Abstract
Since horse’s domestication, physical and physiological capacities have been targeted for selection, leading to the notable functional adaptations to exercise and training of the actual phenotypes registered in equestrian disciplines. During the past ten years, experiments using next generation sequencing attempted to identify specific loci in the equine genome associated with traits of interest (e.g. speed, stamina, strength). Some of these investigations are reviewed in this study within a comprehensive and comparative approach.

Keywords: DNA-sequencing, genomics, exercise physiology, sports science

Introduction
The modern horse presents remarkable physical and physiological capabilities that distinguish them from the other mammals. Natural selection and more recently, artificial selection for speed and stamina provided to the modern horses unique adaptations to exercise and training. The actual phenotypes show high aerobic and anaerobic capacities (up to 200 ml/kg.min-1)2 increased muscle mass to body weight ratio3,5 and efficient locomotion,4,5 however weather these structure and functional changes are related to specific genetic variants remains unclear.

The equine genome sequencing6 provided scientific basis to future analyses of genetic variations underlying superior phenotypes. Genetic variation known as single nucleotide polymorphisms (SNPs) have been described in humans10 and recently in horses undergoing successful performance careers in athleticism.9 Variants of MSTN gene encoding myostatin were related to best race distance among elite flat racing Thoroughbred.16,11 Myostatin down regulates both myoblasts proliferation and differentiation and, therefore acts in the inhibition of muscle growth.12

Further studies reported additional candidate genes for flat racing performance in Thoroughbred (THB),1,10,12 with raising evidence for loci involved in glycolysis,5,14 oxidative phosphorylation15 and fast twitch fibers contractile function.15,16 PDK4 SNP (gene locus involved in the activity of pyruvate dehydrogenase complex) was not different between elite quarter horses (QH) and non-elite QHs engaged in short distances race (401m). However, the frequency for the A allele was higher for the QHs than Arabian horses,3 in agreement with the inverse coupling between muscle fiber type and muscle bioenergetics (e.g slow twitch fibers release less energy through glycolytic pathway).17

Regati et al.14 also observed a new polymorphism for MCT1 gene, which was found only in Arabian horses (18/53), although it was not related to increased endurance performance. MCT1 is involved in the synthesis of monocarboxylate, a Tran’s membrane lactate transporter, highly expressed in horses exposed to long-term endurance training.18 Lactate mediated muscle fatigue is widely described during intense rides of short duration, however it is unlikely a factor for prolonged exercises of moderate to low intensities.19

The aim of this article is to review scientific articles concerning the role of genetic variations in equine exercise physiology, focused on candidate genes for superior performance based on gene sequencing using next generation sequencing (NGS) techniques. A correlation with human genes and athletic performance was also attempted.

Literature review
Origin and phylogeny
Equidae is a family within the odd toed ungulate Order Perissodactyla, composed by a single genus called equus.20 The genus equus originated in North America and probably spread into the Old World through the Bering land bridge. There was also a migration route, despite its minor flux, to the South America in the Late Pliocene. The earliest fossil (Equus simplicidens), found in the North America was morphologically primitive and is likely the ancestor of all specimens of current domestic horse breeds.21

The evolutionary history of the equus started 58 million years ago, from the hyracotherium (also known as Hyrpparion or Eohippos) and unfolded through a branched phylogenetic tree.22 During the early Miocene, high rates of morphological changes occurred in the teeth of the family Equinae, underlying the shift from browsers (Parahyppus) to grazers (Merychippus).23 This new habit of pasture might be considered an essential feature for the adaptation and spread of equus population in the steppes of the Old World. The well-established trend toward increase in body size has also followed the evolutionary history of the equus from the early hyracotherium to the modern domestic horse, although the rate of increase in body mass followed a non-linear pattern. During the first half of horse evolution (57 to 25 million years ago), there was probably no change in body mass, which was followed by a remarkable diversification in body size during the early-middle Miocene.24 Furthermore, some important postural adaptations also have occurred during the Miocene in the primitive specimens of Parahippus and Merychippus. The capability of standing for long periods (23hours per day) is a features currently observed in the modern horse and according to evolutionary studies of fossil horses, it was developed during the middle Miocene (18 to 12million years ago).25
Domestication and spread of de equus

