ks(2)	Ao-aA1a* Ao-aA1a*	
Zobeyni db *Seglawi Jedran Ibn Sbeini* F'daan	Egyptian Foundation horse	Mahruss II, 1893, Egypt	Germany(1)		Ao-aA1a*	
Ilderim db *Saqlawi Jidran*	1900 Poland	Aquinor, 1951, Poland Maharadscha, 1957, Trakehner Doktryner, 1950, Poland	Poland(3) USA(1)	Trak(1), Wb(2) Le(1)	Ao-aA1a* Ao-aA1a4 Ta-b	subline-HT incongruence [4]
Saklawi I *Saklawi Jidran Ibn Sudan*	Egyptian Foundation horse	Ansata Ibn Halima, 1958, Egypt Aswan, 1958, Egypt Galal, 1959, Egypt Habdan Enzahi, 1952, Egypt Morafic, 1952, Egypt	Egypt(1), Qatar(3), Poland(2), Syria(1) Iran(1), Poland(1), Qatar(1), Russia(1) Egypt(1), Germany(1) Austria(1), Egypt(3), Iran(2), Poland(5), Qatar(4)	ShA(1) ShA(1) ShA(2)	Ao-aA1a1 Ao-aA1a1 Ao-aA1a1 Ao-aA1a1 Ao-aA1a1	
El Deree db *Saqlawi Shaifi* Baqqara (Syria/Iraq)	Egyptian Foundation horse	Akhtal, 1968, Egypt	Qatar(2)		Ao-aA1a1	

Table 1. *Cont.*

Foundation Sire [1]	Imported	Line Represented in Dataset via	Registered Arabians Sampled From [2]	Breed Sampled [3]	Y-Chromosomal Haplotype	Remarks
Siglavy db ‚Schwarzenberg' *Saklawi*	1814 Austria	Siglavy Monterosa, 1907, Lipizzaner		Lip(7), Kl(1)	Ao-aA1a2	
		Siglavy Capriola III, 1940, Lipizzaner		Lip(4)	Ao-aA1a2a	subline-HT
		21 Siglavy II, 1909, Shagya Arabian		ShA(1), Trak(2)	Ao-aA1b	private HT
Ibrahim db *Saklawi Faliti* Banu Sakhr	1907 Poland	Negatiw, 1945, Russia	Austria(1), Iran(2), Poland(2), Russia(2)		Ao-aA1a3	
		Ferseyn, 1937, USA	United Arab Emirates(1)	PA(1), Pi(1)	Ao-aA1a3	
Siglavy Bagdady db *Saklawi* Ruala	1902 Hungary	Siglavy Bagdady VI, 1949, Babolna	Germany(1)	ShA(1), PA(1)	Ao-aA1a5	
Dahoman db *Dahman* Djelas Anazeh	1852 Hungary	Dahoman XVI, 1904, Shagya Arabian		ShA(1)	Ao-aA3	
Jamil El Kebir db *Saklawi Jidran* F'daan	Egyptian Foundation horse	Anter, 1946, Egypt	Germany(1)		Ao-aD2	
Krzyzyk db	1876 Poland	Enwer Bey, 1923, Poland	Poland(2)		Ao-aD2	
Mirage db *Seglavi Jedran Dalia* Sbaa (Iraq)	1923 Great Britain	Bey Shah, 1976, USA	Qatar(1), Poland(2)		Ao-aD2	
Kuhailan Haifi db *Koheilan Haifi* Ruala	1931 Poland	Bask, 1956, Poland	Poland(2), Qatar(1)	PA(1)	Ao-aD2	
		Celebes, 1949, Poland	Austria(1), Iran(1), Poland(1), United Arab Emirates(1)		Ao-aD2	
		Wielki Szlem, 1938, POL		WB(1)	Ao-aD2	
Hadban db *Hadban Inzihi* S'baa	1897 Hungary	Habdan XI, 1954, Shagya Arabian		ShA(1)	Ao-aD2	

Table 1. *Cont.*

Foundation Sire [1]	Imported	Line Represented in Dataset via	Registered Arabians Sampled From [2]	Breed Sampled [3]	Y-Chromosomal Haplotype	Remarks
Gazlan db *Kuhaylan Tamri* Would Ali Anazeh (Iraq)	1852 Lipizza	Gazal VII, 1944, Shagya Arabian		ShA(6)	Ta*	
O'Bajan db *Ma'anaqi Sbaili* S'baa Anazeh	1885 Hungary	O'Bajan VII-4 530, 1936, Shagya Arabian O'Bajan X, 1929, Shagya Arabian		ShA(2) ShA(2)	Ta* Ta*	
Shagya db Beni Saher	1836 Hungary	Shagya IV, 1875, Shagya Arabian Shagya VII, 1877, Shagya Arabian Shagya XI, 1886, Shagya Arabian		ShA(3) ShA(2) ShA(3)	Ta-s Ta-s Ta-s	
Mersuch db *Hamdani Semri* Hazaim Pasha (Iraq)	1902 Hungary	Mersuch IV, 1936, Shagya Arabian		ShA(1)	Ta-b	
Bairactar db *Saklawi Jidran* (Syria)	1817 Germany	Arax, 1952, Poland Gwarny,1952	Poland(1), Qatar(1), Russia(1) Germany(2), Poland(1)		Ta-b Ta-bA	subline-HT
Dahman Amir db *Dahman Amir* Sultan Abdulhamid II (Turkey)	1902 Poland	Saludo, 1954, Spain	Qatar(1)	PA(1)	Ta-b	
Seanderich db *Saklawi Jidran* Shammar	1908 Spain	Tabal, 1952, Spain	Germany(1)		T2*	
Kuhailan Afas db Koheilan Afas (Bahrain)	1931 Poland	Comet, 1953, Poland	Germany(1), Poland(1)		Tb-oB1*	
Latif db *Hamdani Semri* Anazeh Fedan	1909 France	Baroud II, 1927, France Kann, 1927, France	Qatar(3), Russia(1) Russia(1)		Tb-oB1* Tb-oB1*	

[1] Name of Foundation Sire, *Strain*, Bedouin tribe, or breeder (country). [2] The number of horses sampled is given in parenthesis. [3] The number of horses sampled is given in parenthesis. Abbreviations: Trakehner (Trak), Shagya Arabian (ShA), Fredriksborg Horse (FH), Knabstrupper (Ks), Warmblood (Wb), Lewitzer (Le), Lipizzaner (Lip), Kladruber (Kl),Partbred Arabian (PA), Pintabian (Pi). [4] Incongruence between tail-male line documentation and HT.

We performed MSY haplotyping as described in the Material and Methods sections and first confirmed that all selected MSY variants were informative (details on typing performance are provided in Table S3). Among the 145 males descending from 26 foundation sires, we detected 16 HTs, and all belonged to the crown haplogroup. Results are shown in Figure 2 and Table 1 (full information about the samples is provided in Table S5). We need to point to a limitation of the genotyping approach. Due to ascertainment bias, we might miss the markers determining potentially private HTs in not yet sequenced lines. As the HTs of those lines are not yet fully resolved, they were placed onto internal branching points of the backbone topology. We indicated those HTs with an asterisk (*) in their HT identifier. For example, the horses carrying HT Ta* clustered definitively into the branch Ta (they carried the derived allele for markers sPZ, fTY, fRL, fXT, fZK, and fZW), but they did not group with Ta-s or Ta-b (they carried the ancestral allele for marker qGB, qGC, and sPY). Sequencing Ta* carriers is needed in order to resolve their topology. A description of the HTs is provided in Table S4 and in the Appendix C Table A3.

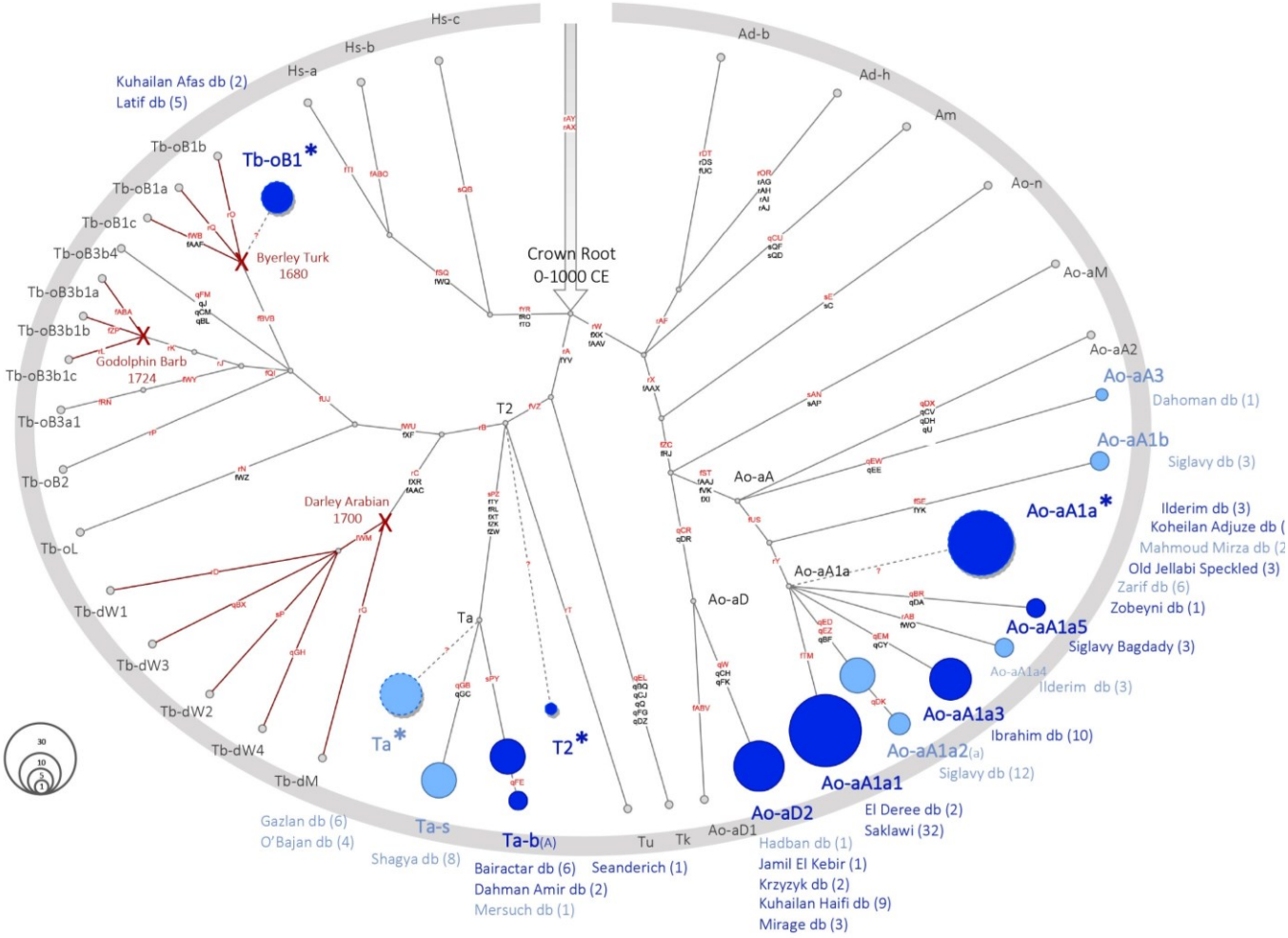


Figure 2. MSY HTs in Arabian sire lines. Simplified crown HT network based on 118 variants. Genotyping results from 145 males are shown in blue circles with size proportional to frequency. Details on samples are given in Table 1, and Table S5 (samples are indicated in Column J; Foundation sires shown in Column AI). Foundation sires are shown for each HT, with the number of samples for each line in parenthesis. HTs/foundation sires that are only active in breeds other than Arabians are given in light blue. Genealogies of the English Thoroughbred founders are outlined with red crosses and the thoroughbred specific subhaplogroups with the branches in red.

Overall, most samples (137 out of 145) were grouped into one of the three HGs anticipated as Arabian by the NGS sequenced samples—Ao-aA, Ta, and sHG Ao-aD2. All three HGs were detected in sire lines active in today's registered Arabians, as well as in lines that survived only in Shagya Arabians or other Arabian influenced breeds (like the Lipizzaner, Kladruber, or Trakehner). Ao-aA was the most common HG seen in 92 samples from 12 sire lines. This HG also had the highest number of accurately defined HTs (9). Among the Ao-aA carriers, most (88) clustered into sHG Ao-aA1a, which determines this sHG as the most pronounced Arabian signature. Six sire lines were placed on the basal *HT Ao-aA1a*, which suggests that we still underestimate HT diversity in this clade.

In addition to the affirmation of the three HGs suggested by the NGS-sequenced horses, genotyping revealed alternative clustering in three lines. A single horse from a Spanish line (foundation sire Seanderich) was grouped to the internal branching point T2*, which indicates a private, not yet resolved, HT in this line. It is worth noting that, in two sire lines (after Kuhailan Afas via Probat, 1975 and Latif via Denouste, 1921; see Table S5), we detected Tb-oB1 (Figure 1 and Table 1). The Tb HG was previously attributed to Turkoman horses [20] and genealogical reconstruction revealed that Tb-oB1 was carried by Byerley Turk in 1680 [22]. In the lines after Byerley Turk, we could only ascertain subline-HTs for lines active in warmbloods (Tb-ob1a-c; [22]). The Byerley Turk Thoroughbred after Djebel, 1937, implemented in the TES set (Y_PR_11_033, Table S1) carried the unaltered foundation HT Tb-oB1. Hence, for the Byerley Turk line, we could not define a long-standing informative lineage-specific marker like, for example, Tb-dW1, which is a subline HT and a unique signature of the Thoroughbred lineage after Whalebone (1807). Considering the early emergence of Tb-oB1 (it was already carried by Byerley Turk in 1680), we expect to underestimate the HT diversity in this sHG in other breeds, due to ascertainment bias. Hence, Tb-oB1* should be the most appropriate term for this sHG at the moment. From the current state of knowledge, we cannot infer explicitly whether 'Kuhailan Afas and Latif Arabians' carry Tb-oB1* because of an undocumented Thoroughbred influence in their tail-male line or through another scenario. Both sire lines are well-known for success in flat racing, and the interbreeding of Thoroughbreds in racing Arabians has been shown recently [27]. In particular, a successful 'Arabian' racing stallion imported from Syria to Egypt was fathered by the Thoroughbred Temeraire, 1905 from the Byerley Turk line born in Ireland [42]. However, Tb-oB1* in Kuhailan Afas and/or Latif lines could also be an autochthonous Arabian HT. Among the strain types bred by the Bedouins (Kuhaylan, Saglawi, Abayyan, etc.), the Muniqi strain was the racing type [42–44]. Muniqi Arabians were phenotypically more similar to Thoroughbreds, and Darley Arabian was described as a Muniqi type Arabian [42]. Hence, Tb-oB1* could have been segregated in Muniqi Arabians without Thoroughbred input and was introduced with horses imported for flat-racing [45].

3.3. Horse MSY Haplotyping in Arabian Horses on the Genealogical Scale

In humans, Y-chromosomal haplotyping is a valued system for combining genetics with ancestry information [17]. Personalized characterization of sire lines can allow the tracing of the breeding influence of particular patrilines, or to check for correctness of pedigree records over several past generations. In Felkel et al. 2019 [22], we genetically redrew the three Thoroughbred sire lines. The power of MSY-analysis to detect wrong assignments in the pedigree that occurred multiple generations back in time was exemplified by addressing a historic dispute concerning the wrong paternity assignment of the Thoroughbred stallion Galopin born in 1872.

In the occidental Arabian collection, paternal line information was available for all 145 individuals (information as a string in Table S5). For 20 foundation sires, we had more than one descendant analyzed. In most lines, MSY clustering followed the expectation, and the members of a particular sire line clustered into the same HT/*HT. Some HTs (Ao-aA1a2, Ao-aA1b, Ao-aA1a3, Ao-aA1a4, Ao-aA1a5, Ao-aA3, and Ta-s) were detected only in a single line, whereas others, in particular the not fully resolved *HTs, harbored several lines. Worth mentioning is HT Ao-aA1a1, which was almost private for the seminal Egyptian

Saklawi line, after Nazeer, 1934. Apart from the Saklawi line, Ao-aA1a1 HT was only found in two males after El Deree, another Egyptian foundation horse.

We further revealed that four variants (qDK, qFE, rAB, and fWO) occurred recently within the timeframe of pedigree documentation. These mutations define three de novo subline-HTs (Ta-bA, Ao-aA1a2a, and Ao-aA1a4), which are unique for sublines of certain sire lines (details about HTs and lines are provided in Table 1).

In some genealogies, the MSY pattern did not fully agree with the recorded paternal lineage. We observed two very distinct HTs in the sire line after Siglavy db (Ao-aA1a2 and Ao-aA1b) and also in the lines after Ilderim db (Ao-aA1a and Ta-b). Those findings cannot be explained by de novo mutations but rather are a clear sign that more than one stallion contributed to those lines.

As shown in Figure 3, we detected two distinct Arabian HTs in horses tracing back to the imported Arabian stallion Siglavy Senior ('Schwarzenberg'), 1810. The Lipizzaner and Kladruber lines after Siglavy Slavina III, 1893 carried Ao-aA1a2 and the subline-HT Ao-aA1a2a, whereas Shagya Arabians and Trakehner after 21 Siglavy II-2, 1909 grouped distantly into HT Ao-aA1b (Figure 3). Both HTs were not detected in any lines other than the Siglavys (Figure 1). Hence, our findings indicate that the Siglavy sire lines in Lipizzaner and Shagya Arabians were founded by paternally unrelated horses. Confusion may have arisen as a result of homonymous names. A second Siglavy line, founded by the Arabian stallion Siglavy IV db, 1819 (imported 1825), was active in parallel to the Siglavy line after Siglavy Senior ('Schwarzenberg'), 1810 in the studs of the Habsburg Monarchy [46].

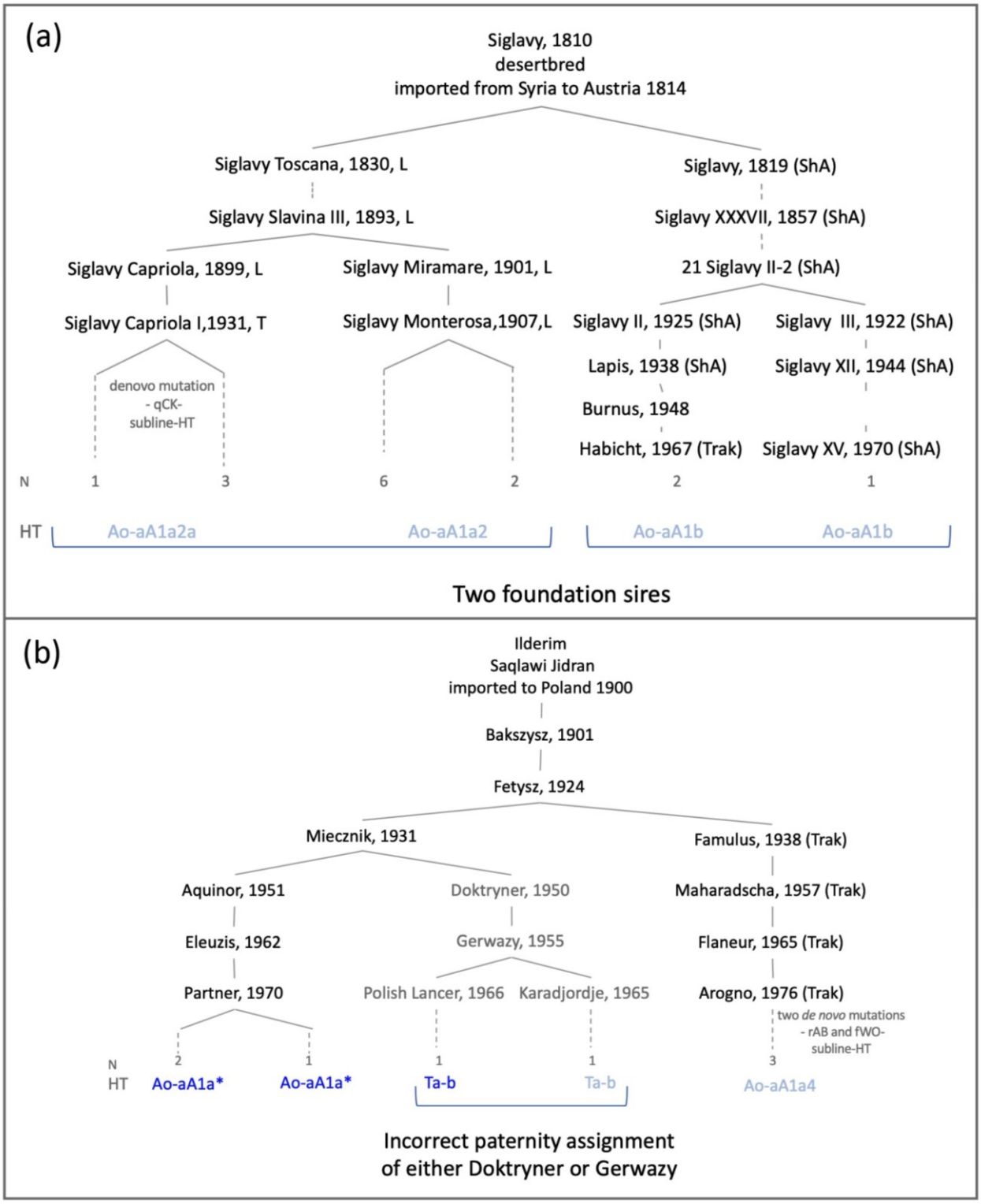


Figure 3. Genealogical cases. (**a**) Paternal genealogies of 15 genotyped male horses after Siglavy, 1810. (**b**) Paternal genealogies of eight genotyped male horses after Ilderim db. Dotted lines indicate that at least one generation is omitted. Abbreviation of horse breeds other than Arabian is given by: L = Lipizzaner, ShA = Shagya Arabian, Trak = Trakehner, AA = Anglo Arabian. The number of genotyped horses and HTs is listed on the bottom (dark HTs were detected in Arabians, light blue HTs in other breeds). The complete tail-male line reconstruction is provided in Table S5.