The domestication of the horse is believed to interact historically with human cultures in the Eurasian steppes. There are archeological evidences that horses’ domestication and the Kurgan Culture spread (Indo-European origins) had shared the same place of birth in the west Eurasia.\textsuperscript{26,27} In addition to the archeological findings, strong support was provided in a genomic DNA study with 322 non-breed horses, sampled in 12 different locations, in order to represent the wide variety of demographic scenarios. Authors proposed a domestication origin in the west central Eurasian (modern territories of Ukraine and Kazakhstan), which is geographically different from the origin of wild progenitor of domestic horses (\textit{E. feraus}), positioned in the east Eurasia.\textsuperscript{28} Several questions emerged from such remarkable difference: why had horses originated in a certain place and been domesticated in another? Was there a priori migration flux of humans from east to west, which started the domestication process? Has the societies of west Eurasia developed the proper knowledge of domestication and equitation?

The genetic structuring of the horse domestication was carried out with such notable particularities, now currently investigated through DNA mitochondrial (m) models. Despite the concept that reduction in genetic variability is likely to occur due to domestication,\textsuperscript{29} the evolutionary process in horses seems to take the opposite way. The modern horse shows greater DNA m diversity than a single wild horse population, such as Alaskan and Przewalski’s horse. The comparative DNA models suggest that several horse population from different geographic origins with minimal within population genetic diversity were recruited during domestication.\textsuperscript{30} When domestication spread out of central west Eurasia, high levels of introgression from wild local horse claims might justify this high genomic diversity of horses during domestication.\textsuperscript{31}

Genes targeted for selection in horses

Genes undergoing recent selection have been studied in species, which are nowadays widely distrusted and derived from a non-cosmopolitan ancestral population. In these species, reduction in population size and patterns of variability often underlies new habits of adapted phenotypes. These changes in the genetic “shape” may be investigated through population-level DNA data, in order to identify genetic basis of adaptation and selection.\textsuperscript{32}

Several approaches are now available to identify genes undergoing recent selection, whereas it requires previous knowledge of gene function and linkage mapping. The simplest method to map these genes is the identification of outlier loci, based on empirical distribution of some chosen features of the data.\textsuperscript{31} It is generally assumed that loci targeted for selection are localized in the extreme tail of an empirical distribution (genome-wide distribution from the summary statistic).\textsuperscript{29} Gu et al.\textsuperscript{13} identified 18 loci in the tail end of the distribution in a study employing hitchhiking mapping approach in several horse breeds. Seventeen loci were found on the negative extreme of the distribution (expected heterozigosity < 3.5), which undergone positive selection. The opposite extreme (positive end) shows the effect of balancing selection.\textsuperscript{32}

In order to identify genomic regions targeted for positive selection in Thoroughbred horses, several breeds were compared using FST and H0 approaches. Clusters of outlier loci with high values of FST (> 0.25) and low H0 were observed only in Thoroughbreds, which reveal evidence of population differentiation. Plotting inter-population FST and THB H0 across each chromosome of the most strongly selected loci, three loci were identified in Ewens-Watterson test: NVHEQ079 and TKY316 (p < 0.05); TKY222 (p < 0.01).