The inconsistency in the lines after Ilderim db suggested a different scenario. As shown in Figure 3, the HT detected in the progeny of the Polish Arabian stallion Doktryner, 1950 via his son Gerwazy, 1955 was Ta-b. This did not concur with the HTs detected in other descendants of Doktryner's grandfather Fetysz, 1924 via Aquinor, 1951 and Maharadscha,1957 in Arabians (Ao-aA1a) and the Trakehner (subline-HT Ao-aA1a4). A recent inaccurate paternity assignment of either Doktryner or Gerwazy is the most probable explanation.

HT Ta-b has so far been detected only in two Arabian sire lines bred in Poland, Bairactar and Dahman Amir (Table 1). Among the Arabian stallions active in 1949 were the Klemensow Stud, where Doktryner was bred, and the three Arabian stallions used in 1954 at Michalow stud, where Gerwazy was bred; the only member from a Ta-b sire line was Amurath Sahib (1932) after Bairactar db [47]. MSY HTs provide an indication that Amurath Sahib, active at Klemensow Stud, was the biological father of Doktryner.

As exemplified here, MSY lineage tracing revealed genealogical insights about Arabian sire lines. However, the explanatory power depends on an accurate assignment of line-specific HTs. The HT structure of several Arabian lines is not yet sufficiently resolved, which impedes the concise delineation of their paternal genealogies. Ascertainment of more informative markers by sequencing the MSY in an extended set of samples will push MSY analysis towards a forensic tool for redrawing paternal lineages of horses in the future. Moreover, implementing markers with divergent mutation rates [48], for example, adding MSY STR markers, should allow a better discrimination of recently emerged lineages and answer questions about relatedness levels that differ in time depth.

3.4. On the History of Arabian Horse Breeding beyond Pedigrees

The combination of MSY topology and pedigrees could enlighten the ancestry and relationship among foundation sires beyond pedigree documentation. From the occidental lineages, we considered Ao-aA, Ao-aD2, and Ta as typical Arabian and Tb-oB1* and T2* as tentative Arabian. However, those populations went through a severe male bottleneck, as our panel originates from only 26 imported stallions. According to several studies, registered Arabian horses in the Middle East show expanded genetic and phenotypic diversity in comparison to the global Arabian bloodlines [27,49–51]. The existence of genetically diverse populations of Arabian horses in the Middle East today could also be reflected in more diverse MSY patterns.

Therefore, we extended our Arabian sampling towards local Middle Eastern Arabian populations and genotyped 35 Arabian males, 8 from Iran, and 27 from Syria, with no recently reintroduced occidental lines included. As shown in Figure 4 (full data in Table S5), the HGs detected in Middle Eastern Arabians overlap with the HGs observed in the occidental lines.

Autochthonous Middle Eastern populations clearly substantiated the Arabian MSY signature (Ao-aA, Ao-aD2, Ta, and, presumably, Tb-oB1*). Mainly basal *HTs were observed in the Middle Eastern samples, and we expect them to carry new HTs not yet ascertained via NGS. We did not detect any of the recently established HTs in Middle Eastern populations, a finding that was expected, as those HTs were ascertained in occidental lines through NGS and, accordingly, are specific markers for those lines. The similar MSY HGs in occidental and Middle Eastern Arabians support the hypothesis of a discrete shared origin with recent divergence, a scenario that was proposed from microsatellite analysis [51,52].

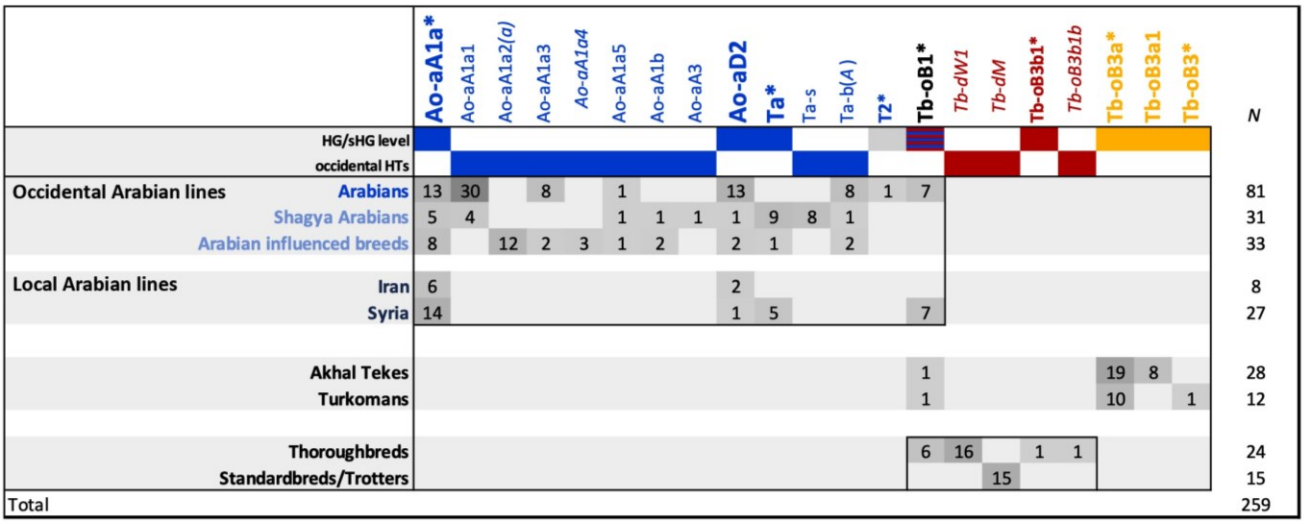


Figure 4. MSY Haplotypes in globally active Arabian lines, Middle Eastern Arabians, and other breeds. Haplogroup (bold) and haplotype distribution in breed or breed groups in absolute numbers (*N* = total number). Thoroughbred HG/HTs are marked in red, Arabian in blue, and Akhal Teke/Turkoman in yellow. The Tb-oB1* subhaplogroup was detected in Thoroughbreds, Akhal Teke/Turkoman, and in a small subset of Arabians, but the ancestry of horses carrying this haplogroup remains unresolved.

The grouping of the 28 Akhal Teke and 12 Turkoman males highlights their clear distinction from the Arabians. Most of the samples from those breeds were grouped distantly from the Arabians and into sHG Tb-oB3. This sHG seems, so far, to be private for Akhal Teke and Turkomans, and it shares a more recent ancestry with the Thoroughbred than all Arabian HGs. However, as shown in Figure 4, sHG Tb-oB1*, which was found in two occidental lines, was detected in all comparison datasets. The prominent occurrence of Tb-oB1*, ranging from globally active lines and Middle Eastern Arabian lines from Syria to Akhal Tekes, Turkomans, and Thoroughbreds fuels reflection about scenarios that could have led to this widespread distribution. Together with the other Arabian HTs (Ao-aA, Ao-aD, and Ta), Tb-oB1* could have been typical of autochthonous Arabian horses from the Nejd highlands. Selection and genetic drift could have led to haplotype frequency differences among different strains and areas. In this scenario, Tb-oB1* may have been distributed from the Arabian plateau during the migrations of the Bedouins. Alternatively, Tb-oB1* could also be of Turkoman origin. The influence of Turkoman horses on Arabian horses, albeit rather undesirable from the standpoint of Arabian horse breeders, may have occurred during the Ottoman Empire (from the 15th to the beginning of the 20th century) and the Wahhabi Wars (during the beginning of the 19th century), through the use of units mounted mainly with stallions [53]. Such admixture may also have happened earlier, at the time of the Crusades, when Turkish soldiers came to Syria together with their horses [54]. In this context, Nissen [55] pointed to a combination of Jilfan mares and Turkoman stallions, from which the Muniqi type might have developed. Admixture with Turkoman horses was also postulated for Muniqi-type Arabians by Carl Raswan [42]. Considering the wide distribution of Tb-oB1*, the least likely, albeit possible, scenario is Byerley Turk bloodlines were recently introduced into the Arabian and other horse bloodlines in the Middle East. With the current state of knowledge, we cannot confirm or reject one or a combination of several scenarios.

Further investigation, based on extended sampling and more sensitive Y-chromosomal markers, is needed to fully resolve the direction of dispersal and influences that happened in the past. We must also keep in mind that the HTs seen in Arabians today are remnants that survived selection through human breeding decisions. Implementation of genetic information from historic remains is now possible [56], and such data will contribute to fully

disentangling the origin of present Arabian lineages, resolving past admixture between and among populations, and uncovering the lineages that have been lost.

The MSY pattern in Arabian horses showed definite Arabian-specific sHGs, which point to linebreeding with selection on males beyond the documented period. On the other hand, the maternal lineages, explained by the mtDNA phylogenetic pattern, have indicated a high maternal diversity of the Arabian breed [28,50,57]. Therefore, the current Arabian horse breed is another example of differing variation of both maternal and paternal history in horses [58]. From a historical and traditional breeding point of view, the matrilineal side was and is still considered the major criterion to determine the purity of the Arabian horses. This work clearly shows that the paternal side is equally important, but a combined systematic analysis of both Y HTs and mtDNA is needed to determine the contrasting patterns of genetic variation in the two genetic compartments in the Arabian horse populations. Given the long history of maternal-oriented breeding practices in the places of breed origin and the small effective population size in the rest of the world, a more detailed comparison of mtDNA and Y HTs would be quite instructional.

4. Conclusions

In this study, we corroborated MSY haplotyping as a meaningful method for addressing population genetic, genealogical, and forensic questions in the recently established modern horse lineages. We present the next milestone for tracing stallion-mediated improvement and fine-scaled sire-line characterization. The determination of the Arabian and Thoroughbred MSY HT signatures now allows the tracing of their recent impact in any breed. However, further investigation is needed, and this should include enlarging the ascertainment panel, determining HT frequency distributions on a broader scale, and implementing faster-evolving markers in the analysis. In the future, MSY lineage determination can substantiate horse breeding in a way that contributes to the understanding of the historic development of breeds, supports decision-making in breed conservation, and can be used to validate the paternal ancestry in pedigree records.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes13020229/s1, Table S1: NGS Data information and MSY HTs; Table S2: Target enrichment bait regions on LipY764; Table S3: Information for haplotyping; Table S4: MSY Haplotypes on the genotyping scale; Table S5: Database with information on samples used for haplotyping; and Table S6: NGS results table. Figure S1: Backbone tree with HT-determining variants on branches.

Author Contributions: Conceptualization of the project was completed by V.R., B.W., and G.B. Funding acquisition was performed by B.W. and G.B. Methodology was determined by V.R., E.B., S.F., and B.W. Data collection was performed by D.R., E.B., S.F., V.R., L.R., G.W., M.S.-S., R.J., L.K., S.R., M.C.P., D.C.M., and B.W. Resources were provided by B.W., V.R., M.S.-S., M.B.-P., A.M.K., G.C., R.J., G.A.B., S.A., M.R., V.V.K., A.M.Z., L.K., S.R., M.C.P., S.B., R.S., D.F.A., and D.C.M. Data analysis was performed by B.W., E.B., S.F., and V.R. Supervision of the project was performed by B.W. Manuscript was prepared by V.R., B.W., G.G.-S., and G.W. All authors contributed to the final editing and review of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Innovation Fund of the Austrian Academy of Sciences (ÖAW), grant number IF_2015_17; the Austrian Federal Ministry of Agriculture, Forestry, Environment and Water Management DAFNE, grant number 101184; and with partial funding from the National Priorities Research Program (NPRP), grant 6-1303-4-023 from the Qatar National Research Fund (a member of Qatar Foundation). E.B. was funded by the Austrian Science Funds (FWF, W1225). Samples provided by the All-Russian Research Institute for Horse Breeding were obtained with the support of the Russian Science Foundation, Project Number 19-76-20058. D.F.A. is an Investigator of the Dorothy Russell Havemeyer Foundation, Inc.

Institutional Review Board Statement: This study was discussed and approved by the Institutional Ethics and Welfare Committee of the University of Veterinary Medicine Vienna in accordance with GSP guidelines and national legislation (ETK-10/05/2016). This research was performed in accordance with relevant guidelines reported in the above-mentioned document. Samples were (i) either hair root samples provided by private horse owners, and these were taken in compliance with the

animal welfare standards in the respective country, or (ii) the material consisted of retained samples from paternity testing (DNA, hair, blood, or sperm), for which we received permission to use from breeding associations. Pedigree information about samples was provided upon availability and informed consent was obtained from horse owners. The biosample and pedigree responsibilities of authors are given for each sample in Table S1 for NGS sequencing samples and Table S5 for the samples used for genotyping. All samples of this project are coded.

Informed Consent Statement: Not applicable.

Data Availability Statement: NGS reads mapped to the LipY764 assembly of all samples analyzed in this paper are available at SRA archive PRJNA430351 (WGS data) and PRJNA787432 (TES data).

Acknowledgments: We thank Astrid Velsen von Zerweck, the Arge Noriker, the Arge Haflinger, the Bundesgestüt Piber, Ines Butler-Wemken, and Suzi Pritchard Jones for their support to obtain access to sample material; and Eva Michaelis, Regina Astl, Simone Reiter, Maria Louise Mühlberger, Christina Hartsleben, Jennifer Zahn, and Judith Strele for their assistance in the lab. Gudrun Waiditschka carried out literature research as part of an unpaid project work at the Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna. Thanks to Gabriella Lindgren for comments on a previous version of the manuscript. Illumina sequencing was completed by the Vienna Biocenter Core Facilities (VBCF) NGS Unit.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results. The statements made herein are solely the responsibility of the authors.

Appendix A

Table A1. Glossary.

Allele	Variant or Alternative form of the DNA Sequence at a Given Locus
clade	a branch on a phylogenetic tree that is formed from individuals of common descent; synonymous to 'monophyletic group'
crown Haplogroup	very recently expanded horse MSY HG, predominant in modern breeds
founder/foundation sire	in this article this term is used to mean the earliest recorded paternal ancestor of an established sire line
genetic marker	a genetic marker is a DNA sequence with a known physical location on a chromosome that is polymorphic and thus informative for differentiating individuals
genotype	alleles possessed by an individual at one locus
haplotype (HT)	a set of DNA polymorphisms that are inherited together due to linkage; a Y-chromosomal haplotype is characterized by the allelic state at several markers
haplotyping	the laboratory process of determining the haplotype of an individual via the genotyping of appropriate markers
haplogroup (HG)	a monophyletic group of MSY HTs
indel	an insertion or deletion of bases in the genome that occurs at a specific genomic position
key variant	in this article, this term entitles the marker selected for haplotyping among markers with tautological readout
locus	a location in the genome; references to any sequence or genomic region, including non-coding regions
mitochondrial DNA (mtDNA)	circular, double-stranded DNA located in the matrix of a mitochondrion; the mtDNA is inherited uniparentally from mother to offspring
modern breed	in this article, this term is used for breeds that were developed or created within the westernized industry of horse breeding, spanning the recent 200 years; here it applies to Arabians, Thoroughbreds, Warmbloods, Central European and British Coldbloods, Central European and British Ponies, Baroque Type, Iberian, and New World breeds

Table A1. *Cont.*

Allele	Variant or Alternative form of the DNA Sequence at a Given Locus
most recent common ancestor (MRCA)	the sequence status from which all HT variations in a clade descend
male-specific region of the Y chromosome (MSY)	the non-recombining region of the Y chromosome which is inherited as a single linkage group from father to sons
patriline	the line of descent traced through the paternal side of the pedigree; synonymous with 'male-tail line'
parsimony	a principle when drawing phylogenetic trees where the most likely tree is the one with the fewest evolutionary changes
pedigree	the record of descent of a horse
phylogeny	a branching tree showing the evolutionary relationships among a set of DNA sequences
sequence assembly	recreation of the original genome from the sequenced reads
short tandem repeats (STRs)	a tract of repetitive DNA in which short DNA motifs (ranging in length from one to six or more base pairs) are repeated; synonymous to 'microsatellites'
single copy Y (scY) regions	well-explored regions of the MSY that are screenable with short-read data
single nucleotide variant (SNVs)	a variant of a single nucleotide that occurs at a specific genomic position
sire line	members of a sire line descend from the same foundation sire in their patriline (male-tail line); hence the breeding influence of a foundation sire is apparent from the sire line distribution range
subline HT	those HTs which emerged recently, within the pedigree supported timeframe, from a new mutation and are therefore unique markers for a particular stallion
target enriched sequencing	NGS sequencing of a specific (desired) regions of the genome
variant calling	identification of variation, mostly SNVs and Indels, present in sequence data by comparing NGS data from individuals with a reference sequence
Y chromosome	one of the two sex chromosomes; determines male sex in horses

Appendix B

Explanatory information for creating a refined Y-chromosomal haplotype phylogeny

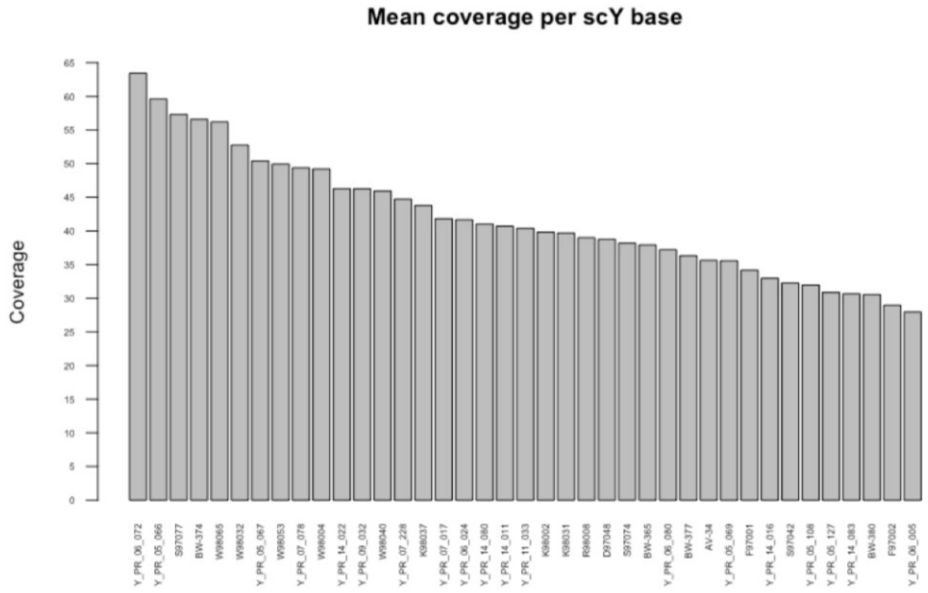


Figure A1. Target-enriched sequencing depth.

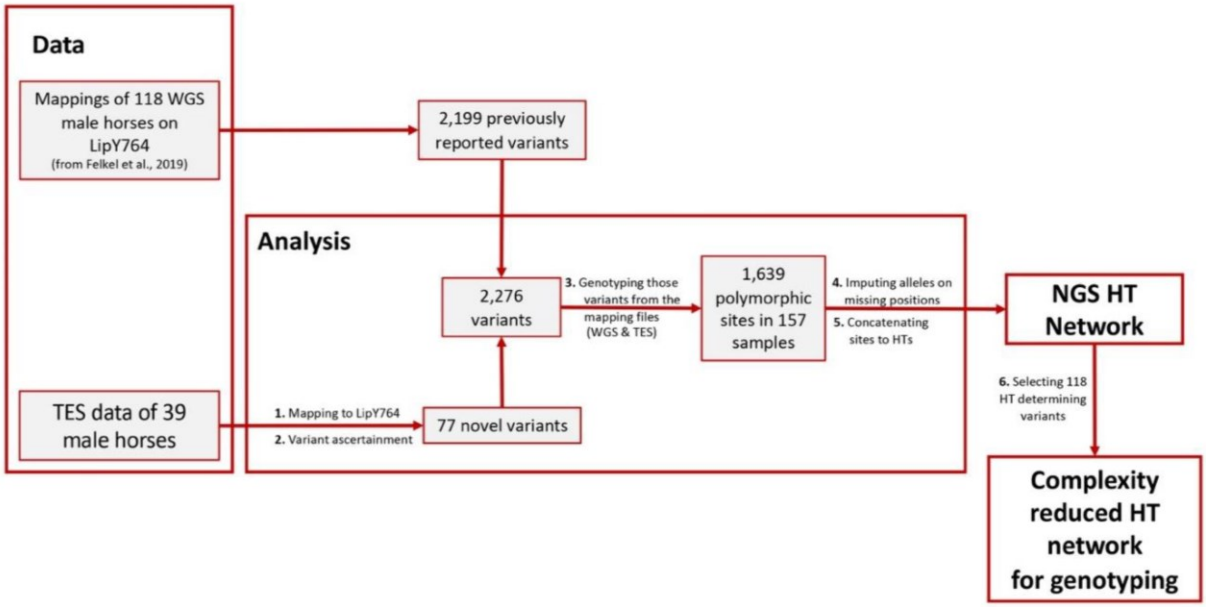


Figure A2. NGS data analysis pipeline. Two NGS data sets were merged to produce the final structure: mappings of 118 WGS males to LipY764 from a previous study and newly generated TES data of 39 males. The 77 novel variants ascertained in the TES dataset were merged with 2199 previously defined variants resulting in a total of 2267 variants. Those variants were genotyped in the mapping files, missing positions imputed, and HTs constructed by concatenating the polymorphic sites. HTs were visualized in a network format. A total of 118 variants was selected for genotyping. Detailed descriptions of the workflow, including programs, and parameters, is provided in the Material and Methods section.

Table A2. Versions of the NGS tools used in data analysis.

Tool	Reference	Version
AdapterRemoval	[30]	2.3.1
ReadTools	[31]	0.2.1.r_716422a3
bwa	[32]	0.7.17
samtools	[33]	1.10
GenomeAnalysisTK	[34]	3.7
bedtools	[35]	2.27.1
IGV	[36]	2.5.3
freebayes	[37]	1.3.2-46-g2c1e395
Network	[39]	10.2

Appendix C

Table A3. Explanation of HG/HT nomenclature. HG/HT-determining variants are provided in Figure 2 and Table S3.

Major Crown Clades (n = 3—A, H, and T)	Haplogroups	Subhaplogroups and Haplotypes Detectable with Genotyping	Comments Regarding Nomenclature
Clade A was first described in an Arabian horse 'A'			
	Ad-b		A draft: british
	Ad-h		A draft: heavy
	Am		A marchador
	Ao-aA		A original: arabian Autochthonous
		Ao-aA1a*, Ao-aA1a1, Ao-aA1a2, Ao-aA1a2a, Ao-aA1a3, Ao-aA1a4, Ao-aA1a5, Ao-aA1b, Ao-aA2, Ao-aA3	
	Ao-aD		A original: arabian Duelmener
		Ao-aD1, Ao-aD2	
	Ao-aM		A original: arabian Marwari
	Ao-n		A original: noriker
Clade H was first described in a Spanish (H) Sorraia horse			
	Hc		H china
	Hs		H spanish
		Hs-a	H spanish: autochthonous
		Hs-b	H spanish: barb
Clade T was first described in Thoroughbreds (T)			
	Ta		T arabian
		Ta*	
		Ta-s	T arabian: shagya
		Ta-b	T arabian: bairactar
		Ta-bA	
	Tb-d		T(horough)bred: darley
		Tb-dM	T(horough)bred: darley Mambrino
		Tb-dW	T(horough)bred: darley Whalebone
		Tb-dW1, Tb-dW2, Tb-dW3, Tb-dW4	
	Tb-o		T(horough)bred: other
		Tb-oB	T(horough)bred: other Byerley/Godolphin
		Tb-oB1*, Tb-oB1a, Tb-oB1b, Tb-oB1c, Tb-oB2, Tb-oB3*, Tb-oB3a*, Tb-oB3a1, Tb-oB3b*, Tb-oB3b1*, Tb-oB3b1a, Tb-oB3b1b, Tb-oB3b1c, Tb-oB4	
		Tb-oL	T(horough)bred: other Lipizzan
	Tk		T kladruber
	Tu		T ubiquitous
Non-crown Haplogroups			
I			Icelandic horse
N			Northern Europe
P			Przewalski

References

1. Kelekna, P. *The Horse in Human History*; Cambridge University Press: New York, NY, USA, 2009; ISBN 978-0521736299.
2. Librado, P.; Khan, N.; Fages, A.; Kusliy, M.A.; Suchan, T.; Tonasso-Calvière, L.; Schiavinato, S.; Alioglu, D.; Fromentier, A.; Perdereau, A.; et al. The origins and spread of domestic horses from the Western Eurasian steppes. *Nature* **2021**, *598*, 634–640. [CrossRef] [PubMed]
3. Librado, P.; Fages, A.; Gaunitz, C.; Leonardi, M.; Wagner, S.; Khan, N.; Hanghøj, K.; Alquraishi, S.A.; Alfarhan, A.H.; Al-Rasheid, K.A.; et al. The evolutionary origin and genetic makeup of domestic horses. *Genetics* **2016**, *204*, 423–434. [CrossRef] [PubMed]
4. Petersen, J.L.; Mickelson, J.R.; Cothran, E.G.; Andersson, L.S.; Axelsson, J.; Bailey, E.; Bannasch, D.; Binns, M.M.; Borges, A.S.; Brama, P.; et al. Genetic Diversity in the Modern Horse Illustrated from Genome-Wide SNP Data. *PLoS ONE* **2013**, *8*, e54997. [CrossRef] [PubMed]
5. Bower, M.A.; Campana, M.G.; Nisbet, R.; Weller, R.; Whitten, M.; Edwards, C.J.; Stock, F.; Barrett, E.; O'Connell, T.; Hill, E.W.; et al. Truth in the bones: Resolving the identity of the founding elite thoroughbred racehorses. *Archaeometry* **2012**, *54*, 916–925. [CrossRef]
6. Raudsepp, T.; Finno, C.J.; Bellone, R.R.; Petersen, J.L. Ten years of the horse reference genome: Insights into equine biology, domestication and population dynamics in the post-genome era. *Anim. Genet.* **2019**, *50*, 569–597. [CrossRef]
7. Fages, A.; Hanghøj, K.; Khan, N.; Gaunitz, C.; Seguin-Orlando, A.; Leonardi, M.; McCrory Constantz, C.; Gamba, C.; Al-Rasheid, K.A.; Albizuri, S.; et al. Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. *Cell* **2019**, *177*, 1419–1435.e31. [CrossRef]
8. Hendricks, B. *International Encyclopedia of Horse Breeds*; University of Oklahoma Press: Norman, OK, USA, 2007; ISBN 9780806138848.
9. Arnason, T.; van Vleck, L.D. Genetic Improvement of the Horse. In *The Genetics of the Horse*; Bowling, A., Ruvinsky, A., Eds.; CABI Publishing: Wallingford, UK, 2000; p. 527. ISBN 9780851999258.
10. Grilz-Seger, G.; Druml, T. *Lipizzaner Hengststämme*; Vehling Medienservice und Verlag GmbH: Graz, Austria, 2011.
11. Petersen, J.L.; Mickelson, J.R.; Cleary, K.D.; McCue, M.E. The american quarter horse: Population structure and relationship to the thoroughbred. *J. Hered.* **2014**, *105*, 148–162. [CrossRef]
12. Weatherby, J. *An Introduction to a General Stud Book*; H. Reynell: London, UK, 1791.
13. Cunningham, E.P.; Dooley, J.J.; Splan, R.K.; Bradley, D.G. Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Anim. Genet.* **2001**, *32*, 360–364. [CrossRef]
14. McGivney, B.A.; Han, H.; Corduff, L.R.; Katz, L.M.; Tozaki, T.; MacHugh, D.E.; Hill, E.W. Genomic inbreeding trends, influential sire lines and selection in the global Thoroughbred horse population. *Sci. Rep.* **2020**, *10*, 1–12. [CrossRef]
15. Zechner, P.; Sölkner, J.; Bodo, I.; Druml, T.; Baumung, R.; Achmann, R.; Marti, E.; Habe, F.; Brem, G. Analysis of diversity and population structure in the Lipizzan horse breed based on pedigree information. *Livest. Prod. Sci.* **2002**, *77*, 137–146. [CrossRef]
16. Pjontek, J.; Kadlečík, O.; Kasarda, R.; Horný, M. Pedigree analysis in four Slovak endangered horse breeds. *Czech J. Anim. Sci.* **2012**, *57*, 54–64. [CrossRef]
17. Calafell, F.; Larmuseau, M.H. The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Hum. Genet.* **2017**, *136*, 559–573. [CrossRef] [PubMed]
18. Jobling, M.A.; Tyler-Smith, C. Fathers and sons: The Y chromosome and human evolution. *Trends Genet.* **1995**, *11*, 449–456. [CrossRef]
19. Radovic, L.; Remer, V.; Reiter, S.; Bozlak, E.; Felkel, S.; Grilz-Seger, G.; Brem, G.; Wallner, B. Y chromosome genetic variation and deep genealogies provide new insights on Lipizzan sire lines. *J. Equine Vet. Sci.* **2021**, *100*, 103501. [CrossRef]
20. Wallner, B.; Palmieri, N.; Vogl, C.; Rigler, D.; Bozlak, E.; Druml, T.; Jagannathan, V.; Leeb, T.; Fries, R.; Tetens, J.; et al. Y Chromosome Uncovers the Recent Oriental Origin of Modern Stallions. *Curr. Biol.* **2017**, *27*, 2029–2035.e5. [CrossRef]
21. Felkel, S.; Vogl, C.; Rigler, D.; Jagannathan, V.; Leeb, T.; Fries, R.; Neuditschko, M.; Rieder, S.; Velie, B.D.; Lindgren, G.; et al. Asian horses deepen the MSY phylogeny. *Anim. Genet.* **2018**, *49*, 90–93. [CrossRef]
22. Felkel, S.; Vogl, C.; Rigler, D.; Dobretsberger, V.; Chowdhary, B.P.; Distl, O.; Fries, R.; Jagannathan, V.; Janečka, J.E.; Leeb, T.; et al. The horse Y chromosome as an informative marker for tracing sire lines. *Sci. Rep.* **2019**, *9*, 1–12. [CrossRef]
23. Han, H.; Wallner, B.; Rigler, D.; MacHugh, D.E.; Manglai, D.; Hill, E.W. Chinese Mongolian horses may retain early domestic male genetic lineages yet to be discovered. *Anim. Genet.* **2019**, *50*, 399–402. [CrossRef]
24. Liu, S.; Yang, Y.; Pan, Q.; Sun, Y.; Ma, H.; Liu, Y.; Wang, M.; Zhao, C.; Wu, C. Ancient Patrilineal Lines and Relatively High ECAY Diversity Preserved in Indigenous Horses Revealed with Novel Y-Chromosome Markers. *Front. Genet.* **2020**, *11*, 1–12. [CrossRef]
25. Castaneda, C.; Juras, R.; Khanshour, A.M.; Randlaht, I.; Wallner, B.; Rigler, D.; Lindgren, G.; Raudsepp, T.; Cothran, E.G. Population genetic analysis of the Estonian native horse suggests diverse and distinct genetics, ancient origin and contribution from unique patrilines. *Genes* **2019**, *10*, 629. [CrossRef]
26. Wallner, B.; Vogl, C.; Shukla, P.; Burgstaller, J.P.; Druml, T.; Brem, G. Identification of Genetic Variation on the Horse Y Chromosome and the Tracing of Male Founder Lineages in Modern Breeds. *PLoS ONE* **2013**, *8*, e60015. [CrossRef] [PubMed]
27. Cosgrove, E.J.; Sadeghi, R.; Schlamp, F.; Holl, H.M.; Moradi-Shahrbabak, M.; Miraei-Ashtiani, S.R.; Abdalla, S.; Shykind, B.; Troedsson, M.; Stefaniuk-Szmukier, M.; et al. Genome Diversity and the Origin of the Arabian Horse. *Sci. Rep.* **2020**, *10*, 1–13. [CrossRef] [PubMed]
28. Glazewska, I. Speculations on the origin of the Arabian horse breed. *Livest. Sci.* **2010**, *129*, 49–55. [CrossRef]
29. About the World Arabian Horse Organization (WAHO). Available online: http://www.waho.org (accessed on 27 October 2021).

30. Schubert, M.; Lindgreen, S.; Orlando, L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **2016**, *9*, 1–7. [CrossRef] [PubMed]
31. Gómez-Sánchez, D.; Schlötterer, C. ReadTools: A universal toolkit for handling sequence data from different sequencing platforms. *Mol. Ecol. Resour.* **2018**, *18*, 676–680. [CrossRef]
32. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
33. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.T.; Abecasis, G.R.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
34. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef]
35. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
36. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **2013**, *14*, 178–192. [CrossRef]
37. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* **2012**, arXiv:1207.3907.
38. Lippold, S.; Xu, H.; Ko, A.; Li, M.; Renaud, G.; Butthof, A.; Schröder, R.; Stoneking, M. Human paternal and maternal demographic histories: Insights from high-resolution Y chromosome and mtDNA sequences. *Investig. Genet.* **2014**, *5*, 1–17. [CrossRef] [PubMed]
39. Bandelt, H.J.; Forster, P.; Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **1999**, *16*, 37–48. [CrossRef] [PubMed]
40. Hammer, M.F. A nomenclature system for the tree of human Y-Chromosomal binary haplogroups. *Genome Res.* **2002**, *12*, 339–348. [CrossRef]
41. Achilli, A.; Olivieri, A.; Soares, P.; Lancioni, H.; Kashani, B.H.; Perego, U.A.; Nergadze, S.G.; Carossa, V.; Santagostino, M.; Capomaccio, S.; et al. Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 2449–2454. [CrossRef] [PubMed]
42. Raswan, C.R.; Seydel, H. *Der Araber und Sein Pferd*; Dritte Nac.; Olms Georg: Hildesheim, Germany, 1998.
43. Forbis, J. *Das klassische arabische Pferd*; Paul Parey: Hamburg, Germany, 1980; ISBN 978-3489609322.
44. Arab Families by Grace Dashiell (Western Horseman June, 1950). Available online: https://wiwfarm.com/GDArabianFamilies.html (accessed on 23 November 2021).
45. Wells, M.L. *The Illustrated Guide to The Morab Horse*, 1st ed.; Holladay Publishing: Manning, NC, USA, 2009; ISBN 0578004658.
46. The two Siglavys with world influence by Furrer, B in Arabische Pferde in the Focus (1/2019). Available online: https://in-the-focus.com/2019/04/ap-1-2019-ist-online/ (accessed on 21 January 2022).
47. Pruski, W. Organizowanie państwowych stadnin arabskich. In *Dwa Wieki Polskiej Hodowli Koni Arabskich (1778–1978) i Jej Sukcesy na Świecie*; Państwowe Wydawnictwo Rolnicze i Leśne: Warszawa, Poland, 1983; pp. 278–322.
48. Jobling, M.A.; Tyler-Smith, C. Human Y-chromosome variation in the genome-sequencing era. *Nat. Publ. Gr.* **2017**, *18*, 485–497. [CrossRef]
49. Almarzook, S.; Reissmann, M.; Arends, D.; Brockmann, G.A. Genetic diversity of Syrian Arabian horses. *Anim. Genet.* **2017**, *48*, 486–489. [CrossRef]
50. Khanshour, A.M.; Cothran, E.G. Maternal phylogenetic relationships and genetic variation among Arabian horse populations using whole mitochondrial DNA D-loop sequencing. *BMC Genet.* **2013**, *14*, 1–12. [CrossRef]
51. Khanshour, A.; Conant, E.; Juras, R.; Cothran, E.G. Microsatellite analysis of genetic diversity and population structure of Arabian horse populations. *J. Hered.* **2013**, *104*, 386–398. [CrossRef]
52. Machmoum, M.; Boujenane, I.; Azelhak, R.; Badaoui, B.; Petit, D.; Piro, M. Genetic Diversity and Population Structure of Arabian Horse Populations Using Microsatellite Markers. *J. Equine Vet. Sci.* **2020**, *93*, 103200. [CrossRef]
53. Flade, J.E. *Das Araberpferd*, 3rd ed.; A. Ziemsen Verlag: Wittenberg Lutherstadt, Germany, 1977; p. 144.
54. Wrangel, G.C.G. *Die Rassen des Pferdes. Ihre Entstehung, Geschichtliche Entwicklung und Charakteristische Kennzeichen. Erster Band*; Verlag von Schickhardt Ebner (Konrad Witter): Stuttgart, Germany, 1908.
55. Nissen, J. *Enzyklopädie der Pferderassen, Europa Band 2*; Franck-Kosmos Verlags-GmbH & Co: Stuttgart, Germany, 1998.
56. Orlando, L. Ancient Genomes Reveal Unexpected Horse Domestication and Management Dynamics. *BioEssays* **2020**, *42*, 1–8. [CrossRef] [PubMed]
57. Bowling, A.; Del Valle, A.; Bowling, M. A pedigree-based study of mitochondrial D-loop DNA sequence variation among Arabian horses. *Anim. Genet.* **2000**, *31*, 1–7. [CrossRef] [PubMed]
58. Warmuth, V.; Eriksson, A.; Bower, M.A.; Barker, G.; Barrett, E.; Hanks, B.K.; Li, S.; Lomitashvili, D.; Ochir-Goryaeva, M.; Sizonovg, G.V.; et al. Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 8202–8206. [CrossRef] [PubMed]

3. Publication III in preparation

Equine Y-chromosomal variants develop into a diagnostic tool for paternal ancestry studies and enlighten the legacy of Spanish horses

Viktoria Remer[1], Barbara Wallner[1], Ivica Međugorac[2], Gottfried Brem[1]

[1] Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna, 1210 Vienna, Austria; vet.remer@gmail.com (V.R.); gottfried.brem@bremkg.de (G.B.); Barbara.wallner@vetmeduni.ac.at (B.R.)

[2] Veterinärwissenschaftliches Department, Ludwig-Maximilians-Universität München, D-82152 Martinsried; ivica.medjugorac@gen.vetmed.uni-muenchen.de (I.M.)

3.1. Abstract

The male-specific region of the Y chromosome (MSY) harbors information on the male ancestry of populations. In horses, historic stallion-mediated breeding wielded influence on many modern breeds, and MSY analyses can illuminate the male demographic history. In previous studies, the MSY haplotype signatures on the Y phylogeny of two recent refiner breeds, namely English Thoroughbred and Arabian horses, were successfully determined by combining MSY haplotype frequency data and breed history information. Once MSY variants are determined as diagnostic for specific ancestry, they can be used as a genetic marker system for migration tracing of paternal lineages. Consequently, the origin of sire lines that survived in modern breeds can be enlightened. The current stage of work focuses on how the MSY haplotype landscape in modern breeds harmonizes with the anecdotal breed formation. Besides Thoroughbred and Arabian stallions, Spanish studs exerted a remarkable influence on modern breeds and their origin and influence are herein assessed by MSY analyses.

3.2. Introduction

The horse (*Equus caballus*) has accompanied humans for the last five millennia and humans have shaped horses for their purposes. Today's 58 million horses counted worldwide (FAO, 2021a) are divided into hundreds of breeds.

Breeding selection is strongly acting on males and stallion-mediated refinement wielded influence on many modern breeds. Since the Y chromosome is inherited uniparentally and almost unmodified, MSY analyses can illuminate the male side of demographic history. In modern domestic horses, the breeding success of a few paternal lineages led to the predominance of one, approximately 2000-year-old 'crown haplogroup' (CH) (Wallner et al., 2017). Within the CH, MSY haplotypes specific for Thoroughbred and Arabian ancestry were characterized in previous studies based on NGS data and MSY haplotyping (Felkel et al., 2019; Remer et al., 2022). Thoroughbreds and Arabian horses were extensively used as refiners in breeding within the recent 200 years. In current work efforts, the legacy of another renowned horse population, 'Spanish horses', is investigated. From the 16th to the 18th century, those horses were disseminated from the Iberian Peninsula throughout Europe and were also translocated to the New World. MSY analyses should have the potency to illuminate the breeding success of patrilineages that derived from Spanish ancestry. In this study, the MSY haplotype spectrum was determined in a variety of modern breeds. The sampled breeds comprise North African, Baroque Type, Iberian, European Coldblood, a collection of New World breeds, as well as Warmblood and Riding Pony breeds of uncertain ancestry. So far, little was known about the MSY landscape in those breeds and collected and genotyped horses are expected to enlighten the paternal legacy of studs that were disseminated from the Iberian Peninsula.

3.3. Material and methods

3.3.1. Sample collection

Samples collected from males were either i) hair root samples provided by private horse owners, taken in compliance with the animal welfare standards in the respective country, or ii) retained samples from paternity testing when breeding associations permitted to use this source (DNA, hair, blood). Informed consent was obtained from horse owners. Pedigree information on samples was provided by owners or breeding associations upon availability. Lipizzan horse blood samples were collected during an EU-INCO Kopernikusproject: 'Biotechnical methods in the maintenance of the genetic diversity in the Lipizzan horse breed' (Project No. IC15CT96-0904). Permission for the scientific use of the samples was granted by all involved stud farms, which were partners in the project. All samples of this project are coded.

The horse Y chromosome haplotype distribution project was discussed and approved by the institutional ethics and welfare committee of the University of Veterinary Medicine Vienna in concordance with GSP guidelines and national legislation (ETK-10/05/2016). The research was performed in accordance with relevant guidelines reported in the above-mentioned document.

3.3.2. Pedigree reconstruction and compilation of the dataset

In order to collect as many paternal lines as possible and to avoid oversampling of highly frequent lines, paternal genealogies were reconstructed from pedigrees in the breeds of interest as described in Remer et al., 2022 before

sampling. Male-tail line reconstruction was stored in a string format for males that could be potentially sampled. Pedigree information was either provided by breeding associations, pertinent literature (Druml, 2006; Grilz-Seger & Druml, 2011; Schweisgut, 1995), or by reconciling results from multiple databases, which are listed in the supplementary section (chapter X., Table 3). Based on this information, two to three present male descendants of each sire in the 1970s were selected. Pedigree information was not only used for the compilation of the dataset – covering a comprehensive collection of the sublines of sire lines active in a breed, while preventing oversampling of certain sire lines - but was also needed for the genealogical application of MSY analyses.

In breeds with no pedigrees available, as many individuals as possible were collected. After haplotype screening, a retrospective downsampling was performed in a way that a maximum of ten breed members were chosen while assuring that all haplotypes detected in the genotyping approach are still represented. In total, the dataset consisted of 508 horses representing 62 breeds from 28 countries, of which 340 samples were hair roots, 45 were blood samples and for 57 horses genomic DNA samples were provided. For 66 samples the haplotype information was derived from NGS data analysis in Remer et al., 2022. The identity of the horses analyzed is not shown to protect the anonymity of the sample.

Out of the 508 samples, pedigree information of more than 4 generations was available for 311 (61%). The majority of samples were from the European Continent (385 samples or 75,78%), some from the Americas (83 samples or 16,34%), and the rest from North Africa (34 samples or 6,69%) and Asia (6 samples or 1,18%). The 72 different breeds were classified into nine categories of interest for the project. The categories are termed as ‘breed groups’ and listed in Table 1, whereas a detailed list of the breed groups can be found in the supplementary section (chapter X., Table 4). Out of the 508 analyzed samples, 20 did not fall into one of the breed categories mentioned above but were included in the analyses in order to represent all CH haplotypes, defined in Remer et al., 2022, in the genotyping tree. Those samples were comprised in the breed group ‘others’. Sampled horses in this group, that had pedigree information available, evidently derived from Arabian or Thoroughbred ancestry.

Table 1: Dataset for extended MSY haplotype analyses. Listed are the nine breed groups enclosing 62 breeds. In total 508 male horses were analyzed of which 311 had pedigree information available. The number of sampled male horses per breed group, as well as the number of those with pedigree availability and the number of samples with NGS data available, is given. A detailed list of horse breeds is given in the supplementary section (chapter X., Table 4). WB = Warmblood

breed group	N enclosed breeds	N samples	N pedigree available	N NGS data available
Akhal Teke	1	12	0	3
Baroque Type breeds	5	73	72	9
British Pony breeds	6	41	40	0
European Coldblood breeds	17	131	101	21

North African breeds	3	72	4	6
New World breeds	15	91	36	1
Iberian breeds	5	41	22	1
WB and Riding Pony breeds (paternal ancestry uncertain)	5	27	20	7
Others	5	20	16	18
Total	**62**	**508**	**311**	**66**

3.3.3. MSY haplotype analyses

In a previous study (Remer et al., 2022) 86 MSY haplotypes were described that are distinguished by 281 CH variants. This CH structure is shown in Figure 4 and was used as a backbone tree for the genotyping approach. Detailed information on variants such as their coordinates on the LipY764 assembly, type of variant, and flanking regions is published in Remer et al. 2022. In the next step, 163 markers on the tree were selected for MSY haplotype genotyping (161 CH variants and the two CH-determining variants rAY and rAX). Out of those variants, 157 were SNVs, eight were short Indels and one was an STR. In Figure 4 the selected markers are underlined and previously defined MSY signatures specific for Thoroughbreds (Felkel et al., 2019) and Arabians (Remer et al., 2022) are indicated with colors. These previous studies evidenced that markers on colored branches are diagnostic for Thoroughbred and Arabian ancestry.