Based on H. Sapiens gene ontology (GO) database, 369 genes with functions on 35 GO Biological Process and 21 Molecular Process may distinguish THB from non-THB horses. According to the author, his main physiological pathways related to these genes were fatty acid metabolism, oxidative phosphorylation and actin cytoskeleton.\textsuperscript{35}

Further population-based studies of selection in horses using microsatellites and Di statistic (designed to detect significance that are near or at fixation within a population) supported the use of these techniques to identify novel variants and selection induced changes in equine genome.\textsuperscript{31} The comparison of three different breeds (THB, QH and Paint Horse - PH) was carried out in order to identify regions of putative selection in horses engaged in different optimal race distance. The highest di values was observed on ECA 18 for QH and PH, which presented 780.7 kb haplotypes composed of 21 SNPs identified in 50%, 91.3% and 100% of THB, QH and PH respectively.\textsuperscript{32} MSTN gene (center of haplotypes found) was sequenced in a cohort with QHs) and without the whole amount of haplotypes (THB) and authors found an SNP in intron 1 of MSTN only in QHs, which had been previously described as a predictor of sprinting ability in Thoroughbreds.\textsuperscript{36}

Heritability of aerobic capacity

Aerobic capacity is nevertheless a complex trait in athletic horses, as well as in other mammals and results from the sum of inherent capabilities and the ability to adapt to training stimulus.\textsuperscript{34} Aerobic capacity is currently estimated through the measurement of maximal oxygen uptake,\textsuperscript{35} which might reach the value of 180 mL/kg/min in trained Thoroughbred horses.\textsuperscript{36} In addition to the intrinsic capacity of an organism to transfer energy through aerobic metabolism in the untrained state, the ability to adapt to physical training is a trait believed to be influenced by certain genetic features.\textsuperscript{3} Furthermore, it is also a concept underlying the a priori hypothesis that superior exercise performance is in part, genetically determined.

Adaptation to aerobic training, heritability patterns and the response to divergent selection were investigated in a founder population of rats exposed to 8 weeks of treadmill running. Total distance covered showed a wide range of variation (from -110 to +430 meter gained) after training. Moreover, first offspring of high responders to aerobic training were able to cover 161 m more than founders, while offspring of low responders showed no difference with founders on their endurance capacity. Thereby, heritability was estimated at 0.43, which means that genetic contribution to the ability to adapt to aerobic training is about 43%.\textsuperscript{37}

Genes of interest

MSTN gene

MSTN gene, composed of three exons and two introns encodes Myostatin, a member of transforming growth factor \(\beta\) super family, which down regulates muscle hyperplasia and hypertrophy in mammals. Myostatin (previously referred to as GDF-8) expression appears to be localized to myotome compartment in embryos of mousse (10.5 days post-coitum) and is likely to be continuously
expressed in several skeletal muscles of adults. The biological function of myostatin was described in mouse through comparison of wild animals with mutant types, produced after disruption of MSTN gene by homologous targeting in embryonic stem cells. Homozygous mutant animals were about 30% larger and showed total muscle cell 86% higher than their heterozygous and wild type littermates. Natural mutations of MSTN gene correlated to increased muscle mass were also identified in cattle, sheep, dogs, fishes, humans, horses.

The very high muscle mass to body weight ratio observed in horses (55%), compared to other species, is probably a trait targeted for artificial selection, especially in Thoroughbred horses. Considering the important contribution of muscle power to flat race ability, the association between MSTN sequence variant and racing performance was investigated in a series of population-based case control study by separating thoroughbreds (n=148) on the basis of retrospective racecourse performance. Genotypes of two single nucleotides polymorphisms attempted (g.6649373737C>T and g.66494218A>C) were not more common among winning thoroughbreds than non-winning thoroughbred horses, however the comparison of horses winning short race (<8 furlongs) with horses winning long distance race (>8 furlongs) revealed significant association. The hypothesis that MSTN SNPs are correlated to best race distance in Thoroughbred horses was also raised in a study using a whole genome association approach. One hundred and eighty-nine elite thoroughbred horses (at least one win), were grouped following the optimal racing distance (<7 f, 8-10 f and >10 f), which were then correlated to sequence variation in specific locus of MSTN gene. Two markers on ECA18 were identified (BIEC2-417274 and BIEC2-417495) when sprinters (<7 f) were compared to middle-distance horses (8 - 10 f). It was noted that MSTN gene is located between the two significant SNPs at 68490208 - 66495180 bp. Sequence variants described on specific locus of MSTN gene seems to be correlated with different morphological types, which may increase the chance for superior performance, especially for sprinter Thoroughbreds, however whether these genomic markers are linked with physiological capabilities such as repeated sprint ability, peak power output and maximal accumulated oxygen deficit (MAOD) remains unknown. In humans an initial effort has been done in order to identify genetic variations of known muscle mass and muscle strength cohorts. Despite the day-to-day variations on indexes of strength, studies with athletes present significant potential to clarify subjects about muscle development in response to training of certain genotypes, due to a greater background of physical tests concerning maximal voluntary contraction. A comparative study on several mammalian species suggested that human MSTN gene have been targeted for natural selection. The five polymorphic sites are remarkably conserved in humans. This high level of conservation indicates that mutations in humans may present functional implications.