All 508 horses listed in Table 1 were genotyped for the selected 163 variants and this was done by different methods. For the 66 males that had NGS data available the allelic state of the defined variants was extracted from the output of the freebayes genotyping file (published in Table S6 in Remer et al., 2022). For the remaining 442 male horses, the STR fBVB was sized using Gene Marker® (Softgenetics, State State College, USA) and the other 162 variants were genotyped by the use of PCR-based KASP^{TM} (Kompetitive Allele-Specific PCR) genotyping technology (KASP^{TM}, lgcgroup.com). The workflow, which included DNA isolation and genotyping was conducted as described in Felkel et al., 2019 and Remer et al., 2022. In those previous studies, the focus of the KASP^{TM} genotyping approach laid mainly on two breeds, Thoroughbred horses and Arabian horses. Here, 45 additional markers were selected in order to cover the full CH MSY haplotype spectrum. The 163 variants used in the KASP^{TM} genotyping approach of the extended dataset are listed in the supplementary section (chapter X., Table 5).

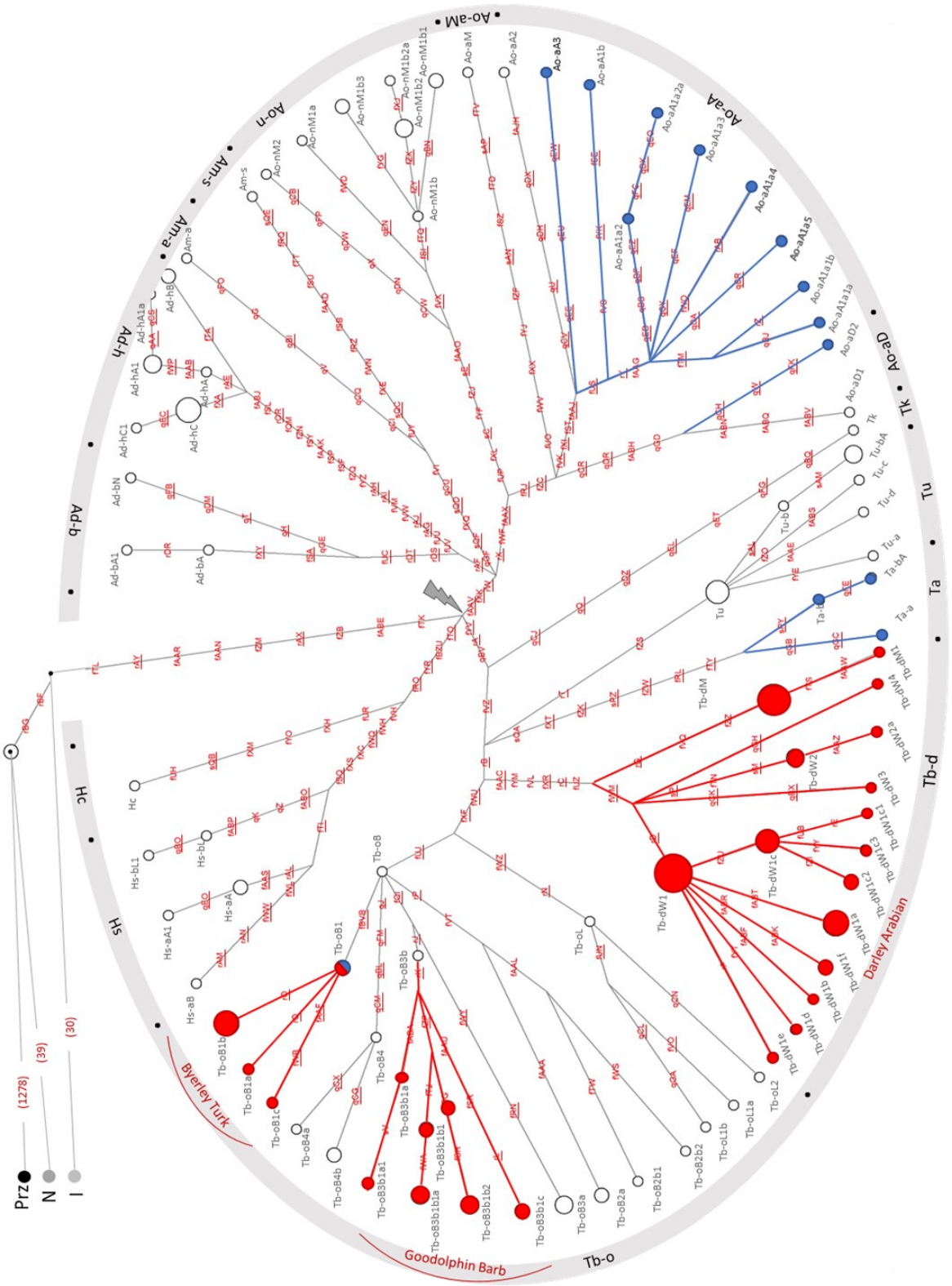


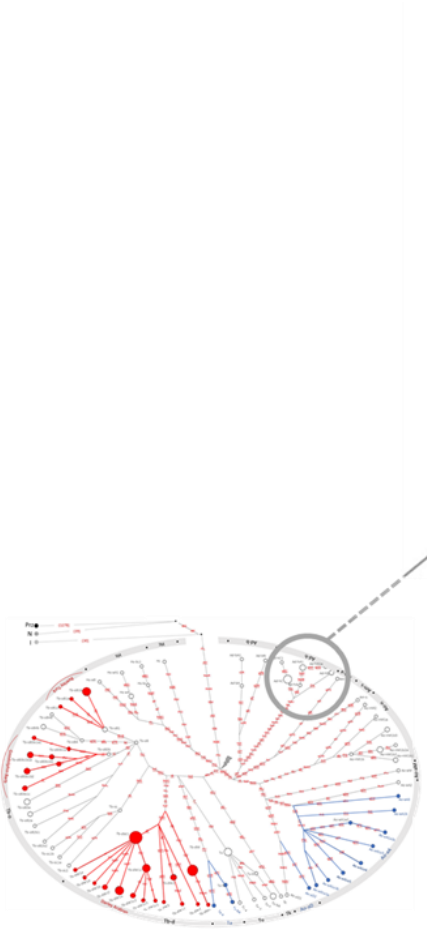
Figure 4: CH tree used as the backbone for genotyping MSY haplotypes, modified after Remer et al., 2022. Haplotypes are indicated as circles, with circle size proportional to frequency. Variants are indicated on branches and underlined when selected for genotyping. The position of the CH MRCA is marked with an arrow. In the outer circle, CH haplogrouping in concordance to Remer et al., 2022 is given. Branches previously assigned to MSY English Thoroughbred lineages are marked in red and those of Arabian lineages in blue.

3.3.4. The KASP genotyping approach

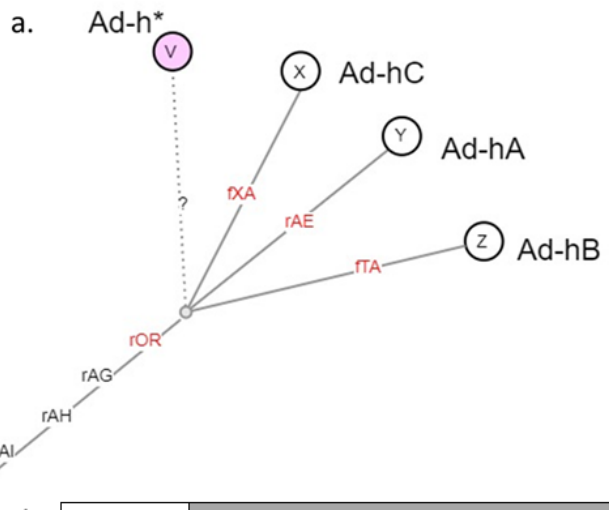
Since little was known about the MSY haplotype landscape of the breeds analyzed, a consecutive analytic approach was performed. For each sample, genotyping started with the CH-determining variants (rAY, rAX) followed by key variants for the three major CH haplogroups T, A, and H (rA, rW, and fYR). After that, the sample was genotyped for the haplotypes of the major haplogroup the sample belongs to on a fine-scaled haplotyping level towards the tip of the branches. If markers gave tautological results in the KASP™ genotyping approach, a representative marker was chosen. In the end, the CH haplotypes were inferred based on 103 variants ('key markers') and those are highlighted in the variant list that is provided in the supplementary section (chapter X., Table 5).

It must be stated here, that the KASP™ genotyping approach is limited by ascertainment bias and thus can only determine the allelic state of previously NGS-ascertained variants. Consequently, haplotypes that may be retained in breeds that were not represented in the NGS approach of previous studies (Remer et al., 2022) might be missed. Therefore, if a sample was placed onto an internal branching point after KASP™ genotyping, the haplotype was marked with an asterisk '*'. The emergence of such asterisk-marked haplotypes '*HTs' and the selection of key variants are illustrated in Figure 5. In the tree of Figure 6, which illustrates the genotyping results, *HTs are outlined with a dashed line branching off at the respective internal node.

Figure 5: Asterisk-marked haplotypes and key variants. Simplified illustration of the KASP genotyping approach and the occurrence of asterisk-marked haplotypes '*HTs'. a) A zoom into the Ad-h branch out of Figure 6. The variants, rAJ, rAI, rAH, rAG, and rOR gave tautological results in all genotyped samples. Thus, a representative variant (rOR) was selected which is marked in red on the branch. b) Variants were genotyped in a consecutive manner, towards the tips of branches on a fine-scaled level. The binary code reveals whether the sample carries the ancestral allele (0) or the derived allele (1) at the representative marker. The sample 'V' was placed on the Ad-h branch, because it carried the derived allele at five markers, but could not be assigned to a specific Ad-h haplotype as it showed the ancestral allele at fXA, rAE, and fTA. This sample is placed onto the internal node, and its haplotype was determined as Ad-h*. The variants determining its potential private haplotype were not yet ascertained by NGS.



a.



b.

	variants	rAJ	rAI	rAH	rAG	rOR	fXA	rAE	fTA
samples									
X		1	1	1	1	1	1	0	0
Y		1	1	1	1	1	0	1	0
Z		1	1	1	1	1	0	0	1
V		1	1	1	1	1	0	0	0

For each analyzed sample the allelic states at the 163 variants were concatenated and missing calls were imputed as described in Remer et al., 2022. The program Network (version 10.2) (Bandelt et al., 1999) was used to infer the genetic network of haplotypes. The network was redrawn with the program draw.io to illustrate the genotyping results as shown in Figure 6.

3.4. Results and Discussion

All 508 males analyzed clustered into the CH and within the CH, 82 haplotypes were detected. Out of those haplotypes, 64, supported by 286 males, fully matched the haplotypes ascertained in the NGS sequencing approach in Remer et al., 2022. However, 222 males (43,70 %) were placed onto internal nodes of the backbone topology and thus were termed as *HTs (in total 18). Most notable, the inner node clustering was most pronounced in horses that were not well represented in the NGS sequencing panel used for generating the haplotype phylogeny (Remer et al., 2022). This was particularly evident in horses from New World and North African breeds. Ten different *HTs (Ad-b*, Ad-h*, Am*, Am-s*, Am-sA*, Ao*, Ao-aA1a*, Cr*, Hs-b*, T1*) were detected in New World breeds and eight (Tb-oB1*, T2*, Hs-b*, Ao-aA1a*, Ao-aA*, Am*, Ad-h*, Ad-bN*) in North African breeds, with some of them overlapping. This suggests that multiple haplotypes are retained in those horse populations that were not represented at all in the CH phylogeny. Consequently, the CH diversity in those breeds is underestimated due to ascertainment bias.

The detected 82 CH haplotypes were united into 45 subhaplogroups (sHGs), which are listed in the supplementary section (chapter X., Table 6). The MSY haplotypes detected in the 508 horses are shown in Figure 6 with the 45 sHGs designated in the outer circle.

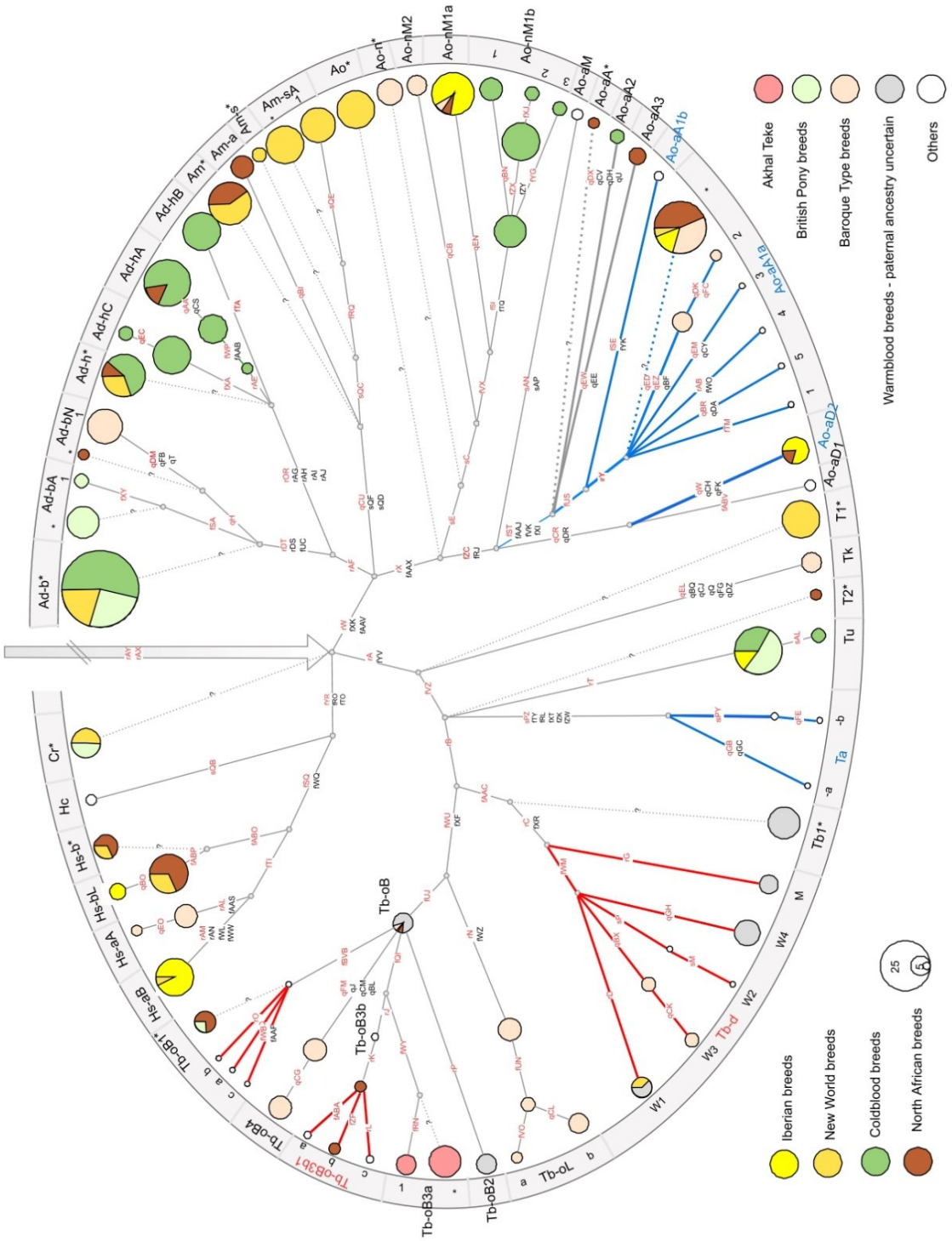


Figure 6: CH haplotypes genotyping results. Simplified CH haplotype network based on 163 variants modified after Remer et al., 2022. Haplotypes detected in the 508 males are shown as circles, with circle size proportional to frequency. The color code of samples is according to breed groups (defined in Table 1). Markers are given on branches with key markers in red. Not yet fully ascertained ambiguous *HTs are labelled with an asterisk (*), their branches are outlined with a dashed line and a '?'. The outer circle designates the 45 sHGs. The sHGs previously stated as Thoroughbred signatures are colour-coded in red, those previously stated as Arabian signatures are color-coded in blue.

3.4.1. MSY signatures of recent refiner breeds

The two recent refiner breeds, Thoroughbreds and Arabians, derive from closed populations and are linebred for multiple generations (Glazewska, 2010; Weatherbys, 1791). Previous studies have shown that both have left unequivocal MSY haplotype signatures on the Y phylogeny. The Thoroughbred signature was stated as sHGs Tb-d and Tb-oB3b1 (Felkel et al., 2019) and the Arabian signature as Ao-aA1a, Ao-aA1b, Ao-aD2, and Ta (Remer et al., 2022). Thus, diagnostic variants for Thoroughbred respectively Arabian ancestry can be applied to trace their influence. Ambiguity has been shown with regards to haplotype Tb-oB1*; which was carried by the Thoroughbred founder stallion Byerley Turk in 1680 (Felkel et al., 2019), but it was also detected in Arabians and many other breeds like Akhal Tekes and Turkomans (Remer et al., 2022).

Here achieved results showed a pronounced recent paternal influence of Thoroughbreds in Warmblood and Riding Pony breeds with uncertain ancestry, as many of those breeds clustered into sHG Tb-d, the evidenced signature of the Thoroughbred founder Darley Arabian (Figure 6). A Thoroughbred influence was also detected, but weaker, in North African breeds (Tb-oB3b1), New World breeds (Tb-d), and Baroque Type breeds (Tb-d). In contrast, no Thoroughbred disseminated sHGs were observed in Iberian and Coldblood breeds, as can be seen in Figure 6.

A prominent Arabian signature showed up in North African breeds (Ao-aA1a, Ao-aD2). The sHGs Ao-aA* and Ao-aA3, detected in North African breeds, are in proximity to the Arabian-specific sHGs. Those two sHGs could be interpreted as a signature of earlier introduced lines of Arabian origin into the North African horse population. The Arabian signature was detectable, but weaker, in Baroque Type breeds (Ao-aA1a) and Iberian breeds (Ao-aA1a, Ao-aD2); further, it was only marginal in New World breeds (Ao-aA1a) and not detected in Warmblood and Riding Pony breeds with unknown ancestry. The Arabian-derived sHGs Ta and Ao-aA1b were not detected in any of the nine breed groups analyzed as can be seen in Figure 6.

The above-mentioned haplotype Tb-oB1* was additionally detected in British Pony and North African breeds. Due to the lack of informative markers, it is currently not possible to distinguish whether the occurrence of Tb-oB1* in those breeds is from Arabian, Turkoman, or Thoroughbred ancestry. Further studies are needed, and the implementation of faster evolving STR markers might allow more fine-scaled tracing (Ballantyne et al., 2014; Court, 2021).

3.4.2. MSY sHGs landscape in modern breeds

Several Warmblood and Riding Pony breeds with uncertain ancestry, Baroque Type breeds, North African and British Pony breeds, as well as the Akhal Tekes clustered into the Tb haplogroup but did not carry one of the Thoroughbred or Arabian derived lines (Tb-d, Tb-oB3b1) within the Tb haplogroup as illustrated in Figure 6. Previous studies attributed the Tb haplogroup to Turkoman origin (Wallner et al., 2017). However, the wide distribution of ‘non-Thoroughbred Tb’ sHGs (Tb1*, Tb-oL, Tb-oB2, Tb-oB4, Tb-oB, Tb-oB1*, Tb-oB3a) is remarkable and further investigation is needed to resolve the origin and dispersal of the ‘non-Thoroughbred-derived’ Tb sHGs. Here, those haplotypes were not considered as informative to circumscribe the paternal legacy of Spanish horses.

In contrast to the discrete MSY haplotype signatures of the recent refiner breeds (Thoroughbreds and Arabians), the broad sHG spectrum observed in North African (16 sHGs), New World (13 sHGs), and Baroque Type breeds (11 sHGs) was remarkable. Notably, North African and Baroque Type horses were scattered across the three CH clades A, H, and T, while New World breeds were mainly detected in clades A and H, as shown in Figure 6. In North African and New World breeds this can be explained by retention of early introduced, private haplotypes. While in the case of Baroque Type breeds, sires from different regions and studs were compiled to consolidate breeds. The results underline the importance of clade A and H to work out the wide dissemination of stallions from historic Spain.

Coldblood breeds clustered into seven sHGs: Tu, Ao-nM1b, Ad-b* and, the close grouping, sHGs Ad-h*, Ad-hA, Ad-hB, and Ad-hC. Remarkable is the joint clustering of British ponies and Coldbloods into sHGs Ad-b* and Tu as shown in Figure 6. Notably, all haplogroups that were prevalent in Coldblood and British Pony breeds (Ad-b, Ad-h, Ao-nM1b, and Tu), were also detected in Iberian and/or New World breeds, whereas the sHGs determined in New World breeds always branch off basally. For instance, sHGs Ad-hA, Ad-hB, and Ad-hC are well-established in Coldblood breeds, while the basally branching off sHG Ad-h* was detected in Coldblood and New World breeds as shown in Figure 6. This arrangement in the phylogenetic tree points to a shared origin of the MSY sHGs preserved in today's Coldblood, British Pony, Iberian, and New World breeds.

3.4.3. MSY signatures of the 'Spanish dissemination'

The traces of Spanish horses were expected to be more diffuse than those of the two recent refiner breeds. Firstly, because of the more unsecured origin of the Iberian horse itself, and secondly the earlier and broad dissemination of the 'Spanish horse' starting from the 15th century, a timeframe far before line-breeding concepts arose (Nissen, 1997). A pronounced 'Spanish' influence is expected in Iberian and New World, and also in Coldbloods and British Pony breeds. In total 46 of such breeds, represented by 304 males, were in the comprehensive dataset and overall, 23 sHGs were detected among those. Only three of the 23 sHGs derive from recent Arabian or Thoroughbred influence (Ao-aA1a, Ao-aD2, Tb-d), another two (Ao-aA2, Tb-oB1*) were detected in very low frequency and are herein not considered as relevant. The remaining 18 sHGs were detected at high frequency in the breed groups listed above and can thus be used as a proxy to circumscribe MSY signatures that could be indicative for Spanish influence. Considering breeding history and pedigree documentation in Baroque Type breeds, five additional sHGs (Ad-bN, Ao-nM2, Hs-aA, Ao-n*, and Tk) could also be assumed to be of Spanish origin. Consequently, the Spanish dissemination signature would be marked by a broad spectrum, in total 23 sHGs (illustrated in Figure 7). Several of those presumably 'Spanish-dissemination sHGs' were also shared with North African breeds, which reflects the pronounced historic interaction between North Africa and the Iberian Peninsula. It can be stated that these preliminary results promise new insights on the presumed origin and dissemination of MSY CH lineages. However, deeper analyses need to be conducted. This should enclose additional NGS sequenced horses, the inclusion of ancient and historic samples, as well as a more comprehensive genotyping panel of breeds of interest. In the case of Spanish-disseminated CH lineages, this would be for example a more comprehensive dataset of Iberian and New World breeds. Overall, carving out MSY spectra

characteristic for Spanish disseminated studs is crucial for a better understanding of the origin and dissemination of sire lines in a large part of the global horse population.

Figure 7: MSY sHGs of presumable Spanish origin. Listed are the 23 sHGs detected at high frequency in Spanish-influenced breeds (Iberian, New World, Coldbood, British Pony and Baroque Type breeds). Shown is the absolute sample number of samples genotyped for each breed group and the number of samples clustering into the listed sHGs. The breed group 'others' comprises mainly horses of Arabian or Thoroughbred ancestry. WB & RP = Warmbloods & Riding ponies of uncertain ancestry.

	N of samples	N in listed sHGs	Ao-nM1a	Ad-b*	Ad-bA	Ad-bN	Ad-h*	Ad-hA	Ad-hB	Ad-hC	Am*	Am-s*	Am-sA	Ao*	Ao-n*	Ao-nM1b	Ao-nM2	Cr*	Hs-aA	Hs-aB	Hs-b*	Hs-bL	Tl*	Tk	Tu
Iberian	41	32	14																	12		3			3
New World	91	89		17			5				7	2	22	13				4			2	5	12		
Coldbloods	131	131		36			10	24	12	12						28									9
British ponies	41	37		13	10													4							10
Baroque Type	73	33	1			12									6		4		6					4	
North African	72	36	1			1	3	4			10										4	13			
Akhal Theke	12	0																							
WB & RP	27	0																							
others	20	0																							

3.4.4. Horse MSY haplotyping as a tool to predict the origin of sire lines

In humans, Y chromosomal haplotyping is a valued system for combining genetics with ancestry information (Calafell & Larmuseau, 2016). In horses, MSY analyses can also be meaningful on a genealogical scale. This was previously proven in the published manuscripts (Felkel et al., 2019; Remer et al., 2022). In the presented panel, pedigree information spanning more than 4 generations was available for 310 (61%) horses. These 310 horses trace back in their male-tail line to 76 foundation sires. When sampled horses were grouped with regard to their paternal lineage information, it could be shown that some MSY haplotypes were even unique for sublines of a certain sire line. For example, all Tb-oL1 carriers trace back to the stallion 'Favory senior', born in 1779, who was the founder stallion of a sire line active in the Lipizzan horse breed. Within this 'Favory' sire line the two subline-haplotypes, Tb-oL1a and Tb-oL1b evolved through *de novo* mutations. Out of the 82 CH haplotypes, 30 (detected in a total of 165 samples) were specific for members of a particular sire line and are listed in Table 2.

Table 2: MSY haplotypes specific to certain sire lines. Listed are 30 haplotypes (and the sHGs they belong to), that are specific for sire lines. In total, those haplotypes were detected in 165 samples. Foundation sires are capitalized in the column 'remarks on sire lines' with their name and birth year, if available. Some haplotypes are even unique for sublines of certain sire lines.

sHG	HT		Breeds (N)	remarks on sire lines
Ad-bA	Ad-bA1		Connemara Pony (2)	Connemara Pony sire line after MOUNTAIN LAD,1928
Ad-bN	Ad-bN1		Friesian (10), Kladruber (2)	Friesian sire line after NEMO,1885
Ad-hA	Ad-hA1a		Ardennes (4), Belgian draft (1), Hungarian Coldblood (1), Murgese (1), Rhenish German Coldblood (8), Saxony Thuringian Coldblood (1)	Coldblood sire line after ORANGE I,1863
Ad-hB	Ad-hB		Noriker (10), Southgerman Coldblood (2)	Coldblood sire line after NERO,1906
Ad-hC	Ad-hC		Haflinger (10)	Detected in the Haflinger sublines B, M, N, and W; tracing back to FOLIE 249,1874
	Ad-hC1		*Haflinger (2)*	*Haflinger subline S after SAPHIR,1989; tracing back to FOLIE 249,1874*
	Am-sA*		Mangalarga Marchador (13)	carriers trace back to FORTUNA II,1848
Am-sA	Am-sA1		Mangalarga Marchador (9)	descendants of RECIFE,1930
Ao-nM2	Ao-nM2		Kladruber (4)	Kladruber sire line after SACRAMOSO RISANOTTA XXVI,1889
Ao-nM1b	Ao-nM1b		Noriker (7), Southgerman Coldblood (1)	Coldblood sire line after DIAMANT,1903 and ELMAR,1896
	Ao-nM1b1		Haflinger (5)	Detected in Haflinger sublines A and N; tracing back to FOLIE 249,1874
	Ao-nM1b2		Noriker (9), Southgerman Coldblood (2)	Coldblood sire line after VULKAN,1887
	Ao-nM1b2a		*Noriker (2)*	*VULKAN,1887 subline*
	Ao-nM1b3		*Noriker (2)*	*VULKAN,1887 subline*
Ao-aA2	Ao-aA2		Haflinger (2)	detected in Haflinger subline ST-line; tracing back to FOLIE 249,1874

Ao-aA1a	Ao-aA1a2a		Kladruber (1), Lipizzaner (3)	Lipizzan sire line after SIGLAVY SENIOR,1810
	Ao-aA1a2a1		*Lipizzaner (1)*	*SIGLAVY SENIOR,1810 subline*
Tk	Tk		Kladruber (4)	Kladruber sire line tracing back to GENERALE XXXIV,
Tb1*	Tb1*		Orlov Trotter (8)	private sHG to Orlov Trotter with the single founder BARS I,1784
Tb-d	*Tb-dW3*		*Lipizzaner (2)*	*Lipizzan sire line after TULIPAN,1860 with origin unclear, now affiliated to Darley Arabian*
	Tb-dW3a		*Lipizzaner (2)*	*TULIPAN,1860 subline*
	Tb-dW4		*Quarter Horse (6)*	*Quarter horse sire line after TRAVELER,1880 with origin unclear, now affiliated to Darley Arabian*
Tb-oL	Tb-oL		Lipizzaner (5)	Lipizzan sire line after NEAPOLITANO,1790
	Tb-oL1		Lipizzaner (2)	Lipizzan sire line after FAVORY SENIOR,1779
	Tb-oL1a		*Lipizzaner (1)*	*FAVORY SENIOR,1779 subline*
	Tb-oL1b		*Kladruber (3), Lipizzaner (1)*	*FAVORY SENIOR,1779 subline*
Tb-oB4	Tb-oB4		Lipizzaner (5)	Lipizzan sire line after MAESTOSO SENIOR,1773
	Tb-oB4b		Lipizzaner (5)	Lipizzan sire line after PLUTO SENIOR,1765
Hs-aA	Hs-aA		Lipizzaner (5)	Lipizzan sire line after CONVERSANO,1767
	Hs-aA1		*Lipizzaner (1)*	*CONVERSANO,1767 subline*

MSY analyses have the potency to deduce the origin of sire lines that survived in modern breeds. The practical application of combining MSY and deep genealogies is herein exemplified in a Baroque Type breed. In the Lipizzan horse breed, a strong tradition in line-breeding is established (Brem, 2011; Kavar et al., 2002), and this is exerted in the naming system. Male Lipizzan foals get traditionally two names, with the first being the line of the sire and the second being the name of the dam. Hence, the 'first name' of a Lipizzan stallion is an explicit label of the paternal lineage - similar to the surname in humans. Out of the 89 foundation sires that initially contributed to the Lipizzan horse, only eight are preserved today (Grilz-Seger & Druml, 2011). Six lines, Siglavy, Pluto, Conversano, Neapolitano, Favory, and Maestoso, descend from classical foundation stallions used in the 18th and 19th centuries by the Lipica stud. Among those, one line was founded by an Original Arabian stallion: 'Siglavy' (born 1810, imported to Lipica 1814). Two additional lines, Tulipan and Incitato, trace back to foundation sires that were never used at Lipica stud, but by other studs within the historic boundaries of the Habsburg Empire (Grilz-Seger & Druml, 2011).

The Lipizzan males included in the genotyping dataset, represented the eight sire lines and pedigree information is available for all individuals. Seven of the eight Lipizzan sire lines carried MSY haplotypes private for the Lipizzan horse breed (Ao-aA1a2a, Tb-dW3, Tb-oL, Tb-oL1, Tb-oB4, Tb-oB4b, Hs-aA1), shown in Table 6. Siglavys grouped, as expected since it is in consistence with pedigree documentation, primarily into the Arabian characteristic sHG Ao-aA1. Conversanos clustered into the Spanish-disseminated sHG Hs-a, a pattern that also coincides with the presumed origin of the foundation sire. Tulipans grouped into a Thoroughbred derived sHG Tb-d. Even though they carried a private haplotype Tb-dW3, see Table 2 and Figure 6; this observation substantiates a Thoroughbred origin of the Tulipan sire line.

The remaining five sire lines grouped into haplogroup Tb and this leaves their origin still open for speculations.

Neapolitanos and Favorys clustered together in sHG Tb-oL, and Maestosos and Plutos jointly at Tb-oB4 as can be seen in Table 2. The Incitatos as the only Lipizzan sire line did not exhibit a private MSY haplotype; they grouped in sHG Tb-oB and this is shown in Figure 6 where Baroque Type breeds cluster in sHG Tb-oB.

3.5. Conclusion

Besides the evidently corroborated MSY signatures of the ‘Arabian’ as well as ‘English’ introgression waves in Felkel et al., 2019 and Remer et al., 2022, the repercussion of horse dissemination from historic Spain on the MSY spectrum of modern breeds is verifiable. It can be disclosed that MSY signatures of Thoroughbreds, Arabians, and Spanish horses explain the history of many sire lines that survived in modern horse breeds.

V. DISCUSSION

The history of the domestic horse is marked by recurrent adjustment to changing socio-economic needs. Over the last centuries most modern breeds were established and this has largely been achieved through the use of a few, very popular sires that have been extensively shifted among breeds (Nissen, 1997). Innumerable myths surround the origin and influences of foundation sires. Scientifically sound genetic deduction of the ancestry and dissemination of prominent sire lines is therefore crucial for our understanding of the historic development of horse breeds. Paternal ancestry inference with Y-chromosomal markers is best established in humans, where deep insights derived from several decades of steady progress in the variant discovery of the MSY and haplotype analyses (Jobling & Tyler-Smith, 2017).

Detailed sire line tracing has been impeded by the low sequence diversity on the horse MSY. The predominance of one, approximately 1,500-year-old, CH in modern horse breeds hindered the ascertainment of informative MSY markers.

1. The MSY as a genetic marker in horses – new insights and open questions

Resolving the CH haplotype structure and characterizing Y-chromosomal haplotypes indicative of influential foundation sires is crucial to disentangle the historic development of modern horse breeds. This knowledge allows the deduction of MSY signatures indicative of certain introgression waves. During this Doctoral research project, the haplotype structure of the CH was resolved in closer detail based on single nucleotide variants from NGS data (up to 5.8 megabases of the MSY were screened in whole-genome or customized target-enriched NGS data) (Felkel et al., 2019; Remer et al., 2022). Establishing the MSY as an informative tool was achieved by a combination of variant ascertainment, to resolve the CH haplotype structure, and comprehensive genotyping for selected, haplotype-determining markers. Within the CH, haplotypes indicative of Thoroughbred (all in haplogroup Tb) (Felkel et al., 2019; Wallner et al., 2013) as well as Arabian ancestry (most in haplogroups Ao-aA1, Ao-aD2, Ta) (Remer et al., 2022) have been determined. This identified CH lineages indicative of two recent introgression waves, namely the ‘Thoroughbred’ and the ‘Arabian’ wave. In an extended dataset, it was shown, that also a third wave (‘Spanish’) is verifiable. It can be disclosed that many sire lines that survived in modern breeds derive from Thoroughbred Arabian or Spanish ancestry.

Furthermore, it was exemplified that MSY analyses in horses are prepared to go practical for addressing genealogical questions (Felkel et al., 2019; Remer et al., 2022) and to deduce the origin of sire lines active in a breed. Acquired results enabled genetic sire line characterization not only in Thoroughbreds and Arabians (Felkel et al., 2019; Remer et al., 2022) but also in other modern breeds with sufficient pedigree information available like Coldblood or Baroque Type breeds.

Overall, MSY haplotyping can be corroborated as meaningful to studying population demography, genealogical and forensic questions in horses. However, the comprehensive genotyping also displayed weaknesses in the approach

system. It should be noted that any custom set of markers will be highly dependent on the samples originally ascertained and sequenced. As the ascertainment panel was focused mainly on European breeds the backbone tree used for MSY haplotyping is quite representative for Thoroughbreds, imported Arabians, Coldbloods, and Lipizzans, but still has deficiencies when studying MSY diversity in other populations. The ascertainment bias becomes evident in New World and North African breeds. In these horses, genotyping revealed many early-branching haplotypes. This evidences, that the haplotype structure is not sufficiently resolved for those breeds, since their private haplotypes are not yet ascertained. In regards to New World breeds, for example, pronounced haplotype diversity can be expected once horses carrying the asterisk-marked subhaplogroups Ad-b*, Ad-h*, Am* Am-s*, Am-sA*, Ao*, T1*, Hs-b* after genotyping become sequenced. High haplotype diversity in New World breeds can be expected for various reasons: first, the early transmission of horses from Europe to the American continent, in the late 15th century (Cortés et al., 2017) should have promoted the segregated development of exported Y-chromosomal lineages; second, also a pronounced autosomal and mtDNA diversity was detected in New World breeds (Luís et al., 2006); and third, a large number of horses and a variety of breeds inhabit the Americas today (FAO, 2021b; Nissen, 1997).

Apart from that, several terminal haplotypes (for example node clustering of various breeds on haplotypes Tb-oB1*, Ad-b*, or Ao-aA1a* as can be seen in Figure 6) were observed in high frequency in modern breeds. This evidences that the resolution of the phylogeny is still too sparse to distinguish all recently established sire lines in modern breeds. Consequently, the incorporation of faster evolving Y-chromosomal short tandem repeat (STR) loci should be intended in future studies. While biallelic SNVs represent an evolutionary stable marker system, STRs mutate much more frequently (Court, 2021). Multiallelic Y-STRs are conventionally used in human studies to infer a person's biogeographic ancestry or for the examination of paternal lineages. The latter approach answers questions of inheritance in forensic as well as civil investigations since they can shed light on paternal relationships (Court, 2021). Y-STR markers were also used in cattle to enlighten the demographic expansion/spread of patrilineages (Ganguly et al., 2020). On the horse MSY, unexpectedly little allelic variation was detected in STRs so far (Kreutzmann et al., 2014). In further studies, to improve the MSY reference, more STR markers will hopefully be detected that are polymorphic. In the future, combined analyses of binary SNVs and multiallelic STRs could then enable to resolve of recent demographic events in closer detail. This could, for example, enlighten the spread of the subhaplogroup Tb-oB1*. This subhaplogroup was introduced into the English Thoroughbred but is also shared among Arabians, Turkomans, Akhal Tekes, North African, and British Pony breeds. At the current stage of knowledge, the historic connections between these breeds are unclear (Remer et al., 2022). In the end, whole-genome or custom-targeted NGS are free of ascertainment bias and could be considered the ultimate method to clarify some open ancestry questions.

It is known that Asian and North European breeds were to some extent protected from CH introgressions and earlier branching haplotypes are retained in those breeds (Castaneda et al., 2019; Felkel et al., 2018; Felkel et al., 2019; Han et al., 2019; Wallner et al., 2017). These 'non-CH' haplotypes were not studied in the context of this Doctoral research project. Hence, resolving the 'basally branching CH and non-CH' haplotypes via carefully conducted variant

ascertainment followed by comprehensive genotyping populations is the next logical step toward a more complete view of the paternal history of horses. A crucial contribution to this future work was the improved resolution of the CH and the identification of diagnostic markers for two refiner breeds, the Thoroughbred and Arabian horse breed (Felkel et al., 2019; Remer et al., 2022). Both are extensively used as refining forces on a global scale and are thus a major impetus for the wide distribution of the CH.

2. Genealogical application of the MSY in modern breeds

Practical implications arise from MSY investigations in forensic analysis and genetic genealogy research. This is best established in humans, where the MSY is used in paternal kinship tests and to combine genetic data with family history (Calafell & Larmuseau, 2016; Jobling et al., 1997; Kayser, 2017; Whiting & Coyle, 2020). Tracing sire lines by the use of MSY variants now enables molecular genetic inference of paternal genealogies also in horses, independent of written pedigree records. This can develop into an interesting tool for horse breeders since with MSY analyses one can detect incongruencies to written pedigree records, investigate relationships among sire lines and decipher the origin of foundation stallions. As shown in Thoroughbreds (Felkel et al., 2019) and Arabians (Remer et al., 2022) MSY analysis can even serve as a forensic tool to detect errors in the paternal lineage, even if they occurred multiple generations previously. Incongruencies to pedigrees can originate from the inadequate or uncertain reconstruction of deep genealogies or cuckoldries. The sequencing panel used in this research was primarily designed to exemplify Thoroughbred and Arabian influences and learn about their signatures in other modern horse breeds. However, sire line tracing on a genealogical scale was also feasible in other breeds that exhibited sufficient written pedigree documentation. In the Lipizzan horse breed, for example, MSY patterns allowed for the genetic characterization of sire lines and the validation of anecdotal breed origins. This offers new insights that were not deducible from autosomal data (Grilz-Seger et al., 2019; Radovic et al., 2021).

3. Reflections on the establishment of the predominant CH

Pedigree records of modern breeds illustrate the remarkable influence that Spanish, Arabian, and Thoroughbred stallions have had on modern horse breeds (Druml, 2011; Grilz-Seger & Druml, 2017b; Nissen, 1997). Presented MSY data indicate that a major impetus for the wide distribution of the CH was, apart from the strong recent amplification of Thoroughbred and Arabian paternal lineages, the dissemination of horses from historic Spain. The traces of Spanish horses exhibit a relatively broader MSY spectrum than the two recent refiner breeds. While only two subhaplogroups (Tb-d, Tb-oB3b1) were stated as Thoroughbred signature (Felkel et al., 2019) and four subhaplogroups (Ao-aA1a, Ao-aA1b, Ao-aD2, Ta) were stated as Arabian signature (Remer et al., 2022), a total of 23 MSY subhaplogroups were identified that could be indicative of a Spanish ancestry. An explanation for this broad haplotype spectrum is the diffuse origin of the Iberian horse itself, which developed from native populations combined with transcontinental influences from North Africa (Druml, 2011; Nissen, 1997). In addition, the dissemination of horses from the Iberian Peninsula

started before line-breeding concepts arose and studbooks were closed in Europe, in the late 18th century (Nissen, 1998). The clustering of breeds within the CH displays the solidification of specific MSY haplotypes in closed, linebred populations. The unequivocal prints of Thoroughbred and Arabian horses, for example, can be explained by pronounced line-breeding in those breeds and their origin from closed populations (Głazewska, 2010; Weatherbys, 2021).

Today, Thoroughbred- and Arabian-inherited sire lines are active on a global scale (Nissen, 1997) and thus still promote the distribution of CH lineages. The dissemination of Spanish studs is a historic phenomenon that led to the establishment of CH lineages in many European and New World breeds. The presented results indicate that the signature of Spanish horses is retained in sire lines of Coldblood breeds, British Pony breeds, Baroque Type breeds, and New World breeds. The discovery of the New World initiated the introduction of CH lineages to the Americas and this is consistent with historical documentation and previous studies that confirmed gene flow from Iberian to New World breeds (Cortés et al., 2017; Jimenez et al., 2012; Luís et al., 2006). Although still preliminary, as there are many New World breeds left to be studied, the MSY haplotypes in the few breeds investigated in this thesis confirmed the 'out of Spain' hypothesis. Previous scientific work revealed that Iberian breeds harbor higher diversity values than South American and North American breeds for several genetic diversity indices, such as nucleotide diversity and mtDNA haplotype diversity (Luís et al., 2006). This trend was not mirrored in the observed MSY spectrum. Instead, haplotype diversity is comparable between New World breeds and North African breeds, which leads to the hypothesis that MSY haplotypes were preserved in New World breeds that went extinct in European horse populations. However, considering ascertainment bias and the large number of horses in the Americas today (FAO, 2021a), this needs to be proven in further studies with improved tools and sample panels.

Overall, the predominance of the CH in modern horse breeds is remarkable and the CH bears the legacy of refinement breeding in historic Europe, from the 16th century onwards; as well as the management shift towards closed studs and line-breeding for breed formation within the recent 200 years.

4. Conclusion

Given the impact of stallions in upgrading and the drop of diversity in present horse breeds, analysis of the male-specific region of the Y chromosome should be further promoted and regularly integrated into horse population genetic screenings. The male-specific Y chromosome haplotype analysis should be conducted in population-representative sample sets to infer how many and which paternal lineages exist in a breed. The information content generated for Thoroughbreds and Arabians (Felkel et al., 2019; Remer et al., 2022) can serve as a basis when screening rural populations for signatures of recent introgression. In addition, the detection of basally branching off haplotypes in rural populations identifies candidate samples for next-generation sequencing and thus for the continuous improvement of the diagnostic marker set. The findings of the thesis provide new insights into the historic development of horse breeds and show how male-specific Y chromosome investigations can support breeding decisions in the future.