PDK4 gene

Pyruvate dehydrogenase kinase isozymes (PDKs) down regulate the activity of pyruvate dehydrogenase complex (PDC), a key enzyme responsible for glucose oxidation by catalyzing the conversion of pyruvate to acetyl-CoA. The up regulation of PDK4 isoforms might impair tricarboxylic acid cycle and then reduce potential energy output of pyruvate processing, thereby acting on the selection of energy source for cellular metabolism. There are evidences that PDK4 expression is involved in the switch from carbohydrate to fatty acid as the main energy substrate during prolonged hibernation. The role of PDK4 on fatty acid metabolism might be an interesting issue for riding horses. In a study investigating energy expenditure and respiratory quotients of Arabian horses undergoing endurance ride, it was observed that fat is likely the main source of energy for muscle metabolism. The effect of fat adaptation on bioenergetics indexes, such as lactate breakpoint have been observed in Arabian horses during endurance training and is also an additional evidence for the metabolic specialization toward ATP production from fat oxidation.

PDK4 isoform have been previously referred as a candidate gene for athletic performance in horses. Hills et al., observed that A:A and A:G genotypes (THB) showed higher performance that G:G genotypes (16.1 - 16.6 lb handicap advantage). Difference in allele frequency of PDK4 gene was also observed between Arabian and Quarter Horses suggesting that it might be involved in the control of energetic pathways during exercise in horses. In humans, PDK4 expression is believed to be regulated by hypoxia and muscle glycogen content prior to exercise.

Next-generation sequencing

NGS is a technology developed to high-throughput biological data acquisition. Sanger sequencing was the first method, which enabled the precise collection of DNA data, by sequencing through a capillary-based, semi automated method. Firstly, an amplified DNA template is produced, using shotgun de novo or target sequencing (PCR simplification). Then the amplified template is stochastically fragmented and submitted to cycles of de-naturation and reconstruction, while fluorescently labeled nucleotides (ddNTPs) are incorporated to the terminal portion of the fragments. The reactions are mediated by DNA-polymerase enzyme, primers containing known DNA sequence and free nucleotides. Thereby, the aim of reconstructing DNA templates with fluorophores is to tag each nucleotide for further identification. In sequencing assay, fragments travel into the gel filled capillary, while ddNTPs are identified according to the laser excitation response. In the last step, DNA sequence is assembled through bioinformatics methods.

Several platforms have been developed over the past 20 years based on Sanger biochemistry and led by industries and institutions such as Roche Applied ScienceTM (454 sequencing), IlluminaTM (Solexa Technology), Applied BiosystemsTM (SOLID), Harvard (Polonator), HelicosTM (Molecule Sequencer Technology). The main difference between these platforms, known as short reads NGS and the pioneer Sanger sequencing is that a sequence platform enables the collection of DNA data from millions of reaction, carried out in parallel, instead of the ordered ddNTPs reading. Short reads sequencing gather two different approaches: sequencing by ligation (SBL) and sequencing by synthesis (SBS) and both enables fast and precise DNA-sequencing.

Acknowledgements

P.B. reviewed all cited papers and wrote this article. W.R. reviewed this article.

Conflict of interest

Author declares there is no conflict of interest in publishing the article.
References