VI. SUMMARY

Paternal lines play an important role in horse breeding and sire lines are especially celebrated in modern horse breeds. The male-specific, non-recombining region of the Y chromosome (MSY), is inherited almost unmodified as a single linkage group, defined as a haplotype from the father to his sons. Hence, the MSY is an ideal genetic marker to study the paternal demography of a population and can be used to conduct male lineage tracing. This has been hampered in modern horses for a long time because of the low sequence diversity observed on the MSY. The majority of MSY haplotypes cluster into the monophyletic crown haplogroup (CH). In the context of the Doctoral thesis, methods were established that enable the comprehensive utilization of the Y chromosome as a genetic marker in horses. This was achieved by combining high-throughput sequencing techniques for variant ascertainment and comprehensive MSY haplotyping in modern breeds.

First, the phylogenetic classification of CH haplotypes was conducted and the results showed, that the CH descends from a most recent common ancestor that lived no longer than approximately 1000 to 2000 years ago. The genetic characterization of sire lines in modern horse breeds was performed by connecting male-tail line information from pedigrees and MSY haplotypes. Major findings were the identification of MSY haplotypes indicative of Thoroughbred and Arabian stallions. In addition, the MSY spectrum in other modern breeds was determined, comprising mainly European horse breeds, but also breeds from North Africa, the Americas, and Asia.

The potential of genetic sire line tracing on a genealogical scale became evident, as controversies between MSY haplotype pattern and pedigree documentation were observed. This makes MSY analysis applicable to address paternity questions. Furthermore, the track record of specific paternal bloodlines and the question of their possible dispersal route through time and space was addressed in the manuscript. Overall, the results show the potential of MSY analysis in horses and how it can improve our knowledge of the influence, origin, dispersal, and relationship of sire lines. MSY information now complements findings based on mtDNA and autosomal information. It has revealed new insights into horse population structure, past demographic phenomena and uncovered that the anecdotal history of some breeds should be reconsidered.

VII. ZUSAMMENFASSUNG

In der Pferdezucht sind väterliche Linien von großer Bedeutung, besonders in modernen Pferderassen sind etablierte Hengstlinien renommiert. Der ‚männlich-spezifische, nicht rekombinierende Teil des Y Chromosoms' (MSY), wird beinahe unverändert als Kopplungsgruppe, definiert als Haplotyp, vom Vater an seine Söhne vererbt. Daher ist der MSY ein idealer genetischer Marker, um die väterliche Demographie einer Population zu untersuchen sowie männliche Abstammungen zu ermitteln. Dies war bei modernen Pferden bislang schwer durchzuführen, da die Variabilität am MSY so gering ist. Die große Mehrheit der MSY Haplotypen gruppiert sich in der monophyletischen Kronhaplogruppe (CH). Im Rahmen der Doktorarbeit wurden Methoden etabliert, die eine umfangreiche Nutzung des Y Chromosoms als genetischen Marker bei Pferden ermöglicht. Dies wurde durch die Kombination von Hochdurchsatz-Sequenzierungstechniken zur Varianten-ermittlung und umfassender MSY Haplotypisierung in modernen Rassen erreicht.

Zunächst wurde die phylogenetische Klassifizierung von CH Haplotypen durchgeführt. Die Ergebnisse zeigten, dass die CH von einem gemeinsamen Vorfahren abstammt, der nicht vor mehr als 1000 bis 2000 Jahren lebte. Durch die Kombinierung von männlichen Stammbäumen mit MSY Haplotypen war die genetische Charakterisierung von Hengstlinien möglich. Wichtige Ergebnisse waren die Identifizierung von MSY Haplotypen, welche indikativ für Vollblut- und Araberhengste sind. Darüber hinaus wurde das MSY Spektrum in anderen modernen Rassen bestimmt, darunter waren hauptsächliche Rassen aus Europa, aber auch welche aus Nordafrika, Amerika und Asien.

Das Potenzial der genetischen Abstammungsermittlung auf genealogischer Ebene wurde deutlich, als Diskrepanzen zwischen MSY Haplotyp Mustern und Stammbaumdokumentationen beobachtet wurden. Dies macht die MSY Analyse anwendbar, um Vaterschaftsfragen zu beantworten. Darüber hinaus wurde im Manuskript die Erfolgsbilanz spezifischer väterlicher Blutlinien und die Frage nach ihrem möglichen Ausbreitungsweg durch Zeit und Raum thematisiert. Insgesamt zeigen die Ergebnisse das Potenzial von MSY Analysen bei Pferden und wie sie unser Wissen über Einfluss, Herkunft, Verbreitung und Verwandtschaft von Hengstlinien verbessern kann.

VIII. LIST OF LITERATURE

Achilli, A., Olivieri, A., Soares, P., Lancioni, H., Kashani, B. H., Perego, U. A., Nergadze, S. G., Carossa, V., Santagostino, M., Capomaccio, S., Felicetti, M., Al-Achkar, W., Penedo, M. C. T., Verini-Supplizi, A., Houshmand, M., Woodward, S. R., Semino, O., Silvestrelli, M., Giulotto, E., … Torroni, A. (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. https://doi.org/10.1073/pnas.1111637109

Almathen, F., Charruau, P., Mohandesan, E., Mwacharo, J. M., Orozco-TerWengel, P., Pitt, D., Abdussamad, A. M., Uerpmann, M., Uerpmann, H. P., De Cupere, B., Magee, P., Alnaqeeb, M. A., Salim, B., Raziq, A., Dessie, T., Abdelhadi, O. M., Banabazi, M. H., Al-Eknah, M., Walzer, C., … Burger, P. A. (2016). Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(24), 6707–6712. https://doi.org/10.1073/pnas.1519508113

AQHA. (2000). *"Thoroughbred horse for breeding".American Quarter Horse Association*. https://www.aqha.com/-/breeding-a-thoroughbred-to-a-quarter-horse?inheritRedirect=true&redirect=https%253A%252F%252Fwww.aqha.com%252Fhome%253Fp_p_id%253Dcom_liferay_portal_search_web_portlet_SearchPortlet%2526p_p_lifecycle%253D0%2526p_p_state%253Dmaximized%2526p_p_mode%253Dview%25

Ballantyne, K. N., Ralf, A., Aboukhalid, R., Achakzai, N. M., Anjos, M. J., Ayub, Q., Balažic, J., Ballantyne, J., Ballard, D. J., Berger, B., Bobillo, C., Bouabdellah, M., Burri, H., Capal, T., Caratti, S., Cárdenas, J., Cartault, F., Carvalho, E. F., Carvalho, M., … Kayser, M. (2014). Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats. *Human Mutation*, *35*(8), 1021–1032. https://doi.org/10.1002/humu.22599

Bandelt, H. J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*(1), 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

Batini, C., Hallast, P., Zadik, D., Delser, P. M., Benazzo, A., Ghirotto, S., Arroyo-Pardo, E., Cavalleri, G. L., De Knijff, P., Dupuy, B. M., Eriksen, H. A., King, T. E., De Munain, A. L., López-Parra, A. M., Loutradis, A., Milasin, J., Novelletto, A., Pamjav, H., Sajantila, A., … Jobling, M. A. (2015). Large-scale recent expansion of European patrilineages shown by population resequencing. *Nature Communications*, *6*(May). https://doi.org/10.1038/ncomms8152

Batini, C., & Jobling, M. A. (2017). Detecting past male-mediated expansions using the Y chromosome. *Human Genetics*, *136*(5), 547–557. https://doi.org/10.1007/s00439-017-1781-z

Begni Cinar Kul, Nuray Bilgen, Johannes A. Lenstra, O. Korkmaz Agaoglu, Bilal Akyuz, and O. E. (2015). Y-chromosomal variation of local goat breeds of Turkey close to the domestication centre. *Journal of Animal Breeding and Genetics*, *132*(6), 449–29. https://doi.org/10.1111/jbg.12154

Bellone, R. R., Forsyth, G., Leeb, T., Archer, S., Sigurdsson, S., Imsland, F., Mauceli, E., Engensteiner, M., Bailey, E., Sandmeyer, L., Grahn, B., Lindblad-Toh, K., & Wade, C. M. (2010). Fine-mapping and mutation analysis of TRPM1: A candidate gene for leopard complex (LP) spotting and congenital stationary night blindness in horses. *Briefings in Functional Genomics and Proteomics*, *9*(3), 193–207. https://doi.org/10.1093/bfgp/elq002

Berber, N., Gaouar, S., Leroy, G., Kdidi, S., Tabet Aouel, N., & Säidi Mehtar, N. (2014). Molecular characterization and differentiation of five horse breeds raised in Algeria using polymorphic microsatellite markers. *Journal of Animal Breeding and Genetics*, *131*(5), 387–394. https://doi.org/10.1111/jbg.12092

Bower, M. A., McGivney, B. A., Campana, M. G., Gu, J., Andersson, L. S., Barrett, E., Davis, C. R., Mikko, S., Stock, F., Voronkova, V., Bradley, D. G., Fahey, A. G., Lindgren, G., MacHugh, D. E., Sulimova, G., & Hill, E. W. (2012). The genetic origin and history of speed in the Thoroughbred racehorse. *Nature Communications*, *3*. https://doi.org/10.1038/ncomms1644

Bowling, A. T., 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. https://doi.org/10.1046/j.1365-2052.2000.00558.x

Bowling, Ann T. (1996a). *Horse Genetics*. CAB International.

Bowling, Ann T. (1996b). Pedigrees and breeding schemes. In *Horse Genetics* (pp. 129–140). CAB International.

Brem, G. (2011). *Der Lipizzaner im Spiegel der Wissenschaft* (1. (ed.)). Verlag der Österreichischen Akademie der Wissenschaften.

Brem, G., Förster, M., Gaillard, C., Gregor, G., Horst, P., Kräußlich, H., Wegner, W., & Zárate, A. V. (1997). *Tierzucht und Allgemeine Landwirtschaftslehre für Tiermediziner*. Ferdinand Enke Verlag.

Brown, S. K., Pedersen, N. C., Jafarishorijeh, S., Bannasch, D. L., Ahrens, K. D., Wu, J. Te, Okon, M., & Sacks, B. N. (2011). Phylogenetic distinctiveness of Middle Eastern and Southeast Asian village dog Y chromosomes illuminates dog origins. *PLoS ONE*, *6*(12). https://doi.org/10.1371/journal.pone.0028496

Burney, D. A., & Flannery, T. F. (2005). Fifty millennia of catastrophic extinctions after human contact. *Trends in Ecology and Evolution*, *20*(7), 395–401. https://doi.org/10.1016/j.tree.2005.04.022

Calafell, F., & Larmuseau, M. H. D. (2016). The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Human Genetics*, *136*(5), 559–573. https://doi.org/10.1007/s00439-016-1740-0

Castaneda, C., Juras, R., Khanshour, A., Randlaht, I., Wallner, B., Rigler, D., Lindgren, G., Raudsepp, T., & Gus Cothran, E. (2019). Population genetic analysis of the Estonian native horse suggests diverse and distinct genetics, ancient origin and contribution from unique patrilines. *Genes*, *10*(8), 1–15. https://doi.org/10.3390/genes10080629

Cieslak, J., Wodas, L., Borowska, A., Cothran, E. G., Khanshour, A. M., & Mackowski, M. (2017). Characterization of the Polish Primitive Horse (Konik) maternal lines using mitochondrial D-loop sequence variation. *PeerJ*, *2017*(8). https://doi.org/10.7717/peerj.3714

Claerhout, S., Van der Haegen, M., Vangeel, L., Larmuseau, M. H. D., & Decorte, R. (2019). A game of hide and seq: Identification of parallel Y-STR evolution in deep-rooting pedigrees. *European Journal of Human Genetics*, *27*(4), 637–646. https://doi.org/10.1038/s41431-018-0312-2

Cortés, O., Dunner, S., Gama, L. T., Martínez, A. M., Delgado, J. V., Ginja, C., Jiménez, L. M., Jordana, J., Luis, C., Oom, M. M., Sponenberg, D. P., Zaragoza, P., Vega-Pla, J. L., Aguirre Riofrío, L., Aguirre Riofrío, L., Brandariz, C., Calderón, J., Cañón Ferreras, J., Rosa Costa, M., … de Souza Sereno, F. T. P. (2017). The legacy of Columbus in American horse populations assessed by microsatellite markers. *Journal of Animal Breeding and Genetics*, *134*(4), 340–350. https://doi.org/10.1111/jbg.12255

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., Grützner, F., & Kaessmann, H. (2014). Origins and functional evolution of y chromosomes across mammals. *Nature*, *508*(7497), 488–493. https://doi.org/10.1038/nature13151

Court, D. S. (2021). The y chromosome and its use in forensic dna analysis. *Emerging Topics in Life Sciences*, *5*(3), 427–441. https://doi.org/10.1042/ETLS20200339

Cunningham, E. P., Dooley, J. J., Splan, R. K., & Bradley, D. G. (2001). Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Animal Genetics*, *32*(6), 360–364. https://doi.org/10.1046/j.1365-2052.2001.00785.x

Devi, K. M., & Ghosh, S. K. (2013). Molecular phylogeny of Indian horse breeds with special reference to Manipuri pony based on mitochondrial D-loop. *Molecular Biology Reports*, *40*(10), 5861–5867. https://doi.org/10.1007/s11033-013-2692-2

Druml, T. (2006). *Das Noriker Pferd* (1. Ausgabe). Vehling Medienservice und Verlag GmbH.

Druml, T. (2011). Historische Herkunft der Rasse. In *Der Lipizzaner im Spiegel der Wissenschaft* (pp. 9–32). Austrian Academy of Science.

Edwards, C. J., Ginja, C., Kantanen, J., Pérez-Pardal, L., Tresset, A., Stock, F., Gama, L. T., Penedo, M. C. T., Bradley, D. G., Lenstra, J. A., & Nijman, I. J. (2011). Dual origins of dairy cattle farming - Evidence from a comprehensive

survey of european Y-chromosomal variation. *PLoS ONE*, *6*(1). https://doi.org/10.1371/journal.pone.0015922

EMBL-EBI. (2021). *Human chromosome Y summary*. http://www.ensembl.org/Homo_sapiens/Location/Chromosome?chr=Y;r=Y:1-57227415

Encyclopaedia Britannica. (1998). *"Breed association".Encyclopedia Britannica*. https://www.britannica.com/science/breed-association

Encyclopaedia Britannica. (2011). *"Studbook".Enyclopaedia Britannica*. https://www.britannica.com/science/studbook

ESSA. (2018). *European State Studs Association e.V.* http://www.europeanstatestuds.org/en/studs.html

Eugene A. Foster, M. A. Jobling, P. G. Taylor, P. Donnelly, P. de Knijff, R. Mieremet, T. Zerjal, C. T.-S. (1998). Jefferson fathered slave`s last child. *Nature*, *396 (6706)*, 27–28. https://doi.org/10.1038/23835

Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., McCrory Constantz, C., Gamba, C., Al-Rasheid, K. A. S., Albizuri, S., Alfarhan, A. H., Allentoft, M., Alquraishi, S., Anthony, D., Baimukhanov, N., Barrett, J. H., Bayarsaikhan, J., Benecke, N., Bernáldez-Sánchez, E., … Orlando, L. (2019). Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. *Cell*, *177*(6), 1419-1435.e31. https://doi.org/10.1016/j.cell.2019.03.049

Fages, A., Seguin-Orlando, A., Germonpré, M., & Orlando, L. (2020). Horse males became over-represented in archaeological assemblages during the Bronze Age. *Journal of Archaeological Science: Reports*, *31*(December 2019), 102364. https://doi.org/10.1016/j.jasrep.2020.102364

Faith, J. T., & Surovell, T. A. (2009). Synchronous extinction of North America's Pleistocene mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(49), 20641–20645. https://doi.org/10.1073/pnas.0908153106

FAO. (2021a). *"Domestic Animal Diversity Information System (DAS-IS)".Food and Agriculture Organization of the United Nations*. http://www.fao.org/dad-is/en/

FAO. (2021b). *"Livestock Systems".Food and Agriculture Organization of the United Nations*. http://www.fao.org/livestock-systems/global-distributions/horses/en/

Felkel, S., Vogl, C., Rigler, D., Dobretsberger, V., Chowdhary, B. P., Distl, O., Fries, R., Jagannathan, V., Janečka, J. E., Leeb, T., Lindgren, G., McCue, M., Metzger, J., Neuditschko, M., Rattei, T., Raudsepp, T., Rieder, S., Rubin, C. J., Schaefer, R., … Wallner, B. (2019). The horse Y chromosome as an informative marker for tracing sire lines. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-42640-w

Felkel, S., Vogl, C., Rigler, D., Jagannathan, V., Leeb, T., Fries, R., Neuditschko, M., Rieder, S., Velie, B., Lindgren,

G., Rubin, C. J., Schlötterer, C., Rattei, T., Brem, G., & Wallner, B. (2018). Asian horses deepen the MSY phylogeny. *Animal Genetics*, *49*(1), 90–93. https://doi.org/10.1111/age.12635

Ganguly, I., Jeevan, C., Singh, S., Dixit, S. P., Sodhi, M., & Ranjan, A. (2020). *Y-chromosome genetic diversity of Bos indicus cattle in close proximity to the centre of domestication*. 1–9. https://doi.org/10.1038/s41598-020-66133-3

Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., Seguin-Orlando, A., Owens, I. J., Felkel, S., Bignon-Lau, O., Damgaard, P. de B., Mittnik, A., Mohaseb, A. F., Davoudi, H., Alquraishi, S., Alfarhan, A. H., Al-Rasheid, K. A. S., Crubézy, E., Benecke, N., … Orlando, L. (2018). Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science*, *360*(6384), 111–114. https://doi.org/10.1126/science.aao3297

Giacomini, J. P. (2014). *The modern "Baroque Breeds."* IALHA - International Andalusian & Lusitano Horse Association. https://web.archive.org/web/20141217041847/http://www.ialha.org/our-breed-2/how-andalusians-friesians-lipizzaners-and-lusitanos-are-called-baroque-horses/

Ginja, C., Gama, L. T., Cortés, O., Burriel, I. M., Vega-Pla, J. L., Penedo, C., Sponenberg, P., Cañón, J., Sanz, A., do Egito, A. A., Alvarez, L. A., Giovambattista, G., Agha, S., Rogberg-Muñoz, A., Lara, M. A. C., Afonso, S., Aguirre, L., Armstrong, E., Vallejo, M. E. C., … Martinez, A. (2019). The genetic ancestry of American Creole cattle inferred from uniparental and autosomal genetic markers. *Scientific Reports*, *9*(1), 1–16. https://doi.org/10.1038/s41598-019-47636-0

Glazewska, I. (2010). Speculations on the origin of the Arabian horse breed. *Livestock Science*, *129*(1–3), 49–55. https://doi.org/10.1016/j.livsci.2009.12.009

Głazewska, I. (2010). Speculations on the origin of the Arabian horse breed. *Livestock Science*, *129*(1–3), 49–55. https://doi.org/10.1016/j.livsci.2009.12.009

Götherström, A., Anderung, C., Hellborg, L., Elburg, R., Smith, C., Bradley, D. G., & Ellegren, H. (2005). Cattle domestication in the Near East was followed by hybridization with aurochs bulls in Europe. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1579), 2345–2350. https://doi.org/10.1098/rspb.2005.3243

Graves, J. A. M. (2006). Sex chromosome specialization and degeneration in mammals. *Cell*, *124*(5), 901–914. https://doi.org/10.1016/j.cell.2006.02.024

Grilz-Seger, G., & Druml, T. (2011). *Lipizzaner Hengststämme*. Vehling Medienservice und Verlag GmbH.

Grilz-Seger, G., & Druml, T. (2017a). Beginn der systematischen Arbeitspferdezucht. In *Kaltblutpferde im Spiegel der Wissenschaft* (pp. 81–140). Austrian Academy of Science.

Grilz-Seger, G., & Druml, T. (2017b). Das schwere Pferd in historischer Annäherung. In *Kaltblutpferde im Spiegel der Wissenschaft* (pp. 35–80). Austrian Academy of Science.

Grilz-Seger, G., Druml, T., Neuditschko, M., Dobretsberger, M., Horna, M., & Brem, G. (2019). High-resolution population structure and runs of homozygosity reveal the genetic architecture of complex traits in the Lipizzan horse. *BMC Genomics*, *20*(1), 1–17. https://doi.org/10.1186/s12864-019-5564-x

Groeneveld, L. F., Lenstra, J. A., Eding, H., Toro, M. A., Scherf, B., Pilling, D., Negrini, R., Finlay, E. K., Jianlin, H., Groeneveld, E., & Weigend, S. (2010). Genetic diversity in farm animals - A review. *Animal Genetics*, *41*(SUPPL. 1), 6–31. https://doi.org/10.1111/j.1365-2052.2010.02038.x

Gršković, B., Mršić, G., Vrdoljak, A., Merkaš, S., & Andelinović, Š. (2010). Population genetic analysis of haplotypes based on 17 short tandem repeat loci on Y chromosome in population sample from eastern Croatia. *Croatian Medical Journal*, *51*(3), 202–208. https://doi.org/10.3325/cmj.2010.51.202

Guimaraes, S., Arbuckle, B. S., Peters, J., Adcock, S. E., Buitenhuis, H., Chazin, H., Manaseryan, N., Uerpmann, H. P., Grange, T., & Geigl, E. M. (2020). Ancient DNA shows domestic horses were introduced in the southern caucasus and anatolia during the bronze age. *Science Advances*, *6*(38), 1–10. https://doi.org/10.1126/sciadv.abb0030

Guirao-Rico, S., Ramirez, O., Ojeda, A., Amills, M., & Ramos-Onsins, S. E. (2018). Porcine Y-chromosome variation is consistent with the occurrence of paternal gene flow from non-Asian to Asian populations. *Heredity*, *120*(1), 63–76. https://doi.org/10.1038/s41437-017-0002-9

Hallast, P., Batini, C., Zadik, D., Delser, P. M., Wetton, J. H., Arroyo-Pardo, E., Cavalleri, G. L., De Knijff, P., Bisol, G. D., Dupuy, B. M., Eriksen, H. A., Jorde, L. B., King, T. E., Larmuseau, M. H., De Munain, A. L., López-Parra, A. M., Loutradis, A., Milasin, J., Novelletto, A., … Jobling, M. A. (2014). The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Molecular Biology and Evolution*, *32*(3), 661–673. https://doi.org/10.1093/molbev/msu327

Hammer, M. (2002). A nomenclature system for the tree of human Y-Chromosomal binary haplogroups. *Genome Research*, *12*(2), 339–348. https://doi.org/10.1101/gr.217602

Han, H., Wallner, B., Rigler, D., MacHugh, D. E., Manglai, D., & Hill, E. W. (2019). Chinese Mongolian horses may retain early domestic male genetic lineages yet to be discovered. *Animal Genetics*, *50*(4), 399–402. https://doi.org/10.1111/age.12780

Han, Haige, Wallner, B., Rigler, D., MacHugh, D. E., Manglai, D., & Hill, E. W. (2019). Chinese Mongolian horses may retain early domestic male genetic lineages yet to be discovered. *Animal Genetics*, *50*(4), 399–402. https://doi.org/10.1111/age.12780

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

Hughes, J. F., & Rozen, S. (2012). Genomics and Genetics of Human and Primate Y Chromosomes. *Annual Review of*

Genomics and Human Genetics, *13*(1), 83–108. https://doi.org/10.1146/annurev-genom-090711-163855

Hurles, M. E., & Jobling, M. A. (2001). Haploid chromosomes in molecular ecology: Lessons from the human Y. *Molecular Ecology*, *10*(7), 1599–1613. https://doi.org/10.1046/j.0962-1083.2001.01314.x

J. R. S. Meadows, O. Hanotte, C. Drögemüller, J. Calvo, R. Godfrey, D. Coltman, J.F. Maddox, N. Marzanov, J. Kantanen, J. W. K. (2006). Globally dispersed Y chromosomal haplotypes in wild and domestic sheep. *Animal Genetics*, *37*(5), 444–453. https://doi.org/10.1111/j.1365-2052.2006.01496.x.

James, D. (1986). *The Arabian Horse at war*. Saudi Aramco World Magazine. https://web.archive.org/web/20120507173157/http://www.saudiaramcoworld.com/issue/198602/the.arabian.horse-at.war.htm

Janečka, J. E., Davis, B. W., Ghosh, S., Paria, N., Das, P. J., Orlando, L., Schubert, M., Nielsen, M. K., Stout, T. A. E., Brashear, W., Li, G., Johnson, C. D., Metz, R. P., Zadjali, A. M. Al, Love, C. C., Varner, D. D., Bellott, D. W., Murphy, W. J., Chowdhary, B. P., & Raudsepp, T. (2018). Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-05290-6

Jiang, Z., Williams, G. A., Chen, J., & Michal, J. J. (2010). Mitochondriomics of Reproduction and Fertility. In *Reproductive Genomics in Domestic Animals* (pp. 167–179). Blackwell Publishing. https://doi.org/10.1002/9780813810898

Jimenez, L. M., Mendez, S., Dunner, S., Cañón, J., & Cortés, Ó. (2012). Colombian Creole horse breeds: Same origin but different diversity. *Genetics and Molecular Biology*, *35*(4), 790–796. https://doi.org/10.1590/S1415-47572012005000064

Jobling, M. A. (2012). The impact of recent events on human genetic diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1590), 793–799. https://doi.org/10.1098/rstb.2011.0297

Jobling, M. A., Pandya, A., & Tyler-Smith, C. (1997). The Y chromosome in forensic analysis and paternity testing. *International Journal of Legal Medicine*, *110*(3), 118–124. https://doi.org/10.1007/s004140050050

Jobling, & Tyler-Smith. (2017). Human Y-chromosome variation in the genome-sequencing era. *Nature Reviews Genetics*, *18*(8), 485–497. https://doi.org/10.1038/nrg.2017.36

K. M. Cliffe, A. E. Day, M. Bagga, K. Siggens, C. R. Quilter, S. Lowden, H. A. Finlayson, C. J. Palgrave, N. Li, L. Huang, S. C, Blott, C. A. S. (2010). Analysis of the non-recombining Y chromosome defines polymorphisms in domestic pic breeds: ancestral bases identified by comparative sequencing. *Animal Genetics*, *41*(6), 619–629. https://doi.org/10.1111/j.1365-2052.2052.2010.02070.x

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

Kayser, M. (2017). Forensic use of Y-chromosome DNA: a general overview. *Human Genetics*, *136*(5), 621–635. https://doi.org/10.1007/s00439-017-1776-9

Kelekna, P. (2009). *The Horse in Human History*. Cambridge University Press.

Khadka, R., Jan Philipsson, S., & Georg Thaller, S. (2010). *European Master in Animal Breeding and Genetics Global Horse Population with respect to Breeds and Risk Status*. *August*.

Khanshour, A. M., & Cothran, E. G. (2013). Maternal phylogenetic relationships and genetic variation among Arabian horse populations using whole mitochondrial DNA D-loop sequencing. *BMC Genetics*, *14*, 10–17. https://doi.org/10.1186/1471-2156-14-83

King, T. E., & Jobling, M. A. (2009a). Founders, drift, and infidelity: The relationship between y chromosome diversity and patrilineal surnames. *Molecular Biology and Evolution*, *26*(5), 1093–1102. https://doi.org/10.1093/molbev/msp022

King, T. E., & Jobling, M. A. (2009b). What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends in Genetics*, *25*(8), 351–360. https://doi.org/10.1016/j.tig.2009.06.003

Kräußlich, H. (1997). Pferde. In *Tierzucht und Allgemeine Landwirtschaftslehre für Tiermediziner* (pp. 450–490). Ferdinand Enke Verlag.

Kreutzmann, N., Brem, G., & Wallner, B. (2014). The domestic horse harbours Y-chromosomal microsatellite polymorphism only on two widely distributed male lineages. *Animal Genetics*, *45*(3), 460. https://doi.org/10.1111/age.12149

Krischke, D. (2011). *Selektion in kleinen Populationen am Beispiel Berberpferd*. *April*, 145.

Kvist, L., Niskanen, M., Mannermaa, K., Wutke, S., & Aspi, J. (2019). Genetic variability and history of a native Finnish horse breed. *Genetics Selection Evolution*, *51*(1), 1–14. https://doi.org/10.1186/s12711-019-0480-8

L. van Dijk, Erwin, Auger, Hélène, Jaszczyszyn, Yan, Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics : TIG*, *30*(9), 418–426. https://doi.org/10.1016/j.tig.2014.07.001

Lahn, B. T., & Page, D. C. (1999). *Four Evolutionary Strata on the*. *286*(October), 964–967.

Larson, G., Karlsson, E. K., Perri, A., Webster, M. T., Ho, S. Y. W., Peters, J., Stahl, P. W., Piper, P. J., Lingaas, F.,

Fredholm, M., Comstock, K. E., Modiano, J. F., Schelling, C., Agoulnik, A. I., Leegwater, P. A., Dobney, K., Vigne, J. D., Vilà, C., Andersson, L., & Lindblad-Toh, K. (2012). Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(23), 8878–8883. https://doi.org/10.1073/pnas.1203005109

Li, G., Davis, B. W., Raudsepp, T., Wilkerson, A. J. P., Mason, V. C., Ferguson-Smith, M., O'Brien, P. C., Waters, P. D., & Murphy, W. J. (2013). Comparative analysis of mammalian y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Research*, *23*(9), 1486–1495. https://doi.org/10.1101/gr.154286.112

Librado, P., Fages, A., Gaunitz, C., Leonardi, M., Wagner, S., Khan, N., Hanghøj, K., Alquraishi, S. A., Alfarhan, A. H., Al-Rasheid, K. A., Der Sarkissian, C., Schubert, M., & Orlando, L. (2016). The evolutionary origin and genetic makeup of domestic horses. *Genetics*, *204*(2), 423–434. https://doi.org/10.1534/genetics.116.194860

Librado, P., Gamba, C., Gaunitz, C., Sarkissian, C. Der, Pruvost, M., Albrechtsen, A., Fages, A., Khan, N., Schubert, M., Jagannathan, V., Serres-armero, A., Kuderna, L. F. K., Povolotskaya, I. S., Seguin-orlando, A., Keyser, C., Marques-Bonet, T., Ludes, B., Crubézy, E., Leeb, T., … Orlando, L. (2017). Ancient genomic changes associated with domestication of the horse. In *Science* (Vol. 356, Issue April).

Librado, P., Khan, N., Fages, A., Kusliy, M. A., Suchan, T., Tonasso-Calvière, L., Schiavinato, S., Alioglu, D., Fromentier, A., Perdereau, A., Aury, J. M., Gaunitz, C., Chauvey, L., Seguin-Orlando, A., Der Sarkissian, C., Southon, J., Shapiro, B., Tishkin, A. A., Kovalev, A. A., … Orlando, L. (2021). The origins and spread of domestic horses from the Western Eurasian steppes. *Nature*, *598*(7882), 634–640. https://doi.org/10.1038/s41586-021-04018-9

Librado, P., & Orlando, L. (2021). Genomics and the Evolutionary History of Equids. *Annual Review of Animal Biosciences*, *9*, 81–101. https://doi.org/10.1146/annurev-animal-061220-023118

Lindgren, G., Backström, N., Swinburne, J., Hellborg, L., Einarsson, A., Sandberg, K., Cothran, G., Vilà, C., Binns, M., & Ellegren, H. (2004). Limited number of patrilines in horse domestication. *Nature Genetics*, *36*(4), 335–336. https://doi.org/10.1038/ng1326

Lippold, S., Matzke, N. J., 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*(1), 328. https://doi.org/10.1186/1471-2148-11-328

Liu, W. (2010). Comparative Genomics of the Y Chromosome and Male Fertility. In *Reproductive Genomics in Domestic Animals* (pp. 129–155). Blackwell Publishing.

Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N., Brockmann, G. A., Castaños, P., Cieslak, M., Lippold, S., Llorente, L., Malaspinas, A. S., Slatkin, M., & Hofreiter, M. (2009). Coat color variation at the beginning of Horse

domestication. *Science*, *324*(5926), 485. https://doi.org/10.1126/science.1172750

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

Makvandi-Nejad, S., Hoffman, G. E., Allen, J. J., Chu, E., Gu, E., Chandler, A. M., Loredo, A. I., Bellone, R. R., Mezey, J. G., Brooks, S. A., & Sutter, N. B. (2012). Four loci explain 83% of size variation in the horse. *PLoS ONE*, *7*(7), 1–6. https://doi.org/10.1371/journal.pone.0039929

Martinez-Cadenas, C., Blanco-Verea, A., Hernando, B., Busby, G. B. J., Brion, M., Carracedo, A., Salas, A., & Capelli, C. (2016). The relationship between surname frequency and Y chromosome variation in Spain. *European Journal of Human Genetics*, *24*(1), 120–128. https://doi.org/10.1038/ejhg.2015.75

McCoy, A. M., Schaefer, R., Petersen, J. L., Morrell, P. L., Slamka, M. A., Mickelson, J. R., Valberg, S. J., & Mccue, M. E. (2014). Evidence of positive selection for a glycogen synthase (gys1) mutation in domestic horse populations. *Journal of Heredity*, *105*(2), 163–172. https://doi.org/10.1093/jhered/est075

Min Zhang, Wei-Feng Peng, Guang-Li Yang, Feng-Hua Lv, Ming-Jun Liu, Wen-Rong Li, Yong-Gang Liu, Jin-Quan Li, Feng Wang, Zhi-Qhiang Shen, Sheng-Guo Zhao, Eer Hehua, Nurbiy Marzanov, Maziek Murawski, Juha Kantanen, M.-H. li. (2014). Y chromosome haplotype diversity of domestic sheep (Ovis aries) in northern Eurasia. *Animal Genetics*, *45*(6), 903–907. https://doi.org/10.1111/age.12214

Navarro-López, B., Granizo-Rodríguez, E., Palencia-Madrid, L., Raffone, C., Baeta, M., & de Pancorbo, M. M. (2021). Phylogeographic review of Y chromosome haplogroups in Europe. *International Journal of Legal Medicine*, *135*(5), 1675–1684. https://doi.org/10.1007/s00414-021-02644-6

Nissen, J. (1997). *Enzyklopädie der Pferderassen*. Franckh-Kosmos Verlags-GmbH&Co.

Nissen, J. (1998). Pferdezucht. In *Enzyklopädie der Pferderassen, Bd.2* (pp. 9–23). Franckh-Kosmos Verlags-GmbH&Co.

Oetjens, M. T., Martin, A., Veeramah, K. R., & Kidd, J. M. (2018). Analysis of the canid Y-chromosome phylogeny using short-read sequencing data reveals the presence of distinct haplogroups among Neolithic European dogs. *BMC Genomics*, *19*(1), 1–9. https://doi.org/10.1186/s12864-018-4749-z

Orlando, L. (2020). The Evolutionary and Historical Foundation of the Modern Horse: Lessons from Ancient Genomics. *Annual Review of Genetics*, *54*, 563–581. https://doi.org/10.1146/annurev-genet-021920-011805

Orlando, L., & Librado, P. (2019). Origin and evolution of deleterious mutations in horses. *Genes*, *10*(9). https://doi.org/10.3390/genes10090649

Ottoni, C., Van Neer, W., De Cupere, B., Daligault, J., Guimaraes, S., Peters, J., Spassov, N., Prendergast, M. E.,

Boivin, N., Morales-Muñiz, A., Bălăşescu, A., Becker, C., Benecke, N., Boroneant, A., Buitenhuis, H., Chahoud, J., Crowther, A., Llorente, L., Manaseryan, N., … Geigl, E.-M. (2017). The palaeogenetics of cat dispersal in the ancient world. *Nature Ecology & Evolution*, *1*(7), 1–7. https://doi.org/10.1038/s41559-017-0139

Outram, A. K., Stear, N. A., Bendrey, R., Olsen, S., Kasparov, A., Zaibert, V., Thorpe, N., & Evershed, R. P. (2009). The earliest horse harnessing and milking. *Science*, *323*(5919), 1332–1335. https://doi.org/10.1126/science.1168594

Pérez-Gutiérrez, L. M., De La Peña, A., & Arana, P. (2008). Genetic analysis of the Hispano-Breton heavy horse. *Animal Genetics*, *39*(5), 506–514. https://doi.org/10.1111/j.1365-2052.2008.01762.x

Pérez-Pardal, L., Sánchez-Gracia, A., Álvarez, I., Traoré, A., Ferraz, J. B. S., Fernández, I., Costa, V., Chen, S., Tapio, M., Cantet, R. J. C., Patel, A., Meadow, R. H., Marshall, F. B., Beja-Pereira, A., & Goyache, F. (2018). Legacies of domestication, trade and herder mobility shape extant male zebu cattle diversity in South Asia and Africa. *Scientific Reports*, *8*(1), 1–8. https://doi.org/10.1038/s41598-018-36444-7

Petersen, J. L., Mickelson, J. R., Cleary, K. D., & McCue, M. E. (2014). The american quarter horse: Population structure and relationship to the thoroughbred. *Journal of Heredity*, *105*(2), 148–162. https://doi.org/10.1093/jhered/est079

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

Petersen, J. L., Mickelson, J. R., Rendahl, A. K., Valberg, S. J., Andersson, L. S., Axelsson, J., Bailey, E., Bannasch, D., Binns, M. M., Borges, A. S., Brama, P., da Câmara Machado, A., Capomaccio, S., Cappelli, K., Cothran, E. G., Distl, O., Fox-Clipsham, L., Graves, K. T., Guérin, G., … McCue, M. E. (2013). Genome-Wide Analysis Reveals Selection for Important Traits in Domestic Horse Breeds. *PLoS Genetics*, *9*(1). https://doi.org/10.1371/journal.pgen.1003211

Poznik, G. D., Henn, B. M., Yee, M., Sliwerska, E., Ghia, M., Lin, A. A., Snyder, M., Quintana-murci, L., Kidd, J. M., Underhill, P. A., & Bustamante, C. D. (2013). Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males versus Females. *Science*, *341*(6145), 562–565. https://doi.org/10.1126/science.1237619

Poznik, G. D., Xue, Y., Mendez, F. L., Willems, T. F., Sayres, M. A. W., Ayub, Q., Mccarthy, S. A., Kashin, S., Chen, Y., Banerjee, R., Rodriguez-flores, J. L., Cerezo, M., Shao, H., Gymrek, M., Malhotra, A., Louzada, S., Desalle, R., Ritchie, G. R. S., Cerveira, E., … Program, S. B. (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nature Genetics*, *48*(6), 593–599.

https://doi.org/10.1038/ng.3559

Pruvost, M., Bellone, R., Benecke, N., Sandoval-Castellanos, E., Cieslak, M., Kuznetsova, T., Morales-Munïz, A., O'Connor, T., Reissmann, M., Hofreiter, M., & Ludwig, A. (2011). Genotypes of predomestic horses match phenotypes painted in Paleolithic works of cave art. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(46), 18626–18630. https://doi.org/10.1073/pnas.1108982108

Radovic, L., Remer, V., Reiter, S., Bozlak, E., Felkel, S., Grilz-Seger, G., Brem, G., & Wallner, B. (2021). 38 Y chromosome genetic variation and deep genealogies provide new insights on Lipizzan sire lines. *Journal of Equine Veterinary Science*, *100*(May), 103501. https://doi.org/10.1016/j.jevs.2021.103501

Ramírez, O., Ojeda, A., Tomàs, A., Gallardo, D., Huang, L. S., Folch, J. M., Clop, A., Sánchez, A., Badaoui, B., Hanotte, O., Galman-Omitogun, O., Makuza, S. M., Soto, H., Cadillo, J., Kelly, L., Cho, I. C., Yeghoyan, S., Pérez-Enciso, M., & Amills, M. (2009). Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Molecular Biology and Evolution*, *26*(9), 2061–2072. https://doi.org/10.1093/molbev/msp118

Raudsepp, T., Finno, C. J., Bellone, R. R., & Petersen, J. L. (2019). Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the post-genome era. *Animal Genetics*, *50*(6), 569–597. https://doi.org/10.1111/age.12857

Raudsepp, T., Santani, A., Wallner, B., Kata, S. R., Ren, C., Zhang, H. Bin, Womack, J. E., Skow, L. C., & Chowdhary, B. P. (2004). A detailed physical map of the horse Y chromosome. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(25), 9321–9326. https://doi.org/10.1073/pnas.0403011101

Remer, V., Bozlak, E., Felkel, S., Radovic, L., Rigler, D., Grilz-eger, G., Stefaniuk-szmukier, M., Bugno-poniewierska, M., Brooks, S., Miller, D. C., Antczak, D. F., Sadeghi, R., Cothran, G., Juras, R., & Khanshour, A. M. (2022). *Y-Chromosomal Insights into Breeding History and Sire Line Genealogies of Arabian Horses*. *1680*.

Royo, L. J., Álvarez, I., Beja-Pereira, A., Molina, A., Fernández, I., Jordana, J., Gómez, E., Gutiérrez, J. P., & Goyache, F. (2005). The origins of Iberian horses assessed via mitochondrial DNA. *Journal of Heredity*, *96*(6), 663–669. https://doi.org/10.1093/jhered/esi116

Savolainen, P., Zhang, Y. ping, Luo, J., Lundeberg, J., & Leitner, T. (2002). Genetic evidence for an East Asian origin of domestic dogs. *Science*, *298*(5598), 1610–1613. https://doi.org/10.1126/science.1073906

Scharnhölz, R. (2017). Die Zukunft der Kaltblutpferdezucht. In *Kaltblutpferde im Spiegel der Wissenschaft* (pp. 697–706). Austrian Academy of Science.

Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., Albrechtsen, A., Dupanloup, I., Foucal, A., Petersen, B., Fumagalli, M., Raghavan, M., Seguin-Orlando, A., Korneliussen, T. S., Velazquez, A.

M. V., Stenderup, J., Hoover, C. A., Rubin, C. J., Alfarhan, A. H., … Orlando, L. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(52), E5661–E5669. https://doi.org/10.1073/pnas.1416991111

Schweisgut, O. (1995). *Haflinger-Pferde. Ursprung, Zucht und Haltung, weltweite Verbreitung.* BLV.

Skaletsky, H., Kuroda-Kawaguchl, T., Minx, P. J., Cordum, H. S., Hlllier, L. D., Brown, L. G., Repplng, S., Pyntikova, T., All, J., Blerl, T., Chinwalla, A., Delehaunty, A., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S. F., Latrielle, P., … Page, D. C. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, *423*(6942), 825–837. https://doi.org/10.1038/nature01722

Tao, R., Zhang, J., Sheng, X., Zhang, J., Yang, Z., Chen, C., Bian, Y., Liu, X., Zhang, S., & Li, C. (2019). Development and validation of a multiplex insertion/deletion marker panel, SifaInDel 45plex system. *Forensic Science International: Genetics*, *41*(April), 128–136. https://doi.org/10.1016/j.fsigen.2019.04.008

Tatjana Kavar, Gottfried Brem, Franc Habe, Johann Sölkner, P. D. (2002). History of Lipizzan horse maternal lines as revealed by mtDNA analysis. *Genetics, Selection, Evolution : GSE*, *34*, 635–648. https://doi.org/10.1051/gse

Teitz, L. S., Pyntikova, T., Skaletsky, H., & Page, D. C. (2018). Selection Has Countered High Mutability to Preserve the Ancestral Copy Number of Y Chromosome Amplicons in Diverse Human Lineages. *American Journal of Human Genetics*, *103*(2), 261–275. https://doi.org/10.1016/j.ajhg.2018.07.007

Thomas Jansen, Peter Forster, Marsha A. Levine, Hardy Oelke, Matthew Hurles, Colin Renfrew, Jürgen Weber, and K. O. (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). https://doi.org/10905-10910

Tomaszkiewicz, M., Medvedev, P., & Makova, K. D. (2017). Y and W Chromosome Assemblies: Approaches and Discoveries. *Trends in Genetics*, *33*(4), 266–282. https://doi.org/10.1016/j.tig.2017.01.008

Tyler-Smith, M. A. J. and C. (1995). The Human Y Chromosome: An Evolutionary Marker Comes of Ages. *Trends in Genetics*, *11*(11), 449–456.

Underhill, P. A., & Kivisild, T. (2007). Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics*, *41*, 539–564. https://doi.org/10.1146/annurev.genet.41.110306.130407

Vidal, O., Drögemüller, C., Obexer-Ruff, G., Reber, I., Jordana, J., Martínez, A., Bâlteanu, V. A., Delgado, J. V., Eghbalsaied, S., Landi, V., Goyache, F., Traoré, A., Pazzola, M., Vacca, G. M., Badaoui, B., Pilla, F., D'Andrea, M., Álvarez, I., Capote, J., … Amills, M. (2017). Differential distribution of Y-chromosome haplotypes in Swiss and Southern European goat breeds. *Scientific Reports*, *7*(1), 1–6. https://doi.org/10.1038/s41598-017-15593-1

Vilà, C., Leonard, J. A., Götherström, A., Marklund, S., Sandberg, K., Lidén, K., Wayne, R. K., & Ellegren, H. (2001). Widespread origins of domestic horse lineages. *Science*, *291*(5503), 474–477. https://doi.org/10.1126/science.291.5503.474

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T. L., Adelson, D. L., Bailey, E., Bellone, R. R., Blöcker, H., Distl, O., Edgar, R. C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J. N., Penedo, M. C. T., Raison, J. M., … Lindblad-Toh, K. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, *326*(5954), 865–867. https://doi.org/10.1126/science.1178158

WAHO. (n.d.). *About the World Arabian Horse Organization*. Retrieved December 29, 2021, from http://www.waho.org/

Wallner, B., Piumi, F., Brem, G., Müller, M., & Achmann, R. (2004). Isolation of Y Chromosome-specific Microsatellites in the Horse and Cross-species Amplification in the Genus Equus. *Journal of Heredity*, *95*(2), 158–164. https://doi.org/10.1093/jhered/esh020

Wallner, Barbara, Palmieri, N., Vogl, C., Rigler, D., Bozlak, E., Druml, T., Jagannathan, V., Leeb, T., Fries, R., Tetens, J., Thaller, G., Metzger, J., Distl, O., Lindgren, G., Rubin, C. J., Andersson, L., Schaefer, R., McCue, M., Neuditschko, M., … Brem, G. (2017). Y Chromosome Uncovers the Recent Oriental Origin of Modern Stallions. *Current Biology*, *27*(13), 2029-2035.e5. https://doi.org/10.1016/j.cub.2017.05.086

Wallner, Barbara, Vogl, C., Shukla, P., Burgstaller, J. P., Druml, T., & Brem, G. (2013). Identification of Genetic Variation on the Horse Y Chromosome and the Tracing of Male Founder Lineages in Modern Breeds. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0060015

Wang, C., Li, H., Guo, Y., Huang, J., Sun, Y., Min, J., Wang, J., Fang, X., Zhao, Z., Wang, S., Zhang, Y., Liu, Q., Jiang, Q., Wang, X., Guo, Y., Yang, C., Wang, Y., Tian, F., Zhuang, G., … Zhong, J. (2020). Donkey genomes provide new insights into domestication and selection for coat color. *Nature Communications*, *11*(1), 1–15. https://doi.org/10.1038/s41467-020-19813-7

Wang, S., Xia, X., Zhang, T., Sun, T., Dang, R., Huang, Y., Lan, X., Chen, H., & Lei, C. (2018). Y-chromosome haplotype analysis revealing multiple paternal origins in swamp buffaloes of China and Southeast Asia. *Journal of Animal Breeding and Genetics*, *135*(6), 442–449. https://doi.org/10.1111/jbg.12364

Warmuth, V., Eriksson, A., Bower, M. A., Barker, G., Barrett, E., Hanks, B. K., Li, S., Lomitashvili, D., Ochir-Goryaeva, M., Sizonovg, G. V., Soyonov, V., & Manica, A. (2012). Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(21), 8202–8206. https://doi.org/10.1073/pnas.1111122109

Warmuth, V. M., Campana, M. G., Eriksson, A., Bower, M., Barker, G., & Manica, A. (2013). Ancient trade routes

shaped the genetic structure of horses in eastern Eurasia. *Molecular Ecology*, *22*(21), 5340–5351. https://doi.org/10.1111/mec.12491

Waters, P. D., Wallis, M. C., & Graves, J. A. M. (2007). Mammalian sex-Origin and evolution of the Y chromosome and SRY. *Seminars in Cell and Developmental Biology*, *18*(3), 389–400. https://doi.org/10.1016/j.semcdb.2007.02.007

Weatherbys. (1791). *History of the General Stud book*. https://www.weatherbys.co.uk/general-stud-book/bloodstock-studbook/history-of-the-general-stud-book

Weatherbys. (2021). *Thoroughbred Horse Breed*. https://www.weatherbys.co.uk/horses-racing/racehorse-owners/bespoke-bloodstock-publications

Wei, W., Ayub, Q., Chen, Y., McCarthy, S., Hou, Y., Carbone, I., Xue, Y., & Tyler-Smith, C. (2013). A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Research*, *23*(2), 388–395. https://doi.org/10.1101/gr.143198.112

Whiting, R., & Coyle, H. M. (2020). Haplotype analysis for Irish ancestry. *Forensic Sciences Research*, *5*(4), 286–291. https://doi.org/10.1080/20961790.2019.1639881

Wutke, S., Benecke, N., Sandoval-Castellanos, E., Döhle, H. J., Friederich, S., Gonzalez, J., Hallsson, J. H., Hofreiter, M., Lõugas, L., Magnell, O., Morales-Muniz, A., Orlando, L., Pálsdóttir, A. H., Reissmann, M., Ruttkay, M., Trinks, A., & Ludwig, A. (2016). Spotted phenotypes in horses lost attractiveness in the Middle Ages. *Scientific Reports*, *6*(November), 1–9. https://doi.org/10.1038/srep38548

Zuckerkandl, E. (1987). On the Molecular Evolutionary Clock. *Journal of Molecular Evolution*, *26*, 34–46. https://doi.org/https://doi.org/10.1007/BF02111280

IX. LISTS OF TABLES AND FIGURES

X. SUPPLEMENTS

Table 3: Webpages for pedigree reconstruction. Given are the web addresses of exploited databases used for the pedigree reconstruction of sampled horses. All were last accessed in May 2020.

numbering of wegpages	web addresses
1	baza.pzhk.pl
2	cpbsonline.cloudapp.net
3	pk.nhkladruby.cz
4	www.ahsa.asn.au
5	www.allbreedpedigree.com
6	www.alterreal.pt
7	www.ancce.es
8	www.bazakoni.pl
9	www.bloodlines.net
10	www.cavalo-lusitano.com
11	www.clydesdalehorse.com.au
12	www.criollo-horse.com
13	www.dartmoorponysociety.com
14	www.denjydskehest.dk
15	www.ecaho.org/
16	www.fellponysociety.org.uk
17	www.frederiksborger.com
18	www.galopp-sieger.de
19	www.haflingerhorse.com
20	www.haststam.se
21	www.haststam.se
22	www.imh.org
23	www.janow.arabians.pl
24	www.kfps.nl
25	www.knab.dk
26	www.lfl.bayern.de/lvfz/schwaiganger
27	www.pedigreequery.com
28	www.percheron.org.uk
29	www.pferdezucht-austria.at
30	www.rimondo.com
31	www.shagya-database.ch
32	www.shirestudbook.com
33	www.sporthorse-data.com
34	www.stavropol-teke.ru
35	www.suffolkhorsesociety.org.uk
36	www.sukuposti.net
37	www.tbheritage.com
38	www.vzap.org
39	www.waho.org
40	www.wpcs.uk.com
41	www.ykk.gov.tr/Home/Index
42	www.zsaa.org

Table 4: Detailed dataset for extended MSY haplotype analysis. Listed are the nine breed groups that comprise 62 breeds. In total 508 male horses were analyzed of which 311 had pedigree information available and 66 had NGS data available. The number of sampled male horses per breed, as well as the number of those with pedigree availability is given. CB = Coldblood; TB = English Thoroughbred.

Breed Groups	enclosed breeds	N samples	N pedigree available	N NGS data available
Akhal Teke				
	Akhal Teke	12	-	3
Baroque Type breeds				
	Fredriksborg Horse	8	8	-
	Friesian	10	10	-
	Kladruber	15	15	2
	Knabstrupper	6	5	-
	Lipizzaner	34	34	7
British Pony breeds				
	Connemara Pony	7	7	-
	Dartmoor Pony	5	5	-
	Exmoor Pony	4	3	-
	Fell Pony	9	9	-
	New Forest Pony	1	1	-
	Welsh A-D	15	15	-
Coldblood breeds				
	Ardennes	5	2	-
	Belgian draft	1	-	-
	Clydesdale	8	7	-
	Franches Montagnes	9	9	8
	Gypsy Cob	10	-	-
	Haflinger	19	19	6
	Hungarian CB	1	-	-
	Jutland Horse	7	7	-
	Murgese	1	1	-
	Noriker	32	32	5
	Percheron	7	0	
	Rhenish German CB	10	9	1
	Saxony Thuringian CB	1	1	
	Schleswig CB	3	3	
	Shire Horse	8	2	
	Southgerman CB	7	7	1
	Suffolk Punch	2	2	
North African breeds				
	Arab Barb	22	3	1
	Barb	47	1	5
	Mogod Pony	3	-	-
New World breeds				
	Baca Chica	1	-	-
	California Vaquero	1	-	-
	Chilean Horse	8	-	-
	Chileno	6	-	-
	Columbian Paso Horse	5	-	-
	Criollo	6	-	-
	Criollo Definitivo	1	1	-
	Fort Polk Feral Horse	10	-	-
	Galiceno	2	-	-
	Mangalarga Marchador	34	29	1
	Mustang	3	-	-
	Paso Fino	2	-	-
	Paso Peruano	6	6	-
	Santra Cruz Island Horse	3	-	-

	Wilbur-Cruce Horse	3	-	-
Iberian				
	Lusitano	15	15	-
	Pura Raza Espanola	5	5	-
	Cartujano	2	2	-
	Sorraia Horse	9	-	1
	Terceira Pony	10	-	-
Warmblood and Riding pony breeds				
(paternal ancestry uncertain)				
	Appaloosa	2	1	-
	Cleveland Bay	3	3	-
	Morgan Horse	6	-	6
	Orlov Trotter	10	10	-
	Quarter Horse	6	6	1
Others				
	Arabian or Arabian ancestry	7	6	8
	Chakouyi Horse	1	-	1
	Duelmen Pony	1	-	1
	Marwari horse	1	-	1
	Thoroughbred or TB ancestry	10	10	7
Total	**62**	**508**	**311**	**66**

Table 5: MSY variants applied in the genotyping approach. Listed are the 163 variants selected for genotyping with the HTs, respectively HGs, they determine. Key variants (in total 103) are listed in bold. Variant IDs are consistent with previous studies and detailed information on variants is published in Remer et al. 2022.

numbering	Variant ID	determined HG or HT	selected as key variant 1= yes 0 = no
1	**rW**	A	1
2	fAAV	A	0
3	fXK	A	0
4	**rAF**	Ad	1
5	**rDT**	Ad-b	1
6	fUC	Ad-b	0
7	rDS	Ad-b	0
8	**fSA**	Ad-bA	1
9	**fXY**	Ad-bA1	1
10	**qH**	Ad-bN	1
11	**qDM**	Ad-bN1	1
12	qEU	Ad-bN1	0
13	qT	Ad-bN1	0
14	**rOR**	Ad-h	1
15	rAG	Ad-h	0
16	rAI	Ad-h	0
17	rAH	Ad-h	0
18	rAJ	Ad-h	0
19	**rAE**	Ad-hA	1
20	**fWP**	Ad-hA1	1
21	fAAB	Ad-hA1	0
22	**qAA**	Ad-hA1a	1
23	qCS	Ad-hA1a	0
24	**fTA**	Ad-hB	1
25	**fXA**	Ad-hC	1
26	**qEC**	Ad-hC1	1
27	**qCU**	Am	1
28	sQD	Am	0
29	sQF	Am	0
30	**qBI**	Am-a	1
31	**sQC**	Am-s	1
32	**fRQ**	Am-sA	1
33	**sQE**	Am-sA1	1
34	**rX**	Ao	1
35	fAAX	Ao	0
36	**fZC**	Ao-a	1
37	fRJ	Ao-a	0
38	**fST**	Ao-aA	1
39	fAAJ	Ao-aA	0
40	fVK	Ao-aA	0
41	fXI	Ao-aA	0
42	**fUS**	Ao-aA1	1
43	**rY**	Ao-aA1a	1

44	**fTM**	Ao-aA1a1	1
45	**rZ**	Ao-aA1a1b	1
46	**qED**	Ao-aA1a2	1
47	**qEZ**	Ao-aA1a2a	1
48	qBF	Ao-aA1a2a	0
49	**qDK**	Ao-aA1a2a1	1
50	**qFC**	Ao-aA1a2a1a	1
51	**qEM**	Ao-aA1a3	1
52	qCY	Ao-aA1a3	0
53	**rAB**	Ao-aA1a4	1
54	fWO	Ao-aA1a4	0
55	**qBR**	Ao-aA1a5	1
56	qDA	Ao-aA1a5	0
57	**fSE**	Ao-aA1b	1
58	fYK	Ao-aA1b	0
59	**qDX**	Ao-aA2	1
60	qCV	Ao-aA2	0
61	qDH	Ao-aA2	0
62	qU	Ao-aA2	0
63	**qEW**	Ao-aA3	1
64	qEE	Ao-aA3	0
65	**qDR**	Ao-aD	1
66	qCR	Ao-aD	0
67	**fABV**	Ao-aD1	1
68	**qW**	Ao-aD2	1
69	qCH	Ao-aD2	0
70	qFK	Ao-aD2	0
71	**sAN**	Ao-aM	1
72	sAP	Ao-aM	0
73	**sE**	Ao-n	1
74	**sC**	Ao-nM	1
75	**fVX**	Ao-nM1	1
76	**qEN**	Ao-nM1a	1
77	**fSI**	Ao-nM1b	1
78	fTQ	Ao-nM1b	0
79	**qBN**	Ao-nM1b1	1
80	**fZX**	Ao-nM1b2	1
81	fZY	Ao-nM1b2	0
82	**fXJ**	Ao-nM1b2a	1
83	**fYG**	Ao-nM1b3	1
84	**qCB**	Ao-nM2	1
85	**rAX**	Crown	1
86	**rAY**	Crown	1
87	**fYR**	H	1
88	fRO	H	0
89	fTO	H	0
90	**sQB**	Hc	1
91	**fSQ**	Hs	1

92	fWQ	Hs	0
93	**fTI**	Hs-a	1
94	**rAL**	Hs-aA	1
95	fAAS	Hs-aA	0
96	**qEO**	Hs-aA1	1
97	**rAM**	Hs-aB	1
98	fWL	Hs-aB	0
99	fWW	Hs-aB	0
100	rAN	Hs-aB	0
101	**fABO**	Hs-b	1
102	**fABP**	Hs-bL	1
103	**qBO**	Hs-bL1	1
104	**rA**	T1	1
105	fYV	T1	0
106	**fVZ**	T2	1
107	**sPZ**	Ta	1
108	fRL	Ta	0
109	fTY	Ta	0
110	fXT	Ta	0
111	fZK	Ta	0
112	fZW	Ta	0
113	**qGB**	Ta-a	1
114	qGC	Ta-a	0
115	**sPY**	Ta-b	1
116	**qFE**	Ta-bA	1
117	**rB**	Tb	1
118	**fAAC**	Tb1	1
119	**rC**	Tb-d	1
120	fXR	Tb-d	0
121	**rG**	Tb-dM	1
122	**fWM**	Tb-dW	1
123	**rD**	Tb-dW1	1
124	**sP**	Tb-dW2	1
125	**sM**	Tb-dW2a	1
126	**qBX**	Tb-dW3	1
127	**qCK**	Tb-dW3a	1
128	**qGH**	Tb-dW4	1
129	**fWU**	Tb-o	1
130	fXF	Tb-o	0
131	**fUJ**	Tb-oB	1
132	**fBVB**	Tb-oB1	1
133	**rQ**	Tb-oB1a	1
134	**rO**	Tb-oB1b	1
135	**fWB**	Tb-oB1c	1
136	fAAF	Tb-oB1c	0
137	**rP**	Tb-oB2	1
138	**fQI**	Tb-oB3	1
139	**fWY**	Tb-oB3a	1

140	**fRN**	Tb-oB3a1	1
141	**rJ**	Tb-oB3b	1
142	**rK**	Tb-oB3b1	1
143	**fABA**	Tb-oB3b1a	1
144	**fZP**	Tb-oB3b1b	1
145	**rL**	Tb-oB3b1c	1
146	**qFM**	Tb-oB4	1
147	qBL	Tb-oB4	0
148	qCM	Tb-oB4	0
149	qJ	Tb-oB4	0
150	**qCG**	Tb-oB4b	1
151	**rN**	Tb-oL	1
152	fWZ	Tb-oL	0
153	**fUN**	Tb-oL1	1
154	**fVO**	Tb-oL1a	1
155	**qCL**	Tb-oL1b	1
156	**qEL**	Tk	1
157	qBQ	Tk	0
158	qCJ	Tk	0
159	qDZ	Tk	0
160	qFG	Tk	0
161	qQ	Tk	0
162	**rT**	Tu	1
163	**sAL**	Tu-b	1
			Total key variants: 103

Table 6: The CH subhaplogroups (sHGs). Listed are the 45 CH sHGs comprising the detected haplotypes as well as their determining key variants with the allelic states (details given in Remer et al., 2022) in square brackets. The sHGs and haplotypes clustering to inner nodes are marked with an asterisk '*'.

numeration	sHG	HTs	key variant
1	**Ad-b***	Ad-b*	rDT[C]
		Ad-bA*	fSA[A]
2	**Ad-bA**	Ad-bA1	fXY[C]
		Ad-bN*	qH[G]
3	**Ad-bN**	Ad-bN1	qDM[T]
4	**Ad-h***	Ad-h*	rOR[A]
		Ad-hA	rAE[G]
		Ad-hA1	fWP[T]
5	**Ad-hA**	Ad-hA1a	qAA[A]
6	**Ad-hB**	Ad-hB	fTA[C]
		Ad-hC	fXA[T]
7	**Ad-hC**	Ad-hC1	qEC[A]
8	**Am***	Am*	qCU[A]
9	**Am-a**	Am-a	qBI[T]
10	**Am-s***	Am-s*	sQC[C]
		Am-sA*	fRQ[G]
11	**Am-sA**	Am-sA1	sQE[T]
12	**Ao***	Ao*	rX[T]
13	**Ao-aA***	Ao-aA*	fST[A]
		Ao-aA1a*	rY[TGTA]
		Ao-aA1a2a	qEZ[A]
14	**Ao-aA1a**	Ao-aA1a2a1	qDK[T]
15	**Ao-aA1b**	Ao-aA1b	fSE[C]
16	**Ao-aA2**	Ao-aA2	qDX[G]
17	**Ao-aA3**	Ao-aA3	qEW[C]
18	**Ao-aD1**	Ao-aD1	fABV[C]
19	**Ao-aD2**	Ao-aD2	qW[T]
20	**Ao-aM**	Ao-aM	sAN[A]
21	**Ao-n***	Ao-n*	sE[A]
22	**Ao-nM1a**	Ao-nM1a	qEN[T]
		Ao-nM1b	fSI[G]
		Ao-nM1b1	qBN[T]
		Ao-nM1b2	fZX[A]
		Ao-nM1b2a	fXJ[A]
23	**Ao-nM1b**	Ao-nM1b3	fYG[A]
24	**Ao-nM2**	Ao-nM2	qCB[T]
25	**Cr***	Cr*	rAX[C]
26	**Hc**	Hc	sQB[T]
		Hs-aA	rAL[G]
27	**Hs-aA**	Hs-aA1	qEO[G]
28	**Hs-aB**	Hs-aB	rAM[G]
29	**Hs-b***	Hs-b*	fABO[A]
		Hs-bL	fABP[A]
30	**Hs-bL**	Hs-bL1	qBO[T]
31	**T1***	T1*	rA[T]
32	**T2***	T2*	fVZ[C]
33	**Ta**	Ta	sPZ[A]
34	**Tb1***	Tb1*	fAAC[T]
		Tb-dM	rG[C]
		Tb-dW1	rD[TACT]
		Tb-dW3	qBX[T]
		Tb-dW3a	qCK[A]
35	**Tb-d**	Tb-dW4	qGH[A]
36	**Tb-oB**	Tb-oB	fUJ[C]
37	**Tb-oB1***	Tb-oB1	fBVB[GATA(15)]

38	**Tb-oB2**	Tb-oB2	rP[C]
		Tb-oB3a*	fWY[A]
39	**Tb-oB3a**	Tb-oB3a1	fRN[C]
40	**Tb-oB3b**	Tb-oB3b	rJ[G]
		Tb-oB3b1	rK[G]
41	**Tb-oB3b1**	Tb-oB3b1b	fZP[T]
		Tb-oB4	qFM[TA]
42	**Tb-oB4**	Tb-oB4b	qCG[G]
		Tb-oL	rN[T]
		Tb-oL1	fUN[A]
		Tb-oL1a	fVO[T]
43	**Tb-oL**	Tb-oL1b	qCL[A]
44	**Tk**	Tk	qEL[G]
		Tu	rT[A]
45	**Tu**	Tu-b	sAL[G]

XI. ACKNOWLEDGEMENT

I would like to express my gratitude to my mentor O. Univ. Prof. DI DDr. habil. Dr. h. c. Gottfried Brem for this research opportunity. Thank you to my supervisor, Priv.-Doz. Dr. habil. Ivica Međugorac, for providing guidance, expertise, and smart feedback. Further, I wish to express my gratitude and appreciation for Priv.-Doz. Dr. med. vet. Barbara Wallner, whose specialized knowledge, patient guidance, support, and encouragement have been invaluable for this thesis. Thanks also to Elif Bozlak, Simone Reiter, Lara Radovic, Sabine Felkel, Claus Vogl, and Doris Rigler for providing cooperation and advice throughout the project; our conversations and meetings were vital and inspiring. Thanks to all collaborators and private horse owners for their important contributions. I would also like to thank Eva Marker, who is a loving grandmother to my children; and of course my parents who set me off on the road to this thesis a long time ago. Above all, I thank my supportive and caring husband for his positive energy. Last but not least, I wish to thank my extraordinary friends and close family; I could not have done this without you, special thanks to our clan!